Estadística
para Administración y Economía

JORGE DOMÍNGUEZ DOMÍNGUEZ
JORGE AXEL DOMÍNGUEZ LÓPEZ

Apoyo en la web

Alfaomega

ESTADÍSTICA
para Administración y Economía

JORGE DOMÍNGUEZ DOMÍNGUEZ
JORGE AXEL DOMÍNGUEZ LÓPEZ

Apoyo en la web

Alfaomega

Director Editorial
Marcelo Grillo Giannetto
mgrillo@alfaomega.com.mx

Jefe de Ediciones
Francisco Javier Rodríguez Cruz
jrodriguez@alfaomega.com.mx

Datos catalográficos

Domínguez Domínguez, Jorge; Domínguez López, Jorge Axel

Estadística para administración y economía

Primera Edición

Alfaomega Grupo Editor, S.A. de C.V., México

ISBN 978-607-707-971-2

Formato: 21 × 24 cm Páginas: 636

Estadística para administración y economía
Jorge Domínguez Domínguez, Jorge Axel Domínguez López

Primera edición: Alfaomega Grupo Editor, México, agosto de 2015

Pitágoras 1139, Col. Del Valle, 03100, México D.F.

Miembro de la Cámara Nacional de la Industria Editorial Mexicana
Registro No. 2317

Pág. Web: **http://www.alfaomega.com.mx**
E-mail: **atencionalcliente@alfaomega.com.mx**

ISBN: 978-607-707-971-2

Esta obra fue compuesta en latex con el compilador MiKTeX 2.8.

Empresas del grupo:

México: Alfaomega Grupo Editor, S.A. de C.V.-Pitágoras 1139, Col. Del Valle, México, D.F.-C.P. 03100.
Tel.: (52-55) 5575-5022 - Fax: (52-55) 5575-2420 / 2490. Sin costo: 01-800-020-4396
E-mail: atencionalcliente@alfaomega.com.mx

Colombia: Alfaomega Colombia S.A. - Calle 62 No. 20-46, Barrio San Luis, Bogotá, Colombia.
Tel.: (57-1) 746 0102/210 0415. E-mail: cliente@alfaomega.com.co

Chile: Alfaomega Grupo Editor, S.A. - Av. Providencia 1443. Oficina 24, Santiago, Chile.
Tel.: (56-2) 2235-4248 - Fax: (56-2) 2235-5786. E-mail: agechile@alfaomega.cl

Argentina: Alfaomega Grupo Editor Argentino, S.A. - Paraguay 1307 P.B. Of. 11, C.P. 1057, Buenos Aires, Argentina.
Tel./Fax: (54-11) 4811-0887 y 4811 7183. E-mail: ventas@alfaomegaeditor.com.ar.

Acerca de los autores

El Dr. Jorge Domínguez es Licenciando en Física y Matemáticas por el IPN, Maestro en Ciencias por la UNAM así como Doctor en Matemáticas por la Universidad Politécnica de Valencia. Es Investigador Titular A del Centro de Investigaciones en Matemáticas y sus principales líneas de investigación son: métodos de optimización multirespuesta, análisis de costos en producción, diseño de experimentos y enseñanza de la probabilidad y la estadística.

Entre sus actividades académicas cuenta con cuarenta años dedicados a la docencia, y diecisiete años en asesoría y capacitación industrial. El Dr. Jorge Domínguez es coautor de seis libros y ha publicado más de treinta artículos de investigación, ha dirigido más de cuarenta tesis de licenciatura, maestría y especialidad, y pertenece al SNI nivel I.

El Dr. Axel Domínguez es Ingeniero en Electrónica y Comunicaciones por la Universidad Iberoamericana, y es Doctor en Ciencias por la Universidad de Southampton. Ha trabajado como investigador en el Centro de Investigación en Matemáticas (CIMAT) en la línea de controles inteligentes para robots usando controles neurodifusos, así como ha desarrollado el programa de cómputo CalEst. Actualmente el Dr. Axel Domínguez es consultor en el desarrollo de tecnología, y en este rubro ha contribuido en diferentes proyectos tecnológicos en empresas multinacionales.

A Sarah, Rodrigo, Maribel, Paulina,
Jorge Xavier, Marielena, Octavio y Mara Elena

Mensaje del Editor

Una de las convicciones fundamentales de Alfaomega es que los conocimientos son esenciales en el desempeño profesional, ya que sin ellos es imposible adquirir las habilidades para competir laboralmente. El avance de la ciencia y de la tecnología hace necesario actualizar continuamente esos conocimientos, y de acuerdo con esta circunstancia Alfaomega publica obras actualizadas, con alto rigor científico y técnico, y escritas por los especialistas más destacados del área respectiva.

Consciente del alto nivel competitivo que debe de adquirir el estudiante durante su formación profesional, Alfaomega aporta un fondo editorial que se destaca por sus lineamientos pedagógicos que coadyuvan a desarrollar las competencias requeridas en cada profesión específica.

De acuerdo con esta misión, con el fin de facilitar la comprensión y apropiación del contenido de esta obra, cada capítulo inicia con el planteamiento de los objetivos del mismo y con una introducción en la que se plantean los antecedentes y una descripción de la estructura lógica de los temas expuestos, asimismo a lo largo de la exposición se presentan ejemplos desarrollados con todo detalle y cada capítulo concluye con un resumen y una serie de ejercicios propuestos.

Además de la estructura pedagógica con que está diseñado el contenido de nuestros libros, Alfaomega hace uso de los medios impresos tradicionales en combinación con las Tecnologías de la Información y las Comunicaciones (TIC) para facilitar el aprendizaje. Correspondiente a este concepto de edición, todas nuestras obras tienen su complemento en una página Web en donde el alumno y el profesor encontrarán lecturas complementarias así como programas desarrollados en relación con temas específicos de la obra.

Los libros de Alfaomega están diseñados para ser utilizados en los procesos de enseñanza aprendizaje, y pueden ser usados como textos en diversos cursos o como apoyo para reforzar el desarrollo profesional, de esta forma Alfaomega espera contribuir así a la formación y al desarrollo de profesionales exitosos para beneficio de la sociedad, y espera ser su compañera profesional en este viaje de por vida por el mundo del conocimiento.

Contenido

Prólogo

La segunda mitad del siglo pasado y el inicio del actual han sido testigos del incremento en el uso del razonamiento estadístico en todos los campos de las ciencias y las humanidades. En consecuencia, también ha sido esencial para los estudiantes de la mayoría de los programas académicos y profesionales el prepararse con los principios y técnicas básicas para el análisis estadístico. Se sabe también que, afortunadamente, la comprensión de una amplia gama de conceptos estadísticos no requiere de un conocimiento matemático avanzado, por lo que el propósito en este libro es contribuir con una propuesta que permita a sus lectores construir tanto el entendimiento de esos conceptos estadísticos, con sus bondades y aplicaciones, mientras introduce el uso de las técnicas que permiten su aplicación, en particular, en el ámbito del análisis de datos en la Administración y la Economía.

La elaboración de este libro ha sido llevada a cabo en un contexto en el que los autores cuentan con una vasta experiencia, el de la enseñanza de la Estadística en instituciones de educación de lengua castellana. Este es uno de los aspectos que pudiera parecer intrascendente, pero al revisar la obra se observa la importancia del manejo del lenguaje como estrategia didáctica que permite que un curso sea exitoso, tanto para los estudiantes como para los profesores.

Los autores también se han propuesto hacer comprender los conceptos bajo los paradigmas actuales del pensamiento y la práctica estadística. Aunque existen libros, tanto en lengua inglesa como en castellano, similares en mostrar a los lectores las explicaciones de métodos estadísticos y los razonamientos probabilísticos que hay detrás de estos, cada capítulo de este texto presenta una estructura sistemática y consistente, compuesta de: un contenido disciplinar con fundamentos teóricos y el desarrollo del modelo o procedimiento estadístico; el sustento vinculado con la teoría de Probabilidad que se requiera y las prácticas diseñadas, a partir de la experiencia en la consultoría en Estadística de los autores; un resumen que proporciona conceptos, expresiones clave del tema tratado y explicaciones sintéticas, tan apreciadas por los profesores y estudiantes cuando se requiere de un texto de referencia; un complemento didáctico basado en problemas y ejercicios que apoyan al estudiante en la consolidación de los contenidos; y una serie de ejercicios de aplicación que fomentan el desarrollo de habilidades y la competencia estadística.

Con la estructura que caracteriza este libro es de esperarse que los estudiantes difícilmente pierdan el interés en la materia, como suele suceder en libros tipo recetario, que a menudo llevan a cometer errores en el uso de las técnicas estadísticas. Aquí, antes de presentar formalmente una técnica estadística, se presenta al lector el ámbito y situaciones que lo motiven al estudio y a logro del aprendizaje de la técnica y el razonamiento correspondiente. Además, de acuerdo a la experiencia en estudios con profesores, la secuencia de motivación, razonamiento estadístico, método y ejemplo es muy efectiva en cursos introductorios.

La obtención de inferencias a partir del análisis de datos, es el tema central en Estadística, sin

embargo, esta obra hace un énfasis inicial tanto en la preparación y organización de los tipos de datos, como en el potencial descriptivo de los mismos, lo que les previene de obtener resultados sin sentido y les permite apreciar la utilidad de los modelos, las técnicas y los procedimientos estadísticos, tanto de los básicos o más simples, como de los avanzados. Un panorama general de los contenidos temáticos nos permite identificar que esta obra aborda los temas de: el análisis exploratorio de datos; el sustento teórico de la Estadística y la Probabilidad; el análisis de la variación y los conceptos del azar y aleatoriedad; el tratamiento de la inferencia estadística; el vínculo con el uso de las distribuciones de probabilidad; el uso de la estimación estadística, el poder de las pruebas de hipótesis; la modelación estadística; y una selección de temas selectos, pertinentes al ámbito de la Administración y la Economía.

Se espera que el potencial de esta obra sea descubierto y aprovechado por sus lectores, ya que en este se reúnen años de experiencia exitosa en la práctica de la enseñanza de la Estadística, la cual ha sido armónicamente compilada y cuidadosamente organizada para ofrecer un texto que apoye la compleja labor de la enseñanza y el aprendizaje de una disciplina que hoy, más que nunca, resulta vital en la formación de profesionales de las ciencias económico-administrativas, que sean críticos y capaces de entender los datos que describen el funcionamiento de las organizaciones y aportar, competente y profesionalmente, en su gestión y toma de decisiones basadas en el tratamiento científico de los datos y su información derivada.

José Paúl Carrasco Escobar
Aguascalientes, México, mayo de 2014

Introducción

Modelo del estudio de la estadística

Frecuentemente se visualiza a la estadística como la actividad de recopilación, presentación, análisis e interpretación de los datos recabados en una muestra. Sin embargo, ésta puede ser una de las etapas finales de esta materia, la cual va más allá. El concepto de estadística se refiere a la posibilidad de establecer un diálogo entre lo que se puede percibir del mundo real y lo que se observa a través de los datos. El aprender y comprender sobre el entorno de la administración y la esencia de la economía, parte de ciertas ideas de conocimientos iniciales; luego, mediante la información generada por un adecuado proceso de los datos, esas ideas darán lugar a otra reflexión, y a su vez este resultado motivará a una nueva comunicación con la realidad, lo que al final se convertirá en fuente de sabiduría.

Objetivo

El objetivo fundamentral de esta obra es ubicar al lector en el quehacer de la estadística al abordar problemas de interés en administración y economía en un ámbito general. En cada uno de los capítulos se presentan casos en los que se propone la estadística en acción, estos se enmarcan en lo que se ha llamado el mundo de la información. Con este tipo de planteamientos además de motivar al estudio de la estadística, se pretende propiciar las capacidades para aplicar el conocimiento en la práctica, así como generar conocimiento sobre el área de estudio.

Materiales auxiliares para la enseñanza y el aprendizaje

En cada capítulo se ha incorporado un complemento didáctico en el que se plantean problemas con la finalidad de proponer una serie de ideas de apoyo pedagógico para los profesores asi como para que el lector realice proyectos por sí mismo. Además de esto se completan ideas relevantes sobre el material expuesto en el capítulo, en algunos casos se presentan anécdotas e ideas sobre curiosidades de la aplicación de la estadística, se destacan situaciones de las limitaciones y abusos de la estadística, en particular en este apartado se plantea la necesidad de que el usuario desarrolle la capacidad de trabajo en equipo, que sea creativo y adquiera habilidad para identificar, plantear y resolver problemas.

Programa CalEst

Como una útil herramienta se presenta el programa **CalEst** el cual contiene una serie de desarrollos computacionales con un novedoso enfoque visual y práctico para apoyar la enseñanza y el aprendizaje de

la estadística y probabilidad. El propósito general del **CalEst** es auxiliar al usuario en la comprensión de una serie de conceptos técnicos en estas materias, dado que en la práctica de la enseñanza y aprendizaje estos conocimientos han mostrado un grado de dificultad en su entendimiento. Por ejemplo, se ilustran varias distribuciones de probabilidad con la finalidad de conocer y comprender su forma, el cálculo de probabilidades y la aplicación de éstas en varios temas tales como los procedimientos de inferencia estadística, entre otras. También se utiliza para generar y crear muchas situaciones típicas de los juegos de azar para explicar el cálculo de probabilidades. Este material le permitirá al profesor motivar una buena cantidad de conocimientos estadísticos y de probabilidad, y en esa dirección crear habilidad para buscar procesar y analizar información procedente de diversas situaciones.

Contenido

En general el contenido de la obra comprende los temas básicos y clásicos de un curso de estadística para la administración y economía, sin embargo se han propuesto una serie de modelos para hacer énfasis en el papel que desempeña la estadística para conocer diferentes problemáticas y proyectos de interés en estas áreas. De la misma manera, se proponen algunas ideas para presentar los contenidos temáticos y sugerencias para exposición de distintos conceptos. En el capítulo 1 se da un panorama general de los elementos principales de la estadística. En relación a la estadística descriptiva para una variable, se expone en los capítulos 2, 3 y el caso de dos variables se explican en la primera parte del capítulo 10. Los conceptos importantes de probabilidad para comprender la parte esencial de la estadística, se describen en el capítulo 4. Se continúa en los capítulos 5, 6 y 7 con las distribuciones de probabilidad básicas, para variables discretas, continuas y muestrales respectivamente. La parte fundamental de la inferencia estadística se trata en los capítulos 7, 8 y 9, para el caso de dos variables se muestra en la segunda parte del capítulo 10. Finalmente una parte elemental de la estadística no paramétrica se introduce en el capítulo 11.

Enfoque pedagógico

Con el fin de adquirir una mayor claridad sobre los conceptos se han propuesto algunas estrategias didácticas para explicar y presentar diferentes temas. Estos consisten en modelos que describen la relación entre la población y la muestra; también se emplean diferentes gráficas y situaciones para motivar la construcción de fórmulas, el cálculo de probabilidades, la prueba de hipótesis, la construcción del modelo lineal, el análisis de la varianza entre otros. La mayoría de los ejemplos han servido para ilustrar los temas tratados, y se ha procurado relacionarlos con situaciones cercanas al contexto de nuestro entorno, a nuestras experiencias de la vida cotidiana.

Ejercicios

Al final de cada capítulo se han elaborado una serie de ejercicios que contemplan cuatro actividades de aprendizaje en las que se busca su desarrollo integral de superación profesional del estudiante. Estas labores van desde las básicas que comprenden acciones de memorización y reproducción, las de razonamiento, las de resolución de problemas hasta las metas cognitivas. Los ejercicios propuestos son un recurso que el docente tiene para la evaluación.

Material Web

Como parte del material Web del libro se incluyen una presentación Power Point y un mapa conceptual de cada uno de los capítulos, lecturas complementarias y solución de problemas seleccionados, además de esto se cuenta con acceso en línea a módulos seleccionados del programa **CalEst**.

Plataforma de contenidos interactivos

Para tener acceso al material de la plataforma de contenidos interactivos del libro: Estadística para Administración y Economía, siga los siguientes pasos:

1. Ir a la página: http://libroweb.alfaomega.com.mx
2. Ir a la sección Catálogo y seleccionar la imagen de la portada del libro, al dar doble clic sobre ella tendrá acceso al material descargable.

NOTA: Se recomienda respaldar los archivos descargados de la página web en un soporte físico.

CalEst: Calculador Estadístico

Introducción

Con el fin de contribuir en la solución de la problemática de la enseñanza y aprendizaje de la estadística y probabilidad, se ha realizado un proyecto de desarrollo tecnológico que se denomina **CalEst**. En éste se plantean una serie de propuestas didácticas con un enfoque visual con el fin de proporcionar una mejor comprensión de conceptos básicos en estadística y probabilidad: estos consisten en la presentación mediante animación de los diferentes juegos clásicos de azar para introducir las ideas principales de la probabilidad; en otro módulo se ilustra el cálculo directo de probabilidades de diferentes modelos probabilísticos, tal que mediante una graficación animada de estos modelos se alcanza un mayor conocimiento en los conceptos de función densidad y distribución de probabilidad, en lugar del trillado uso de las tablas de probabilidad. Este procedimiento se tiene para una gama amplia de distribuciones. Con este enfoque gráfico se motiva de manera dinámica los conceptos de prueba de hipótesis. Se ha generado el esquema de prueba de hipótesis mediante la animación de un sistema de seguridad a partir de la identificación del iris. Este planteamiento permite ilustrar las hipótesis nula y alternativa, así se crean una serie de escenarios virtuales diferentes para ejemplificar otras pruebas estadísticas.

Presentación

La experiencia muestra que existen varias dificultades para transmitir y comprender diferentes conceptos en el proceso de enseñanza y aprendizaje en las áreas de estadística y probabilidad. Con el fin de contribuir en la solución de esta problemática se ha realizado un proyecto de desarrollo tecnológico denominado **CalEst**. La investigación en este trabajo se plantea en dos direcciones: en la primera se han desarrollado una serie de ideas y propuestas didácticas con el propósito de facilitar la presentación y comprensión de conceptos en las dos áreas citadas; la segunda tiene la finalidad de evaluar el impacto en el conocimiento de nociones estadísticas usando el material de este proyecto, ésta contempla la parte operativa y consiste en realizar de manera sencilla la descripción y el análisis de datos generados en un estudio. Varios de los resultados operativos se pueden visualizar utilizando el **CalEst**. El ambiente general del calculador se muestra en la figura 1.

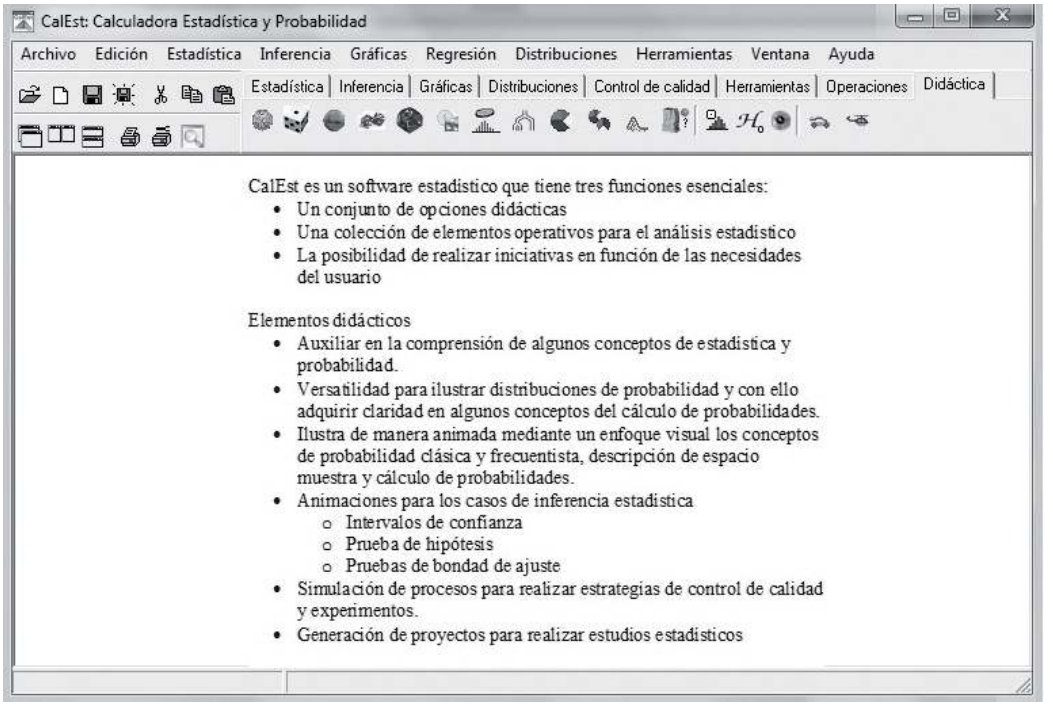


Figura 1 Ambiente del **CalEst**.

Los estudiantes necesitan tener un buen conocimiento y dominio de varios conceptos estadísticos y de probabilidad, tales como cálculo de probabilidades, distribuciones de probabilidad, pruebas de hipótesis y elaboración de modelos estadísticos. Considerando esta situación, se ha elaborado un novedoso y atractivo material didáctico con efectos animados asistidos por computadora. El material tiene la finalidad de contribuir para mejorar las actividades de enseñanza-aprendizaje en la educación superior. El desarrollo de este enfoque visual para la educación contempla la idea de aprender jugando y haciendo el trabajo por sí mismo. Las nociones de probabilidad desempeñan un papel esencial en el análisis e interpretación de los datos estadísticos, con esa finalidad en este proyecto se han creado una serie de animaciones que consideran los juegos clásicos de probabilidad tales como el lanzamiento de monedas, dados, moneda-dado, ruletas, las urnas para extraer canicas, laberintos. Una extensión a la comprensión de estos conceptos son las distribuciones de probabilidad, y se han expuesto de tal manera que se tenga una percepción clara sobre las funciones de densidad y de probabilidad acumulada. Estas animaciones permitirán que los estudiantes incrementen su interés en esta área, así como buscar nuevas ideas interactuando con el material.

El desarrollo tecnológico propuesto mediante el **CalEst** ayuda de manera animada a calcular directamente probabilidades, además de ilustrar los conceptos de función densidad y distribución de probabilidad. Esta técnica se extiende a una gama de distribuciones, en el caso de variables discretas están Bernulli, Binomial, Poisson. En el caso variables continuas se tienen la Normal, t-Student, Ji cuadrada,

la F, Weibull, Gama Exponencial, Uniforme, Beta, Lognormal, Logística y Valores Extremos. Estas se describen en la figura 2.

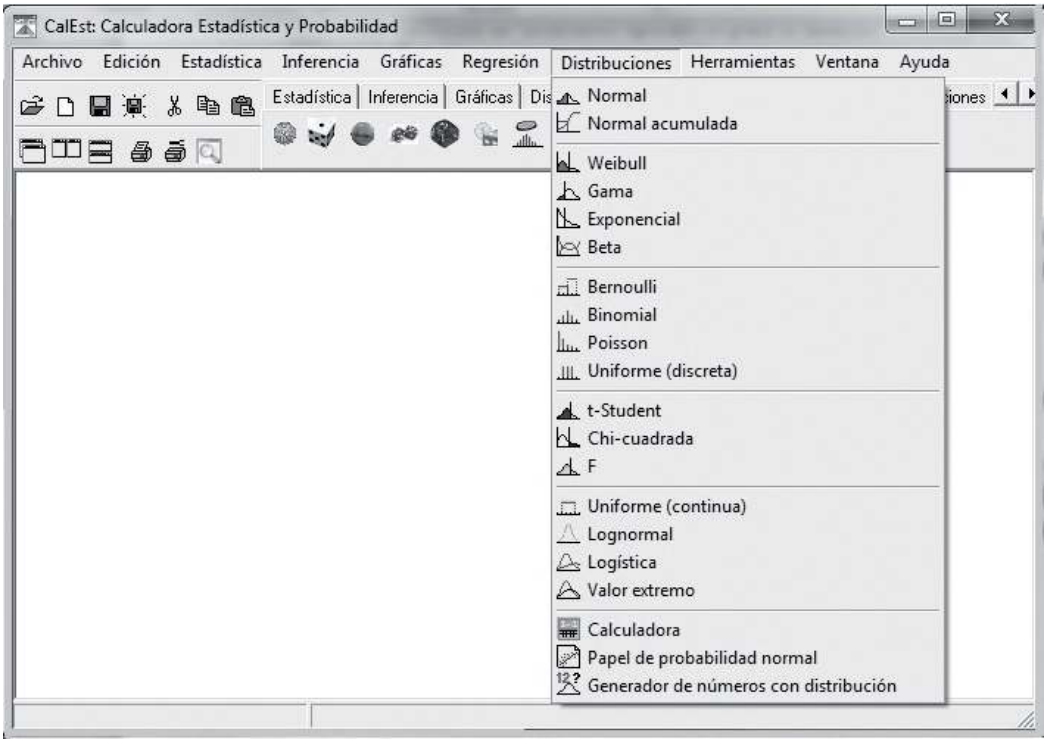


Figura 2 Descripción de las distribuciones de probabilidad.

La inferencia estadística, es el proceso de elaborar conclusiones razonables acerca de los valores poblacionales a partir del conocimiento de la información que proporcionan los datos de una muestra. Se ilustra este concepto mediante el planteamiento gráfico de las hipótesis nula y alternativa, en éste se crean una serie de escenarios virtuales diferentes para ejemplificar otras pruebas estadísticas. Con el objetivo de familiarizar al estudiante con la inferencia estadística, se ha elaborado una animación de un sistema de seguridad mediante la identificación del iris.

La integración de éste desarrollo tecnológico asistido de manera visual y de las estrategias para producir la información, crean una dinámica que proporcionan una mayor facilidad en la comprensión de conceptos y motivan el aprendizaje de la estadística y probabilidad. Con el fin de mostrar esta situación, a continuación se exponen sólo tres dispositivos didácticos que han sido desarrollados en el proyecto **CalEst** para ilustrar y motivar diferentes conceptos en la estadística y probabilidad.

Material Didáctico

Partimos del hecho de que la disciplina de la estadística es el proceso de descubrir más sobre el mundo real mediante la colección, análisis e interpretación de datos. En esa dirección los estudios en estadística

se diseñan como un procedimiento de búsqueda, en el que se plantea un problema y a partir de ahí se derivan una serie de cuestiones, las cuales se responderán y explicarán con una apropiada recolección y análisis de datos. Sin embargo, en la práctica es común dar datos para que los estudiantes hagan cálculos, por lo general no se hacen interpretaciones de los resultados. Por otro lado, el estudio de la estadística se fundamenta en conceptos de la teoría de probabilidad, por lo general, en la práctica existe una cierta dificultad para enseñar y aprender nociones sobre estos temas.

Muestra del Material Didáctico 1: Cálculo de Probabilidades

El **CalEst** cuenta con 10 opciones para calcular probabilidades de eventos y aplicar las reglas para calcular la probabilidad de eventos compuestos; véase la figura 3.

Figura 3 Descripción del material didáctico con respecto a juegos de probabilidad.

Aquí únicamente se presenta la opción de las ruletas, en la figura 4 se muestra la oportunidad de dividir dos ruletas cada una con un número diferente de grupos, en particular se consideraron hasta cuatro. Esta posibilidad crea diferentes escenarios usando los colores de las ruletas o los números y letras. A partir de ellos se generan diferentes espacios muestra para calcular probabilidades, un ejemplo es el siguiente:

1. ¿Cuál es la probabilidad de que ambas ruletas tengan el mismo color? Se tiene que se repiten 4 colores de 16 resultados posibles entonces.

2. ¿Cuál es la probabilidad de que al menos una de las ruletas sea azul? En el espacio muestra se

cuenta en el número de renglones que tienen al menos un azul, de estos hay 7.

3. Considere los números y las letras en la primer y segunda ruleta respectivamente, ahora se definen los eventos E como los números E={1,2,3,4}, y el F como las letras F={A,B,C,D}. ¿Cuál es la probabilidad de que al girar ambas ruletas la flecha marque un número par y una consonante? R: rojo y A: amarillo corresponden a los números pares y R: rojo, V: verde y A: amarillo se asocian a las consonantes, de esa manera se tiene H={(R,V), (A,A), (R,A), (A,R), (R,R), (A,V)} de manera equivalente el evento H se escribe por H={(2,C), (4,D), (2,D), (4,B), (2,B), (4,C)}. Entonces la probabilidad es 0.375, o sea P(H)=0.375

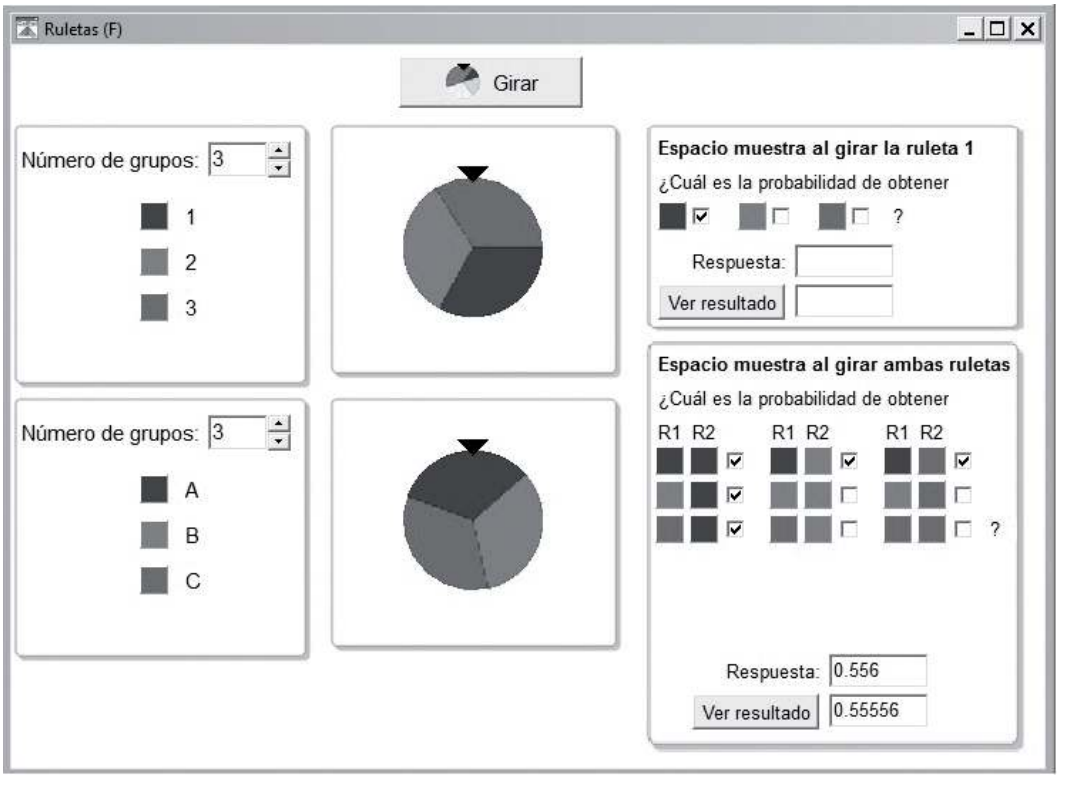


Figura 4 Cálculo de probabilidades utilizando las ruletas.

Muestra del Material Didáctico 2: Distribución de Probabilidad

Mediante un ejemplo se ilustra la sencillez para calcular probabilidades de una función de densidad usando el material didáctico desarrollado. Consideremos un estudio sobre el cerebro, en éste se desea conocer el tiempo de respuesta de una persona ante un estímulo visual. Para fijar ideas, se supone que la variable tiene una distribución de probabilidad normal con media y desviación estándar. En este caso se desea conocer el porcentaje de personas que tardan menos de 30 segundos en responder, o estimar el porcentaje de individuos en responder entre 70 y 180 segundos. Esta situación se plantea en términos de

probabilidad, para el primer caso y en el otro caso. A partir de este planteamiento se generan una serie de cuestiones tales como la de comprender el concepto de probabilidad, el de variable aleatoria, función de densidad, distribución de probabilidad.

Para abordar este tipo de problemas en cursos básicos de estadística y probabilidad se supone la distribución normal. Lo primero que se le explica al estudiante, antes de hacer el cálculo de estas probabilidades dados los parámetros para la media y desviación estándar de esta distribución, es el proceso de estandarización. Situación que distrae de la finalidad principal, toma tiempo y de alguna manera complica al estudiante. Además de las dificultades que genera al alumno comprender estas funciones de probabilidad y distinguir la diferencia entre ellas. Puesto que el objetivo principal es mostrar un efecto visual de conceptos, en esta presentación será difícil ilustrar el potencial del proyecto, así que sólo se describirán mediante figuras los resultados del ejemplo descrito anteriormente. En la figura 5 se ilustra el cálculo de probabilidad descritos por gráficas de la densidad y distribución de probabilidad normal.

Se pueden calcular diferentes probabilidades mediante el movimiento de los umbrales, cambiando los valores o usando el ratón. Es importante resaltar que aquí se calculan las probabilidades de la normal declarando el valor de la media y desviación estándar, sin necesidad de estandarizar. Esta última queda implícita al escribir. En la parte superior de la pantalla se tiene un calculador de la normal, el cual permite obtener diferentes probabilidades dados los valores de la variables aleatoria, la media y la desviación estándar. También aplica la inversa, es decir, dada una probabilidad se obtiene el valor de la variable aleatoria.

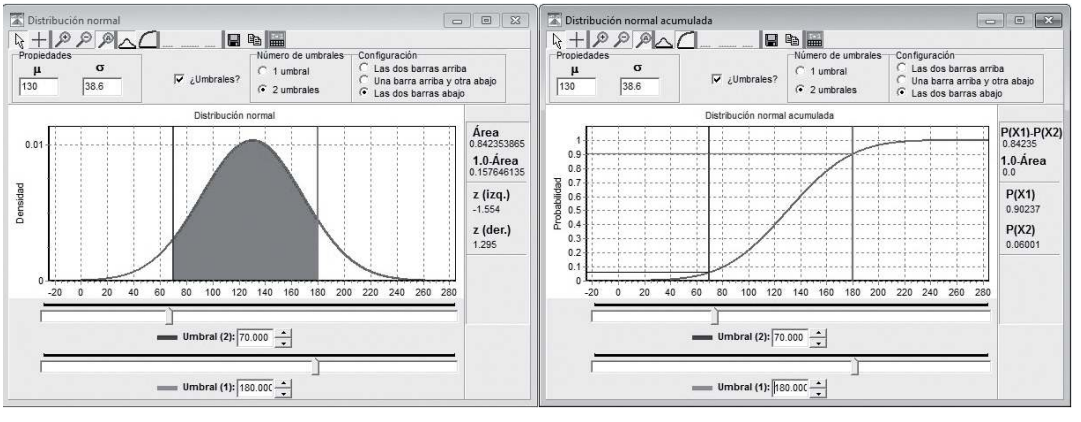


Figura 5 Efecto visual para el cálculo de probabilidades.

El desarrollo tecnológico propuesto mediante el **CalEst**, ayuda de manera animada a calcular directamente esas probabilidades, además de ilustrar los conceptos de función densidad y distribución de probabilidad. Esta técnica se extiende al conjunto de distribuciones mencionadas en el primer apartado.

Muestra Material Didáctico 3: Prueba de Hipótesis

En el ejemplo descrito también se tiene el interés en verificar la hipótesis de que las personas no reaccionan de manera lenta, en términos estadísticos la media debe ser menor a 130 segundos. Descripción gráfica

del procedimiento, figura 6.

El planteamiento estadístico formal es: la verificación de esta prueba estadística se fundamenta en conceptos de la teoría de probabilidad. Así es necesario comprender lo que significa la probabilidad de rechazar la hipótesis, cuando ésta es verdadera, o la probabilidad de no rechazarla cuando ésta es falsa, así como la potencia de la prueba.

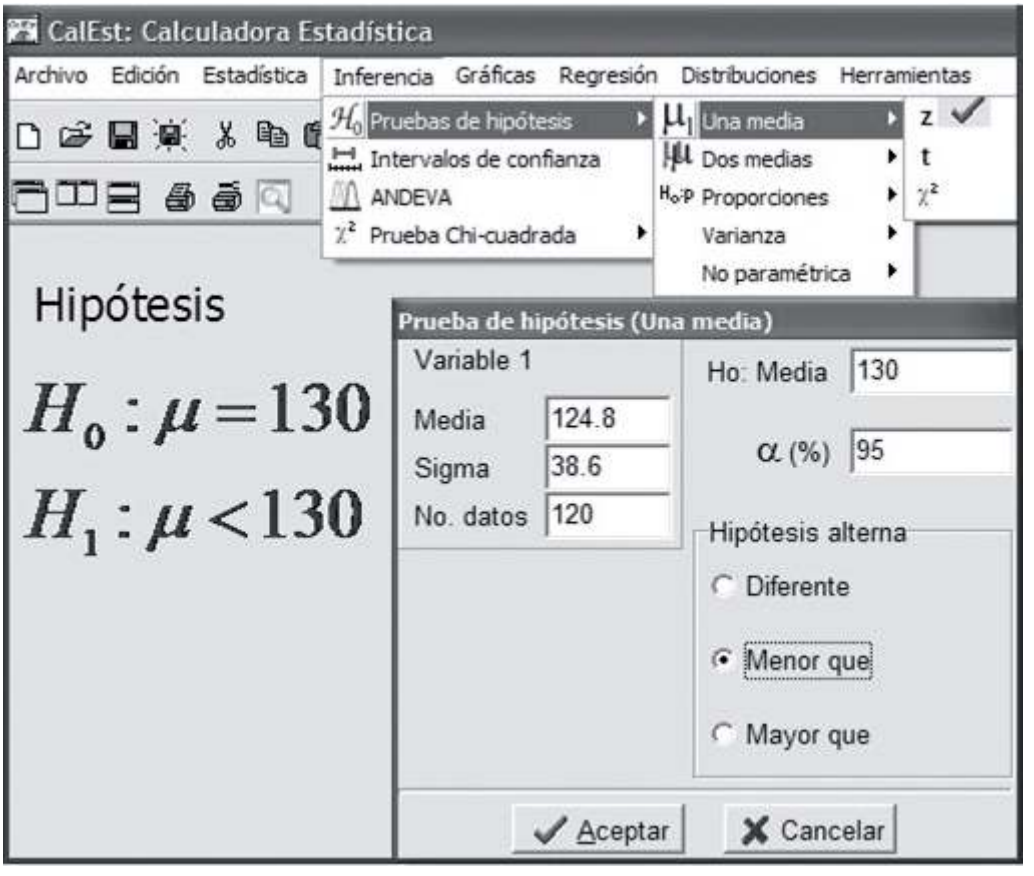


Figura 6 Procedimiento operativo para la prueba de hipótesis.

Estos conceptos se muestran de manera animada mediante una gráfica y se pueden simular diferentes escenarios del problema aplicando el **CalEst**. Esta temática cae en la parte de inferencia estadística, un resultado en el que se sustenta esta teoría es el conocido teorema de límite central, éste se ilustra de manera visual con el fin de entender cómo al variar el tamaño de muestra la distribución de probabilidad del estadístico tiende a ser simétrica. De igual forma se ilustra el concepto de prueba de hipótesis por medio de la animación de un sistema de seguridad mediante la identificación del iris. El segundo caso que se presenta corresponde a la prueba de hipótesis, como se muestra en la figura 7. En referencia a la expresión indicada en el ejemplo inicial, la gráfica muestra el caso para una posible hipótesis alterna, en este caso se planteó una de ellas: aquí se pueden usar diferentes escenarios para ver las probabilidades del

error tipo I y el error tipo II en el contexto de esta temática. Se observa que a partir de esta se consigue motivar los conceptos de prueba de hipótesis ante distintos problemas. En este caso también se anexa el calculador para la distribución normal.

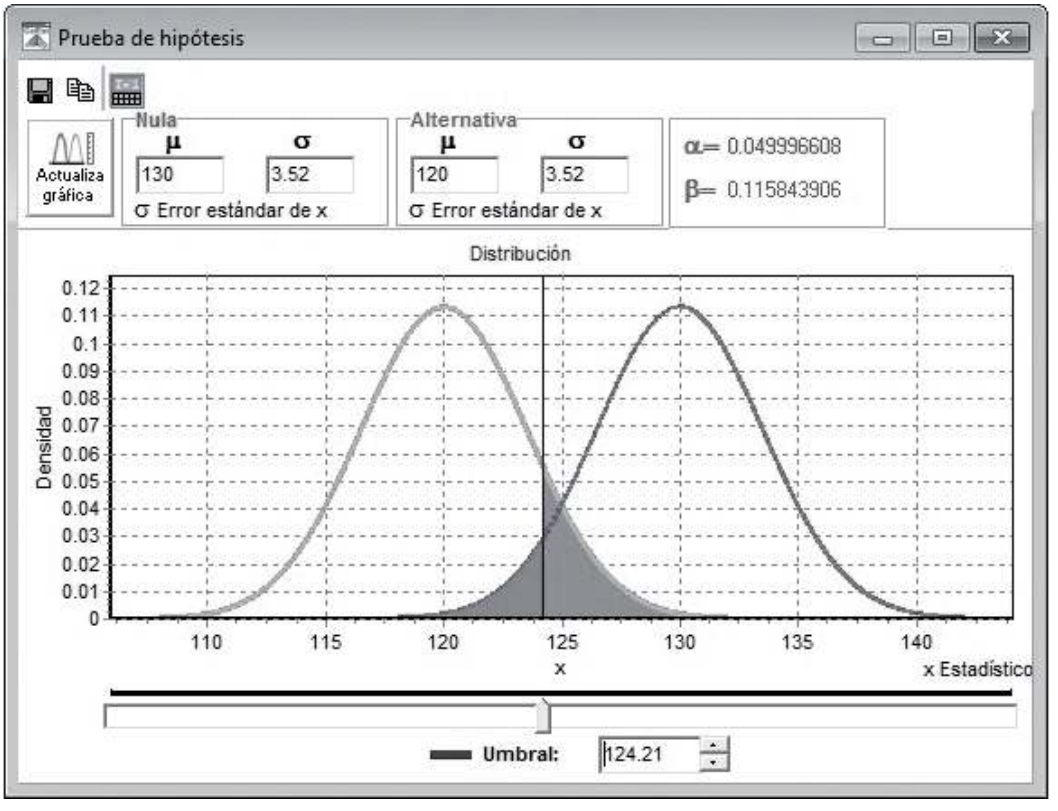


Figura 7 Animación del procedimiento para la prueba de hipótesis.

Material operativo

Finalmente, una vez que se han obtenido los datos, el **CalEst** cuenta con una variedad de herramientas para el cálculo, análisis, manipulación de datos, gráficas y está ilustrado con una guía didáctica. El trabajo que se realiza en esta parte cubre los temas de cursos a nivel bachillerato y licenciatura en varias carreras. Tales como: medidas de tendencia central, dispersión y posición, histograma, polígono de frecuencia, distribución empírica, diagrama de pastel, diagrama de puntos, diagrama de tallo y hoja, diagrama de caja, gráfica de dispersión, papeles de probabilidad, pruebas de hipótesis para 1 o 2 medias, proporciones, para la varianza, no paramétricas, intervalos de confianza, análisis de varianza, regresiones, pruebas de bondad de ajuste. También cuenta con varias opciones que son de utilidad para el reporte de los resultados gráficos y analíticos, por ejemplo grabar en archivos, imprimir, modificar los escenarios en la pantalla, así como la importación y exportación de datos.

Conclusiones

Esta propuesta usa como recurso tecnológico la animación asistida por computadora y se presenta como un paquete. Así este desarrollo resulta novedoso, visualmente atractivo y es una herramienta complementaria que beneficia en el aprendizaje, enseñanza y aplicación de los conceptos de estadística y probabilidad. Por un lado, da elementos a los profesores para explicar diferentes temáticas de una manera más amena y fácil de entender. Asimismo, le da al profesor la opción de profundizar ampliamente en los temas. Su entorno visual y animaciones no sólo permite, sino además alienta, que el estudiante explore y aprenda por si mismo utilizando el material de prácticas auxiliar al paquete. El material contribuye a que los estudiantes entiendan claramente los conceptos, se motiven a conocer más y a explorar por sí mismos.

CalEst sirve como material de apoyo para comprender mejor algunos conceptos y resolver problemas de diferentes libros de estadística cuya temática se enfoca a los planes de estudio en los bachilleratos, tecnológicos regionales, licenciaturas e ingenierías. Inclusive, por sus animaciones y graficas, también puede ser utilizado en primaria o secundaria para adentrar a los estudiantes en el tratamiento de la información.

Una parte fundamental del desarrollo de este trabajo se da en el conocimiento y habilidad de cómputo aplicada para explicar y desarrollar conceptos y resultados en la enseñanza de la estadística y probabilidad. Este proyecto se ha elaborado con alto desarrollo en programación avanzada y cómputo especializado dedicado como apoyo integral a la educación con la presentación de imágenes animadas y visuales para comprender los conceptos básicos en estadística y probabilidad. **CalEst** corre en los sistemas operativos Windows XP y Vista.

Capítulo 1

Estadística: el mundo de la información

La ciencia es conversación. Aprender y comprender también es conversación. Aprender y comprender quizá sea una actividad íntima. Se aprende y comprende en la soledad de la reflexión, pero siempre al final de algún tipo de conversación. Siempre he creído que algo extraño ocurre en las escuelas y universidades, en las que el alumno escucha mucho y conversa poco.

Jorge Wagensberg

Competencia general

Adquirir conocimiento acerca del mundo de la información en temas relacionados en administración y la economía, con la finalidad de aprender a realizar aplicaciones de los métodos estadísticos apoyados en modelos de probabilidad, así descubrir más sobre el mundo real mediante la colección, análisis e interpretación de los datos.

Competencias específicas

- Aprender a identificar problemas reales o curiosidades para crear la necesidad de generar información a través de datos.
- Plantear preguntas sobre la naturaleza de los estudios con la finalidad de tener ideas sobre problemas en la administración y la economía, y contestarlas con datos.
- Comprender los conceptos de población y muestra para evaluar la importancia de generar información para el conocimiento de temas en administración y economía.
- Estudiar los diferentes tipos de datos y su clasificación, con el fin de adquirir práctica para obtener información de una población por medio de una encuesta o un experimento.
- Conocer la importancia de aplicar una encuesta o realizar un experimento.
- Realizar estudios pequeños sobre diferentes temas propuestos con la finalidad de aplicarlos en encuestas.

1.1 Elementos básicos en Estadística

1.1.1 Introducción

La estadística es una materia que ha adquirido una importante presencia durante los últimos 20 años en diferentes áreas como las ciencias sociales, biológicas, bioquímicas y ciencias básicas. El impacto del estudio de problemas estadísticos en estas y otras áreas representa una enorme aportación al conocimiento humano y al desarrollo tecnológico.

La administración y la economía son otras de las disciplinas que no podrían sustraerse a los beneficios metodológicos de la estadística. Así, por ejemplo, observar, analizar e interpretar los precios de un producto dará lugar a que un administrador tome alguna decisión para no afectar la economía de su empresa. Una actividad importante que realizan los gerentes de una compañía es mantenerla competitiva en el negocio, por ello deben planear diferentes estrategias de ventas con base en métodos estadísticos como son los modelos de probabilidad, para tener una mayor presencia en el mercado. La información que obtengan mediante la colección, investigación y entendimiento de los datos, les permitirá evaluar si las tácticas empleadas resultaron eficientes. Otra de las funciones de los administradores en una agrupación es seleccionar personal, para lo cual realizan pruebas de habilidades e inteligencia, entre otras, cuyos resultados los presentan y analizan con técnicas estadísticas.

Como podrá verse en los ejemplos anteriores, los datos constituyen la "materia prima" para la generación de información, la cual a su vez se transformará en conocimiento a partir de su análisis e interpretación. Y es en este proceso donde la estadística desempeña una trascendente función en prácticamente todas las disciplinas.

Por lo tanto, los datos también se pueden generar mediante la realización de experimentos, como en el diseño de nuevos productos para que una corporación tenga más ganancias. Por ejemplo, en la elaboración de medicamentos, la dirección de un laboratorio puede experimentar con diferentes variedades de plantas para producir nuevas sustancias. Un investigador en medicina estudia a través de un experimento los efectos de la penicilina y otros antibióticos en seres humanos; las investigaciones acerca de medicamentos para enfermos mentales se realizan comparando el efecto con grupos de control y experimentales; también por medio de experimentos con grupos de control se prueban nuevos medicamentos.

Estudio estadístico

Cuando los datos no están disponibles en algún registro para una aplicación particular entonces se lleva a cabo un estudio estadístico. Éstos se clasifican como observacionales o experimentales.

Muchos profesionistas en diferentes áreas de estudio emplean la estadística para analizar sus datos. Algunos expertos, que han comprendido la importancia del análisis de sus datos, han hecho suya la frase del escritor inglés George Herbert Wells, quien dijo: "para cualquier persona, el pensamiento estadístico será tan necesario para ser eficiente como lo es la habilidad para leer y escribir ".

Muestreo

Los trabajos que realizamos en diferentes actividades requieren de investigación que puede ir desde sencilla hasta avanzada, pero en ellas se requiere de aplicación estadística. Una guía para esta actividad la señala Mario Tamayo Tamayo: "Tanto las ciencias sociales como las ciencias naturales inician estudios o realizan investigación de una forma controlada, sistematizada, crítica, con el fin primordial de aprobar o desechar hipótesis como explicativas de los fenómenos del comportamiento humano".

"Su punto de partida está también en la observación de los hechos, en la formulación del problema, en la estructuración de la hipótesis y en la búsqueda de pruebas para confrontar esas hipótesis, con el fin de establecer una ley o norma explicativa de la conducta social de los individuos".

El mundo de la información 1. Administración y economía familiar

En este escenario, se puede comenzar abordando el tema de la economía familiar, a partir de la siguiente interrogante: ¿cómo se distribuye el gasto de una familia al mes? Siguiendo en esta línea, también es de interés considerar la manera de administrar el gasto, así como plantearse si existen algunas estrategias para ahorrar en diferentes rubros del gasto. Véase la tabla 1.1.

Tabla 1.1 Porcentajes del gasto familiar.

Vivienda	**13 %**
Alimentos, bebidas y tabaco	**23 %**
Transporte	**13 %**
Educación y esparcimiento	**12 %**
Salud	**9 %**
Varios	**7 %**
Vestido	**6 %**
Domésticos	**4 %**

Preguntas sobre la naturaleza del estudio

Para determinar las características de la naturaleza del estudio, se puede formular la siguiente pregunta: ¿cuánto gastan en pesos, las familias en vivienda, en alimentos, en transporte, en educación, en salud, en diversión, entre otros? A la vez, en cada una de las categorías citadas se plantean otras preguntas, por ejemplo ¿en qué tipo de alimentos gastan? ¿Qué tan nutritivos son? ¿Cómo se genera la información?

Estrategias estadísticas sobre el estudio

El gasto de la familia es una variable aleatoria que en una primera aproximación se divide en categorías. En cada una de éstas, del total del ingreso se asocia el porcentaje del gasto, en referencia al INPC (índice nacional de precios al consumidor), como se muestra en la figura 1.1

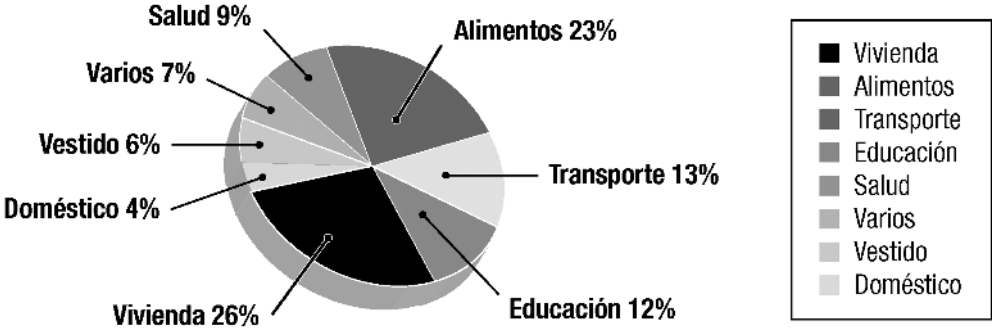


Figura 1.1 Distribución del gasto familiar.

El mundo de la información 2. Salario mínimo

Otro tema relevante de la economía familiar son los ingresos, los cuales están regulados por los salarios mínimos que se establecen de manera anual. En particular en México, se dan por tres zonas geográficas A, B y C, a partir del 2013 se reducirán a dos zonas A y B. En este sentido puede plantearse: ¿cómo se relaciona este salario con el precio de los alimentos para saber si dicho salario es competitivo?

Preguntas sobre la naturaleza del estudio

En relación con la naturaleza del estudio, es pertinente preguntarse: ¿el salario mínimo crece de manera proporcional al aumento real en tortillas, en frutas y verduras, entre otros? ¿Cuántos salarios mínimos, debe ganar una persona para no caer a la definición de pobreza? ¿Cuántos salarios mínimos gana un funcionario? ¿Los precios que los administradores de restaurantes suben a los alimentos, se relacionan con el incremento del salario mínimo? ¿Las casetas de peaje, suben en la misma proporción que el salario mínimo? ¿Existe una recompensa real al ingreso de un trabajador con respecto al salario mínimo?

Estrategias estadísticas sobre el estudio

Al cambiar de año el salario mínimo sufre una modificación tal y como se observa en la figura 1.2; ésta describe el historial de los últimos años de la zona B.

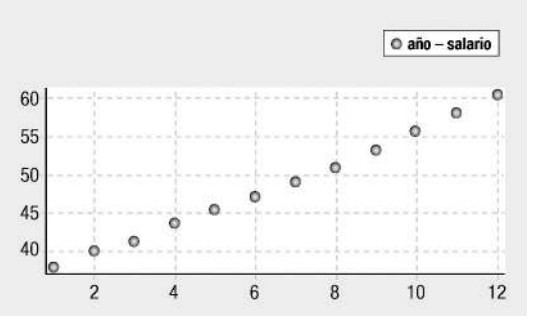


Figura 1.2 Salario mínimo en México del 2001 al 2012.

Datos: el mundo de la información

En el marco de la administración y la economía, a continuación se refiere de manera general una lista de situaciones en la vida que necesita la generación de datos para tomar alguna medida correctiva.

- La administración de una industria que produce alimentos necesita tener la información adecuada para conocer la vida de anaquel de un alimento.
- En el contexto de la economía social, un país: desea observar el número de años de estudio que tienen diferentes sectores de la población para establecer estrategias de planes educativos.
- En una región urbana, el 3 % de las parejas se divorcian antes de cumplir un año de casados y 35 % se divorcia antes de cumplir cinco años. ¿Cuál es el impacto económico que produce esta situación?
- Economía en la salud: en una ciudad, con un nuevo tratamiento médico disminuyó el número de diabéticos.
- Economía familiar: en un determinado país, el poder adquisitivo de las familias disminuye año con año.
- Administración y economía en la empresa: tener una descripción mensual del personal ocupado y horas trabajadas en la industria maquiladora.

Los ejemplos anteriores pretenden hacer evidente que el conocimiento de diferentes problemáticas en las áreas de administración y economía se obtiene mediante la generación de datos, los cuales una vez procesados se traducen en información. Este proceso da lugar a un mejor conocimiento sobre los estudios, lo que permitirá a su vez plantear estrategias de solución y toma de decisiones. La necesidad de tener una buena comprensión del mundo que nos rodea, implica aprender y practicar la estadística. Por otro lado, para obtener la información estadística y su interpretación, es fundamental tener habilidad en la operación de los conceptos básicos de la probabilidad. Por ejemplo, en varias áreas del ámbito laboral, ya sea en los negocios, la medicina, la educación, las ciencias sociales o la física, por mencionar algunas disciplinas, las personas deben poseer cierta habilidad para leer, interpretar y aplicar los resultados de un análisis estadístico de datos.

Ejemplos del papel de la estadística

Como hemos visto, *la estadística* desempeña un papel clave en muchas otras actividades. A continuación se presenta una lista con algunas de éstas:

- Evaluar el efecto de diferentes métodos del rendimiento de trabajo aplicados en una administración.
- Interpretar los censos económicos de la población.
- Aplicar y analizar las encuestas de opinión, para que una administración pueda llevar a cabo mejoras en determinada empresa.
- Llevar a cabo el monitoreo del impacto de la lluvia ácida en la producción agrícola y evaluar su repercusión económica en las cosechas.
- A la gerencia de una compañia le interesa llevar a cabo la aplicación de las técnicas de control de calidad para diferentes productos.
- En estudios económicos se calculan las tasas de desempleo y ocupación en América Latina.
- Analizar e interpretar diferentes indicadores, para establecer el índice de pobreza en el mundo.
- Para planear estrategias de venta, se desea determinar y clasificar las 40 canciones más escuchadas en el radio cada semana.
- Existe un efecto económico en la determinación de garantías, por ello es necesario predecir la confiabilidad de diferentes productos: calentador, estufa, televisores, por mencionar algunos.
- Establecer si la edad o el género son factores de discriminación para conseguir empleo.
- Es importante emprender estrategias médicas para precisar el impacto que tiene el desarrollo físico de un bebé al nacer cuando la madre es fumadora o consume alguna otra droga.
- Entre otros efectos, la prevención de accidentes automovilísticos tiene consecuencias económicos, por ello es relevante establecer la eficacia de los cinturones de seguridad en la prevención de accidentes.
- Una de las metas en las administraciones del sector de gobierno es estimar la cantidad de drogas ilegales que entran de contrabando en un país.
- En el contexto de inversión, la administración de un hospital desea adquirir equipos de ultrasonido para estudiar y modelar el desarrollo de un feto durante la gestación.
- Con el propósito de obtener recursos económicos, es importante determinar la incidencia de una enfermedad contagiosa.
- La dirección escolar requiere conocer las necesidades de los jóvenes y evaluar su escala de valores.

- Un sociólogo tiene interés en evaluar si una de las causas de la desintegración familiar se debe a la situación económica.

Un problema estadístico se involucra con el estudio de alguna característica asociada con un grupo de objetos, a la cual suele denominarse *unidad de observación* o *experimental*.

La estadística es una rama de las matemáticas que trata del análisis e interpretación de un conjunto de datos, de manera que ésta requiere de métodos en probabilidad, y así, resolver problemas estadísticos.

Ejemplo 1.1

Con frecuencia los periódicos contienen reportes de estudios estadísticos. El lunes 3 de noviembre de 2012, en uno de los diarios de circulación nacional se publicó la siguiente noticia de indicadores de salud: "El número de unidades médicas públicas de salud pasó de 19 099 en el año 2000 a 21 973 en el 2011, con una tasa media anual de crecimiento de 1.28 %. Del total de 2011, 20 705 son unidades de consulta externa y 1 268 de hospitalización".

Ejemplo 1.2

En la actualidad, existen muchos organismos -nacionales o internacionales- que realizan estudios estadísticos con fines comparativos, así como para evaluar el avance y desarrollo de un país. Por ejemplo, en las publicaciones de los estudios que lleva a cabo la Organización para la Cooperación y el Desarrollo Económico (OCDE), se reportó que México ocupa los siguientes lugares: la última posición de 21 países miembros de la OCDE en gasto en educación por estudiante; la última posición de los 27 países en el número de investigadores por cada mil habitantes empleados, y la segunda posición entre 28 países en el porcentaje de población que padece obesidad.

Consulte la siguiente página para conocer diferentes estadísticas reportadas por esta organización.

http://es.wikipedia.org/wiki/Organización_para_la_Cooperación_y_el_Desarrollo_Económico

1.1.2 Población y muestra

Los conceptos de población y muestra son sumamente importantes para la generación de información y conocimiento en general. En temas de administración y economía también son fundamentales, tal como se muestra en el ejemplo siguiente de El mundo de la información.

El mundo de la información 3. Mejorar una medicina

Elaborar medicinas más eficientes para curar una gripe es una tarea que involucra a muchas personas, tales como los bioquímicos que las formulan, los dueños de laboratorios que las producen, los médicos que las prescriben y los enfermos que las toman. La idea es considerar una población de personas mayores de 16 años y enfermas de gripe. Se observará el tiempo en días que tarda una persona en restablecerse después de haber iniciado un tratamiento. Se puede decir que la **estadística es el proceso de descubrir más sobre el mundo real** mediante la caracterización de los problemas, haciendo preguntas sobre su naturaleza y luego a través de la formulación de una adecuada colección, análisis e interpretación de los datos. La figura 1.3 presenta un resumen de la actividad de los proyectos de corte estadístico.

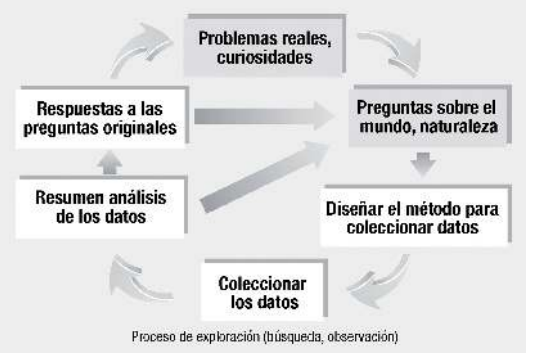


Figura 1.3 Modelo que caracteriza a los proyectos estadísticos en el descubrimiento del conocimiento.

Preguntas sobre la naturaleza del problema

Por experiencia todos sabemos que una gripe dura varios días. En muchos casos, las personas acuden a medicinas que se venden sin receta, y por uso y costumbre se evalúa cuál es más efectiva. Pero otras personas siguen algún tratamiento médico. ¿Cómo valoramos la eficiencia de un medicamento? Una respuesta sencilla a esta pregunta es saber en cuántos días se alivian las personas de los síntomas. También se puede averiguar si con el tratamiento seguido disminuyeron sus molestias. ¿Cómo conseguir toda esta información?

Estrategias estadísticas sobre el estudio

Nuestra meta es producir información acerca del efecto de un tratamiento para curar la gripe. En principio, se sabe que el mercado día a día incorpora nuevos medicamentos para disminuir los síntomas y malestares de esa enfermedad. Un procedimiento natural para generar información en este caso, es aplicar un tratamiento con la medicina X en un grupo de personas con gripe. Luego se cuenta y registra el número de días en que logra aliviarse una persona.

Con el propósito de ampliar nuestro conocimiento del tema, se anotará si disminuyeron las molestias, y si el medicamento produjo sueño.

Sin embargo, el reto está aún presente y se plantean las siguientes preguntas: ¿cómo se planea rea-

lizar el estudio? ¿Qué personas recibirán el medicamento? ¿Cómo se escogen a las personas que seguirán el tratamiento? ¿Cómo medir el efecto de una medicina? Si sólo a unas cuantas personas se les da la medicina y se observan los resultados, ¿éstos serán igualmente válidos para más personas?

1.2 Métodos para la colección de datos

1.2.1 Procedimiento para recabar información

El primer paso es identificar a la *población* que será objeto de estudio. La población está constituida, en sentido amplio, por individuos u objetos. A éstos se les observa algún atributo, el cual es la finalidad de cualquier estudio. En resumen, estas observaciones generan un conjunto de datos y se convierten en la información que se requiere para conocer sobre un tema específico. En ocasiones, los datos también son producto de resultados experimentales.

En muchas áreas de aplicación, las poblaciones son fáciles de identificar, pero en otras no es tan sencillo. La población debe estar bien definida antes de iniciar un estudio. En el mundo de la información 1 la población de búsqueda fueron las personas enfermas de gripe. Pero, ¿cómo la definimos? Se considera a las personas con gripe y la población puede limitarse a una escuela, una colonia, una delegación o un hospital.

Una vez definida la población, el siguiente paso es obtener una *muestra* de dicha población. Una muestra consiste en seleccionar una parte de la población para realizar el estudio. A continuación se hace un breve análisis de la información generada a partir de la muestra.

Las ventajas de seleccionar una muestra radican en reducir el tiempo y el esfuerzo para obtener la información. Una adecuada planeación para seleccionar la muestra proporcionará una información aproximada sobre las características de la población.

Población

Una *población* consta de una colección de individuos u objetos a los que se les observa una característica particular que será objeto de estudio.

Muestra

La *muestra* es una parte de la población que se estudiará para conocer las características de la población.

øPor qué tomar una muestra?

Se ha dicho que la población representa el todo y la muestra es sólo una parte de la población. Una pregunta que surge al respecto es ¿por qué procurar examinar una muestra cuando lo que realmente se desea es estudiar la población? La mayoría de las veces no se puede estudiar a la población, por lo que

podemos usar la muestra como una guía. Las razones principales para pensar así son:

- Podría tomar mucho tiempo estudiar una población.
- Resulta muy costoso, en cuanto a esfuerzo, estudiar una población.
- En ocasiones es difícil identificar a todos los miembros de una población.
- Si estudiamos a toda la población, no debemos dejar a alguien fuera del estudio.

Variable

En el mundo de la información 1, se busca el tiempo que tarda una persona en recuperarse de una gripe. Esta información se precisa por el número de días, y resulta una característica relevante para el estudio, la cual se conoce como *variable de respuesta*. La variable de respuesta, o simplemente variable, desempeña un papel fundamental en el trabajo estadístico.

Los datos que se observan en la muestra que se tomó del mundo de la información 3 son el número de días. Si, por ejemplo, en la muestra hay cinco personas, 4, 10, 3, 8 y 6 podrían ser los días que esas personas tardan en recuperarse de una gripe.

Variable y censo

Una variable es una característica que tiene cada miembro de la población. Una variable tiene o toma un valor para cada uno de los miembros o individuos.

Un censo es el estudio sobre toda la población.

Nociones intuitivas sobre población y muestra

Algunas vivencias que nos son familiares pueden ayudarnos a comprender los conceptos de población y muestra. Por ejemplo, cuando vamos a un laboratorio para que nos practiquen un análisis de sangre para evaluar nuestro estado de salud, sólo toman una muestra de sangre y no toda la sangre, esto es, la población. Por supuesto, se sabe lo que ocurriría si se sacara toda la sangre, por lo que con una muestra se tendrá una buena aproximación al diagnóstico del estado de salud. Sin embargo, puede darse el caso de que la información proporcionada por la muestra no coincida con la situación real. Para evaluar esta situación se recurre a conceptos de probabilidad, los cuales se tratarán en otro capítulo.

Cuando queremos comprar nueces en un tianguis, le pedimos al vendedor que nos permita tomar una muestra. No permitimos que él nos dé la muestra, y sacamos del costal un par de nueces. Si están bien y nos gustan, compramos un kilo o más. Así también, cuando pedimos una probadita de algún producto, como un helado, el vendedor nos dará un poco de helado con una cucharita para que uno se decida hacer o no la compra. Lo que nos dan es una muestra, pues el vendedor no nos dará todo el bote de helado, que en este caso es la población.

Ronald Aylmer Fisher (1890 - 1962)

Este científico, matemático, estadístico, biólogo evolutivo y genetista inglés, realizó muchos avances en la estadística, siendo una de sus más importantes contribuciones la inferencia estadística creada por él en 1920.

Hablando de Fisher, Efron expresó: "incluso los científicos necesitan sus héroes, y R. A. Fisher fue sin duda el héroe de la estadística del siglo XX. Sus ideas transformaron nuestra disciplina de tal forma que hasta un César o un Alejandro hubieran envidiado".

Cometarios de interés. El proceso de seleccionar una muestra se conoce como muestreo. El muestreo no es una selección caprichosa, sino que incorpora reglas definidas para la selección de la muestra. Dichas reglas se originan en la probabilidad y es un tema que se verá más adelante, en el capítulo 7.

Para que los estudios estadísticos sean eficientes y proporcionen un buen conocimiento de la población, se requiere de una adecuada selección de los datos. Por ello, resulta importante aprender las técnicas para seleccionar una muestra. A veces, tenemos necesidad de estudiar a la población entera. Cuando este es el caso se dice que se practica un censo. [1]

Ejemplos de población y muestra

1. Elaborar productos más eficientes en la búsqueda de nuevos remedios para tratar la caspa, la calvicie, la obesidad, etcétera. En el tratamiento de la caspa muchas personas están involucradas (como en el mundo de la información 3), esto es, los bioquímicos que hacen fórmulas para diferentes champús u otros productos, los dueños de los laboratorios que los producen, y las personas que requieren el tratamiento. Ahora bien, aquí la población queda integrada por las personas que tienen caspa. Si bien en este caso puede resultar complicado identificar a dichas personas, será necesario limitar el estudio a un universo más específico. En consecuencia, la muestra será una parte de las personas con caspa.

2. El área de control de calidad busca mejorar los productos. El control de calidad es una actividad importante en la mayoría de las empresas que manufacturan diferentes productos. El seguimiento de estos productos está a cargo de administradores, ingenieros industriales, químicos, mecánicos, electrónicos, por mencionar algunos. Un fabricante de "chips" para computadoras desea monitorear la calidad del producto. Dado que se produce una gran cantidad de chips cada día, sólo se tomará

[1] A lo largo del libro propondremos algunas actividades para que el lector obtenga algunos conjuntos de datos, ello con el fin de que pueda desarrollar la habilidad y el conocimiento para generar su propia información y pueda analizarla. Esto limita la inferencia de los resultados a la población, sin embargo se plantearán algunas prácticas para simular el proceso en la selección de una muestra.

una muestra de éstos. La población consta de todos los chips fabricados por el productor en ese día o en un lote de producción. Entonces, ¿cuál sería una muestra? La respuesta es: una parte de los chips de ese lote de producción. Existen mecanismos estadísticos para seleccionar los chips de la muestra, los cuales se verán más adelante.

3. Encuestas para las elecciones. En la actualidad, diferentes empresas encuestadoras realizan sondeos para conocer la preferencia de los votantes por algún candidato. Ellas seleccionan una muestra del directorio telefónico para conocer la opinión de la gente. ¿Cuál es esa población? La población es el número de personas con teléfono y que aparezcan en ese directorio telefónico. ¿Cuál es la muestra? Una muestra será escoger unos cuantos números que hay en ese directorio.

Estudio observacional

Un estudio observacional, es aquel que observa a las unidades de la población o muestra y mide las variables de interés. Una respuesta es una variable que mide el resultado de un estudio. El objetivo de un estudio observacional es describir algún grupo o situación. Un estudio observacional puede ser descriptivo o analítico.

1.2.2 Procedimiento para seleccionar una muestra

Con el propósito de mostrar una aproximación al concepto de la relación población-muestra, una idea inicial se esboza en la figura 1.4, la cual además ejemplifica la manera en que se escoge una muestra, y también plantea una práctica de la misma. En general, el tamaño de la población es el número de individuos de la población. Se simboliza con N. El tamaño de la muestra es el número de individuos de la muestra. Se simboliza con n.

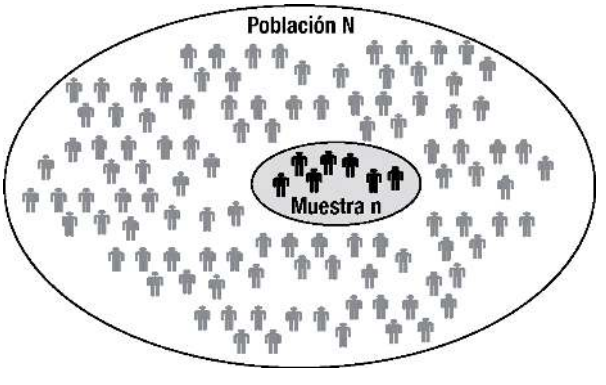


Figura 1.4 Relación población de tamaño N y muestra tamaño n.

Práctica 1

Procedimiento para la selección de una muestra. Proponemos una población de $N = 50$ personas. El 50 representa el tamaño de la población y la meta es seleccionar una muestra de 5 individuos: $n = 5$. Con referencia al problema 3, tenemos una población de jóvenes que asisten a una escuela de nivel medio superior, y de éstos, 50 tienen gripe e iniciaron un tratamiento. El conjunto de datos es:

Tabla 1.2 El número de días que una persona se recupera totalmente de una gripe.

3	7	2	8	9	4	4	6	9	9
7	5	6	9	5	6	5	2	6	5
6	7	8	6	5	7	6	3	7	6
7	3	5	4	4	7	5	9	4	6
6	8	9	8	7	2	8	6	8	2

Solución: Procedimiento 1 para seleccionar una muestra

Se cortan 50 papelitos y en cada uno se anota el número de días que cada joven tardó en aliviarse. Se doblan los papeles y se depositan en una caja o bolsa, se revuelven y se seleccionan 5 papeles, anotando en una tabla el valor que indica la respuesta.

Tabla 1.3 Selección de una muestra de 5 personas.

Selección 1	9
Selección 2	5
Selección 3	6
Selección 4	8
Selección 5	4

Repita el mismo procedimiento. ¿Cuáles son sus valores? Una tabla de números aleatorios es una lista de números generados y listados en el orden en el que se generaron.

Tabla de números aleatorios

Mediante el uso de una tabla de números aleatorios se puede seleccionar la muestra. En la tabla 1.4 presentamos una porción de la tabla de números aleatorios. En el apéndice Tablas se tiene una más completa. Una tabla de números aleatorios debe contener de manera muy aproximada tantos ceros (0) como unos (1), y así sucesivamente hasta 9.

¿Cómo funciona la tabla? Para usarla, primero debe identificarse el tamaño de la cifra que nos servirá de referencia. Para fijar esta idea, pensemos que tenemos el número de los expedientes de los alumnos en una escuela y que su expediente tiene cinco números. El primer número aleatorio que se escoja con cinco números corresponde al número de expediente de un alumno y ése es el que estará en la muestra. También

el número de credencial de los alumnos podría servir como referencia para identificar al individuo en la población.

Tabla 1.4 Porción de una tabla de números aleatorios con 7 columnas y 10 renglones.

1	75154	93670	85909	81250	70667	73814	81475
2	85824	83677	39169	75348	96200	96112	17203
3	97363	65182	53409	81156	11548	91729	63780
4	60922	63727	98853	47346	27716	30801	68153
5	80022	81475	72039	95152	68374	43012	29796
6	16719	40197	17148	73965	57929	45691	86499
7	54553	69711	68295	29969	14810	36859	64576
8	40024	52842	96948	38807	16625	51388	74773
9	47290	38044	52860	35749	15608	29880	43268
10	90836	21503	95996	35601	12460	46380	93888

El camino que se sigue para entrar a la tabla es cerrar los ojos y apuntar a un número. Por ejemplo, considere que en su intento seleccionó el número que está en el renglón 3 y la columna 5. Entonces el número es 11548, los cuatro números que faltan se escogen siguiendo el mismo renglón pero cambiando de columna. Entonces esos números son:

91729, 63780, 60922 y 63727.

Los cinco alumnos seleccionados son los que tienen estos números en su expediente. También es válido el proceso de selección de una muestra aleatoria, si se pasa de un renglón al siguiente. Por ejemplo para el número 11548, bajando por el renglón los 4 números restantes son:

27716, 68374, 57929 y 14810.

A una muestra seleccionada mediante este procedimiento se le llama muestra aleatoria simple, y quiere decir que cada elemento de la población tiene la misma oportunidad de ser seleccionado. Una muestra aleatoria simple es una muestra que ha sido seleccionada de tal manera que todos los miembros de la población tengan la misma oportunidad de ser escogidos.

Procedimiento 2 para seleccionar una muestra

Primero identificamos a los 50 jóvenes asignándoles un número y se elabora una lista. Luego, para seleccionar a 5 personas se usa la tabla 1.5 de números aleatorios. De la tabla 1.4, hemos elegido con los ojos cerrados la columna 1 y el renglón 5. Nos fijamos en los dos últimos números de la cifra 80022, en este caso 22, entonces en la tabla 1.5, se busca el ID correspondiente a 22, lo que indica que la persona tardó 9 días en curarse. Así bajamos a la siguiente cifra, ésta es 16719 y el número correspondiente es 19;

la cifra que sigue es 54553, y el número es 53 pero como la población es de 50 números descartamos el 53 y pasamos a la siguiente cifra, es decir, 40024. En este caso, es el 24, y como se terminaron los números de esa columna pasamos a la siguiente, y seguimos la misma operación hasta completar el tamaño de la muestra. La selección de 5 números y el número de días en recuperarse se presenta en la siguiente tabla:

Selección 1	22	9
Selección 2	19	6
Selección 3	24	4
Selección 4	27	5
Selección 5	11	7

Tabla 1.5 Identificación de las personas de la población y los días que tarda en recuperarse.

ID: 01	3	ID: 11	7	ID: 21	8	ID: 31	4	ID: 41	6
ID: 02	7	ID: 12	5	ID: 22	9	ID: 32	6	ID: 42	2
ID: 03	6	ID: 13	7	ID: 23	6	ID: 33	7	ID: 43	3
ID: 04	7	ID: 14	3	ID: 24	4	ID: 34	7	ID: 44	9
ID: 05	6	ID: 15	8	ID: 25	8	ID: 35	2	ID: 45	6
ID: 06	2	ID: 16	9	ID: 26	9	ID: 36	4	ID: 46	9
ID: 07	6	ID: 17	5	ID: 27	5	ID: 37	5	ID: 47	6
ID: 08	8	ID: 18	6	ID: 28	5	ID: 38	6	ID: 48	7
ID: 09	5	ID: 19	6	ID: 29	4	ID: 39	5	ID: 49	4
ID: 10	9	ID: 20	2	ID: 30	7	ID: 40	8	ID: 50	8

Procedimiento 3 para seleccionar una muestra, mediante el uso de CalEst

El paquete estadístico CalEst brinda la posibilidad de generar una muestra aleatoria. En el menú Herramientas, a la izquierda en la figura 1.5, se muestra la opción para generar números aleatorios cuyo nombre es precisamente **Genera números aleatorios**.

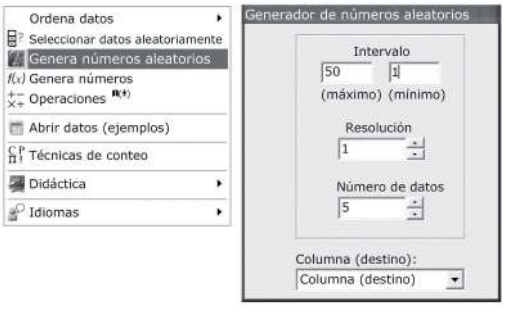


Figura 1.5 Opciones que se obtienen a partir del módulo Herramientas. Entre ellas está el Generador de números aleatorios.

Una vez que se selecciona dicha opción, aparece una pantalla como la que se encuentra a la derecha en la figura 1.5. En la configuración determinamos el intervalo de números en que se desea tener un número aleatorio; para el ejemplo los valores son: de 50 a 1. A continuación se pasa al cuadro Número de datos, en cuyo caso se ha propuesto una muestra de 5 individuos, es decir, sólo hay una variable. Finalmente, se indica la columna, en la hoja de datos, en que se desea que aparezcan los números, y se oprime el cuadro Aceptar. Un resultado de muchos posibles se describe en la variable 1.

A continuación, se reproducen los resultados en los cuales se observa el número de días en que se restablecen de la gripe las personas que pertenecen a la muestra.

Selección 1	8	8
Selección 2	4	7
Selección 3	13	7
Selección 4	12	5
Selección 5	11	7

Se recalca el procedimiento para obtener una muestra:

1. Poner en una caja o bolsa trozos de papel con el nombre de cada uno de los miembros de la población, y luego sacar el número de miembros que se desea que esté en la muestra.
2. Seleccionar la muestra mediante la tabla de números aleatorios.
3. Obtener una muestra aleatoria usando el paquete estadístico.

1.2.3 Proceso experimental

Mediante la experimentación también se genera información para el análisis estadístico. Además, en este caso se puede estudiar una relación de causa efecto entre variables.

La ciencia

La ciencia se basa en dos actividades: percepción y reflexión. Las dos cosas tienen que ver con la realidad de este mundo y las dos son, en el fondo, dos formas de conversación. La percepción de la realidad empieza por ver, mirar (detener la vista) y observar (detener la mirada), pero suele acabar en algo más comprometido: experimentar. Jorge Wangensberg.

El mundo de la información. Prevención de la gripe

En el contexto de la economía familiar, las administraciones de los centros de salud a nivel nacional realizan campañas para aplicar vacunas con el objeto de prevenir enfermedades. En particular, en los

últimos años ha sido común que durante los meses de octubre y noviembre se administre una vacuna para procurar que las personas no enfermen de gripe, ya que en los meses de mayor frío suele ser frecuente esta enfermedad. En este caso interesa conocer si hay un efecto favorable en vacunarse para prevenir la gripe. El estudio se hace mediante una estrategia experimental en que se consideran dos grupos, uno que recibirá la vacuna y otro que no.

Preguntas sobre la naturaleza del problema

¿A quiénes se les aplicará la vacuna? ¿Cómo se dará seguimiento para observar si las personas se contagiaron o les dio gripe a pesar de estar vacunados?

Estrategias estadísticas sobre el estudio.

En este caso al consultar a una persona, la variable es contrajo gripe, y la respuesta es "sí" o "no". Los datos se tienen para dos grupos: uno de ellos es el número de personas vacunadas que no contrajeron gripe; el otro, el que corresponde al número de personas no vacunadas que no contrajeron gripe. El interés médico aquí es ver si esta diferencia es realmente importante.

Complemento técnico: Los elementos esenciales que se han visto en estos dos problemas se sintetizan a continuación. Éstos desempeñan un papel preponderante para llevar a cabo el resumen y el análisis de los datos.

1. Una unidad experimental es la entidad más pequeña que es de interés en un estudio estadístico.
2. Una variable es una característica que se mide en cada unidad experimental en un estudio estadístico.
3. Una observación es el valor que toma la variable para una unidad experimental.
4. La colección de observaciones que toma una o más variables conforma el conjunto de datos.

Objetivo de un experimento

Un experimento consiste en aplicar un tratamiento a una unidad o individuo con el propósito de observar su respuesta. La intención de un experimento es determinar si el tratamiento produce un cambio en la respuesta.

Comentarios:

- Un objetivo del análisis de datos es organizar y resumir la información del conjunto de datos con el propósito de hacerla más comprensible; es lo que se conoce como estadística descriptiva, y se presentará en el capítulo 2.
- En las secciones El mundo de la información 1 y 2 se han generado los datos: uno mediante un proceso de muestreo y el otro con un estudio experimental. Éstas son dos áreas del proceso estadístico que ayudan a generar información.

1.3 Tratamiento de la información

1.3.1 Tipo de datos

Se ha visto que existen dos maneras de generar datos. En este apartado se verán y clasificarán los diferentes tipos de datos.

El mundo de la información 4. Reporte sobre una muestra de estudiantes

Se le preguntó a un grupo de estudiantes por el tipo de carrera profesional que desean cursar en la universidad. Además se les preguntó sobre otras características, tales como el desempeño académico medido por el promedio general de calificaciones, el género, si estudia otro idioma además del español y el número de horas que la persona estudia a la semana.

Preguntas sobre la naturaleza del problema

Como se observa, se preguntan varias características de interés. La idea del problema es hacer una presentación de las respuestas que se dan, de acuerdo con el tipo de variable, para clasificar las variables y el tipo de datos que se generan. **Datos:** En la tabla 1.6 se presentan los resultados que respondieron 8 estudiantes.

Tabla 1.6 Diferentes características para individuos y tipos de variables.

Género	Carrera	Idioma	Promedio	Horas de E.
Femenino	Robótica	Inglés	8.1	3
Femenino	Mecatrónica	Inglés	7.8	2
Masculino	Diseño	Inglés	8.8	7
Femenino	Biología	Francés	8.2	5
Masculino	Leyes	Inglés	7.3	1
Masculino	Medicina	Inglés	8.5	4
Femenino	Matemáticas	Francés	9.0	6
Femenino	Comunicación	Inglés	8.0	3

Una colección de datos es un conjunto de observaciones o mediciones que se toman de un grupo de objetos o individuos. En el problema se han considerado unas variables, tales como el género, el tipo de carrera, el idioma, el promedio general y el número de horas que estudia a la semana. Las observaciones correspondientes a esas variables en el primer renglón son: Femenino, Robótica, Inglés, 8.1 y 3, por citar una de las posibilidades de que ocurra. Como aquí se han considerado varias variables, se habla de un conjunto de datos multivariado, en el sentido de que se han medido varias características a cada unidad.

Conjunto de datos

Los datos se *clasifican* en (univariados, bivariados y multivariados):

1. Un conjunto univariado de datos es una colección de datos al que en cada unidad experimental se le mide una característica (variable).
2. Un conjunto bivariado de datos es una colección de datos al que en cada unidad experimental se le miden dos características (variables).
3. Un conjunto multivariado de datos es una colección de datos al que en cada unidad experimental se le miden varias características (variables).

Tipos de variables

1. Una *variable numérica* es una variable cuyos valores son números, los cuales se obtienen mediante un conteo o medición. A este tipo de variable también se le llama variable cuantitativa.
2. Una *variable categórica* es una variable con clasificaciones o categorías. También se le conoce como variable cualitativa, y se clasifica en nominales y ordinales.
3. Una variable discreta es una variable numérica que toma un número finito o infinito contable de valores. Contable significa que podemos asociar a los valores con el conteo de los números: 1, 2, 3,. . . , es decir, valores que pueden contarse.
4. Una variable continua es una variable numérica que puede tomar un número infinito de valores entre dos números o valores asociados a un intervalo de la recta numérica.

Como se advierte en la tabla 1.6, hay diferencias entre los valores de las variables. Por ejemplo, las que aparecen en las tres primeras columnas son no numéricas (cualitativas) y las dos últimas son numéricas (cuantitativas).

Ejemplos de variables numéricas o cuantitativas

- Promedio en el grado escolar.
- Peso.
- Edad.
- Ingreso familiar.
- Razón de crecimiento.
- Número de hijos por familia.

- Número de suicidios en un año determinado.
- Número de pacientes operados con éxito.
- Número de votantes a favor de un candidato.

Las variables numéricas son discretas o continuas.

Por lo general, las variables discretas son resultado de un proceso de conteo. Las últimas cuatro variables citadas en los ejemplos son discretas. Así, el número de votantes es: 1, 2, 3,...

Tipos de variables continuas son las primeras cinco variables escritas en los ejemplos de variables numéricas, es decir: promedio en el grado escolar, peso, estatura, ingreso familiar, razón de crecimiento.

Ejemplos de variables categóricas nominales

• Género.	• Estado civil.
• Afiliación política.	• Estado de salud.
• Ocupación.	• Tipo de lectura.
• Preferencia religiosa.	• Género de película que más le gusta.

Ejemplos de variables categóricas ordinales

- Grados de estudio.
- Severidad del estado de salud de una persona.
- Nivel de preferencia por un producto.
- Calidad de colorido en una foto.
- Frecuencia con que se realiza ejercicio.

En la figura 1.6 se muestran unos ejemplos del tipo de variable y su relación con el tipo de dato. En resumen, la clasificación de variables se muestra en la figura 1.7.

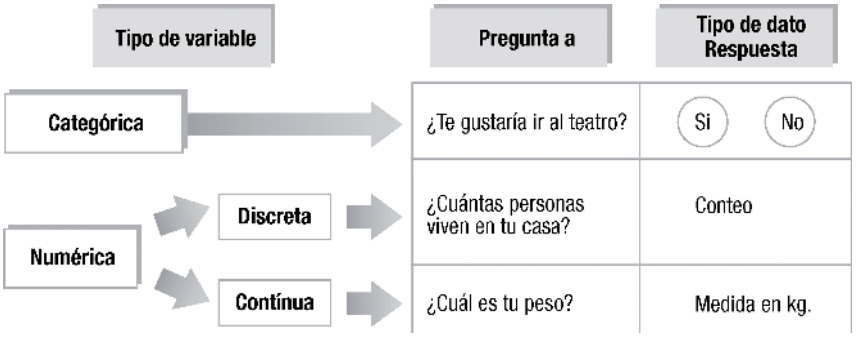


Figura 1.6 Ejemplos de tipo de variable y tipo de dato.

- Una variable categórica nominal se obtiene clasificando en categorías las características de la unidad observada o experimental.
- Una variable categórica ordinal se obtiene asignando las unidades observadas o experimentales en categorías. Cada categoría tiene una jerarquía y un orden.

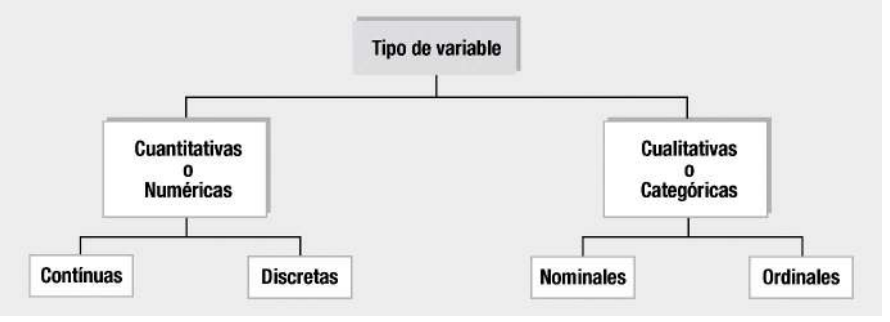


Figura 1.7 Clasificación del tipo de variables.

Aplicación de escalas para tener información. La administración de los establecimientos que prestan servicios, como en restaurantes, hospitales, cines, hoteles, escuelas, entre muchos otros, nos entregan hojas, cuestionarios, para evaluar el servicio. Debido a que los dueños de esos negocios desean estar a la vanguardia en los servicios, contratan personas para que analicen esa información. Con los datos obtenidos los dueños pretenden realizar las mejoras necesarias, que tendrán un impacto económico.

Temas de interés: escala para medir Linker: Un tipo de preguntas en una encuesta que maneja datos ordinales, se conoce como la escala Linker. Existen varias líneas en las que se puede emplear este tipo de escala, a saber, por convenio o acuerdo, frecuencia, importancia, calidad, confianza. A continuación se presenta el tipo de preguntas para la escala Linker. Tipo de escala 1, referida como Acuerdo.

Tipo de escala 1: Acuerdo

Totalmente de acuerdo	
De acuerdo	
Indeciso	
En desacuerdo	
Totalmente en desacuerdo	

Ejemplo para el tipo de escala 1

Se realiza un estudio con respecto a la aprobación de la píldora del día después, marque con **X** una de la siguientes opciones, está:

1. Totalmente de acuerdo

2. De acuerdo
3. Indeciso
4. En desacuerdo
5. Totalmente en desacuerdo

Ejemplo para el tipo de escala 2

Se presenta una segunda forma: Tipo de escala 2, referida como frecuencia. En la tabla de abajo se presenta el esquema con dos opciones. En un estudio para saber el comportamiento social de los jóvenes se les pregunta: ¿cuando asiste a una fiesta acostumbra a tomar bebidas alcohólicas? Marque con **X** una de la siguientes opciones.

Tipo de escala 2: Frecuencia

Opción 1		Opción 2	
Muy frecuente		Siempre	
Frecuentemente		Muy frecuente	
Ocasionalmente		Ocasionalmente	
Rara vez		Rara vez	
Muy raro		Muy raro	
Nunca		Nunca	

Ejemplo para el tipo de escala 3

Una tercera opción es: Tipo de escala 3, referida como Importancia. Marque con **X** una de la siguientes opciones. "En relación con un futuro inmediato, terminar la carrera profesional para usted es":

Tipo de escala 3: Importancia

Muy importante	
Importante	
Moderadamente importante	
De Poca importancia	
Sin importancia	

Ejemplo para el tipo de escala 4

Finalmente el tipo de escala 4, se refiere a opinar, esta puede ser sobre la calidad de un servicio, así como externar un punto de vista sobre algún tema. **Nota:** en la tabla de abajo se puede cambiar las últimas dos opciones: Regular por Pobre, y Muy pobre por Mala. En un sondeo para conocer la relación que lleva con sus padres, marque con **X** una de la siguientes opciones si dicha relación es:

Tipo de escala 4: Opinión

Excelente	
Muy buena	
Buena	
Regular	
Mala	

Encuestas

Una encuesta es un estudio o investigación que se realiza a una muestra de individuos representativa seleccionada de una población, mediante el empleo de procedimientos de interrogación cuyo objetivo es obtener mediciones cuantitativas y cualitativas de una diversidad de características objetivas y subjetivas de una población.

Ejemplo 1.3

Encuestas. En este ejemplo se describe una encuesta que se realiza en un conjunto de salas cinematográficas, las cuales las utilizan para evaluar el servicio que ofrece la empresa.

	Excelente	Bueno	Regular	Malo	Pésimo
Limpieza del loby					
Limpieza de las salas					
Limpieza de los baños					
Comodidad de las butacas					
Sonido					
Proyección					
Servicio en taquillas					
Servicio en dulcería					
Estacionamiento					

En la escala del 1 al 10, ¿qué calificación le dá a este conjunto? ¿Tiene algún comentario o sugerencia para mejorar?

Esquema básico que sigue de un estudio estadístico: Frecuentemente se visualiza a la estadística como la actividad de recopilación, presentación, análisis e interpretación de los datos recabados en una muestra. Sin embargo, ésta puede ser una de las etapas finales de esta materia, la cual va más allá. El

concepto de estadística se refiere a la posibilidad de establecer un diálogo entre lo que se puede percibir del mundo real y lo que se observa a través de los datos. El aprender y comprender sobre el entorno de la administración y la esencia de la economía, parte de ciertas ideas de conocimientos iniciales; luego, mediante la información generada por un adecuado proceso de los datos, esas ideas darán lugar a otra reflexión, y a su vez este resultado motivará a una nueva comunicación con la realidad, lo que al final se convertirá en fuente de sabiduría. Esta idea general se ilustra en la figura 1.8.

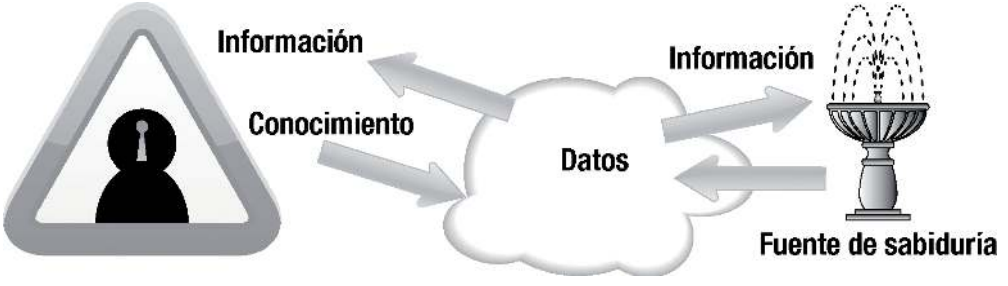


Figura 1.8 Esquema básico que sigue de un estudio estadístico.

1.4 Resumen

Población	Colección de individuos u objetos en los que se observa una característica particular que será objeto de estudio.
Tamaño de la población	Número de miembros de la población. Su símbolo es N.
Censo	Estudio de toda la población.
Muestra	Parte de la población que se estudiará para conocer las características de la población.
Tamaño de la muestra	Número de miembros de la muestra. Su símbolo es n.
Variable	Característica de cada miembro de la población.
Unidad experimental	Entidad más pequeña que es de interés en un estudio estadístico.
Observación	Valor que toma la variable para una unidad experimental.
Datos	Colección de observaciones que toma una o más variables.
Datos cualitativos	Describen una característica particular de una unidad experimental. Con frecuencia, éstos son no numéricos.
Datos cuantitativos	Esencialmente son numéricos.
Datos discretos	Toman sólo ciertos valores, que a menudo son números enteros.

Datos continuos	Toman un número infinito de valores entre dos números, y son valores asociados a un intervalo de la recta numérica.
Datos ordinales	Se crean asignando las unidades experimentales en categorías. Cada categoría tiene una jerarquía y un orden.
Datos nominales	Se crean clasificando en categorías las características de la unidad experimental.
Muestra aleatoria simple	Muestra que ha sido seleccionada de tal manera que todos los miembros de la población tengan la misma oportunidad de ser escogidos.

1.5 Complemento didáctico

Estrategias para recabar información

En el complemento didáctico de este capítulo se presenta una serie de ejercicios de aplicación práctica con la finalidad de que los estudiantes realicen actividades que les permita hacer encuestas, entrevistas, así como aplicar algunas estrategias para recabar información.

1.6 Ejercicios

Elementos básicos en estadística

1.1 Una compañía farmacéutica desarrolla una nueva medicina que se considera eficiente para tratar la calvicie. La firma debe contar con el permiso de las autoridades correspondientes antes que el producto salga al mercado. Ellos deben demostrar la seguridad y efectividad del tratamiento en una muestra de individuos, seleccionados de la población a la que se desea llegar.

1. ¿Cuál es esa población?
2. ¿Cuál podría ser una muestra?
3. De manera intuitiva, ¿cómo se podría evaluar la eficiencia del tratamiento?

1.2 Unos estudiantes desean saber qué tan bueno es el servicio que ofrece la cafetería de la escuela y

están interesados en todos los clientes de la cafetería.

1. ¿Cuáles son los posibles clientes?
2. ¿Qué personas deben incluirse en la población?
3. ¿Cuál es la muestra?
4. ¿Cómo podría evaluar el servicio que ofrece la cafetería?

1.3 La dirección de una escuela desea estudiar la comprensión de la lectura de sus estudiantes antes y después de un programa de capacitación en lectura. Si la escuela tiene 5000 estudiantes, podría ser virtualmente imposible medir la compresión de lectura de todos ellos.

1. ¿Cuál es la población?
2. ¿Cuál podría ser la muestra?

1.4 El secretario académico de una universidad desea conocer el tipo de actividades y trabajo que realizan sus graduados, después de 5 años de haberse recibido.

1. ¿Cuál es la población de interés?
2. Identifique qué razones se pueden considerar para tomar una muestra.
3. Indique dos variables o características de los miembros de esta población que se deben considerar.

1.5 "Estoy tan estresado" es una afirmación frecuente entre los estudiantes. ¿Cuál es su estrés?

1. ¿Cuál es la población de interés?
2. Identifique qué razones se pueden considerar para tomar una muestra.
3. Indique dos variables o características de los miembros de esta población que se deben considerar.

Métodos para la colección de datos

1.6 Realice la encuesta que se propone al final de este ejercicio. Seleccione una muestra de 20 compañeros siguiendo los tres procedimientos propuestos. En el segundo procedimiento recuerde proponer una estrategia de identificación. Considere las siguientes poblaciones:

1. Su grupo
2. Su escuela

Preguntas para realizar la encuesta a las muestras en los incisos a y b.

- Indique cinco habilidades que debe tener un administrador.
- Indique cinco actitudes que debe desarrollar una persona para ser un buen administrador
- ¿Un administrador debe ser innovador? ¿por qué?

¿Qué diferencias hay entre las respuestas de su grupo y de la escuela?

1.7 Repita el ejercicio anterior, pero ahora proponga una encuesta con temas que le gustaría saber de sus compañeros. Coordine este trabajo con su profesor.

1.8 Una estación de radio desea saber qué tan aceptable es su programación. Se selecciona una muestra de 1000 personas de una ciudad en la que se trasmite esa estación. ¿Cómo podría obtener una muestra aleatoria?

1.9 La dirección de una escuela le pide a sus profesores que, a lo largo del ciclo escolar, inculquen en sus alumnos la exclusión de refrescos en su dieta alimenticia y que les expliquen por qué. Al final del semestre se desea conocer la proporción de alumnos que dejaron de consumir refresco. Se elige una muestra de 100 alumnos de la escuela.

1. ¿Qué preguntaría y cuándo?
2. ¿Cómo obtendría una muestra aleatoria?

1.10 Usando el Generador de números aleatorios del paquete estadístico seleccione:

1. 15 números entre 1000 y 0
2. 50 números entre 2500 y 100

Tratamiento de la información

1.11 A continuación se propone como ejercicio que realice una encuesta. Existe una amplia variedad de encuestas y, por lo general, pretenden conocer las características de una población con el fin de obtener información, como podría ser: saber qué opinan las personas sobre algo, distinguir sus preferencias, entre muchas otras. Aunque también se realizan encuestas para resolver problemas, como en la fabricación de productos o errores de producción.

1. Seleccione una muestra de compañeros de su escuela (población) y aplique la encuesta.
2. Identifique el tipo de variables en cada pregunta y describa sus posibles respuestas.

La finalidad es que aplique esta encuesta para que conozca las opiniones sobre este tema y luego pueda usarla en otra parte del curso. Busque en revistas y periódicos artículos o reportajes que contengan información estadística. Describa cinco variables de cada tipo.

1.- Edad					
2.- Sexo	Hombre		Mujer		
3.- Materia que más le gusta					
4.- Materia que menos le gusta					
5.- Materia que le resulta más fácil:					
6.- Materia que le resulta más complicada:					
7. Materia que quitaría:					
8. Materia que pondría:					
9. ¿Cuánto tiempo dedica a estudiar?					

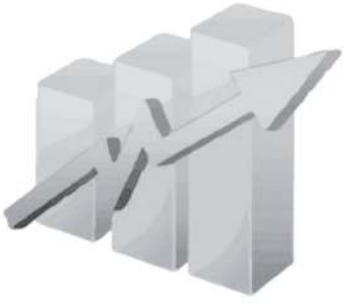
Más de 2 horas diarias:	
2 horas diarias:	
De 1 a 2 horas diarias:	
1 hora o menos diaria:	
Cuando puedo:	
Sólo en época de exámenes:	

El nivel de enseñanza en la escuela a la que asiste es							
Alto		Medio		Bajo			
11. En general, ¿cómo considera la relación profesor alumno?							
Buena		Regular		Mala		No existe	
12.- ¿Lee periódicos todos los días?	Sí		No				
13.- ¿Cuántos libros lee al semestre?							
14.- ¿Qué tipo de programas de TV le gustan más?							

Deportivos		Series policíacas	
Películas		Musicales	
Novelas		Noticieros	
Cómicos		Documentales	

1.7 Evaluación

En la evaluación de este capítulo se plantean siete cuestiones en las que se pide que el alumno muestre su comprensión de los conceptos expuestos.



Capítulo 2

Presentación, organización y descripción de datos

Cualquier persona que quiera desempeñarse adecuadamente en el mundo, debería tener al menos una noción elemental de lo que es la estadística. La estadística tiene injerencia en la práctica profesional de disciplinas tan variadas como la medicina, agronomía, veterinaria, biología, administración, economía, psicología, sociología o finanzas. Incluso, en una amplia gama de artículos de investigación científica, estos no pueden ser aceptados en nuestros días si carecen de técnicas y conceptos estadísticos en sus etapas de planteamiento, diseño y análisis de datos.

Ignacio Méndez Ramírez

Competencia general

Tener habilidad para organizar y presentar de manera lógica la información estadística descriptiva, con un enfoque visual mediante gráficas cuya finalidad sea interpretar y explicar sus ejemplos y proyectos.

Competencias específicas

- Conocer las estrategias necesarias para agrupar los datos que se obtienen de un estudio.
- Aprender a graficar los datos generados en un estudio.
- Conocer que el polígono de frecuencias es otra herramienta útil para describir la distribución numérica de los datos, y aprender a elaborar la gráfica e interpretarla.
- Interpretar datos a partir del polígono de frecuencia acumulado, y de la ojiva.
- Comprender cómo se construye un diagrama de puntos y evaluar su utilidad como herramienta gráfica.
- Elaborar un diagrama para agrupar datos que los represente de manera tabular y gráfica a la vez.
- Comprender la forma en que se hace el análisis descriptivo de los datos usando diferentes tipos de gráficas de manera integral.
- Conocer algunas de las herramientas gráficas para describir datos cualitativos.
- Realizar pequeños estudios o proyectos mediante encuestas o entrevistas para generar datos y analizarlos con el uso de técnicas gráficas de estadística descriptiva.

2.1 Introducción

Se dice que una imagen vale más que mil palabras. En esta unidad se mostrarán los métodos para desplegar las observaciones de una variable con el propósito de evaluar en forma rápida los rasgos principales de los datos. Una vez que se han coleccionado los datos, se les debe dar una mirada general para obtener una percepción de sus características principales y de algún rasgo sorprendente, antes de intentar contestar cualquier pregunta formal. Esta etapa del estudio de los datos generalmente se conoce como la fase exploratoria.

En la siguiente etapa, primero se propone la manera en que los datos deben presentarse, particularmente en tablas; después se muestra el tipo de gráficas que son apropiadas para describir los datos, tales como el histograma, el polígono de frecuencias, el diagrama de puntos, el diagrama de tallo y hoja. En este punto se puede plantear la siguiente pregunta: ¿qué pueden decir las gráficas acerca de la información proporcionada por los datos? Por último, se presentan algunas gráficas para las variables de tipo cualitativo.

Uno de los objetivos de la estadística es obtener información sobre diversos temas, situaciones y fenómenos, entre muchas otras acciones que lleven a descubrir nuevos paradigmas o hechos relevantes. Para ello es necesario organizar esa información de alguna manera. La tabla 2.1 representa un ejemplo que se refiere a categorías de trabajo, que muestra la relación de años de servicio y el salario según la actividad.

Tabla 2.1 Resumen de salarios en la industria para diferentes actividades y años de servicio

Actividad	Años de servicio	Salario
Empleado	22	16,000
Empleado	12	12,500
Asistente	8	15,000
Asistente	5	11,500
Jefe de administración	12	24,000
Jefe de administración	8	19,000
Subgerente	9	25,000
Gerente	7	38,000
Gerente	3	30,500

De acuerdo con este ejemplo es de esperar que al pasar los años, una persona gane más dinero. Si se observa cierta ocupación se pueden plantear varias cuestiones como ¿existe una tendencia explicable mediante un modelo? El poder adquisitivo de diferentes categorías de trabajo ¿contribuye a mantener la economía de una región? ¿Qué factores resultan importantes para sostener una economía en apogeo? ¿Existe una relación entre el salario y la inflación? ¿Existe algún modelo que explique esta relación?

Algunas de estas preguntas se pueden responder mediante la presentación de la información en forma gráfica, y a partir de este enfoque visual darle una interpretación. Estas ideas se plantearán en este capítulo y a lo largo del libro.

Modelo de estudios estadísticos

Se propone estudiar un problema de interés en diferentes áreas del conocimiento, en particular en la administración y en la economía. En situaciones reales, algunos le llaman la estadística en acción, aplicación de la estadística, escenario, casos prácticos, pero nosotros aquí le hemos llamado "el mundo de la información".

El punto de partida es definir la población que será objeto de estudio, la cual es finita, es decir, se pueden contar todas las unidades que la componen. También se presenta el caso de la población hipotética, la cual ocurre porque no se pueden considerar todas las unidades.

Una vez que se ha establecido el problema, surgen preguntas sobre su naturaleza, que están relacionadas con la composición de la población. Por ello, en un gran número de situaciones se plantean hipótesis sobre ciertas características de la población, y como resulta costoso obtener información acerca de ella, se recurre a seleccionar una muestra aleatoria de la población a través de los métodos de muestreo con el fin de ganar conocimiento.

En la figura 2.1 se muestra el modelo del proceso que sigue un estudio estadístico, en particular para el análisis gráfico.

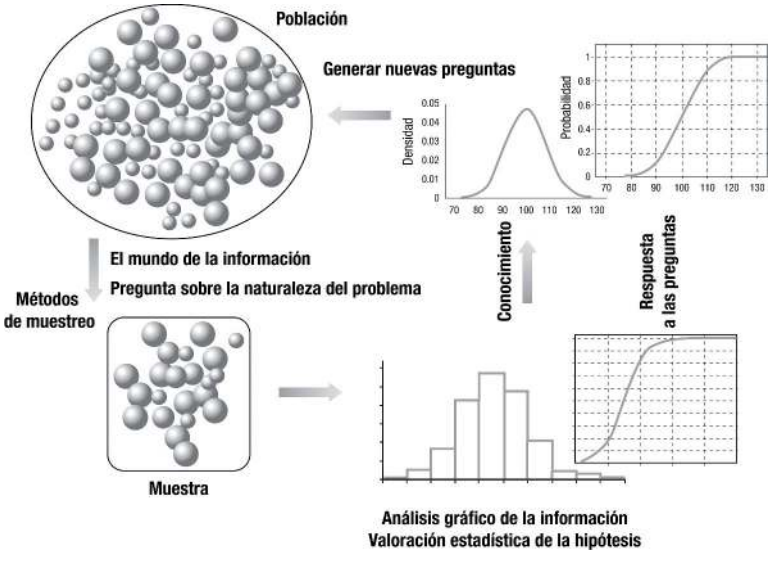


Figura 2.1 Descripción del modelo del procedimiento estadístico para generar conocimiento.

A partir de la información que genera la muestra, se pueden construir modelos gráficos; en particular aquí se presentan un histograma y el polígono de frecuencia acumulado, cuyas figuras describen una o algunas cuestiones planteadas en el problema. La interpretación de estas gráficas que se obtuvieron de la muestra

pretende contestar las preguntas propuestas acerca de la naturaleza del problema. La característica que se estudia en la población se representa mediante una variable aleatoria, que se verá más adelante. Los temas que se van a tratar pertenecen al contexto de la administración y economía. Finalmente, la descripción representada por el histograma y el polígono de frecuencia acumulado, proporciona conocimiento sobre el problema que se investiga y da lugar a elaborar nuevas preguntas.

Muestreo

El muestreo no es una simple sustitución de una cobertura total por una parcial. El muestreo es la ciencia y el arte de controlar y medir la confiabilidad de la información estadística útil a través de la teoría de la probabilidad.

W. E. Deming

2.2 Estadística descriptiva: variables numéricas

2.2.1 Tabla de distribución de frecuencias

El mundo de la información 1. Consumo de energía

La generación de energía resulta costosa para el gobierno federal. Con el fin de ahorrar en ese concepto, se han propuesto diferentes estrategias, entre ellas destacan los mensajes a la población para que economicen energía apagando los focos que no estén utilizando o cambiando los focos convencionales por ahorradores. Existen otras estrategias para ahorrar energía, entre las medidas adoptadas por los gobiernos está el uso del horario de verano.[1]

Preguntas sobre la naturaleza del problema

¿Cómo se puede obtener información sobre el consumo de energía? ¿Cómo puede usar una oficina de gobierno la información disponible sobre el consumo de energía para elaborar planes de ahorro? ¿Cuál es la utilidad de la información del consumo de energía? Varias preguntas más se pueden formular para adquirir conocimientos en torno al consumo de energía. Ahora bien, se sabe que cada bimestre se debe

[1] Para saber más sobre la generación de energía, consulte la página

www.conae.gob.mx/wb/CONAE/CONA_9_desde_el_hogar

pagar el recibo de consumo de energía de nuestra casa. ¿Cuánta energía se ha consumido en los últimos dos meses? En las casas de los compañeros del grupo, ¿se consumió más luz que el bimestre anterior?

Alternativas de energía

En http://energiasalternativasdb.blogspot.mx se hace referencia a una página en la que podrá encontrar ideas sobre alternativas de energía. Desde un punto de vista administrativo y económico es de interés y de utilidad ver ese documento, porque incluye un análisis estadístico en tablas sobre este tipo de opciones energéticas.

Procedimiento para seleccionar los datos

Para contestar algunas de estas preguntas o tener una idea más clara acerca del consumo de energía en los hogares, se requiere de la información correspondiente. Para obtenerla, se pidió a un grupo de 38 alumnos que cada uno de ellos llevara a la clase su último recibo de luz. En la facturación aparece el consumo en KWh. A continuación se muestra la información de los 38 consumos:

299	308	335	330	317	330	327	346	315	320	301	312	320	324
319	314	309	326	314	311	322	325	300	322	312	307	311	334
322	298	308	312	336	314	312	328	305	315	301	327		

La información tal como se muestra es difícil de interpretar. En ese sentido, el objetivo es presentar de manera simplificada el conjunto de los datos obtenidos de los recibos. El procedimiento que se propone para resumir los datos es agruparlos. Esta agrupación consiste en construir una tabla de frecuencias.

La tabla de frecuencias contiene un determinado número de clases de igual tamaño. *Al número de datos que cae dentro del intervalo de clase* se le denomina *frecuencia.* Para elaborar una tabla de este tipo se deben responder las siguientes preguntas: ¿cómo se define el número de clases?, ¿cómo determinamos el ancho o tamaño del intervalo de clase? El ancho del intervalo de clase que se busca tiene dos límites (extremos). Una vez definido lo anterior, ¿cómo se determinan esos límites del intervalo de clase?

El resumen de los datos se presenta en una tabla de frecuencias. Una *tabla de frecuencias* o *distribución de frecuencias* divide los datos en clases de valores o categorías, y registra el número de veces que cada dato ocurre en esa clase. A continuación se ejemplifica el procedimiento que permitirá construirla. Clase es un intervalo en el que se recopilan los datos, por ejemplo el intervalo [298,304) son los datos que son iguales o mayores a 298 y menores a 304. Cada intervalo de clase es mutuamente excluyente, esto es la propiedad de un conjunto de clases que permite incluir a un individuo, objeto o medida en una sola categoría.

Elaboración de la tabla de frecuencias

El procedimiento que permite agrupar los datos se describirá por pasos. Cada uno de estos pasos se

ejemplificará con los 38 datos de consumo de luz. La presentación final de los datos en forma agrupada se describe en la tabla 2.2. El procedimiento es el siguiente.

1. Se determina la lectura mayor (máxima) y la menor (mínima) en los datos, y se calcula la distancia entre estos valores. Así la distancia es:

$$distancia = máximo - mínimo$$

Consideremos los datos de consumo de luz del problema, donde los números mayor y menor son 346 y 298, respectivamente; entonces la distancia es:

$$distancia = máximo - mínimo = 346 - 298 = 48$$

2. Se divide la distancia entre el número k de clases que se deseen, todas de igual ancho.

$$\text{ancho (aproximado del intervalo de clase)} = \frac{distancia}{k}$$

Note que el 48 es un múltiplo de 6 y 8. En este sentido, estos números son candidatos para determinar el *número de clases*. Suponga que se decide utilizar 8 intervalos, por tanto el ancho de los intervalos de clase es:

$$\text{ancho} = \frac{48}{8} = 6$$

El *ancho del intervalo* de clase se determina en función del número de clases que se desee, también se conoce como ancho de clase. Aunque esto resulta arbitrario, casi siempre se proponen entre 5 y 15 intervalos, pero ello depende del número de datos. Las observaciones agrupadas sacrifican información dependiendo de la manera en que las observaciones se distribuyan dentro de cada clase. Por un lado, un número limitado de intervalos (menos de 5) daría muy poca información de la distribución y variación de los datos y, de otra manera, demasiados intervalos podrían provocar que la información esté dispersa.

3. Se establecen los intervalos de clase, para lo cual se requiere obtener el primero de ellos. Se elige el menor de los datos y se suma el valor del ancho. Al valor resultante se le vuelve a sumar el ancho, y así sucesivamente hasta obtener el número de clases establecidos.

En el problema, el valor menor es 298 y a éste se le suma 6. De modo que 298 y 304 es el primer intervalo de clase. Los siguientes intervalos se muestran en la tabla 2.2. Cada intervalo de clase, como se puede observar, presenta dos valores que se pueden superponer. Para evitar que ello ocurra, deben contarse las observaciones iguales o mayores al valor de la izquierda del intervalo y las menores al valor de la derecha. Se denota con X el valor de la observación, por lo que la representación formal de un intervalo de clase es:

$$\text{valor izquierdo} \leq X < \text{valor derecho}$$

Esta última expresión se simplifica mediante la siguiente notación:

valor izquierdo, valor derecho

4. Se cuenta el *número de observaciones* que caen dentro del intervalo de clase. A este número se le conoce como frecuencia. Así, para el problema del consumo de luz, el primer intervalo de clase es:

$$298 \leq X < 304$$

o también,

$$298, 304$$

De los 38 datos de consumo hay 4 números entre estos valores; por lo tanto, *la frecuencia en ese intervalo de clase es 4.* Las demás frecuencias aparecen en la tabla 2.2, a la cual se le conoce como tabla de frecuencias.

La tabla de frecuencias es una clasificación de datos y permite tener una mejor idea acerca de los datos iniciales. Además, los valores de las frecuencias individuales se ven influidos por el tamaño de la muestra, pues cuando las muestras son grandes, las frecuencias individuales serán mayores que cuando son pequeñas. La comparación entre diferentes muestras se vuelve complicada; también se torna difícil de interpretar y comprender la problemática planteada, por ello se recurre al concepto de frecuencia relativa.

La *frecuencia relativa* de una clasificación de datos es el número de veces que una observación cae sobre una clase, y representa una proporción del número total de datos. Por esta razón, la frecuencia relativa se expresa en fracciones, decimales o porcentajes.

El cálculo de la frecuencia relativa se obtiene empleando la expresión:

$$\textit{frecuencia relativa} = \frac{frecuencia}{n}$$

Donde n número total de datos. Por ejemplo, para la clase 1, frecuencia relativa $= \frac{4}{38} = 0.105$
En el problema del consumo de luz, los valores de la frecuencia relativa aparecen en la última columna de la tabla 2.2.[2]

Interpretación

Mediante la tabla de frecuencias puede observarse que en 9 hogares consumen menos de 310 KWh. En

[2]En relación con el consumo de energía, consulte la página

www.conae.gob.mx/wb/CONAE/CONA_9_desde_el_hogar

términos de frecuencia relativa los valores menores a 310 KWh se suman, esto es:

$$0.105 + 0.132 = 0.237$$

Se obtiene una mejor interpretación de esta cantidad, si se expresa en porcentaje, es decir, 23.7 % del total de hogares consume menos de 310 KWh. Podemos subrayar que la mayoría de los hogares consume entre 310 y 334 KWh,

$$0.263 + 0.184 + 0.158 + 0.080 = 0.685$$

Esta cantidad, expresada en porcentaje, representa 68.5 % de las 38 observaciones.

En un número pequeño de hogares se consume más de 334; en este caso:

$$0.053 + 0.026 = 0.079$$

cantidad que no logra llegar a 1 % del total de la muestra.

Así se puede ver que el análisis descriptivo de estos datos, presentados de manera resumida en una tabla, da cierta información sobre el consumo de energía en esos 38 hogares.

Los datos registrados en la tabla 2.2 permiten obtener conclusiones con un margen de error[3] para la población, y contestar a las preguntas planteadas. Suponga que esta muestra de 38 hogares se obtuvo de una población que comprende varios fraccionamientos de interés social. Así en una población de 800 hogares se tendrá que (800)(0.158)=126.4, si se considera el intervalo de clase 5. Es decir, el consumo de luz en 126 hogares estará entre 322 y 328. Recuerde que una muestra se obtiene de una población que tiene una característica determinada.

Tabla 2.2 Tabla de frecuencias para el consumo de energía en 38 hogares.

Clase	Intervalo de clase	Conteo	Frecuencia	Frecuencia relativa
1	298-304	||||	4	0.105
2	304-310	|||||	5	0.132
3	310-316	||||| |||||	10	0.263
4	316-322	||||| ||	7	0.184
5	322-328	||||| |	6	0.158
6	328-334	|||	3	0.080
7	334-340	||	2	0.053
8	340-346	|	1	0.026

Complemento técnico

Existen algunas reglas para establecer el número de clases. Entre esas reglas destacan las dos siguientes:

[3] El concepto de margen de error se expone en el capítulo 7.

1. El número de clases se determina por:
$$k = \sqrt{n}$$
donde n es el número de datos.

2. También puede utilizarse la fórmula de Sturges:
$$k = 1 + 3.322 \log_{10}(n)$$
donde n es el número de datos.

   Observe que al aplicar estas expresiones matemáticas a los 38 datos del consumo de energía se tiene que:
$$\sqrt{38} = 6.16$$
y para la segunda:
$$k = 1 + 3.322 \log_{10}(38) = 6.248$$
Ambas cantidades se redondean al entero más próximo, que en este caso es 6. Para el caso del ejemplo, en lugar de 8 intervalos se obtendrían sólo 6 intervalos.

3. Regla empírica para determinar el número de clases.

En ocasiones, los datos extremos resultan engorrosos debido a que no terminan en un número fácil de manejar y porque generan intervalos de clase que dificultan el uso de la información. La solución que conviene adoptar en estas situaciones es redondear los números extremos a valores que no generen decimales. De esa manera, pueden aproximarse los extremos a números terminados en 0 o 5. Un ejemplo para mostrar esta situación se verá en el apartado del histograma.

Ejemplo 2.1

En la actualidad, con el objeto de financiarse su carrera, muchos estudiantes laboran durante sus horas libres en diferentes tipos de trabajos. Se tomó una muestra de 30 jóvenes y se les preguntó el salario que perciben a la semana. Los datos reportados en el estudio se describen en la tabla de abajo, con ellos construya la tabla de frecuencias.

2740	1810	2610	1760	1110	1710	2390	1150	1500	1050
4190	2100	2460	1070	850	2020	2110	1560	1970	510
2240	2370	1450	1090	860	3470	3920	1680	1350	1860

Solución operativa clásica

El valor máximo y el mínimo en este conjunto de datos es de 4190 y 510, respectivamente. Para construir los intervalos de clase primero se calcula la distancia de estos dos valores:

$$máximo - mínimo = 4190 - 510 = 3680$$

Dado que el número de datos es pequeño, se proponen *cinco intervalos de clase*; así:

$$ancho = \frac{3680}{5} = 736$$

Entonces la tabla de frecuencias para este ejemplo se presenta en la tabla 2.3.

Tabla 2.3 Tabla de frecuencias para el salario semanal que perciben los estudiantes.

Clase	Intervalo de clase	Frecuencia	Frecuencia relativa
1	510-1246	8	0.267
2	1246-1982	10	0.333
3	1982-2718	8	0.267
4	2718-3454	1	0.033
5	3454-4190	3	0.100

Interpretación: Sólo 0.133 **gana más** de 2718 pesos, lo que equivale a 13.3 %. Quienes ganan entre 1246 y 2718 pesos se ubican en 0.6, esto es, 60 % de los estudiantes, lo que corresponde a la mayoría. Por último, 27 % gana menos de 1246 pesos.

Solución mediante el uso de CalEst

La solución anterior puede obtenerse usando el programa de estadística y probabilidad CalEst en la opción de Gráficas. La imagen del módulo Gráficas se describe en la figura 2.2, donde la alternativa para la tabla de frecuencias se encuentra al final. Al aplicar la opción se tiene un cuadro como el que se observa a la derecha de la figura, en Datos se señala la columna, nombre de la variable, que contiene el conjunto de datos a estudiar, luego se escribe el Número de intervalos que se desea, y a continuación se escoge la Frecuencia relativa, o la segunda, si los datos aparecen como porcentajes. Por último, se da aceptar y aparece la tabla de frecuencias de la figura 2.3.

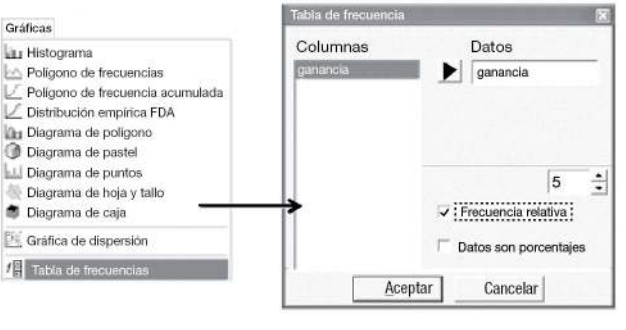


Figura 2.2 Descripción del contenido de la opción de Gráficas del conjunto módulos en **CalEst**.

La figura 2.3 contiene el número de clases, los valores principales de cada clase, en seguida se muestra la tabla de frecuencias en que se señalan tanto las frecuencias como las *frecuencias relativas* (Frec. relativa), además de las columnas de *frecuencia acumulada* y *frecuencia relativa acumulada*. Ambos tipos de frecuencias se verán más adelante.

Clases	Mayor que	Punto intermedio	Menor que	Frecuencia	Frec. relativa	Frec. acumulada	Frec. rel. acum.
1	160.00000	563.00000	966.00000	4.00000	0.1333	4.00000	0.1333
2	966.00000	1369.00000	1772.00000	11.00000	0.3666	15.00000	0.5
3	1772.00000	2175.00000	2578.00000	10.00000	0.3333	25.00000	0.8333
4	2578.00000	2981.00000	3384.00000	2.00000	0.0666	27.00000	0.8999
5	3384.00000	3787.00000	4190.00000	3.00000	0.1	30.00000	0.9999

Figura 2.3 Descripción de la tabla de frecuencia generada por **CalEst**.

2.2.2 Tabla de distribución de frecuencias e histograma

El mundo de la información 2. Tiempo de juego

La evolución tecnológica ha puesto al alcance de las personas una gran cantidad de pasatiempos, entre los que se encuentran los juegos para equipos electrónicos. En muchas ocasiones, esta proliferación de juegos ha dado lugar a la relajación. Sin embargo, existe una importante cantidad de juegos educativos, creativos y de habilidades motoras para los jugadores. Algunos de estos juegos recreativos requieren de un tiempo límite para completarse, el cual depende de las habilidades de cada persona.

Preguntas sobre la naturaleza del problema

¿Por qué es importante que una persona desarrolle este tipo de habilidades? Una persona con destreza en juegos educativos ¿desempeñará mejor una tarea práctica en su desarrollo profesional? Un psicólogo investiga el desarrollo de ciertas habilidades de un grupo de individuos cuyas edades oscilan entre los 18 y los 24 años de edad. En la parte inicial de su investigación, la intención del psicólogo es conocer el tiempo que tardan 50 individuos en terminar un juego. El investigador toma como referencia un juego diseñado para computadoras personales, aunque dicho juego también puede encontrarse en versiones para otros equipos electrónicos, como celulares, agendas electrónicas y computadoras portátiles. Los datos registrados son:

11.52	9.25	7.21	9.01	10.75	8.43	8.58	9.03
10.26	7.53	12.13	9.37	9.55	10.01	9.21	10.11
10.42	9.56	11.33	9.24	8.06	8.45	9.50	11.18
10.12	11.07	11.05	8.46	12.10	7.14	11.39	8 .09
11.47	9.06	9.22	9.20	7.19	10.10	10.12	9.26
15.21	8.36	10.34	8.26	9.58	10.55	12.40	7.35
13.42	7.11						

Tabla de frecuencias

Como se ha visto, la tabla de frecuencias es una manera de presentar un resumen de los datos. Al construir la tabla de este problema se aplicó la regla empírica; entonces el ancho de clase es:

$$\frac{16-7}{9} = 1.0$$

Para que el cálculo de los intervalos de clase sea más fácil de manejar, el valor del ancho se aproximará a 1. Luego se fija, como valor izquierdo en la clase 1, el valor de 7. Nótese que 7 no es el mínimo, ni 16 el máximo. A partir de ese valor se construye la tabla. Sin embargo, también pudo haberse tomado el 6.5 como valor inicial izquierdo. La tabla de frecuencias de los tiempos para finalizar un juego se aprecia en la tabla 2.4.

Aparte de este resumen estadístico, el psicólogo quiere tener una información gráfica de los datos. El *histograma* es una herramienta que satisface este deseo.

Tabla 2.4 Frecuencias del tiempo necesario para finalizar un juego.

Clase	Intervalo de clase	Frecuencia	Frecuencia relativa
1	7-8	6	0.12
2	8-9	8	0.16
3	9-10	14	0.28
4	10-11	10	0.20
5	11-12	7	0.14
6	12-13	3	0.06
7	13-14	1	0.02
8	14-15	0	0.00
9	15-16	1	0.02
Total		50	1

Construcción del histograma

El *histograma* es una gráfica de barras que consta básicamente de un conjunto de rectángulos. Su forma está determinada por tres elementos: *el número de rectángulos, el ancho y la altura de éstos*. En esencia, el histograma es una representación visual de la tabla de frecuencias. En ese sentido, el número de barras, rectángulos, corresponde al número de clases. El ancho del rectángulo corresponde al intervalo de clase y la altura es la frecuencia. Por ello, suele decirse que el histograma es el complemento gráfico de la tabla de frecuencias. Para elaborar el histograma primero se debe construir la tabla de frecuencias del conjunto de datos de una muestra, tal y como se indica en el apartado anterior, en la tabla 2.4.

Se usa la información descrita en la tabla 2.4 para construir su representación gráfica. La columna 3 de dicha tabla representa la frecuencia, y en la gráfica izquierda de la figura 2.4 se muestra el *histograma de frecuencias*. Se puede observar que el intervalo de clase, señalado en el eje horizontal, corresponde al ancho del rectángulo, y la frecuencia, marcada en el eje vertical, a la altura.

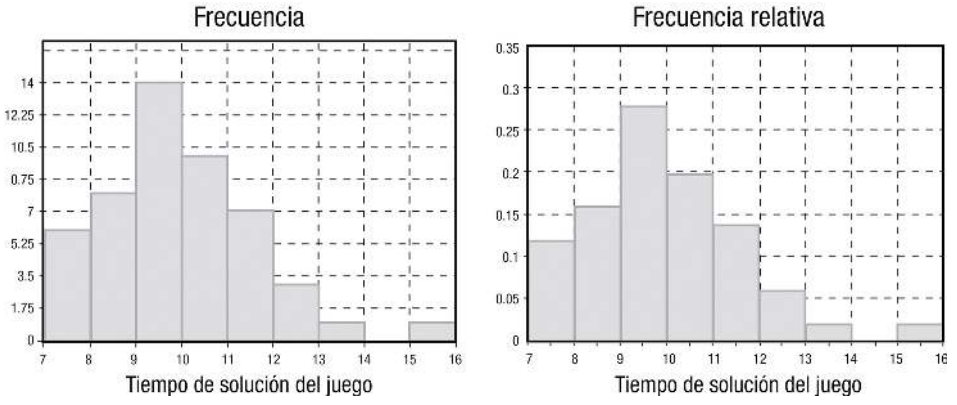


Figura 2.4 Histogramas de frecuencias y frecuencias relativas de tiempos para finalizar un juego.

Si en el eje vertical del histograma se cambia la *marca de frecuencia* por el de la *frecuencia relativa*, se tiene el que se conoce como *histograma de frecuencia relativa*, el cual se presenta en la gráfica que aparece a la derecha en la figura 2.4.

Dado que todas las barras de un histograma tienen el mismo ancho, el área de una barra es proporcional a la frecuencia relativa de la clase correspondiente. Por ejemplo, si 25 % del área bajo la distribución queda sobre cierto intervalo, entonces 25 % de las observaciones caen en ese intervalo. En ese sentido, el área total de las barras es igual a 1.

La *frecuencia relativa* en un intervalo de clase es la proporción del número total de observaciones que caen dentro de ese intervalo de clase, y es proporcional al área de la barra correspondiente a ese intervalo.

Ambos histogramas son idénticos pero su escala es diferente. El de la izquierda representa la frecuencia, y también hace referencia a la frecuencia absoluta. El histograma de la derecha, en cambio, describe la frecuencia relativa; esto es importante porque facilita la interpretación de los datos, como se indica en el siguiente párrafo.

Interpretación

Se puede observar que los participantes que terminaron el juego en menos de 9 minutos caen dentro de las frecuencias de 8 y 6, y la suma de estas frecuencias es de 14, lo que equivale a la suma de las frecuencias relativas que, en este caso, es igual a 0.28. Este resultado, expresado en porcentaje, corresponde a 28 % de la muestra. Las personas cuya habilidad para terminar el juego es mayor a 9 minutos, se ubican en 72 %. Observe que la mayoría logra finalizar el juego entre 8 y 12 minutos. Por último, unos cuantos participantes tardan más de 12 minutos y, en consecuencia, son los menos hábiles.

Ejemplo 2.2

Por lo general, las administraciones de gobierno controlan los precios de los productos llamados básicos. En particular el de las tortillas de maíz, se realizó un muestreo en expendios de tortillas en 56 ciudades

de México. Los datos se escriben en la tabla siguiente.

12.33	16.14	13.55	13.75	14.25	14.00	12.10	12.30	14.00	12.00	12.00	14.33
13.22	11.13	10.90	11.63	12.00	14.00	11.50	10.50	14.00	13.00	11.50	11.69
11.80	10.57	11.86	14.67	14.00	13.94	13.89	12.33	10.85	8.80	14.00	15.00
14.00	11.67	14.00	14.00	16.50	15.00	15.00	14.20	13.00	15.00	13.50	13.25
12.33	10.00	13.50	15.00	13.75	10.25	13.83	11.25				

Usando **CalEst**, elabore el histograma. Primero capture la información. Una vez que se adquiere la habilidad para elaborar tablas de frecuencia e histogramas, sólo será importante que se planteen preguntas sobre temas y problemas reales, además de interpretar los datos mediante estas técnicas. ¿Qué tanto discrepan estos precios del oficial? En los últimos 10 años, ¿el precio de las tortillas y el salario mínimo son compatibles?

Solución mediante el uso de CalEst

Una vez capturados los datos, se aplica la opción Graficar y se escoge el tipo de gráfica que se necesite. El CalEst cuenta con las básicas. En particular, el *histograma* que se genera mediante el paquete se presenta en la figura 2.5.

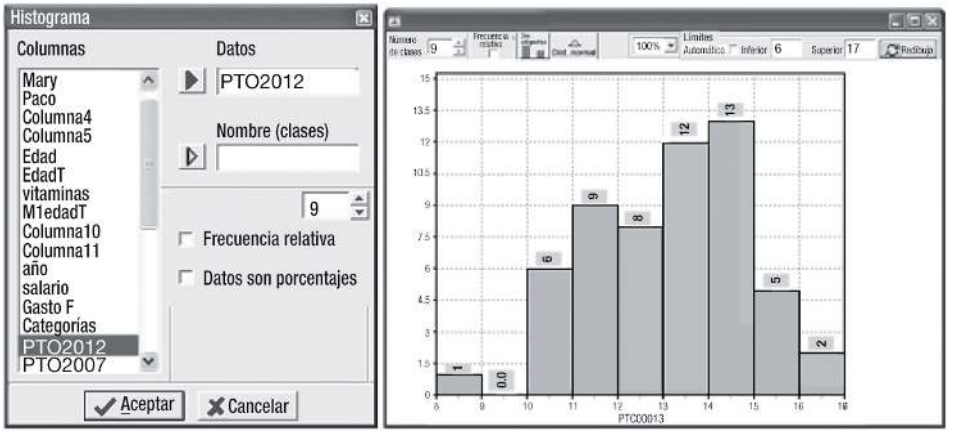


Figura 2.5 Histograma generado con la opción Gráficas de **CalEst**.

2.2.3 Polígono de frecuencias

Para elaborar el polígono de frecuencias se pueden aplicar los procedimientos aprendidos en las actividades de aprendizaje del apartado anterior.

El mundo de la información 3. Condición física

Con el propósito de obtener una evaluación global del sistema educativo se busca un índice que señale la condición física de los alumnos cuyas edades están entre 11 y 13 años. Para alcanzar ese objetivo es necesario conocer datos sobre medidas antropométricas, pruebas de velocidad, agilidad, fuerza, flexibilidad, resistencia y habilidad. Se espera que el resultado, tanto de las pruebas físicas como de las de habilidad, esté relacionado con las medidas antropométricas. Por ello, primero se describirá la distribución de una muestra de peso de los alumnos que participan en el estudio.

Preguntas sobre la naturaleza del problema

Un problema de actualidad en la salud en los niños es la obesidad. Desde luego este tipo de problema tiene un enfoque económico, por las repercusiones que genera en la inversión de gastos médicos, entre muchos otros. Así los administradores tienen una actividad relevante en realizar proyectos para resolver la cuestión de la obesidad. El control del peso es un medio para estar al pendiente de que los niños no rebasen un límite. ¿Cuál es la forma de la distribución de los pesos de esta muestra de niños? ¿Qué porcentaje de niños supera los 55 (kg)? ¿A partir de qué peso los niños pierden habilidad?

Los pesos en kilogramos (kg) observados en 120 niños son:

48	47	43	45	47	55	42	54	46	66
51	54	37	34	44	47	48	33	36	54
37	43	48	60	42	45	58	38	52	52
48	52	35	45	54	39	42	48	45	40
44	39	30	37	57	28	47	48	44	46
56	52	44	50	62	30	45	47	44	55
43	40	61	42	40	57	40	49	35	40
39	37	39	48	42	58	36	53	43	27
64	40	35	43	56	39	47	54	65	44
47	50	54	41	45	37	44	46	61	33
47	52	48	56	59	60	49	37	52	47
42	33	43	43	31	37	41	46	42	37

Tabla 2.5 Frecuencia para peso (en kg) de 120 niños.

Clase	Intervalo de clase	Frecuencia	Frecuencia acumulada	Frecuencia relativa	Frecuencia relativa acumulada
1	27-31	4	4	0.0333	0.0333
2	31-35	5	9	0.0416	0.0749
3	35-39	14	23	0.1166	0.1916
4	39-43	20	43	0.1666	0.3583
5	43-47	24	67	0.2000	0.5583
6	47-51	21	88	0.1749	0.7333
7	51-55	14	102	0.1166	0.8500
8	55-59	9	111	0.0749	0.9250
9	59-63	6	117	0.0500	0.9750
10	63-67	3	120	0.0250	1.0000

El polígono de frecuencias es otra herramienta útil para describir la distribución de los datos, y su construcción depende del histograma. A partir de la experiencia ganada en los problemas anteriores se puede construir la tabla de frecuencias, tabla 2.5, y el histograma de frecuencias, figura 2.6. El histograma se convierte fácilmente en un polígono de frecuencias uniendo mediante líneas rectas las alturas de las barras del histograma que corresponden al punto medio del intervalo de clase, conocida también como marca de clase.

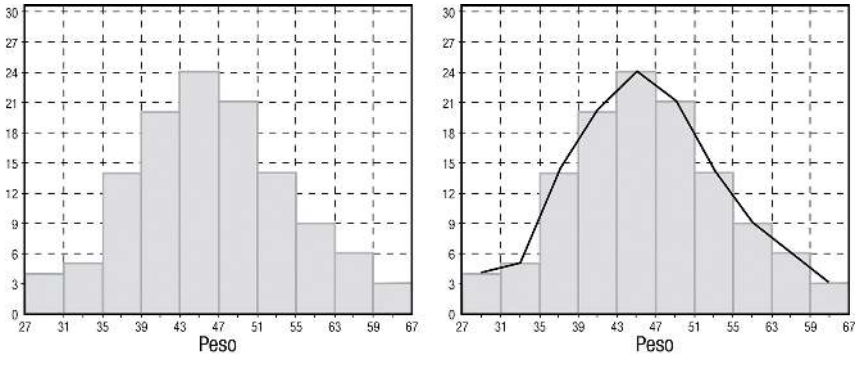


Figura 2.6 Histograma, a la izquierda, e histograma y polígono de frecuencia sobrepuestos, a la derecha.

Modelos estadísticos

El *polígono de frecuencias* que resulta de modelar el contorno del histograma se conoce como el *modelo empírico*. Éste da la idea de la forma aproximada que tiene la distribución, simétrica, sesgada a la

derecha o a la izquierda. En el estudio de la condición física -peso-, la distribución tiene una forma aproximadamente simétrica y permite tener una noción de cómo podría ser la distribución de la población que es objeto de estudio.

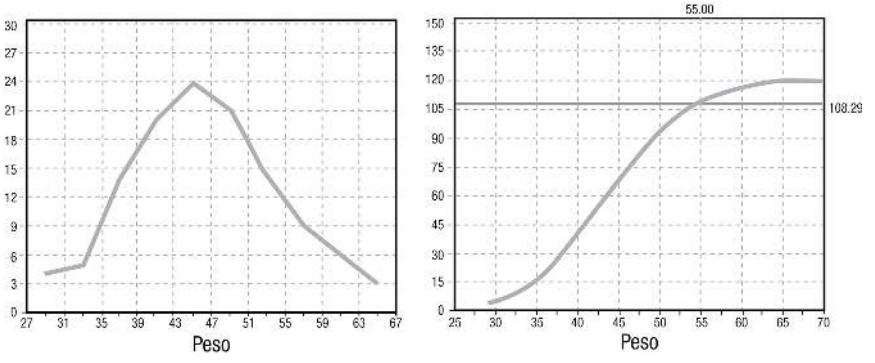


Figura 2.7 Polígonos de frecuencias, a la izquierda, y frecuencias acumuladas, a la derecha.

Otro modelo que resulta de interés para la interpretación es el de frecuencias acumuladas, el cual se obtiene del contorno que deja el ir sumando las barras del histograma. El modelo resultante es el *polígono de frecuencias acumuladas,* observe la cuarta columna de la tabla 2.5. En las gráficas de la figura 2.7 se presentan el modelo empírico y el de frecuencias acumuladas, en particular para este último se indica de manera aproximada a qué dato de los 120 corresponden el peso 55. Una mejor interpretación en la práctica es cuando se tienen las gráficas relacionadas con la frecuencia relativa, lo que se verá a continuación, en el siguiente apartado.

Polígono de frecuencia relativa

Sin embargo otra gráfica que puede ayudarnos a representar la tabla de frecuencias es el polígono de *frecuencias relativas.* Éste se construye tomando como referencia el *histograma frecuencias relativas.* Su procedimiento de construcción es similar al de polígono de frecuencias.

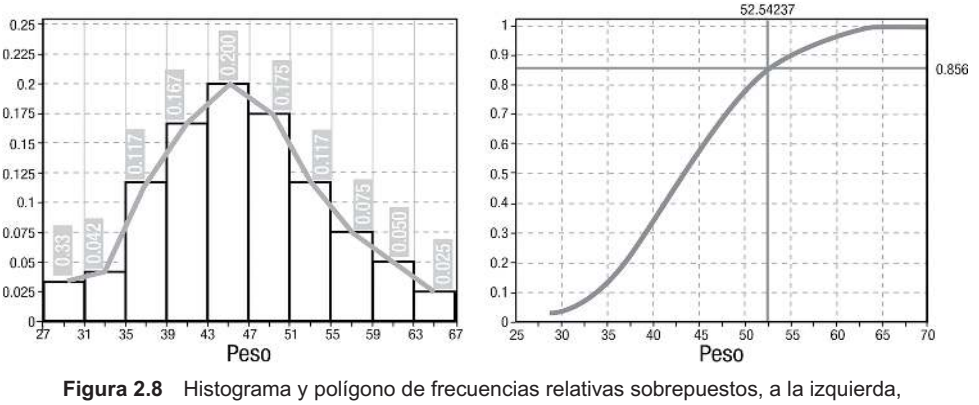


Figura 2.8 Histograma y polígono de frecuencias relativas sobrepuestos, a la izquierda, y distribución acumulada, a la derecha.

La primera y última líneas que representan el polígono de frecuencias inicia en los intervalos de clase adyacentes a la primera y última clases. Es claro que en esos la frecuencia es cero. Así, la primera línea irá de 0 a 0.0333; la segunda, de 0.0333 a 0.0417, y así sucesivamente, situación que se muestra en la gráfica de la izquierda en la figura 2.8.

Como se observa, el polígono de frecuencias relativas tiene en común con el histograma que las áreas de las gráficas sobre un intervalo son idénticas. Éstas se representan en la figura 2.8. A la gráfica sobrepuesta en el histograma, como se indicó antes, se le conoce como *densidad empírica de la distribución de datos.*

La interpretación que se da a la información proporcionada por el peso de los niños es similar a la discutida en el caso del histograma.

En resumen, el polígono de frecuencias es una descripción pictórica que permite captar la forma del histograma. Como se verá más adelante, esta descripción gráfica es una herramienta muy útil para delinear varias características de la distribución de los datos, como son la simetría, el sesgo y la variabilidad de los datos.

Frecuencia relativa acumulada

En los datos que se presentan en la tabla 2.5 se puede observar que están ordenados de menor a mayor. En ese contexto, se pueden sumar las frecuencias relativas y, ante esa situación, pueden plantearse algunas preguntas que resultan interesantes para realizar una interpretación de una muestra de datos. Por ejemplo, podríamos preguntarnos ¿a qué valor de los datos le corresponde 75 %? En particular, en la figura 2.8 se presenta el caso para el 85.6 %. Así, 85.6 % de niños estudiados pesan 52.54 kilos o menos.

Con la información que se proporciona en la tabla 2.5, se facilita la construcción del *histograma de frecuencias acumuladas y polígono de frecuencias relativas en forma acumulada.* Véase la última columna de la tabla 2.5, donde se observa cómo se han utilizado estos datos para elaborar las gráficas de la figura 2.8. Ahí se muestra el histograma de frecuencias relativas donde se ha ido sumando el tamaño de las barras en cada una de las clases. La otra gráfica muestra el polígono que describe las frecuencias relativas acumuladas; éste es el que más se utiliza en la práctica.

El *polígono de frecuencias relativas acumuladas* también recibe el nombre de *ojiva.* Así, cuando le pidan construir la ojiva de un conjunto de datos, se le pide trazar el polígono de frecuencias relativas acumuladas. Véase la figura 2.9.

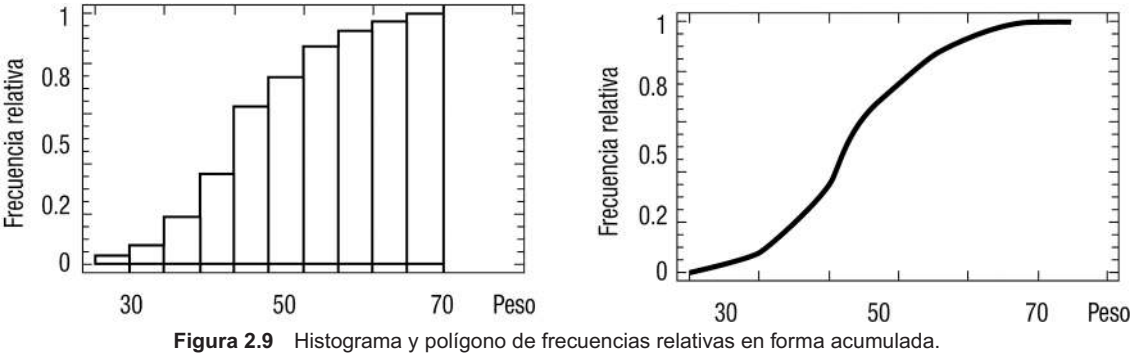


Figura 2.9 Histograma y polígono de frecuencias relativas en forma acumulada.

La frecuencia relativa acumulada de una clase C_i es la suma de las frecuencias relativas de todas las clases anteriores a C_i. Ésta se expresa como fracción, decimal o porcentaje.

Complemento técnico

La gráfica de frecuencias relativas acumuladas es útil para identificar el valor de la variable de interés al que corresponde un porcentaje. Los valores típicos, expresados en porcentaje, a los que se recurre en una gráfica de este tipo son 1 %, 5 %, 10 %, 25 %, 50 %, 75 %, 90 %, 95 %, 99 %. En el eje horizontal de la gráfica de la figura 2.10 se indica los pesos de los niños. En este caso nos interesa conocer a qué pesos equivalen 25 %, 50 % y 75 %, del eje vertical. Los pesos relativos a esos porcentajes son: 40.0, 45.0 y 52.0, respectivamente.

No sólo se pueden obtener los porcentajes señalados, sino lo que se busca es obtener el valor para cualquier percentil. A continuación se define el percentil.

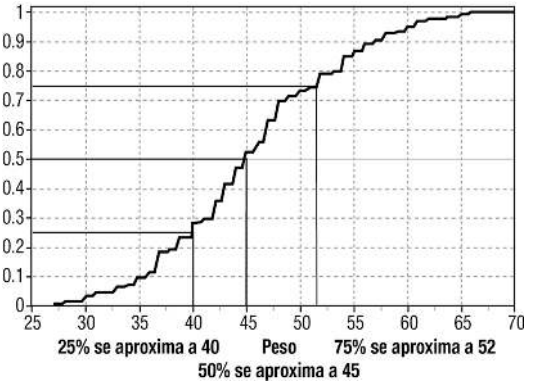


Figura 2.10 Polígono de frecuencias relativas en forma acumulada, aquí se resaltan los porcentajes 25 %, 50 % y 75 %, así como los pesos correspondientes.

El 100p-ésimo percentil de un conjunto de datos es un número en el eje vertical del polígono de frecuencia relativa y el valor correspondiente en el eje horizontal es el 100p % de los datos. Este valor corresponde al área debajo de la distribución de frecuencia relativa para los datos que quedan a la izquierda de este percentil.

2.2.4 Diagrama de puntos

El diagrama de puntos es otra herramienta gráfica para representar la distribución de datos. En el siguiente caso se aplica este criterio al analizar la información asociada con la resistencia de una bobina.

El mundo de la información 4. Resistencia de una bobina

La administración y la economía desempeñan un papel muy importante en la industria, inclusive en varios centros de educación superior existen licenciaturas en esa dirección. Se plantea un problema que

es de interés en la industria, tal que la calidad de este producto impacta en la economía de una empresa y en general en el mercado. Éste consiste en determinar la resistencia en ohms de las bobinas. Una bobina está formada por un alambre conductor con el que se han hecho espiras a manera de resorte. Si se aplica corriente continua (corriente que no varía con el tiempo) a un inductor, éste se comporta como un corto circuito y dejará pasar la corriente a través de ella sin ninguna oposición.

Preguntas sobre la naturaleza del problema

Se busca conocer la resistencia en una producción de bobinas que se emplearán para unos transformadores. ¿Qué tan eficiente es la resistencia de una bobina?

Este problema pretende determinar la distribución de la resistencia y averiguar si existen grupos de puntos, huecos y valores extremos. Para lograrlo se toma una muestra de 100 bobinas de un lote de producción.

Los datos de resistencia registrados para las 100 bobinas son los siguientes:

3.35	3.31	3.40	3.35	3.37	3.31	3.31	3.37	3.33	3.33
3.33	3.36	3.38	3.40	3.36	3.35	3.35	3.33	3.27	3.38
3.33	3.36	3.30	3.33	3.34	3.37	3.34	3.36	3.41	3.33
3.35	3.28	3.31	3.38	3.33	3.37	3.33	3.28	3.40	3.32
3.38	3.36	3.33	3.39	3.36	3.34	3.30	3.37	3.33	3.28
3.30	3.32	3.33	3.37	3.31	3.32	3.30	3.31	3.34	3.38
3.37	3.34	3.32	3.34	3.38	3.35	3.30	3.33	3.25	3.31
3.34	3.30	3.33	3.31	3.34	3.31	3.33	3.31	3.33	3.34
3.30	3.38	3.30	3.38	3.36	3.35	3.42	3.33	3.33	3.31
3.40	3.35	3.36	3.34	3.31	3.40	3.32	3.36	3.32	3.40

Con el objetivo de ejemplificar la distribución de los datos en este problema se recurre a un *diagrama de puntos*. Este es un tipo de gráfica sencilla, mediante la cual se puede representar un conjunto de datos en una escala de números. Para el ejemplo se ha definido la escala en puntos individuales; en este caso se incrementan los números en una centésima, es decir, 3.23, 3.24, y así sucesivamente, hasta llegar a 3.47 (véase la gráfica derecha en la figura 2.11). En la figura 2.11 se ha puesto de manera alternativa un histograma para comparar un histograma con el diagrama de puntos.

La conclusión a que se llega con esta representación gráfica es que la mayor parte de los datos se encuentra entre las resistencias 3.29 y 3.38. Asimismo, se observa un punto extremo en 3.25. En el contexto del proceso industrial puede ser una bobina que se deseche por tener baja resistencia. También destacan dos patrones de puntos: el primero entre 3.27 y 3.28; el segundo en las resistencias 3.39 y 3.42. En un proceso industrial real, ambas situaciones se interpretan en relación con la calidad de la bobinas.

Este tipo de gráfico es útil cuando el conjunto de datos es pequeño; sin embargo, en este tipo de casos

(cuando el conjunto de datos es pequeño) el histograma resulta inadecuado.

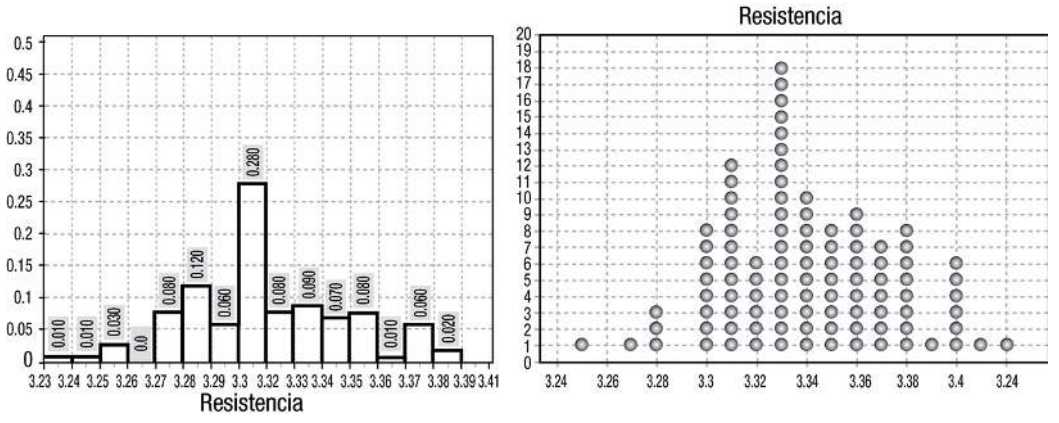


Figura 2.11 Diagrama de puntos para la resistencia eléctrica.

Un *diagrama de puntos* es una gráfica en que cada observación se señala en el eje horizontal. Este eje se marca de tal forma que cada punto se localice de manera única en el eje. El eje vertical sirve como referencia para indicar el valor más grande de la frecuencia.

2.2.5 Diagrama de tallo y hoja

Naturaleza del problema. Índice de contaminación

La contaminación es un tema de actualidad por los efectos que trae consigo. En el ámbito mundial existe una preocupación latente sobre la problemática de la contaminación. Por ello, varios países han firmado el tratado de Kioto. Este problema constituye una línea que comprende la administración y la economía, en esa dirección se genera y se plantean la convocatoria de proyectos derivados hacia este entorno. El estudio y la solución de problemas relacionados con el medio ambiente reciben apoyo económico para su realización. En ellos participan técnicos e investigadores de diferentes disciplinas.

De los varios tipos de contaminación que existen, aquí nos referiremos sólo a la contaminación del aire, la cual guarda relación con los gases, las pequeñas gotas y partículas que flotan en el ambiente y que reducen la calidad del aire.

El aire contaminado no es un problema exclusivo de las ciudades, sino que también puede encontrarse en el campo. En los centros urbanos, los causantes de la contaminación del aire son los autos, autobuses, aviones, así como la industria y las construcciones. Por su parte, en el campo la contaminación pueden causarla, entre otros factores, el polvo que levantan tanto los tractores (cuando aran la tierra) como los camiones y los autos que se manejan en caminos de terracería o grava; también la explotación de las minas, el humo de la madera y las quemas de pastizales o de basura son fuentes de contaminación.

El ozono al nivel de la tierra es el compuesto que causa la mayor parte de la contaminación en casi todas las grandes ciudades. Este tipo de ozono se forma cuando los gases provenientes de motores que

ya han sido expuestos al ambiente reaccionan con la luz solar. Los niveles de ozono en las ciudades se incrementan cuando el aire se encuentra estancado, el sol brilla fuertemente y las temperaturas son más altas.

En resumen, el ozono (de fórmula O_3) es un gas irritante, incoloro, tóxico y muy oxidante; es un alótropo del oxígeno molecular. Cuando es producto de las reacciones fotoquímicas que ocurren en la troposfera, es un contaminante secundario de importancia.

El mundo de la información 5. Índice de contaminación

En las ciudades existen estaciones para medir varios contaminantes, sobre todo los índices de ozono. A partir de esa información, las autoridades encargadas del ambiente toman las medidas pertinentes. En el histograma de la figura 2.12 se muestra el reporte de una estación ubicada en el norte de una gran ciudad durante 50 días.

Preguntas sobre la naturaleza del problema

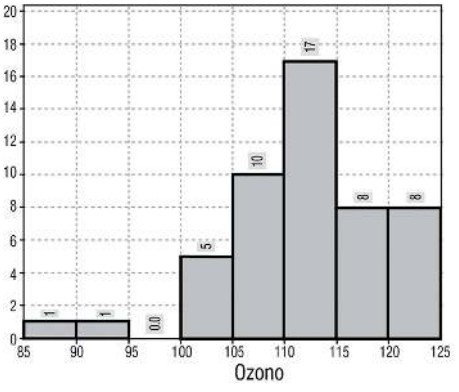
Sobre esta problemática se dio una explicación en la introducción, sin embargo es de interés mantener los índices de contaminación bajo cierta norma.[4] En este caso, ¿se cumple la norma establecida? Los datos observados en esa estación son:

105	104	110	121	122	117	101	111	93	119
108	114	118	111	122	111	108	116	120	110
102	110	124	124	89	107	114	119	110	112
112	110	108	116	120	114	110	101	112	103
108	119	105	108	107	114	113	113	124	106

El *diagrama conocido como tallo y hoja* es de mucha utilidad para exhibir la distribución de los datos. Se utiliza ampliamente cuando el conjunto de datos es pequeño. Una ventaja de los diagramas de tallo y hoja sobre las gráficas de puntos es que se conservan los datos originales, es decir, se reconstruyen los valores originales de los datos. Se puede considerar que el diagrama de tallo y hoja representa, en una sola gráfica, tanto a la tabla de frecuencias como al histograma, lo cual es una gran ventaja.

Un *diagrama de tallo y hoja* tiene una forma como la que se muestra en la gráfica derecha de la figura 2.12. El diagrama de esta figura en particular representa la distribución de los datos del ozono. El siguiente paso será describir el procedimiento para construir una imagen de los datos.

[4]Para conocer más acerca de este tratado puede verse la página:
http://archivo.greenpeace.org/Clima/Prokioto.htm



Tallo	Hoja	Frecuencia
8	9	1
9	3	1
9		0
10	11234	5
10	5567788888	10
11	00000011122234444	17
11	66678999	8
12	00122444	8

Figura 2.12 Histograma comparación con el diagrama de tallo y hoja para el ozono.

Construcción del diagrama de tallo y hoja

Ilustraremos la construcción del diagrama de tallo y hoja considerando los nueve datos correspondientes a la muestra de la producción de maíz en un área de 100 m^2. El peso, en kg., del maíz para cada muestra es: 312, 324, 310, 314, 322, 328, 316, 314, 324. Antes de elaborar este diagrama, se dará una idea general de cómo se forman el tallo y la hoja.

La idea principal es partir cada número para formar primero el tallo y luego la hoja. En general, donde se hace la partición depende del rango de los datos. Para fijar ideas considere el número 310, el cual se parte en dos cifras, 31 y 0. El 31 representará al tallo y el 0 a la hoja. En la representación se escribe el 31 a la derecha separándolo con un espacio, que aquí se resalta con una línea vertical, y el 0 del lado izquierdo. Esto es:

$$31 \mid 0$$

El procedimiento para formar el diagrama de tallo y hoja se indica en los siguientes cuatro pasos:

1. Se ordenan los datos de menor a mayor:

$$310, 312, 314, 314, 316, 322, 324, 324, 328$$

2. Se consideran todos los datos de manera integral. Los números candidatos para el tallo en estos datos son el 31 (para los números 310, 312, 314, 316) y 32 (para los números 322, 324, 328); así

$$\begin{array}{c|} 31 \\ 32 \end{array}$$

3. Se incorpora la otra parte del número partido; así, se anexan 0, 2, 4, 4 y 6, que son la parte correspondiente del 31, o sea:

$$\begin{array}{c|c|} 31 & 02446 \\ 32 & \end{array}$$

4. Se agrega la parte dividida correspondiente al número 32; así la gráfica queda como sigue:

31	02446
32	2448

Cabe observar que en este ejemplo, el ancho de los intervalos de clase es de 10 unidades (de 310 a 320). Además podemos resaltar que la realización del paso 1 facilita la construcción del diagrama.

Supongamos que el ancho de intervalos sea de tan sólo 5 unidades (310 a 315, 315 a 320, y así sucesivamente). Para construir el diagrama de tallo y hoja con 5 unidades, se pone el 31 dos veces en el tallo. Así, en la hoja del primer 31 del tallo, se consideran los números menores que 5, que en este caso serán 310, 312, 314 y 314. En el segundo tallo los números mayores o iguales que 5 y menores o iguales que 9, esto es, el 316 del ejemplo. De manera análoga se procede para el otro número del tallo. En este caso el diagrama es:

31	0244
31	6
32	244
32	8

Con el auxilio de esta gráfica se tendrá una idea general de la distribución de los datos.

Construcción del diagrama de tallo y hoja para el índice de ozono

Tallo	Hoja	Frecuencia	Frecuencia relativa
8	9	1	0.02
9	3	1	0.02
9		0	0.00
10	11234	5	0.10
10	5567788888	10	0.20
11	00000011122234444	17	0.34
11	66678999	8	0.16
12	00122444	8	0.16

Figura 2.13 Diagrama de tallo y hoja para el ozono con la frecuencia y frecuencia relativa.

Retomando el problema del índice de ozono, observe que los datos de los valores del índice están en centésimas, salvo dos casos. Así, en el tallo se pueden escribir por décimas y en la hoja por unidades. Por ejemplo, las cantidades 107 y 122, se pueden representar por:

10	7
12	2

A partir de esta idea se construye la figura 2.13, pero se debe resaltar que el ancho de clase es de 5 unidades. El primer renglón del tallo y hoja comprende los valores entre 85 y 89; el segundo, entre 90 y 94, y así hasta el último renglón que corresponde a los datos de ozono entre 120 y 124.

A partir de estos datos puede interpretarse que el índice de ozono en dos de los 50 días estuvo por debajo de las 100 unidades. El índice más alto apareció durante 8 días. Con el propósito de tener un panorama completo y de la utilidad del diagrama de tallo y hoja, a la derecha del diagrama se han incorporado las frecuencias absoluta y relativa, respectivamente.

Unidad en un diagrama de tallo y hoja

Establecer una unidad de ayuda en la construcción de un diagrama de tallo y hoja. Por lo general, ésta debe aparecer en la parte superior del diagrama. Por ejemplo:

$$\text{Unidad: } 8|3 = 83000$$

En este caso la unidad que se lee indica que la partición se da entre diez de miles y los miles. De este modo, si el número con tallo fuera 9 y el de la hoja 6, se leería como 96000. En contraste, si dice:

$$\text{Unidad: } 8|3 = 0.083$$

Esto significa que el número con tallo 9 y hoja 6 se leería como 0.096.

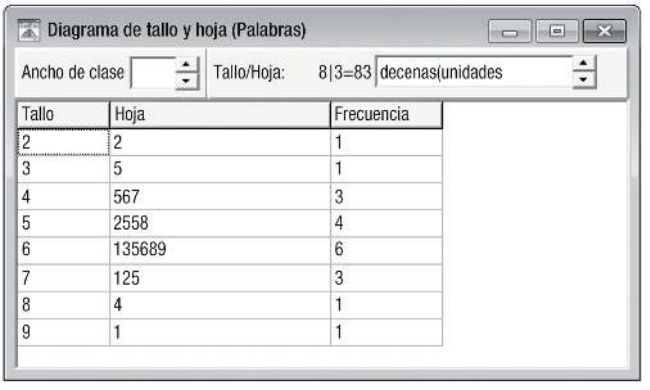
Ejemplo 2.3

Un administrador tiene que evaluar la habilidad en mecanografía en un grupo de trabajadores de su empresa, para llevar a cabo un proyecto que acaban de solicitarles. Con el propósito de conocer el desempeño de los empleados, toma una muestra de 20 de ellos. La medición consiste en escribir el mayor número de palabras en 2 minutos, como se muestra en la tabla de abajo. En este caso, hay que construir el diagrama de tallo y hoja para este conjunto de datos.

68	72	91	47	52	75	63	55	65	35
84	45	58	61	69	22	46	55	66	71

Solución mediante el uso de CalEst

Para la construcción del diagrama de tallo y hoja mediante **CalEst**, seleccione en la opción de gráficas el Diagrama de Tallo y Hoja (programa representado por una hoja verde). En la hoja de captura se deben tener los datos que se van a analizar, en este caso 20. Al seleccionar la opción el programa le pide el nombre de la variable, nombrada en la hoja, para la que quiere el diagrama de tallo y hoja. A continuación se debe completar las opciones del número de tallos que desea y las unidades relacionadas con el ejemplo; el inicio para modificar el diagrama y el resultado se muestran en la figura 2.14.



Tallo	Hoja	Frecuencia
2	2	1
3	5	1
4	567	3
5	2558	4
6	135689	6
7	125	3
8	4	1
9	1	1

Figura 2.14 Diagrama de tallo y hoja para el ejemplo 2.3.

2.2.6 Interpretación del histograma de tallo y hoja

El mundo de la información 6. Tiempo de falla de balatas

Un fabricante de balatas para el sistema de frenado de automóviles quiere saber cuánto tiempo tardarán en fallar las balatas que manufactura. El tiempo lo mide según el número de kilómetros recorridos por un auto, de modo que la prueba la efectúa en colaboración con una empresa que produce automóviles. De la muestra utilizada para realizar el estudio, se logró el registro en 55 autos.

Preguntas sobre la naturaleza del problema

El interés de este estudio consiste en establecer en cuántos kilómetros se tiene 10 % de los datos. Esta información se requiere para establecer periodos de garantía y de revisión de frenos. Los datos son:

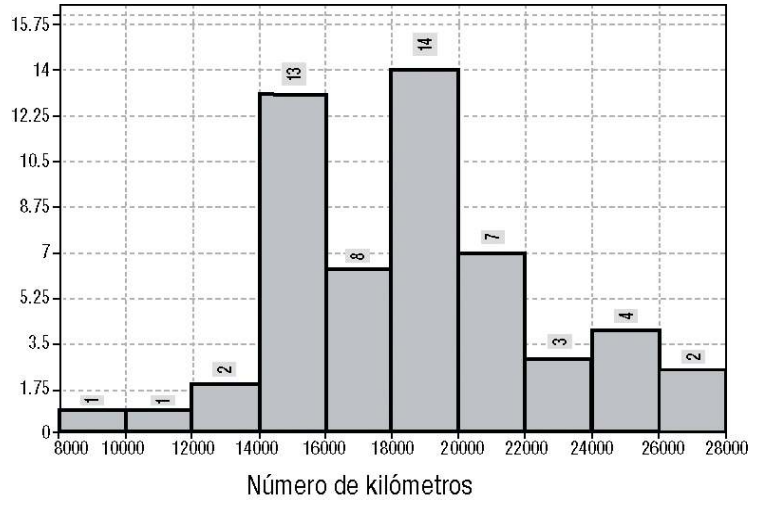
9500	14214	14951	16223	17980	19036	19451	21382	23659	27423
10512	14404	15117	16481	18508	19126	19611	21599	24165	
12824	14520	15520	16622	18624	19165	19708	21789	25731	
13514	14689	15555	16626	18699	19274	20066	21978	25961	
14096	14766	15912	16689	18719	19414	20546	22386	25991	
14128	14856	16037	16935	18773	19429	20610	23592	26533	

El resumen de estos datos se presenta en la tabla 2.6, donde se puede apreciar la relevancia del diagrama de tallo y hoja, pues proporciona tanto la información de la tabla de frecuencias como la forma en que se

distribuyen los datos. En el contexto del problema del tiempo de fallas de las balatas, las fallas reiteradas inician a partir de los 14 000 km. Así que el servicio o revisión de éstas debe ser antes de que un auto llegue a los 14 000 km, por ejemplo, a los 10 000 km.

El histograma de este conjunto de datos se muestra en la figura 2.15. Debemos resaltar que hay un parecido entre el histograma y el diagrama tallo y hoja. Por esa razón, se dice que éste es un diagrama que representa la tabla de frecuencias y, al mismo tiempo, al histograma. Interpretando con el histograma del ejemplo se llega a una conclusión similar a la que se llegó con el diagrama de tallo y hoja, esto es, que entre los kilómetros 14 000 y 22 000 el tiempo de falla es mayor.

Además, observe que el ancho de clase en el histograma equivale a 2000 km, mientras que en el diagrama de tallo y hoja, lo que corresponde a la clase es de 1000 km. Por lo tanto dos líneas del diagrama de tallo y hoja se iguala con una barra del histograma. En el contexto de la interpretación del problema, se tienen los elementos económicos para establecer el periodo de garantía de las balatas, es claro que ésta no debe pasar los 12 000 km. Considerando la información de la tabla de frecuencias e histograma, la garantía se puede extender hasta los 14 000 km.



Tallo	Hoja	Frecuencia
9	5	1
10	5	1
11	---	0
12	8	1
13	5	1
14	012456789	9
15	1559	4
16	0246669	7
17	9	1
18	56677	5
19	011244467	9
20	056	3
21	3579	4
22	3	1
23	56	2
24	1	1
25	799	3
26	5	1
27	4	1

Figura 2.15 Histograma para el tiempo de falla de las balatas, y su comparación con el diagrama de tallo y hoja.

Ante esta situación el fabricante de balatas considera retrasar el tiempo de falla para tratar de ser más competitivo. Para ello analizará los factores que ayudarán a mejorar la calidad; una técnica apropiada para dicho análisis es el principio de Pareto, el cual se verá más adelante cuando se estudien las gráficas para variables cualitativas.

Puede observar que el análisis se puede completar comparando las gráficas del histograma y el de puntos. Quizá una ventaja del diagrama de puntos en este ejemplo, es su presentación adecuada a la

medición del rendimiento de las balatas, ésta proporciona una idea clara de la distribución de los datos. En la figura 2.16 complemente la interpretación con su percepción de las gráficas.

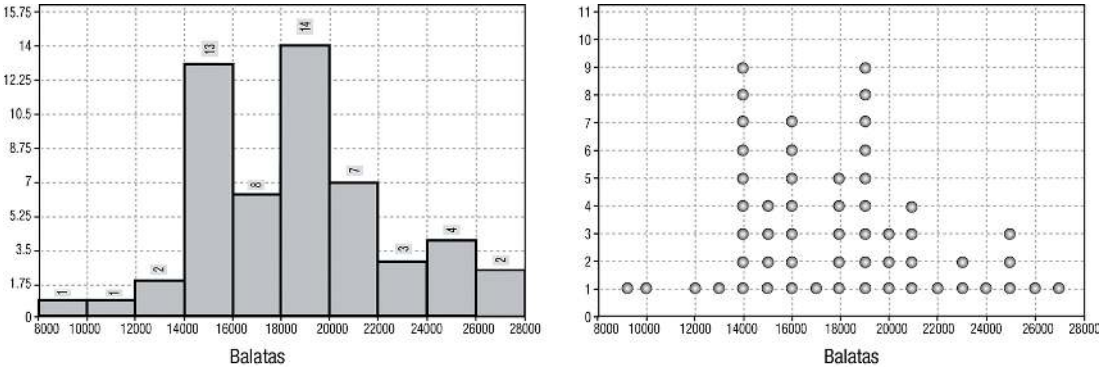


Figura 2.16 Histograma para el tiempo de falla de las balatas y su comparación con el diagrama puntos.

Tabla 2.6 Frecuencias para el tiempo de falla (en km) de las balatas.

Clase	Intervalo	de clase	F.	F.R.
1	8000.0	10000.0	1	0.0182
2	10000.0	12000.0	1	0.0182
3	12000.0	14000.0	2	0.0364
4	14000.0	16000.0	13	0.2364
5	16000.0	18000.0	8	0.1455
6	18000.0	20000.0	14	0.2545
7	20000.0	22000.0	7	0.1273
8	22000.0	24000.0	3	0.0545
9	24000.0	26000.0	4	0.0727
10	26000.0	28000.0	2	0.0364

Polígono de frecuencia y su relación con el percentil

Con auxilio del polígono de frecuencias relativas se logra un mayor conocimiento del tiempo de falla de las balatas. Así, en la figura 2.17 se observa que 50 % de las balatas fallan a los 18 699 km. Antes de llegar a los 30 000 km fallaron todas balatas de la muestra. A la derecha de esta gráfica se presentan 9 de los porcentajes (percentiles) para el ejemplo. En cada uno de ellos se presenta el kilometraje de falla, por ejemplo 90 % de las balatas fallaron a los 24 165 km.

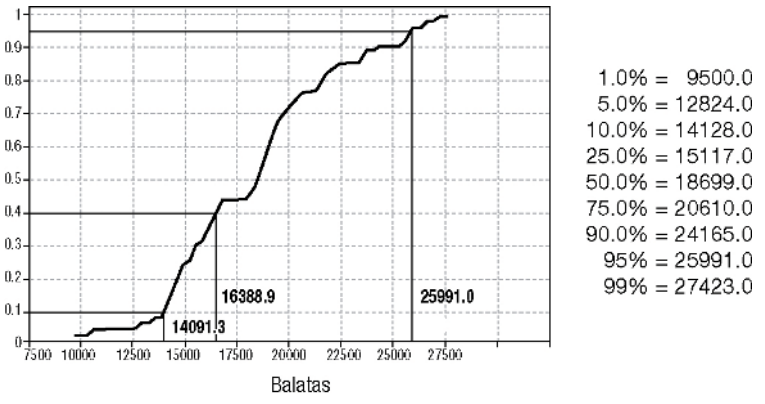


Figura 2.17 Polígono de frecuencia y su relación con el percentil.

2.2.7 Gráficas para datos cualitativos

Para describir observaciones cualitativas, se definen categorías de modo que cada observación pertenezca a una categoría única. De esta manera, el conjunto de datos se podrá representar numéricamente contando el número de veces que una observación cae dentro de una de las categorías definidas.

Las gráficas que, por lo general, se emplean en observaciones cualitativas son los tradicionales *diagramas de barras* y los denominados tipo pastel o circulares. Como se verá en este apartado, mucha información que se usa en la vida cotidiana se ejemplifica mediante estos tipos de gráficas.

Diagramas de barras: El mundo de la información 7. Hábito de fumar

Un grupo de padres de familia se interesa en conocer el hábito de fumar entre los adolescentes. La comprensión de este hecho permitirá proponer algunas medidas preventivas y con ello disminuir la tendencia a fumar entre la juventud, ya que esta conducta provoca serios daños a la salud. Para obtener la información sobre este tema se aplicó una encuesta a 150 jóvenes con tres preguntas:

1.	Fumas	Sí	No	
2.	Género	Mujer	Hombre	
3.	Si fumas, indica cuántos cigarros consumes al día, entre			
	1 y 5	6 y 10	11 y 15	Más de 15

En este estudio se pueden relacionar la pregunta 1 con la 2, y la pregunta 2 con la 3. En las tres tablas siguientes se presenta un resumen de los resultados, esto es, se realiza un registro de datos.

Clase Fumas	Frecuencia	Frecuencia relativa (%)
Sí	48	32.0
No	102	68.0
Total	150	100.0

Clase Fumas	Frecuencia	Frecuencia relativa (%)
Mujer	69	46.0
Hombre	81	54.0
Total	150	100.0

Clase Cigarros	Frecuencia	Frecuencia relativa (%)
1 y 5	8	16.67
6 y 10	21	43.75
11 y 15	12	25.00
más de 15	7	14.58
Total	48	100.0

Procedimiento para construir un diagrama de barras

1. Trazar el primer cuadrante del plano cartesiano.
2. Dividir el eje horizontal en intervalos del mismo ancho e igualmente espaciados para representar las categorías que aparecen descritas en la tabla de frecuencias.
3. Marcar las divisiones pertinentes en el eje vertical para denotar las frecuencias o frecuencias relativas.
4. A partir de las divisiones realizadas en el paso 2, se dibuja el rectángulo o la barra (las bases del rectángulo tienen el mismo ancho). La altura de las barras representa la frecuencia con que aparece la categoría.
5. Finalmente, escribir las etiquetas de las categorías y el título del diagrama.

En la figura 2.18 se reproduce el diagrama de barras tomando en cuenta tanto las frecuencias como las frecuencias relativas para la encuesta, considerando el género.

Una interpretación sencilla para esta representación gráfica es señalar que el porcentaje de mujeres encuestadas fue menor al de los hombres.

Sin embargo, es importante destacar cuál de los géneros es el que más fuma y en qué cantidad, según la encuesta. De los resultados mostrados en la primera tabla, se ve que hay 48 fumadores, y el equivalente en porcentaje es 32 %. De esos 48, ¿cuántos son mujeres? Para contestar esta pregunta es necesario utilizar otra tabla en que se detalle aún más la información. La información por género de los 48 fumadores y la cantidad de cigarros fumados al día, se muestra en la tabla 2.7.

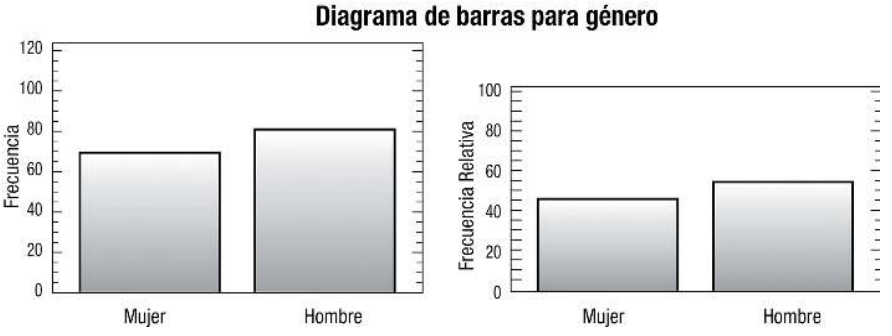


Figura 2.18 Diagramas de barras según el género para ambos tipos de frecuencias.

Tabla 2.7 Resumen de la información proporcionada por la encuesta sobre el hábito de fumar entre los jóvenes.

Fuma Cigarros	Mujer 16	Hombre 32	Total 48	Frecuencia relativa
1-5	2	6	8	0.16
6-10	5	16	21	0.44
11-15	3	9	12	0.25
más de 15	2	5	7	0.12

En la figura 2.19 las barras de la gráfica de la izquierda describen el número de fumadores y no fumadores. En ambas también se representa el género. Una primera conclusión indica que hay menos adolescentes que fuman de los que no fuman. Además la parte sombreada de la barra muestra que los hombres fuman más que las mujeres, y hay más mujeres que hombres que no fuman. Esta descripción en porcentaje se señala en la gráfica que está del lado derecho en la figura 2.19.

Por último, la figura 2.20 muestra cuántos cigarros fuman al día los adolescentes. Por ejemplo, en los extremos se observa que un número reducido de personas fuman menos de 5 o más de 15 cigarros. La mayor parte de los que fuman están entre 6 y 15 cigarros diarios. La sombra de las barras distingue entre mujeres y hombres.

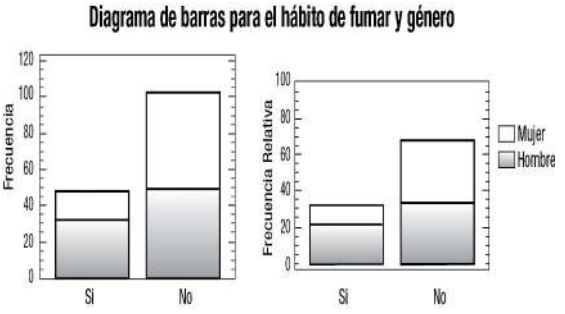


Figura 2.19 Diagramas de barras que describen en frecuencias absoluta y relativa el hábito de fumar por género.

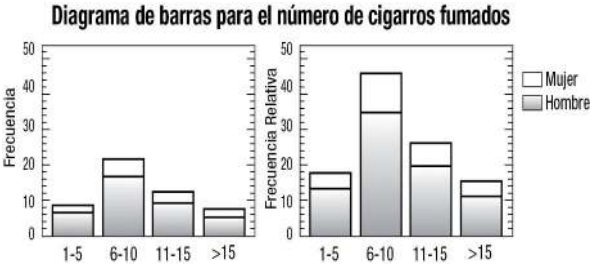


Figura 2.20 Diagramas de barras que describen en frecuencias absoluta y relativa el número de cigarros fumados.

Otras gráficas de barras

Una aplicación adicional de este tipo de gráficas es la de *barras múltiples*. Por ejemplo, en las facturas que emite bimestralmente una compañía encargada del suministro de energía aparece una historia del consumo de energía (ver figura 2.21). Esa información permite que el usuario se dé cuenta en qué periodo consumió más energía para tratar de averiguar la razón. Establecer comparaciones con otras personas, saber si durante el horario de verano se consume más o menos energía que en el horario normal, son algunos de los aspectos que pueden conocerse mediante una gráfica de barras del recibo de luz. A continuación se reproduce el diagrama de un recibo.

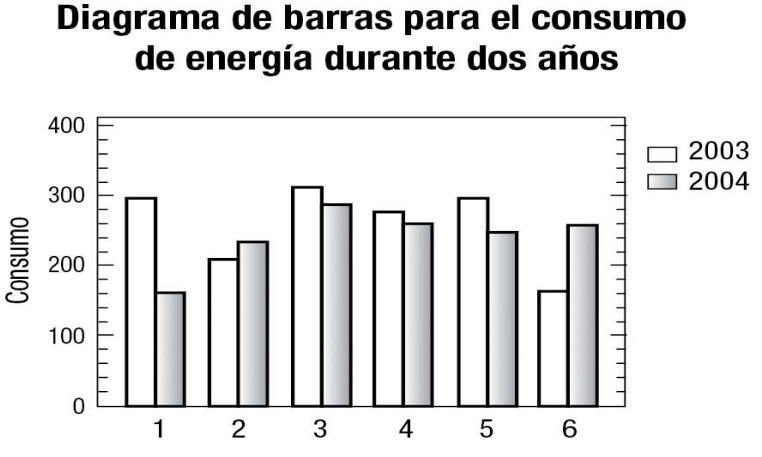


Figura 2.21 Diagrama de barras múltiple que describe el consumo de energía en cada bimestre durante dos años.

El Instituto Nacional de Estadística Geografía e Informática (INEGI) reporta información estadística sobre censos nacionales y de la aplicación de diferentes tipos de encuesta, por ejemplo las encuestas nacionales de ocupación y empleo. La figura 2.22 muestra la tasa media del crecimiento entre los años

de 1950 y 2000.

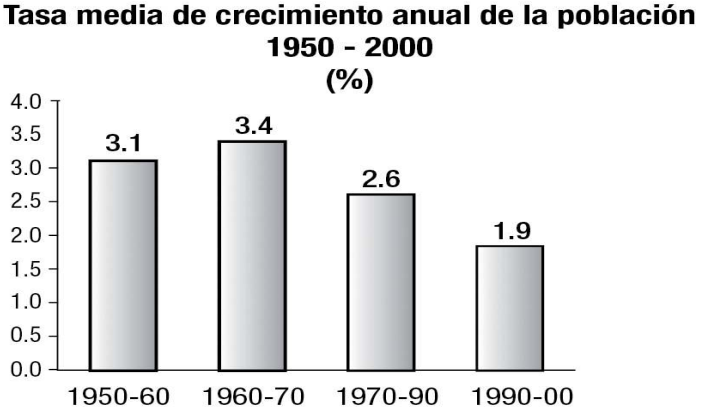


Figura 2.22 Diagrama de barras que describe la tasa media de crecimiento de la población.

La tasa media de crecimiento anual (TMCA) se calculó con la siguiente fórmula:

$$\left[(Pf/Pi)^{1/T} - 1\right]^{*} 100$$

donde Pf representa la población al final del periodo en estudio; Pi representa la población al iniciar el periodo y T la magnitud de dicho periodo. Otro ejemplo de una gráfica de *barras múltiple* se muestra en la figura 2.23, donde se describe la estructura de la población por sexo. ¿Qué conclusiones puede señalar de esta gráfica?

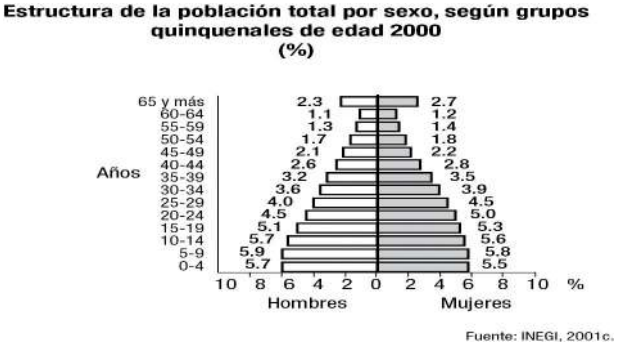


Figura 2.23 Otra forma de diagrama de barras múltiple que describe la estructura de la población en el año 2000.

Diagramas tipo pastel

El diagrama tipo pastel o circular es apropiado sólo cuando se quiere mostrar las proporciones en forma global. Sin embargo, debe tenerse en cuenta que una gráfica de este tipo no da toda la información. La

información proporcionada por este diagrama tiene un grado de subjetividad y está influenciado por el tamaño de cada rebanada, por lo que el diagrama de barras puede ser más eficiente. Este tipo de gráfica se utiliza más para que las personas puedan leer de manera gráfica la información. Por ello, muchas organizaciones dedicadas al levantamiento de encuestas y estudios de mercado recurren a diagramas circulares.

Ejemplo 2.4

Para mostrar la presentación de este tipo de gráfica, se presenta un reporte del Instituto Nacional de Nutrición en que se describe el porcentaje del suministro de energía alimentaria mediante un diagrama circular, el cual señala lo que una persona debe consumir: 3159 calorías por día. Véase la figura 2.24. A partir de este esquema gráfico se puede interpretar que los cereales (donde también se ubica la cerveza) son los que más aportan calorías. Otra fuente de calorías está en los edulcorantes. Un diagrama tipo pastel caracteriza los datos en forma de rebanadas o secciones de un círculo. Cada rebanada representa una categoría, y su tamaño corresponde de manera proporcional a la frecuencia relativa.

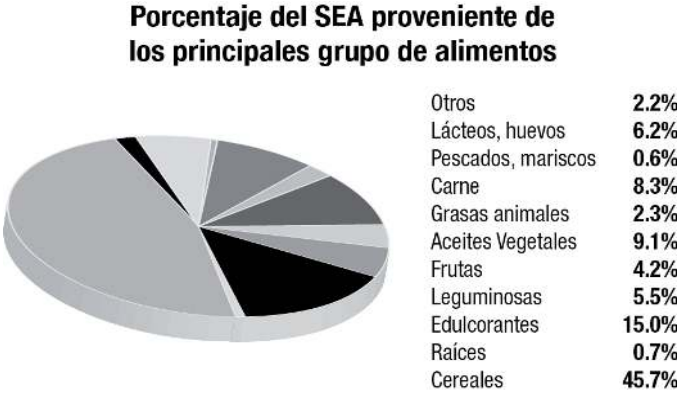


Figura 2.24 Diagrama tipo pastel sobre suministro de energía alimentaria que proviene de los principales alimentos. Fuente: Publicaciones del INN.

Diagrama tipo pastel usando CalEst

Para mostrar cómo se usa el CalEst en la elaboración de un diagrama de pastel, se empleará la distribución del gasto de las familias. En la opción gráfica del CalEst está el diagrama tipo pastel. Primero capture dos variables en la hoja de datos, una columna corresponde al porcentaje de las categorías, mientras que el nombre de éstas se encuentra en una segunda columna. Esta descripción aparece en la figura 2.25, en el cuadro de la izquierda, después se aprieta el botón aceptar y aparece el diagrama.

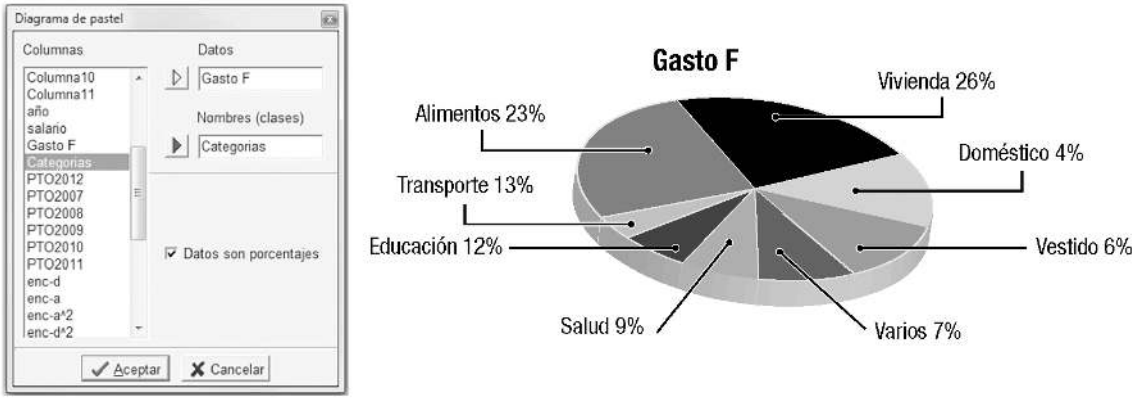


Figura 2.25 Diagrama tipo pastel sobre la distribución del gasto en la economía familiar.

Procedimiento para construir un diagrama tipo pastel

1. Trazar un círculo.
2. Dividir el círculo en espacios para representar las categorías que aparecen descritas en la tabla de frecuencias. Para obtener el ángulo de cada categoría, se multiplica la frecuencia relativa por 360 grados, lo cual corresponde al círculo completo.
3. Hacer las divisiones pertinentes en el eje vertical para denotar las frecuencias o frecuencias relativas.
4. Por último, escribir las etiquetas de las categorías, con sus porcentajes y el título de la gráfica.

Un diagrama tipo pastel caracteriza los datos en forma de rebanadas o secciones de un círculo. Cada rebanada representa una categoría, y su tamaño corresponde de manera proporcional a la frecuencia relativa.

Diagrama de Pareto

El diagrama de Pareto es una herramienta para tomar decisiones respecto a qué causas hay que resolver de manera prioritaria para lograr mayor efectividad en la solución de problemas.

En el contexto de la calidad, la administración de una empresa de manufactura es consciente de que las situaciones y los problemas que debe enfrentar tienen una importancia distinta. Los problemas de calidad se presentan en forma de pérdidas (artículos defectuosos y los costos asociados a ello). Es frecuente observar que 80 % de los defectos es producido solamente por 20 % de las causas posibles.

En 1897 el economista italiano V. Pareto mostró que 80 % del ingreso en Italia lo gozaba 20 % de la población. En general este principio dice que aproximadamente 80 % de los problemas se deben a tan sólo un 20 % de causas. Es decir, un mínimo porcentaje de causas originan un gran porcentaje del problema.

Esta distribución no homogénea se conoce actualmente como Principio de Pareto, el cual aparece en distintas situaciones de la vida cotidiana. Por ejemplo:

- La riqueza mundial la acumula solamente unos pocos países.
- La contaminación ambiental la producen sólo algunas empresas.
- El 80 % de las ventas se realizan con el 20 % de los clientes (clientes clave).
- En un proceso industrial, 80 % de los tiempos muertos en un proceso industrial se deben a 20 % de las causas posibles.
- El ausentismo del personal en una empresa se concentra en ciertas personas.

Este principio, conocido también como Ley 80-20, igualmente se observa en el terreno del control de calidad como pocos vitales, muchos triviales, es decir, los problemas de calidad que enfrenta una empresa se derivan de unas cuantas causas. El diagrama de Pareto permite identificar ese pequeño porcentaje de causas relevantes sobre las que se debe actuar primero.

En la producción de bienes y servicios hay varios aspectos que pueden ser mejorados, como los tiempos, los costos, los volúmenes, la cantidad de defectos, etc. Cada problema, a su vez, consta de pequeños problemas que en la práctica resulta imposible eliminar o disminuir. Puesto que los recursos de cualquier empresa son limitados, es necesario utilizarlos para obtener los mejores resultados posibles. La estrategia es identificar los principales problemas que han de enfrentarse, de modo que las posibilidades de éxito sean máximas, y el instrumento para identificar las causas que dan lugar a los problemas en estos casos es el análisis de Pareto.

El mundo de la información 8. Tipo de defectos en un radio

En una compañía de la rama electrónica, una línea de radios presenta problemas. Se realizó una inspección a 100 % de un lote de 3000 radios, y 531 fueron rechazados por estar defectuosos. Entre las distintas causas de los defectos, los más frecuentes fueron: mala sintonización, presentación exterior dañada, fallas en la fuente, problemas con el volumen, indicadores luminosos defectuosos, nitidez del audio y otros problemas menores. En la tabla 2.8 se presentan los defectos por tipos, la frecuencia con que se presentaron, el porcentaje que representan dentro del lote, el porcentaje de la causa dentro de los defectuosos y el porcentaje acumulado.

La información gráfica es muy valiosa en estos casos, por lo que el análisis de Pareto se apoya en lo que se conoce como diagrama de Pareto. Éste consta de un diagrama de barras, donde cada una de las causas consideradas se representa por una barra y su altura corresponde a la frecuencia o al porcentaje de incidencias del defecto. Un diagrama de Pareto presenta casi toda la información generada en el análisis. La figura 2.26 muestra el diagrama asociado a los datos del ejemplo descritos en la tabla 2.8.

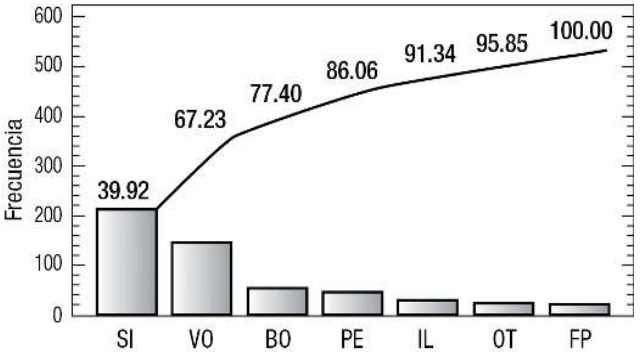


Figura 2.26 Diagrama de Pareto para el tipo de defectos.

Tabla 2.8 Registro en una muestra de radios.

	Fecha 21 de abril			
Tipo de defecto	Número de defectuoso	Porcentaje de defectuosos	Distribución de defectuosos	Distribución acumulada
PE: Presentación exterior	46	1.53	8.66	8.66
SI: Sintonización	212	7.07	39.92	48.58
EP: Fuente de poder	22	0.73	4.14	52.72
VO: Volumen	145	4.83	27.31	80.30
IL: Indicadores luminosos	28	0.93	5.27	85.30
BO: Bocinas	54	1.80	10.17	95.47
OT: Otros	24	0.80	4.52	99.99
Total	531	17.69		100

Construcción de un diagrama de Pareto

Al realizar un análisis de Pareto es importante la construcción del diagrama correspondiente. Se recomienda seguir los siguientes pasos:

1. Determinar el problema que se desea estudiar.
2. Determinar el periodo que abarcará el análisis.
3. Determinar la frecuencia total, la frecuencia en cada una de las categorías y los porcentajes correspondientes.
4. Elaborar una tabla. Dejar como última categoría Otros, si es el caso (tabla 2.8). Acumular los porcentajes.

5. Para elaborar el diagrama de Pareto hay que trazar dos ejes perpendiculares. Luego debe determinarse la escala que se usará en el eje vertical, usando como base las unidades utilizadas.

6. Ordenar por categorías, de modo decreciente según porcentajes y dibujar un diagrama de barras. En el extremo izquierdo se dibuja la barra correspondiente a la categoría con mayor incidencia, y hacia la derecha, en orden decreciente, se dibujan las demás barras.

7. Marcar escalas.

8. Dibujar la curva de Pareto (polígono de frecuencias acumuladas). En el extremo derecho de cada clase hay que marcar la altura que le corresponde a la frecuencia acumulada. En el último punto correspondiente a Otros, debe coincidir con la marca de 100 %. Luego, se unen los puntos por líneas rectas.

9. Nombrar por categorías a cada barra según la categoría que representa.

10. Interpretar el diagrama de Pareto.

Interpretación

En la gráfica puede verse que la sintonización fue la causa de defectos más frecuente, después el volumen, luego las bocinas, etc. La última barra presenta la categoría Otros que incluye todas las causas no estudiadas explícitamente en el análisis. Este diagrama considera la información principal del análisis y, como se observa, al presentarse de manera gráfica puede verse con claridad el problema.

Así, un programa de mejoramiento (disminución de defectos) de la compañía de artículos electrónicos deberá concentrarse sobre todo en los problemas de sintonización y volumen.

Práctica

Copie la siguiente tabla 10 veces y realice la encuesta a 10 vecinos o parientes.

Encuesta	Respuesta
1. Identificación número	
2. ¿Cuál fue el consumo de energía que marcó el último recibo de luz de tu casa?	
3. ¿Cuántos metros de construcción tiene tu casa?	
4. ¿Cuántas personas viven en tu casa?	
5. ¿Cuántos aparatos eléctricos hay en tu casa?	
¿Cuál es el consumo histórico de los seis bimestres de los dos últimos años?	
Año B1 B2 B3 B4 B5 B6	

1. La pregunta 1 servirá para identificar el número de encuesta y no se repita la información cuando se vacíen los datos en una tabla. La respuesta aparece en la columna de la derecha, excepto en la última pregunta.

a) Cada pregunta es una variable, así la 2 es consumo de energía, la 3 tamaño de la casa, la 4 el número de personas y la 5 el número de aparatos. Clasifique esas variables.

b) Realice un diagrama de puntos para la pregunta 2.

2. Con respecto a la variable sobre el consumo de energía, reúna su información con la de otros 10 compañeros de clase. Elabore una tabla de frecuencia, un histograma, el diagrama de tallo y hoja. Luego dibuje el polígono de frecuencias acumuladas y estime el 5, 25, 50, 75 y 95 percentil.

3. Compare la información con la de otros compañeros y discuta los resultados con ellos.

2.3 Resumen

Término	Definición
Tabla de frecuencias	Una tabla de frecuencias o distribución de frecuencia es una tabla que registra categorías, valores o clases de frecuencias valores que una variable debe tener y el número de veces que cada una ocurre.
Histograma	El histograma es, en esencia, una gráfica de barras en que las categorías son clases. En un histograma de frecuencias, la altura de las barras está determinada por la frecuencia de clase.
Frecuencia relativa	La frecuencia relativa de una clasificación es el número de veces que una observación cae sobre la clasificación, la cual está representada como una proporción del número total de observaciones.
Frecuencia relativa acumulada	La frecuencia relativa acumulada de una clase C_i es la suma de las frecuencias relativas de todas las clases anteriores a C_i. Ésta se expresa como fracción, decimal o porcentaje.
Diagrama de puntos	Un diagrama de puntos es una gráfica que se emplea para un conjunto pequeño de datos, y cada observación se traza como un punto en el eje horizontal.
Diagrama tallo y hoja	El diagrama de tallo y hoja se utiliza para el análisis exploratorio de datos y representa tanto la tabla de frecuencias como el histograma. Cada observación del conjunto de datos se divide en dos partes, el tallo y la hoja.
Diagrama de barra	Una gráfica de barras representa la frecuencia de una tabla de frecuencias.

Término	Definición
Diagrama de pastel	Un diagrama de pastel es una gráfica en que se usa un círculo para representar el todo y cada rebanada se refiere a una categoría. El tamaño de la rebanada es proporcional a la frecuencia relativa de la categoría.
Diagrama de Pareto	Un diagrama de Pareto es una gráfica de barras de frecuencias en que las barras se presentan en orden de altura, iniciando con la más alta.

2.4 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

2.5 Ejercicios

Variables numéricas 1

2.1 Con el propósito de mejorar las condiciones económicas y nutritivas en una comunidad rural, se construyeron algunos estanques para el cultivo de peces. El programa consiste en producir peces con el tamaño y peso apropiados para el consumo. Además, mediante la alimentación, los biólogos encargados de este proyecto trataron de controlar otros aspectos como el color; también midieron los niveles de mercurio en los peces para que no rebasaran los límites permitidos. En la pesca se espera que los peces atrapados alcancen una longitud mayor a 35 centímetros y un peso mayor a 600 gramos. Mediante un mecanismo apropiado de pesca se sacan en una primera etapa 60 pescados. ¿Tendrán esos pescados las dimensiones propuestas? Datos observados:

53	36.3	42	56.9	44	37.1	46.4	55.9	46.3	49
61.8	39.6	42.3	51.1	49.4	49.5	24.5	45.3	43.1	47.6

37.4	39.2	43.1	45.6	40	46.9	33.5	53.5	42.6	43.3
44.8	41.8	35.9	50.2	36.8	49	53.5	49.9	44	39.1
45.6	42.8	39.7	35.7	56	53.2	46.7	43.4	37.7	42
39.4	40.8	41.8	46.4	43.1	45.1	54.6	44	41.2	36

Construya la tabla de frecuencias. Luego, indique qué porcentaje de pescados está por debajo de 35 cm y qué porcentaje entre 36 y 50 cm.

2.2 En una segunda etapa del problema anterior se pescó y pesó 90 peces. En la siguiente tabla aparecen los pesos de dichos animales. Use el CalEst para obtener la tabla de frecuencias e interprete los resultados de la tabla generada.

1424	1352	1616	1269	1471	1469	1568	1526	1438	1421
1381	1385	1358	1412	1376	1435	1572	1551	1494	1423
1530	1463	1610	1544	1458	1477	1539	1385	1420	1508
1462	1512	1339	1273	1373	1329	1519	1346	1428	1346
1571	1397	1596	1460	1378	1493	1479	1472	1600	1453
1577	1360	1487	1447	1372	1519	1531	1475	1407	1431
1341	1456	1458	1465	1535	1359	1435	1389	1578	1532
1384	1348	1433	1461	1459	1486	1432	1503	1511	1363
1461	1552	1478	1462	1435	1533	1400	1723	1455	1355

2.3 Se llevó a cabo un estudio para saber cuántas horas a la semana dedican los jóvenes a platicar por internet ("chatear") con otros amigos. La respuesta dada por 30 jóvenes fue:

0.0	4.1	11.9	15.5	17.8	2.9
0.0	4.2	13.9	16.7	17.9	8.6
0.1	6.9	13.4	15.5	18.4	15.1
1.4	8.0	13.5	15.5	18.7	16.8
2.5	8.2	14.9	15.5	19.5	19.7

Elabore una tabla de frecuencias para el tiempo que emplean los 30 jóvenes en platicar por internet. Incluya la frecuencia relativa e interprete los resultados.

2.4 La estatura de 100 personas es la siguiente:

170	171	168	175	154	156	174	159	172	152	173	178
159	157	158	161	168	172	175	170	166	164	176	161
160	164	170	161	169	163	173	170	164	175	171	173
171	170	165	161	162	159	151	162	159	168	172	166
167	173	164	169	177	173	175	148	169	165	172	172
164	167	166	169	153	171	160	166	169	158	169	156

170	180	172	171	162	169	167	159	157	161	169	159
172	175	182	172	152	163	151	187	170	165	162	168
159	161	171	164								

Elabore una tabla de frecuencias para la estatura de esas 100 personas. Incluya en ella la frecuencia relativa e interprete los resultados. ¿Qué porcentaje de personas miden más de 160 cm?

2.5 La edad en que contrajeron matrimonio 120 personas se muestra a continuación:

24.9	41.6	40.1	24.0	40.8	32.0	32.4	35.1	28.2	23.3	30.7	24.6
27.7	34.8	58.7	32.3	32.5	45.4	28.5	38.8	29.4	32.3	37.8	41.1
33.9	40.7	36.4	34.7	35.9	27.1	31.5	42.1	24.9	31.0	40.7	26.8
24.9	50.2	48.0	34.7	25.5	29.0	47.1	34.7	26.0	36.4	35.6	22.8
25.1	21.5	42.7	31.5	30.3	33.7	32.6	41.9	25.2	30.0	31.3	27.0
44.8	24.2	33.5	32.1	26.7	35.2	41.7	30.6	30.0	20.9	30.4	29.6
33.7	28.0	35.4	30.1	32.8	47.0	37.7	23.0	32.0	34.5	27.0	53.5
29.1	39.9	40.9	51.7	22.7	29.5	36.6	29.1	37.8	30.3	41.3	27.7
42.9	29.9	30.8	26.9	25.9	33.3	38.7	35.3	35.4	34.8	41.6	22.0
28.3	42.4	18.7	31.6	39.6	27.2	42.5	28.6	29.6	28.6	27.1	37.5

Elabore una tabla de frecuencias para la edad en la que contrajeron matrimonio esas 120 personas. Incluya la frecuencia relativa e interprete los resultados. ¿Qué porcentaje de personas se casan a los 20 años o menos? ¿Cuántos se casan entre los 20 y los 25 años? ¿Cuántos se casan siendo mayores de 40 años?

2.6 El peso en gramos de 80 bebés al nacer es:

2932	2870	2837	3010	2845	2892	2907	2833
3107	2824	2345	2744	2905	3275	3162	2782
3151	2985	3290	2979	3098	2766	2676	2845
3162	3010	2959	2938	2965	2733	2489	3196
2875	2937	3012	2802	2929	2925	3087	3103
2552	2936	3019	3049	2927	2999	2997	3275
2994	2965	2965	3060	2934	3211	2576	2468
2582	3065	2758	3184	2457	2709	2647	2998
3171	2988	2946	3016	3080	3021	3050	2778
3151	3051	3049	2572	3205	2892	2751	2623

Construya la tabla de frecuencias correspondiente al peso de los 80 bebés. Incluya la frecuencia relativa e interprete los resultados. ¿Qué porcentaje de bebés pesa menos de 3 kilogramos?

2.7 Se preguntó a 40 personas cuánto tiempo dedican a la semana para limpiar sus casas, sin considerar

el tamaño de éstas. Las respuestas obtenidas se presentan en la siguiente tabla.

0.6	1.2	1.3	1.9	2	2.1	2.2	2.4
2.6	2.6	2.7	2.9	3	3	3.1	3.1
3.3	3.3	3.3	3.4	3.6	3.8	3.8	4.1
4.2	4.4	4.4	4.4	4.5	4.7	4.7	4.9
5	5	5.6	5.7	5.8	6	6	7.4

Elabore una tabla de frecuencias para el tiempo que las personas dedican a la limpieza de sus casas. Incluya la frecuencia relativa e interprete los resultados. ¿Cuál es el porcentaje de personas que emplea más de 3 horas en la limpieza? ¿Cuál es el que emplea entre 2 y 5 horas? ¿Cuál porcentaje de personas emplea más de 6 horas en la limpieza de sus casas?

2.8 El precio de un producto (que bien puede ser una medicina, una bebida o una fruta) cambia en cada establecimiento. Existe una dependencia que, con el lema "quién es quién en los precios", se especializa en averiguar los precios de diferentes productos en distintas épocas del año. Dicha dependencia gubernamental realizó un estudio para conocer el precio de diferentes medicinas. Aquí se presenta la información sobre el precio para un tipo particular de medicamento. La presentación de esta medicina viene en una caja con 14 tabletas de 50 mg. La encuesta se realizó en 45 comercios en una ciudad grande y los precios (en pesos) recabados se muestran a continuación:

279	356	346	342	357	320	342	351	332
295	326	327	352	344	347	372	329	375
348	300	355	329	335	358	364	341	380
335	351	353	343	328	325	334	335	405
355	335	334	333	331	381	313	368	390

Elabore una tabla de frecuencias para el precio de dicha medicina. Incluya la frecuencia relativa e interprete los resultados. ¿Qué porcentaje de establecimientos la venden a menos de 310 pesos? ¿Qué porcentaje la vende a más de 360 pesos? ¿Qué porcentaje de establecimientos vende la medicina entre 310 y 360 pesos?

2.9 Por lo general, los jóvenes invierten tiempo para ir de sus casas al colegio. Para saber cuánto tiempo tardan en llegar a la escuela, el director le pidió a uno de los profesores que preguntara a 100 estudiantes cuánto tiempo tardan en ir de su casa a la escuela. Los resultados registrados en minutos son como siguen:

35	26	27	34	54	30	28	43	34	28
44	36	31	53	29	37	47	27	31	61
34	34	39	35	33	53	31	50	44	46
36	57	61	41	29	36	53	52	29	26
36	42	49	36	29	39	39	45	38	47

44	36	47	51	53	43	42	42	28	34
40	23	71	38	48	58	38	57	47	24
46	26	36	36	47	24	25	34	31	26
37	37	35	36	32	34	41	35	32	35
50	39	31	39	27	43	47	53	62	53

Construya una tabla de frecuencias a partir de los datos anteriores. Incluya la frecuencia relativa e interprete los resultados. ¿Qué porcentaje de estudiantes emplean más de 60 minutos en llegar a la escuela? ¿Qué porcentaje tarda entre 30 y 60 minutos en ir de su casa a la escuela? ¿Qué porcentaje de estudiantes emplea menos de 30 minutos en llegar a la escuela desde sus casas?

Variables numéricas 2

2.10 Construya los histogramas de frecuencia y frecuencia relativa de todas las actividades de aprendizaje del apartado anterior. Interprete el histograma resultante en el contexto de cada uno de los problemas.

2.11 Se registra el tiempo, en segundos, de 1000 llamadas telefónicas. Luego se muestra la tabla de frecuencias para la distribución de llamadas.

Clase	Intervalo de clase	Frecuencia	Frecuencia relativa	Frecuencia relativa acumulada
1	0 - 100	6		
2	100 - 200	28		
3	200 - 300	88		
4	300 - 400	180		
5	400 - 500	247		
6	500 - 600	260		
7	600 - 700	133		
8	700 - 800	42		
9	800 - 900	11		
10	900 - 1000	5		

1. Construya el histograma de frecuencias.

2. Complete la tabla de frecuencias relativas y elabore el histograma de frecuencias relativas.

Variables numéricas 3

2.12 Construya los histogramas de frecuencias relativas y el correspondiente polígono de frecuencias relativas para las actividades de aprendizaje de los apartados anteriores. Interprete todos los histogramas resultantes de la unidad en el contexto del problema.

2.13 Elabore el polígono de frecuencias y frecuencias relativas para los datos de la actividad de aprendizaje 2 del apartado anterior. Indique el porcentaje de llamadas entre 300 y 600 segundos. ¿Qué porcentaje de llamadas ocurren después de 700 segundos?

Variables numéricas 4

2.14 Use el paquete estadístico para elaborar el polígono de frecuencias relativas acumuladas del ejemplo 2.1 y señale los salarios para 25 %, 50 % y 75 %. Interprete los resultados.

2.15 Elabore el polígono de frecuencias para el problema 2.1, Consumo de energía. ¿Cuál es el valor de la energía equivalente a 25 %, a 50 % y a 75 %?

2.16 Construya el polígono de frecuencias relativas acumuladas para las actividades de aprendizaje del apartado Tabla de distribución de frecuencias. Encuentre los valores correspondientes a 5 %, 25 %, 50 %, 75 % y 95 %; además interprete los resultados en el contexto del problema.

Variables numéricas 5

2.17 Escoga un libro, puede ser de historia o literatura, y ábralo en cualquier página; luego, seleccione un párrafo. Considere las primeras 50 palabras y de cada una de ellas anote el número de letras, esto es, el tamaño de la palabra. Con estos 50 datos construya un diagrama de puntos. ¿Hay algún patrón en la distribución? ¿Cuál es el tamaño más frecuente? ¿De qué tamaño es la letra más grande y con qué frecuencia aparece?

2.18 En el taller de lectura y redacción se les ha recomendado a los estudiantes un libro para que lo lean en una semana. Se les preguntó a 35 estudiantes la cantidad de páginas leídas el día en que iniciaron la lectura de ese libro. Los resultados fueron los siguientes:

17	7	11	11	15	14	12	15	20	8	12	21	15	8	8	7
8	14	16	17	6	19	13	13	16	11	13	12	17	9	8	23
15	14	12													

Con base en los datos obtenidos, elabore un diagrama de puntos.

2.19 Considere las siguientes palabras:

Piedra	Cama	Llavero	Alarma
Tapa	Libro	Tijeras	Reloj
Silla	Bote	Pera	Perico
Mochila	Verde	Coche	Árbol

Escriba las palabras en una tarjeta, luego muéstresela a una persona por 20 segundos, deje pasar un minuto y pídale que le diga las palabras que recuerda, pero no puede tardarse más de dos minutos.

Anote el número de palabras que recuerda, y repita esta actividad con otras 25 personas.

1. Con los 25 datos obtenidos de la actividad elabore un diagrama de puntos.
2. ¿Qué porcentaje de personas recuerda más de 12 palabras?
3. ¿Qué porcentaje de personas recuerda menos de 8 palabras?

2.20 El número de días en que 30 personas se aliviaron totalmente de una gripe se muestra en el siguiente conjunto de datos.

3	7	6	7	6	8	6	8	5	11	7	5	7	10	8	15	14
9	5	5	4	7	4	6	7	7	4	8	9	8				

1. Dibuje un diagrama de puntos para estos datos.
2. ¿Qué porcentaje se alivió en 10 días o más?

Variables numéricas 6

2.21 Un especialista en biotecnología estudia el tiempo que puede durar un producto alimenticio en el estante de una tienda de autoservicio, también conocido como tiempo de vida de anaquel. Hace algunas pruebas en el laboratorio considerando un tipo de conservador para establecer el tiempo en días que puede durar. Con la información que le brindan 30 pruebas puede fijar la fecha de caducidad. Los datos son:

99	62	45	49	51	53	74	87	51	75	43	60	93	49	47
40	46	44	52	64	41	90	58	45	39	68	73	89	63	89

Elabore un diagrama de tallo y hoja. A partir de los datos observados, ¿qué tiempo recomendaría para la caducidad?

2.22 Elabore un diagrama de tallo y hoja para el ejercicio 9 de las actividades de aprendizaje 2.1.

2.23 ¿Cómo se leerá un número en un diagrama de tallo y hoja con (i) tallo 15 y hoja 7 y (ii) tallo 6 y hoja 7. Considere que las unidades establecidas son: a) 12|9=1.29; b) 12|9=129; c) 12|9=0.00129.

2.24 Construya un diagrama de tallo y de hoja para el nivel de colesterol, medido en moles por litro, de una muestra de 38 personas mayores de 40 años, que arrojó los siguientes resultados.

67	78	66	50	94	53	77	66	63	50
61	72	77	83	68	93	83	64	88	87
51	69	51	85	61	65	70	83	86	70
48	70	52	47	51	47	91	47		

2.25 Determine un diagrama de tallo y hoja apropiado. Constrúyalo para los siguientes conjuntos de datos, y en cada caso establezca la unidad.

1. Diámetro, en centímetros, de unos cilindros: 1.85, 1.63, 2.14, 2.08, 2.12, 1.80, 2.02, 1.74, 1.81, 2.08, 1.64, 1.96, 1.88, 1.92, 1.79, 1.75, 1.92, 1.96, 1.77, 1.82.

2. Altura, en centímetros, de unos árboles: 543, 625, 651, 659, 746, 528, 664, 705, 589, 588, 499, 478, 569, 519, 538, 517, 746, 434, 591, 772, 482, 695, 505, 458, 683, 423, 589, 554, 556, 545, 520, 697, 545, 509, 603, 534, 605, 700, 470, 669, 637, 464.

3. La longitud en metros: 16.321, 17,144, 16.492, 17.526, 16.144, 17.633, 17.835, 15.835, 16.463, 17.237, 16.871, 16.815, 16.960, 17.387, 17.422, 17.467, 17.527, 17.976, 15.893, 15.902, 15.718.

Variables numéricas 7

2.26 Construya el diagrama de tallo y hoja para los datos del problema 2.3 "Condición física", y compárelo con el histograma de la figura 2.6.

2.27 Use los datos del ejercicio 9 de las actividades de aprendizaje 2.1 para realizar un comparativo entre el histograma y el diagrama de tallo y hoja.

2.28 La cantidad de lluvia que se reportó, durante los últimos 40 días del verano, en una ciudad del sureste se presenta en la siguiente tabla:

86	25	87	28	53	78	86	74
21	78	48	33	23	101	62	118
43	15	51	88	34	100	102	65
59	66	62	70	84	75	60	66
80	58	53	68	108	26	50	85

1. Elabore la tabla de frecuencia y el histograma correspondiente.

2. Construya el diagrama de tallo y hoja y compárelo con el histograma.

3. ¿Cuál es el porcentaje de días en que la cantidad de agua fue superior a los 100 mm?

2.29 En un bachillerato se aplica una prueba de conocimientos generales al finalizar el ciclo escolar. Esta prueba permitirá que la dirección de la escuela sepa el nivel de conocimientos que poseen sus alumnos y, además, pueda estimar el éxito de sus estudiantes en el examen de admisión a la universidad. La prueba consta de 200 preguntas, y los datos que se muestran a continuación representan el número de respuestas correctas.

135	136	129	132	142	138	141	122	121	114	126	114
117	141	118	148	123	122	112	116	147	112	145	139
106	127	132	143	146	111	153	130	112	123	141	120
115	145	132	140	129	115	123	140	123			

1. Preparare una tabla de frecuencias con su respectivo histograma.

2. Trace un diagrama de tallo y hoja, después compárelo con el histograma.

3. Dibuje el polígono de frecuencias relativas acumuladas y señale a qué calificaciones corresponde 25 %, 50 % y 75 % de los datos.

2.30 Siguiendo con la actividad anterior, la dirección de la escuela consideró que el nivel de conocimientos es bajo, y puso en práctica algunos cambios en la capacitación de los alumnos en la siguiente generación. Los resultados de la prueba de conocimientos generales del siguiente año fueron:

136	151	155	148	158	151	164	145	157	152	148	142
150	153	134	156	146	133	157	140	148	137	152	152
148	149	146	152	158	164	162	142	147	140	160	160
153	142	154	167	147	148	148	136	135			

1. Elabore el diagrama de tallo y hoja y compárelo con el diagrama de la actividad 4. ¿Qué diferencias observa?

2. Grafique el polígono de frecuencias relativas acumuladas y encuentre el número de respuestas correctas para el 25 %, 50 % y 75 %. Nuevamente, compare los resultados con los de la actividad 4; ¿qué puede concluir?

Variables categóricas

2.31 El gerente de un supermercado desea conocer el estado que guardan en el almacén las cajas de cereales. Un empleado recoge la siguiente información:

Abierta	Aplastada	Abollada	Sin def.	Aplastada
Sin def.	Aplastada	Sin def.	Sin def.	Sin def.
Aplastada	Sin def.	Abollada	Sin def.	Aplastada
Aplastada	Abierta	Abollada	Sin def.	Abollada

Construya el diagrama de barras que le corresponda.

2.32 El profesor de sociales le preguntó a sus 35 alumnos por el género de película que más les gustaba ver.

Ficción	CF
Musical	M
Terror	T
Comedia	C
Drama	D
Cómica	CO
Fantasía	F

Los datos son:

D	D	C	C	D	C	C
C	C	C	T	CF	C	C
CF	CO	D	F	M	C	CF
D	F	C	C	C	D	CO
CO	C	T	D	D	C	D

1. Dibuje un diagrama de barras, en que se incluya la frecuencia relativa. Use el paquete estadístico para elaborar este diagrama.
2. ¿Cuál es el género de película que tiene mayor preferencia entre los alumnos de ese grupo?

2.33 En la actualidad, hay bebidas de leche y jugo de soya para alcanzar una dieta balanceada. Un equilibrio en la alimentación se establece en 38 % de proteínas, 14 % de humedad, 18 % en grasas, 15 % en carbohidratos insolubles y 15 % de carbohidratos solubles: (estaquiosa, rafimosa). Elabore un diagrama de barras para la información anterior.

2.34 Una empresa desea averiguar el medio en el que se transportan sus empleados para saber el plazo en el que debe ampliarse el estacionamiento. Se entrevistó a 25 personas y la información que se registró fue:

Público	Público	Público	Público	Público
Publico	Público	Caminando	Público	Público
Coche	Caminando	Público	Público	Público
Coche	Público	Bicicleta	Coche	Caminando
Coche	Coche	Público	Público	Coche

1. Elabore una tabla de frecuencia para las respuestas.
2. Dibuje un diagrama de barras.
3. ¿Cuál es el medio de trasporte más utilizado?

2.35 Construya un diagrama tipo pastel para el ejercicio 3 de las actividades de aprendizaje 2.8.

2.36 Se preguntó a 350 estudiantes por el tipo de carrera que tenían pensado estudiar. Luego, se hizo una clasificación en categorías, como la que se muestra en la siguiente tabla:

Área	Frecuencia
Biológicas	89
Administrativas	124
Sociales	96
Ingenierías	25
Ciencias exactas	4
Artes	12
Total	350

1. Calcule la frecuencia relativa.
2. Dibuje un diagrama tipo pastel. Use el paquete estadístico para hacer este diagrama.
3. ¿Qué conclusiones puede obtener?

2.37 En el primer semestre del año, un profesor que revisa las tareas de sus alumnos encontró un incremento en el número de faltas de ortografía. Decidió analizar la situación y aplicó a 175 estudiantes un cuestionario para tratar de detectar las causas. Las causas se resumen en la siguiente tabla:

Causas	Frecuencia
No conocen las reglas de ortografía	65
No leen con regularidad	48
No revisan las notas de clase	23
No hacen la tarea	4
Escriben poco	35
Total	175

1. Elabore el diagrama de Pareto.
2. Interprete los resultados.

2.38 Los datos de la tabla de abajo registran la calidad con que llegó un cargamento de melocotón al supermercado.

Causas	Melocotones perdidos
Dañados	100
Demasiado grandes	87
Podridos	235
Muy maduros	9
Mal clasificados	7
Con gusanos	3
Total	441

1. Realice el estudio de Pareto, es decir, construya la tabla y el gráfico.
2. Interprete los resultados.

2.39 En un estudio para identificar las causas de falla en una flotilla de microbuses se encontró que las causas y las frecuencias de las descomposturas durante un periodo de tres meses fueron: sistema de frenos 30 veces, motor 15 veces, sistema eléctrico 10 y suspensión 5 veces.

1. Describa la tabla de frecuencias
2. Elabore el diagrama de Pareto

El mundo de la información

Vea las páginas que a continuación se indican y escoja un tema de interés y realice actividades relacionadas con este capítulo: 1.-www.inegi.org.mx/rne/docs/Pdfs/Mesa2/20/JavierSalas.pdf. 2.-www.inegi.org.mx/ 3.- www.eumed.net/cursecon/ecolat/mx/2008/mggg2.htm

2.6 Evaluación

En esta evaluación se plantea una serie de cuestiones en las que a partir de un caso concreto se busca que el alumno muestre su comprensión de los conceptos expuestos.

Capítulo 3

Caracterización y resumen de los datos

Pensar es el trabajo más difícil que existe, tal vez por ello tan pocos lo hacen.

Henry Ford

Competencia general

Desarrollar habilidad para determinar las propiedades estadísticas de los datos con el propósito de explicar su significado, así como adquirir práctica en la comparación de conjuntos de datos e interpretación de las características de los datos en la toma de decisiones.

Competencias específicas

- Aprender a calcular la media de un conjunto de datos e interpretarla.
- Ejercitar el cálculo de la mediana de un conjunto de datos e interpretarla.
- Calcular la media y la mediana en una distribución de datos para realizar un comparativo de ambas medidas y así conocer la forma de la distribución.
- Adquirir habilidad en el cálculo de la moda y compararla con la media y la mediana.
- Tener habilidad práctica en el cálculo de los cuartiles de un conjunto de datos.
- Comprender la importancia de las medidas de dispersión en un conjunto de datos, y aprender a calcularlas e interpretarlas.
- Hacer prácticas para el cálculo de la desviación media, varianza y desviación estándar; comprender su importancia para la interpretación de la distribución de datos y la utilidad práctica que estas medidas tienen en el conocimiento de los procesos naturales.
- Explicar la construcción de un diagrama de caja e interpretarlo.

3.1 Introducción

La estadística en acción. Si bien es cierto que la estadística es una herramienta necesaria en muchas áreas experimentales, también nos ayuda a comprender situaciones de nuestra vida real. ¿Cuál es la edad promedio en la que se casan hombres y mujeres? ¿Qué tanto varía una medicina en cumplir la fecha de caducidad? Otras estadísticas importantes son: número promedio de ocupantes por vivienda, promedio de escolaridad, esperanza de vida.

3.2 Medidas de tendencia central

En ocasiones, en la vida cotidiana nos encontramos con que los datos proporcionan una gran cantidad de información. Cuando por ejemplo se dice que el costo promedio de una habitación en diferentes hoteles es de 1000 pesos, se está resumiendo todos los precios obtenidos. Existen, como en el ejemplo anterior, una gran cantidad de situaciones reales en que se da información con la media, lo cual nos permite especificar un valor central de los datos. Sin embargo, la media no es la única referencia de los valores centrales de los datos, también existen otras medidas tales como la mediana, moda, la media geométrica y media armónica para obtener dichos valores. En la primera parte de este capítulo se trata el tema de los valores centrales de los datos. A continuación, en la segunda parte se estudian las medidas de posición relativa tales como los cuartiles y percentiles que permiten comparar valores dentro de la misma colección de datos. Finalmente, se aborda el estudio de un conjunto de medidas que son muy importantes en el análisis de datos, y son relevantes ya que evalúan qué tan dispersos son los datos en torno a la media. Estas se conocen como medidas de dispersión o variación.

A partir del problema que se proponga para estudiar a una población , el procedimiento es seleccionar una muestra de ésta, y con la información que proporcione dicha muestra se puede inferir sobre la población, de tal manera que se tiene un conocimiento aproximado de ésta. A este procedimiento se le conoce como muestreo, y se explicará en el capítulo 7.

Población $\Longrightarrow$	*Selección*	$\Longrightarrow$ **Muestra**
Variable: X		Variable: X
Valores de la variable		Valores de la variable
$x_1, x_2, ..., x_N$		$x_1, x_2, ..., x_n$
N número de unidades		n número de unidades
en la población.		en la muestra.
Parámetros Medidas $\Longleftarrow$	*Inferencia*	$\Longleftarrow$ *Estadísticos Medidas*

Estadística es la asignatura para descubrir más sobre el mundo real mediante la colección, análisis e interpretación de datos.

El estudio de la población parte de establecer la característica o las características que se desea conocer de una población, éstas se plantean mediante variables : X. Por ejemplo, se desea estudiar el gasto, en pesos, semanal en comidas de las familias del barrio de la estación. La variable es X: el gasto en pesos, los valores de esta corresponde a lo que se conoce como datos.

Las medidas de tendencia central pueden calcularse tanto para una población como para una muestra de ésta. Las medidas numéricas que se calculan sobre los datos de la población se conocen como parámetros , y estadísticos cuando se refieren a la muestra.

Parámetro y Estadístico

Un **parámetro** es la medida numérica que se calcula a partir de los datos observados en una población.

Un **estadístico** es la medida numérica que se calcula a partir de los datos observados en una muestra.

Media

El mundo de la información 1. Precipitación pluvial

Si bien llevar un registro de la precipitación pluvial es muy importante en muchas actividades del ser humano, es particularmente valiosa para los agricultores. Por ello, resulta de interés elaborar todos los años un historial de cuánta lluvia ha caído. En México hay una temporada de lluvias que ocurre entre los meses de mayo y agosto. Pero, aunque se tiene el registro de lluvias de varios años, aquí sólo se considerará un periodo de 8 años, correspondientes al mes de julio. Lo descrito en este problema tiene que ver con la economía, así que se cita una página web[1] para que el lector pueda tomar ideas respecto de la producción agrícola. Desde luego que contar con una administración efectiva y eficiente ayudará a utilizar adecuadamente los recursos.

Preguntas sobre la naturaleza del problema

Llevar un registro de la precipitación pluvial nos permite planear muchas actividades como la siembra, la limpieza de bordos para captar agua, el desasolve del alcantarillado en grandes ciudades para evitar inundaciones, etcétera. La media aritmética es una medida adecuada para saber la cantidad de lluvia que cae en cierta zona. También nos permite conocer si en una región llueve más que en otra o comparar entre varios años. La variable X es *la cantidad de lluvia que cae en un año.*

Registro de datos (la cantidad de lluvia se mide en milímetros)

Año	1992	1993	1994	1995	1996	1997	1998	1999
Julio	125.3	98.0	119.2	87.4	92.7	108.0	162.6	149.8

[1] http://cuentame.inegi.org.mx/economia/default.aspx?tema=E

La media se obtiene sumando *los ocho valores* que indican el registro de la precipitación pluvial durante el mes de julio. El total se divide entre las observaciones realizadas, que en este caso es de ocho. La media es:

$$media = \frac{125.3 + 98.0 + 119.2 + 87.4 + 92.7 + 108.0 + 162.6 + 149.8}{8} = \frac{943}{8} = 117.875$$

Por lo general, la media se denota con la letra x con una barra arriba, es decir, $\overline{x} = 117.875$ mm. Se interpreta diciendo que en promedio la lluvia que cae en esa región es de 117.875 mm. En general, la media se expresa de la siguiente manera:

$$\overline{x} = \frac{\text{La suma de todos los valores de la muestra}}{\text{El número total de observaciones}} \tag{3.1}$$

Características de interés para la media

- Algunos se refieren a ella como promedio; desde luego se usa de manera frecuente y resulta un valor de referencia.
- Siempre existe y esta medida toma en cuenta cada valor.
- La media se ve afectada por valores extremos.
- Debido a que satisface ciertas propiedades matemáticas, esta medida tiene un buen desempeño en muchos métodos estadísticos.

Ejemplo 3.1

La administración de una institución financiera selecciona una muestra de 5 asesores, vendedores de estos fondos, con el fin de saber cuál es la media de trámites de afores[2] que han realizado en el último mes. La variable X es el número de trámites de afores durante un mes. Los valores son: 67, 70, 53, 79, 61.

Solución

Usando la expresión (3.1), la media es:

$$\overline{x} = \frac{67 + 70 + 53 + 79 + 61}{5} = 66$$

[2] Administradoras de fondos para el retiro.

Interpretación gráfica de la media: La media de la muestra se puede visualizar como el centro de la balanza de un conjunto de datos. Para ilustrar esta idea, se consideran las cinco observaciones del ejemplo.

En la figura 3.1 se representan estos 5 números en una recta y se coloca un punto de apoyo que balancea los datos. Esta idea intuitiva es importante para comprender otros conceptos estadísticos, por ejemplo que tan cerca o lejos están los datos del punto de equilibrio. ¿Qué ocurre si se quita un valor? La *media de la muestra* (media muestral) es una medida de tendencia central de un conjunto de datos. Ésta se obtiene sumando todos los valores de la muestra divididos entre el número de observaciones

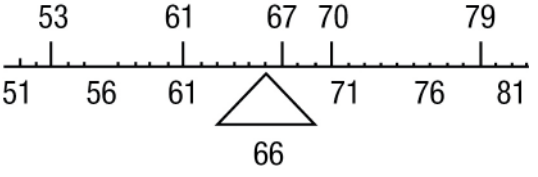


Figura 3.1 Interpretación física de la media.

Suma abreviada

Sumar los valores x_1, x_2, x_3, esto es, SUMAR $(x_1, x_2, x_3) = x_1 + x_2 + x_3$ simbólicamente se representa de la siguiente manera:

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

Notación básica

1. La letra griega mayúscula sigma (Σ) se usa como fórmula para abreviar la suma. Por ejemplo, si tenemos 3 datos con valores x_1, x_2 y x_3 se escribe como suma: $x_1 + x_2 + x_3$. Si se utiliza la notación Σ, entonces la suma se expresa como en la fórmula que sigue. Donde, el número 3 arriba del símbolo Σ indica el número total de datos y el índice i recorrerá los números desde 1 a 3. Así, $\sum_{i=1}^{n} x_i$ es la suma de todos los valores x observados en la muestra.

$$\sum_{i=1}^{3} x_i = x_1 + x_2 + x_3$$

Para el problema de la precipitación pluvial tenemos que $x_1 = 125.3$, $x_2 = 98.0$, $x_3 = 119.2$, $x_4 = 87.4$, $x_5 = 92.7$, $x_6 = 108.0$, $x_7 = 162.6$, $x_8 = 149.8$. En la notación Σ, en resumen, la media para n datos: $x_1, x_2, ..., x_n$, se expresa por la fórmula:

$$\sum_{i=1}^{8} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 = 943$$

En resumen, la media para n datos, $x_1, x_2, \ldots, x_n$, se expresa con la fórmula

$$\text{Media de la muestra:} \qquad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.2}$$

2. La media de la muestra $\overline{x}$ es un *estadístico* , en tanto que la media correspondiente a la población se denota por la letra griega (μ) y es un *parámetro* . En la práctica, es difícil obtener este parámetro por cuestiones económicas y porque, en general, es poco viable y por lo tanto μ es desconocido. Así $\overline{x}$ se puede calcular, puesto que los datos de la muestra están disponibles. Una de las tareas principales de la estadística es estimar el parámetro μ y evaluar ciertas propiedades estadísticas del estimador $\overline{x}$ pero ese es tema de la inferencia estadística, tema que abordaremos en el capítulo 7.

$$\text{Media de la población:} \qquad \mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.3}$$

En resumen, considerando las expresiones (3.2) y (3.3) se tiene:

Población	*Parámetro* μ	Medida: media poblacional:	μ
Muestra	*Estadístico* $\overline{x}$	Medida: media muestral:	$\overline{x}$

La media ponderada

La *media ponderada* es una medida de interés, inicialmente el cálculo de la media considera que cada valor de la variable es de igual importancia. Sin embargo, en diferentes situaciones es de interés darle mayor peso a alguno de los valores observados x_i. En ese sentido cada valor se multiplica por el peso, w_i, que se le dé a la observación, se suman estos valores y se divide entre la suma de los pesos. Esto se obtiene mediante la siguiente expresión:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{3.4}$$

donde w_i es el peso o ponderación asignada a cada valor.

Ejemplo 3.2

Un administrador desea darle mayor peso a la comisión de sus vendedores debido al tipo de cliente que visitan. En este caso tienen determinadas cuatro zonas, en la zona 1 tendrá el doble, la zona 2 el triple, en la zona 3 no se tiene comisión extra y en la zona 4 se quintuplica. La variable X es las ventas, en miles de pesos, realizadas en cada zona, los valores observados son 82, 60, 140, 35. ¿Cuál es la media de ventas?

Solución

En el siguiente cuadro, se describen las operaciones para obtener la media ponderada.

x_i	w_i	$\sum_{i=1}^{n} w_i x_i$
82	2	164
60	3	180
140	1	140
35	5	175
	11	659

Por lo tanto la media ponderada se obtiene aplicando la expresión (3.4)

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{659}{11} = 59.909$$

Media ponderada

La *media ponderada* es una medida de tendencia central que toma en consideración la importancia relativa de los valores observados.

Mediana

Como se vio, la media es una medida que especifica el centro de los datos, pero ésta suele estar afectada por valores extremos: pequeños y grandes. Debido a esta condición, la media puede no representar adecuadamente los datos. Por ello, existen otras medidas que permiten determinar la tendencia central

de los datos. Entre estas medidas destaca la *mediana*.

El mundo de la información 2. Tiempo de espera

Con frecuencia, los tiempos de espera en muchos servicios (transportes, bancos, consultas médicas, restaurantes, etcétera) son un inconveniente y resultan engorrosos para quienes utilizan estos servicios. Para resolver este tipo de situaciones se requiere que la persona que se dedica a actividades de administración proponga soluciones eficientes. Este problema se enfocará a estudiar el tiempo de espera para una consulta médica. Variable: *tiempo de espera.*

Preguntas sobre la naturaleza del problema

Una de las tareas más importantes cuando una empresa se dedica a dar servicio al público es reducir los tiempos de espera. La mediana puede permitirnos conocer esta problemática y ayudar a ejecutar acciones para que la resuelvan.

Registro de datos

Once pacientes necesitaron esperar varios minutos para que los atendieran en la clínica de su localidad, aunque hicieron una cita previa. Para cada paciente, el plazo respectivo de espera en minutos, *los valores de la variable* son:

$$38,\ 42,\ 36,\ 40,\ 44,\ 45,\ 42,\ 36,\ 43,\ 39 \text{ y } 42$$

El procedimiento para calcular la *mediana* del tiempo de espera se divide en dos etapas.

- **Etapa 1.** Se ordenan los datos de menor a mayor. Como se observa en la tabla que aparece abajo, en el segundo renglón se ordenan los datos y en el primero se describe la posición en que quedan éstos después de ordenarlos.

						↓					
Posición	1	2	3	4	5	6	7	8	9	10	11
Orden	36	36	38	39	40	42	42	42	43	44	45

- **Etapa 2**. Se localiza la observación que está exactamente a la mitad de la posición. En este caso, como es número impar, el número que está a la mitad se obtiene por la expresión: $(11+1)/2=6$. La observación que está en la posición 6, en este caso 42, es la mediana de los datos, esto es:

$$\widetilde{m} = 42$$

La interpretación de este problema es que la mitad de los pacientes esperan menos de 42 minutos para entrar a la consulta y la otra mitad más de 42 minutos, lo que en términos reales resulta una demora considerable.

De lo antes expuesto se desprende que la *mediana* de la muestra (mediana muestral) es el valor de la observación ubicada a la mitad en un conjunto ordenado de datos.

Formalización del procedimiento para obtener la mediana

1. Se ordenan los datos de menor a mayor.
2. Se localiza la observación que está exactamente a la mitad de la posición.
   - *a*) Si el número n de observaciones es impar, la mediana es el valor correspondiente a la posición $(n+1)/2$.
   - *b*) Si el número n de observaciones es par, la mediana es la media de los valores que están en las posiciones $n/2$ y $(n/2)+1$.

Ejemplo 3.3

Un profesor desea conocer la mediana de retraso de sus alumnos en su clase. A quienes estén por debajo de la mediana les aplicará un ejercicio más de tarea y a quienes se ubiquen por arriba, dos ejercicios más. Con estas medidas busca recuperar el tiempo perdido por la demora. El primer día registró el tiempo que 8 alumnos de uno de sus grupos tardaron en llegar a clase. Variable: *tiempo de retraso*, los *valores* de la variable son: 10, 11, 12, 14, 16, 17, 17 y 20 minutos.

Solución

					↓				
Posición	1	2	3	4		5	6	7	8
Orden	10	11	12	13		16	17	17	20

Como hay un número par de observaciones, se seleccionan los valores que corresponden a los datos que están en las posiciones:

$$\frac{n}{2} \text{ y } \frac{n}{2}+1$$

respectivamente, que para el ejemplo son 8/2=4 y 8/2+1=5. Los valores que corresponden a las posiciones 4 y 5 son 14 y 16. La media en este caso es la media de 14 y 16, es decir:

$$\frac{14+16}{2}=15$$

Los alumnos que lleguen con un retraso menor a los 15 minutos harán un ejercicio más de tarea, y quienes lleguen con un retraso mayor a 15 minutos tendrán dos ejercicios más de tarea.

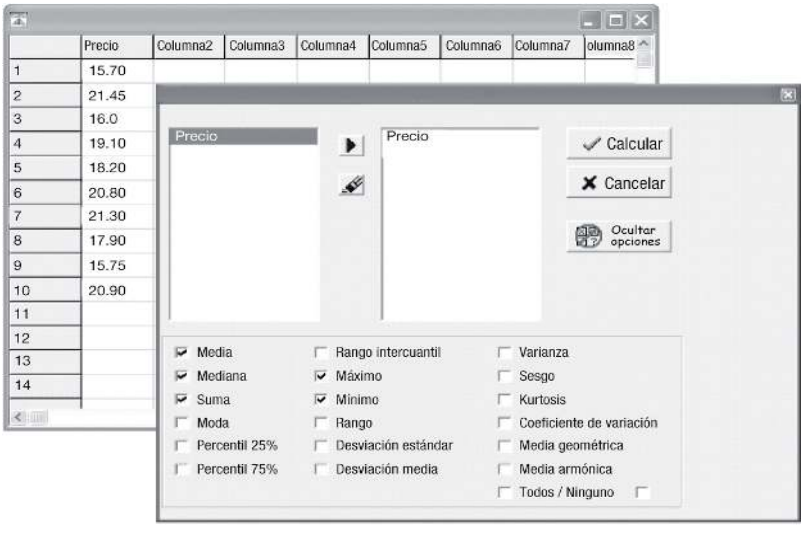


Figura 3.2 Hojas de captura, y de registro para el cálculo de estadísticas de un conjunto de datos. En este caso la media, mediana, suma, y valores máximo y mínimo.

Solución mediante el uso de CalEst

Usando **CalEst** se ingresan los datos de cualquier muestra para generar la media, la mediana y otras medidas estadísticas casi de manera inmediata. Los cálculos se ilustran con el ejemplo 3 y el procedimiento se muestra en los siguientes pasos:

1. A partir del paquete estadístico **Calculadora Estadística** en **CalEst** se abre el módulo **Estadística**.
2. En **CalEst** aparecerá una pantalla como la que se observa en la figura 3.2.
3. En la columna **Datos** se anotan los datos que se generen de cualquier estudio. Una vez anotados todos los datos, con el botón izquierdo del mouse se oprime la tecla **Calcular** y entonces aparecen los valores de la media y la mediana, entre otras medidas estadísticas.

Ejemplo 3.4

En un estudio de precios al consumidor, se reporta el precio en pesos, de un litro de aceite comestible de una marca específica, variable: *precio del litro de aceite*. Los precios de una muestra aleatoria en 10 tiendas diferentes son: 15.70, 21.45, 16.0, 19.10, 18.20, 20.80, 21.30, 17.90, 15.75, 20.90. Usando **CalEst**,

obtener la media y la mediana. Comparar los resultados de los valores de estas medidas. ¿Los valores de las medidas de tendencia central que se obtuvieron en este caso pueden orientar al consumidor respecto dónde comprar una botella de aceite?

Solución

En el menú de programas del paquete está el de Estadística y en éste se encuentra la opción **CalEst**. Aplicando los tres pasos para el uso de **CalEst**, se capturan los datos del ejemplo en la columna de datos. Se oprime el botón izquierdo del mouse en **Calcula** y se obtienen, entre otras medidas estadísticas que se verán más adelante, la media y la mediana (figura 3.3).

Resumen de Medidas

	Precio
n	10
Media	18.71000
Mediana	18.65000
Suma	187.10000
Máximo	21.45000
Mínimo	15.70000

Figura 3.3 Descripción del reporte de salida del **CalEst**.

Interpretación

Como se puede observar de la figura 3.3, la media ($\overline{x} = 18.71$) es ligeramente mayor a la mediana ($\widetilde{m} = 18.65$), lo que indica que el precio de al menos una botella de aceite es alto y esto provoca que la media se cargue a la derecha. Por su parte, la mediana indica que menos de 50 % de botellas cuesta menos de 18.65 pesos. Este valor orienta al consumidor en saber que existen lugares donde es más barato el aceite. La interpretación en función de la media indica que hay lugares donde el aceite es más barato a 18.71. Para tener una información más integral de lo que indican estas medidas, existen las medidas de dispersión que se verán más adelante.

Población	*Parámetro* Me	Medida: mediana poblacional:	Me
Muestra	*Estadístico* $\widetilde{m}$	Medida: mediana muestral:	$\widetilde{m}$

Comparación de la media y mediana

Hemos visto que la *media y la mediana* nos proporcionan diferente información sobre el conjunto de datos, y asimismo, con el propósito de ejemplificar su cálculo, ambas se obtuvieron para un conjunto pequeño de datos. En el capítulo 2 se estudió, para una muestra, la distribución de los datos para conjuntos grandes y pequeños de datos. Ahora, además de estudiar la distribución para un conjunto de

datos, también se puede calcular su media y mediana.

Observación

En general no se da una expresión para la mediana de la población, aquí la consideramos con el fin de tener presente la idea de la relación población y muestra.

Entre las características de interés para la mediana se encuentran las siguientes:

- tiene un uso frecuente;
- siempre existe y no toma en cuenta cada valor;
- no se ve afectada por valores extremos, y es apropiada en estos casos.

El mundo de la información 3. Competencia atlética

A pesar de que en nuestro país una gran cantidad de personas vive obsesionada con el futbol, muy pocos individuos practican deporte. De las actividades deportivas menos ejercitadas está el atletismo. Por ello, algunos profesores de educación física interesados en estimularlo, organizan encuentros en diferentes disciplinas del atletismo. El departamento de actividades deportivas de un bachillerato organizó, entre otras, una carrera de 100 metros con 112 alumnos para evaluar su potencial atlético.

Preguntas sobre la naturaleza del problema

La información que se obtenga de los resultados, en particular de la prueba de 100 metros, permitirá elaborar programas de acondicionamiento físico-atlético en el bachillerato en cuestión. Los resultados de esta actividad determinan cuál es realmente la condición física de los jóvenes y podrían motivar a que los estudiantes mejoren su rendimiento físico. Los datos registrados en segundos mediante un cronómetro, esto es, los *valores* de la variable, fueron:

17.59	17.04	15.85	15.07	15.86	16.18	16.88	15.14	14.92	16.89	17.43
15.15	15.51	17.49	17.12	15.67	17.07	15.72	18.13	16.52	16.33	17.31
18.29	17.43	16.36	17.47	17.75	20.48	18.79	17.88	15.97	16.98	17.07
18.74	16.42	18.69	19.85	15.01	15.54	16.88	16.16	16.47	14.34	16.93
15.1	14.82	16.93	17.95	18.09	18.44	18.0	16.65	16.79	15.79	15.78
18.36	15.61	16.39	17.51	17.36	17.47	17.16	17.14	15.52	14.51	15.51
15.44	15.7	16.45	19.78	14.49	16.09	16.08	16.01	16.78	18.11	15.94
19.34	16.48	16.63	18.88	15.77	16.43	16.45	15.93	16.61	15.86	15.68
15.96	15.55	15.22	19.75	15.63	18.17	16.1	15.54	15.74	15.66	17.34
15.51	17.94	15.74	15.27	20.70	16.37	16.65	17.55	16.94	15.62	15.76
17.95	16.97									

La media de los 112 datos es:

$$\overline{x} = \frac{1871.81}{112} = 16.71$$

y la mediana es:

$$\widetilde{m} = 16.475$$

Ambas medidas de tendencia central son diferentes; sus diferencias pueden verse claramente en las gráficas de la figura 3.4, donde se muestra la distribución de los datos mediante el polígono de frecuencias relativas (izquierda) y el polígono de frecuencias relativas acumuladas (derecha).

Análisis usando la tecnología

Use la calculadora estadística para comprobar los valores de la media, mediana y para hacer el polígono de frecuencia con respecto a los resultados de la competencia atlética.

La mediana, como puede observarse, está a la izquierda, por lo tanto es menor que la media. Ello ocurre porque la media está influenciada por los valores grandes a la derecha de la distribución. En el contexto del problema, esto puede interpretarse diciendo que algunos participantes tardaron más en recorrer 100 metros.

La relación entre media y mediana puede formalizarse con la distribución de los datos. El polígono de frecuencia relativa en la gráfica izquierda de la figura 3.4 describe la distribución para la muestra de 112 corredores.

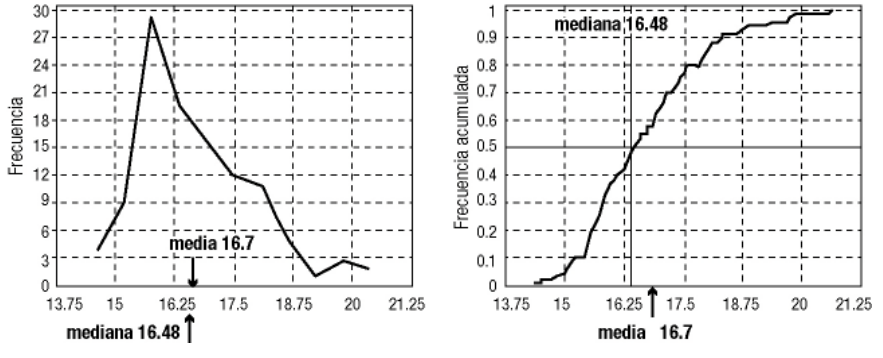


Figura 3.4 Polígonos de frecuencia relativa y acumulada para los tiempos registrados por los participantes en una prueba de 100 metros.

Relación entre la distribución de una muestra y la de una población

La media y la mediana permiten ganar información sobre la distribución muestral de los datos, tal y como se refleja en la figura 3.4. La relación entre los valores de la media y mediana proporcionarán una idea de la forma de la distribución. Recordemos que la muestra es parte de una población, y el objetivo

principal de la estadística es describir características relevantes sobre la población de datos coleccionados de una muestra.

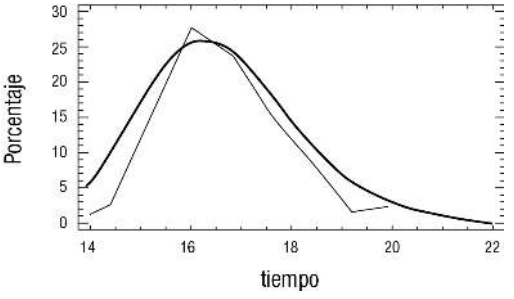


Figura 3.5 Polígono de frecuencias relativas para la muestra y distribución de la población en los tiempos registrados por los participantes en una prueba de 100 metros.

Si la prueba atlética se aplicara a más estudiantes del bachillerato, la muestra crecería (se obtiene mayor información sobre la población), y el ancho del intervalo de clase se reduciría; de esa manera se suavizan los picos en el polígono de frecuencias. Si se consideraran a todos los estudiantes de bachillerato del estado al que pertenece la escuela, se estaría hablando de una población. El polígono de frecuencias relativas se aproximaría al de la figura 3.5. En este caso, se tendría la distribución de una población, que serían los tiempos realizados por todos los estudiantes de bachillerato en una carrera de 100 metros.

En general, la distribución de datos se interpreta de tres maneras:

1. Si la media y la mediana son iguales, se dice que la distribución es simétrica y no tiene sesgo (gráfica central de la figura 3.6).

2. Si la media es menor que la mediana y los datos están cargados a la derecha, la distribución tiene un sesgo negativo (gráfica a la derecha en la figura 3.6).

3. Finalmente, si la media es mayor que la mediana y la mayoría de los datos están a la izquierda, la distribución tiene sesgo positivo (gráfica a la izquierda de la figura 3.6).

En resumen,

Si	Distribución	De la figura 3.6:
Media = Mediana	Simétrica	Al centro
Media <Mediana	Sesgada a la izquierda	A la izquierda
Media >Mediana	Sesgada a la derecha	A la derecha

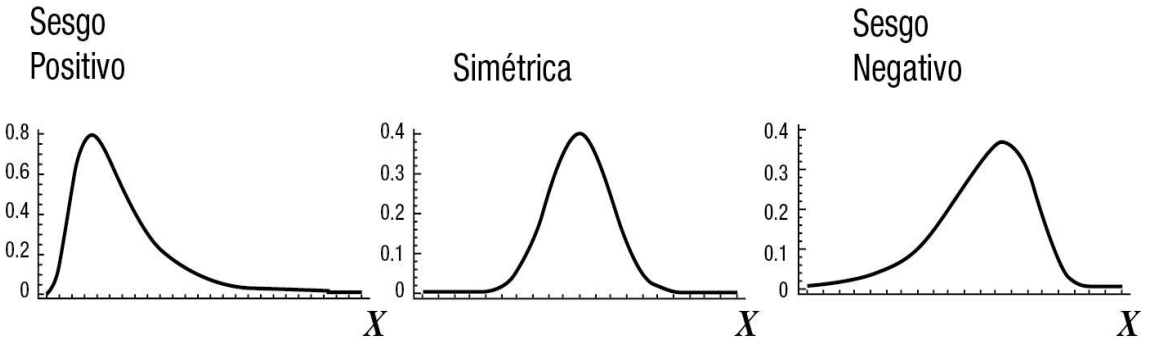


Figura 3.6 Ilustra el sesgo y simetría de la distribución de la variable X.

Moda

En esta parte se ahondará en el cálculo de las medidas estadísticas que, junto con la media y la mediana, permitirán completar la información que proporcionan los datos, comprender mejor el problema que es objeto de estudio y tener una mayor interpretación de los datos.

La moda es otra medida de tendencia central y ésta se define de como se muestra en el siguiente ejemplo.

La *moda* es otra medida de tendencia central y ésta se define como el valor de la observación que más veces se repite en la muestra (o en la población); es decir, el de mayor frecuencia, y se denotará por m_o.

Ejemplo 3.5

La administración de una universidad quiere evaluar la calidad de la comida que ofrece la cafetería a sus estudiantes, para decidir si renueva el contrato a la empresa de alimentos que tiene contratada. Se aplica una encuesta a una muestra de estudiantes que acuden a comer con regularidad a la cafetería. A los estudiantes se les pidió que calificaran la calidad de la comida en una escala de 1 a 5, donde 1 es extremadamente mala y 5 extremadamente buena. Los resultados son:

1	1	1	1	1	1	1	1	2	2	2	3	3	3	3
3	3	3	3	3	3	3	3	4	4	4	4	4	4	4

1. ¿Cuál es la media de la calificación?
2. ¿Cuál es la mediana de la calificación?
3. ¿Cuál es la moda de esta calificación?
4. Con la información, ¿se puede decir que los estudiantes entrevistados tienen una opinión favorable, neutral o desfavorable para la calidad en la comida?

Solución

1. La media es:

$$\overline{x} = \frac{\sum_{i=1}^{30} x_i}{30} = \frac{78}{30} = 26$$

2. Como se tienen 30 datos, la mediana es la media de las observaciones que están en las posiciones: $\frac{30}{2} = 15$ y $\frac{30}{2} + 1 = 16$, es decir: $\frac{3+3}{2} = 3$.

3. La tabla de frecuencias es:

Evaluación	1	2	3	4
Frecuencia	8	3	12	7

La moda es 3.

Interpretación

Tanto la mediana como la moda permiten concluir que los estudiantes tienen una posición neutral para la calidad de la comida. En términos de la media, la opinión es ligeramente desfavorable.

Población	*Parámetro* Mo	Medida: moda poblacional:	Mo
Muestra	*Estadístico* m_o	Medida: moda muestral:	m_o

Observación

Como ocurre con la mediana, no se propone un símbolo especial para la moda de la población. Algunas características de interés de la moda on las siguientes:

- podría no existir;
- puede haber más de un valor;
- no toma en cuenta cada valor;
- tiene un mayor uso para datos en escala nominal.

3.3 Medidas de posición

Cuartiles

Los *cuartiles* se conocen como medidas de posición relativa, dan información sobre la posición de una observación en la muestra. Éstos se consideran medidas adicionales a la media y mediana.

El cuartil es un indicador que permite completar un resumen estadístico de los datos. En el capítulo anterior se relacionó la gráfica de frecuencias relativas acumuladas con los cuartiles, ver figura 3.4. En esa descripción la mediana es el 50 percentil. Otros números de interés para obtener los cuartiles son los correspondientes al 25 percentil y al 75 percentil. Estos tres percentiles (25, 50 y 75 percentil) se denominan cuartil inferior, medio y superior, respectivamente, y se denotan por C_1, $\widetilde{m}$ y C_3.

Construcción del cuartil

- **Paso 1**. Ordenar los datos de menor a mayor y encontrar la mediana.
- **Paso 2**. El primer cuartil (C_1) de todas las observaciones es la mediana del conjunto de datos que está por debajo de la mediana.
- **Paso 3.** El primer cuartil (C_3) de todas las observaciones es la mediana del conjunto de datos que está por arriba de la mediana.

Segundo cuartil

Es frecuente referirse a la mediana como el segundo cuartil, es decir: $C_2 = \widetilde{m}$

Ejemplo 3.6

En una competencia mundialista, las diferencias en décimas de segundo de la corredora que ganó con respecto a sus contrincantes en los 400 metros planos fueron: 15, 48, 56, 59, 78, 96, 124. Encontrar la mediana, el primer y tercer cuartiles.

Posición	1	2	3	4	5	6	7
Valores	15	48	56	59	78	96	124
		↑		↑		↑	
		C_1		$\widetilde{m}$		C_3	

Solución

- **Paso 1.** Ordenar los datos de menor a mayor.
- **Paso 2.** Si el número de datos es impar, la mediana es el dato que está en la posición $(n+1)/2$, que en el ejemplo sería $(7+1)/2 = 4$. Así la mediana es:

$$\widetilde{m} = 59$$

- **Paso 3.** Para determinar el primer cuartil, se toma la mitad de los datos inferiores (todos los valores ubicados por debajo de la mediana) y se encuentra la mediana de éstos. En el caso del ejemplo, la mediana de 15, 48 y 56 es el primer cuartil, esto es:

$$C_1 = 48$$

- **Paso 4**. De manera análoga, el tercer cuartil es la mitad de los datos superiores, es decir, todos los valores mayores que la mediana. En el ejemplo son 78, 96 y 124, donde el tercer cuartil es:

$$C_3 = 96$$

Finalmente, cabe destacar los valores mínimo y máximo (15 y 124 en el ejemplo de la corredora). Estos cinco números son la base que permitirá construir el diagrama de caja, que se verá más adelante.

Pensando en los cuartiles, éstos dividen los datos en cuatro partes. La mediana divide al conjunto de datos a la mitad, y si tomamos la mitad de una mitad se obtiene un cuarto; éstos son los que deseamos.

Interpretación

El primer cuartil $C_1 = 48$ indica que menos de 25 % de las corredoras están por debajo de ese registro, lo que señala que esas competidoras estuvieron cerca de la victoria. El valor $C_3 = 96$ corresponde a 75 % de las corredoras y revela que 25 % de ellas quedaron lejos del primer lugar. En la práctica, esta situación sirve de referencia a las competidoras para mejorar sus marcas.

Aprender a calcular un cuartil es importante porque permite ganar un mayor conocimiento sobre un problema o tema de interés a través de la información proporcionada por los datos.

Ejemplo 3.7

En la actualidad se efectúan pruebas sicológicas para medir la tendencias a la agresividad con el fin de ayudar a las personas a reducir esas inclinaciones. Se aplicó una prueba psicológica a 11 jóvenes que están bajo un tratamiento médico para medir la tendencia a la agresividad. La prueba evalúa de 10 a 50,

donde 50 indica el grado mayor de agresividad. La variable es: *calificación de tendencia a la agresividad*, los valores de la variable están descritos por los siguientes datos:

38, 27, 44, 39, 41, 26, 35, 45, 39, 28 y 16.

Solución

1. $$\text{La media :} \quad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \text{la suma :} \quad \sum_{i=1}^{n} x_i = 378, \text{entonces} \quad \overline{x} = \frac{378}{11} = 34.36.$$

2. La mediana, ordenar los datos:

Posición	1	2	3	4	5	6	7	8	9	10	11
			C_1			$\Downarrow$			C_3		
	16	26	27	28	35	38	39	39	41	44	45
			$\uparrow$			$C_2 = \widetilde{m}$			$\uparrow$		

Como el número de datos es impar, la mediana es el número que está en la posición:

$$\frac{n+1}{2} = \frac{11+1}{2} = 6; \text{entonces} : \widetilde{m} = 38.$$

3. La moda: para este conjunto de datos el valor más frecuente es el 39.

4. El primer cuartil C_1 es la mediana del conjunto de datos que están por debajo de la mediana, es decir: 16, 26, 27, 28, 35 entonces la mediana de este conjunto es $C_1 = 27$.

5. El tercer cuartil C_3 es la mediana del conjunto de datos que están por arriba de la mediana, es decir: 39, 39, 41, 44, 45. Entonces la mediana de este conjunto es $C_3 = 41$.

El resumen estadístico se presenta en la tabla 3.1.

Tabla 3.1 Resumen estadístico de las calificaciones de tendencia a la agresividad.

Estadística descriptiva								
n	$\overline{x}$	$C_2 = \widetilde{m}$	m_0	C_1	C_3	$Mín$	$Máx$	$Suma$
11	34.36	38	39	27	41	16	45	378

Se puede observar que las medidas de tendencia central dan valores altos para la calificación de agresividad, lo que indica que los jóvenes tienden a ser agresivos.

Resumen técnico para los cuartiles

Anteriormente se presentó la mediana de un conjunto de datos, la cual es una medida que corresponde exactamente al punto medio de los datos ordenados de menor a mayor, es decir, el 50 % de los datos son menores o iguales a la mediana y el 50 % son mayores o iguales a ésta. De esa manera la mediana divide a los datos en dos partes iguales. En ese mismo sentido los tres cuartiles, denotados por C_1, C_2 y C_3, dividen a los datos en cuatro partes iguales.

1. **Primer cuartil** C_1 : *El primer cuartil* es el valor de los datos ordenados que representa al 25 %, es decir: al menos el 25 % de los valores de datos ordenados son menores o iguales a C_1, y al menos 75 % de los valores son mayores o iguales a C_1.

2. **Segundo cuartil** C_2 : *El segundo cuartil* es el valor de los datos ordenados que corresponde a la mediana.

3. **Tercer cuartil** C_3 : *El tercer cuartil* es el valor de los datos ordenados que representa al 75 %, es decir *:* al menos el 75 % de los valores de datos ordenados son menores a C_3, y al menos 25 % de los valores son mayores a C_3.

Ejemplo 3.8

Caso 1: En un sistema educativo ¿qué tan preparados están los estudiantes? Un profesor de historia aplica un cuestionario a sus alumnos sobre temas de historia universal y nacional, el cuestionario consta de 20 preguntas de opción múltiple. La variable es el número de respuestas correctas, los resultados de una muestra de esos 15 alumnos son: 16, 9, 13, 15, 16, 19, 8, 11, 12, 6, 20, 17, 10, 18, 5. Encuentre los cuartiles C_1, $C_2 = \widetilde{m}$, C_3.

Posición	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
				↓				↓				↓			
Valores	5	6	8	9	10	12	12	13	15	16	16	17	18	19	20
				C_1				C_2				C_3			

Así, menos un cuarto de alumnos tiene 9 o menos aciertos. La mitad tiene 13 aciertos y el 75 % tiene 17 o más aciertos.

Caso 2: En un proceso de administración bancaria, la atención en minutos para una muestra de 8 clientes, la variable tiempo de atención. Los valores se registran a continuación:

Posición	1	2		3	4		5	6		7	8
			↑			↑			↑		
Valores	2	3		6	12		17	21		28	34
			C_1			C_2			C_3		

Encuentre los cuartiles C_1, $C_2 = \widetilde{m}$, C_3.

Solución

Primero se determina la mediana, se divide el grupo de datos en dos, puesto que el número de datos es par, la mediana esta entre la posición 4 y 5. Entonces se calcula la media entre los valores que ocupan esas posiciones respectivamente, esto es: $(12 + 17)/2 = 14.5$, así $C_2 = \widetilde{m} = 14.5$. El primer cuartil se calcula considerando la mediana de los datos que están abajo de la mediana $\widetilde{m}$. La mediana correspondiente es la que se encuentra entre las posiciones 2 y 3 de esa manera se obtiene el promedio de los valores 3 y 6, por lo tanto $C_1 = (3 + 6)/2 = 4.5$. Analógamente se encuentra el tercer cuartil, ahora calculando la mediana de los valores arriba de la mediana $\widetilde{m}$ en este caso se obtiene la media entre los valores 21 y 28, así $C_3 = (21 + 28)/2 = 24.5$. Observación: el cálculo de estos valores usando los paquetes estadísticos, tienen, por decirlo así, una mejor precisión. Aunque en algunos varían ligeramente. En particular el cálculo usado en **CalEst**, aparece en el último ejemplo del presente capítulo.

En el siguiente ejemplo se considera un conjunto de datos mucho mayor, con la finalidad de considerar la descripción tratada en el capítulo anterior y los cálculos de las medidas de tendencia central y posición.

Ejemplo 3.9

Las compañías de seguros les piden a sus posibles clientes que se realicen pruebas para verificar su estado de salud. Esta acción va en el sentido de que éstas no tengan pérdidas económicas. Un centro escolar realiza un estudio médico a una muestra de 62 profesores. Entre los resultados del estudio se encuentran la variable X que son los niveles de colesterol, que tiene los valores:

167, 184, 192, 198, 200, 202, 210, 211, 212, 215, 216, 217, 218, 220, 225
225, 226, 230, 230, 230, 230, 231, 232, 232, 232, 234, 234, 236, 236, 238,
240, 243, 246, 247, 248, 254, 254, 254, 256, 256, 258, 263, 264, 267, 267,
268, 268, 270, 270, 272, 278, 278, 283, 285, 300, 300, 309, 327, 332, 336,
355, 394.

Mediante un diagrama de polígono de frecuencia ilustrar el cálculo de los cuartiles. Presentar un resumen estadístico de las medidas de tendencia central. Un nivel de colesterol mayor a 250 requiere ya de atención médica, ¿qué porcentaje de los profesores tienen que tratarse por rebasar el nivel de colesterol establecido? ¿Qué información proporcionan los cuartiles para ubicar problemas en los niveles de colesterol alto?

Solución

En la figura 3.7 se reproduce la gráfica del polígono de frecuencias acumuladas para los datos del colesterol y en éste se señalan los 3 cuartiles. Estos valores aproximados, interpretación a ojo, son:

$$C_1 = 224, C_3 \ = 271 \text{ y } \widetilde{m} = 242.$$

El máximo y mínimo son medidas que se integran a los tres indicadores anteriores para darnos una buena idea de la distribución de los datos. En el ejemplo el $mínimo = 394$ y el $máximo = 167$. Hay que observar que en este caso a la gráfica le falta precisión, pero da una idea aproximada de por donde están los cuartiles.

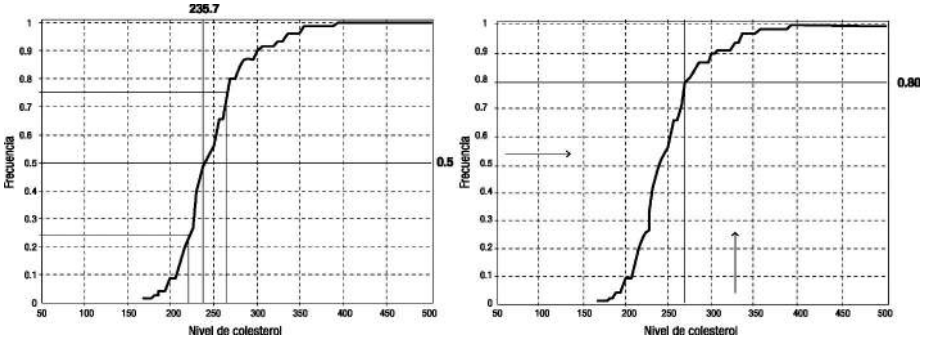


Figura 3.7 Polígono de frecuencias acumuladas expresado en porcentajes.

En la tabla 3.2 se presenta un resumen estadístico del cálculo de las medidas de tendencia central correspondientes a los niveles de colesterol. Por lo general, el resumen estadístico es una manera de entregar un reporte sobre el estudio de un tema de interés, ya que sintetiza la información que genera el conjunto de datos.

Tabla 3.2 Resumen estadístico del nivel de colesterol.

n	$\overline{x}$	$C_2 = \widetilde{m}$	m_0	C_1	C_3	$Mín$	$Máx$	$Suma$
62	250.08	241.5	230	225	268.5	167	394	15505

Interpretación

El tercer cuartil es $C_3 = 268$, lo que significa que 25 % (aproximadamente 16 profesores) tiene un nivel de colesterol por arriba de 268. Al menos estos profesores necesitan un tratamiento para reducir el colesterol.

En la gráfica izquierda de la figura 3.7 trace, a ojo, una línea "más clara" que va desde el nivel de colesterol 250 al polígono de frecuencia y luego al eje vertical. Ahí se observa un valor cercano a 56 %, es decir, aproximadamente 35 profesores de los 62 satisface el nivel de colesterol. En caso contrario 27 profesores rebasan el límite establecido, que resulta alto, por lo que tienen que someterse a un tratamiento médico, con las respectivas condiciones económicas. La mediana y el primer cuartil muestran valores que están por debajo del nivel de colesterol establecido, y sirven de referencia para saber qué tan alto está el colesterol del 50 % y 25 % de los profesores.

Complemento técnico: El procedimiento para la obtención de los cuartiles no es un resultado universalmente uniforme puesto que diferentes programas estadísticos dan distintos valores para los cuartiles

ya que usan diferentes técnicas para encontrarlos. El método que hemos propuesto para calcular cuartiles no es exacto, pero resulta más sencillo y se obtienen valores muy cercanos a los que se llegarían por métodos más exactos. Por lo tanto habrá ligeras diferencias cuando usemos distintos paquetes; un paquete estadístico es un conjunto de programas de cómputo para calcular las estadísticas, elaborar diversas gráficas y hacer análisis estadístico.

Solución mediante el uso de CalEst

Los cuartiles se pueden obtener utilizando **CalEst**. En particular los resultados del ejemplo 3.9 usando este medio aparecen en la tabla 3.2. El método empleado por **CalEst** se muestra al final del resumen de este capítulo.

Población	*Parámetro*	Medidas de posición para poblacional:
Muestra	*Estadístico*	Medidas de posición para la muestral:

Observación. Las medidas de posición también existen para la población, es decir, los parámetros de posición, y corresponden al modelo teórico de la distribución de los datos.

Percentiles

Con la finalidad de completar la exposición sobre las medidas de posición en los datos, a continuación se presenta el tema de los percentiles. En la distribución de los datos va desde la observación 1 hasta la observación n, por ejemplo, en el caso de los niveles de colesterol se tiene desde la observación 1 hasta la observación 64. Existe el interés de conocer un determinado percentil de la información, digamos el 0.10, 0.15, 0.90, 0.95, éstos se pueden expresar en porcentajes, es decir, 10 %, 15 %, 90 % o el 95 % de los datos. Se denotan por P_{10}, P_{15}, P_{90} y P_{95}, en general para m es P_m y se lee el emésimo percentil. El m emésimo percentil divide al grupo de datos, el m % de los datos que están por abajo y el (100-m) % que está por arriba. ¿Estos porcentajes a qué valor del nivel de colesterol corresponden? En notación el valor que corresponde a un percentil se denota con x_p.

El planteamiento es encontrar el emésimo percentil. La idea es encontrar la posición entre los n datos ordenados que corresponde a P_m. El punto es evaluar la expresión $\left(\frac{m}{100}\right) \times n$ esto es:

Procedimiento para encontrar los percentiles
1. Si $\left(\frac{m}{100}\right) \times n$ es entero, entonces
La posición entre los datos de P_m es $\left(\frac{m}{100}\right) \times n + 0.5^{(*)}$
2. Si $\left(\frac{m}{100}\right) \times n$ no es entero, entonces
La posición entre los datos de P_m es el siguiente entero.
Así el percentil P_m es valor del dato en esa posición

Interpretación

El P_m es la media entre los valores que están entre la posición $\left(\frac{m}{100}\right) n$ entera y el de la siguiente posición. En resumen: Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales. Nota: Los resultados son aproximados.

Ejemplo 3.9a

Para lo datos de los niveles de colesterol entre los profesores (ejemplo 3.9), encontrar 1. P_{80} y 2. P_{50}.

Solución

1. Se tiene $m = 50$, por tanto

$$\left(\frac{m}{100}\right) n = \left(\frac{50}{100}\right) 62 = 31$$

   La posición es $31 + 0.5$, entonces el valor es la media entre los valores que están en la posición 31 y 32. Éstos corresponden a 240 y 243, de esa manera la media de esos datos es el valor 241.5. Observe que este último valor corresponde a la mediana en el ejemplo 3.9.

2. En el caso $m = 80$ se sigue que:

$$\left(\frac{m}{100}\right) n = \left(\frac{80}{100}\right) 62 = 49.6$$

   Así, la posición corresponde al dato en el lugar 50, entonces el 80 percentil $P_{80} = 272$, véase la figura 3.7, gráfica derecha. Nuevamente se resalta que esta figura es una gráfica empírica, porque contiene los datos de la muestra, y por lo tanto sólo da una idea aproximada de la distribución de los datos. Obsérvese que del eje vertical se puede señalar el percentil que una persona desee, llegar a la línea y bajar de manera perpendicular al eje horizontal. Es de interés notar que el procedimiento también es recíproco, es decir, ir del eje horizontal, esto es dar un valor de la variable y contrastar el porcentaje que le corresponde en el eje vertical, tal como lo destacan las flechas horizontal y vertical. A continuación se explica la técnica inversa de los percentiles.

Proceso inverso en percentiles

Es de utilidad dado un valor del conjunto de datos, saber a qué percentil pertenece. En el ejemplo se vio que 272 fue el 80 percentil. Dicho de otra manera, se dice que 80 es el rango percentil del nivel

de colesterol 272. Se puede decir que el rango percentil es el proceso inverso, ahora se da un nivel del colesterol y hay que determinar cuál es el percentil Se obtiene a partir de la siguiente regla.

Regla para determinar el rango percentil

El rango percentil m del valor de un dato x es:

$$m = \left(\frac{Número\ de\ datos\ menores\ que\ x}{Número\ total\ de\ datos} \right) \times 100$$

Hay que destacar que los resultados son aproximados.

Ejemplo 3.9b

A partir de los datos del ejemplo 9, encontrar el rango percentil para el dato 226.

Solución

Observe que hay 16 datos antes que 226 de la lista de 62 observaciones. El cálculo es:

$$m = \left(\frac{16}{62} \right) 100 = 25.8 \approx 26$$

Entonces el 226 es el 26 percentil P_{26}, muy cercano al primer cuartil.

La media armónica y la media geométrica

La *media armónica* y la *media geométrica* completan el panorama de medidas de tendencia central. En la práctica, éstas suelen usarse menos y en general para aplicaciones específicas, por ejemplo en comercio y economía.

La media armónica

La *media armónica* (que se denota con la letra H) de un conjunto de datos $x_1, x_2, \ldots, x_n$ es el recíproco de la media aritmética del recíproco de esos datos:

$$H = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_I}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_I}}$$

La media armónica se emplea cuando se desea promediar velocidades, tiempos, rendimiento, etc., es decir, cuando influyen los valores pequeños. Pero es necesario ser cuidadoso en estos casos, ya que cuando algún dato es cero o muy cercano a cero, no se puede calcular.

Ejemplo 3.10

La administración de una empresa está interesada en conocer la producción por hora de las cinco máquinas que tiene operando, las cuales se usan para producir un componente para el asiento de un automóvil. Cada una de las cuatro máquinas lo hace en 3.5, 2.0, 1.5, 5.5 y 4 minutos. ¿Cuál es la velocidad promedio? ¿Cuántos componentes producen las máquinas en una hora?

Para encontrar la solución a este ejemplo recurrimos a una medida de tendencia central que se conoce como media armónica. Con el fin de facilitar su definición, primero se desglosan algunos términos.

Solución

Para encontrar la media armónica se aplica la fórmula a los cinco tiempos registrados por las máquinas.

$$H = \frac{1}{\frac{1}{5}\left(\frac{1}{3.5} + \frac{1}{2.9} + \frac{1}{1.5} + \frac{1}{5.5} + \frac{1}{4}\right)} = 2.65\text{minutos}$$

Este resultado indica que la producción de las máquinas en una hora, 60 minutos, es (5*60)/2.65=113 componentes.

Promedio de los recíprocos

- El recíproco de un número x es $\frac{1}{x}$
- El recíproco de n números $x_1, x_2, \ldots, x_n$ es $\frac{1}{x_1}$, $\frac{1}{x_2}$, …, $\frac{1}{x_n}$
- El promedio de n recíprocos es $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_I}$
- El recíproco de la media de n recíprocos es $\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_I}}$

La media geométrica

Si algunos valores son muy grandes en magnitud y otros son pequeños, entonces la media geométrica es una medida que representa los datos mejor que la media.

Si hay n observaciones $x_1, x_2, \ldots, x_n$, la *media geométrica* G de un conjunto de datos es la raíz enésima del producto de esos datos.

$$G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

Por lo general, la media geométrica se utiliza cuando los valores de la variable siguen una progresión geométrica, o cuando se necesitan promediar porcentajes, tasas, índices, etc., siempre que vengan dados en porcentajes.

Ejemplo 3.11

Los registros estadísticos de la Secretaria de Salud señalan que el número de personas que se enferman a causa de su hábito de fumar aumenta año con año. El hábito de fumar tiene repercusiones económicas en el sector salud.

El registro de un hospital indica que el número de personas de 25 años que se enferma a causa de su hábito de fumar aumenta año con año. ¿Cuál es el porcentaje promedio de incremento en la cantidad de enfermos que ocurren por año? A continuación se muestra una tabla donde se reportan los datos:

Año	Número de enfermos	Razón al valor del año anterior
1990	5	x
1991	12	2.40
1992	16	1.30
1993	25	1.56
1994	40	1.60
1995	65	1.63

Solución

Aplicando la fórmula para el promedio geométrico tenemos:

$$G = \sqrt[n]{x_1 x_2 ... x_n} = \sqrt[5]{2.4 \times 1.33 \times 1.56 \times 1.6 \times 1.63} = 1.669$$

es decir, un aumento medio por año de 66.9 %. Este resultado indica que en los últimos cinco años en promedio, el número de enfermos por el hábito de fumar ha crecido. Con esta información las autoridades encargadas de la salud deben aplicar planes para remediar esta situación.

3.4 Medidas de dispersión

Con seguridad, en alguna ocasión habrá visto o practicado el tiro al blanco. El objetivo del juego es atinarle a ese blanco. Si se hacen diez tiros es casi seguro que no todos den en el blanco; aunque si se es un buen tirador los tiros llegarán muy cerca del blanco. ¿Qué entenderíamos por ser un tirador bueno, uno regular, uno malo y uno pésimo? ¿Cómo se llamaría al hecho de que no siempre se dé en el blanco?

De manera análoga, la media desempeña el papel del blanco de modo que el análisis estadístico de los datos estaría incompleto si sólo nos fijamos en ese punto. Por eso hay una variación en torno al blanco cuando se intenta atinarle.

Además de conocer los valores centrales de la muestra de datos, es importante comprender qué tan diferentes son los datos de la muestra, es decir, necesitamos saber qué tan dispersos están los datos en torno a la media. Entender la varianza de un conjunto de datos es un tema fundamental en estadística. En esta parte se expondrán algunas *medidas de variación*, como *el rango, la desviación estándar y el rango intercuartil*.

Rango

El *rango* es una de las medidas más sencillas para expresar la dispersión de los datos. Tan sólo se requiere considerar los valores máximo y mínimo de las observaciones. El rango muestral (R) es la diferencia entre el máximo y el mínimo de las observaciones de la muestra:

$$R = máximo - mínimo.$$

Ejemplo 3.12

En el contexto del impacto económico, se tomó una muestra de 8 farmacias para conocer el precio de una medicina. Es de interés conocer si existe uniformidad en los precios. En ese sentido las medidas de tendencia central no pueden aportar esa información, por lo cual recurrimos a las medidas que se llaman de dispersión. El rango es una medida de dispersión que sólo requiere del conocimiento de los valores máximos y mínimos.

La variable X es el precio de la medicina, los datos recogidos para el precio en estas 8 farmacias son:

Precio	87.4	92.7	98.0	108.0	119.2	125.3	149.8	162.6

El valor máximo es 162.6 y el mínimo 87.4; entonces, el cálculo del rango para precio de la medicina es:

$$R = máximo - mínimo = 162.6 - 87.4 = 75.2$$

En la figura 3.8 se ilustra el valor del rango de manera gráfica. Cabe resaltar que en esa misma figura se ve cómo los puntos están dispersos alrededor de la media. Donde la media es:

$$\overline{x} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{943}{8} = 117.875,$$

La línea vertical describe este valor.

La mediana se obtiene a partir de la media de los datos ordenados que están en las posiciones: $\left(\frac{8}{2} \text{ y } \frac{8}{2}+1\right)$, es decir: $\widetilde{m} = \frac{108+119.2}{2} = 113.6$.

La información generada por el rango complementa el reporte que se tiene de la media y la mediana, pues la media es $\overline{x} = 117.875$ y la mediana $\widetilde{m} = 113.6$, donde la media se ve afectada por el valor extremo más grande.

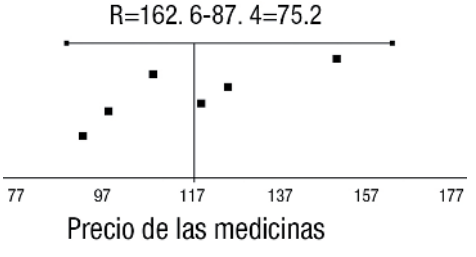


Figura 3.8 Dispersión de los datos alrededor de la media muestral, y el valor del rango entre los valores extremos.

Interpretación

En principio, este valor del rango indica una amplia dispersión en los precios de la medicina. Esto significa que en algunas farmacias tienen precios económicos y otras muy caros. Por ello conviene tener presente estudios donde se reporte los precios de las medicinas.

Rango intercuartil

El *rango intercuartil* es otra medida de dispersión y a continuación se especifican sus características. A menudo, en los datos aparecen observaciones con valores muy pequeños o demasiado grandes. En el contexto estadístico a estos valores se les llaman datos anómalos, y el rango es una medida sensible a

este tipo de datos. Con el propósito de tener una medida más adecuada ante ese tipo de situaciones se propone como alternativa el rango intercuartil. Esta medida es útil porque describe la magnitud de la dispersión de 50 % de las observaciones. Representa la distancia entre el primer y tercer cuartil.

El *rango intercuartil muestral* (RIC) es la diferencia entre el tercer cuartil C_3 y el primer cuartil C_1
$RIC = C_3 - C_1$

Ejemplo 3.13

La administración de un hospital ha adquirido una serie de ultrasonidos, en particular a continuación se verá la utilidad de uno de ellos. Se midió, mediante ultrasonido, la circunferencia abdominal a 8 bebés en la semana 34 de gestación. Este tipo de mediciones se realizan para que el médico pueda evaluar el desarrollo del bebé. La variable X es circunferencia del abdomen, los valores de ésta se presentan ordenados de menor a mayor:

Posición	1	2	C_1	3	4	$\widetilde{m}$	5	6	C_3	7	8
Orden	304	308		320	321		332	333		352	380
			↑			↑			↑		

El primer cuartil corresponde a la mediana del valor por debajo de la mediana. En este caso, como el número de datos es par, la media entre los números situados en las posiciones 2 y 3 son el primer cuartil, y se expresa de la siguiente forma:

$$C_1 = (308 + 320)/2 = 314$$

De manera análoga se obtiene el tercer cuartil, pero en este caso corresponde al valor que está entre la mediana de la mediana y el máximo, es decir,

$$C_3 = (333 + 352)/2 = 342.5$$

Por lo que el rango intercuartil es:

$$RIC = C_3 - C_1 = (342.5 - 314) = 28.5$$

La diferencia en el 50 % de los bebés es de 28.5 y está dentro de la normalidad.

Rango intercuartil

Se han calculado el rango y el rango intercaurtil para la muestra con el fin de obtener una idea de la variación de los datos; si los datos corresponden a una población, también se obtienen estas medidas. En algunos estudios en administración o economía pueden ser de utilidad estudiarlos, pero corresponde a un mayor entrenamiento en estadística.

Desviación media, varianza y desviación estándar

El mundo de la información 4. Costo de la vida (inflación)

El costo de las cosas sube con cada año que pasa, y a esta alza también se le conoce como inflación. De alguna manera los salarios deben ajustarse al incremento de las cosas para no perder el llamado poder adquisitivo. La dependencia del gobierno encargada de difundir los aumentos en los precios, dio a conocer el incremento en porcentaje de los productos de la canasta básica. En este caso consideramos una muestra de 10 de esos productos. Por lo general, se maneja el promedio para indicar cuál es la inflación.

Preguntas sobre la naturaleza del problema

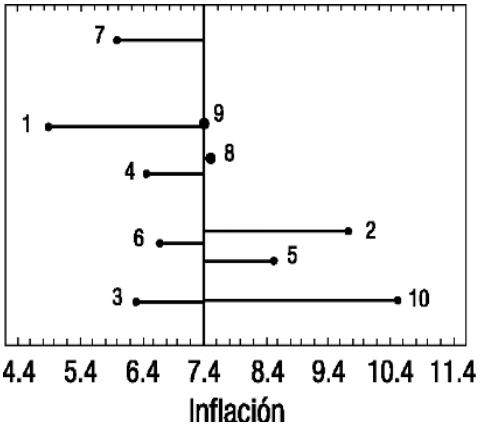
Una cuestión general es determinar qué productos suben más que otros con respecto al año de referencia. Como se sabe, casi todos los productos incrementan su costo de un año al otro. Pero ¿este incremento guarda la misma proporción?, ¿existe mucha variación entre los porcentajes de precios?

Los datos registrados, en porcentaje de la muestra de 10 artículos de la canasta básica son:

4.9, 9.7, 6.3, 6.5, 8.5, 6.7, 6.0, 7.5, 7.4, 10.5.

En la figura 3.8 se aprecia que cada observación está lejos del centro de los datos, pero el rango no mide qué tan lejos está del centro cada dato. Una medida que evalúa esa dispersión es la conocida como desviación estándar. Pero esta medida de dispersión no se calcula por sí sola, sino que se deriva de la varianza. Para fines prácticos, ésta se debe calcular primero. Siguiendo la idea que se presenta en la figura 3.8, primero debe medirse la desviación de cada observación con respecto a la media.

En la figura 3.9 se ejemplifica la dispersión de los datos y se detalla la desviación respecto a la media. ¿Qué se distingue en este esquema gráfico? La gráfica de dispersión muestra que las observaciones 1, 2 y 10 son las que están más alejadas de la media, y por lo tanto son las que más contribuyen a la varianza. Al comparar este hecho en la tabla adjunta, se ve que las desviaciones con mayor magnitud son las citadas.



	$\chi_i - \overline{\chi}$
1	4.9 – 7.4 = - 2.5
2	9.7 – 7.4 = 2.3
3	6.3 – 7.4 = - 1.1
4	6.5 – 7.4 = - 0.9
5	8.5 – 7.4 = 1.1
6	6.7 – 7.4 = - 0.7
7	6.0 – 7.4 = - 1.4
8	7.5 – 7.4 = 0.1
9	7.4 – 7.4 = 0
10	10.5 – 7.4 = 3.1

Figura 3.9 Dispersión de los datos alrededor de la media muestral, y desviaciones de las observaciones con respecto a la media.

Cálculo de la desviación media

Finalmente, en la última columna de la tabla 3 se inicia el procedimiento para el cálculo de la desviación media. Se puede observar que al sumar los valores de las desviaciones $x_i - \overline{x}$, es cero. Este resultado no permite evaluar la desviación de los datos exhibida en la figura 3.9. Se pueden ignorar los signos para calcular la variabilidad promedio; para eliminar el signo se toma el valor absoluto, se suman esos valores y se divide entre 10 para obtener la desviación media muestral. En la tabla 3.3 se describe el procedimiento para obtener la suma del valor absoluto de las desviaciones.

- La *desviación media de la muestra* es:

$$D_m = \frac{\sum_{i=1}^{10} |x_i - \overline{x}|}{10} = \frac{13.2}{10} = 1.32$$

- La *desviación media muestral* es:

$$D_m = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

- La *desviación media de la población* es

$$D = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N}$$

Población	Parámetro Δ	Medida: *Desviación media poblacional:*	Δ
Muestra	Estadística D	Medida: *Desviación media muestral:*	D

En la literatura estadística, no existe un símbolo especial para el parámetro que identifique a la desviación media, aquí se ha referido por Δ.

Valor absoluto

Se usa el símbolo $|\ \ |$ para representar el valor absoluto de un número.
En general se define el valor absoluto de cualquier número real x como sigue:

- $|x| = x$, si x es positivo;
- $|x| = -x$, si x es negativo.

Por ejemplo: $|x| = 0$, si $x = 0$. Si $x = -3, |-3| = -(-3) = 3$

La *desviación media muestral* (D_m) es la media del valor absoluto de las desviaciones de cada observación con respecto a la media muestral.

Tabla 3.3 Desarrollo del cálculo de la desviación media.

Datos	$x_1 - \overline{x}$	$x_1 - \overline{x}$	$\lvert x_1 - \overline{x}\rvert$
1	$4.9 - 7.4$	-2.5	2.5
2	$9.7 - 7.4$	2.3	2.3
3	$6.3 - 7.4$	-1.1	1.1
4	$6.5 - 7.4$	-0.9	0.9
5	$8.5 - 7.4$	1.1	1.1
6	$6.7 - 7.4$	-0.7	0.7
7	$6.0 - 7.4$	-1.4	1.4
8	$7.5 - 7.4$	0.1	0.1
9	$7.4 - 7.4$	0	0
10	$10.5 - 7.4$	3.1	3.1
Suma	0	0	13.2

Observaciones

1. La suma en la columna 2 es: $\sum_{i=1}^{10}(x_i - \overline{x}) = 0$ es decir la suma de las desviaciones de cada valor con respecto a la media es cero. En general

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0 \tag{3.5}$$

2. En valor absoluto en la columna 3 se expresa el resultado, esto es: $\sum_{i=1}^{10} |x_i - \overline{x}| = 13.2$

Justificación de la expresión 3.5

Recuerde	$\overline{x} = \frac{x_1+\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Despejando	$n\overline{x} = \sum_{i=1}^{n} x_i$
Suma de las desviaciones	$\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x}$
Vea que $\overline{x}$ se suma n veces	$\sum_{i=1}^{n} \overline{x} = n\overline{x}$
Finalmente	$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$
Se sigue que:	$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

Cálculo de la varianza

De manera similar, en la última columna de la tabla 3.4 se inicia el procedimiento para el cálculo de la varianza. Ahora en este caso, cada desviación se eleva al cuadrado y luego se suma.

Por lo tanto, la *varianza muestral* es:

$$S^2 = \frac{26.84}{10-1} = \frac{26.84}{9} = 2.982$$

Una varianza pequeña indica que los datos están cerca de la media y que por lo tanto hay poca dispersión. En caso contrario, si la varianza es grande, existirá dispersión en los datos. En el escenario del problema de la inflación, el valor de la varianza indica que algunos productos subieron más que el promedio, mientras que otros están por debajo del promedio. Eso indica que efectivamente sí hay una variación (aunque parezca moderada) en el porcentaje con que suben los precios en relación con el año anterior. La expresión general para la *varianza muestral* es:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

Tabla 3.4 Desarrollo del cálculo de la varianza.

Dato	$x_i - \overline{x}$		$(x_i - \overline{x})^2$
4.9	$4.9 - 7.4 =$	-2.5	6.25
9.7	$9.7 - 7.4 =$	2.3	5.29
6.3	$6.3 - 7.4 =$	-1.1	1.21
6.5	$6.5 - 7.4 =$	-0.9	0.81
8.5	$8.5 - 7.4 =$	1.1	1.21
6.7	$6.7 - 7.4 =$	-0.7	0.49
6.0	$6.0 - 7.4 =$	-1.4	1.96
7.5	$7.5 - 7.4 =$	0.1	0.01
7.4	$7.4 - 7.4 =$	0	0
10.5	$10.5 - 7.4 =$	3.1	9.61
SUMA		0	26.84

La varianza de la muestra S^2 es la media del cuadrado de las desviaciones de cada observación con respecto a la media muestral.

$$Varianza \text{ de la población} \quad \sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}$$

Población	Parámetro σ^2	Medida: varianza poblacional:	σ^2
Muestra	Estadística S^2	Medida: varianza muestral:	S^2

Complemento técnico

Para encontrar el promedio de la suma de cuadrados, se ha dividido entre $n-1$ y no exactamente entre n. Esto se debe a dos razones. La primera es que, como se ve en la tercer columna de la tabla 3.4, la suma de las desviaciones es cero. Esto quiere decir que cualquier desviación se puede encontrar a partir de las nueve desviaciones restantes. De modo que el valor de la suma de cuadrados depende sólo de nueve desviaciones que son la libertad de variar de una muestra a la siguiente. En general, se dice que la suma de cuadrados tiene $n-1$ grados de libertad. La segunda razón es tema de la inferencia estadística, y se puede decir que dividir la suma de cuadrados entre $n-1$ hace que la varianza muestral (estadístico) sea un mejor estimador de la varianza poblacional (parámetro). Esta última se expresa mediante la letra griega sigma minúscula elevada al cuadrado, σ^2.

Debe resaltarse que debido a sus propiedades matemáticas, la varianza y la desviación estándar se usan con mayor frecuencia que la desviación media.

Desviación estándar

- La desviación estándar muestral S es la raíz cuadrada positiva de la varianza.
- La desviación estándar de la población se denota por σ.

Población	Parámetro σ	Medida: desviación estándar poblacional:	σ
Muestra	Estadística S	Medida: desviación estándar muestral:	S

$$\text{Desviación estándar de la población} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}}$$

Cálculo de la desviación estándar

A partir de la varianza, de manera directa se calcula la *desviación estándar* . La fórmula para la desviación estándar es:

$$S = \frac{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}{n-1}, \qquad \text{o bien} \qquad S = \frac{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}}{n-1}$$

En la práctica, se usa más la desviación estándar porque tiene las mismas unidades que la desviación de los datos originales.

Ejemplo 3.14

Una compañía vende cursos de inglés mediante llamadas telefónicas, y su gerente desea conocer el tiempo que un cliente permanece en el teléfono durante una llamada, variable X: tiempo de llamada. La información recabada del tiempo (en minutos) de las llamadas con 9 clientes diferentes los valores de la variable fueron:

15, 23, 34, 16, 25, 37, 19 ,28 y 44

A estas 9 mediciones se le calcula la desviación media y la desviación estándar.

Solución

La desviación media se obtiene aplicando la fórmula:

$$D_m = \frac{\sum_{i=1}^{n}|x_i - \overline{x}|}{n};$$

así:

$$D_m = \frac{\sum_{i=1}^{9}|x_i - \overline{x}|}{9} = \frac{|-11.8|+|-3.8|+|-10.8|+|-1.8|+|10.2|+|-7.8|+|1.2|+|17.2|+|7.2|}{9} = 7.975$$

donde la media es:

$$\overline{x} = \frac{15+23+34+16+25+37+19+28+44}{9} = \frac{241}{9} = 26.778$$

Para calcular la desviación estándar, primero se obtienen los valores de la suma de cuadrados y de la media, y a continuación se sustituyen estos valores en la fórmula de la desviación estándar. Entonces:

$$\sum_{i=1}^{n} x_i^2 = 15^2 + 23^2 + 34^2 + 16^2 + 25^2 + 37^2 + 19^2 + 28^2 + 44^2 = 7241$$

$$S = \frac{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}}{n-1} = \sqrt{\frac{7241 - 9 \times (26.778)^2}{9-1}} = \sqrt{\frac{7241 - 6453.552}{8}} = \sqrt{98.431} = 9.921$$

De modo que $S = 9.921$. Éste resultado indica la variación alrededor del tiempo medio (26.778) que permanece un cliente en la línea telefónica.

Ejemplo 3.15

En las investigaciones que se realizan a nivel de laboratorio en busca de medicamentos para la cura de enfermedades, se utilizan como prueba animales pequeños. Este tipo de estudios requieren inversión inicial, ésta es una de las actividades en que las personas dedicadas a la administración tienen que realizar gestiones para conseguir recursos. Los siguientes datos representan el tiempo que sobrevivieron una muestra de 20 ratas de laboratorio expuestas a un nivel de radiación muy alto. La variable X es el tiempo de supervivencia, los valores en días son:

i	1	2	3	4	5	6	7	8	9	10
x_i	37	55	60	67	71	75	80	86	92	97
i	11	12	13	14	15	16	17	18	19	20
x_i	103	105	107	114	115	127	127	127	133	137

donde i es la posición y x es el valor. Calcular las medidas de tendencia central, los cuartiles y las medidas de dispersión, y finalmente interpretar los resultados.

Solución

Se aplican las expresiones matemáticas y los procedimientos para calcular las medidas de tendencia central y dispersión.

- La media se obtiene sumando las x_i y dividiendo entre 20:

$$\overline{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{1915}{20} = 95.75$$

- Como el número de datos es par, la mediana se obtiene mediante la media de los valores que están en las posiciones , es decir, $\frac{n}{2}$ y $\frac{n+1}{2}$, así:

$$\widetilde{m} = \frac{97 + 103}{2} = 100.$$

- El primer cuartil es la mediana del grupo de datos que están por debajo de la mediana. En este

caso los 10 primeros datos ordenados, componen ese grupo; así la media de 71 y 75 son el primer cuartil

$$C_1 = \frac{71 + 75}{2} = 73.$$

Tabla 3.5 Resumen estadístico para el tiempo que sobreviven las ratas de laboratorio a la radiación, Mín: Mínimo, Máx:Máximo

n	$\overline{x}$	$\widetilde{m}$	m_0	C_1	C_3	D_m	S^2	S
20	95.75	100	127	73	121	23.875	821.355	28.659
Mín	Máx	*Suma*	*Rango*	*RIC*				
37	137	1915	100	48				

- De manera análoga se obtiene el tercer cuartil, sólo que ahora considerando el grupo de datos que son mayores que la mediana; por lo tanto:

$$C_3 = \frac{115 + 127}{2} = 121.$$

- El rango es $R = mximo - mnimo$=137-37=100.

- El rango intercuartil es:

$$RIC = C_3 - C_1 = 121 - 73 = 48.$$

- La desviación media se obtiene mediante la media del valor absoluto de la diferencia de cada valor y la media ($|x_i - \overline{x}|$) es decir:

$$D_m = \frac{\sum_{i=1}^{20} |x_i - \overline{x}|}{20} = 23.875.$$

- La varianza es:

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1} = \frac{198967 - 20\left((95.75)^2\right)}{19} = \frac{15607.75}{19} = 821.35$$

- La desviación estándar es la raíz cuadrada de la varianza, por lo cual $S = \sqrt{821.35} = 28.659$.

La tabla 3.5 presenta un resumen de estos cálculos. El tiempo medio de vida de estos animales de laboratorio es de 96 (redondeado) días. Éste difiere de la mediana porque existen datos de tiempo de vida pequeños y éstos influyen en el valor de la media. Se puede observar que en torno a la media existe una importante variabilidad de los datos, lo que se puede interpretar que al recibir la dosis de radiación varias ratas vivirán mucho menos de 96 días, pero también varias sobrepasarán los 96 días.

Resumen estadístico de un conjunto de datos

En la tabla 3.1 se presentó el resumen estadístico para un conjunto de datos, la cual contiene únicamente las medidas de tendencia central. A esta tabla se integran también las medidas de dispersión para contar con un panorama completo de las medidas estadísticas en la muestra. Éstas, junto con la imagen gráfica de la distribución, permiten abarcar una información integral de cualquier problema que sea objeto de estudio.

En este sentido, se retoma el problema de la competencia atlética, y se le anexa a la tabla 3.1 la información de las medidas de dispersión.

Tabla 3.6 Resumen estadístico completo para la competencia atlética.

n	$\overline{x}$	$C_2 = \widetilde{m}$	m_0	C_1	C_3	S^2	S
112	16.71	16.48	15.51	15.73	17.47	1.69	1.3
Mín	Máx	*Suma*		*Rango*	*RIC*		
14.34	20.7	187.81		6.36	1.74		

3.5 Medidas estadísticas de datos agrupados

Las fórmulas y procedimientos indicados para calcular tanto las medidas de tendencia central como las de dispersión se aplican a cualquier conjunto de datos donde la variable sea cuantitativa. Si el número de datos es grande, el esfuerzo se presenta en la captura de éstos. En la actualidad existe una gran cantidad de calculadoras que permiten estimar con facilidad las medidas expuestas en esta unidad, con el paquete estadístico **CalEst**.

Sin embargo, a veces se puede recurrir a calcular las medidas desde las tablas de frecuencia. La que se presenta hoy en día con mayor regularidad es la de tabla de frecuencias para datos numéricos discretos. Si ese fuera el caso, las fórmulas para la media , la desviación media y la varianza son:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i U_i}{n}, D_m = \frac{\sum_{i=1}^{k} f_i \left| U_i - \overline{x} \right|}{n} \text{ y } S^2 = \frac{\sum_{i=1}^{k} f_i \left(U_i - \overline{x} \right)^2}{n-1}$$

donde, k número de clase f_i es la frecuencia de los valores de la clase $U_i, n = \sum_{i=1}^{k} f_i$ y $||$ es el valor absoluto.

Ejemplo 3.16

Muchos estudios revelan que la eficiencia de los estudiantes se ve reducida por el número de horas que le dedican a ver televisión, al "Chat", al internet, entre otras actividades. Las administraciones de los

centros escolares se ven obligadas a realizar estudios en esta dirección. En particular, a una muestra de 100 alumnos que cursan el tercer año de preparatoria se les preguntó cuántas horas veían televisión los fines de semana. Los datos se escriben a continuación:

6	2	3	8	1	11	3	6	8	2	7	3	8	2	8	2	6	5	4	3
3	4	6	4	4	2	2	7	5	4	2	4	6	3	2	2	9	4	3	3
2	9	2	4	2	5	4	5	11	6	5	10	2	3	4	2	4	4	6	9
4	5	7	5	3	9	7	3	3	3	3	4	7	8	5	8	4	2	3	6
7	4	5	4	7	5	4	2	5	6	2	4	2	2	2	2	6	5	4	5

Solución

Los datos en este ejemplo son numéricos discretos, así que los valores para U_i son discretos. En la tabla 3.7 los valores de la clase representan el número de horas y para cada uno se expresa la frecuencia. Las horas que ven televisión los alumnos comprende desde 1 hasta 11 horas.

Tabla 3.7 Frecuencias para el número de horas frente al televisor.

Valor U_i	1	2	3	4	5	6	7	8	9	10	11
f_i	1	21	15	20	13	10	7	6	4	1	2

Se calcula la media con la expresión para la media escrita en este apartado:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i U_i}{n} =$$

$$\frac{1}{100}(1 \times 1 + 2 \times 21 + 3 \times 15 + 4 \times 20 + 5 \times 13 + 6 \times 10 + 7 \times 7 + 8 \times 6 + 9 \times 4 + 10 \times 1 + 11 \times 2) = 4.58$$

Para realizar el cálculo de la desviación media y de la varianza se requiere de la expresión $U_i - x_i$. Ésta se reproduce a continuación:

U_i	1	2	3	4	5	6	7	8	9	10	11
$U_i - x_i$	-3.58	-2.58	-1.58	-0.58	0.42	1.42	2.42	3.42	4.42	5.42	6.42

La desviación media se obtiene aplicando la expresión:

$$D_m = \frac{\sum_{i=1}^{k} f_i \left| U_i - \overline{x} \right|}{n} = \frac{\sum_{i=1}^{k} f_i \left| U_i - 4.58 \right|}{100} = \frac{186.12}{100} = 1.861$$

La varianza se obtiene por:

$$S^2 = \frac{\sum_{i=1}^{k} f_i (U_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{k} f_i (U_i - 4.58)^2}{99} = \frac{520.36}{99} = 5.256$$

Ejemplo 3.17

La finalidad de este ejemplo es aprender a calcular la media, la mediana, la moda y la varianza a partir de una tabla de frecuencias. Es necesario aclarar que si bien el procedimiento para obtener estas medidas es operativo y puede resultar un poco mecánico, lo que importa es la interpretación. La interpretación de estas medidas es similar a la que se ha realizado en los ejemplos y problemas vistos con anterioridad. La tabla de frecuencias para el problema de la competencia atlética, se muestra en la tabla 3.8.

Tabla 3.8 Frecuencias para la competencia atlética.

Clase	Intervalo de clase	Punto medio U_i	Frecuencia
1	14.0, 14.8	14.4	3
2	14.8, 15.6	15.2	17
3	15.6, 16.4	16.0	31
4	16.4, 17.2	16.8	27
5	17.2, 18.0	17.6	17
6	18.0, 18.8	18.4	10
7	18.8, 19.6	19.2	2
8	19.6, 20.4	20.0	3
9	20.4, 21.2	20.8	2
10	21.2, 22.0	21.6	0

Los datos en este caso son numéricos continuos, por lo que el valor de U_i corresponderá al punto medio del intervalo de clase. Aplicando las expresiones vistas en este apartado se puede tener una aproximación a la media y la desviación estándar.

Solución

La tabla 3.9 reproduce los cálculos requeridos para aplicar las fórmulas que dan lugar a la media, desviación media y varianza.

La media es:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i U_i}{n} = \overline{x} = \frac{\sum_{i=1}^{k} f_i U_i}{112} = \frac{1874.4}{112} = 16.736$$

Los cálculos presentados en la tabla 3.7 se sustituyen en la fórmula para la desviación media y se obtiene:

$$D_m = \frac{\sum_{i=1}^{k} f_i \left| U_i - \overline{x} \right|}{n} = \frac{\sum_{i=1}^{k} f_i \left| U_i - 16.736 \right|}{112} = \frac{11.84}{112} = 0.999$$

La varianza se obtiene mediante la siguiente expresión:

$$S^2 = \frac{\sum_{i=1}^{k} f_i(U_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{k} f_i(U_i - 16.736)^2}{111} = \frac{190.898}{111} = 1.72$$

Tanto la media como la varianza dan valores muy aproximados a los que se reportaron en la tabla 3.6.

Si contemplamos los datos agrupados, entonces la moda está en la clase de mayor frecuencia. La moda se presenta en el intervalo de clase: 15.6, 16.4 (véase el polígono de frecuencia relativa de la figura 3.4). En este caso, se toma de manera aproximada el punto medio de ese intervalo como la moda, esto es, 16.

Tabla 3.9 Cálculo auxiliar para obtener medidas estadísticas de datos agrupados.

Clase	U_i	f_i	f_iU_i	$U_i - x_i$	$f\lvert U_i - \overline{x}\rvert$	$f_i\,(U_i - \overline{x})^2$
1	14.4	3	43.2	-2.336	7.008	16.371
2	15.2	17	258.4	-1.536	26.112	40.108
3	16.0	31	496	-0.736	22.816	16.793
4	16.8	27	453.6	0.064	1.728	0.111
5	17.6	17	299.2	0.864	14.688	12.69
6	18.4	10	184	1.664	16.64	27.689
7	19.2	2	38.4	2.464	4.928	12.143
8	20.0	3	60	3.264	9.792	31.961
9	20.8	2	41.6	4.064	8.128	33.032
10	21.6	0	0	4.864	0	0
Total		112	1874.4		11.84	190.898

Interpretación de la desviación estándar

Una manera de interpretar la variabilidad de un conjunto de datos es indicar el porcentaje de éstos que caen dentro de un número específico de desviación estándar de la media. Es decir, si a la media se le resta y suma una desviación estándar se tendrán dos valores: $\overline{x} - S$ y $\overline{x} + S$. Entonces la idea es encontrar el porcentaje entre estos dos números. De manera análoga se puede preguntar, ¿qué porcentaje de datos están dentro de dos desviaciones estándar de la media, es decir, entre los números $\overline{x} - 2S$ y $\overline{x} + 2S$? La respuesta depende de la forma de la distribución. Si la distribución de frecuencia muestra una forma más o menos simétrica alrededor de la media, la regla empírica da en forma aproximada los porcentajes de la distribución.

Regla empírica para datos de una muestra

Si la descripción de los datos en un histograma o diagrama de tallo y hoja presenta una distribución más o menos simétrica, entonces aproximadamente:

1. De las observaciones, 68 % caerán dentro de una desviación estándar de la media, es decir:

$$(\overline{x} - S, \overline{x} + S).$$

2. De las observaciones, 95 % estarán dentro de dos desviaciones estándar de la media, es decir:

$$(\overline{x} - 2S, \overline{x} + 2S).$$

3. Todas las observaciones caerán dentro de tres desviaciones estándar de la media, es decir:

$$(\overline{x} - 3S, \overline{x} + 3S).$$

En la figura 3.10 se describe esta *regla empírica* de manera gráfica.

Regla de Chebyshev

Una regla útil cuando la distribución no es simétrica, es la regla de Chebyshev .

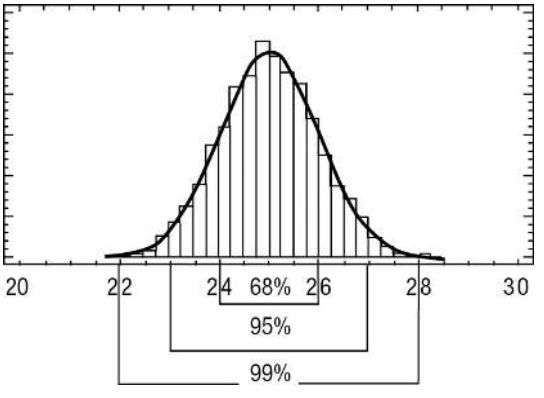


Figura 3.10 Descripción gráfica de la regla empírica.

Sin importar la forma de distribución.

1. Al menos 3/4 de las observaciones caen dentro de dos desviaciones estándar de la media, es decir:

$$(\overline{x} - 2S, \overline{x} + 2S)$$

2. Al menos 8/9 de las observaciones caen dentro de tres desviaciones estándar de la media:

$$(\overline{x} - 3S, \overline{x} + 3S)$$

Ejemplo 3.18

Con fines de planeación estratégica, la administración del sector salud de una comunidad tiene interés en conocer la edad en que las mujeres tienen a su primer hijo. Un trabajador social registra la edad a una muestra de mujeres recabada en varios hospitales, la información se describe a continuación:

30, 18, 35, 22, 23, 22, 36, 24, 23, 28, 19, 23, 25, 24, 33, 21, 28, 21, 23, 15, 20, 26
21, 24, 19, 33, 23, 19, 32, 21, 18, 36, 21, 25, 17, 21, 24, 20, 29, 24, 38, 16, 24,39
39, 22, 23, 18, 22, 28, 18, 15, 25, 21, 23, 26, 38, 24, 20, 14, 25, 26, 42, 22, 24, 22
36, 27, 21, 28, 26, 22, 28, 33, 18, 17, 21, 15, 20, 16, 21, 20, 20, 17,24, 20, 17, 19

Solución

La tabla 3.10 describe el resumen estadístico de estos datos.

Aquí se verá una aplicación de la regla empírica, aunque los datos no son exactamente simétricos. Para verificar esta regla se necesita observar qué porcentaje de los datos está entre una desviación estándar de la media. En la tabla 3.10 que presenta el resumen estadístico de estos datos; se observa que la media y la desviación estándar son, respectivamente: $\overline{x} = 23.92$ y $S = 6.23$, por lo tanto:

$$\overline{x} - S = 23.93 - 6.23 = 17.7 \text{ y } \overline{x} + S = 23.93 + 6.23 = 30.16$$

A partir de la gráfica de frecuencia acumulada (figura 3.11) se observa que aproximadamente 62 de 88 datos están entre 17.7 y 30.2; esto es: 62/88=0.705,

$$\overline{x} - 2S = 23.93 - 12.46 = 11.47 \text{ y } \overline{x} + 2S = 23.93 + 12.46 = 36.39$$

Esto significa que la totalidad de los datos está aproximadamente dentro de dos desviaciones estándar de la media. Se puede ver que a la izquierda del valor existen 5 valores, esto es, cerca de 6 % de mujeres tienen a su primer hijo cuando son mayores de 36 años.

Tabla 3.10 Resumen estadístico para la edad en que las mujeres tienen a su primer hijo.

n	$\overline{x}$	$\widetilde{m}$	m_0	C_1	C_3	D_m	S^2	S
88	23.93	23	21	20	26	4.695	38.85	6.23
Mín	Máx	Suma	Rango	RIC				
14	42	2106	28	6				

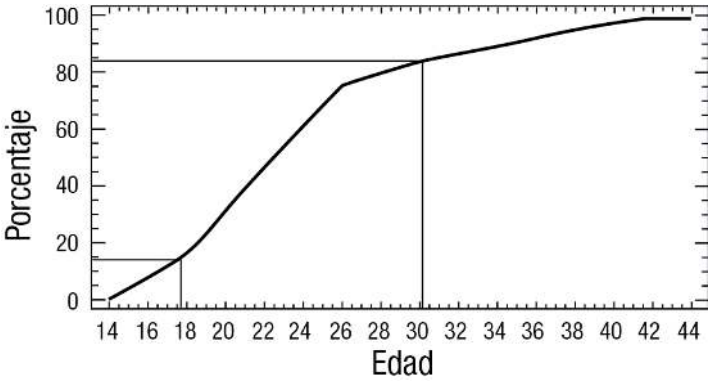


Figura 3.11 Polígono de frecuencia acumulada para la edad de las mujeres en tener su primer hijo.

3.5.1 Mediciones para el sesgo y el coeficiente de variación

Medición del sesgo

En la figura 3.6 se caracterizó el sesgo de una distribución, sin embargo, no se presentó ninguna expresión que indicara una *medida del sesgo.* Ahí se observó que una distribución es simétrica si la media y la mediana son iguales, que hay un sesgo positivo en la distribución si la media excede a la mediana y que, por el contrario, hay un sesgo negativo si la media es menor que la mediana. Una fórmula austera que permite obtener un valor para el sesgo de manera aproximada es:

$$\text{medición de sesgo} = \frac{3\,(\text{media - mediana})}{S} = \frac{3\,(\overline{x} - \widetilde{m})}{S}$$

Esta expresión permite efectuar mediciones del sesgo, donde se tienen como valores el cero, el positivo y el negativo. Una idea aproximada de la distribución depende del valor del sesgo. En caso de que la medida sea mayor que 1 se puede decir que la distribución presenta un sesgo marcadamente positivo. De igual manera, si el sesgo es -1, entonces el sesgo es claramente negativo.

Ejemplo 3.19

Con el objetivo de mejorar la puntualidad del servicio, la administración de una empresa de autobuses foráneos quiere conocer los tiempos, en minutos, de retraso en la salida de sus unidades. Los datos registrados para ese propósito son:

4.6	8.6	9.6	10.3	10.8	11.5	11.9	12.4	12.9	13.6	8.5	9.5	10.1
4.9	8.7	10	10.3	10.9	11.5	12	12.4	13.2	14.1	10.7	11.2	11.8
5.4	9.2	10	10.6	10.9	11.6	12.1	12.4	13.2	14.2	12.3	13.4	
7.8	9.5	10.1	10.6	11	11.6	12.2	12.4	13.4	14.7	12.9	15.1	

Solución

En la tabla 3.11 se reporta el resumen estadístico de los datos del problema.

De ahí podemos conseguir los valores para la media, la mediana y la desviación estándar, respectivamente, que son: sustituyendo estos valores en la fórmula de la medición del sesgo tenemos:

$$medida\ de\ sesgo = \frac{3\,(11.05 - 11.35)}{2.27} = -0.4$$

Podemos decir que en este ejemplo, la distribución muestra un ligero sesgo negativo. Esto se debe a que la media se ve influida por salidas con un retraso de menos de 5.5 minutos.

A partir de las medidas de tendencia central se observa que los autobuses salen con un retraso alrededor de 11 minutos. Considerando la desviación estándar a esta información, se observa que aproximadamente 68 % de los autobuses retrasan su salida en 2.27 minutos en torno a la media, es decir, entre 8.73 y 13.27 minutos. Esta demora, en la práctica, puede considerarse importante.

Coeficiente de variación

Otra medida apropiada para medir la variación de un conjunto de datos es el llamado : CV. La fórmula involucra la media y la desviación estándar y casi siempre, se expresa como porcentaje; ésta es:

$$CV = \frac{S}{\overline{x}} \times 100$$

Esta medida es útil para comparar las variabilidades de dos conjuntos de datos cuando existe una clara diferencia en la magnitud tanto en la media como en la desviación estándar. Otra característica importante del CV es que es independiente de las unidades de medición. Por ejemplo, podría servir para comparar la variación de peso entre elefantes y ratones.

A continuación se presenta un resumen del sesgo y coeficiente de variación para su cálculo con los datos de la población.

Población	Parámetro $Sesgo_{po}$	$Sesgo_{po} = \frac{3(\mu - m_{po})}{\sigma}$
Población	Parámetro CV_{po}	$CV_{po} = \frac{\sigma}{\mu} \times 100$

Tabla 3.11 Resumen estadístico del tiempo de retraso de la salida de autobuses.

n	$\overline{x}$	$\widetilde{m}$	m_0	C_1	C_3	D_m	S^2	S
50	11.05	11.35	12.4	10	12.4	1.71	5.17	2.27
Mín	Máx	Suma	Rango	RIC				
4.6	15.1	552.6	10.5	2.4				

Ejemplo 3.20

Para comparar la variación de la estatura entre un grupo de niños de 5 años de edad y otro de 15 años, se tomó una muestra de 25 personas de cada grupo de edad. Las muestras arrojaron los siguientes resultados: para el primer grupo se tiene que $\overline{x}_{g5} = 100$ cm, $S_{g5} = 6$ cm, para el segundo, $\overline{x}_{g15} = 150$ cm, $S_{g5} = 9$ cm. ¿Cuál de estos grupos de edad tiene una mayor variación?

Solución

Ahora es necesario encontrar el coeficiente de variación de cada grupo. A partir de la desviación estándar se observa que hay mayor variación en el segundo grupo, pero es necesario recordar que estamos comparando grupos de personas de diferentes magnitudes debido a su edad. Desde un punto de vista relativo se verá que las estaturas de ambos grupos están muy aproximadas. Determinemos el coeficiente de variación para cada grupo:

$$CV_{g5} = \frac{6}{100} \times 100 = 6\,\% \text{ y } CV_{g15} = \frac{9}{150} \times 100 = 6\,\%$$

Como se ve, ambos coeficientes son iguales, por lo que la dispersión relativa de ambos grupos es igual.

Uso de la calculadora estadística para obtener el resumen estadístico

Hasta el momento se han calculado los estadísticos relevantes que describen la tendencia central y la variación de los datos. Mediante la calculadora estadística se podrá obtener el resumen completo de todas estas medidas. En esta etapa ya se adquirieron las habilidades desarrolladas para usar en forma adecuada la calculadora estadística de las medidas estudiadas. En este breve apartado se completará la presentación de todos los resultados que se obtienen mediante la calculadora. Para ilustrar esta situación retomamos el ejemplo 2.

Ejemplo 3.21

Mediante el uso de la calculadora estadística obtener el resumen estadístico del precio de un litro de aceite comestible. Datos presentados en el ejemplo 3.2, es decir: 15.70, 21.45, 16.0, 19.10, 18.20, 20.80, 21.30, 17.90, 15.75, 20.90.

Solución

En la figura 3.12 se reproduce la figura 3.3, sólo que ahora ya se conocen todas las medidas de estadísticas. Con éstas se puede reproducir el resumen estadístico como el que se expone en la tabla 3.12.

Tabla 3.12 Resumen estadístico correspondiente al precio de una botella de aceite en 10 tiendas.

n	$\overline{x}$	$\widetilde{m}$	C_1	C_3	D_m	S^2	S	Sesgo	CV
10	18.71	18.65	16	20.9	2	5.51	2.35	0.08	7.96
Mín	Máx	Suma	Rango	RIC					
15.7	21.45	187.1	5.75	4.9					

Resumen de Medidas

	Precio		
n	10		
Media	18.71000	Rango	5.75000
Mediana	18.65000	Desviación estándar	2.34791
Suma	187.10000	Desviación media	2.00000
Moda	No existe	Varianza	5.51267
Percentil 25%	15.93750	Sesgo	- 0.12466
Percentil 75%	21.00000	Kurtosis	1.17394
Rango Intercuantil	5.06250	Coeficiente de variación	12.54894
Máximo	21.45000	Media geográfica	18.57494
Mínimo	15.70000	Media armónica	18.43843

Figura 3.12 Salida que reproduce el resumen estadístico usando CalEst.

3.6 Diagrama de caja

El *diagrama de caja* es una técnica de graficación relativamente moderna y sencilla, que permite apreciar las características principales de los datos y con ello tener una idea aproximada de la distribución. Asimismo, nos permite en una sola gráfica comparar la distribución de varios conjuntos de datos, ello con el objetivo de determinar las diferencias principales de los datos.

El mundo de la información 5. Índice de masa corporal

Economía en la salud. El control de peso siempre ha sido un asunto de interés para la salud. En la actualidad, en muchos medios se trata el tema y todos los días aparecen nuevos procedimientos, dietas y medicamentos que permiten vigilar el peso. Una medida aceptada para evaluar el estado de salud de las personas es el que se conoce como índice de masa corporal (IMC). Éste considera el peso y la estatura mediante la siguiente relación: IMC = peso/estatura^2. Así que el departamento de deportes de una universidad midió este índice en 98 de sus estudiantes para saber en qué estado de salud se encontraban.

Preguntas sobre la naturaleza del problema

La meta del estudio del departamento de deportes es conocer si los 98 estudiantes están en el peso ideal, en sobrepeso o si son obesos. El valor del índice permitirá establecer las condiciones de salud de los estudiantes. Se dice que la relación entre peso y estatura de una persona es normal si el índice se ubica entre 18 y hasta menos de 25; se presenta sobrepeso entre 25 y menos de 30; y hay obesidad cuando el índice marca 30 en adelante. Lo que se desea averiguar es cuántos de estos estudiantes están en alguno de estos tres grupos del IMC. La información proporcionada por el estudio, ¿podría extenderse a la población de alumnos universitarios?

Los datos se obtuvieron usando una báscula y una cinta métrica. Así los valores del IMC son:

24.85	22.48	25.56	24.42	23.49	23.7	25.68	24.68	25.43	23.89
26.73	26.96	23.94	31.58	22.27	24.05	25.36	30.14	32.01	30.22
26.28	24.49	29.5	29.76	23.92	28.45	29.03	25.21	24.6	26.29
23.36	27.55	27.89	23.12	28.53	22.09	26.13	25.03	23.27	30.05
27.86	25.23	25.7	26.06	31	26.01	27.93	25.39	25.57	24.24
25.05	23.66	27.33	26.71	25.74	29.35	22.4	24.91	31.04	25.35
23.57	23.21	24.05	25.81	24.79	24.62	27.22	26.07	24.22	29.9
27.47	26.07	26.44	28.03	24.9	25.91	32.53	22.98	22.43	27.53
25.18	28.75	27.7	24.15	28.34	25.84	23.02	24.38	25.74	
24.27	37.02	25.5	23.35	25.51	26.11	28.72	30.23	23.2	

El reporte estadístico de este conjunto de datos se presenta en la tabla 3.13, los cuales se obtuvieron utilizando **CalEst**.

Tabla 3.13 Resumen estadístico completo para el índice de masa corporal (IMC).

n	$\overline{x}$	$\widetilde{m}$	C_1	C_3	S^2	S	Sesgo	CV
98	26.16	25.63	24.24	27.7	7.07	2.66	0.597	10.17 %
Mín	Máx	Suma	Rango	RIC				
22.09	37.02	2563.28	14.93	3.46				

Cinco medidas básicas para construir el diagrama de caja

En el resumen estadístico hay cinco medidas principales para elaborar un diagrama de caja, las cuales

son los valores que corresponden al mínimo, C_1(primer cuartil), la mediana C_2 (el segundo cuartil) y el máximo C_3 (el tercer cuartil). Para el problema estos cinco números son:

1. El mínimo 22.09.
2. El primer cuartil $C_1 = 24.24$.
3. La mediana 25.63.
4. El tercer cuartil $C_3 = 27.7$
5. El máximo 37.02.

En la figura 3.13 se describen estos puntos y el rectángulo que representará al diagrama de caja, mientras que en la tabla 3.13 se presentan los cinco elementos principales en la construcción del diagrama de caja.

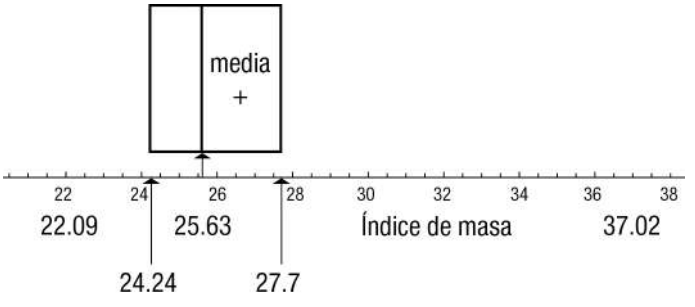


Figura 3.13 Cinco elementos para construir el diagrama de caja.

Construcción del diagrama de caja

- **Paso 1**. Trazar una línea horizontal que incluya los valores mínimo y máximo.
- **Paso 2**. Arriba de la línea dibujar un rectángulo (caja) cuyos lados queden en los puntos C_1 y C_3.
- **Paso 3**. Dentro de la caja, trazar una línea en el punto que corresponde a la mediana.
- **Paso 4**. Se traza una línea de cada lado de la caja cuya extensión es: 1.5 (RIC). Es decir para la izquierda:

$$L_1 = C_1 - 1.5(RIC)$$

y para la derecha

$$L_2 = C_3 + 1.5(RIC)$$

- **Paso 5.** Si hay puntos más allá de estas líneas, éstas se marcan con un asterisco (*). Estas observaciones corresponden a datos anómalos.
- **Paso 6**. Si no existen datos anómalos, al final de las líneas hay unas líneas pequeñas. A tales líneas se les conoce como "bigotes".

En la figura 3.13 se presentan todos los detalles que caracterizan el diagrama de caja. A continuación se indica el procedimiento para construir el diagrama de caja.

Los límites L_1 y L_2 que marcan la distancia de las líneas que salen de la caja para el problema tienen los valores $L_1 = 19.05$ y $L_2 = 32.89$. Es importante advertir que L_1 no se extiende hasta su valor mínimo porque este es de 22.09, entonces esa línea se queda hasta ese valor. En el caso de L_2 no se alcanza a cubrir el máximo, por lo que este es un dato anómalo.

En la figura 3.15 se compara el diagrama de caja con el histograma y polígono de frecuencias. Ahí se puede observar la relevancia del diagrama de caja, pues estos cinco datos brindan muy buena información de la distribución de la muestra. Haciendo un análisis minucioso de este diagrama, comparándolo con los otros dos, se verá que el diagrama de caja indica que la distribución tiene sesgo positivo. En la gráfica resaltan las medidas de la media y mediana, y es evidente cuál es su posición. Desde luego, también se identifican rápido los datos anómalos y también puede compararse con el diagrama de tallo y hoja.

Interpretación

En el contexto del problema del índice de masa corporal, cabe notar que 25 % de la muestra está en el grupo donde prácticamente no hay dificultades con el peso (figura 3.14). Se puede decir que básicamente la mayoría de los alumnos padecen sobrepeso y unos cuantos muestran obesidad.

En este problema no puede aplicarse la regla empírica porque la distribución no es simétrica, por ello se aplica la regla de Chebyshev. Así que para los datos de IMC, de dos desviaciones estándar de la media se tiene:

$$26.16 \pm \ 2 * (2.66) = (20.84, 31.48)$$

Entre los valores de 20.84 y 31.48 está al menos $\frac{3}{4}$ de los datos.

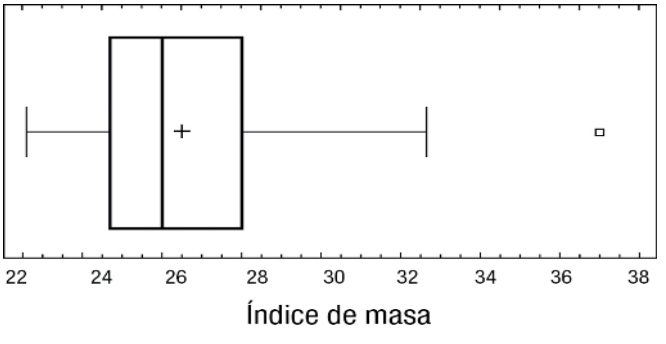


Figura 3.14 Diagrama de caja completo para el índice de masa corporal.

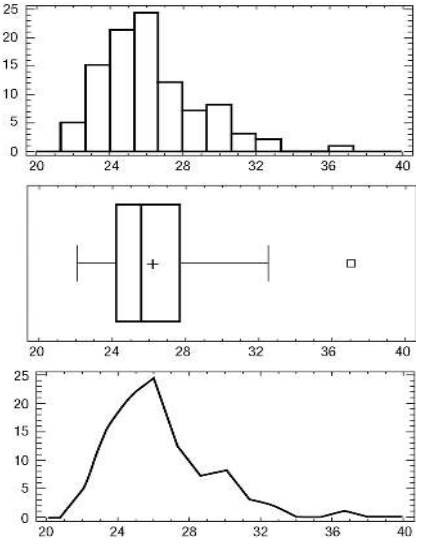


Figura 3.15 Diagrama de caja y su comparación con el histograma y el polígono de frecuencias.

Construcción del diagrama de caja mediante el uso de CalEst

En la opción Estadística en el menú del CalEst aparece la opción que se llama Diagrama de caja en el módulo de gráficas. Hay que seleccionarlo para que aparezca la pantalla mostrada en la figura 3.10. El procedimiento para usar esta hoja de cálculo es como sigue:

- **Paso 1.** Seleccione el número de variables para las cuales se dibujará la caja.
- **Paso 2**. Alguna variable, o algunas de ellas, presentarán mayor número de datos. Use el valor y escríbalo en la casilla referente al número de datos. Se oprime la tecla, usando el ratón, regresar y se desplegará una tabla; las columnas corresponden a las variables y los renglones a los datos.
- **Paso 3**. Escriba los valores de las variables en la tabla.
- **Paso 4**. Oprima donde dice Graficar. Entonces aparecerá el diagrama o los diagramas de caja para las variables seleccionadas.

Diagrama de caja información complementaria

Estas medidas de localización tienen una aplicación que resulta relevante en el análisis descriptivo de los datos. El llamado *diagrama de caja* recoge la información de las medidas de posición. Observe la figura 3.16. La opción 8 de gráficas en el paquete contiene el mecanismo para elaborar los diagramas de caja.

Ejemplo 3.22

Véase en el **Complemento didáctico** el problema rompecabezas de la República Mexicana. La variable de respuesta es el tiempo que tarda en resolver el rompecabezas de la República Mexicana una muestra de 120 estudiantes de licenciatura de una universidad. Los datos se muestran a continuación. Del reporte estadístico generado por **CalEst** se consideran 5 valores, éstos son los tres cuartiles, el máximo y el mínimo, y se muestran en la siguiente tabla:

505	499	513	482	493	493	513	509	498	508	504	497
501	500	481	494	512	514	489	480	510	501	481	480
491	499	486	495	494	491	509	498	508	497	499	500
477	513	509	509	485	494	489	514	521	514	512	494
477	507	495	506	485	499	504	510	494	488	508	498
494	491	475	486	524	474	498	479	508	520	509	488
488	499	500	499	518	516	518	523	493	497	519	479
485	500	499	512	511	494	498	497	501	501	494	501
498	492	514	509	481	502	518	504	478	508	494	518
510	499	515	495	494	485	503	480	509	487	491	495

Variable	Min	Max	C_1	C_2	C_3
Tiempo	474	524	491.25	499	509

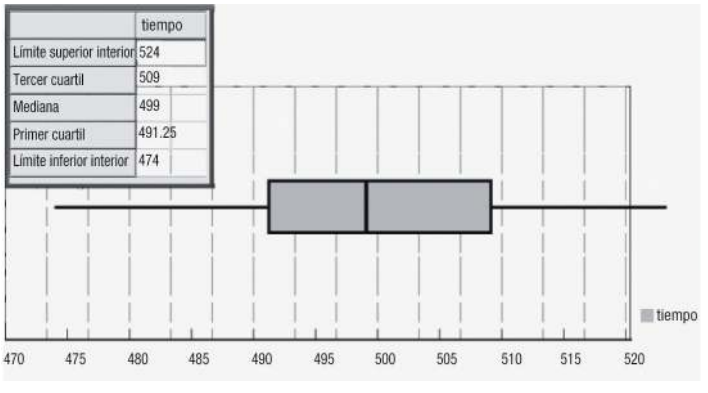


Figura 3.16 Diagrama de caja para los datos de la resolución del rompecabezas.

Con este reporte se elabora el diagrama de caja como se describe en la figura 3.16.

Ejemplo 3.23

Una aplicación de interés del diagrama de caja es que permite la comparación entre grupos, años, dependencias, entre otros casos. Aquí se reportará el precio de la tortilla de 2007 a 2012 en 53 ciudades de México. La figura 3.17 describe esa situación y la figura 3.18 muestra el reporte de las estadísticas de posición. Ahora interprete los resultados.

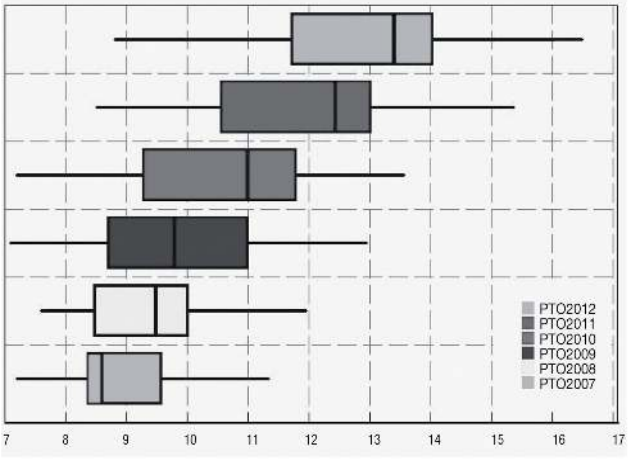


Figura 3.17 Diagrama de caja del precio de la tortilla de 2007 a 2012.

	PTO2012	PTO2011	PTO2010	PTO2009	PTO2008	PTO2007
Límite superior interior	16.5	15.4	13.6	13	12	11.4
Tercer cuartil	14	12	11.775	11	10.035	9.6
Mediana	13.375	12.43	11	9.8	9.5	8.63
Primer cuartil	11.7175	10.565	9.305	8.73	8.5	8.4
Límite inferior interior	8.8	8.5	7.2	7.1	7.6	7.2

Figura 3.18 Reporte de las estadísticas de posición.

3.7 Resumen

Existen varias formas para caracterizar los datos usando medidas de la muestra, la cual está representada por los estadísticos. Ningún número o medida hará el trabajo, pues no existe un camino fácil para obtenerlo. No obstante, las medidas que seleccionemos reflejarán las características de los datos. El resumen estadístico nos permite una mejor descripción de los datos y de ellos se desprenderán las conclusiones sobre un estudio.

Para finalizar, cabe resaltar que el análisis de los datos no es una herramienta estática, ya que se debe mirar el conjunto de datos desde muchos ángulos para obtener la mejor información posible a través de ellos. Algunas veces diferentes métodos nos conducen a las mismas conclusiones y en otras, un método dará mejor resultado que otro.

Diagrama de caja	Gráfica que exhibe un resumen de la distribución de una muestra usando los cuartiles y la mediana
Regla empírica	Indica que, para una distribución simétrica, de manera aproximada: de las observaciones, 68 % están a una desviación estándar de la media. de las observaciones, 95 % están a dos desviaciones estándar de la media. más de 99 % de las observaciones están a una desviación estándar de la media.
Primer cuartil, C_1	Valor en la muestra que representa 25 % de los datos.
Tercer cuartil, C_3	Valor en la muestra que representa 75 % de los datos.
El rango intercuartil	Diferencia entre el tercer cuartil y el primer cuartil, $C_3 - C_1$
Clase modal	Intervalo de clase en una tabla de frecuencias o en un histograma que muestra la mayor frecuencia.
Parámetro	Descripción numérica que se usa para describir a una población.
Dato anómalo	Valor que cae más allá de las líneas en un diagrama de caja.
La media muestral	Centro de balance de un conjunto de datos, se obtiene sumando todos los valores y dividiendo la suma entre el número de observaciones.
La mediana muestral	Valor que está a la mitad de las observaciones en un conjunto de datos ordenados de menor a mayor.
La moda muestral	Valor que más ocurre en la muestra.

El rango muestral, R	Diferencia entre las observaciones máxima y mínima de una muestra.
Desviación media	Media del valor absoluto de las desviaciones de los valores observados con respecto a la media.
Desviación estándar muestral	Raíz positiva de la varianza muestral.
Varianza muestral	Promedio del cuadrado de las desviaciones de los valores observados con respecto a la media.
Estadístico	Descripción numérica que se calcula a partir de la muestra y se usa para describir a ésta.

Población		$\Longrightarrow$ Selección $\Longrightarrow$	Muestra	
Variable:	X		Variable:	X
Valores de la variable	$x_1, x_2, ..., x_N$		Valores de la variable	$x_1, x_2, ..., x_n$
Parámetros	*Medidas*	$\Longleftarrow$ Inferencia $\Longleftarrow$	Estadísticos	*Medidas*
Media	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$		Media	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Mediana	m		Mediana	$\widetilde{m}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}$		Varianza	$S = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$
Desviación Estándar	$\sigma = \frac{\sqrt{\sum_{i=1}^{N}(x_i-\mu)^2}}{N-1}$		Desviación Estándar	$S = \frac{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{n-1}$
Desviación Media	$D = \frac{\sum_{i=1}^{N}\lvert x_i-\mu\rvert}{N}$		Desviación Media	$D_m = \frac{\sum_{i=1}^{n}\lvert x_i-\overline{x}\rvert}{n}$

Método para calcular cuartiles usando CalEst

Con la letra ℓ se designa la posición del dato ordenado. Así, si hay 8 datos $n = 8$, lo que indica que se tienen 8 lugares para cada dato. Por ejemplo, si $\ell = 3$ indica que el dato está en la posición 3.

$$\ell_1 = \frac{n+1}{4} \qquad \ell_3 = \frac{3(n+1)}{4}$$

Las expresiones ℓ_1 y ℓ_3 determinan las posiciones para el primer y tercer cuartil. Es decir, ℓ_1 divide de manera aproximada el 25 % de los datos abajo del cuartil C_1 y el 75 % arriba. De manera análoga ℓ_3 divide a los datos cerca del 75 % debajo de C_3 y el 25 % arriba. En algunos casos resulta que tanto ℓ_1 como ℓ_3 no son enteros y dificulta la ubicación del lugar exacto del dato, entonces habrá que estimarlo. Par fijar ideas consideremos un caso del ejemplo 7.

Ejemplo 3.24

Los datos son: 2, 3, 6, 12, 17, 21, 28, 34

ℓ : *posición*	1	2	C_1	3	4	C_2	5	6	C_3	7	8
datos		2	3		6	12		17	21		28	34
			↑			↑			↑		

¿Cómo representar en símbolos los datos y la posición? $x_{(1)} = 2$, $x_{(2)} = 3$, $x_{(3)} = 6$, $x_{(4)} = 12$, $x_{(5)} = 17$, $x_{(6)} = 21$, $x_{(7)} = 28$, $x_{(8)} = 34$.

Solución

Las posiciones

$$\ell_1 = \tfrac{8+1}{4} = \tfrac{9}{4} = 2.25 \qquad \ell_3 = \tfrac{3(8+1)}{4} = \tfrac{27}{4} = 6.75$$

Interpretación, $\ell_1 = 2.25$ indica que a partir de ese valor, 25 % de los datos están por debajo de C_1 y 75 % arriba. ¿Cuál es el valor de X en ese punto, es decir $x_{(2.25)}$? $x_{(2.25)}$ se lee el valor de la variable en la posición 2.25, para tener una número aproximado se aplica la siguiente expresión:

$$x_{(2.25)} = x_{(2)} + 0.25(x_{(3)} - x_{(2)})$$
$$x_{(2.25)} = 3 + 0.25(6 - 3) = 3.75$$

Por lo tanto el valor que corresponde al primer cuartil es $C_1 = 3.75$.

De manera similar se procede para estimar el valor del tercer cuartil, en este caso $\ell_3 = 6.75$ ¿cuál es el valor para $x_{(\ell_3)} = x_{(6.75)}$?

$$x_{(6.75)} = x_{(6)} + 0.75(x_{(7)} - x_{(6)})$$
$$x_{(6.75)} = 21 + 0.75(28 - 21) = 26.25$$

Finalmente el tercer cuartil es $C_3 = 26.25$.

3.8 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

3.9 Ejercicios

Medidas de tendencia central:Media

3.1 En una zona rural de la región central de México se registró durante el mes de junio de los últimos doce años la siguiente precipitación pluvial: 91.2, 85.7, 90.8, 87.2, 85.7, 91.1, 88.3, 93.5, 88.4, 87.2, 87.3, 86.8. Calcule la media para este conjunto de datos.

3.2 En la sierra norte de un estado de la República Mexicana la cantidad de lluvia registrada durante los últimos nueve meses fue: 54.6, 66.4, 31.8, 59.6, 51.7, 67.2, 62.5, 40.3, 78.1. Calcule la media de las precipitaciones. Con la información obtenida de las actividades 1 y 2, ¿podría decir dónde llueve más en promedio: en la zona rural o en la sierra?

3.3 Una empresa dedicada a elaborar material para la enseñanza de idiomas utiliza un material que hace atractivo y entretenido el aprendizaje; además, pretende que las personas a las que está dirigido desarrollen sus habilidades auditivas. Dichos materiales, presentados en discos compactos, fueron diseñados para usarse con un video y una computadora. Una parte de esos materiales consiste en realizar unos ejercicios en los que se escucha una palabra y la persona debe relacionarla con una figura. Toma cierto tiempo relacionar las palabras y las figuras de cada ejercicio. El tiempo se registra en la pantalla. Después de una etapa de entrenamiento, 10 personas de una clase de francés realizan un ejercicio en la computadora. Los tiempos en segundos registrados en acertar a todas las figuras fueron: 39, 38, 39, 41, 40, 41, 37, 39, 40 y 42. Elabore un diagrama de puntos e identifique la media.

3.4 Encuentre la media del siguiente conjunto de datos:

1. 1, 3, 6, 9, 10, 14
2. 1, 1, 2, 4, 5, 6, 8
3. 3.0, 4.0, 2.1, 3.5, 2.6, 3.0, 4.8, 2.5

3.5 Aplicación *media ponderada.* Una mensajería le cobra a una empresa con base en la media del peso de los paquetes que enviará. De los 60 paquetes, 12 pesaron 250 gramos, 25 pesaron 500 gramos, 18 llegaron a 750 gramos, y el resto registraron 900 gramos. Si se debe pagar 10 por cada gramo que resulte de la media de pesos. ¿Cuánto debe pagar la empresa por cada paquete?. Use la expresión

$$\overline{x} = \frac{\sum_{i=1}^{4} w_i x}{\sum_{i=1}^{4} w_i}$$

Media y mediana

3.6 Una compañía que manufactura un pesticida estudia el número de insectos que aniquila una dosis específica del insecticida. Se realizan 10 pruebas. En cada una de ellas el número de insectos muertos de 40 son:

19 22 34 28 18 16 25 27 31

Calcule la media y la mediana. Con base en esos datos, señale qué tan efectivo es el insecticida.

3.7 Un administrador realiza un estudio para conocer las destrezas de medición entre dos grupos de empleados. Para ello aplica un reactivo de 12 preguntas en un primer grupo que está compuesto por 20 empleados entrenados y en otro que tiene 17 no entrenados. Las destrezas por el primer grupo son:

54.2 39.6 52.3 48.4 35.9 30.4 25.2 45.4 48.9 48.9
45.8 44.0 52.5 48.3 59.9 51.7 38.6 39.1 49.9 38.3

Las del segundo grupo son:

30.3 43.0 25.7 26.7 27.3 31.9 53.7 32.9 19.4 23.7
23.3 23.3 37.8 39.5 33.5 30.4 28.5

1. Calcule la mediana para el primer grupo. Realice una gráfica de puntos y ubique la mediana.
2. Calcule la mediana para el grupo 2. Realice una gráfica de puntos y ubique la mediana.
3. Junte ambos conjuntos de datos y calcule la mediana.
4. Calcule la media de los incisos anteriores.
5. Interprete los valores de estas dos medidas.

Capture los datos de los siguientes dos puntos y use el **CalEst**.

3.8 El salario que percibieron 20 albañiles por una semana de trabajo fue:

1210, 1440, 2050, 1530, 1250, 1260, 1280, 1300, 2130, 1480
1310, 1410, 1390, 1340, 1350, 1470, 1410, 1440, 1330, 1370.

Con la calculadora estadística obtenga la media y la mediana. Interprete los resultados.

3.9 Una muestra de 35 estudiantes que compraron material escolar al inicio de un semestre arrojó que se gastaron las siguientes cantidades en pesos:

827	839	800	775	736	755	834	885	772	823	680	834
639	769	809	818	819	786	775	725	793	804	775	715
800	778	674	761	756	711	740	854	844	768	785	

Use la calculadora estadística para calcular tanto la media como la mediana de los datos anteriores.

3.10 El administrador de juegos en una feria considera que el tiempo de espera para subirse a la rueda de la fortuna debe estar entre 5 y 10 minutos. La información de una muestra del tiempo que esperaron 12 personas fue: 5.5, 9.6, 5.1, 9.3, 9.5, 13.6, 14.1, 9.7, 15.0, 8.6, 6.5, 9.1. ¿Cuál es la mediana de los tiempos de espera? ¿Se cumplen las expectativas del administrador?

Aplicación: medidas de tendencia central

3.11 Responda las siguientes preguntas: ¿Por qué es preferible usar la media muestral o la mediana muestral si los datos están sesgados? ¿Qué se puede decir sobre la media muestral de una variable cualitativa?

3.12 El tiempo (en minutos) que esperan 6 personas para realizar un trámite en una oficina gubernamental fue: 15, 3, 46, 432, 126, 64. Calcule la media y la mediana. ¿A qué se puede atribuir la evidente diferencia en los valores de estas medidas?

3.13 Elabore un diagrama de puntos para los siguientes dos conjuntos de datos y luego calcule la media y la mediana.

1. 12, 14, 17, 18, 22, 21, 24, 25.

2. 12, 14, 17, 18, 22, 21, 34, 40.

¿Qué observa de estos conjuntos de datos? ¿Qué puede concluir?

3.14 El problema se refiere al tamaño (en cm) y peso (g) de 143 pescados (véanse tablas de abajo); la administración de una tienda de autoservicio estudia la talla y peso de los pescados para fijar los precios. Mediante la calculadora estadística obtenga los datos e interprete los resultados. Elabore los polígonos de frecuencia relativa y frecuencia relativa acumulada, y ubique tanto la media como la mediana en ambos polígonos. En cada caso indique si la distribución es simétrica, sesgada positiva o sesgada negativa.

Longitud

55.8	50.7	53.3	56.6	50.8	59.8	60.5	49.7	59.4	58.3	62.7
55.5	42.4	55.3	61.3	55.5	31.8	45.2	52.6	50.2	61.7	48.2
54.3	58.4	51.1	55.6	47	51	55.5	61.9	51.1	60.8	55.9
59.1	54.6	57.2	44.4	57.5	50.1	59.4	50.7	54.8	56.8	50
55	52.2	51.9	48.7	58.9	44.1	61.9	50.7	57.6	52.7	50.5
62	43.2	55	64.5	55.8	53.6	48.5	51.7	49.3	45.3	57.6
56.7	61.1	42.2	52.7	56.8	51.9	59.5	53.7	56.7	38.7	57.5
38.1	51.1	50	51.4	50.8	46.7	44.4	52.3	53.6	58.7	39.5
65.6	63.1	64.1	45.8	56.3	68.2	57.7	50.3	54.5	59	37.7
51.5	54.8	58.5	48.4	63.2	60.2	53.8	45.4	63.3	41.9	47.7
52.9	59.5	55.7	38.6	50.4	44.6	53.8	52.7	60.6	59.6	52.8
50.7	50.4	53.6	50.2	42.1	59.2	47.6	51.3	60.8	58.9	52.3
57.9	52.1	54.6	61.2	61.1	43.9	42.4	53.9	57.5	57.1	56.1

Peso

1493	1457	1245	1424	1357	1401	1324	1290	1370	1322	1181
1257	1380	1393	1325	1289	1370	1380	1420	1151	1326	1335
1188	1224	1480	1214	1278	1316	1442	1296	1414	1259	1143
1398	1469	1449	1253	1376	1429	1391	1406	1278	1465	1459
1401	1385	1447	1468	1372	1241	1464	1426	1423	1488	1260
1332	1366	1224	1282	1392	1338	1404	1487	1453	1471	1399
1355	1454	1454	1391	1332	1412	1532	1356	1365	1310	1412
1372	1409	1406	1430	1459	1439	1398	1345	1260	1472	1328
1331	1410	1108	1363	1443	1472	1460	1348	1303	1264	1474
1444	1344	1315	1528	1452	1290	1371	1316	1425	1386	1383
1408	1339	1361	1362	1304	1425	1243	1238	1433	1489	1374
1022	1334	1268	1472	1349	1341	1421	1271	1300	1433	1277
1350	1437	1273	1347	1302	1445	1357	1226	1420	1435	1311

3.15 Se investiga la absorción de plomo en niños que viven alrededor de una zona industrial. Los datos que siguen indican el nivel de plomo en la sangre (microgramos por decilitro: μg/dl) de 33 niños.

38	23	41	18	37	23	62	31	34	24	36
14	21	17	16	20	15	10	45	39	35	22
49	48	44	35	43	39	34	13	73	25	27

1. Calcule la media y la mediana para el nivel de plomo en la sangre e interprete los resultados.
2. Elabore los polígonos de frecuencia relativa y frecuencia relativa acumulada del nivel de plomo en la sangre y señale si la distribución es simétrica, sesgada positiva o sesgada negativa.
3. Dibuje un diagrama de tallo y hoja para este conjunto de datos.

3.16 Para proponer una compensación al salario, el administrador de una escuela desea conocer la distancia del desplazamiento diario de su personal. Toma una muestra aleatoria de 40 profesores y se les pregunta cuál era la distancia en kilómetros que recorrían todos los días de su casa a la escuela. Calcule la media y la mediana de este conjunto de datos. Dibuje un diagrama de tallo y hoja para observar la simetría o sesgo de la distribución. Interprete el valor de la media para estos datos. ¿Qué distancia recorre 50 % de los profesores? Considerando la distribución, ¿qué puede concluir?

8	18	10	14	22	26	7	12	41	24	27	6	4	5
3	8	7	16	48	10	4	40	9	19	17	14	17	
9	48	29	5	6	11	20	21	8	4	22	9	28	

3.17 Suponga que usted es el gerente de una empresa, y desea calcular las medidas de tendencia central para los niveles de utilidad de su firma durante los últimos 12 meses. Además comente los beneficios relativos de usar cada estadístico. Considere las siguientes utilidades mensuales dadas en miles de pesos.

125	141	257	214	214	232
190	213	-54	197	232	-28

Media armónica y media geométrica

3.18 Una compañía que entrega paquetes tiene cuatro mensajeros que utilizan un trayecto de un lugar A, a otro lugar B en 3, 5, 4 y 7 horas, respectivamente. Calcule el tiempo promedio que utilizará un mensajero para hacer este trayecto. Si el costo por hora a la empresa le cuesta 50 pesos, ¿cuál es el costo total de las entregas?

3.19 Se realizó el recorrido de una ciudad A, a una ciudad B y debido al tipo de carretera se realizaron tres velocidades diferentes: 80 km/h, 110 km/h y 130 km/h. Obtenga la velocidad promedio de ese recorrido.

3.20 La producción de un artículo ha experimentado un incremento de 10 % del primero al segundo año y de 50 % del segundo año al tercer año. ¿Cuál es la tasa promedio durante estos dos años?

3.21 El número de reconocimientos que una empresa otorga a sus trabajadores durante los últimos seis años consecutivos se muestra a continuación:

Año	1997	1998	1999	2000	2001	2002
Número	6	9	15	22	30	38
Razón al año anterior		1.5	1.67	1.47	1.36	1.19

¿Cuál es el porcentaje medio de incremento en el número de reconocimientos otorgados por año?

Medidas de posición 1

3.22 Los tiempos registrados de nueve personas que esperaron para comprar un boleto en una línea de autobuses fueron los siguientes: 2, 4, 5, 6, 6, 8, 10, 10 y 12 minutos. Encuentre la mediana, así como el primer y el tercer cuartiles.

3.23 Los tiempos que esperaron 8 personas para comprar un boleto en otra línea de autobuses fueron: 3, 4, 7, 8, 11, 13, 21 y 29 minutos. Encuentre la mediana, el primer y el tercer cuartiles.

3.24 Compare los tiempos de tardanza de ambas líneas de autobuses en función de las medidas estadísticas que calculó en los dos ejercicios anteriores. ¿Qué observa con respecto a estas medidas?

3.25 El número de años que necesitaron 40 personas para concluir el bachillerato en el sistema de preparatoria abierta son:

4,	3,	4,	3,	3,	5,	5,	6,	4,	4,	4,	3,	3,	4,	3,	3,	6,	4,	5,	3,
6,	3,	2,	4,	1,	4,	4,	4,	4,	4,	5,	3,	4,	3,	1,	2,	2,	5,	2,	4,

Encuentre la moda, la media y la mediana y compárelos en el contexto del problema.

3.26 En una tienda que vende ropa para mujer, se registró la talla de blusas que vendió durante las últimas dos semanas. Los datos son:

6,	8,	8,	10,	10,	10,	10,	10,	10,	10,
10,	10,	10,	10,	10,	12,	12,	12,	12,	12,
12,	12,	12,	12,	12,	14,	14,	14,	14,	16,

Calcule la moda. ¿Qué tan adecuadas pueden resultar la media y la mediana en este ejemplo?

3.27 ¿Qué medida de tendencia central utilizaría para discutir el salario en una compañía grande, desde el punto de vista de la gerencia, del sindicato y de los empleados?

3.28 El gasto (pesos) de transporte que realizan una muestra de 40 familias se presenta a continuación:

43	52	63	55	66	78	79	87	42	43	44	49	52	53	53	53	64	58	76	78
70	78	86	87	90	81	84	85	97	79	88	96	91	92	89	94	60	85	85	97

1. Trace un polígono de frecuencia acumulado y estime los cuartiles.

2. Calcule las medidas de tendencia central e interprete los resultados en el contexto del problema.

3.29 La pérdida de calcio es un problema que se presenta principalmente en las mujeres mayores. Un médico investiga, en una muestra de 40 mujeres, la pérdida de calcio a lo largo de un año. A cada mujer le hace una medición inicial de calcio y luego al año siguiente una nueva medición. Los datos de pérdida de calcio al año son:

4	4	5	10	11	8	2	5	8	5	17	2	15	10	8	7	13	7	1	16
3	2	0	1	11	7	1	9	6	9	11	7	16	12	3	3	11	8	7	11

1. Encuentre la media y mediana muestral.

2. ¿Cuál de las dos medidas dan mejor indicación de la pérdida de calcio?

Medidas de posición 2

3.30 Para estimar el número de árboles de café en una granja, el agrónomo divide la granja en 1000 pequeñas parcelas. Él selecciona de manera aleatoria 20 de estas parcelas y cuenta el número de árboles. Los resultados son:

41	56	47	59	24	37	23	53	44	43
62	28	54	41	30	44	52	69	34	46

Calcule el rango y el rango intercuartil para este conjunto de datos. ¿Qué información adicional le proporciona este nuevo cálculo, unido al de la media, la mediana y al primer y tercer cuartil?

3.31 Al inicio del semestre un profesor realiza una prueba para evaluar la lectura de comprensión a una muestra de 20 alumnos. Se califica sobre 100, y los resultados de la prueba se describen a continuación:

24	31	54	62	36	28	37	55	18	27
58	32	37	41	55	39	56	42	29	35

Durante el semestre el profesor aplicó un método para mejorar la lectura de comprensión. Al finalizar el semestre se utilizó una prueba similar en una muestra de 25 alumnos. Los resultados son:

64	71	81	43	69	75	86	58	63	66	82	62	79
91	83	55	68	74	48	66	84	77	73	59	55	

1. Calcule el rango, el rango intercuartil y elabore un diagrama de dispersión, y trace la media para el primer conjunto de datos.

2. Calcule el rango, el rango intercuartil y elabore un diagrama de dispersión, y trace la media para el segundo conjunto de datos.

3. ¿Resultó efectivo el método del profesor para la lectura de comprensión?

Medidas de dispersión

3.32 Calcule la desviación media, la varianza y la desviación estándar de los siguientes conjuntos de datos. En cada caso reproduzca una gráfica similar a la figura 3.9:

1. La estatura de cinco personas es: 157, 183, 163, 152 y 157 centímetros.
2. El salario, en miles de pesos, que reciben ocho profesores de secundaria al mes son: 8.75, 6.9, 8.2, 9.4, 10.7, 7.1, 8.3, 6.7.
3. El salario, en miles de pesos, que reciben 9 jefes de departamento (funcionarios) al mes son: 35.8, 36.4, 38.6, 29.4, 30.8, 39.6, 28.9, 37.3, 36.1.
4. El salario, en miles de pesos, que reciben 10 líderes sindicales al mes son: 75.6, 55.9, 46.2, 79.4, 57.2, 67.8, 80.3, 57.7, 66.9, 70.5.
5. En una muestra de 8 días al mes, el administrador de una empresa grande anota el ausentismo de los trabajadores, este es: 8,12, 10, 5, 11, 8, 3, 9.

3.33 A un grupo de estudiantes se les aplicó un examen de opción múltiple de geografía económica. Las calificaciones que obtuvieron se dividieron en dos grupos: en el primero están los que sacaron entre 6 y 8; en el otro, los que obtuvieron una calificación mayor que 8.

1. El tiempo que emplearon los 22 estudiantes del primer grupo en contestar las preguntas se muestra a continuación:

95	85	87	82	98	92	97	103	92	94	94	98	90	100	92	100
92	91	92	100	87	94	75									

   Calcule la desviación media, la varianza y la desviación estándar del tiempo. Luego, complete el resumen estadístico. Elabore el diagrama de tallo y hoja para estos datos.

2. El tiempo que emplearon los 23 estudiantes del segundo grupo en contestar las preguntas se muestra a continuación:

122	116	120	121	120	115	118	115	113	112	117	115	122	119
119	126	117	118	120	123	129	125	112					

   Calcule la desviación media, la varianza y la desviación estándar del tiempo. Luego complete el resumen estadístico. Elabore el diagrama de tallo y hoja para estos datos.

3. Compare los resultados de los incisos a y b. ¿Qué puede concluir?

3.34 Una empresa que elabora alimentos procesados realiza pruebas para determinar la vida de anaquel de un nuevo producto. En el estudio se consideran 21 productos. El número de días que duraron los productos sin descomponerse se presenta en el siguiente cuadro.

152	152	115	109	137	88	94	77	160	165	125	40	128	136
62	153	83	69	132	120	101							

1. Calcule desviación media, la varianza y la desviación estándar.
2. Complete el resumen estadístico para este conjunto de datos.
3. En cada caso interprete su valor.

3.35 Con el fin de estudiar la eficiencia de un medicamento, la administración de una clínica probaron dos tratamientos para reducir los niveles de colesterol. El tratamiento 1 se aplicó a 13 pacientes. El nivel de colesterol se midió antes de la aplicación y después de ella. Los datos que indican la reducción son:

54	39	44	53	56	66	34	61	36	67	32	22	40

El segundo tratamiento se empleó en 11 personas y los resultados fueron:

40	31	50	40	52	44	74	38	81	64	66

Con base en los datos responda lo siguiente:

1. Calcule desviación media, la varianza y la desviación estándar para el primer tratamiento.
2. Calcule desviación media, la varianza y la desviación estándar para el segundo tratamiento.
3. En qué tratamiento hay mayor variación.
4. Complete un resumen estadístico para cada tratamiento.
5. Intuitivamente mencione cuál tratamiento es mejor. Interprete sus resultados.

3.36 Realice la siguiente encuesta a una muestra aleatoria de 25 compañeros de su grupo o de su escuela. Para seleccionar la muestra consiga la lista de su grupo o un número de referencia para la escuela, por ejemplo número de credencial. Pregunte lo siguiente:

1.
   - ¿Cuánto gasta a la semana en transporte?_____
   - Aproximadamente en un día ¿cuánto tiempo emplea para trasladarse de su casa a la escuela?_____
   - ¿Cuántas horas a la semana dedica a ver televisión?_____
   - ¿Cuál es su peso?_____
   - ¿Cuál es su estatura?_____

   Calcule su índice de masa corporal, dividiendo el peso entre el cuadrado de la estatura: peso/estatura_____

2. En cada situación cuestiónese por qué puede ser de interés contar con la información sobre estos puntos.

3. Realice un resumen estadístico para cada punto.

4. Interprete sus resultados estadísticos en términos del planteamiento en el inciso a). Obtenga sus conclusiones.

3.37 Por internet se anuncia la venta de coches usados del 2011, el precio por mil pesos, de 9 coches de una marca específica de tipo compacto son:

149 170 154 109 195 145 199 178 165

1. Clasifique las siguientes variables como cuantitativas o cualitativas y describa valores posibles de éstas.

   a) Marca de coche

   b) Tipo de coche

   c) Precio del coche

2. Calcule las medidas de tendencia central

3. Calcule las medidas de posición, primer y tercer cuantil.

4. Calcule las medidas de dispersión

5. ¿Cómo interpreta la existencia de los valores extremos 109, 195 y 199? ¿En qué medidas se ven reflejadas estos tres datos?

3.38 Se toma una muestra de 12 coches del mismo tipo como los descrito en el punto anterior, pero éstos se vende en los llamados tianguis. Los precios son:

178 138 163 122 147 145 154 208 153 146 188 169

1. Realice la actividad de los incisos b, c y d del ejercicio anterior. Discuta las diferencias.

2. Haga los diagramas de caja y discuta las diferencias.

3.39 Encuesta: Entreviste a 10 compañeros: Con la finalidad de tener una idea y poder hacer algunas estimaciones sobre algunas características de interés de los estudiantes para poder finalizar sus estudios en alguno de los niveles escolares. Así como alguna información relacionada con algunos puntos económicos cercanos al entorno del ingreso familiar. Solicitamos nos permitan realizar algunas preguntas que serán de interés únicamente estadístico. Actividad totalmente anónima. En cada caso calcule las medidas de tendencia central, posición y dispersión.

1. ¿Cuál es el gasto semanal en transporte?
2. ¿Cuánto gastan en material educativo al semestre?
3. ¿Cuántas horas estudian a la semana?
4. ¿Cuántos libros compran al semestre?
5. ¿Qué tiempo realizan de su casa a su centro educativo?
6. ¿Cuánto tiempo le dedica a buscar temas escolares en internet?
7. ¿Cuántas veces a la semana está en internet por distració?
8. ¿Cuál es el consumo de energía, medida en Kwh, en su hogar en el último bimestre?
9. De la anterior en los últimos 5 o 6 bimestres

3.40 Del ejemplo 20 de este capítulo tome tres muestras de tamaño 7. En cada caso calcule las medidas de tendencia central, posición y dispersión. Compare sus resultados.

Medidas de datos agrupados

3.41 El número de elotes de una muestra aleatoria de 60 plantas se presenta a continuación:

4	3	4	3	3	5	5	6	4	4	4	3	4	3	3	6	4	5	4	5
4	4	4	4	4	5	3	4	3	1	2	2	2	4	3	5	5	3	5	5
5	6	4	4	3	3	3	5	6	3	1	3	4	2	3	7	6	3	3	3

1. Construya la tabla de frecuencias.
2. Grafique un diagrama de puntos.
3. Estime la media y la varianza usando las expresiones para datos agrupados.

3.42 A continuación se presenta 50 mediciones de lluvia ácida medida en una región determinada. La acidez se mide en una escala de pH, donde 1 es muy ácido y 7 es básico.

3.58	4.05	4.27	4.35	4.45	4.51	4.58	4.62	4.7	5.07	4.05	4.5	4.62
3.8	4.12	4.28	4.35	4.5	4.52	4.6	4.65	4.72	5.2	4.21	4.6	5.48
4.01	4.18	4.3	4.41	4.5	4.52	4.61	4.7	4.78	5.26	4.33	4.8	
4.01	4.2	4.32	4.42	4.5	4.52	4.61	4.7	4.78	5.41	4.33	4.7	

1. Estime la media y la varianza usando las expresiones para datos agrupados.
2. Calcule la mediana y los cuartiles.
3. Calcule la media y la desviación estándar y compare estos resultados con los obtenidos en el inciso a.

4. Determine los intervalos

5. ¿Qué proporción de los datos están en estos intervalos?

3.43 El ozono es un indicador del aire en ciudades grandes y en las zonas industriales, lo cual tiene efectos en la economía. A continuación se muestran 78 registros máximos de la concentración (ppm) de ozono en una zona industrial.

65	98	54	98	85	105	88	64	61	74	83	96	147
44	55	60	47	41	92	61	88	77	87	65	67	124
96	84	86	83	81	90	46	106	68	75	71	96	106
90	74	77	77	86	88	88	44	89	67	97	86	87
72	84	95	104	85	58	55	67	63	124	69	70	98
74	77	60	90	44	91	111	50	92	92	71	64	96

1. Dibuje el polígono de frecuencia acumulado y estime la mediana, el primer y tercer cuartil.

2. Estime la media y la varianza usando las expresiones para datos agrupados.

3. Calcule la media y la varianza de estos datos y compare los resultados con el inciso anterior.

4. Estudie la regla empírica

3.44 Considerando los datos del mundo de la información 3: Competencia atlética. Haga los puntos de a-d del ejercicio anterior.

1. Tome dos muestras de tamaño 9 y calcule la media, la mediana y la desviación estándar, para cada muestra.

2. Luego elabore un diagrama de caja, apartado siguiente, para todos los datos y para las dos muestras discuta sus resultados.

3.45 Usando los datos del ejemplo 6 haga los puntos de a-d del ejercicio anterior.

1. Tome dos muestras de tamaño 12 y calcule la media, la mediana y la desviación estándar, para cada muestra.

2. Luego elabore un diagrama de caja, apartado siguiente, para todos los datos y para las dos muestras discuta sus resultados.

Diagramas de caja

3.46 El peso, en gramos, de 20 ratones para pruebas de laboratorio seleccionados de manera aleatoria son:

15.7	14.8	13.7	16.1	15.2	13.1	16.9	15.8	16.3	14.9
13.9	14.2	16.7	14.1	16.2	13.4	14.7	15.6	17.0	16.0

1. Use la calculadora estadística para elaborar un resumen estadístico.

2. Trace el diagrama de caja.

3.47 Una institución gubernamental, realizó un estudio sobre quién es quién en los precios de un medicamento. De entre varios de ellos, se encuentra el reporte de Avelox 7d, de la compañía Bayer, caja de 400 mg. Se seleccionó una muestra aleatoria de 22 farmacias, que arrojó los siguientes precios:

247.5	259.0	262.0	290.0	300.0	370.0	280.0	420.5	388.0	426.5	450.0
253.0	268.0	350.5	420.5	283.5	362.0	434.0	298.5	386.0	278.5	358.5

1. Use la calculadora estadística para elaborar un resumen estadístico.

2. Trace el diagrama de caja.

3.48 Datos del ejercicio 3.1. El salario, en pesos, que percibieron 20 albañiles por una semana de trabajo fue: 1210, 1440, 2050, 1530, 1250, 1260, 1280, 1300, 2130, 1480, 1310, 1410, 1390, 1340, 1350, 1470, 1410, 1440, 1330, 1370.

1. Escriba un resumen estadístico de este conjunto de datos. ¿Son simétricos los datos? ¿La variación de estos datos es grande?

2. Dibuje un diagrama de caja, ¿qué observa?

3.49 Se seleccionó una muestra aleatoria de 40 estudiantes, a quienes se les aplicó un cuestionario. Entre las preguntas había dos que se referían al género y a la estatura. A continuación se describen los datos; al género masculino le corresponde el 1 y al femenino, el 2; la estatura se pidió en cm.

Sexo	Estatura	Sexo	Estatura	Sexo	Estatura	Sexo	Estatura
1	183	2	166	1	169	2	157
2	163	2	157	2	168	2	169
2	152	2	168	2	165	2	177
2	157	2	167	2	166	2	174
2	157	2	156	2	164	1	183
2	165	2	155	2	163	1	181
1	173	1	178	2	161	1	182
1	180	2	169	2	157	1	171
2	164	2	171	1	181	1	184
2	160	1	175	2	163	1	179

1. Calcule la media y la mediana para la estatura (sin distinción de género) e interprete los resultados.

2. Elabore los polígonos de frecuencia relativa y frecuencia relativa acumulada de la estatura y señale si la distribución es simétrica, sesgada positiva o sesgada negativa.

3. Dibuje un diagrama de caja para las estaturas de las mujeres y otro para la estatura de los hombres. ¿Qué puede concluir?

4. Calcule los coeficientes de variación para las estaturas de mujeres y hombres, respectivamente. ¿Qué observa?

3.50 Se desea estudiar el efecto de la televisión en el comportamiento de los adolescentes. Para ello, se escoge una muestra aleatoria de 26 y se pregunta por el número de horas por día que ven televisión. Los datos registrados son:

3.6	3.7	3.7	3.8	3.9	3.9	3.9	3.9	3.9	3.9	4.2	4.3	4.6	5.0	5.3	5.6	5.7
5.8	6.0	6.0	6.0	6.0	6.0	6.0	6.3	6.9								

1. Dibuje el diagrama de caja para este conjunto de datos.

2. ¿Este diagrama de caja es una buena representación para este tipo de datos?

3. ¿Entre qué horas está aproximadamente el 68 % de los adolescentes?

3.51 La dieta en carbohidratos (mg por día) para una muestra de 30 mujeres mayores de 40 años es:

199	162	327	145	149	351	453	374	287	151
201	375	223	230	193	229	206	144	152	164
121	190	158	145	129	168	173	189	589	247

1. Construya un diagrama de caja para estos datos. ¿Qué se puede concluir?

2. Estime la proporción de mujeres que están en el intervalo.

3.52 Se estudia la cantidad de monóxido de carbono (en gramos por kilómetro) emitido por una muestra aleatoria en 44 automóviles similares. El objetivo del estudio es producir productos que ayuden a disminuir estas emisiones. Los datos registrados son:

5.01	14.67	8.6	4.42	4.95	7.24	7.51	15.13	5.04	3.95	6.02
12.3	14.59	7.98	11.53	4.1	5.21	12.10	3.38	4.12	23.53	3.99
19.0	22.92	11.2	3.81	3.45	1.85	4.10	2.26	4.74	4.29	5.22
5.36	14.83	5.69	6.35	6.02	5.79	2.03	4.62	6.78	8.43	14.97

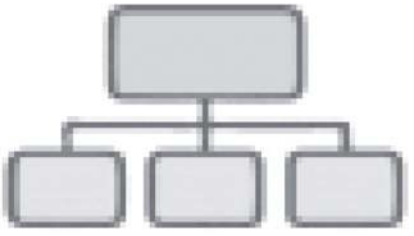
1. Construya un diagrama de caja para estos datos. ¿Qué se puede concluir?

2. Estime la proporción de mujeres que están en el intervalo: y

3. Estime la simetría de esta distribución.

3.10 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.



Capítulo 4

Estadística y probabilidad

Hay cosas inciertas para nosotros, cosas más o menos probables, y nosotros tratamos de comprender la imposibilidad de conocerlas por el procedimiento de establecer sus diversos grados de probabilidad. En consecuencia, debemos a la debilidad de la mente humana una de las teorías matemáticas más delicadas e ingeniosas, la ciencia del azar y la probabilidad.

Pierre Simon Laplace

Competencia general

Comprender los conceptos básicos de la probabilidad para evaluar la importancia que tiene esta materia en la explicación de muchos problemas de la vida real, en particular en administración y economía, así como adquirir habilidad en el cálculo de probabilidades.

Competencias específicas

- Conocer y comprender los conceptos de ensayo, experimento y espacio muestra.
- Comprender el procedimiento para asignar una probabilidad a sucesos elementales, conocer y aplicar la definición de probabilidad clásica y empírica.
- Conocer tres relaciones básicas de eventos: complemento, unión e intersección. A partir de éstas comprender las leyes de complemento y adición para adquirir habilidad en el cálculo de probabilidades.
- Aplicar las leyes de probabilidad a situaciones con mayor grado de dificultad en el cálculo de probabilidades.
- Aprender las técnicas de multiplicación, permutación y combinación para obtener el número de posibles arreglos en conjuntos de objetos.

Continúa

Competencias específicas. Continuación

- Adquirir habilidad en identificar los elementos que componen un espacio muestra y asignarles una probabilidad.
- Describir diferentes situaciones o problemas donde se puedan plantear fenómenos de tipo aleatorio.
- Escribir con la mayor claridad posible las reglas de probabilidad y explicarlas mediante ejemplos.
- Adquirir habilidad para dominar el vocabulario relacionado con las ideas esenciales en probabilidad.
- Identificar las relaciones básicas de eventos y aplicarlas en el cálculo de probabilidades.
- Conocer las técnicas de multiplicación, permutación y combinación para obtener el número de posibles arreglos en conjuntos y objetos.
- Aplicar los conceptos y las leyes de probabilidad para realizar ejercicios en diferentes situaciones.
- Establecer la relación entre la estimación de intervalos de confianza y pruebas de hipótesis.
- Leer un artículo de divulgación en temas de administración o economía para identificar aplicaciones de probabilidad.

4.1 Introducción

4.1.1 Elemento básicos y nociones de probabilidad

Como sabemos, uno de los principales objetivos de la estadística es ayudar o apoyar en la toma de decisiones en condiciones de incertidumbre. La probabilidad es el camino para cuantificar los resultados que no se pueden predecir con certidumbre. Así, en la operación de variables aleatoria se requiere del conocimiento de las distribuciones de probabilidad. Éstas tienen una amplia aplicación en inferencia estadística, además permiten modelar diferentes problemas que son de utilidad en diversas áreas del conocimiento para explicar e interpretar el comportamiento de la variable aleatoria.

Las nociones de probabilidad desempeñan un papel esencial en el análisis e interpretación de los datos estadísticos. Por ello, este capítulo contiene un complemento didáctico con la opción de usar un material

animado para motivar la enseñanza y comprensión de algunos conceptos de probabilidad. El material incluye juegos clásicos con monedas, dados, extracción de canicas con reemplazo y sin reemplazo, las ruletas con tamaño fijo y aleatorio, entre otros. En el marco teórico de este trabajo se indican algunos conceptos relevantes de la probabilidad, los cuales se expondrán en algunos ejemplos. Muchos fenómenos de la naturaleza están gobernados por el mundo del azar, y la probabilidad es un camino para entenderlos; el efecto visual ayudará para su cálculo e interpretación.

Pierre-Simon Laplace (1749-1827)

Fue uno de los grandes impulsores del desarrollo de la teoría de la probabilidad, a la cual denominó "el sentido común reducido al cálculo" En 1767 empezó a dar clases de matemáticas en la Escuela Militar de París; uno de sus alumnos fue Napoleón. Durante la revolución francesa ¡salvó la cabeza! y a cambio de ello lo pusieron a calcular trayectorias para la artillería. Entre otras, tuvo contribuciones relevantes en astronomía, física y matemáticas, donde destacan sus conceptos de la transformada de Laplace y la ecuación de Laplace.

El mundo del azar. Ideas preliminares

En los noticiarios que se transmiten por radio y televisión se ofrecen reportes acerca del estado del tiempo y en los principales diarios nacionales se proporciona una descripción detallada del clima. Es común escuchar o leer frases como éstas: "Hay poca probabilidad de lluvia" o "Habrá lluvias ocasionales" o "Hay alta probabilidad de lluvia".

En nuestro lenguaje habitual nos encontramos expresiones como las siguientes: "Llegué a esa dirección por casualidad", "Se notaba tan sano que fue inesperado verlo enfermo", "Por suerte obtuve el premio que se rifó en la kermés", "El descubrimiento de la penicilina se debió a un hecho accidental", "Le ganamos al equipo campeón por pura casualidad". Para ilustrar los temas relacionados con el azar en la vida, se puede escuchar la canción "es caprichoso el azar" de Joan Manuel Serrat.

Existen muchas situaciones en la vida cotidiana cuyos resultados se manifiestan de manera fortuita. Una gran cantidad de nuestras experiencias de carácter aleatorio están relacionadas con los juegos de azar.

Como sabemos, nuestra vida está rodeada por la presencia de fenómenos imprevisibles los cuales aparecen en los campos biológico (la fracción de insectos que mueren por la acción de un pesticida), médico (un tratamiento es mejor que otro), agrícola (la cantidad de un fertilizante para mejorar la producción de un fruto), social (número de errores en un inventario), político (la preferencia de un votante por un candidato), entre muchos otros más. Es decir, vivimos en un mundo donde la naturaleza está gobernada por la incertidumbre, y la probabilidad es un procedimiento para medirla.

En resumen, diremos que los *modelos de probabilidad* son parte fundamental de un segmento de la teoría estadística, por lo que resulta primordial aprender la teoría probabilística para alcanzar una adecuada comprensión de los métodos estadísticos. Estos últimos son una herramienta útil en la formación de todo estudiante universitario, ya que los métodos estadísticos son imprescindibles en el quehacer científico, industrial, social y profesional.

4.2 Eventos y espacio muestral

El mundo de la información 1. Genes

La determinación del sexo en los seres humanos está dispuesta por los cromosomas sexuales: una mujer siempre tiene dos cromosomas X, mientras que el hombre tiene un cromosoma X y uno Y. La transmisión de los caracteres genéticos sexuales se debe a un cromosoma X que el bebé recibe de su madre, y de su padre puede recibir un cromosoma X o Y, lo que define si es niño o niña. La sustitución de un cromosoma X por un Y representa la diferencia básica entre los sexos; así, la mitad de los recién nacidos podrían ser hombres y, por supuesto, la otra mitad mujeres.

Preguntas sobre la naturaleza del problema

Si bien, en el campo de la genética el asunto planteado en el problema anterior puede ser motivo de muchas especulaciones, aquí sólo queremos rescatar el hecho de que tenemos dos posibles resultados. Esto es, cuando en el proceso de fecundación un gameto masculino (célula de esperma) se une a un gameto femenino, hay igual probabilidad de que la unión dé como resultado una célula con dos cromosomas X o con un cromosoma X y uno Y. En el primer caso el recién nacido será niña y en el segundo, niño.

En principio, la variable género del recién nacido tendrá dos valores niña o niño, nuestros datos serán dos. Para tratar esta información desde el punto de vista de la probabilidad se ahondará en los conceptos de probabilidad, espacio muestral, experimentos y eventos.

Noción de probabilidad

¿Qué es probabilidad? Cuando se nos plantea esta pregunta, en general tenemos cierta idea de lo que significa esa palabra. En el caso que estamos tratando, la determinación del sexo de un ser humano, consideramos que cuando nace un bebé hay un 50 % de oportunidades de que sea niña y un 50 % de que sea niño, es decir, hay una oportunidad mitad–mitad (esto es, 50-50) de que así ocurra. De manera intuitiva, pensamos en la probabilidad como un valor numérico que se asocia con algún resultado e indica el grado de certeza de que pueda ocurrir el resultado.

Espacio muestral

Para entender lo que es un *espacio muestral* es necesario definir el concepto experimento. Un *experimento* es cualquier proceso que genera una observación (por ejemplo, en *El mundo de la información 1*, al nacer

un bebé se registra el sexo). Sin embargo, cabe aclarar que el concepto de experimento que mencionamos es más amplio que el empleado en las ciencias físicas, donde se usan diferentes equipos como tubos de ensayo, etc. Otros ejemplos de experimentos son: anotar la preferencia de un cliente por la marca de un teléfono; registrar la opinión de una persona respecto a la píldora del día después; medir la concentración de oxígeno en un río contaminado; lanzar un dado y anotar el número de la cara que queda arriba.

Experimento aleatorio

Un *experimento aleatorio* es el resultado de un proceso que genera una observación que no puede predecirse.

Se llamarán sucesos aleatorios a los resultados posibles de un *experimento aleatorio*. Por ejemplo, una experiencia aleatoria consiste en preguntar a una persona, elegida al azar de un grupo de 10 clientes, si es partidaria o no de consumir determinado producto. Los sucesos aleatorios en este caso son dos: es partidaria, no es partidaria.

Entre los resultados, se distingue a los *sucesos elementales o simples*, pues éstos no pueden descomponerse en otros más simples, mientras que los compuestos son los que se componen de dos o más sucesos elementales.

Se puede considerar cada suceso elemental asociado a un experimento como un elemento del conjunto formado por todos los sucesos elementales posibles asociados a ese experimento.

Ejemplo 4.1

Las calificaciones de 50 estudiantes para la materia de literatura se dividen en 5 categorías A, B, C, D y E. El experimento aleatorio consiste en seleccionar de manera aleatoria a un estudiante y observar en qué categoría está su calificación. Identificar los sucesos elementales y escribir el espacio muestral. Proponer un suceso compuesto.

Solución

Los sucesos elementales son 5 resultados posibles: $e_1 = A, e_2 = B, e_3 = C, e_4 = D, e_5 = E$.

Todos los resultados posibles para los 50 estudiantes se representan en el conjunto:

$$M = \{e_1, e_2, e_3, e_4, e_5\}$$

Un suceso compuesto es $\{e_4, e_5\}$, que consta de dos sucesos elementales. En el contexto del ejemplo se puede referir a que los alumnos que obtengan una calificación en las categorías D y E tienen que entregar un trabajo extra para aprobar el curso.

En resumen, se dice que un *experimento es aleatorio* si se cumplen los siguientes puntos:

1. Se repite bajo condiciones idénticas.
2. El resultado observado no se puede predecir.
3. El resultado que se obtiene, pertenece a un conjunto conocido previamente de resultados posibles. A este conjunto de resultados posibles se le denomina *espacio muestral.*

Espacio muestral

El *espacio muestral* es el conjunto de todos los posibles resultados de un *experimento aleatorio*. A cada resultado se le llama resultado *elemental* o elemento del espacio muestral.

Un *evento A* o suceso aleatorio es un resultado (evento simple) o conjunto de resultados que son de interés para el experimentador.

Ejemplo 4.2

Se describen varios casos de *experimentos aleatorios*, señalando todos sus posibles resultados.

Experimento aleatorio	M: Muestral
Conocer el estado de salud de una persona.	{Sano, enfermo}
Observar el tiempo de vida de una lámpara.	$[0,+\infty)$
Observar el tiempo de vida de un virus.	$[0,+\infty)$
Contar el número de vehículos que pasan por una caseta durante un intervalo de 15 minutos.	{0,1,2...}
Observar el número de viajeros que usarán el autobús.	{ 0,1,2...}
Contestar al azar un examen.	{Verdadero, falso}
Cobrar una póliza de seguro.	{Si, no}
Pesar a una persona.	(0,200)*
En un juego de azar tirar un dado de seis caras.	{1,2,3,4,5,6}

*Suponiendo que el peso no es mayor que 200kg

Ejemplo 4.3

1. En un estudio sobre el hábito de lectura de jóvenes entre los 20 y 25 años se estimó que 50 % no lee un libro (ciencia ficción, divulgación, novela, literario, otros géneros atractivos) durante las vacaciones de verano. Al regreso a clases se seleccionó de manera aleatoria a tres estudiantes y se les preguntó si habían leído en las vacaciones. En la línea responde "Sí " o "No" a la siguiente pregunta: ¿Leyó un libro en estas vacaciones?_____

   - Describa el espacio muestral y escriba una lista de los siguientes eventos:

   a) Evento A: Exactamente dos de tres leyeron.

   b) Evento B: Sólo uno leyó.

   c) Evento C: Los tres leyeron.

   d) Evento D: Al menos uno de los tres leyó.

2. La administración de una empresa realiza una serie de proyectos en sus diferentes departamentos; en la primera etapa elabora el diseño, en la segunda corresponde al desarrollo. En la primera etapa los diseños tardan en plantearse durante 2, 3 y 4 meses. En la segunda tardan en completarse 6, 7 y 8 meses; al final un proyecto se completa juntando las dos etapas.

   - Elabore el diagrama de árbol para describir los meses en que se completan los proyectos, y a partir de ahí describa el espacio muestral.
   - ¿Cuál es el evento A: que el proyecto termina en menos de 10 meses?
   - ¿Cuál es el evento B: que el proyecto termina en 10 o más meses?

Solución 1

Una de las metas en el estudio de la probabilidad es alcanzar el dominio y la habilidad en la elaboración de la lista de los resultados del experimento. El diagrama de árbol es una gráfica que resulta útil para alcanzar ese fin.

Para comprender el ejemplo se construye un diagrama de árbol. Cada uno de los alumnos responderá que sí leyó con una (s) y que no leyó por medio de una (n). En la figura 4.1 se describen las posibilidades sobre la práctica de lectura de los tres estudiantes.

A partir del diagrama del árbol tendremos definido el espacio muestral, el cual queda descrito por:

$$M = \{(sss), (ssn), (sns), (snn), (nss), (nsn), (nns), (nnn)\}$$

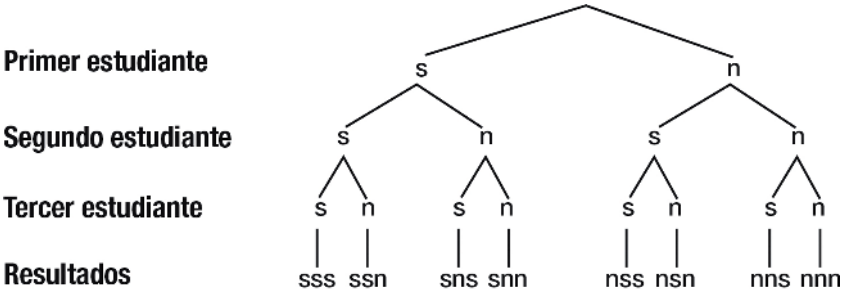


Figura 4.1 Descripción de los posibles resultados en el hábito por la lectura.

La segunda parte del ejemplo consiste en obtener los diferentes eventos, los cuales se obtienen a partir del espacio muestral. De los resultados se construye cada uno de los eventos. Así:

$$A = \{(ssn), (sns), (nss)\}$$
$$B = \{(snn), (nsn), (nns)\}$$
$$C = \{(sss)\}$$
$$D = \{(sss), (ssn), (sns), (snn), (nss), (nsn), (nns)\}$$

Solución 2

La figura 4.2, reporta la solución de la primera parte del inciso 2.

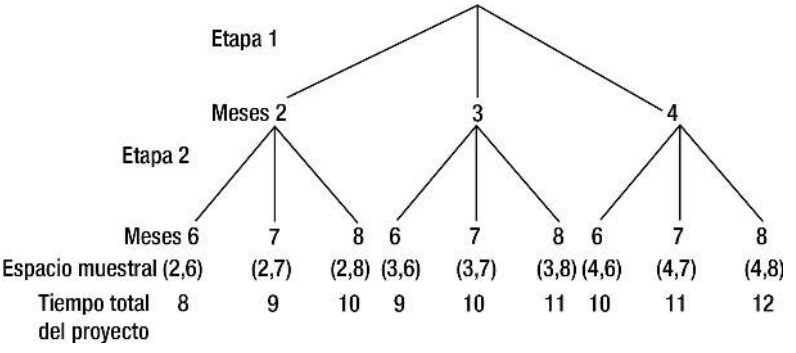


Figura 4.2 Descripción del diagrama de árbol para el tiempo en que se realiza un proyecto en dos etapas.

Los eventos para los casos señalados son:

$$A = \{(2,6), (3,6), (2,7)\}$$
$$B = \{(2,8), (3,7), (3,8), (4,6), (4,7), (4,8)\}$$

4.3 Probabilidad de un evento

El mundo de la información 2. Representante de una escuela

En una escuela existen 10 grupos y cada uno tiene un jefe de grupo. De los jefes de grupo 4 son hombres y 6 son mujeres. Los jefes de grupo se reunieron para elegir quién será el representante de la escuela ante el distrito escolar. Decidieron seleccionar a esa persona de manera aleatoria. Para asegurarse de que así fuera, colocaron en una bolsa 10 papeles, de los cuales 4 correspondían a hombres y los 6 restantes a mujeres. La pregunta por contestar es: ¿cuál es el espacio muestral? Enumere los resultados de los eventos: ¿cuál es la probabilidad de que la persona seleccionada sea mujer?

El espacio muestral es:

$$M = \{h_1, h_2, h_3, h_4, m_1, m_2, m_3, m_4, m_5, m_6\}$$

.

Los eventos A y B se pueden enlistar como sigue: $A = \{h_1, h_2, h_3, h_4\}$ y $B = \{m_1, m_2, m_3, m_4, m_5, m_6\}$.
Expresado en términos del problema, lo que se quiere es encontrar la probabilidad del evento B.

Asignar probabilidades a sucesos elementales

La descripción de un experimento asegura que cada resultado elemental es tan probable que ocurra como cualquier otro. La lista del espacio muestral es:

$$M = \{h_1, h_2, h_3, h_4, m_1, m_2, m_3, m_4, m_5, m_6\}$$

donde hay 4 hombres y 6 mujeres. La expresión de "manera aleatoria" significa que cualquiera de los 10 alumnos tendrán la misma probabilidad de ser electo representante. En este caso, se dice que los 10 resultados son igualmente probables, es decir, 1/10. Esta es la probabilidad de seleccionar a uno de los diez estudiantes, un suceso elemental. Éste se denota por:

$$P(\text{seleccionar un estudiante}) = \frac{1}{10}$$

Así se tiene que:

$$P(h_1) = P(h_2) = P(h_3) = P(h_4) = P(m_1) = P(m_2) = P(m_3) = P(m_4) = P(m_5) = P(m_6) = \frac{1}{10}$$

En general, la *probabilidad de un suceso elemental* es 1 entre el número de resultados posibles, esto es:

$$P(\text{un suceso elemental}) = \frac{1}{\text{número de resultados elementales en M}}$$

La *probabilidad* es una medida de incertidumbre, cuyo valor está entre 0 y 1 porque si es 0, se dice que es un resultado que no puede ocurrir; en cambio, si es igual a 1, se piensa en un evento que siempre ocurrirá (véase la figura 4.3).

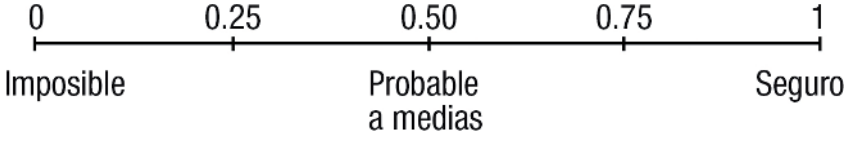


Figura 4.3 Escala de la probabilidad.

Asignar probabilidades de los sucesos elementales en un espacio muestral, debe satisfacer dos condiciones:

1. La probabilidad de cada resultado debe estar entre 0 y 1, inclusive.
2. La probabilidad de todos los resultados en el espacio muestral deben sumar 1.

Una vez que las probabilidades se asignan a todos los sucesos elementales del espacio muestral, podemos encontrar las *probabilidades de los eventos.*

A continuación se plantea la solución al problema de "El mundo de la información 2". Como se indicó anteriormente, la probabilidad de cada suceso elemental es de 1/10. El evento B, que corresponde al hecho de ser mujer, tienen seis sucesos elementales, por lo tanto la suma de las probabilidades de esos sucesos nos proporcionará la probabilidad de que la persona seleccionada sea mujer.

Una interpretación animada de la escala de probabilidades se muestra en la figura 4.4, aparece en la figuras públicas de probabilidad en internet.

$$P(mujer) = \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{6}{10}$$

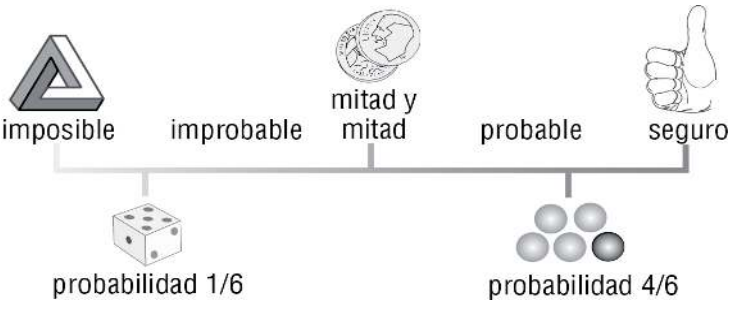


Figura 4.4 Escala de la probabilidad animada.

Definición de probabilidad clásica

En general, si el espacio muestral tiene N sucesos elementales, la *probabilidad de un evento* A que consta de r sucesos elementales, se puede expresar por:

$$P(A) = \frac{r}{N} = \frac{\text{número de resultados elementales en A}}{\text{número de resultados elementales en M}}$$

Entonces, la probabilidad de seleccionar a un hombre es:

$$P(hombre) = \frac{4}{10}$$

El total de las dos probabilidades es:

$$P(mujer) + P(hombre) = \frac{6}{10} + \frac{4}{10} = 1$$

y no existe otro resultado posible. A continuación se expresa la regla de probabilidad que en general se denomina probabilidad clásica a priori.

Probabilidad clásica

Considere que A es un evento de un espacio muestral M, donde cada resultado elemental en M es igualmente probable. Entonces la probabilidad de A es:

$$P(A) = \frac{\text{Número de formas en las que puede ocurrir el evento } A}{\text{número total de posibles resultados}}$$

Observación: Es importante destacar que para calcular probabilidades aplicando la definición, es necesario que todos los sucesos elementales sean igualmente probables.

Ejemplo 4.4

Beatriz (B), Jaime (J) y Luisa (L) son finalistas en un concurso de ortografía que se realizó en un distrito escolar. El ganador y el segundo lugar irán a la competencia estatal. ¿Cuál es la probabilidad de que Luisa gane el concurso local? ¿Cuál es la

probabilidad de que Beatriz no vaya al concurso estatal?

Solución

La lista del espacio muestra es $M = \{BJ; BL, JB, JL, LB, LJ\}$ donde BJ son Beatriz ganó y Jaime segundo, así los demás. El evento "Luisa ganó" se escribe: $A = \{\text{Luisa ganó}\} = \{LB, LJ\}$. Por consiguiente la probabilidad del evento A es: $P(A) = \frac{2}{6}$. El evento "Beatriz no va" es: $B = \{\text{Beatriz no va}\} = \{JL, LJ\}$. Entonces la probabilidad del evento B se escribe como: $P(B) = \frac{2}{6}$.

Teoría clásica. Laplace

El primer intento de presentar una definición con rigor matemático de la noción de probabilidad se debió a Laplace. En el año de 1812, Laplace dio la definición que se conoce como clásica de la probabilidad de un evento que puede ocurrir solamente un número finito de formas, y ésta dice que: la proporción del número de casos favorables al número de casos posibles, siempre que todos los resultados sean igualmente probables. Es decir, si w es el número de casos favorables de un evento A y m el número de casos posibles todos igualmente probables, entonces $P(A) = w/m$.

Ejemplo 4.5

Experimento con dados. Un jugador gana si en el primer tiro la suma de dos dados es de 7 u 11. Calcule la probabilidad de que el jugador gane en el primer tiro.

Solución

El espacio muestral para el par de dados tirados es:

M= {(1,1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),
(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),
(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),
(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),
(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),
(6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)}

En el supuesto de que los dados no estén cargados, se tienen 36 sucesos elementales que son igualmente probables; esto es, cada uno tiene una probabilidad de $\frac{1}{36}$. El evento de interés es que la suma sea 7 u 11 y la lista de éste es:

$$A = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1), (5,6), (6,5)\}$$

Sumando las probabilidades de los sucesos elementales en el evento A, se tiene:

$$P(A) = 8\left(\frac{1}{36}\right) = \frac{8}{36} = \frac{2}{9}$$

Así, la probabilidad de que el jugador gane es $\frac{8}{36}$ o $\frac{2}{9}$

Cálculo de la probabilidad de un evento

1. Definir el experimento y hacer la lista de los resultados del espacio muestral.
2. Asignar probabilidades a los sucesos elementales, de modo que cada uno esté entre 0 y 1.
3. Que la suma de las probabilidades de los sucesos elementales sea 1.
4. Listar los resultados que corresponden a un evento.
5. Sumar las probabilidades de los resultados que están en el evento.

Definición de probabilidad de frecuencias o empírica

La probabilidad de un resultado en un experimento puede describirse como la *frecuencia relativa* con la que un resultado ocurrirá si se repite el experimento un número grande de veces. Por ejemplo, si se toman los expedientes de los últimos 1000 nacimientos de un hospital, se puede decir que la probabilidad de que el recién nacido sea niña es $\frac{1}{2}$ o 50 %. Esta situación sería difícil de observar en 10 nacimientos.

La probabilidad se puede convertir a porcentaje si multiplicamos su valor por 100. Por ejemplo, una probabilidad de $\frac{3}{4} = 0.75$ en porcentaje es 75 %.

También podemos pensar en la probabilidad como una proporción. En el mismo sentido, si la probabilidad es $\frac{3}{4}$, la proporción de veces que un evento ocurrirá es $\frac{3}{4}$. Por ejemplo, en términos de germinación o de apuestas, la probabilidad $\frac{3}{4}$ equivale a la razón $\frac{3}{4}$ a $\frac{1}{4}$, que se expresa como 3 a 1 de que así ocurra. Como interpretación de este último resultado podría decirse que las apuestas están 3 a 1 a favor del equipo A sobre el B.

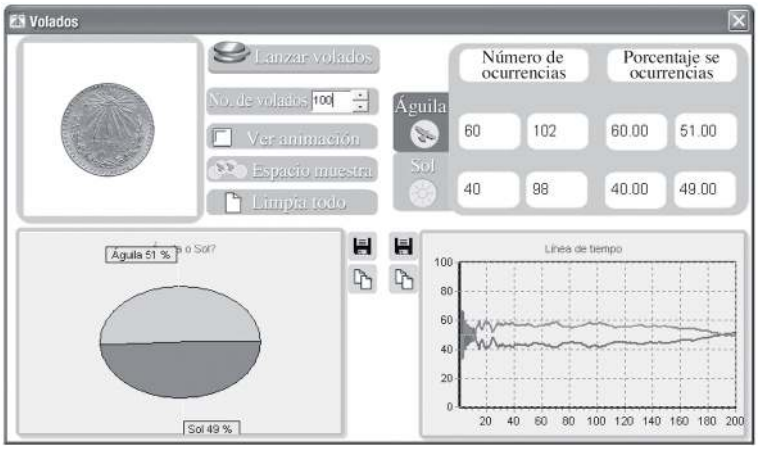


Figura 4.5 Lanzamiento de las monedas, para estudiar la probabilidad frecuentista y teórica.

Comentarios

- La palabra independiente significa que los resultados de cualquiera de los 1000 nacimientos no depende de los resultados de un nacimiento previo.

- El lanzamiento de una moneda permite visualizar con mayor claridad el cálculo de la frecuencia relativa. En la figura 4.5 se aprecia la alternativa visual del CalEst para el cálculo de probabilidades empírica y teórica, ésta aparece en el módulo de probabilidad. Ahí se tiene la opción de usar el simulador de lanzamiento de monedas, estimar la probabilidad de Águila o Sol. En éste se puede ver cómo la frecuencia relativa se estabiliza cuando el número de lanzamiento de la moneda crece. El espacio muestra incial consiste en $A = \{Águila, Sol\}$, si lanza más monedas puede ver la opción espacio muestra para obtener éste.
- En la figura 4.6 del módulo Probabilidad aparecen dos opciones con el juego de los dados. Con éstas puede ver la frecuencia relativa en cada lanzamiento y las probabilidades de ganar al jugar con los dados.
- En el complemento didáctico se integran una serie de dinámicas visuales para estudiar y comprender algunos conceptos de probabilidad.

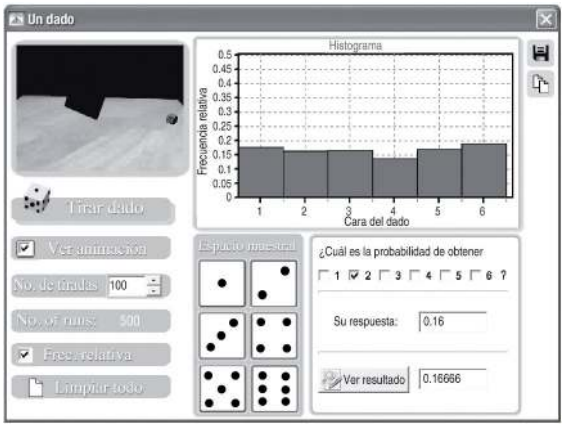


Figura 4.6 Se han lanzado 500 dados para observar la probabilidad frecuentista, en la otra opción se ha calculado la probabilidad de que al lanzar el dado salga un dos.

Probabilidad de un evento

Si en un número N grande de ensayos independientes, k de estos ensayos son resultados de un evento A, entonces la probabilidad del evento A se define como k/N y se expresa como:

$$\text{P(A)}=\frac{k}{N} = \frac{\text{Número de veces que ocurre A}}{\text{Número total de observaciones}}$$

Ejemplo 4.6

Debido a la posición geográfica de algunos países (por ejemplo, México), el cambio de horario, conocido como horario de verano, realmente parece no tener un impacto significativo en el ahorro de energía debido a que no hay mucha variación de la luz solar por día en las diferentes etapas del año. En apariencia, esto genera descontento entre algunos sectores de la población. En una encuesta realizada a 700 personas se les preguntó su punto de vista sobre el horario de verano. Se les pidió que marcaran con una X alguna de las siguientes opciones para manifestar su postura con respecto al cambio de horario.

A: Totalmente de acuerdo	26
B: Acuerdo	82
C: Sin decisión	181
D: Desacuerdo	177
E: Totalmente en desacuerdo	234
Total	700

Solución

A esta tabla se le agrega la columna de frecuencia relativa y se pueden evaluar algunas probabilidades de eventos que son de interés. Por ejemplo, estimar la probabilidad de las personas que no tienen decisión o saber si el porcentaje de las que están de acuerdo es mayor que las que no lo están. La tabla con la frecuencia relativa es:

A: Totalmente de acuerdo	26	26/700=0.04
B: Acuerdo	82	82/700=0.12
C: Sin decisión	181	181/700=0.26
D: Desacuerdo	177	177/700=0.25
E: Totalmente en desacuerdo	234	234/700=0.33
Total	700	1

Así las personas que estén en desacuerdo o totalmente en desacuerdo forman el evento S. Entonces la probabilidad de que una persona opine de esa manera es:

$$P(S) = 0.25 + 0.33 = 0.58$$

Ello significa que más de la mitad de las personas desaprueban el horario de verano.

Probabilidad subjetiva

Hemos visto dos definiciones de probabilidad, aunque también es común que usemos en nuestra vida diaria un concepto de probabilidad, que llamaremos *probabilidad subjetiva,* para pronosticar eventos futuros.

Este concepto no descansa en la repetibilidad de ningún suceso. En ese sentido, no podemos dar un valor para estimar la probabilidad. Se puede evaluar la probabilidad de futuros eventos que ocurrirán una sola vez, por ejemplo la probabilidad de que se descubra un medicamento que cure el sida en el próximo año, cuando tratamos de predecir si lloverá mañana o intentamos evaluar la reacción de otros ante nuestras acciones u opiniones. En la valoración subjetiva de la probabilidad podemos tomar en cuenta los datos experimentales de eventos pasados, pero nuestra verosimilitud adquiere un tono de subjetividad al depender de nuestra personalidad.

4.4 Relación entre eventos y leyes de probabilidad

Los juegos de azar tales como lanzar monedas o dados, las barajas (naipes), las rifas y la ruleta, entre varios más, desempeñan un papel muy importante en el cálculo de probabilidades. En esta parte se presentarán las reglas básicas de la probabilidad, las cuales estarán ilustradas mediante juegos de azar. Ello permitirá dominar y adquirir habilidad en el *cálculo de probabilidades,* así como entender algunos conceptos importantes sobre este tema. Cuando una persona domina el cálculo de probabilidades, posteriormente podrá aplicar lo que sabe a diferentes situaciones de la vida real.

El mundo de la información 3. Lanzamiento de monedas y dados

Al lanzar una moneda (de 1, 2, o 5 pesos) se tiene un espacio muestral con dos resultados posibles: águila (a) y sol (s), como tradicionalmente en México se hace referencia al lanzamiento de una moneda para decidir la suerte; a esto se le conoce como "volado". Es común en otros lados denominar al resultado de lanzar una moneda como: "cara y cruz".

Así, el espacio muestral es $M = \{a, s\}$ donde ambos tienen la misma posibilidad de ocurrir. En ese caso decimos que la moneda no está cargada. Con el lanzamiento de monedas se pueden plantear varios problemas para establecer la relación entre eventos y a partir de éstos derivar algunas leyes de probabilidad.

Preguntas sobre la naturaleza del problema

Al lanzar dos monedas se puede preguntar: ¿qué eventos se pueden obtener? ¿Qué relaciones existen entre esos eventos? ¿Cómo se pueden calcular probabilidades a partir de esas relaciones? Como actividad opcional o alternativa se puede usar la opción descrita en la figura 4.5.

En caso de un dado tenemos 6 posibilidades al lanzarlo; así, el espacio muestral es: $M = \{1, 2, 3, 4, 5, 6\}$. La finalidad de lanzar un o dos dados es similar al de las monedas. Estos serán de utilidad para ilustrar cálculos de probabilidades. Ver la figura 4.6.

Ejemplo 4.7

Una pareja no ha decidido sobre si van a ir a una fiesta o no. Sofía desea ir, pero Víctor prefiere ir al cine. Deciden dejarlo a la suerte. Cada quien lanzará una moneda, si ambas caen iguales irán a la fiesta y si son diferentes entonces saldrán al cine. ¿Cuál es la probabilidad de ir a la fiesta? ¿Qué probabilidad hay de que vayan al cine?

Solución

El espacio muestral del lanzamiento de dos monedas es: $M = \{aa, as, sa, ss\}$. El evento para que se cumpla lo que quiere Sofía es $A = \{aa, ss\}$, y para Víctor es $B = \{as, sa\}$. La probabilidad para el evento A es:

$$P(A) = P(aa) + P(ss) = \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$$

De manera análoga, la probabilidad de B es:

$$P(B) = P(as) + P(sa) = \frac{1}{4} + \frac{1}{4} = \frac{2}{4} = \frac{1}{2}$$

En ese sentido, es un acuerdo justo para ambos porque resulta que hay un 50 % de probabilidades para cada uno de obtener lo que desean.

4.4.1 Relación de eventos

En el ejemplo se han descrito el espacio muestral M y los eventos A y B. Con ellos se pueden derivar varias operaciones. A continuación se definen tres *operaciones básicas con eventos* y se introducen los símbolos correspondientes.

El evento **no** A, o *complemento* de un evento A , denotado por A^c, es el conjunto de todos los sucesos elementales que no están en A. La ocurrencia de A^c significa que A no ocurre.

En referencia al ejemplo 4.7, el evento $A = \{aa, ss\}$ que son los sucesos elementales que apoyan a que Sofía y Víctor vayan a la fiesta. El complemento de este evento es $A^c = \{as, sa\}$, éste coincide con el evento B.

El evento A o B es *la unión* de los eventos A y B, denotado por A o B, es el conjunto de todos los sucesos elementales que están en A, B, o en ambos. La ocurrencia de A o B significa que al menos uno de los dos eventos ocurre. El símbolo matemático para expresar esta relación es: A o $B = A \cup B$.

El evento A y B, es *la intersección* de los eventos A y B, denotado por A y B, es el conjunto de todos los sucesos elementales que están en A y B. La ocurrencia de A y B significa que ambos eventos ocurren. El símbolo matemático para expresar ésta relación es: A y $B = A \cap B$.

Ejemplo 4.8

Una empresa otorgó cuatro becas de manera aleatoria a un grupo de ocho estudiantes, a los cuales se les denotará con las letras a, b, c, d, e, f, g y h. Se distinguen los siguientes eventos: a los estudiantes que obtuvieron la beca, $A = \{a, c, f, g\}$; B los estudiantes que tuvieron 10 de promedio el último año, $B = \{d, f, h\}$, y C los que obtuvieron 9 de promedio el último año, $C = \{a, b, c, e, g\}$. Describir una lista de los siguientes eventos:

1. A^c
2. $A \cup A^c$
3. $B \cup C$
4. $A \cap B$
5. $B \cap C$
6. $A^c \cap B$

Solución

A^c son los estudiantes que no obtuvieron la beca, es decir,

$$A^c = \{b, d, e, h\}.$$

La unión de evento A y su complemento es:

$$A \cup A^c = \{a, b, c, d, e, f, g, h\}$$

Se puede observar que este evento coincide con el espacio muestral. Por consiguiente se tiene la siguiente relación:

$$A \cup A^c = M.$$

$$B \cup C = \{a, b, c, d, e, f, g, h\},$$

es el evento que corresponde a los estudiantes que tuvieron el promedio de 10 o de 9.

$$A \cap B = \{f\},$$

es el evento de los estudiantes que obtuvieron la beca y además tuvieron 10 de promedio.

$$B \cap C = \{\},$$

este evento no tiene ningún suceso elemental.

$$A^c \cap B = \{d, h\}$$

es el evento de los estudiantes que no obtuvieron la beca pero tuvieron 10 de promedio.

Observe que $A \cup B$ es un conjunto más grande que contiene tanto A como B, mientras que $A \cap B$ es la parte común de los conjuntos A y B. También las definiciones $A \cup B$ o $B \cup A$ representan el mismo evento. Mientras que $A \cap B$ y $B \cap A$ son ambas expresiones para A y B. Las operaciones se pueden extender a dos o más eventos.

Dos eventos A y B son mutuamente excluyentes si su intersección $A \cap B$ no tiene elementos en común, es decir, es vacía, porque los eventos mutuamente excluyentes no tienen sucesos elementales en común; ellos no pueden ocurrir simultáneamente.

Los diagramas de Venn son un auxiliar gráfico para representar un evento. En la figura 4.7 se utilizan los diagramas de Venn para representar la relación de eventos.

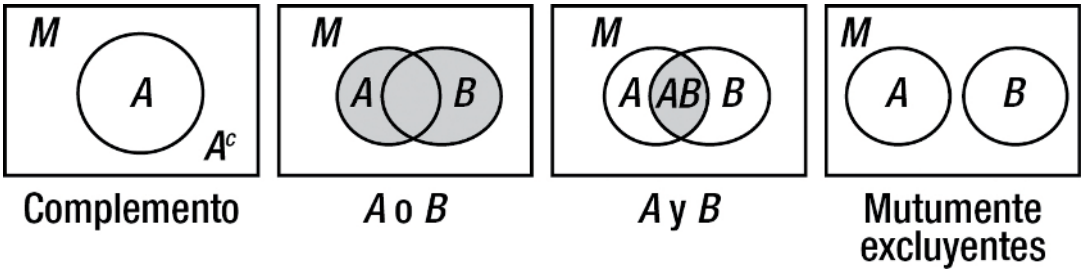


Figura 4.7 Representación gráfica de las relaciones básicas entre dos eventos.

Creadores de la probabilidad

La escuela francesa de mediados del siglo XVII contribuyó a la probabilidad tal y como se le conoce hoy en día. Entre los creadores que destacan se encuentran un noble de elevada posición, Chevalier de Méré y dos mátemáticos, Blaise Pascal y Pierre Fermat.

4.4.2 Dos leyes de probabilidad

A continuación se estudiará cómo se obtienen las probabilidades cuando las operaciones *complemento, unión, intersección* se aplican a eventos. Recuerde que $P(A)$ es la suma de las probabilidades de los sucesos elementales que están en A, y $P(M) = 1$.

Vea como $P(A^c)$ se relaciona con $P(A)$. La suma $P(A)+P(A^c)$ es la suma de las probabilidades, los sucesos elementales que están en A más la suma de las probabilidades los sucesos elementales que no están en A. Juntando estos dos eventos forman el espacio muestral M y $P(M) = 1$. En consecuencia $P(A) + P(A^c) = 1$. Esto da lugar a la primera ley de probabilidad que involucra el complemento de un evento.

Ejemplo 4.9

Una persona participó en un concurso dentro de un programa televisivo. Le plantearon cuatro preguntas a las que debía

contestar verdadero (v) o falso (f). Debido a la rapidez y el nerviosismo, la persona contestó sin pensar. Si la persona contesta 3 o más respuestas, se lleva un premio. ¿Cuál es la probabilidad de que no se lleve premio? El espacio muestral es:

$$M = \{(vvvv), (vvvf), (vvfv), (vfvv), (fvvv), (vvff), (vfvf), (vfff)\}$$
$$\{(fvfv), (ffvv), (fvvf), (vfff), (fvff), (ffvf), (fffv), (ffff)\}$$

Consideremos como evento A que la persona contesta 3 o más respuestas. Así, la lista del evento A es:

$$A = \{(vvvv), (vvvf), (vvfv), (vfvv), (fvvv)\}$$

La probabilidad de A es $P(A) = \frac{5}{16}$. Por lo tanto, la probabilidad que la persona no gane un premio se obtiene aplicando la ley del complemento, esto es:

$$P(A^c) = 1 - P(A) = 1 - \frac{5}{16} = \frac{11}{16}$$

Ley del complemento

Consideremos a A un evento con probabilidad $P(A)$. Entonces:

$$P(A^c) = 1 - P(A)$$

o bien:

$$P(A) = 1 - P(A^c)$$

4.4.3 Eventos combinados

Ahora consideramos el caso donde los eventos A y B se pueden combinar de dos maneras diferentes. La primera es A o B se pueden componer de sucesos elementales que están en A en B o en ambos. Por lo que $P(A \cup B)$ es la suma de las probabilidades asignadas a cada uno de los sucesos elementales en A y B. Por otro lado, la suma $P(A) + P(B)$ incluye todos los resultados elementales tanto de A como de B. Si uno o más resultados elementales están en ambos eventos, quiere decir que se ha contado doble. Para ajustar este doble conteo se resta $P(A \cap B)$ de $P(A) + P(B)$. Así la ley aditiva es:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Ejemplo 4.10

Tenemos que el espacio muestral al lanzar dos monedas es $M = \{aa, as, sa, ss\}$. Ahora si A es el evento "al menos un águila" y B "al menos un sol", se tienen las siguientes listas para los eventos A y B (véase la figura 4.8)

$$A = \{aa, as, sa\}, B = \{as, sa, ss\}$$

Entonces, el evento $A \cup B$:

$$A \cup B = \{aa, as, sa, ss\} = M$$

y el evento:

$$A \cap B = \{as, sa\}$$

Si dos eventos A y B son mutuamente excluyentes, su intersección no tienen sucesos elementales en común, es vacía, así $P(A \cap B) = 0$.

La frase "al menos", quiere decir que cae un "águila" o más de una, análogamente para "sol".

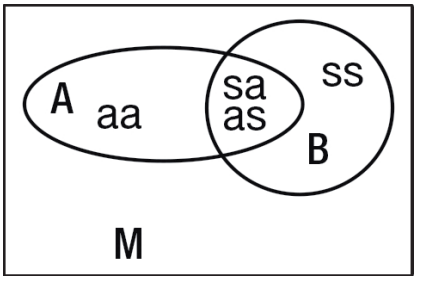


Figura 4.8 Diagrama de Venn que representa los eventos en el ejemplo 4.7.

Ley aditiva para eventos mutuamente excluyentes

Si los eventos A y B son mutuamente excluyentes, entonces

$$P(A \cup B) = P(A) + P(B)$$

Se puede observar que los eventos A y su complemento A^c son mutuamente excluyentes. Con este resultado podemos formular la ley aditiva para eventos mutuamente excluyentes.

Ejemplo 4.11

En la lista de abajo se indica la escuela y la edad de cuatro alumnos que llegaron a la final de un concurso.

Alumno	Escuela	Edad
1	Escuela 1	17
2	Escuela 1	19
3	Escuela 6	17
4	Escuela 6	17

Se seleccionaron de manera aleatoria dos alumnos y a cada uno de ellos se les otorgará el primer premio.

1. Considerar todos los resultados posibles.
2. Escribir una lista de los eventos A, B y C que se describen a continuación:

   A Los alumnos seleccionados son de la misma escuela.

   B Los alumnos seleccionados tienen la misma edad.

   C Los alumnos seleccionados son de diferente escuela.

3. Indicar si tienen elementos en común.
4. Anotar los resultados para los eventos A^c, $A \cup B$, $A \cap B$, $B \cap C$.
5. ¿Cuál es la probabilidad de $P(A^c)$, $P(A \cap B)$, $P(A \cup B)$ y $P(B \cap C)$?

Solución

1. Los posibles resultados son el par de respuestas al seleccionar un alumno: $\{1, 2, 3, 4\}$. Por ejemplo, alumno 3 escuela 6 y 17 años. A continuación se presenta la lista de estos pares, y como referencia se denotan por e_1, e_2, e_3, e_4, e_5, e_6.

$$\begin{array}{ll} e_1 = (1,2) & e_4 = (2,3) \\ e_2 = (1,3) & e_5 = (2,4) \\ e_3 = (1,4) & e_6 = (3,4) \end{array}$$

2. Así el par (1,2) corresponde a los alumnos de la escuela 1 pero de edad diferente. De manera análoga (3,4) son de la escuela 6 y de la misma edad. Los pares procedentes de la misma edad son: (1,3), (1,4) y (3,4). Los de escuelas diferentes son: (1,3), (1,4), (2,3) y (2,4). En consecuencia los eventos están formados por los resultados siguientes:

$$A = \{e_1, e_6, \},\ B = \{e_2, e_3, e_6\}\ y\ C = \{e_2, e_3, e_4, e_5\}$$

3. Los eventos $A \cap B$ tienen en común el resultado $\{e_6\}$. Los eventos $B \cap C$ cuentan en común con $\{e_3\}$. Los eventos A y C no tienen resultados en común.
4. Los resultados de A^c, A o B, A y B, B y C son:

$$\begin{array}{l} A^c = \{e_2, e_3, e_4, e_5\} \\ A \cup B = \{e_1, e_2, e_3, e_6\} \\ A \cap B = \{e_6\} \\ B \cap C = \{e_2, e_3\} \end{array}$$

5. La probabilidad para cada evento es:

$$P(A) = \frac{2}{6}, P(B) = \frac{3}{6}, P(C) = \frac{4}{6}$$

Aplicando las leyes de probabilidad se tiene lo siguiente.

La ley de complemento:

$$P(A^c) = 1 - P(A) = 1 - \frac{2}{6} = \frac{4}{6}$$

La ley de la adición:

$$P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B) = \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6}$$

Solución mediante el uso de CalEst

En el bloque Didáctica de *CalEst* está la opción de lanzar dos dados. En la figura 4.9 se suguiere hacer prácticas con esta alternativa didáctica. Ahí aparece la descripción del espacio muestral, la estimación de probabilidad empírica, el cálculo de probabilidad clásica para los eventos simples, y el casino donde se pueden generar varios cálculos de probabilidades aplicando las reglas.

Por ejemplo, ¿cuál es la probabilidad de que al lanzar dos dados ambos sean iguales, o de que la suma sea 7? En el evento A, ambos dados, marque el mismo número, $A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$. Observe el espacio muestra, figura 4.9. En el evento B los datos suman 7: $B = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$, donde $A \cup B = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6),(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$. Así: $P(A \cup B) = P(A) + P(B) = \frac{6}{36} + \frac{6}{36} = \frac{1}{3}$. Los eventos A y B son mutuamente excluyentes puesto que se cumple $P(A \cup B) = P(A) + P(B)$.

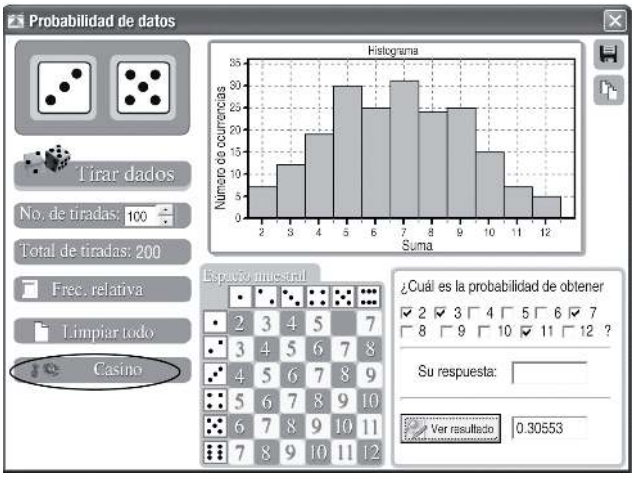


Figura 4.9 Cálculo de probabilidades, frecuentista, clásica y la acción de evaluar las ganancias en un juego al lanzar dos dados.

4.4.4 Leyes de probabilidad y tablas de contingencia

En unidades anteriores se han presentado tablas que describen los conteos o frecuencias de dos o más variables, denominadas tablas de contingencia.

Éstas nos ayudan a comprender de manera sencilla cómo determinar las probabilidades en problemas de la vida real. En la actualidad, hay una cantidad importante de estudios en los que se recurre a este tipo de tablas. En las actividades de aprendizaje y en los ejercicios de este libro se pondrán varios ejemplos encaminados en esa dirección.

La meta ahora es mostrar cómo se da la relación de una tabla de dos variables con los eventos para aplicar las leyes de probabilidad. Además, éstas se pueden aplicar a problemas de interés social y médico, por mencionar dos casos, entre muchos otros. El siguiente ejemplo presenta un caso sencillo con información del desempeño académico al finalizar el bachillerato.

Ejemplo 4.12

Del registro escolar del año 2000 en un Instituto Tecnológico, un sociólogo revisa los expedientes de esa generación y escribe el porcentaje como se describe en la tabla. Anota el promedio (por arriba de 8, por debajo de 8), el turno (matutino, vespertino) y el género (mujer M, hombre H).

	Matutino		Vespertino		Total
Promedio	M	H	M	H	0
Arriba de 8	18	10	21	10	59
Debajo de 8	12	20	7	2	41
Total	30	30	28	12	100

De esta información seleccionó un alumno al azar y definió los eventos A, B y C.

A Está en el turno vespertino.

B Tiene un promedio mayor a 8.

C Tiene un promedio mayor a 8 y es mujer.

Encontrar las probabilidades $P(A), P(B), P(B \cap C)$ y $P(A \cap B \cap C)$.

Solución

El porcentaje de alumnos del turno vespertino es 40 %, por lo tanto, la probabilidad es:

$$P(A) = \frac{40}{100} = 0.4$$

El porcentaje de alumnos que corresponde al promedio mayor a 8 es 59 %. Así:

$$P(B) = \frac{59}{100} = 0.59$$

$B \cap C$ significa que tiene promedio mayor que 8 y es mujer; de esa manera la probabilidad $P(B \cap C)$ es:

$$P(B \cap C) = \frac{39}{100} = 0.39$$

Finalmente, $A \cap B \cap C$. Ya sabemos algo sobre $B \cap C$.

Ahora incorporamos la información del evento A a la de los eventos $B \cap C$. Esto indica que el alumno seleccionado está en el turno vespertino, con promedio mayor que 8 y es mujer.

$$P(A \cap B \cap C) = \frac{21}{100} = 0.21$$

Interpretando el problema se observa que la probabilidad de un rendimiento académico arriba de 8 es mayor en referencia a los que están por debajo. La probabilidad del promedio arriba de 8 en ambos turnos es casi igual. Es más probable que

una mujer, con respecto a un hombre, tenga un promedio arriba de 8.

Independencia y probabilidad condicional

Ahora consideramos el caso en que dos eventos pueden ocurrir de manera simultánea, y si uno sucede pero que no afecta la probabilidad de ocurrencia del otro. Es decir, la probabilidad de que ocurra el evento A no afecta la probabilidad de que suceda B. En este caso, se dice que los eventos A y B son *independientes*.

Por ejemplo, si se lanza una moneda de 1 peso y otra de 2 pesos, el hecho de que el peso caiga águila no afecta a lo que sucede al lanzar la moneda de 2 pesos. Sus resultados son independientes.

Eventos independientes

Los eventos A y B son independientes si el hecho de que uno ocurra no afecta la probabilidad de que ocurra el otro.

La probabilidad de A como la de B es $\frac{1}{2}$ de que caiga "águila". En ese sentido la probabilidad de B permanece como $\frac{1}{2}$ sin importar lo que pase al lanzar la otra moneda. El espacio muestral $M = \{aa, as, sa, ss\}$ y $\frac{1}{4}$ es la probabilidad de que ambas monedas caigan "águila"$A \cap B : \{aa\}$. Si multiplicáramos la probabilidad A por la de B, es decir $P(A)\ P(B)$, se tiene que es $\frac{1}{4}$.

Ley de multiplicación para dos eventos independientes

Consideremos los eventos A y B independientes. Entonces,

$$P(A \cap B) = P(A)P(B)$$

Ejemplo 4.13

Desde un punto de vista administrativo y económico, a menudo se considera tener mejoras en varios campos, en particular en medicina. Se desarrolla una vacuna para curar una enfermedad. La vacuna se prueba en tres animales de laboratorio. Después de un cierto tiempo de aplicada se espera que el animal sobreviva (s) o muera (m). En este caso el espacio muestral es:

$$M = \{(sss), (ssm), (sms), (smm), (mss), (msm), (mms), (mmm)\}$$

El biólogo piensa que esta vacuna da un 70 % de sobrevivir, con lo que se mejora la expectativa anterior. Con esa información

hay que asignar la probabilidad a cada resultado elemental del espacio muestral M.

Solución

Consideremos el primer resultado de M, es decir, $\{sss\}$. Entonces, la probabilidad de $\{sss\}$ considerando que 70 % sobreviva es:

$$\begin{aligned} P\{sss\} &= P(\text{1ro sobrevive})P(\text{2o sobrevive})P(\text{3er sobrevive}) \\ &= (0.7)(0.7)(0.7) = 0.343. \end{aligned}$$

Entonces la probabilidad de que sobrevivan los tres animales es 0.343. Asimismo, se calculan los otros resultados elementales. Por ejemplo,

$$P(msm) = (0.3)(0.7)(0.3) = 0.063.$$

En la tabla 4.1 se reproducen los cálculos para los 8 resultados del espacio muestral.

Tabla 4.1 Probabilidad de sobrevivir después de aplicar la vacuna

Resultado	Probabilidad
{sss}	0.343
{ssm}	0.147
{sms}	0.147
{smm}	0.063
{mss}	0.147
{msm}	0.063
{mms}	0.063
{mmm}	0.027
Total	1.000

Ejemplo 4.14

A partir de la tabla 4.1, a continuación se definen cuatro eventos.

- Evento A: Exactamente dos sobreviven.
- Evento B: Sólo uno sobrevive.
- Evento C: Los tres sobreviven.
- Evento D: Al menos uno de los tres sobrevivió.

Si 70 % de los animales sobrevive, debe determinarse las probabilidades de que ocurran los cuatro eventos.

Solución

El evento se forma mediante los resultados siguientes. $A = \{(ssm), (sms), (mss)\}$. Aplicando la tabla 4.1, la probabilidad de A se tiene aplicando la ley aditiva para eventos mutuamente excluyentes.

$$P(A) = 0.147 + 0.147 + 0.147 = 0.441.$$

De manera análoga, se calcula la probabilidad para B; éste consiste de $B = \{(smm), (msm), (mms)\}$. Entonces,

$$P(B) = 0.063 + 0.063 + 0.063 = 0.189.$$

Para el evento $C = \{(sss)\}$, se tiene $P(C) = 0.343$.

Finalmente, al evento D le falta un resultado elemental para completar el espacio muestral.

$$D = \{(sss), (ssm), (sms), (smm), (mss), (msm), (mms)\}.\text{Así,}$$

$$P(D) = 1 - 0.027 = 0.973.$$

Diferencia entre eventos mutuamente excluyentes e independientes

- Si los eventos A y B son mutuamente excluyentes, entonces

$$P(A \cap B) = 0$$

- Si los eventos A y B son independientes, entonces

$$P(A \cap B) = P(A)P(B)$$

El mundo de la información 4. Misión con robots

Dos robots son enviados independientemente a recolectar objetos del RMS Titanic a más de cuatro mil metros bajo el nivel del mar. La probabilidad de éxito en su misión del primer robot es de 0.9 mientras que la probabilidad de éxito del segundo robot es 0.8.

Preguntas sobre la naturaleza del problema

Es razonable asumir que el éxito o fracaso del primer robot es independiente del éxito o fracaso del segundo robot. Con esta consideración, ¿cuál es la probabilidad de que al menos un robot realice una misión exitosa?

Información: Se definen los dos eventos:

A={El primer robot tiene éxito}, B={El segundo robot tiene éxito}

El evento de interés es entonces A o B. Claramente, A y B pueden ocurrir al mismo tiempo porque los dos eventos no son mutuamente excluyentes. Por medio de la ley de adición:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Por el texto del problema, sabemos que $P(A) = 0.9$ y $P(B) = 0.8$. Dado que A y B son independientes, el valor de $P(A \cap B)$ es igual al producto de $P(A)$ y $P(B)$:

$$P(A \cap B) = P(A)P(B) = (0.9)(0.8) = 0.72$$

Por lo tanto,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.9 + 0.8 - 0.72 = 0.98$$

La probabilidad de éxito es más grande si se envían a dos robots en vez de a uno solo.

Otra forma de resolver este problema es considerar el complemento del evento. Sea $C = A \cup B$. Entonces C^c es el evento que $A \cup B$ no ocurra. Es el caso de que [el primer robot fracase en su misión] y que [el segundo robot fracase en su misión]. Esto es,

$$C^c = A^c \cap B^c$$

Dado que los robots trabajan independientemente, A^c y B^c son eventos independientes. Por la definición de independencia,

$$P(A^c \cap B^c) = P(A^c)P(B^c)$$

También, por la ley de probabilidad del complemento:

$$P(A^c) = 1 - P(A) = 1 - 0.9 = 0.1$$
$$P(B^c) = 1 - P(B) = 1 - 0.8 = 0.2$$

entonces $P(C^c) = P(A^c \cap B^c) = P(A^c)P(B^c) = (0.1)(0.2) = 0.02$. Por último, por medio de la ley de probabilidad de complemento,

$$P(A \cup B) = P(C) = 1 - P(C^c) = 1 - 0.02 = 0.98$$

lo cual es idéntico a la solución por el otro método.

Este ejemplo muestra cómo las leyes de probabilidad pueden ser utilizadas en diversas maneras para resolver un problema. ¿Qué solución prefiere? ¿Qué pasaría si en vez de dos robots, enviamos a tres robots a recolectar independientemente y estamos interesados en encontrar la probabilidad de que al menos un robot tenga éxito? Para este caso, el segundo método es más fácil.

Ejemplo 4.15

Un estudiante ha organizado una fiesta en su casa y ha invitado a sus compañeros. La invitación incluye un mapa de cómo llegar a la casa. Rumbo a la fiesta, dos estudiantes, quienes fueron cada uno por su lado, han perdido el mapa. En el camino, cada uno por separado encontró una bifurcación, donde un sendero conduce a Valle del Ciprés (correcto) y el otro a Praderas del Ciprés (equivocado). Los amigos seleccionan al azar el camino que los llevaría a la casa donde era la fiesta.

1. Encuentre la composición de los eventos: A ={El primer estudiante escoge el camino correcto}, B ={El segundo estudiante escoge el camino equivocado}
2. Encuentre las probabilidades de A, B, y $A \cap B$.
3. Verifique la siguiente relación $P(A \cap B) = P(A)P(B)$ en este ejemplo.

Solución

1. El espacio muestral es $S = \{cc, ce, ec, ee\}$ y los eventos son: $A = \{\text{cc, ce}\}$ $B = \{\text{ce, ee}\}$, c = correcto y e = equivocado.
2. El evento $A \cap B = \{ce\}$ contiene el resultado común para A y B. Dado que la probabilidad de escoger cualquiera de los dos caminos es igual, los cuatro resultados en S son igualmente probables. Así:

$$P(A) = \frac{2}{4} = 0.5 \quad P(B) = \frac{2}{4} = 0.5 \quad P(A \cap B) = \frac{1}{4} = 0.25$$

3. En este caso $P(A)\ P(B) = (0.5)(0.5) = 0.25$, lo cual es igual a $P(A \cap B)$. Es sólo una coincidencia que $P(A)P(B) = P(A \cap B)$.

4.4.5 Probabilidad condicional

Si los eventos A y B se relacionan, la información que nos proporciona cuando B ha ocurrido es importante para mejorar la evaluación de la probabilidad de A. La probabilidad corregida de A, esto es, cuando se sabe que B ha ocurrido, se llama *probabilidad condicional* de A dado B y se denota por $P(A \mid B)$.

La *probabilidad condicional* de un evento A dado que un evento B ha ocurrido se establece dividiendo la probabilidad de que A y B ocurren por la probabilidad de que B ocurrió, esto es,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

se debe verificar que $P(B)$ es diferente de cero.

También se observa que los eventos son independientes si:

$$P(A \mid B) = P(A)$$

El mundo de la información 5. Influencia de un comercial

Vivimos en una época en que estamos sometidos a la presencia de la publicidad para comprar determinado producto. Con el interés de conocer el impacto que produce un comercial en la compra de un producto, una agencia tomó una muestra de 550 personas. Se les formuló dos preguntas: 1) ¿compró el producto porque vio el comercial? y 2) ¿no compró el producto porque no vio el comercial?

Preguntas sobre la naturaleza del problema

Aquí el objetivo consiste en evaluar la influencia de un comercial en la compra del nuevo producto. De modo que el motivo principal es conocer las posibilidades de compra, expresados por la aplicación de las leyes de probabilidad. La idea es encontrar si ver el comercial (evento B) afecta la probabilidad de que una persona compre el producto (evento A). Es necesario conocer $P(A)$ *y* $P(A \mid B)$. Los datos de la muestra aleatoria se describen abajo y se indica la proporción entre paréntesis.

	Vió el comercial	No vió el comercial	Total
Compraron	198(0.36)	55(0.10)	253(0.46)
No Compraron	110(0.20)	187(0.34)	297(0.54)
Total	308(0.56)	242(0.44)	550(1.00)

Solución

De la tabla se obtiene: $P(A) = 0.46$, lo que equivale a decir que 46 % de las personas compraron el producto.

La probabilidad de las personas que compraron el producto dado que vieron el comercial es:

$$P(A \mid B) = \frac{P(A \text{ y } B)}{P(B)} = \frac{0.36}{0.56} = 0.643$$

Esta probabilidad es mayor que $P(A)$. Este resultado indica que ver el comercial influye en la compra.

Ejemplo 4.16

El profesor encargado del departamento de deportes de un bachillerato señala que 35 % de los estudiantes practica futbol, 15 % practica atletismo y 6 % practica ambos deportes. Si se escoge al azar a un estudiante que ejercita futbol, ¿cuál es la probabilidad de que practique atletismo?

Solución

Denotamos con F el evento de que un estudiante practique futbol y con A que practique atletismo. Así, las probabilidades de F, de A y de $F \cap A$.

$$P(F) = 0.35 \quad P(A) = 0.15 \quad P(F \cap A) = 0.06$$

Entonces, la probabilidad de que practique atletismo dado que juega futbol es:

$$P(A \mid F) = \frac{P(F \cap A)}{P(F)} = \frac{0.06}{0.35}$$

Detalle operativo en las operaciones de probabilidad

Relación de la tabla con las probabilidades. Ahora se muestra la tabla que relaciona la información a la operación de eventos y la notación que se ha venido utilizando:

Evento A: comprar	Evento A^c: no comprar
Evento B: ver el comercial	Evento B^c: no ver el comercial

	B	B^c	Total
A	$P(A \cap B)$	$P(A \cap B^c)$	$P(A)$
A^c	$P(A^c \cap B)$	$P(A^c \cap B^c)$	$P(A^c)$
Total	$P(B)$	$P(B^c)$	1

La expresión $P(A^c \cap B^c)$ se interpreta como la probabilidad de las personas que no compraron y no vieron el comercial. ¿Cómo se interpretan $P(A^c \cap B)$ *y* $P(A \cap B^c)$?

Ejemplo 4.17

Un nutriólogo clasifica a un grupo de jóvenes con respecto a su peso y su actividad deportiva. La proporción en diferentes categorías aparece en la tabla de abajo.

	Sobrepeso	Peso normal	Bajo peso	Total
Hace deporte	0.04	0.08	0.18	0.30
No hace deporte	0.21	0.44	0.05	0.70
Total	0.25	0.52	0.23	1.00

¿Cuál es la probabilidad de que un joven seleccionado al azar practique deporte? Si un joven seleccionado al azar de este grupo padece sobrepeso, ¿cuál es la probabilidad de que también practique deporte? ¿Los eventos A hacer deporte y B sobrepeso son independientes?

Solución

Como 30 % de los jóvenes practica deporte y el joven es seleccionado al azar de este grupo, se concluye que $P(A) = 0.3$. Cuando nos referimos al sobrepeso nos ubicamos en la primera columna. Ésta clasifica a los alumnos en el subgrupo de sobrepeso; la proporción de los que practican deporte es $0.04/0.25 = 0.16$.

Por lo tanto, dada la información de que un joven está en este subgrupo, la probabilidad de que practique deporte es:

$$P(A \mid B) = \frac{0.04}{0.25} = 0.16$$

El cálculo de esta expresión se pudo haber realizado, aplicando la expresión de la probabilidad condicional. Observe que $P(A \cap B) = 0.04$ y $P(B) = 0.25$. Entonces,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.04}{0.25} = 0.16$$

En otras palabras, $P(A \mid B)$ es la proporción de la población que tiene la característica A entre aquella que tiene la característica B.

De los cálculos anteriores se obtuvo $P(A) = 0.3$ y $P(A \mid B)$. Aplicando la regla para la independencia, se observa que

$$0.3 = P(A) \neq P(A \mid B) = 0.16$$

Se concluye, entonces, que los eventos A y B no son independientes.

Ley de multiplicación

Consideremos los A y B con probabilidad diferente de cero. Entonces,

$$P(AyB) = P(A \mid B)P(B)$$

Ejemplo 4.18

En una urna hay 25 bolas, de las cuales 20 son blancas y 5 están marcadas con un premio. Se extraen dos bolas al azar de la urna y se observa su característica. Calcular la probabilidad de que:

- Ambas estén premiadas.

- Una esté premiada y la otra sea blanca.

Solución extracción sin reemplazo

Primera respuesta. Usamos la letra A para indicar que son blancas y B para las premiadas. Un subíndice identifica el orden en que se extrajo la bola. Así B_1 y A_2 se refieren a que primero salió una bola premiada y enseguida una bola blanca.

Nuestro problema es calcular $P(B_1 \cap B_2)$. En ese sentido se puede aplicar la ley de multiplicación. Es decir,

$$P(B_1 \cap B_2) = P(B_2 \mid B_1)P(B_1)$$

Podemos emplear el siguiente razonamiento: si para calcular $P(B_1)$ se necesita escoger una entre las 20 bolas blancas y 5 bolas premiadas, es claro que

$$P(B_1) = \frac{5}{25}$$

El siguiente paso es calcular $P(B_2 \mid B_1)$. Dado que se ha seleccionado una bola premiada, quedan 20 bolas blancas y 4 premiadas para efectuar la segunda selección; por lo tanto, la probabilidad condicional de B_2 dado B_1 es

$$P(B_2 \mid B_1) = \frac{4}{24}$$

Multiplicando estas dos probabilidades se obtiene

$$P(\text{ambas premiadas}) = P(B_1 \mid B_2) = \frac{5}{25} \times \frac{4}{24} = \frac{1}{30} = 0.033$$

Segunda respuesta. Indica que el evento de extraer 1 bola premiada en las dos posibilidades es la unión de los eventos excluyentes $(B_1 \mid A_2)$ y $(A_1 \mid B_2)$. La probabilidad de cada uno de éstos se calcula aplicando la ley de la multiplicación, de manera similar a la respuesta anterior.

$$P(B_1 \cap A_2) = P(B_1)P(A_2 \mid B_1) = \frac{5}{25} \times \frac{20}{24} = \frac{1}{6}$$

$$P(A_1 \cap B_2) = P(A_1)P(B_2 \mid A_1) = \frac{20}{25} \times \frac{5}{24} = \frac{1}{6}$$

Así la probabilidad que se plantea es

$$P(B_1 \cap A_2) + P(A_1 \cap B_2) = \frac{2}{6} = 0.333$$

Complemento técnico: Existen dos maneras de extraer objetos para tener una muestra de un conjunto dado de elementos. A dicho procedimiento se le conoce como muestreo de una población. *Muestreo con reemplazo* significa que el objeto que se extrajo al azar se coloca de nuevo en el conjunto dado, se vuelve a mezclar y se procede a obtener al azar el siguiente objeto. En cambio, el *muestreo sin reemplazo* indica que el objeto que se extrajo se deja aparte.

Solución mediante muestreo con reemplazo

En el contexto del ejemplo 18 consideremos la urna con las 25 bolas. Una de ellas se elige al azar, y enseguida se regresa a la urna. Se revuelve la urna y se extrae otra bola. Este procedimiento se conoce como muestreo con reemplazo.

La probabilidad de sacar una bola premiada en la primera extracción es $P(B) = 5/25$. Al regresarla a la urna y volver a tomar otra bola, la probabilidad de que esta sea una bola premiada es $P(B) = 5/25$. Los resultados para las dos extracciones

son independientes y, por lo tanto, se puede aplicar la propiedad de independencia para calcular la probabilidad de que ambas extracciones estén premiadas.

$$P(B_1B_2) = P(B_1)P(B_2) = \frac{5}{25} \times \frac{5}{25} = 0.04$$

Notas

- En este caso no se tuvo que trabajar con la probabilidad condicional.
- Este procedimiento se puede extender para calcular la probabilidad de cualquier número de extracciones que sean independientes. Por ejemplo, consideremos que sacamos 3 bolas con reemplazo, donde las primeras 2 sean blancas y la tercera esté premiada. Entonces,

$$P(A_1A_2B) = P(A_1)P(A_2)P(B) = \frac{20}{25} \times \frac{20}{25} \times \frac{5}{25} = 0.128$$

Confiabilidad

La confiabilidad es una denominación alternativa de probabilidad, y es un tema que hoy día tiene mucha aplicación en la industria, en el desarrollo de nuevos productos, en medicina, por mencionar algunos campos de aplicación. Como es un área con oportunidades de trabajo en muchas profesiones, citamos algunos ejemplos:

- Estudiar la fatiga del acero en la industria siderúrgica.
- Evaluar la degradación de un sistema de refrigeración en la manufactura de aparatos electrodomésticos.
- Estudiar el tiempo de falla de componentes electrónicos.
- Determinar la vida de anaquel (fecha de caducidad) de una medicina o de un alimento.
- Conocer el periodo de vida de una persona enferma que es sometida a un tratamiento médico.

Ejemplo 4.19

Un sistema mecánico consta de dos componentes mecánicos que funcionan de manera independiente. Después de un periodo largo de prueba, se estableció que la componente 1 tiene una confiabilidad de 0.94 y la componente 2 tiene una confiabilidad de 0.98. Si el sistema funciona sólo si ambas componentes funcionan, ¿cuál es la confiabilidad del sistema?

Consideremos los eventos:

A_1: Funcionamiento de la componente 1.

A_2: Funcionamiento de la componente 2.

S : Funcionamiento del sistema.

El evento S es igual A_1 y A_2. Las componentes 1 y 2 operan de manera independiente. En consecuencia, se puede aplicar la ley de multiplicación, es decir:

$$P(S) = P(A_1 \cap A_2) = P(A_1)P(A_2) = (0.94)(0.98) = 0.92$$

De modo que el sistema tiene una confiabilidad de 0.92.

Confiabilidad

Siguiendo la línea del ejemplo 4.19, sobre la confiabilidad, existen muchos problemas en la vida real a los que se les puede estudiar de manera similar. Por ejemplo, el número de autos que transitan por la carretera durante intervalos de tiempo establecido, estudiar la probabilidad del tiempo de vida en Latinoamérica, los resultados de experimentos médicos para determinar el efecto de nuevas medicinas o evaluar la calidad de los productos industriales elaborados en una fábrica.

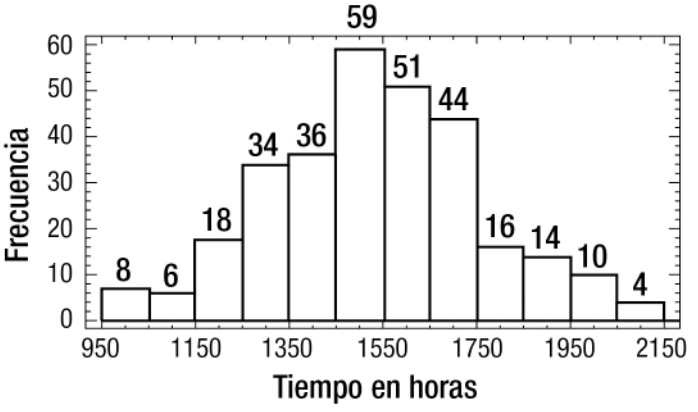


Figura 4.10 Histograma del tiempo de vida de una muestra de focos.

Ejemplo 4.20

El tiempo de vida (en horas) de los focos de luz también es un tema relacionado con la confiabilidad . A continuación se presenta el histograma que describe la distribución del tiempo de vida de una muestra de 300 focos, ver figura 4.10, donde el ancho de clase es de 100. En cada barra del histograma se indica la frecuencia.

- ¿Cuál es la probabilidad de que un foco dure más de 1750 horas?
- ¿Cuál es la probabilidad de que un foco que ya ha durado 1650 horas dure 1750 horas?

Solución

Denotemos con A el evento "dura más de 1750 horas". Sumamos las probabilidades de que un foco dure entre las clases de tiempo de vida de 1750 a 2150. Así,

$$P(A) = 0.053 + 0.047 + 0.033 + 0.0133 = 0.146$$

El evento B lo definimos como "dura más de 1650 horas". Siguiendo el proceso anterior $P(B)$ es:

$$P(B) = 0.147 + 0.053 + 0.047 + 0.033 + 0.0133 = 0.293$$

A y B tienen en común las cuatro últimas clases, por lo tanto, $P(A \cap B) = 0.146$. Finalmente, la probabilidad de que un foco dure 1750 horas dado que ya duró 1650, así $P(A \mid B)$ es:

$$P(A \mid B) = \frac{P(A \text{ y } B)}{P(B)} = \frac{0.146}{0.293} = 0.498$$

4.5 Temas selectos: Fórmula de Bayes y diagrama de árbol

Una parte importante en el cálculo de probabilidades es cuando se tiene una nueva información. En particular, para algún evento se puede iniciar con la estimación de una probabilidad inicial, también referida como probabilidad a priori. En situaciones tales como una muestra o la prueba de un producto, se tiene una información adicional sobre los eventos. A partir de esta nueva información se actualiza la probabilidad a priori mediante el cálculo de probabilidades anteriores. El teorema propuesto por Bayes permite realizar el cálculo de estas probabilidades. Aquí se presentará como la *fórmula de Bayes.* En esta parte se verán dos situaciones que son apropiadas para el cálculo de probabilidades. Estas son la fórmula de Bayes y el diagrama de árbol. Descripción del procedimiento de Bayes: Probabilidad a priori $\longrightarrow$ Nueva información $\longrightarrow$ Aplicación de la fórmula de Bayes $\longrightarrow$ Probabilidad. El cálculo de esta última probabilidad se le conoce como probabilidad a posteriori.

Ejemplo 4.21

Una caja contiene dos bolas rojas. Una segunda caja de idéntica apariencia contiene una bola roja y una blanca. La estrategia es la siguiente y es importante fijarse bien.

Procedimiento: se selecciona una caja al azar y de esa caja se extrae de manera aleatoria una bola. ¿Cuál es la probabilidad de que la primera caja sea la seleccionada, si se ha sacado de ahí una bola roja?

Solución

El ejemplo se plantea para usar las experiencias sobre las leyes de probabilidad. Luego, ilustra la aplicación de la fórmula de Bayes. Definamos los eventos A y A^c por

A : Seleccionar la primera caja.

A^c: Seleccionar la segunda caja.

Los eventos B y B^c por

B: Extraer una bola roja.

B^c: Extraer una bola blanca.

El problema consiste en calcular la probabilidad condicional:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(A \cap B)$ lo calculamos usando la expresión:

$$P(A \cap B) = P(A)P(B \mid A) = \frac{1}{2} \times 1 = \frac{1}{2}$$

Observe que la probabilidad de sacar una bola roja dado que se ha seleccionado la caja 1.

Para calcular la probabilidad de seleccionar una bola roja, debemos considerar ambas cajas. Caja 1: A y B, Caja 2: A^c y B. Esto es, calcular la suma de probabilidades de eventos ajenos: $P(B) = P(A \cap B) + P(A^c \cap B)$, pero de aquí no conocemos $P(A^c \cap B)$; por lo tanto:

$$P(A^c \cap B) = P(A^c)P(B \mid A^c) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Entonces,

$$P(B) = P(A \cap B) + P(A^c \cap B) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

Ahora sustituyendo en $P(A \mid B)$ tenemos

$$P(A \mid B) = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

Fórmula de Bayes

El mundo de la información 6. Emisión de contaminantes

En cierto estado de un país, los automóviles se deben verificar para controlar la emisión de contaminantes. Se encontró con que 23 % de los autos verificados emitían una cantidad excesiva de contaminantes. El 98 % de éstos fallaron cuando se probaron, pero el 17 % de los autos no contaminantes también fallaron.

Preguntas sobre la naturaleza del problema

¿Cuál es la probabilidad de que un automóvil de los que no pasan la prueba emita excesiva cantidad de contaminantes?

A es el evento de que un auto no pasó la prueba y el evento B es el automóvil que emite excesiva cantidad de contaminantes. Se expresan lo porcentajes dados en términos de probabilidad, es decir:

$$P(B) = 0.23, P(A \mid B) = 0.98 \; y \; P(A \mid B^c) = 0.17$$

Para dar respuesta al problema se requiere calcular la siguiente probabilidad: $P(B \mid A)$ a continuación se obtiene esta expresión en términos de las probabilidades de A y B, y de la condicional de A dado B.

La probabilidad condicional de A dado B se puede expresar como: $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$, y así: $P(A \mid B)\ P(B) = P(A \cap B)$. Ahora, la probabilidad condicional de B dado A se puede expresar por $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$, entonces, $P(B \mid A)\ P(A) = P(A \cap B)$, por lo tanto: $P(B \mid A)P(A) = P(A \mid B)P(B)$, y finalmente: $P(B \mid A) = \frac{P(A|B)\ P(B)}{P(A)}$.

En la figura 4.11 se muestra el diagrama de árbol para obtener primero la probabilidad de A. Ésta se obtiene sumando las probabilidades siguiendo la rama de arriba y luego la de abajo, es decir, $P(A) = 0.225 + 0.131 = 0.356$ se sustituye en la fórmula para $P(B \mid A)$; así:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)} = \frac{(0.98)(0.23)}{0.356} = 0.633$$

Esta es la probabilidad de que un auto que no pasa la prueba emite una cantidad excesiva de contaminantes.

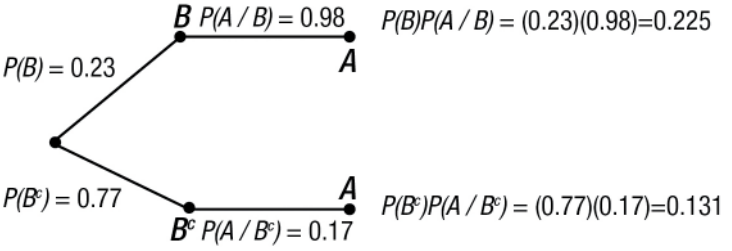


Figura 4.11 Diagrama de árbol para la emisión de contaminantes.

Fórmula de Bayes para dos eventos

$$P(A_1 \mid B) = \frac{P(A_1)P(B \mid A_1)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.1}$$

$$P(A_2 \mid B) = \frac{P(A_2)P(B \mid A_2)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2)} \tag{4.2}$$

Ejemplo 4.22

Considerando el ejemplo 21 se aplicará la conocida fórmula de Bayes.

Solución

Con la fórmula de Bayes en el ejemplo 21, los eventos H_1 y H_2 son los correspondientes a los de las cajas 1 y 2, respectivamente. En la figura 4.12 se muestra en un diagrama de Venn las características del ejemplo y el planteamiento de Bayes. Definimos con A el evento de tener una bola roja. Seleccionar una caja al azar es

$$P(H_1) = P(H_2) = \frac{1}{2}$$

Además la probabilidad de seleccionar una bola roja de la caja 1 $P(A \mid H_1) = 1$ y la roja de la caja 2 $P(A \mid H_2) = \frac{1}{2}$. Entonces la fórmula de Bayes:

$$P(H_1 \mid A) = \frac{P(H_1)P(A \mid H_1)}{P(H_1)P(A \mid H_1) + P(H_2)P(A \mid H_2)}$$

Sustituyendo queda como sigue:

$$P(H_1 \mid A) = \frac{\frac{1}{2} \times 1}{\frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2}} = \frac{2}{3}$$

Extensión de la fórmula de Bayes al caso de n eventos $A_1, A_2, ..., A_n$ mutuamente exclusivos, la unión de estos eventos está en el espacio muestral completo. El cálculo de la fórmula de Bayes para la probabilidad a posteriori $P(A_i \mid B)$ es:

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + ... + P(A_n)P(B \mid A_n)} \tag{4.3}$$

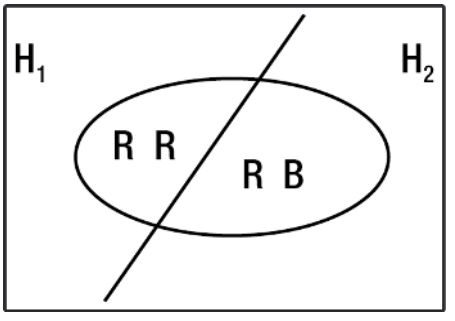


Figura 4.12 Diagrama de Venn para caracterizar la fórmula de Bayes.

4.5.1 Probabilidad con diagrama de árbol

El mundo de la información 7. Opinión sobre el aborto

En una escuela, un grupo de jóvenes ven un documental sobre el aborto en una de sus clases. Al final de la proyección, la maestra realiza una encuesta para saber cuál es la opinión de ellos sobre el aborto. Cuando les pregunta: ¿apoyarías en alguna circunstancia el aborto?, esta pregunta genera dos eventos, A: posición (en contra, a favor) con respecto al aborto y B: género (mujer, hombre).

Preguntas sobre la naturaleza del problema

Como se sabe, el aborto es un problema real y muchas personas tienen dudas, sobre todo cuando deben enfrentar una situación así. Muchas mujeres se arriesgan a practicarse uno sin contar con la información suficiente y suelen practicárselo en la clandestinidad, pues está penado por las leyes. Por otro lado, hay quienes se oponen totalmente y otros que lo apoyan cuando la mujer es víctima de violación o hay alguna consecuencia fatal para la madre o para el producto (algún defecto).

¿Cuál es la posición de los jóvenes ante el problema del aborto? Si seleccionamos una persona al azar. ¿Cuál es la probabilidad de que una mujer esté en contra del aborto? ¿Cuál es la probabilidad de que un hombre esté en contra del aborto? ¿Cuál es la probabilidad de que sea mujer y esté a favor del aborto?

En la siguiente tabla se muestran los datos, en porcentaje, sobre la opinión de los jóvenes.[1]

Aborto	Mujeres	Hombres	Total
En contra	38	41	79
A favor	9	12	21
Total	47	53	100

[1] Para conocer más sobre el tema, se puede consultar http://www.crlp.org/esp_pub_fac_saludpub.html

Se utiliza el diagrama de árbol para mostrar las probabilidades de cada resultado. Esta situación se describe en la figura 4.13. En la primera rama aparece la clasificación por género y sus probabilidades suman 1.

De las ramas primarias se desprende un conjunto de ramas secundarias, las cuales están clasificadas para los elementos de los eventos en contra y a favor del aborto.

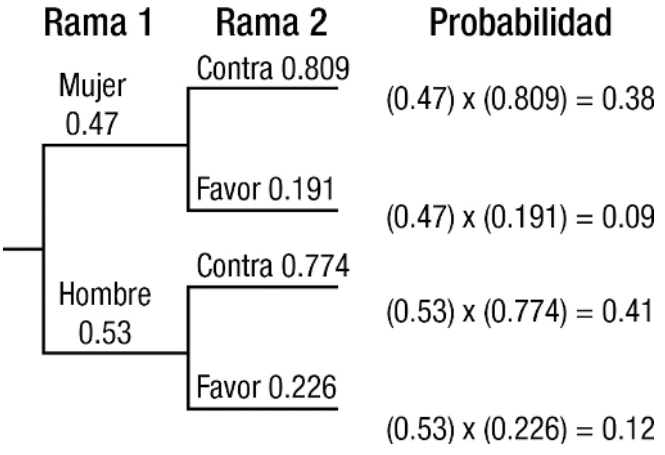


Figura 4.13 Diagrama de árbol para el problema posición frente al aborto.

Cada una de éstas corresponde a la probabilidad condicional

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Por ejemplo, la probabilidad de que esté en contra del aborto dado que es mujer es:

$$P(contra \mid mujer) = \frac{P(\text{contra y mujer})}{P(\text{mujer})} = \frac{0.38}{0.47} = 0.809$$

Del mismo modo se calculan las demás probabilidades de la segunda rama.

Complemento técnico: Cada rama determina que se juntan dos resultados elementales de cada evento; por ejemplo, una primera rama describe los eventos, ser mujer y estar en contra del aborto. Entonces, en cada caso se multiplican las probabilidades y su resultado es la probabilidad de la rama. Esto es:

$$P(\text{mujer y en contra}) = P(\text{mujer})P(\text{en contra}|\text{mujer}) = (0.47)(0.809) = 0.38$$

Otro caso considera los eventos es hombre y está a favor del aborto.

$$P(\text{hombre y a favor}) = P(\text{hombre})P(\text{a favor}|\text{hombre}) = (0.53)(0.226) = 0.12$$

La suma de las probabilidades de la rama es 1.

Interpretación: Se puede observar en el diagrama de árbol que hay dos probabilidades en la rama secundaria, que tienen un valor "cercano a uno". Eso nos indica que entre mujeres y hombres existe una alta probabilidad de estar en contra del aborto.

El diagrama de árbol para un caso general

El diagrama de árbol es una estrategia conveniente para organizar la información de probabilidad condicional. La probabilidad del evento en la rama 1 representa la información de la probabilidad incondicional. Para introducir esta idea,

consideremos los eventos B_1 y A_1. Enseguida la información de la probabilidad en la rama 2 depende (es decir, está condicionada a los eventos B_2 y A_2) de lo que sucedió en la rama 1. Entonces, el segundo conjunto de ramas se expresa en términos de la probabilidad condicional. Supongamos que suceda el evento B_2 dado que ocurrió el evento $B_1, P(B_2 \mid B_1)$. Así, para los demás casos se diría $P(A_2 \mid B_1)$. La figura 4.14 ayuda a comprender en forma total esta presentación.

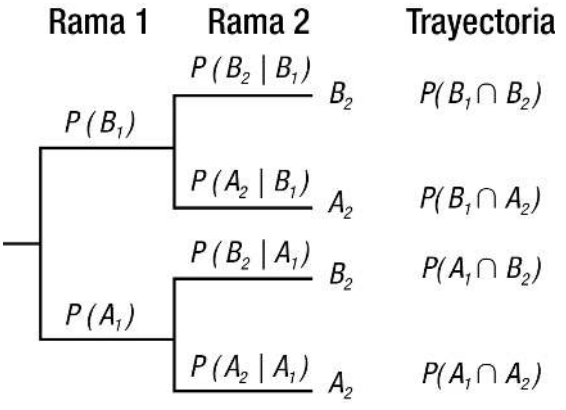


Figura 4.14 Diagrama de árbol considerando dos ramas y cuatro eventos.

4.6 Técnicas de conteo

Un reto al que se han enfrentado muchas generaciones es encontrar el número de formas en las que un conjunto de objetos se puede ordenar. Por ejemplo, ¿cuántas placas se pueden obtener tomando tres letras y cuatro números? ¿De cuántas maneras se pueden ordenar 12 personas frente a la taquilla de un cine? En este apartado se consideran los principios generales que permiten encontrar el número de arreglos posibles de un conjunto de objetos.

La solución de muchos problemas depende de la enumeración de todas las posibilidades de un suceso. Por ello, el simple hecho de contar es importante para poder determinar el número de elementos en un espacio muestral. Con ejemplos se mostrarán en este apartado tres técnicas de conteo.

Regla de multiplicación

Ejemplo 4.23

Consideremos un equipo de trabajo integrado por cinco personas:

$$E = \{\text{Sarah, Dolores, Beatriz, Roberto, Raúl}\}$$

Formemos un comité de supervisores que conste de un hombre y de una mujer elegidos entre el equipo de trabajo. ¿Cuántos comités se pueden formar?

Solución

El diagrama de árbol es de nuevo una herramienta apropiada para contestar a la pregunta (figura 4.15).

Existen dos posibilidades para elegir un hombre y hay tres posibilidades para escoger una mujer. Por consiguiente, en el diagrama de la figura 4.15 se observa que pueden formarse seis comités, es decir:

Roberto-Sarah	Raúl-Sarah
Roberto-Beatriz	Raúl-Beatriz
Roberto-Dolores	Raúl-Dolores

Considérese el siguiente ejercicio: describa el diagrama de árbol suponiendo ahora que primero se elige a una mujer. ¿Cuántos comités se pueden integrar?

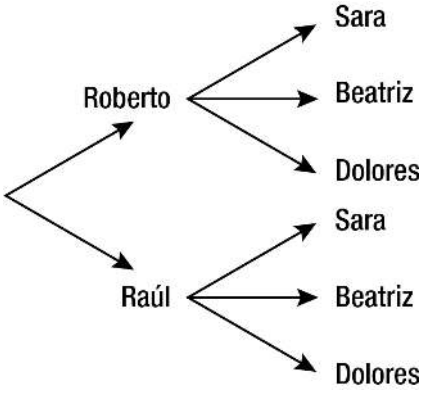


Figura 4.15 Diagrama de árbol para formar un comité.

Ejemplo 4.24

Del equipo de trabajo E del ejemplo anterior, se debe nombrar un coordinador y un secretario. ¿De cuántas maneras se pueden elegir ambos representantes?

Solución

Existen cinco posibles elecciones para el puesto de coordinador. De los restantes quedan cuatro posibles candidatos para escoger al secretario. Como se puede ver en la figura 4.l6 existen 20 posibles formas diferentes para elegir a los dos comisionados.

Observemos que en la primera selección existen 5 formas diferentes, y en la segunda elección hay 4 modos diferentes. Entonces, las dos elecciones juntas se pueden efectuar de 5×4 maneras diferentes.

Este procedimiento puede extenderse a más elecciones, y da lugar a lo que se conoce como *principio de multiplicación* Abajo se expresa en su generalización.

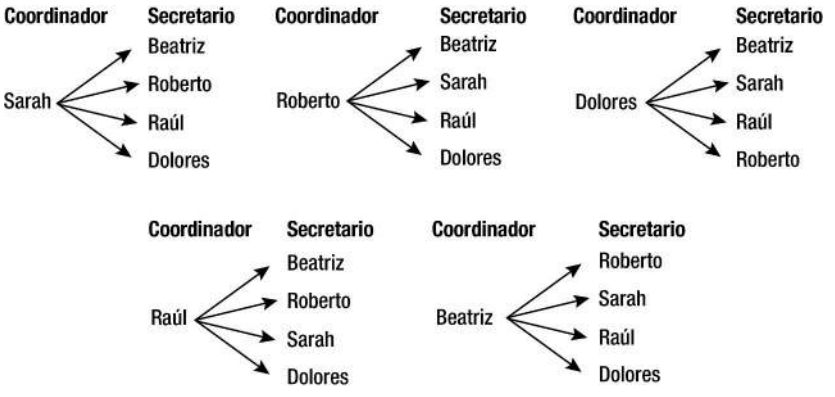


Figura 4.16 Diagramas de árbol para elegir un coordinador y un secretario.

El principio de la multiplicación

Si una operación se puede ejecutar de n_1 maneras, y después se efectúa en cualquiera de esas maneras una segunda operación, se puede ejecutar de n_2 maneras, y después se efectúa en cualquiera de esas formas, y así sucesivamente hasta para k operaciones. Entonces k operaciones se pueden ejecutar conjuntamente en $n_1 \times n_2 \times n_3 \times ... \times n_k$ maneras.

Ejemplo 4.25

Del equipo de trabajo E del ejemplo 4.23, se debe enviar un representante a dos reuniones de trabajo que se realizaran en dos días diferentes. ¿De cuántas maneras se puede seleccionar al representante si cada persona del equipo puede ser elegido?

Solución

Para la primera reunión existen 5 posibilidades diferentes de elegir a una persona. A la siguiente reunión puede ir nuevamente cualquiera de los cinco, esto es, otra vez hay 5 posibles elecciones. Por consiguiente, aplicando el principio de multiplicación, existen 5×5, es decir, 25 posibilidades.

Ejemplo 4.26

¿Cuántas placas se pueden hacer usando 3 letras seguidas por 4 dígitos?

Solución

Observe que en este ejemplo debemos llenar 7 espacios.

El primer espacio se puede llenar con 25 letras posibles (sin la I), luego en el segundo y tercer espacios se colocan 26 letras; en el cuarto espacio se colocan 9 números; en el quinto, sexto y séptimo espacios se colocan 10 números.

Por el principio de multiplicación se tiene:

$$25 \times 26 \times 26 \times 9 \times 10 \times 10 \times 10 = 152100000$$

Ejemplo 4.27

Permutación. Con el propósito de integrar el comité que representa a la escuela, se propusieron a Alberto (A), Bernardo (B) y Carmen (C). El comité debe estar formado por un presidente, un secretario y un tesorero. ¿De cuántas formas posibles se puede formar ese comité con estas tres personas?

Solución

Opción 1 (se ha planteado con anterioridad): recurrir al diagrama que se muestra en la figura 4.l7.

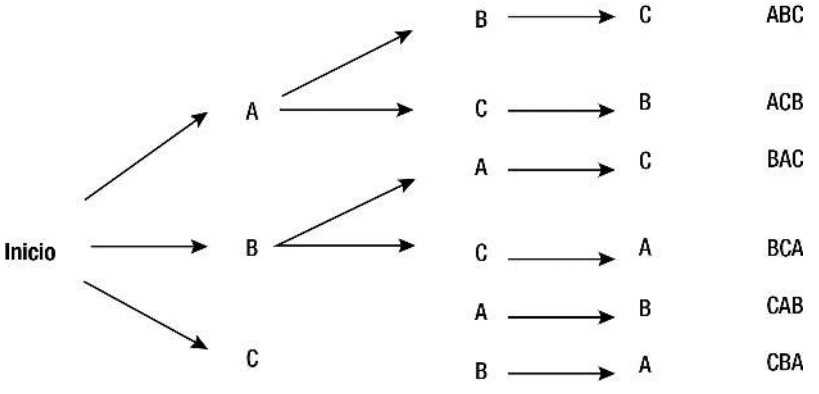


Figura 4.17 Diagrama de árbol para formar un comité.

El primer arreglo posible (ABC), indica que Alberto es el presidente, Bernardo el secretario y Carmen la tesorera. ¿Cuáles son los otros comités?

Opción 2. Una solución más conveniente va más allá del diagrama de árbol y el razonamiento es como sigue: el problema requiere que llenemos tres espacios.

En el primer espacio podemos poner A, B o C, por lo que ese espacio puede llenarse de 3 maneras:

$$3$$

Esta cifra se indica en el diagrama de árbol con las 3 ramas que salen de I. Para llenar el siguiente espacio, ahora tenemos 2 posibilidades porque uno de ellos ya es presidente, es decir:

$$3 \quad 2$$

Así tenemos llenos los primeros 2 espacios 3×2 de 6 maneras diferentes. Para llenar el último espacio, sólo nos queda una persona. Por lo tanto, se pueden llenar los espacios de 6×1 o de 6 maneras diferentes. El número de formas para completar cada uno de los tres espacios es:

$$3 \quad 2 \quad 1$$

Comparamos las dos alternativas y observamos que ambas proporcionan el mismo número de arreglos. Es decir el total de números de arreglos se tiene por medio de la multiplicación:

$$3 \times 2 \times 1 = 6$$

Permutación es una palabra más precisa para referirnos a los arreglos de objetos. En referencia a este ejemplo que estamos tratando, decimos que hay 6 permutaciones de las 3 personas, considerando las 3 a la vez con el fin de formar el comité.

Permutación

Una permutación de un número de objetos es cualquier arreglo de estos objetos en un orden definido.

Fórmulas para permutaciones

El principio de multiplicación proporciona un método general para encontrar el número de permutaciones de un conjunto de objetos. Este método puede abreviarse por medio de algunos símbolos y fórmulas para cierta clase de problemas.

Como se ha visto, el principio de multiplicación permite que se establezcan relaciones como las siguientes:

1. Se pueden ordenar 7 personas en una fila en

$$7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 \text{ maneras diferentes.}$$

2. Se pueden acomodar 20 libros en un librero en

$$20 \times 19 \times 18 \times \ldots \times 3 \times 2 \times 1 \text{ modos diferentes.}$$

3. Se pueden arreglar n objetos en una línea en

$$n(n-1)(n-2)\ldots 3 \times 2 \times 1 \text{ formas diferentes.}$$

Los puntos indican que se comienza multiplicando con un número n y se continúa multiplicando con otro número menor que el precedente hasta llegar a 1.

Complemento técnico: factorial El producto de todos los números desde 1 hasta n se llama *n factorial,* y se denota por $n!$ Ésta se escribe:

$$n! = n(n-1)(n-2)...3 \times 2 \times 1 = n(n-1)!$$

Ejemplos:

$$\begin{aligned} 1! &= 1 \\ 2! &= 2 \times 1 = 2(1)! = 2 \\ 3! &= 3 \times 2 \times 1 = 3(2)! = 6 \\ 4! &= 4 \times 3 \times 2 \times 1 = 4(3)! = 24 \\ 5! &= 5 \times 4 \times 3 \times 2 \times 1 = 5(4)! = 120 \end{aligned}$$

Ejemplo 4.28

Del conjunto E del equipo de trabajo descrito en el ejemplo 4.23, se eligieron un coordinador y un secretario. Ahí se determinó que había 20 posibles elecciones para obtener a los representantes mediante la expresión 5×4, es decir, 20 *permutaciones* del conjunto E tomando los elementos de 2 en 2. En cada uno de los 20 casos, la primera persona debe ocupar el puesto de coordinador y la segunda, el de secretario. Por ello, es importante el orden en que se consideren las personas. ¿Cómo obtener la permutación?

Solución

Para resolver el problema es necesario que recordemos que la permutación de un conjunto de elementos es una ordenación específica de algunos elementos del conjunto. En este caso, la permutación de cinco elementos tomados de dos en dos es 20. Esta se expresa en símbolos por la expresión:

$$_5P_2 = 20$$

La expresión $_5P_2$ se lee de la siguiente manera: la permutación de 5 elementos tomados de 2 en 2.

Complemento técnico: permutación: Con el propósito de obtener un valor numérico para la *permutación* se puede utilizar la siguiente fórmula:

$$_nP_r = \frac{n!}{(n-r)!}$$

En el ejemplo anterior era $n = 5$ y $r = 2$, por tanto,

$$_5P_2 = \frac{5!}{(5-2)!} = 5 \times 4$$

Observemos que $r < n$ y que $0! = 1$.

Ejemplo 4.29

¿Cuántas ternas pueden formarse con las 26 letras del alfabeto si cada letra sólo puede emplearse una vez?

Solución

En este caso, se desea determinar el número de permutaciones de 26 elementos tomados de 3 en 3. Considerando la fórmula se tiene:

$$_{26}P_3 = \frac{26!}{(26-3)!} = 26 \times 25 \times 24 = 15\ 600$$

Ejemplo 4.30

Combinación. La *combinación* es otra técnica que permite la selección de objetos sin fijarse en su orden. Asimismo, conviene distinguir entre la permutación y la combinación; para ello considérese la siguiente situación: ¿De cuántas formas un lector puede seleccionar tres libros, sin fijarse en el orden de un conjunto de 4 libros denotados por *A, B, C* y *D*?

Solución

Se ha visto que el número de permutaciones de 4 libros diferentes, tomados 3 a la vez es:

$$_4P_3 = 4 \times 3 \times 2 = 24$$

En esta permutación el orden de los libros cuenta.

No obstante, el problema es completamente diferente cuando deseamos hacer una selección de 3 libros, de 4 libros *A, B, C* y *D*, sin considerar el orden. Estas sólo son cuatro posibles selecciones:

$$ABC\ ABD\ ACD\ y\ BCD$$

Como puede verse, *ACB* no está en la lista, pues la selección de *ACB* es la misma que *ABC*, puesto que el orden no cuenta.

Se llama combinación a la lista *ABC*, *ABD, ACD* y *BCD* de 4 libros que se tomaron 3 a la vez. El número total de combinaciones se denota por:

$$_4C_3 \text{ o por } \binom{4}{3}$$

$\binom{4}{3}$ se lee como el número de combinaciones de 4 cosas tomando 3 a la vez.

La fórmula para $\binom{4}{3}$ es:

$$\binom{4}{3} = \frac{4!}{3!(4-3)!} = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times 1} = 4$$

Relación entre una permutación y una combinación. La diferencia entre una permutación y una combinación es que en una permutación el orden cuenta, mientras que en una combinación el orden no cuenta.

Una *combinación* es la selección de n objetos considerados sin fijarse en su orden. Un subconjunto de r objetos seleccionados sin fijarse en su orden de un conjunto de n objetos diferentes, se llama una combinación de los n objetos tomados r a la vez.

El número total de combinaciones se denota por $_nC_r$ o por $\binom{n}{r}$, con $r \leq n$.

Consideremos los cuatro libros *A, B, C* y *D* y la lista de las posibles selecciones de 3 libros de 4. En la tabla 2, se señala, en la primera columna, la lista de los posibles resultados en una combinación. Pero con un nuevo arreglo, se obtienen 6

permutaciones de cada una de las soluciones de la columna 1 de la tabla 4.2.

Tabla 4.2 Comparación entre la relación de la combinación y permutación.

Combinaciones	Permutaciones					
ABC	ABC	ACB	BAC	BCA	CAB	CBA
ABD	ABD	ADB	BAD	BDA	DAB	DBA
ACD	ACD	ADC	CAD	CDA	DAC	DCA
BCD	BCD	BDC	CBD	CDB	DBC	DCB

La fórmula de la combinación de n cosas r a la vez, esto es, el número de combinaciones de un conjunto de n objetos diferentes tomando r a la vez, es:

$$\binom{n}{r} = \frac{n!}{r!(n-)r!}$$

Ejemplo 4.31

En una fuente de sodas hay una mesa con 5 sillas(véase la figura , llegan 3 personas y se sientan. Si las personas se sientan de manera aleatoria, la lista de todos los posibles arreglos de 3 sillas ocupadas y 2 vacías es la combinación de 5 sillas, tomadas 3 a la vez.

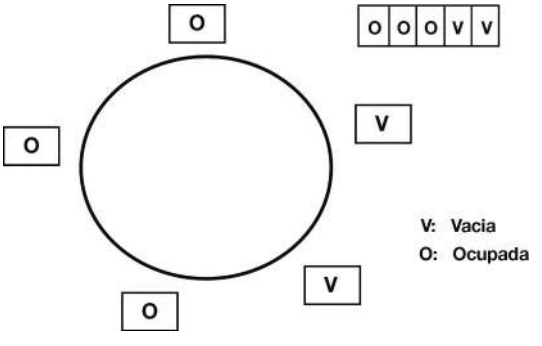


Figura 4.18 Mesa con 5 sillas.

Solución

Asientos	1	2	3	4	5	1	2	3	4	5
	O	O	O	V	V	O	V	O	O	V
	O	O	V	O	V	V	O	O	O	V
	O	O	V	V	O	V	O	O	V	O
	O	V	O	V	O	V	O	V	O	O
	O	V	V	V	O	V	V	O	O	O

$$\binom{5}{3} = \frac{5!}{3!2!} = 10$$

Ejemplo 4.32

¿De cuántas maneras se puede formar un comisión de 3 personas que son seleccionadas de 4 matrimonios

1. Si todos tienen la misma posibilidad de ser elegidos?
2. Si el comité consiste de 2 mujeres y un hombre?
3. Si el esposo y la esposa de un matrimonio no pueden estar en la comisión?

Solución

1. Si en la comisión el orden no cuenta, el problema es seleccionar 3 personas de 8, de todas las formas posibles. Aplicando la fórmula de combinatoria se tiene:

$$\binom{8}{3} = \frac{8!}{3!(8-3)!} = \frac{8!}{3!5!} = \frac{8 \times 7 \times 6}{1 \times 2 \times 3} = 56$$

2. La mujer se puede seleccionar de $\binom{4}{2}$ maneras, y después seleccionar a un hombre y esto puede ser mediante $\binom{4}{1}$ formas. Por el principio de multiplicación el número de maneras para seleccionar 2 mujeres y 1 hombre es:

$$\binom{4}{2} \times \binom{4}{1} = 6 \times 4 = 24$$

4.7 Resumen

La probabilidad es el puente entre el estudio de la estadística descriptiva vista en las unidades anteriores y la inferencia estadística (en la unidad 7 se exponen principios básicos de este tema). Los conceptos que son relevantes para el desarrollo de la probabilidad se resumen a continuación.

Evento	Un evento A es un resultado o conjunto de resultados que son de interés para el experimentador.

Experimento	Resultado de un proceso que genera una observación que no se puede predecir.
Probabilidad	Medida numérica asociada con algún resultado e indica qué tan probable es que el resultado pueda ocurrir.
Espacio muestral	Conjunto de todos los posibles resultados de un experimento. Si en un número N grande de ensayos independientes
Probabilidad empírica	k de estos ensayos son resultado de un evento A, la probabilidad del evento A es k/N y se expresa por: $P(A) = \frac{k}{N}$
Complemento	El complemento de un evento A es la colección de resultados en el espacio muestral que no está en A.
A o B	Contiene todos los resultados en A o B, o en ambos
A y B	Contiene todos los resultados que están en los eventos A y B
Mutuamente excluyentes	Los eventos A y B son mutuamente excluyentes si no tienen elementos en común.
Eventos Independientes	Los eventos A y B son independientes si el que uno ocurra no afecta la probabilidad de que ocurra el otro.
Probabilidad de un evento, $P(A)$	Es una medida de la verosimilitud de que el evento A ocurrirá.

Lista de fórmulas.

Frecuencia relativa	$P(A) = \frac{Nmero\ de\ veces\ en\ que\ el\ evento\ ha\ ocurrido}{nmero\ total\ de\ observaciones}$
Probabilidad clásica	$P(A) = \frac{\text{Número de formas en las que puede ocurrir un evento}}{\text{número total de posibles resultados}}$
Complemento	$P(A) = 1 - P(A^c)$ o $P(A^c) = 1 - P(A)$
Ley aditiva	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Ley aditiva eventos excluyentes	$P(A \cup B) = P(A) + P(B)$
Probabilidad Condicional	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$

Lista de fórmulas.

Ley de la multiplicación	$P(A \cap B) = P(B)P(A \mid B)$
Ley de la multiplicación eventos independientes	$P(A \cap B) = P(A)P(B)$

Técnicas de conteo.

Principio	Descripción	Fórmula
Principio básico de conteo	Si un evento puede ocurrir de n_1 maneras distintas y un segundo evento ocurre de n_2 maneras distintas, el número de resultados posibles es $n_1 \times n_2$	$n_1 \times n_2$
Permutaciones	El número de arreglos ordenados de n objetos distintos. El número de permutaciones de n objetos distintos tomados de k a la vez, donde $k \leq n$	$n!$ ${}_nP_k = \frac{n!}{(n-k)!}$
Combinaciones	El número de combinaciones de k objetos seleccionados de un grupo de n objetos sin considerar el orden	${}_nC_k = \frac{n!}{(n-k)!k!}$

4.8 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

4.9 Ejercicios

Nociones básicas de probabilidad

4.1 Describa cinco situaciones donde intervenga el azar.

4.2 A partir del ejemplo 5.1, los tres estudiantes seleccionados fueron Joaquín, Darío y Silvia. Se vio que Joaquín y Darío sí leyeron, en cambio, Silvia, no. Indique cuáles de los cuatro eventos *A, B, C y D* ocurrieron, y cuáles no.

4.3 En un estudio de salud, los asistentes son asignados al azar a uno de cinco grupos diferentes de ejercicios. Liste los posibles resultados en el espacio muestral y los siguientes eventos:

1. Asignado al grupo 3.
2. Asignado a uno de los primeros tres grupos.
3. Asignado al grupo 4 o 5.
4. Asignado a un grupo entre 2 y 5 inclusive

4.4 En un grupo, 60 % de los alumnos son mujeres. Se selecciona al azar tres alumnos para representar a la escuela en un concurso literario. Liste los resultados del espacio muestral.

4.5 A una persona le presentan tres palabras asociadas con su significado. Junto a cada una están escritas dos opciones, una correcta y otra incorrecta. La persona no entiende claramente el significado y responde por el significado de cada palabra adivinando. Use un diagrama de árbol para generar el espacio muestral.

4.6 En un restaurante sirven tres tipos de sopa de pasta: (espagueti, macarrón y fettuccini) y lo sirven con una de dos salsas (tomate y blanca). Un cliente ordena un plato de sopa de pasta.

1. ¿Cuál es el espacio muestral de las posibles selecciones del cliente?
2. Liste los resultados de los siguientes eventos: $A = \{\text{Un plato de espagueti}\}$ y $B = \{\text{Un plato con salsa de tomate}\}$

Probabilidad de un evento

4.7 Calcule las probabilidades de los eventos B y D del ejemplo 4.6.

4.8 Un grupo de personas está compuesto por 2 niños menores de 12 años, 3 adolescentes y 5 adultos. Se debe seleccionar a una persona al azar. ¿Cuál es la probabilidad de que la persona sea un adulto? ¿Cuál es la probabilidad de que la persona sea mayor de 12 años?

4.9 En un estudio de ejercicio aeróbico, las personas son asignadas de manera aleatoria a cinco grupos diferentes de ejercicio. Escriba una lista de los sucesos elementales de:

1. El espacio muestral
2. Cada uno de los siguientes eventos: A : se asigna al grupo 3, B: asignada a uno de los tres primeros grupos, C: asignada al grupo 4 o 5, D: asignada del grupo 2 al 5.
3. Asigne la probabilidad a cada uno de los sucesos elementales del espacio muestral y calcule la probabilidad de cada evento.

4.10 Una persona está realizando una encuesta en hogares para conocer la preferencia sobre una telenovela. Él pregunta en cada casa si le gusta o no les gusta la telenovela. Hasta el momento lleva la información de 4 casas. Describa una lista de todos los posibles sucesos elementales que puede tener el entrevistador:

1. Del espacio muestral.
2. De los eventos: A: a nadie le gusta la novela, B: en tres casas les gusta la novela, C: sólo en una casa gusta la novela.
3. Asigne la probabilidad a cada uno de los sucesos elementales del espacio muestral y calcule la probabilidad de cada evento.

4.11 Se lanza un dado de seis caras.

1. Describa el espacio muestral.
2. ¿Cuál es la probabilidad de obtener un número mayor que 3?

4.12 Susana y Martha están rifando un reproductor de discos compactos (discman). Susana debe vender 35 boletos a 30 pesos, mientras que Martha hizo 50, pero los vende a 20 pesos. Roberto tiene 180 pesos, pero no decide a cuál de las dos comprar el boleto porque no sabe con quién tiene mayor probabilidad ganar. ¿A quién le debe comprar Roberto para tener mayor probabilidad de ganar?

4.13 Una persona se interesa en instalar un café internet. Para ello efectúa un estudio de mercado. Entrevista a 460 adolescentes en escuelas cercanas al lugar donde piensa establecerse. Les pregunta cuántas horas dedican a navegar por internet a la semana. Las respuestas son:

A: menos de 1 hora	23
B: de 1 a 4 horas	55
C: de 4 a 7 hora	69
D: de 8 a 10 horas	147
E: de 10 a 15 horas	120
F: más de 15 horas	46
Total	460

Calcule el porcentaje en la última columna de la tabla anterior. ¿Cuál es la probabilidad de que una persona usuaria del internet seleccionada al azar dure entre 8 y 10 horas? ¿Conviene instalar internet si existe alta posibilidad de que las personas estén en el internet más de 8 horas?

Relación entre eventos y leyes de probabilidad

4.14 Use la información del ejemplo 5.12 para calcular las probabilidades que se señalan en la interpretación, es decir:

1. La probabilidad de que el alumno seleccionado obtenga un rendimiento académico arriba de 8 y la probabilidad de que el alumno obtenga un rendimiento académico debajo de 8. Compárelas.
2. La probabilidad de que el alumno seleccionado esté en el turno vespertino y con rendimiento académico arriba de 8. La probabilidad de que el alumno seleccionado esté en el turno matutino y con rendimiento académico arriba de 8.
3. La probabilidad de que el alumno seleccionado sea mujer con promedio arriba de 8. La probabilidad de que el alumno seleccionado sea hombre con promedio arriba de 8. Contraste estas probabilidades.

4.15 Se lanzan dos dados, uno blanco y otro verde. Describa el espacio muestral M de todos los posibles resultados. Use un plano de coordenadas para presentar en una gráfica los elementos de M.

1. Identifique los eventos A ={suma siete}, B ={suma impar}, C ={suma once}, D ={suma seis}, E ={el mismo número en cada dado}.
2. Si los dados no están cargados, asigne probabilidades a cada resultado elemental.
3. Calcule las probabilidades $P(A)$, $P(B)$, $P(C)$, $P(D)$, $P(E)$.
4. Obtenga la probabilidad de $P(A \cup C)$.
5. ¿Cuál es la probabilidad de que cada, uno de ellos muestre al menos cinco puntos al caer?

4.16 Raúl y Omar juegan a lanzar dos dados de seis caras considerando las siguientes reglas:

- Calcular la diferencia de los puntos entre el mayor y menor.
- Si la diferencia es de 0,1 o 2, entonces Omar se anota un punto.
- Si la diferencia es de 4,5 o 6 es Raúl el que gana un punto.
- Juegan a 25 puntos.

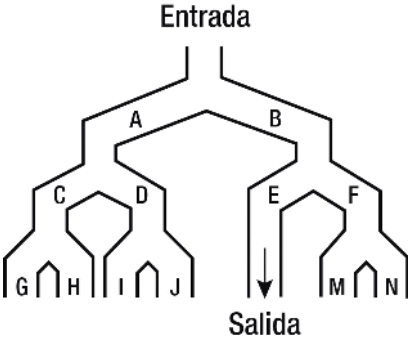
¿Parece este juego equitativo? ¿Qué jugador conviene ser?

1. Juegue con un compañero, tiren los dados unas 15 veces y anoten los resultados en cada lanzamiento. ¿Qué observan en función de la frecuencia relativa?
2. ¿Cuál es el espacio muestral?
3. Enumere los sucesos elementales de cada uno de los siguientes eventos.

A :{la diferencia de los puntos es 5}, B :{la diferencia de los puntos es 4}
C :{la diferencia de los puntos es 3}, D :{la diferencia de los puntos es 2}
E :{la diferencia de los puntos es 1}, F :{la diferencia de los puntos es 0}.

4. Calcule las probabilidades de D, E, F y $A \cup B \cup C$.
5. ¿Cuál es la probabilidad de que el juego lo gane Omar?

4.17 Una persona se desplaza por el laberinto que se muestra en la figura:



¿Cuál es la probabilidad de que pueda salir del laberinto si cada camino tiene la misma probabilidad de ser elegido por una persona?

4.18 Una tienda vende cachuchas (gorras) de varios colores: rojas (r), naranjas (n), amarillas (a), verde (v), café (c), morado (m) y blanco (b). Para la próxima venta de una cachucha, considere los dos eventos $A = \{n, v, c, b\}$ y $B = \{r, n, c\}$

1. Especifique el espacio muestral.
2. Dibuje un diagrama de Venn que muestre los eventos A y B.
3. Determine la composición de los siguientes eventos:

   a) $A \cap B$

   b) B^c

   c) $A \cap B^c$

   d) $A \cup B$

4. Suponga que las probabilidades de los sucesos elementales para la próxima compra de una cachucha son: $P(r) = P(n) = P(a) = 0.1$, $P(v) = P(c)0.15$, $P(m) = 0.05$, $P(b) = 0.35$.
5. Encuentre $P(A)$, $P(B)$ y $P(A \cap B)$
6. Utilizando las leyes de probabilidad y los resultados del apartado anterior, calcule $P(A^c)$ y $P(A \cup B)$.

7. Verifique sus respuestas del apartado anterior sumando las probabilidades de los resultados elementales en A^c y $A \cup B$.

4.19 Ocho repartidores de pizzas $r_1, r_2, \ldots, r_8$ están trabajando cerca de su escuela. ¿Cuál de ellos repartirá el siguiente pedido?

1. Especifique el espacio muestral.
2. Considere tres eventos $A = \{r_1, r_2, r_5, r_6, r_7\}$, $B = \{r_2, r_3, r_6, r_7\}$ y $C = \{r_6, r_8\}$. Dibuje el diagrama de Venn que muestre estos tres eventos.
3. Considere que las probabilidades de que un repartidor sea el siguiente en entregar una pizza son:
   $P(r_1) = 0.16$, $P(r_2) = P(r_3) = P(r_4) = 0.08$, $P(r_5) = P(r_6) = P(r_7) = P(r_8) = 0.15$.
   Dé la composición y determine la probabilidad de:

   a) $B^c \cap C$

   b) $B \cap C$

   c) $A \cup C$

   d) $A^c \cup C$

   e) Escriba el evento en notación de conjuntos, dé la composición y encuentre la probabilidad de:

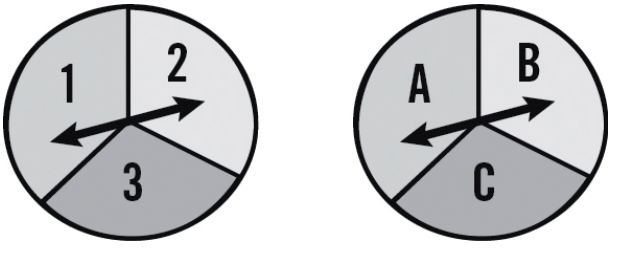
   C no ocurre
   A y B ocurren
   A ocurre pero B no ocurre
   Tanto A como C no ocurren

4.20 En su ciudad se van a estrenar tres películas muy populares. Una es de aventura, otra es de suspenso y la tercera es de terror. Las letras A, S y T definen los eventos de conseguir boletos para las películas. Exprese los siguientes eventos en notación de conjuntos.

1. Obtener boletos para las películas de Aventura y la de Suspenso.
2. Obtener boletos para las películas de Aventura y la de Suspenso pero no para la película de Terror.
3. No consigue boletos para las películas de Aventura y la de Suspenso.

4.21 En una bolsa hay 12 bolas rojas y seis negras. Si se sacan dos en forma consecutiva, sin reponer la primera, ¿cuál es la probabilidad de que la primera sea roja y la segunda sea negra?

4.22 Un experimento consiste en dos pequeñas ruletas con una flecha giratoria al centro. La gráfica es la que se muestra abajo. Considere tanto los números como las letras y escriba el espacio muestral. El evento A consiste en que la primera flecha marca número impar. Encuentre $P(A)$. B es el evento donde la segunda flecha cae en vocal. Calcula $P(B)$ y $P(A \cup B)$.



4.23 En la actualidad, aprender inglés se ha convertido en una necesidad. Unos estudiantes quieren registrarse para tomar un curso durante el verano. Se abrieron 6 grupos, pero por políticas de la escuela se asigna al azar a cada estudiante a uno de esos grupos.

1. Describa el espacio muestral y conteste lo siguiente: ¿cuál es la probabilidad de cada resultado elemental?
2. Un estudiante tiene doble oportunidad de ser asignado a un grupo impar. Asigne las probabilidades a cada posible resultado elemental.
3. Si a Teresa la colocaron en un grupo marcado con número par, y a Juan en uno cuyo número es menor que 3, ¿cuál es la probabilidad de que Teresa y Juan estén en el mismo grupo?

4.24 Se lanza un dado de ocho caras, el espacio muestral es $M = \{1, 2, 3, 4, 5, 6, 7, 8\}$, $A = \{3, 5, 7\}$, $B = \{2, 4, 6\}$, $C = \{3, 4, 5, 6, 7\}$. Liste los resultados de los siguientes eventos.

1. $A \cup B$
2. $A \cap B$
3. $A \cap C$
4. C^c
5. $A^c \cup C$
6. $A \cup B \cup C$
7. $A \cap B \cap C$
8. ¿A y B son mutuamente excluyentes?

4.25 En el lanzamiento de un dado de seis caras considere los siguientes eventos

A: *que* el dado marque un número impar $= \{1, 3, 5\}$
B: *que* el dado marque un número mayor que $4 = \{5, 6\}$
C: *que* el dado marque un número primo $= \{2, 3, 5\}$

Vea la opción lanzamiento de un dado, figura 4.6. ¿Cuál es la probabilidad de $A \cup B$, $A \cup C$ y $B \cup C$?

4.26 Para dos eventos, se especifican las siguientes probabilidades: $P(A) = 0.52$, $P(B) = 0.36$, $P(A \cap B) = 0.20$.

1. Con estas probabilidades complete la siguiente tabla.
2. Determine las probabilidades de $A \cap B^c$, $A^c \cap B$ *y* $A^c \cap B$, y llene la tabla.
3. Exprese los siguientes eventos en notación de conjuntos y encuentre sus probabilidades.

   a) B ocurre y A no ocurre.

   b) Ni A ni B ocurren.

   c) Sea que A ocurra o que B no ocurra.

4.27 El historial médico reportado en una clínica durante un año, arroja los siguientes porcentajes:

	Caso Ligero		**Caso grave**	
Edad (años del paciente)	Antecedentes de diabetes		Antecedentes de diabetes	
	Si	No	Si	No
Debajo de 40	15 %	10 %	8 %	2 %
Arriba de 40	15 %	20 %	20 %	10 %

Supongamos que un paciente es seleccionado de manera aleatoria de este grupo, los eventos A, B y C se definen como sigue: A =\{Tiene un caso serio\}, B =\{su edad está por debajo de 40\} y C =\{Con antecedentes en diabetes\}

1. Encuentre las probabilidades $P(A)$, $P(B)$, $P(B \cap C)$ *y* $P(A \cap B \cap C)$.
2. Describa verbalmente los siguientes eventos y encuentre sus probabilidades:

   a) $(A \cap B)^c$,

   b) $A^c \cup C^c$,

   c) $A^c \cap B \cap C^c$

Temas selectos en el cálculo de probabilidades

4.28 Una empresa produce dos tipos de zapatos denominados A: de vestir y B: casual. Las probabilidades de que A tenga cero defectos es $P(A) = 0.68$, B cero defectos es $P(B) = 0.55$ y de que no haya ningún defecto en ambos es $P(A \cap B) = 0.32$.

1. Encuentre la probabilidad condicional de que B ocurra dado que A ocurrió.
2. Encuentre la probabilidad condicional de que B no ocurra dado que A ocurrió.
3. Encuentre la probabilidad condicional de que B ocurra dado que A no ocurrió.

4.29 La probabilidad de que una planta germine ante ciertas condiciones adversas es $P(A) = 0.4$ y la probabilidad de que otra variedad de planta germine ante las mismas condiciones es $P(B) = 0.6$ y la $P(A \cup B) = 0.7$.

1. Encuentre la probabilidad condicional de que A ocurra dado que B ocurrió.
2. Encuentre la probabilidad condicional de que B ocurra dado que A no ocurrió.

4.30 Considere las probabilidades $P(A) = 0.55$, $P(B) = 0.68$ *y* $P(A^c \cap B) = 0.35$.

1. Determine las probabilidades para completar la tabla:

	B	B^c	
A			0.55
A^c	0.35		
	0.68		

2. Encuentre la probabilidad de A dado que B no ocurrió.

4.31 Una empresa encuesta a 120 personas de las cuales 2/3 son hombres y de ellos 2/5 fuman. Se sabe que 1/3 de las mujeres fuman, calcule:

1. La probabilidad de que al elegir un encuestado al azar sea un hombre que fume.
2. La probabilidad de elegir un encuestado que no fume sabiendo que es mujer.

4.32 Los promotores de una editorial realizaron una encuesta en las facultades de economía, fe, y administración, fa, con la finalidad de saber si los estudiantes prefieren comprar libros en papel o tabletas. El número de estudiantes encuestados fueron 200. Los resultados se describen en la tabla:

	Papel (P)	**Tabletas (T)**	**Total**
E: fe	60	56	116
A: fa	44	40	84
Total	104	96	200

Determine la probabilidad de que el estudiante:

1. Considere comprar libros en papel, dado que estudia economía.
2. Sea economista dado que prefieren la tabletas.

4.33 Con base en información recabada sobre la calidad de un servicio, un administrador conoce que la probabilidad de que el servicio que presta sea malo, regular, bueno, muy bueno, excelente es: 0.18, 0.20, 0.37, 0.19, 0.06, respectivamente. Estime las probabilidades de que el servicio prestado se clasifique como:

1. Al menos regular.
2. Bueno o muy bueno.
3. Ni malo ni excelente.
4. A lo más, bueno.

4.34 $P(A) = 0.5$ es la probabilidad de que una persona padezca sobrepeso, $P(B) = 0.25$ es la probabilidad de que una persona sea hipertensa y $P(A|B) = 0.8$. Aplicando las reglas de probabilidad calcule:

1. $P(A^c)$
2. $P(A \cap B)$
3. $P(A^c \cap B)$

4.35 Si A es el evento de que un estudiante obtendrá una beca, B el evento de que el estudiante encontrará un trabajo de medio tiempo, y C es el estudiante se graduará, exprese en términos de símbolos la probabilidad de que:

1. Un estudiante que obtuvo la beca se graduará.
2. Un estudiante que no obtuvo la beca encontrará trabajo de medio tiempo.
3. Un estudiante que no obtuvo la beca no encontrará trabajo ni se graduará.
4. Un estudiante que obtuvo la beca y trabaja de medio tiempo no se graduará.

4.36 E es el evento de que una mujer aplica para un puesto en ventas y que tiene experiencia previa, C es el evento de que ella tiene coche, y G el evento de que es graduada de la universidad. Explique, de manera verbal, que expresan las siguientes probabilidades:

1. $P(C \mid G)$
2. $P(E \mid C^c)$
3. $P(C^c \mid E)$
4. $P(G^c \mid C^c)$
5. $P(C \mid E \cup G)$
6. $P(E \cap C^c \mid G)$

4.37 Se aplica una encuesta a 400 estudiantes de diferentes niveles escolares para conocer si les da miedo, o no, visitar al dentista. Las respuestas se registran en la siguiente tabla.

	Primaria	**Secundaria**	**Preparatoria**
Miedo	48	32	20
Sin miedo	112	108	80

Consideremos los siguientes eventos A ={miedo}, B ={Secundaria}. Si un estudiante es seleccionado al azar, encuentra las siguientes probabilidades:

a). $P(A)$, *b)* $P(B)$, *c)* $P(A^c \cap B)$, *d)* $P(A \cap B)$, *e)* $P(A \mid B)$, *f)* $P(B^c \mid A)$
Conteste a la siguiente pregunta: ¿Son A y B independientes?

4.38 Hay dos cajas. En la primera hay dos letras A y una L; en la otra, cuatro letras A y dos L. Se debe elegir una de las dos cajas y a continuación extraer, al azar, tres letras, una a una sin reemplazo. Si el resultado es ALA se gana un premio. ¿Qué caja se debe elegir?

4.39 La tabla de longevidad de un país indica que la probabilidad de llegar a los 30 años es de 0.94, mientras que la probabilidad de llegar a los 70 años es de 0.62. Si una persona tiene 30 años, ¿cuál es la probabilidad de que llegue a los 70 años?

Fórmula de Bayes y diagrama de árbol

Estas actividades de aprendizaje se proponen como una práctica. Se trata de reconstruir las siguientes tablas y diagramas de árbol a partir de un ejemplo.

4.40 Un psicólogo estudia la memoria a corto plazo en dos grupos de diferente edad: niños entre 8 y 10 años y jóvenes entre 18 y 20 años. A cada niño o joven se le muestra una tarjeta con 16 palabras durante 25 segundos, se deja pasar un minuto y se le pide que en 2 minutos mencione las palabras que recuerde. El psicólogo se auxilia de varias personas para completar una muestra de 100 personas. En su reporte divide el hecho "recordar palabras" en dos categorías: la primera los que recuerdan menos de 10, y la otra 10 o más. Y hay dos grupos de edad, los cuales son: entre 8 y 10 años, y entre 18 y 20 años.

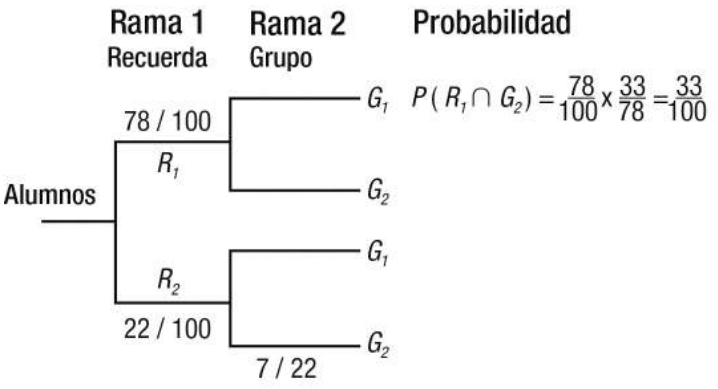
Recuerde	**G_1: grupo1**	**G_2: grupo 2**	**Total**
	8 y 10 años	18 y 20 años	
R_1: menos de 10 palabras	33	45	78
R_2: 10 o más palabras	15	7	22
Total	48	52	100

A partir de la tabla anterior complete el cálculo de las siguientes frecuencias:

	FA*	FR^0
menos de 10 palabras recordadas	78	78/100
10 o más palabras recordadas		
total del grupo 1		
total del grupo 2	48	48/100
menos de 10 palabras recordadas en el G_1		
menos de 10 palabras recordadas en el G_2	33	33/100
10 o más palabras recordadas en el G_1		
10 o más palabras recordadas en el G_2		

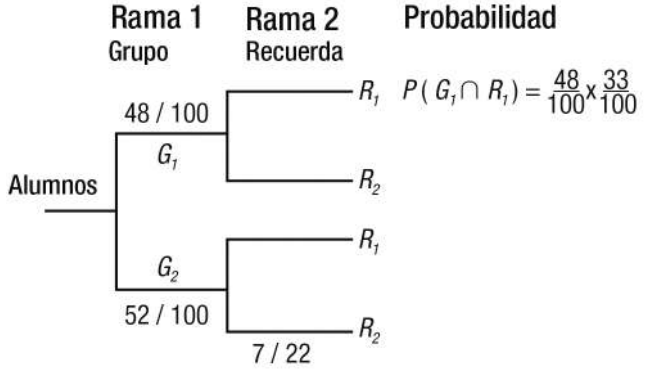
* Frecuencia absoluta, °Frecuencia relativa.

4.41 A partir de la actividad anterior, complete el siguiente diagrama de árbol, escribiendo las frecuencias relativas correspondientes.



4.42 A partir de los datos de la tabla anterior, se ha decidido cambiar el orden de los eventos en la rama 1 y la rama 2. Complete el siguiente diagrama de árbol. Considerando este diagrama de árbol, determine las siguientes frecuencias relativas:

- Personas del grupo 1 que recuerdan menos de 10 palabras.
- Personas del grupo 1 que recuerdan 10 o más palabras.
- Personas del grupo 2 que recuerdan 10 o más palabras.

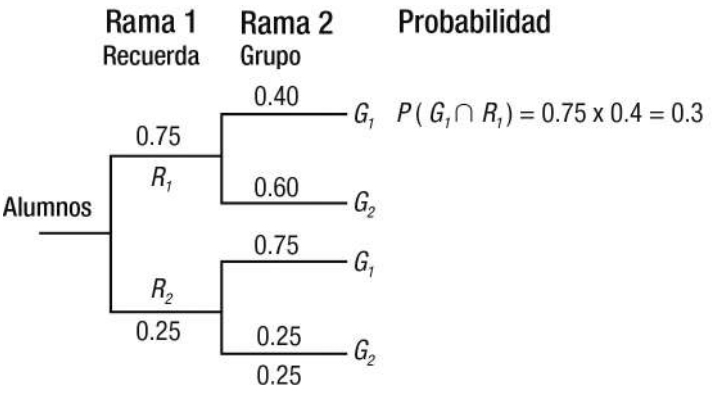


4.43 El psicólogo juntó la información de los resultados en otros estudios que realizó y la presentó en el siguiente diagrama.

1. Calcule las probabilidades de los siguientes sucesos:

   - Una persona que recuerda menos de 10 palabras está en el grupo 1.
   - Una persona que recuerda menos de 10 palabras.
   - Una persona que recuerda 10 o más palabras está en el grupo 1.
   - Una persona que es del grupo 1.

2. Complete las probabilidades de la trayectoria.
3. Calcule las probabilidades de G_1 y G_2. Tenga en cuenta que G está formado por los eventos excluyentes (R_1 y G_1) y (R_2 y G_1). De manera análoga, G_2 está formado por dos eventos excluyentes.



4. Use la fórmula de Bayes y calcula $P(R_1 \mid G_1)$.

4.44 Se realiza un examen de laboratorio para diagnosticar la presencia de tifoidea en una población. Se sabe que 5 de cada 10 000 personas en que la prueba es negativa, padece tifoidea, y 3 % en la que la prueba fue positiva también la padece. Si 3 % de las pruebas administradas fue positiva, se definen los siguientes eventos: $H_1 = \{$la prueba resultó positiva$\}$, $H_2 = \{$la prueba resultó negativa$\}$, $A = \{$la persona padece tifoidea$\}$.

1. Represente estos eventos en un diagrama de Venn.
2. ¿Cuál es la proporción de tifoidea en la población?

4.45 Una empresa que vende sistemas contra robo instala uno en una fábrica. De existir intento de robo, la alarma se activa el 98 % de las veces. También puede producirse una falsa alarma con una probabilidad de 0.007, en el caso de no haber robo. Sabiendo que la probabilidad de que se produzca un robo es de 0.010 y habiéndose detectado la alarma, ¿cuál es la probabilidad de que ésta sea infundada?

Técnicas de conteo

1. ¿De cuántas maneras se pueden ordenar en un librero 3 libros denotados por A, B y C?
2. Si se tienen al menos 3 copias de cada uno de los 3 libros A, B y C. ¿De cuántas maneras (distinguibles) se pueden arreglar 3 de los libros en el librero? Considere las copias como indistinguibles.
3. Tres botellas contienen diferentes químicos con etiquetas A, B y C. Durante el traslado a un laboratorio, las etiquetas se cayeron. Si el conductor decide pegar estas etiquetas de manera aleatoria, dado que es inexperto en química. ¿Cómo pudo ordenar las etiquetas para cada botella?

   a) Elaborar una lista de todos los posibles resultados.

   b) Encuentre la composición de los eventos. $A = \{$Sólo una etiqueta coincide$\}$ $B = \{$Todas las botellas tienen la etiqueta equivocada$\}$

4. Utilice el ejercicio anterior para:

   a) Asignar las probabilidades a cada resultado elemental.

   b) Encontrar $P(A)$ y $P(B)$.

5. Una persona tiene 5 camisas, 4 pantalones y 3 pares de zapatos. Estas prendas se pueden combinar de cualquier modo. ¿De cuántas maneras diferentes puede vestirse esa persona?
6. Un equipo de béisbol tiene 6 lanzadores y 2 receptores. ¿Cuántas parejas pueden formarse con un lanzador y un receptor?

7. Un club consta de 20 integrantes. ¿Cuántos conjuntos diferentes se pueden formar para elegir la mesa directiva, la cual debe estar compuesta por un presidente, un secretario y un tesorero? Cada persona sólo puede ocupar uno de esos puestos.
8. ¿Cuántas matrículas pueden formarse utilizando 2 letras del alfabeto seguidas por 4 dígitos, de los cuales el primero no puede ser cero?
9. Calcule a) 8! b) $\frac{11!}{7!}$ c) ${}_7P_2$ d) ${}_{10}P_{10}$ e) ${}_{12}P_3$ f) ${}_5C_2$ g) ${}_{11}C_3$ h) ${}_9C_4$
10. Una empresa planea comprar tres automóviles compactos. Dentro de sus posibilidades tiene consideradas tres marcas 7 tipos de compactos y 6 colores. ¿Cuántos automóviles distintos puede contemplar la empresa para comprar?
11. Responda los siguientes incisos.

    a) ¿Cuántas señales diferentes se pueden formar con 7 banderas si cada señal se hace con 3 banderas colocadas una debajo de otra?

    b) ¿De cuántas maneras diferentes se puede realizar un programa musical con 7 discos?

    c) ¿De cuántas formas diferentes puede disponerse el orden de bateo de los 9 jugadores de un equipo de béisbol?

    d) Explique por qué un candado de los llamados de combinación podría darse el nombre de candado de permutación.

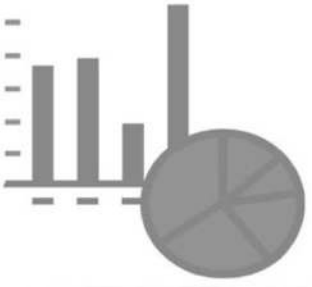
12. Una caja azul (A) contiene 5 pelotas y una caja blanca (B) tiene 10 pelotas. ¿De cuántas formas se pueden seleccionar diez pelotas si deben tomarse 3 de la caja A y 7 de la caja B?
13. ¿De cuántas maneras se pueden sentar 4 pasajeros en los 4 asientos restantes después de que el conductor se ha sentado en un coche con 5 asientos?
14. Se va seleccionar al azar un comité de 2 personas del conjunto $E = \{\text{Sarah, Rosa, Berta, José y Javier}\}$. ¿Cuál es la probabilidad de que 2 de los integrantes del comité sean mujeres?
15. Una caja contiene 2 pelotas rojas R_1 y R_2 y 2 pelotas blancas B_1 y B_2. Muestre el espacio muestral de los sucesos al sacar 2 pelotas una tras otra, sin devolverlas a la caja. Determine la probabilidad de que ambas sean rojas.
16. Se me olvidó el número secreto para acceder a la cuenta bancaria la cuál consta de 8 caracteres. Las letras que tengo consideradas son j, j, s, r, r, 4, 4, 4, 9, 9. ¿Cuántas claves pude haber realizado?
17. La administración de una empresa constructora prevé contratar a dos proveedores de azulejos. Consulta a 10 consorcios, de estos seis están afiliados a una cámara de industriales. ¿De cuántas maneras se puede seleccionar a los dos proveedores

    a) afiliados a la cámara?

    b) sin considerar la afiliación?

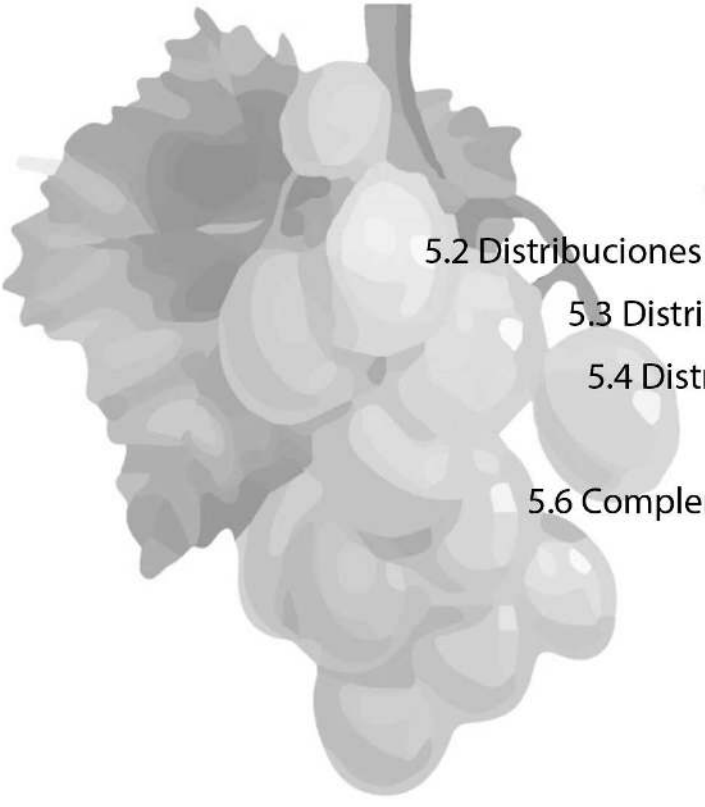
4.10 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.



Capítulo 5

Distribuciones de probabilidad: variables aleatorias discretas



Gran parte de las dificultades por las que atraviesa el mundo se deben a que los ignorantes están completamente seguros y los inteligentes llenos de dudas.

Bertrand Russell

Competencia general

Conocer los conceptos de la distribución de probabilidad para una variable aleatoria discreta, así como sus aplicaciones. Estudiar las distribuciones binomial y Poisson.

Competencias específicas

- Comprender el significado de una variable aleatoria y su distribución para el caso discreto y continuo.
- Comprender lo que se entiende por valor esperado y aprender a calcular la media y la desviación estándar de una distribución de probabilidad discreta.
- Comprender lo que se entiende por ensayos Bernoulli y aplicarlos para definir la distribución de probabilidad binomial.
- Estudiar las características matemáticas de la distribución Poisson.
- Aplicar las distribuciones binomial y Poisson en diferentes problemas en la práctica.

5.1 Introducción

En general, resulta poco práctico estudiar un proceso o investigar un tema específico en toda una población, sobre todo cuando los componentes de ésta son muy numerosos. Por ejemplo, supongamos que se desea saber el estado civil (soltero, casado, divorciado, unión libre) de las personas que trabajan de meseros en restaurantes de una ciudad, entonces recurrimos a examinar una muestra e inferir de la población entera a partir de ésta. Debido a que las predicciones o decisiones que formulamos sobre una población con apoyo de la información muestral, genera un grado de incertidumbre, ésta se expresa en probabilidades. Para el caso de los meseros, puede ser de interés estudiar si la probabilidad de divorcio es alta.

El principio de la distribución de probabilidad resulta al considerar un experimento aleatorio, y al preguntarnos acerca de los eventos posibles y sus respectivas probabilidades. Hay dos tipos de distribución de probabilidad los cuales son de gran importancia en muchas áreas del conocimiento: las distribuciones discretas y las continuas. Las primeras se derivan de las variables aleatorias discretas, por ejemplo el número de llamadas telefónicas que se reciben por quejas, el número de respuestas contestadas de manera correcta en un examen, el número de artículos defectuosos, etcétera. Las distribuciones continuas se generan a partir de variables aleatorias continuas; ejemplos de este tipo de variables son el periodo de vida de una batería, el promedio de las calificaciones de los alumnos al finalizar un ciclo escolar, o el coeficiente intelectual de una persona.

En el capítulo anterior se presentó el concepto de espacio muestral y tanto los problemas como ejemplos que se explican en esta unidad toman en cuenta esa idea para indicar cómo surge la variable aleatoria y su distribución de probabilidad.

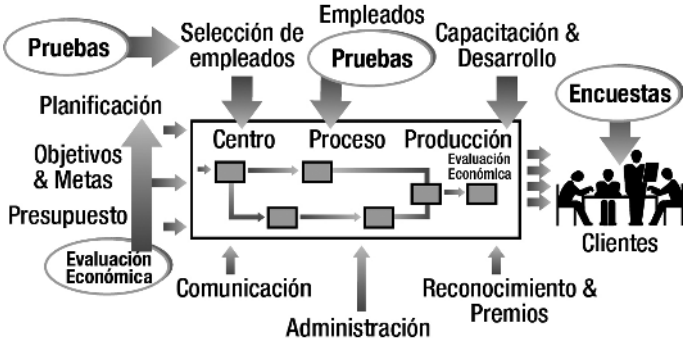


Figura 5.1 Descripción global de un proceso en que se destacan algunas características de interés.

Escenario variable aleatoria: eźito y fracaso

Se puede considerar una gran cantidad de escenarios en que destacan las variables aleatorias discretas. En particular, en la figura 5.1 se describe la relación de un proceso de producción. Estas actividades están relacionadas con la esencia administrativa y el impacto económico de decisiones adecuadas, lo que

nos lleva a una serie de temas de interés para la planeación. Por ejemplo en la economía los trabajadores esperan contar con un ingreso para la seguridad social una vez que se jubilan. Entre otras cuestiones, en una encuesta se les pregunta a los trabajadores si esperan un apoyo médico en la jubilación, y qué tan seguros están de permanecer en su empleo. Otro ejemplo es cuando la administración hace estudios para evaluar el desempeño de sus empleados, tal como estimar el costo de la eficiencia: "los empleados procuran un mejor uso de los materiales y equipos". Por otra parte, en la producción, la administración espera reducir el número de defectos y así aumentar la productividad. En esa misma dirección, los clientes esperan que el producto que se obtiene del proceso no sea defectuoso.

En todos estos escenarios es importante destacar un elemento: que las respuestas cumplan o no con una característica determinada, *variable Bernoulli.* Por ejemplo:

1. Los empleados esperan contar con un ingreso para la seguridad social cuando se jubilan: SÍ o NO.
2. Después de aplicar una serie de pruebas, el desempeño de un empleado fue: BUENO o MALO.
3. En el proceso de producción, los clientes evalúan un producto como: ADECUADO o IMPROPIO.
4. En la realización de un trámite, el resultado se puede clasificar como: EXITOSO o FALLIDO

En general, estas preguntas se pueden escribir utilizando la variable aleatoria X, como:

$$X = \begin{cases} 1 & \text{Si cumple con la característica} \\ 0 & \text{Si no cumple con la característica} \end{cases}$$

Considerando los valores de esta variable para representar a una población de N unidades se tiene la proporción de los que cumplen con la característica:

$$p = \frac{X}{N}$$

Una *población* consta de una colección de individuos u objetos en los que se observa una característica particular que será objeto de estudio. Para estudiar a la población se requiere tomar una muestra y calcular la proporción de la muestra:

$$\widehat{p} = \frac{X}{n}$$

Procedimiento y relación entre la población y muestra, probabilidad de seleccionar un elemento de la población, inferir usando la distribución de probabilidad del estadístico $\widehat{p}$.

Población	$\Longrightarrow$	Selección	$\Longrightarrow$	Muestra
$p = \frac{X}{N}$	$\Longleftarrow$	Inferencia	$\Longleftarrow$	$\widehat{p} = \frac{X}{n}$

De una muestra de n elementos se pregunta ¿cuál es la probabilidad de que X cumpla la característica?

Una descripción de posibles referencias de la característica sería:

p	Éxito	Ganó	Funciona	Bueno	Pasa	Presente	Si	Abierto
$q = 1 - p$	Fracaso	Perdió	Defectuoso	Malo	Falla	Ausente	No	Cerrado

Así, el lanzamiento de una moneda es un auxiliar para comprender los conceptos de la distribución de una variable aleatoria. En el caso que se trate de una moneda mexicana los resultados serían: águila (éxito) o cara (fracaso).

Apoyo tecnológico en distribuciones de probabilidad variable discreta

Con el propósito de conocer y estudiar las distribuciones, **CalEst** muestra varias distribuciones que son muy importantes en el estudio cuantitativo de las investigaciones. En la figura 5.2 se describe esta serie de distribuciones. En todas se **cuenta con un calculador estadístico** que resulta relevante puesto que le permite el cálculo de probabilidades, y de modo inverso dada una probabilidad se establece el valor de la variable. En este capítulo se expondrán las distribuciones de probabilidad Bernoulli, Binomial y Poisson, mientras que en el complemento didáctico se agrega una breve guía de trabajo.

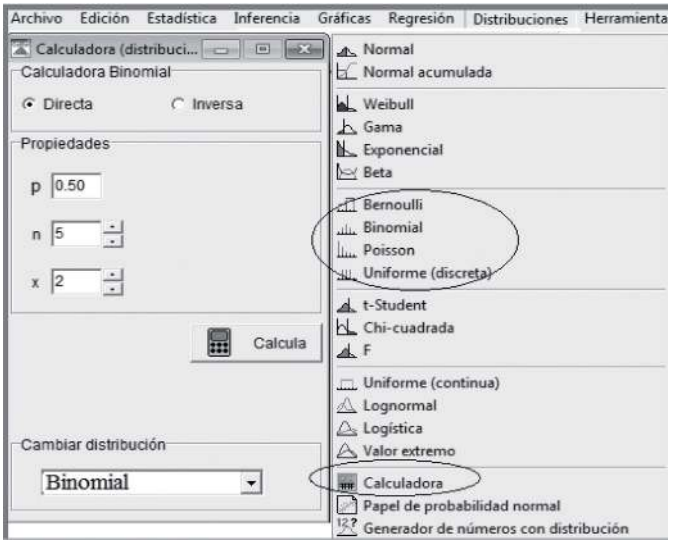


Figura 5.2 Galería de distribuciones de probabilidad disponibles en **CalEst**.

5.2 Distribuciones de probabilidad

Variables aleatorias

Los resultados de un experimento aleatorio pueden ser de dos tipos: cuantitativos o cualitativos. Por lo general, es más fácil tratar dichos resultados cuando se les asigna una medida numérica. Sin embargo, en

ocasiones los resultados no tienen necesariamente un valor numérico, por lo que en muchas encuestas los resultados se miden en una escala cualitativa. Por ejemplo, para conocer el efecto que produce tomar una medicina, las respuestas que se darían son: sin síntoma, náusea moderada, náusea severa, entre otras.

A los números se les puede registrar y contar; además su información puede resumirse con facilidad en un reporte. La regla que asigna números a los posibles resultados de un experimento se le conoce como variable aleatoria.

La variable aleatoria se denota con letras mayúsculas X y sus valores con minúsculas x.

Variables

Recordemos que las variables se dividieron en dos categorías: las cuantitativas y las cualitativas. Las cuantitativas o numéricas son discretas o continuas.

Por otro lado, una variable aleatoria es una regla que representa los posibles valores numéricos asociados con los resultados de un experimento.

Variable aleatoria discreta

Las variables aleatorias que se observan en la naturaleza poseen ciertas características y se pueden clasificar según su tipo, por ejemplo el número de insectos que se mueren al aplicar una dosis de un insecticida. Los valores de esa variables son: $X = 0, 1, 2, \ldots$ A esta variable se le clasifica como discreta. A continuación se estudiarán ese tipo de variables y se deducirán las probabilidades asociadas a sus posibles valores.

El mundo de la información 1. Examen de opción múltiple

Por lo general, se considera que los exámenes de opción múltiple son sencillos y a veces los alumnos los prefieren porque se contestan según el criterio. Inclusive, después de la aplicación de un examen de este tipo, es común escuchar que los estudiantes dicen que estuvo facilísimo. Sin embargo, también es usual que al recibir sus resultados se encuentren con la sorpresa de que a pesar de que parecía fácil el examen, la calificación obtenida no es tan buena como la que esperaban.

Preguntas sobre la naturaleza del problema

En este caso consideraremos un examen de tres preguntas las cuales se contestan al azar. ¿De cuántas maneras diferentes se puede contestar el examen? ¿Cuántas respuestas correctas se pueden obtener? ¿Cómo organizar la información según el número de respuestas correctas?

Información: Cada pregunta tiene dos respuestas: I (si se responde en forma incorrecta) y C (si se responde en forma correcta).

Análisis: Los sucesos elementales de este experimento, al contestar el examen, indican las diferentes respuestas al examen que se describen en la columna 1 de la tabla 5.1. El primer suceso elemental, III, indica que no hubo respuestas correctas. En la segunda columna de la tabla 5.1 se presenta el número de respuestas correctas.

Tabla 5.1 Resultados del número de respuestas correctas.

Resultado	Valor de X
III	0
IIC	1
ICI	1
CII	1
ICC	2
CIC	2
CCI	2
CCC	3

Esta información se interpreta de la siguiente manera: $X = 0$ indica que las tres preguntas se contestaron de manera incorrecta. Si $X = 1$ significa que una de las tres respuestas es correcta, y si $X = 2$ señala que dos respuestas fueron correctas y $X = 3$ que las tres respuestas son correctas; esta situación se resume en la tabla 5.2.

Tabla 5.2 X se asocia a un valor numérico con cada suceso elemental.

Valor numérico de X con un evento	Composición
$[X = 0]$	{III}
$[X = 1]$	{ICI, IIC, CIII}
$[X = 2]$	{ICC, CIC, CCI}
$[X = 3]$	{CCC}

Entonces X es la variable aleatoria que asocia el número de respuestas correctas, resultado del experimento. En este caso X puede tomar los valores de 0, 1, 2 o 3. La variable se define con el número de respuestas correctas que se obtienen en las tres preguntas. A este tipo de variables se le conoce como variable aleatoria discreta.

Distribución de probabilidad de una variable aleatoria discreta

Siguiendo con el problema de *El mundo de la información 1*, la información sobre los posibles números de respuestas correctas y sus probabilidades se recogen en la tabla 5.3. La probabilidad de obtener exactamente dos respuestas correctas ICC, CIC y CCI, es de $\frac{3}{8}$. De igual manera se obtienen las otras posibilidades.

Tabla 5.3 Probabilidades para el número de respuestas correctas.

Número de respuestas	0	1	2	3
Probabilidad	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Una variable aleatoria discreta es una variable numérica, que toma un número de valores que se pueden contar. La variable X representa el número de respuestas correctas; la tabla 5.3 describe los posibles valores que X puede tener, y la probabilidad de cada valor.

La notación que se usa para la distribución de una variable aleatoria es similar a la notación para la probabilidad de un evento. Por ejemplo $X = 0$ se puede definir como el evento $A =$\{ninguna respuesta es correcta\}. Por ejemplo, $X = 0$ correspondería a un evento (supongamos que es A), ninguna respuesta es correcta $P(A) = P(X = 0)$. En general, la probabilidad para la variable aleatoria X se expresa por:

$$p(x) = P(X = x) \text{ para todos los valores posibles de } X.$$

Ahora bien, retomando el problema se tiene que las probabilidades son:

$$p(0) = PX = 0) = \frac{1}{8}, p(1) = P(X = 1) = \frac{3}{8}, p(2) = P(X = 2) = \frac{3}{8}, p(3) = P(X = 3) = \frac{1}{8}$$

La gráfica (diagrama de barras) de esta distribución de probabilidad se describe en la figura 5.3.

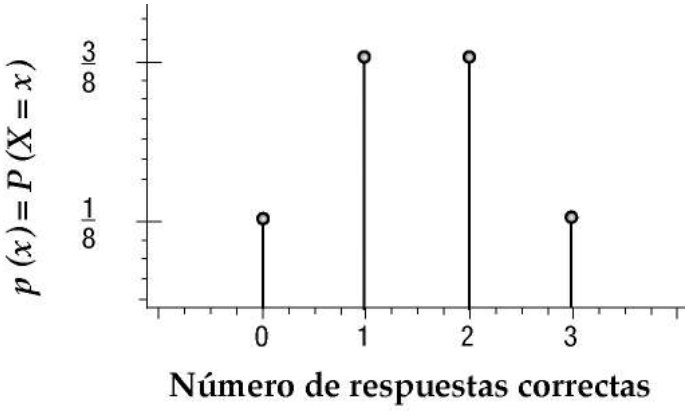


Figura 5.3 Distribución de probabilidad para el número de respuestas correctas.

Ejemplo 5.1

Este ejemplo pretende relacionar los hechos históricos con las fechas en que éstos ocurrieron. En muchas ocasiones nos preguntan algo sorpresivamente y debemos responder sobre sucesos que ocurrieron en la cultura, el deporte, la historia, entre otros, y contestamos por pura intuición. A una persona joven se le pide que relacione tres hechos (la expropiación petrolera en México, el fin de la segunda guerra mundial, y el nacimiento de John Lennon) con tres fechas (1940, 1938, 1945). Si el joven adivina las respuestas correctas, ¿cuál es la distribución de probabilidad del número de respuestas que tendría correctas?

Una tabla o función que liste todos los posibles valores de una variable aleatoria discreta y su probabilidad asociada, se conocen como distribución de probabilidad de la variable aleatoria.

Solución

El espacio muestral M para la situación descrita corresponde a las siguientes seis permutaciones de los tres datos. Denotaremos con la letra "e" el resultado elemental,

$e_1 = 1938, 1940, 1945$	$e_2 = 1938, 1945, 1940$
$e_3 = 1940, 1938, 1945$	$e_4 = 1940, 1945, 1938$
$e_5 = 1945, 1938, 1940$	$e_6 = 1945, 1940, 1938$

Si la respuesta fue estrictamente adivinando, entonces la probabilidad de cada permutación es $\frac{1}{6}$.

Enseguida, se asocia cada elemento de M al número de respuestas correctas X. Para facilitar este procedimiento ordenamos los hechos históricos: (1) la expropiación petrolera, (2) nacimiento de John Lennon y (3) fin de la segunda guerra. Si el joven contestó e_1 tiene las tres respuestas correctas. Si respondió e_2, e_3 o e_6 tiene una respuesta correcta. Finalmente, si contestó e_4 o e_5 ninguna respuesta fue correcta; véase la tabla 5.4. En la tabla de la figura 5.4 se detalla la función de probabilidad de la variable aleatoria X: número de respuestas correctas.

Tabla 5.4 Parejas de hechos históricos con fechas.

Permutación	e_1	e_2	e_3	e_4	e_5	e_6
Probabilidad	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
No. de respuestas correctas X	3	1	1	0	0	1

Número de respuestas correctas	0	1	3
Probabilidad	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{1}{6}$

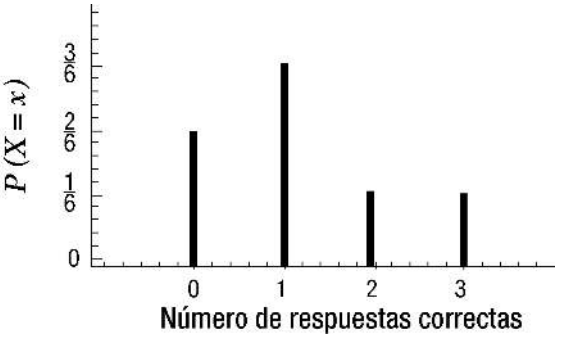


Figura 5.4 Distribución de probabilidad para el ejemplo de hechos históricos.

Ejemplo 5.2

El número de partidos de fútbol que los estudiantes de preparatoria ven a la semana en la televisión se denota con X. En la tabla 5.5 se presenta la distribución de frecuencias de una encuesta realizada a 500 estudiantes. Se aproxima la distribución de probabilidad de la variable aleatoria X.

Tabla 5.5 Distribución de frecuencias del número X de partidos de fútbol.

Partidos de fútbol (x)	0	1	2	3	4	Total
Frecuencia	75	190	135	70	30	500
Frecuencia relativa	0.15	0.38	0.27	0.14	0.06	1.0

Solución

Considerando la frecuencia relativa como una estimación empírica de las probabilidades, esencialmente se ha obtenido una aproximación de la distribución de probabilidad de X. Se tendría la verdadera distribución de probabilidad si se escogiera una muestra de estudiantes muy grande de la preparatoria, idealmente la población.

Complemento técnico: Con base en los ejemplos discutidos, se puede presentar los resultados en forma general usando notación $x_1, x_2, \ldots, x_k$ para designar los distintos valores de la variable aleatoria X. La probabilidad de que un valor particular x_i ocurra se denotará con $p(x)$. La distribución de probabilidad de X se describe como se muestra en la tabla 5.6.

Tabla 5.6 Distribución de probabilidad para la variable aleatoria discreta X.

Valores x	x_1	x_2	...	x_k	Total
Probabilidad	$p(x_1) =$	$p(x_2) =$		$p(x_k) =$	1.0
	$P(X = x_1)$	$P(X = x_1)$		$P(X = x_1)$	

Puesto que $p(x_1)$ representa probabilidades, sus valores para cada i deben estar en 0 y 1, es decir:

$$0 \leq p(x_1) \leq 1.$$

La distribución de probabilidad de la variable aleatoria discreta X se describe mediante la expresión:

$$p(x_i) = P(X = x_i)$$

la cual indica la probabilidad para cada valor y satisface las siguientes condiciones:

1. $p(x_i) \geq 0$, para cada valor x_i de X.
2. $0 \leq p(x_i) \leq 1$
3. $\sum_{i=1}^{k} p(x_i) = 1$

El mundo de la información 2. Germinación

El amaranto fue uno de los principales alimentos de los antiguos pueblos mesoamericanos. Cuando los españoles llegaron en el siglo XVI, el cultivo de amaranto casi fue eliminado. En la actualidad se considera un cultivo prometedor debido a sus características agroclimáticas. Un bioquímico estudió una variedad de estas semillas para evaluar su tolerancia en condiciones ambientales adversas. Usó un papel de filtro húmedo para poner a germinar 10 semillas de amaranto en una charola. En el experimento se emplearon 80 charolas para ver si existe una alta probabilidad de germinación.

Preguntas sobre la naturaleza del problema

La variable aleatoria X es el número de semillas que germinarán en una muestra de 80 charolas; en cada una se pusieron 10 semillas. X es una variable aleatoria discreta que toma los valores 0, 1, 2,. . . , 10. Si el evento la charola seleccionada no tiene semilla germinada se expresa como $X = 0$, es decir, el número de semillas germinadas es igual a 0. Será de interés estudiar la distribución de probabilidad para establecer si la probabilidad de germinación es alta.

Datos: El número de semillas germinadas se anota en la siguiente tabla.

Tabla 5.7 Distribución de la variable número de semillas germinadas.

x	0	1	2	3	4	5	6	7	8	9	10
Frecuencia	0	1	1	4	3	2	5	4	32	18	10

La probabilidad de que una semilla germine al seleccionar al azar una charola 1 es 0.0125, es decir, $p(1) = P(X = 1) = 0.0125$. En la tabla 5.8 se observan las probabilidades para los otros valores de la variable X.

Tabla 5.8 Cálculo de la probabilidad para la germinación.

$p(0) = P(X = 0) = 0$	$p(6) = P(X = 6) = 0.0625$
$p(1) = P(X = 1) = 0.0125$	$p(7) = P(X = 7) = 0.050$
$p(2) = P(X = 2) = 0.0125$	$p(8) = P(X = 8) = 0.400$
$p(3) = P(X = 3) = 0.005$	$p(9) = P(X = 9) = 0.225$
$p(4) = P(X = 4) = 0.0375$	$p(10) = P(X = 10) = 0.125$
$p(5) = P(X = 5) = 0.025$	

En resumen en la tabla 5.9 se muestra la distribución de probabilidad para la variable aleatoria X: número de semillas germinadas.

Tabla 5.9 Distribución de probabilidad para la germinación[1]

x	0	1	2	3	4	5	6	7	8	9	10
$p(x)^*$	0	0.013	0.013	0.05	0.038	0.025	0.063	0.05	0.4	0.225	0.125

En la figura 5.5 se presenta la distribución de probabilidad para la germinación. La intención del bioquímico es evaluar las condiciones adversas a que se sometieron unas semillas para su germinación. Por ello, está interesado en evaluar la probabilidad de que la geminación sea mayor que 7. En la tabla 5.9 puede verse que la probabilidad de que germinen más de 7 semillas corresponde a los eventos donde la variable aleatoria es $X = 8, X = 9$ o $X = 10$. Estos eventos son mutuamente excluyentes, por lo tanto la probabilidad de que sea mayor que 7 equivale a:

$$\begin{aligned} P(X > 7) &= P\left((X = 8) \cup (X = 9) \cup) (X = 10)\right) \\ &= p(8) + P(9) + p(10) = 0.4 + 0.225 + 0.125 = 0.75 \end{aligned}$$

Interpretación: Se tiene una probabilidad "alta", por lo que se dice que el proyecto del bioquímico puede considerarse exitoso.

[1] Probabilidad redondeada a milésimas.

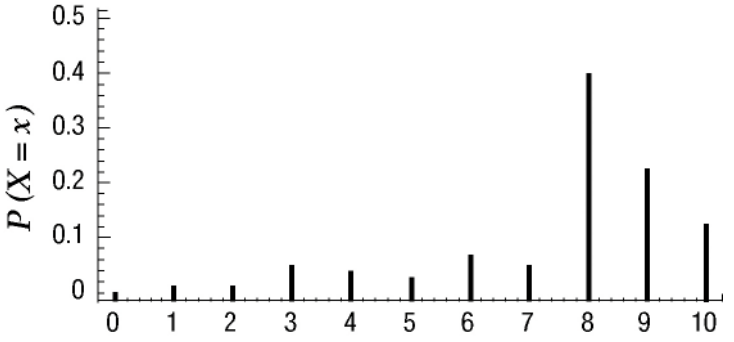


Figura 5.5 Distribución de probabilidad para el número de semillas germinadas.

Notación para el cálculo de probabilidades: En el problema que acabamos de ver se puede advertir que se calculó una probabilidad $P(X > 7)$ y se evaluó ésta indicando el conjunto de sus posibles valores de manera clara. A menudo, estamos interesados en el procedimiento para calcular probabilidades en que la variable aleatoria toma un valor. Cuando se realiza este tipo de cálculos se utiliza un lenguaje con el que debemos familiarizarnos, por lo cual decimos que X es:

"al menos x",	"menos de x",
"más que x o mayor que x",	"entre x_1 y x_2" o
"a lo más x",	"entre x_1 y x_2 inclusive"

En la tabla 5.10 se resume esta notación y se indica el significado y su relación con la probabilidad.

Tabla 5.10 Notación que resume el cálculo para encontrar probabilidades.

	Interpretación	Notación
La probabilidad de que	Todos los valores de la variable	
X tome un valor que es	aleatoria	
Al menos x	que son los valores de x o mayores que x	$P(X \leq x)$
Más que x	que son mayores que el valor x	$P(X > x)$
A lo más x	que son los valores de x o menores que x	$P(X \geq x)$
Menos de x	que son menores que el valor x	$P(X < x)$
Entre x_1 y x_2	que son mayores que los valores x_1 y menores que los valores x_2	$P(x_1 < X < x_2)$
Entre x_1 y x_2 y que les incluye	que inicia con el valor x_1 y termina con el valor x_2 incluyendo los valores x_1 y x_2	$P(x_1 \leq X \leq x_2)$

Ejemplo 5.3

A un grupo de personas entre 15 y 30 años de edad se les pregunta si les gustan las películas de terror. La respuesta es "sí" o "no". El cálculo de probabilidad reportada para este estudio se muestra en la tabla 5.11; la correspondiente gráfica de esta distribución de probabilidad se describe en la figura 5.6 y es la probabilidad de que exactamente a 4 personas les guste una película de terror. Es de interés calcular la probabilidad de que entre 4 y 8 personas inclusive tengan gusto por una película de terror, ya que si ésta resulta alta querrá decir que un buen porcentaje de personas entre 15 y 30 años verán películas de ese género.

Tabla 5.11 Distribución de probabilidad del gusto por las películas de terror.

$p(0) = PX = 0) = 0$	$p(6) = P(X = 6) = 0.226$
$p(1) = P(X = 1) = 0.002$	$p(7) = PX = 7) = 0.194$
$p(2) = P(X = 2) = 0.016$	$p(8) = P(X = 8) = 0.121$
$p(3) = P(X = 3) = 0.054$	$p(9) = P(X = 9) = 0.054$
$p(4) = P(X = 4) = 0.121$	$p(10) = P(X = 10) = 0.016$
$p(5) = P(X = 5) = 0.194$	$p(11) = P(X = 11) = 0.002$
	$p(12) = P(X = 12) = 0.0$

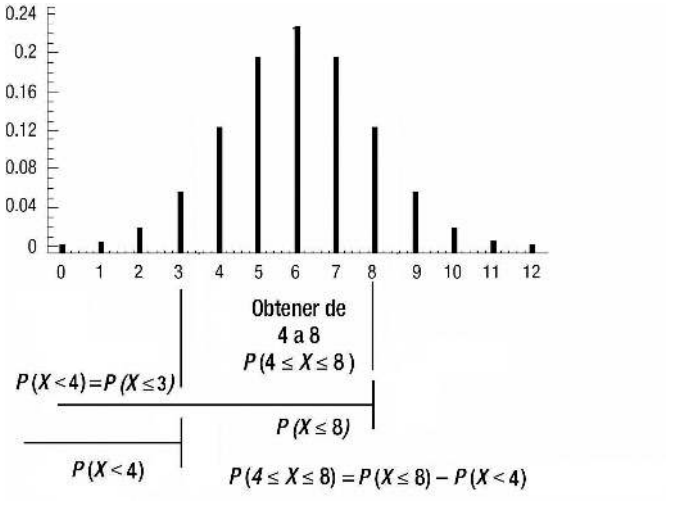


Figura 5.6 Cálculo de probabilidades a partir de la distribución de probabilidad.

Solución

Recurrimos a la tabla 5.11 y a la figura 5.6 para mostrar el cálculo correspondiente, es decir:

$$P(4 \leq X \leq 8) = P(X \leq 8) - P(X < 4)$$
$$P(X \leq 8) = 0.002 + 0.016 + 0.054 + 0.121 + 0.194 + 0.226 + 0.194 + 0.121$$
$$= 0.928$$
$$P(X < 4) = P(X \leq 3) = 0.002 + 0.016 + 0.054 = 0.072$$
$$P(X \leq 8) - P(X < 4) = 0.928 - 0.072 = 0.856$$

Función de distribución acumulada de una variable aleatoria discreta

En el capítulo 2 se vio el polígono de frecuencias acumuladas, y de manera similar las probabilidades asociadas pueden expresarse con la variable aleatoria. Éstas se acumulan y su resultado es la función de distribución de probabilidad acumulada (o fdpa). La fdpa recolecta todas las probabilidades de valores menores o iguales que el valor de la variable X. La notación para la fdpa es $P(X \leq x)$. La función de probabilidad acumulada crece de 0 a 1 tal y como x va creciendo, vea la figura 5.7.

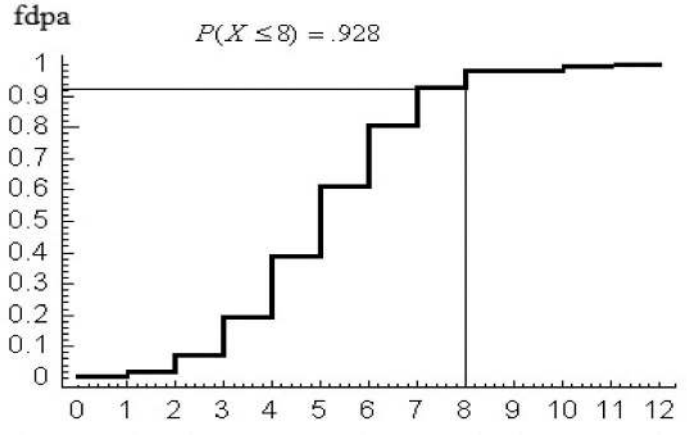


Figura 5.7 Función de distribución de probabilidad acumulada.

La gráfica de una fdpa se parece a una escalera. La altura de cada escalón es igual a la probabilidad asociada con el valor de x en ese escalón.

Ejemplo 5.4

Encontrar probabilidades a partir de una distribución de probabilidad. El número de horas diarias que estudian los jóvenes de una preparatoria a la semana es una variable aleatoria X cuya distribución de probabilidad es la que se reporta en la tabla 5.12.

Tabla 5.12 Distribución de probabilidad número de horas de estudio diarias.

Variable X	×	0	1	2	3	4	5
Probabilidad	$p(x)$	0.10	0.12	0.25	0.30	0.20	0.03
Probabilidad acumulada	$P(X \leq x)$	0.10	0.22	0.47	0.77	0.97	1

Un sociólogo está interesado en conocer la probabilidad de que un joven estudie exactamente dos horas. Ésta podemos encontrarla en la tabla anterior:

$$p(2) = P(X = 2) = 0.25$$

1. ¿Cuál es la probabilidad de que un joven estudie dos o tres horas cualquier día de la semana?
2. ¿Cuál es la probabilidad de que un joven estudie al menos tres horas diarias?

Solución

1. Primero debemos reconocer que se quiere encontrar la probabilidad $P\left((X = 2) \cup (X = 3)\right)$. Puesto que los valores de las variables aleatorias son los eventos $A = \{X = 2\}$ y $B = \{X = 3\}$ y como los eventos A y B son mutuamente excluyentes, entonces simplemente sumamos las probabilidades:

   $$P\left((X = 2) \cup (X = 3)\right) = P\left(X = 2\right) + P\left(X = 3\right) = p(2) + p(3) = 0.25 + 0.30 = 0.55$$

   Con esto se interpretaría que de 100 jóvenes 55 estudian dos o tres horas diarias.

2. El sociólogo considera que estudiar al menos tres horas diarias debe dar resultados redituables a los jóvenes; entonces:

   $$\begin{aligned} P(X &\geq 3) = P\left(X = 3\right) \cup \left(X = 4\right) \cup \left(X = 5\right)) \\ &= p(3) + p(4) + p(5) = 0.30 + 0.20 + 0.03 = 0.53 \end{aligned}$$

Se pueden tener bajas calificaciones si se estudian menos de tres horas diarias. La probabilidad: $P(X < 3)$. El evento en que la variable X sea menor que 3 es complemento del evento en el que $X \geq 3$, por lo tanto,

$$P(X < 3) = 1 - P(X \leq 3) = 1 - 0.53 = 0.47$$

Función distribución de probabilidad acumulada

La función distribución de probabilidad acumulada (fdpa) de una variable aleatoria X es la que da la probabilidad de que X sea menor o igual que x, y su notación es $P(X \leq x)$.

Ejemplo 5.5

Interpretación de la distribución de probabilidad discreta en términos de porcentaje. Cuando una persona ingresa a un hospital para una cirugía, el número de días que permanece hospitalizada dependerá de varias circunstancias. No se sabe exactamente en cuánto tiempo se recuperará la persona. En la tabla 5.13 se registra el resumen sobre el número de días que 285 personas estuvieron hospitalizadas. En el segundo renglón se da la probabilidad de las personas según el periodo de estancia y ésta se puede interpretar como el porcentaje de pacientes que se quedan en el hospital X días.

Por ejemplo, el porcentaje de pacientes que se quedarán siete días en el hospital es[2]:

$$86/285 = 0.302 \ (30.2\,\%).$$

Tabla 5.13 Distribución de probabilidad estancia en un hospital.

Días de estancia	x	4	5	6	7	8	9	10
Frecuencia		12	34	120	86	24	7	2
Probabilidad	$p(x)$	0.042	0.119	0.421	0.302	0.084	0.025	0.007
Prob. A.	$P(X \leq x)$	0.042	0.161	0.582	0.884	0.968	0.993	1

Estos porcentajes se tratarán como la distribución de la probabilidad para una variable aleatoria X, donde X es el número de días en que un paciente permaneció en el hospital, el último renglón de la tabla 5.13: Prob. A., representa la probabilidad acumulada. Así:

$$P(X = 7) = 0.302$$

1. ¿Cuál es la probabilidad de estar más de una semana (8 días o más)?

[2]Las proporciones se redondearon a milésimas.

2. ¿La probabilidad de estar menos de 6 días?

3. ¿Cuál es la probabilidad de estar entre 5 y 7 días inclusive?

Solución

$$a)\ P(X \geq 8) = p(8) + p(9) + p(10) = 0.084 + 0.025 + 0.007 = 0.116$$

En términos de porcentajes el valor 0.116 (probabilidad) se interpreta como el 11.6 % de pacientes que están 8 días o más.

$$b)\ P(X < 6) = P(X \leq 5) = p(4) + p(5) = 0.042 + 0.119 = 0.161$$

De manera análoga 16.1 % de pacientes están menos de seis días.

$$c)\ P(5 \leq X \leq 7) = p(5) + p(6) + p(7) = 0.119 + 0.421 + 0.302 = 0.842$$

Observe que el 84.2 % de pacientes permaneció entre 5 y 7 días. Esto indica que la mayoría está entre esos días.

Valor esperado y desviación estándar

El mundo de la información 3. Número de materias aprobadas

La administración de una escuela de nivel secundaria lleva el registro del número de materias que reprobaron los estudiantes de la última generación.

Preguntas sobre la naturaleza del problema

¿Cuál es la distribución de probabilidad del número de materias reprobadas? ¿Cuál es la media del número de materias reprobadas? ¿Cuál es la desviación estándar para esta distribución de probabilidad?

Datos: El registro que ha realizado acerca del número de materias reprobadas se presenta en la tabla 5.14.

Se tiene el interés por conocer la media para el número de materias reprobadas. Como recordará en el Capítulo 3, se vieron las medidas de tendencia central de un conjunto de datos (como la media), las medidas de dispersión, como la varianza y desviación estándar. Para el caso de la distribución de probabilidad la media se verá como un parámetro y se denota por la letra griega μ. En este caso la media de la población es una media ponderada y se obtiene sumando los valores que resultan de la

multiplicación de cada valor de X por su probabilidad. En el último renglón de la tabla 5.14 se tiene la multiplicación indicada y la suma es la media:

$$\mu = 0.10 + 0.40 + 0.54 + 0.88 + 0.50 + 0.48 + 0.21 + 0.32 = 3.43$$

El valor del parámetro μ es 3.43 de materias reprobadas y representa la media del número de materias reprobadas en esa generación. Este parámetro es una medida del centro de la distribución de probabilidad.

Tabla 5.14 Distribución de probabilidad del número de materias reprobadas.

x	0	1	2	3	4	5	6	7	8
$p(x)$	0.05	0.10	0.20	0.18	0.22	0.10	0.08	0.03	0.04
$xp(x)$	0	0.10	0.40	0.54	0.88	0.50	0.48	0.21	0.32

Valor esperado

Parámetro es un valor numérico que describe una característica de la población.

La media o valor esperado de la variable aleatoria X es:

$$\mu = E(X) = \sum(\text{valor}) \times \text{probabilidad} = \sum_{i=1}^{k} x_i p(x_i) \qquad (5.1)$$

La suma es sobre todos los valores distintos x_i de X

Complemento técnico: A la media de una variable aleatoria discreta X se le llama valor esperado y alternativamente se denota con $E(X)$. Así la media μ y el valor esperado $E(X)$ son la misma cantidad. Al valor esperado también se le conoce como esperanza matemática o simplemente esperanza.

Ejemplo 5.6

El costo para entrar a participar en un juego de azar es de 15 pesos para cada partida. El jugador gana 10, 20 o 30 pesos con las respectivas probabilidades de 0.6, 0.3 y 0.1. ¿Es éste un juego atractivo para ganar?

Solución

Se efectúa un análisis intuitivo: supongamos que el jugador interviene en 100 partidas con la oportunidad de ganar 10 pesos en 60 de los 100 juegos. Similarmente, esperaría 20 pesos en 30 juegos y 30 pesos en 10 juegos. La ganancia total en pesos es de $600 + 600 + 300 = 1500$ pesos; la media de ganancia es de

$1500/100 = 15$ pesos por juego. El juego parece justo porque después de muchos juegos la ganancia final es la misma que la cuota de entrada. En la Tabla 5.15 se reproduce esta situación en términos de la distribución de probabilidad.

La ganancia es la suma del último renglón de la Tabla 5.15 y la media se obtiene mediante la siguiente operación: $suma/100 = (10 \times 0.6 + 20 \times 0.3 + 30 \times 0.1)/100 = 0.15$.

Observemos que si este juego se reproduce un número muy grande de veces, la media de la ganancia será 15.

Tabla 5.15 Análisis para la oportunidad de ganar.

Precio	x	10	20	30
Probabilidad	$p(x)$	0.60	0.30	0.10
No. de juegos ganados	$xp(x)$	0.6×100	0.3×100	0.1×100
Ganancias		$10 \times 0.6 \times 100$	$20 \times 0.3 \times 100$	$30 \times 0.1 \times 100$

Varianza de una distribución de probabilidad

El concepto de valor esperado permite medir de manera numérica la dispersión de una distribución de probabilidad, es decir, la desviación estándar. Para obtener el valor de esta medida se hará un razonamiento similar al que se realizó en el capítulo 3 para calcular, primero, la varianza y luego la desviación estándar.

Puesto que μ es el centro de la distribución de X, se expresa la variación de X en términos de la desviación $X - \mu$. Se define la varianza de X como el valor esperado de la desviación al cuadrado: $(X - \mu)^2$. El cálculo del valor esperado de $(X - \mu)^2$ se obtiene multiplicando cada valor de $(x_i - \mu)^2$ por la probabilidad $p(x_i)$ (tabla 5.16), y luego se suman estos productos.

Así la varianza se define por:

$$\text{Varianza de } X = \sum \text{ (desviación)}^2 \times \text{ (probabilidad)} = \sum_{i=1}^{k} = (x_i - \mu)^2 p(x_i)$$

En resumen:

La varianza de la variable aleatoria X es:

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)$$

La desviación estándar de la variable aleatoria X es:

$$\sigma = DE(X) = \sqrt{Var(X)}$$

Tabla 5.16 Desviación y probabilidad para la variable aleatoria discreta X.

Valores de x	$(x_1 - \mu)^2$	$(x_2 - \mu)^2 \ldots$	$(x_k - \mu)^2$
Probabilidad	$p(x_1)$	$p(x_2)$	$(p(x_k)$

Ejemplo 5.7

En la actualidad existen normas para proteger a la ciudadanía contra problemas de incendio o fenómenos naturales. Una de las medidas que establecen las normas es realizar pruebas de evacuación. Una ciudad lleva a cabo varias pruebas de este tipo. Se observa que el tiempo que requerirán para desalojar a las personas en las oficinas de gobierno está entre 12 y 17 minutos con las probabilidades que se muestran en la tabla 5.17.

1. Usando las fórmulas indicadas, calcule la media y la desviación estándar de la distribución de probabilidad.

2. Mediante el paquete estadístico resuelva el inciso a.

Solución

1. Se calcula el valor esperado mediante la fórmula $\mu = \sum_{i=1}^{n} x_i p(x_i)$. En el tercer renglón de la tabla 5.17 se ha realizado el producto y en la última columna se hace la suma. Esto es: la media o valor esperado es:

$$\mu = \sum_{i=1}^{n} x_i p(x_{i)} = 14.11$$

   De manera análoga se aplica la fórmula para la desviación estándar y se obtiene:

$$\sigma = \sqrt{\mu = \sum_{i=1}^{n} (x_i - 14.11)^2 p(x_i)} = \sqrt{1.38} = 1.17$$

   Los cálculos se reproducen en la tabla 5.17.

2. Ahora se resolverá el problema utilizando **CalEst**. La salida se muestra en la figura 5.8.

Tabla 17 Distribución de probabilidad para el tiempo de desalojo en las oficinas de gobierno.

x	12	13	14	15	16	17	Suma
$p(x)$	0.06	0.25	0.38	0.18	0.09	0.04	
$xp(x)$	0.72	3.25	5.32	2.7	1.44	0.68	14.11
$(x-\mu)^2p(x)$	0.27	0.31	0.01	0.14	0.32	0.33	1.38

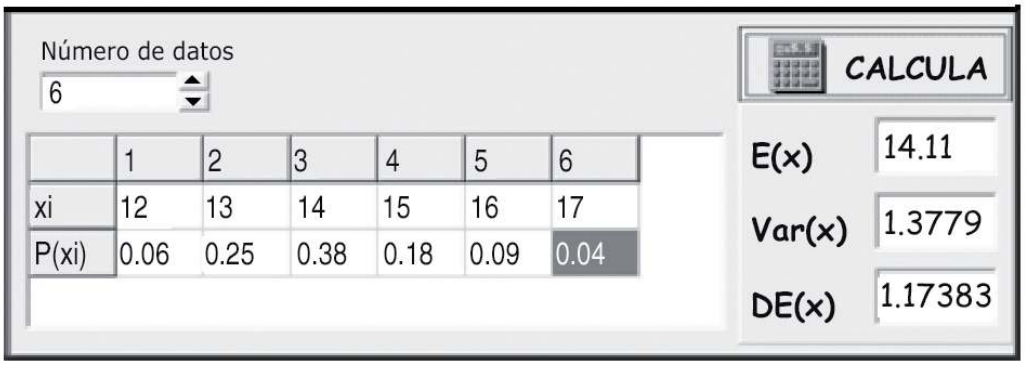


Figura 5.8 Cálculo del valor esperado, la varianza y la desviación estándar de la distribución de probabilidad.

5.3 Distribución binomial

La distribución binomial se genera mediante una serie de procesos Bernoulli. Por esa razón este apartado se inicia con los conceptos básicos de este tipo de distribución.

Distribución Bernoulli

Los ensayos Bernoulli se definen como una acción en que se tiene dos posibles resultados mutuamente excluyentes. Por lo general, estos resultados se refieren como éxito y como fracaso.

Las probabilidades respectivas son p para éxito y $(1-p)$ para fracaso. Entonces la variable aleatoria (número de éxitos Bernoulli) tiene una distribución Bernoulli con parámetro p. En la tabla 5.18 se describe lo que se comprende por un ensayo Bernoulli.

Tabla 5.18 Número de éxitos en un ensayo Bernoulli.

Número de éxito	0	1
Probabilidad	$(1-p)$	p

El siguiente ejemplo muestra el procedimiento para obtener una distribución binomial.

La distribución de Bernoulli se expresa como una función de probabilidad $P(x)$ como sigue:

$$P(x) = p^x (1-p)^{1-x} \quad \text{para } x = 0, 1$$

La expresión $P(x)$ es la probabilidad de tener x éxitos en una prueba Bernoulli. Claramente x puede ser 0 o 1, así:

$$P(x) = \begin{cases} p & \text{para } x = 1 \\ 1-p & \text{para } x = 0 \end{cases}$$

La media y la varianza para la distribución Bernoulli son:

$$\text{media } \mu = p \quad \text{varianza } \sigma^2 = p(1-p)$$

Jakob Bernoulli (1654-1705)

Nació el 27 de diciembre de 1654. Uno de los trabajos más importantes de Jacob Bernoulli fue *The Art of Conjeture,* se trata de una obra póstuma que su sobrino edita en 1713, un documento de la mayor importancia dentro de la teoría de probabilidades. De éste se deriva su trabajo sobre el estudio de la distribución binomial y la ley de los grandes números. Por cierto Jakob fue hermano del también matemático Johann Bernoulli y tío de Nicolás Bernoulli. Se sabe que estudió filosofía y teología en la Universidad de Basilea, alternativamente se educó, por su entusiasmo, en matemáticas y astronomía. A partir de 1683 regresó a esta universidad como profesor de mecánica. Murió el 16 de agosto del 1705 en Basilea.

El mundo de la información 4 Información sobre el SIDA

El síndrome de inmunodeficiencia adquirida (SIDA) es una condición deficiente del sistema inmunológico humano (el sistema encargado de defender al cuerpo del ataque de enfermedades) causada por un virus identificado como virus de inmunodeficiencia humana: VHI. Aunque existe información sobre este tema,

aparentemente muy pocas personas están enteradas.

Se preguntó a un grupo de jóvenes si sabían qué era el SIDA. La respuesta era "sí" o "no". La variable de respuestas se definió como el número de jóvenes que están informados sobre el SIDA. Claramente la variable aleatoria toma únicamente los valores $0, 1, 2, \ldots, n$, donde n es el número de jóvenes en la muestra. Por lo tanto X es una variable aleatoria.

Preguntas sobre la naturaleza del problema

Se quiere saber qué tan bien están informados los jóvenes sobre el tema del sida , pero también se quiere averiguar cuál es el modelo de probabilidad para X en este estudio.

Antes de generar los datos para estudiar la distribución de probabilidad binomial se presentan algunas propiedades referentes a esta distribución.

Variable aleatoria binomial

El estudio se plantea desde el punto de vista experimental, el cual consiste en n ensayos o pruebas independientes.

1. Las pruebas son repeticiones en condiciones idénticas. Por ejemplo, en el estudio acerca de la información sobre el SIDA, cada prueba es la respuesta de un joven y la respuesta se supone que es independiente, esto es, la respuesta de cada joven no está influenciada por la respuesta de otro.
2. Cada joven da una respuesta dicotómica, esto es, "sí" o "no".
3. La probabilidad de obtener un "sí " como respuesta en una muestra seleccionada de manera aleatoria no cambia.

La distribución binomial se usa como modelo de probabilidad para X en el problema del SIDA. Una variable aleatoria cuya distribución es binomial se llama variable aleatoria binomial.

Características que definen una variable binomial

1. Existe un número n fijo de ensayos Bernoulli.
2. Cada ensayo tiene un resultado de dos posibilidades, las cuales se definen como éxito o como fracaso.
3. La probabilidad p de éxito en una prueba es constante.
4. Los ensayos son independientes, es decir, la probabilidad de éxito en cualquiera de los ensayos no se ve afectada por el resultado de un ensayo previo.

Desarrollo del modelo de probabilidad para un experimento binomial

En esta parte se presentan los datos que se obtienen para la información sobre el sida. La muestra para el estudio quizá incluya a muchos jóvenes, por ejemplo 500. Sin embargo, para simplificar la presentación se considera la respuesta de tan sólo cuatro jóvenes, $n = 4$.

Se ha indicado que X es la variable aleatoria y representa el número de respuestas afirmativas entre estos 4 jóvenes. Los posibles valores de X son 0, 1, 2, 3 y 4.

Supongamos, como conocimiento previo, que el porcentaje de jóvenes que están informados sobre el SIDA es 25 %. Entonces, la probabilidad de que un joven responda "sí" es:

$$p = P(\text{sí}) = 0.25$$

Observaciones

1. Usualmente no se conoce el valor de p.

2. El hecho de que un joven responda "sí", diremos que es el caso de ÉXITO, y "no" el caso de FRACASO.

Si la probabilidad de que el primer joven entrevistado diga "sí" es $p = 0.25$ y la probabilidad de que diga "no " es:

$$P(\text{no}) = 1 - P(\text{sí}) = 1 - p = 1 - 0.25 = 0.75$$

Por lo general, se denota $1 - p$ o por q, es decir: $q = 1 - p$, entonces:

$$q = P(\text{no}) = 0.75$$

Dado que se ha supuesto independencia entre las respuestas, recuerde que p y q son las probabilidades de éxito y fracaso respectivamente.

El cálculo de la probabilidad se hace aplicando la ley de la multiplicación. Por ejemplo, razonemos el caso en el que dos jóvenes respondan: el primero contesta "no" y el segundo "sí ". Por lo tanto, la probabilidad en este caso se plantea con la expresión:

$$P(\text{no y sí}) = P(\text{no}) \times P(\text{sí}) = q \times p = (0.75)(0.25) = 0.188$$

Pensemos el caso de $n = 4$, suponemos por ejemplo, que los 4 entrevistados respondieron {no, sí, no, no}. La probabilidad en esta situación es:

$$P(\text{no, sí, no, no}) = q \times p \times q \times q = p \times q^3 = (0.25)(0.75)^3 = 0.105$$

La variable de respuesta en este caso es $X = 1$, puesto que sólo un joven dijo "sí".

¿Cuántas maneras más existen para obtener $X = 1$ o un "sí" y tres respuestas "no"? Hay otras 3 situaciones donde un entrevistado puede responder "sí "; a saber:

sí, no, no, no
no, no, sí, no
no, no, no, sí

Las probabilidades para cada caso son:

$$\begin{aligned} P(\text{sí, no, no, no}) &= p \times q \times q \times q = p \times q^3 = (0.25)(0.75)^3 = 0.105 \\ P(\text{no, no, sí, no}) &= q \times q \times p \times q = p \times q^3 = (0.25)(0.75)^3 = 0.105 \\ P(\text{no, no, no, sí}) &= q \times q \times q \times p = p \times q^3 = (0.25)(0.75)^3 = 0.105 \end{aligned}$$

En total, existen cuatro maneras diferentes de obtener exactamente una respuesta "sí". De modo que la probabilidad de $X = 1(P(X = 1))$ es:

$$P(x = 1) = 4p \times q^3 = 4(0.25)(0.75)^3 = 0.422$$

El modelo de probabilidad binomial

Si la variable aleatoria satisface los requisitos de una variable aleatoria binomial, ésta puede tomar uno de los posibles valores $x = 0, 1, 2, \ldots, n$, (número de pruebas). La probabilidad asociada con cada posible valor se denota con $p(x)$ y está dada por:

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{para } x = 0, 1, 2, ..., n \tag{5.2a}$$

donde al término $\binom{n}{x}$ se le conoce como coeficiente binomial. Ello es así porque es la combinación de n pruebas seleccionadas x veces, y resulta una técnica útil de conteo.

Regla de combinación (combinatoria)

El número de maneras en que x éxitos pueden ocurrir en n pruebas se le conoce por la combinación de n pruebas seleccionadas x veces, y se calcula como sigue: el número $n!$, o n factorial, representa el producto.

$$n! = n(n-1)(n-2)...(2)(1)$$

Observación: n es entero y 0! se define como $1(0! = 1)$.

Ensayos Bernoulli

Si se realiza una serie de n ensayos Bernoulli independientes, entonces *la variable aleatoria número de éxitos* en n ensayos tiene una *distribución Binomial con parámetros* n y p. Una descripción de esta distribución es $B(n,p)$. Donde $\binom{n}{x}=\frac{n!}{x!(n-x)!}$.

Como se ve, la fórmula de combinación es útil porque se usa con el modelo binomial para contar el número de maneras en las que exactamente x éxitos pueden ocurrir en n pruebas de una variable aleatoria. En la tabla 5.19 están los datos del ejemplo.

Tabla 5.19 Uso del modelo binominal.

Éxitos	Entrevistados y respuestas	Combinación	Probabilidad
0	No No No No		$p(0) = 1 \times (0.25)^0(0.75)^4 = 0.316$
	Sí No No No		
1	No Sí No No		$p(1) = 4 \times (0.25)^1(0.75)^3 = 0.412$
	No No Sí No		
	No No No Sí		
	Sí Sí No No		
	Sí No Sí No		
2	Sí No No Sí		$p(2) = 6 \times (0.25)^2(0.75)^2 = 0.211$
	No Sí Sí No		
	No Sí No Sí		
	No No Sí Sí		
	Sí Sí Sí No		
3	Sí Sí No Sí		$p(3) = 4 \times (0.25)^3(0.75)^1 = 0.047$
	Sí No SíSí		
	No Sí Sí Sí		
4	Sí Sí Sí Si		$p(4) = 1 \times (0.25)^4(0.75)^0 = 0.004$

Modelo binomial en el problema de información sobre el SIDA

Aquí describiremos las cinco situaciones con detalle para el problema de *El mundo de la información 4*. A continuación las enumeramos, pero es conveniente seguir la tabla 5.19.

1. No hay éxitos: única situación posible. Combinación de cuatro pruebas y sin éxitos.
2. Un éxito: cuatro situaciones. Combinación de cuatro pruebas y sólo se tiene un éxito.
3. Dos éxitos: seis situaciones posibles. Combinación de cuatro pruebas y se tienen dos éxitos.
4. Tres éxitos: cuatro situaciones posibles. Combinación de cuatro pruebas y se tienen tres éxitos.
5. Cuatro éxitos: una situación. Combinación de cuatro pruebas y se tienen cuatro éxitos.

Por ejemplo:

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2!2!} = \frac{4 \times 3 \times 2 \times 1}{(2 \times 1)(2 \times 1)} = \frac{24}{4} = 6$$

Es importante observar que el término p^x en el modelo de probabilidad binomial representa la probabilidad de x éxitos en x pruebas. El término q^{n-x} representa la probabilidad de $n - x$ fallas en $n - x$ pruebas. Se multiplican p^x y q^{n-x} porque las pruebas son independientes.

Representación gráfica de la distribución binomial

Retomando la información presentada en la tabla 5.19, se observa que la distribución de probabilidad para una binomial con $n = 4$ y $p = 0.25$ es:

Tabla 5.20 Número de éxitos en un ensayo Bernoulli.

	0	1	2	3	4
$p(x)$	0.316	0.422	0.211	0.047	0.004
$P(X \leq x)$	0.316	0.738	0.949	0.996	1

La representación gráfica de esta distribución se muestra en la figura 5.9.

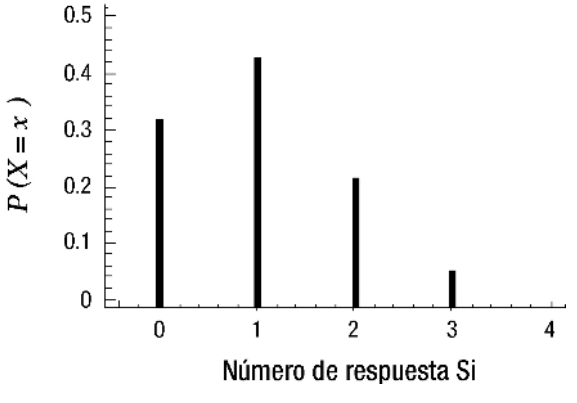


Figura 5.9 Distribución de probabilidad binomial para el número de respuestas "sí".

Ejemplo 5.8

Se tiene un nuevo tratamiento para curar un dolor muscular en el 60 % de los casos. Si éste se prueba en 15 pacientes, encontrar la probabilidad de que:

1. A lo más 6 se curen
2. El número que se curen no sea menos de 7 ni más de 9.
3. Diez o más se curarán.

Solución mediante el uso de CalEst

En el menú principal se selecciona la opción Probabilidad, enseguida se accede a la distribución de probabilidad binomial y aparece un esquema como el que se describe en la figura 5.10. Observe que en el extremo superior izquierdo se tiene un cuadro que contiene los parámetros de la distribución binomial, los que el usuario puede operar con las flechas indicadas, en este caso $p = 0.6$ y $n = 15$. El umbral permitirá calcular las probabilidades de esta distribución.

Observación sobre el uso de **CalEst**. Al seleccionar la distribución en la pantalla original aparece en la parte superior un cuadro que simula el teclado de una calculadora, éste lo puede usar para hacer cálculos directos de probabilidades de la binomial. Esta alternativa no aparece en esta gráfica pero se muestra en la figura 5.2.

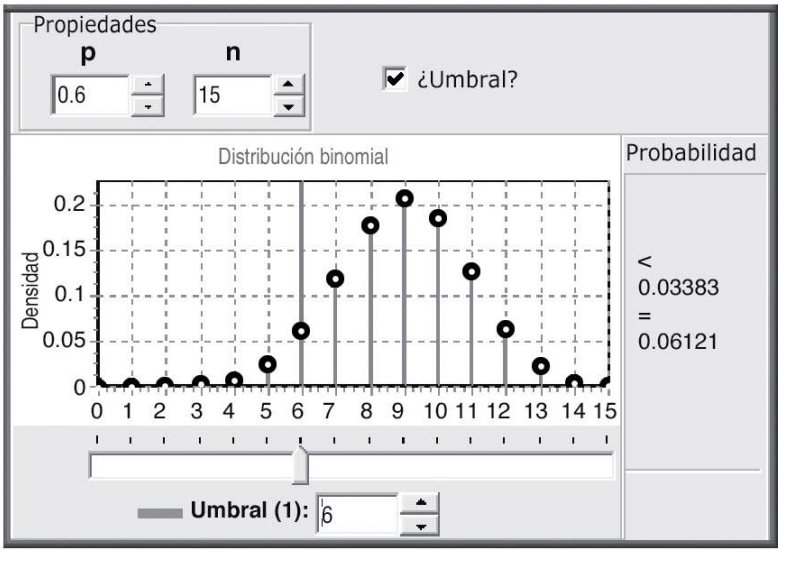


Figura 5.10 Descripción de la entrada para el cálculo de probabilidades en la binomial y el reporte de salida.

1. En este caso se requiere calcular la probabilidad de que a lo más 6 pacientes se curen, es decir, $P(X \leq 6)$. Para encontrar esta probabilidad en la pantalla de la figura 5.10 se ejemplifica cómo hacerlo: con el botón izquierdo del mouse se pone el umbral en 6. En el cuadro Probabilidad aparecen tres probabilidades, las cuales son: la probabilidad de que X sea menor, igual o mayor que 6; en símbolos: $P(X < 6), P(X = 6)$ y $P(X > 6)$. Así la probabilidad buscada en este problema es:

$$P(X \leq 6) = P(X < 6) + P(X = 6) = 0.0338 + 0.0612 = 0.095$$

Una solución alternativa es aplicar la ley de probabilidad del complemento, donde el evento A es: $A = \{X \leq 6\}$ y $A^C = \{X > 6\}$; entonces, $P(A) = 1 - P(A^c)$. Así:

$$P(X \leq 6) = 1 - P(X > 6) = 1 - 0.9049 = 0.095$$

2. En este caso se requiere la probabilidad $P(7 \leq X \leq 9)$. Para realizar este cálculo, primero se pone el umbral en 9 y luego en 7. En ambas situaciones se recurre al CalEst y se encuentran esquemas similares a las de la figura 5.10. El resultado es:

$$\begin{aligned} P(7 \leq X \leq 9) &= P(X \leq 9) - P(X \leq 7) \\ &= (P(X < 9) + P(X = 9)) - (P(X < 7) + P(X = 7)) = 0.384 \end{aligned}$$

donde:

$$\begin{aligned} P(X < 9) + P(X = 9) &= 0.3901 + 0.2065 = 0.597 \\ P(X = 7) + P(X = 7) &= 0.095 + 0.118 = 0.213 \end{aligned}$$

Finalmente, diez o más se curan:

$$P(X \geq 10) = 1 - P(X < 10) = 1 - 0.5967 = 0.4033$$

Nota. Los cálculos están hasta diez milésimas sin haber redondeado ese número, por lo que hay errores de redondeo después de las milésimas.

La distribución binomial acumulada

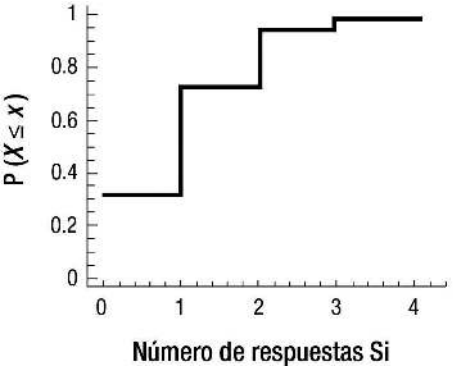
Finalmente, se obtiene la distribución de probabilidad binomial acumulada, es decir $P(X \leq x)$. Continuando con la información que se generó en el problema de *El mundo de la información 4*, se observa el último renglón de la tabla 5.20, donde se encuentran los cálculos para la distribución.

Interpretación: La probabilidad de obtener dos o menos respuestas "sí " (dos o menos éxitos) es:

$$P(X \leq 2) = 0.949$$

Este resultado significa que más de las veces el número de respuesta sí será 0, 1 o 2, lo que se traduce en que prácticamente los jóvenes tienen poca información sobre el SIDA.

La gráfica de la distribución binomial acumulada correspondiente se muestra en la figura 5.11. En está se ha puesto a un lado la tabla de distribución asociada.



X	p (X)	P (X ≤ x)
0	0.316	0.316
1	0.422	0.738
2	0.211	0.949
3	0.047	0.996
4	0.004	1

Figura 5.11 Distribución de probabilidad binomial acumulada para el número de respuestas sí y tabla de distribución.

Comentario. En el apartado Complemento didáctico se muestran las opciones para realizar el cálculo de la distribución binomial usando **CalEst**.

Las probabilidades en la distribución binomial están dadas por

$$P(X = x) = p(x) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, ..., n \tag{5.3}$$

donde: X es número de éxitos en n pruebas, n es el número de pruebas, $x = 0, 1, 2, \ldots, n$, p es la probabilidad de éxitos en una prueba, $q = 1 - p$ la probabilidad de fracasos en una prueba.

La distribución acumulada

$$P(X \leq x) = p(0) + p(1) + ... + p(x)$$

La binomial se denota : $binomial(n, p)$.

Distribución de probabilidad acumulada

$$P(X \leq c) = \sum_{x=0}^{c} f(x) = \sum_{x=0}^{c} \binom{n}{x} p^x (1-p)^{n-x}$$

La media y la varianza para la distribución binomial son:

$$\text{media} \quad \mu = np \qquad \text{varianza} \quad \sigma^2 = np(1-p)$$

Tablas para la distribución binomial

Para calcular las probabilidades en un experimento binomial necesitamos conocer el número de pruebas y la probabilidad de éxito. Se dispone de tablas de distribución binomial apropiadas para tal fin. A continuación, en la tabla 5.21, se reproduce una porción de la tabla binomial relacionada con $n = 4$ y diferentes probabilidades de éxito p. Esta tabla se puede reproducir usando el paquete estadístico en la opción Probabilidad y luego en Distribuciones de probabilidad. En el ejemplo 8 se ilustró la técnica de cálculo.

En la tabla 5.21 aparece el valor de n, x y p y los valores correspondientes que se refieren a la distribución binomial acumulada $P(X \leq x)$. Si queremos calcular la probabilidad de $P(X = 2)$ y $p = 0.20$, entonces:

$$P(X = 2) = P(X \leq 2) - P(X \leq 1) = 0.9728 - 0.8192 = 0.1536$$

donde $P(X \leq 2) = 0.9728$, se obtiene de la tabla 5.21

Tabla 5.21 Distribución de probabilidad binomial acumulada. para n = 4 y p = 0.05, 0.10, 0.20, 0.25, 0.50.

$n = 4$			p		
x	0.05	0.10	0.20	0.25	0.50
0	0.8145	0.6561	0.4096	0.3164	0.0625
1	0.9860	0.9477	0.8192	0.7386	0.3125
2	0.9995	0.9963	0.9728	0.9492	0.6875
3	1.000	0.9999	0.9984	0.9961	0.9375
4	1.000	1.000	1.000	1.000	1.000

Verifique la columna con $p = 0.25$ con la tabla 5.20 o figura 5.11 para el problema de información sobre el SIDA. Se ha tratado este caso por el impacto económico que representa.

Ejemplo 5.9

Un profesor de historia quiere saber si los estudiantes leen el periódico al menos una vez a la semana. Le interesa conocerlo porque considera que los periódicos son una buena fuente de información acerca de, entre otros temas, acontecimientos sociales, políticos, deportivos, económicos y culturales. Con los resultados podrá deducir el nivel de información e interés que los alumnos muestran hacia su entorno.

Les pregunta a 4 estudiantes de su clase, de manera independiente, si leen el periódico una vez a la semana. La variable aleatoria es el número de alumnos que leen el periódico una vez a la semana. Se sabe que p=0.10, la cual es la probabilidad de éxito y la respuesta es "sí".

Solución

1. Usando la tabla 5.21 con $n = 4$ y $p = 0.10$ se tiene:

$$P(X = 3) = P(X \leq 2) = 0.9999 - 0.9963 = 0.0036$$

2. Usando el paquete estadístico, con $n = 20$ y $p = 0.10$, tenemos que:

$$P(X \leq 6) = P(X < 6) + P(X = 6) = 0.9887 + 0.0088 = 0.997$$

3. Después de obtener que 6 alumnos o más leen el periódico, nuevamente usando el paquete estadístico, donde se ve que $P(X \leq 5) = P(X < 5) + P(X = 5) = 0.988$ se sigue que:

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - 0.9887 = 0.011$$

4. Finalmente,

$$P(X = 6) = P(X \leq 6) - P(X \leq 5) = 1 - 0.9976 - 0.9887 = 0.011$$

Interpretación: Con los valores de las probabilidades podemos concluir que la mayoría de las veces el número que dijo "sí " será 0, 1, 2, 3, 4, 5 o 6. Esto significa que pocos alumnos leen un periódico al menos una vez a la semana.

Valor esperado y varianza de la distribución binomial

La media o el valor esperado para la distribución binomial se puede obtener directamente de la definición que se dio con anterioridad en este capítulo, es decir:

$$\mu = E(X) = \sum_{i=1}^{k} x_i p(x_i)$$

Se puede demostrar matemáticamente de esta definición una forma más sencilla y simple para expresar la media de la distribución binomial; esta es:

$$\mu = np$$

donde n y p son los parámetros de la distribución binomial.

Después de algún desarrollo matemático sencillo la expresión para la varianza es:

$$\sigma^2 = Var(X) = mp(1-p)$$

De esta expresión se deduce la desviación estándar:

$$\sigma = DE(X) = \sqrt{np(1-p)}$$

Ejemplo 5.10

En la producción de transmisiones de autos se tiene la probabilidad de que una sea defectuosa es de 4 %. Encontrar:

1. El número de transmisiones defectuosas esperado en un lote de 1000.
2. La varianza y la desviación estándar.

Solución

1. De la información del ejemplo se tiene $n = 1000$ y $p = 4/100 = 0.04$. El valor esperado medio se obtiene mediante la expresión $\mu = np$. Así:

$$\mu = np = 1000 \times 0.04 = 40 \text{ transmisiones}$$

2. Usando esa misma información se obtiene la varianza y la desviación estándar.

$$\sigma^2 = np(1-p) = 1000 \times 0.04(1-0.04) = 38.4 \qquad \sigma = \sqrt{38.4} = 6.19$$

5.4 Distribución Poisson

En diferentes estudios es frecuente encontrarse con problemas donde una información se deriva de la ocurrencia aleatoria de eventos durante un periodo de tiempo establecido o de la longitud determinada en un segmento.

Considere el número de ocurrencias de fenómenos naturales como terremotos o huracanes en algún intervalo de tiempo, por ejemplo un año. Se considera como una variable aleatoria con una distribución de Poisson .

La distribución de probabilidad está dada por:

$$P(x) = \frac{\exp(-\lambda)\lambda^x}{x!} \quad \text{para } x = 0, 1, 2, ...; \lambda > 0, \tag{5.4a}$$

donde $P(x)$ es la probabilidad de que x eventos aleatorios ocurrirán por unidad de tiempo o espacio; λ es la razón de ocurrencias (la media del número de eventos aleatorios) por unidad de tiempo o espacio.

La media y la varianza para la distribución Poisson son:

$$\text{media } \mu = \lambda \quad \text{varianza } \sigma^2 = \lambda$$

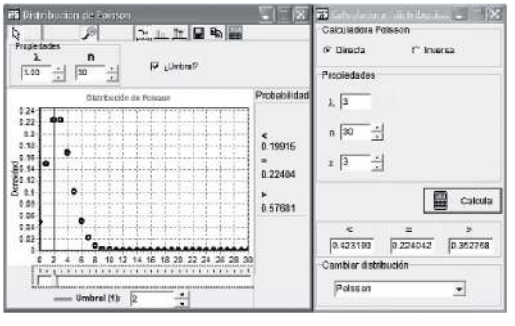


Figura 5.12 Distribución de Poisson con parámetros $\lambda = 3$ en una muestra $n = 30$.

Ejemplo 5.11

Cálculo de probabilidades usando CalEst

La probabilidad de que un artículo producido por una máquina durante cada periodo de revisión sea defectuoso es: 0.1. Determine la probabilidad de que en una muestra de 30 artículos no haya más de dos defectuosos.

Solución

Se desea estimar $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$, se sustituyen en la expresión (5.4a) los valores de x y $\lambda = 30(0.1) = 3$. Usando la distribución Poisson en el grupo de distribuciones en **CalEst** se obtiene el valor deseado (véase la figura 5.12). Observe que en la gráfica se indican tres valores con respecto a dónde se puso el cursor. En el caso de la figura 5.12 el cursor está en 2, entonces se indican la probabilidad de: X menor a 2, X igual a 2 y X mayor a 2. Para el cálculo deseado se tienen dos posibilidades:

Primero, si se pone el cursor en 2 la probabilidad $P(X \leq 2)$ se obtiene sumando los valores de 0.19915 y 0.22404, es decir:

$$P(X \leq 2) = 0.19915 + 0.22404 = 0.42319$$

La otra es poner el cursor en 3; el valor es el anterior porque se ha calculado la probabilidad de que X sea menor que 3, lo que resulta equivalente a que x sea menor e igual a 2 ($X \leq 2$). Empleando la calculadora Poisson también se obtienen los valores de las probabilidades; en la tabla a la derecha en la figura 5.4 se ha ilustrado la segunda situación que se explicó anteriormente.

5.5 Resumen

Variable aleatoria	Regla que representa los posibles valores numéricos asociados con los resultados de un experimento.
Variable aleatoria discreta	Variable numérica que toma un número de valores que se pueden contar.
Puesto que $p(x_1)$ representa probabilidades, sus valores para cada i deben estar en 0 y 1,	$0 \leq p(x_1) \leq 1$.
La distribución de probabilidad de la variable aleatoria discreta X se describe mediante la expresión:	$p(x_i) = P[X = x_i]$
La probabilidad para cada valor satisface las siguientes condiciones:	$p(x_i) \geq 0$, para cada valor x_i de X. $0 \leq p(x_i) \leq 1$ $\sum_{i=1}^{k} p(x_i) = 1$

Notación que resume el cálculo para encontrar probabilidades.

	Interpretación	**Notación**
La probabilidad de que	Todos los valores de la variable	
X tome una valor que es	aleatoria	
Al menos x	que son los valores de x o mayores que x	$P(X \geq x)$
Más que x	que son mayores que el valor x	$P(X > x)$
A lo más x	que son los valores de x o menores que x	$P(X \leq x)$
Menos de x	que son menores que el valor x	$P(X < x)$
Entre x_1 y x_2	que son mayores que los valores x_1 y menores que los valores x_2	$P(x_1 < X < x_2)$
Entre x_1 y x_2 y que les incluye	que inicia con el valor x_1 y termina con el valor x_2 incluyendo su valor	$P(x_1 \leq X \leq x_2)$

Notación que resume el cálculo para encontrar probabilidades.

	Interpretación	**Notación**
El valor esperado para una variable aleatoria X	.....	Se denota por y su fórmula: $\mu = E(x) = \sum_{i=1}^{k} x_i p(x_i).$
La varianza de la variable aleatoria X es: La desviación estándar de la variable aleatoria X es:	.....	$\sigma^2 = \textbf{Var}(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(x_i)$ $\sigma = DE(X) = \sqrt{Var(X)}$

Distribución de Probabilidad	Una tabla o función que liste todos los posibles valores de una variable aleatoria discreta, y su probabilidad asociada es la distribución de probabilidad de la variable aleatoria.
Distribución de Probabilidad acumulada	La función de la distribución de probabilidad acumulada. (fdpa) de una variable aleatoria X es la que da la probabilidad de que X sea menor o igual que x, y se escribe como: $P(X \leq x)$.
Propiedades de una variable binominal	Las características que definen una variable binomial: 1. Existe un número n fijo de ensayos Bernoulli. 2. Cada ensayo presenta un resultado de dos posibilidades, que se refieren como éxito y fracaso 3. La probabilidad p de éxito en una prueba es constante. Los ensayos son independientes, es decir, la probabilidad de éxito en cualquiera de los ensayos no se ve afectada por el resultado de un ensayo previo.
Distribución de una probabilidad binominal	Las probabilidades en la distribución binomial están dadas por la probabilidad $P(X = x) = p(x) = \binom{n}{x} p^x q^{n-x}$ $x = 0, 1, 2, ..., n$ donde: X es número de éxitos en n pruebas. n es el número de pruebas. p es la probabilidad de éxitos en una prueba. $q = 1 - p$ la probabilidad de fracaso en una prueba. La distribución acumulada $P(X \leq x) = p(0) + p(1) + ... + P8x)$ La binomial se denota: binomial (n, p)
Distribución de probabilidad Poisson	$P(x) = \frac{\exp(-\lambda)\lambda^x}{x!}$ para $x = 0, 1, 2, ...; \lambda > 0$, $P(x)$ la probabilidad de x eventos en un área, espacio, lapso de oportunidad $\lambda =$ número de eventos esperados $\exp = e$ una constante matemática aproximadamente igual a 2.71828 $x =$ número de eventos

5.6 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

5.7 Ejercicios

Distribuciones de probabilidad

5.1 ¿Qué es una variable aleatoria? ¿Qué es una variable aleatoria discreta?

5.2 ¿Cuál es el principio general para encontrar $p(x)$?

5.3 ¿Cuáles son las propiedades generales que debe satisfacer la probabilidad para que sea una función de probabilidad?

5.4 Para referirse a la expresión $p(x) = P(X \leq x)$, indique dos maneras de enunciarlo.

5.5 En una distribución de probabilidad acumulada, escriba la expresión para indicar la probabilidad de la cola a la derecha.

5.6 El cálculo indicado por la expresión: $P(X \leq 10) - P(X \leq 6)$, indique a qué probabilidad nos referimos.

5.7 A continuación se describe la distribución de probabilidad de una variable aleatoria X, encuentre el valor que falta.

x	1	2	3	4	5
$p(x)$	0.08		0.12	0.31	0.39

5.8 Una compañía vende bolígrafos en paquetes grandes a una tienda. Según los registros en los libros de la tienda, se tiene el número de plumas defectuosas, que es una variable aleatoria discreta. Los valores y su distribución de probabilidad se describen a continuación.

x	0	1	2	3	4	5	6
$p(x)$	0.30	0.21	0.12	0.10	0.10	0.09	0.08

1. Averigue la probabilidad de que un paquete de bolígrafos contenga al menos una defectuosa.
2. Encuentre la probabilidad de que el paquete contenga entre 2 y 5 bolígrafos defectuosos.
3. Obtenga la probabilidad de que el número de bolígrafos defectuosos sea a lo más de 2.
4. Elabore un diagrama de barras para ilustrar el número de bolígrafos defectuosos.

5.9 Una bolsa contiene 2 canicas blancas y 6 negras, se extraen tres canicas con remplazo. Encuentre la distribución de probabilidad para el número de canicas de color blanco.

5.10 De una encuesta a 200 adultos sobre cuántos televisores hay en su casa, resultó que 176 tienen 1; 22 tienen 2, y 2 cuentan con 3 de estos aparatos.

Si una persona se selecciona al azar, ¿cuál es la función de probabilidad para X el número de televisores que tiene una persona?

5.11 A cuatro jóvenes se les preguntó si tomaban medidas para no embarazarse cuando tenían relaciones sexuales. Cada una respondía con un "sí " o un "no".

1. Defina una variable aleatoria en el espacio muestral de resultados posibles.
2. Escriba una lista del espacio muestral de la variable aleatoria.

5.12 De acuerdo con un estudio en una escuela de nivel bachillerato se sabe que 50 % de los estudiantes practica algún deporte. Indique con la letra S si hace deporte y con N si no lo hace. Se seleccionan a 6 estudiantes de un espacio muestral de 64 posibles resultados diferentes. La variable aleatoria X denota el número de estudiantes que practica deporte. Anote la lista de los valores de la variable X e interprete los valores.

5.13 Se realiza un experimento que consiste en lanzar una moneda hasta que aparezca "águila". Defina una variable aleatoria que describa el experimento y escriba la lista del espacio muestral correspondiente.

5.14 En una familia con 3 niños, la variable aleatoria es X : número de niñas en la familia.

1. Encuentre la distribución de probabilidad para X.
2. Describa la gráfica de la distribución.

5.15 ¿Cuál es el principio general que se usa para encontrar $P(X \leq x)$?

5.16 ¿Cómo interpreta la probabilidad $P(X \leq x)$?

5.17 El número de periódicos que se venden todos los días en un quiosco es una variable aleatoria X que tiene la siguiente distribución.

x	0	1	2	3	4	5
$p(x)$	0.07	0.15	0.31	0.23	0.18	0.05
$P(X \leq x)$						

1. ¿Cuál es la probabilidad de que se venda exactamente 1 periódico?
2. ¿Cuál es la probabilidad de que se vendan menos de 2 periódicos?
3. El éxito del quiosco de periódicos se da si la probabilidad es de 3, es decir, si se venden 3 o más periódicos. Encuentre la probabilidad. ¿Considera que es redituable el negocio?

5.18 Un número en la siguiente tabla para la función de probabilidad de una variable aleatoria X es incorrecto. ¿Cuál es éste, y qué se podría hacer para que sea un valor correcto? Justifique su respuesta.

x	1	2	3	4	5
$p(x)$	0.07	0.10	0.10	0.32	0.40
$P(X \leq x)$					

5.19 Supongamos que X es una variable aleatoria discreta con la siguiente distribución de probabilidad:

x	1	3	5	7	9
$p(x)$	1/15	2/15	3/15	4/15	5/15
$P(X \leq x)$					

1. ¿Cuál es el valor de x que es más probable que ocurra?
2. ¿Cuál es la probabilidad de que x sea par?
3. ¿Cuál es la probabilidad de que x sea menor que 7?

5.20 El 40 % de empleados en una compañía están a favor de una nueva propuesta de pago de incentivos, describa con F los que están a favor y con NF a los que no están a favor. Mediante un diagrama de árbol elabore la distribución de probabilidad para una muestra de

1. dos empleados,

2. tres empleados,

3. Considerando el inciso b, calcule las probabilidades para los valores de $X = 0$, 1, o 2.

Distribución binomial

5.21 Una empresa fabrica chips que se utilizarán en los automóviles. La distribución de probabilidad de chips defectuosos en un lote de producción es como sigue:

x	0	1	2	3	4	5
$p(x)$	0.31	0.41	0.21	0.05	0.009	0.001

Encuentre el valor esperado y la desviación de esta distribución de probabilidad.

5.22 Con base en los registros, una empresa que pinta automóviles tiene determinada la siguiente distribución de probabilidad para el número de clientes por día.

x	0	1	2	3	4	5
$p(x)$	0.05	0.20	0.30	0.25	0.15	0.05

1. Haga una gráfica para representar esta distribución.

2. ¿Cuál es el valor medio de clientes esperado al día?

3. ¿La dispersión en torno al valor esperado es grande?

5.23 Un carpintero se inscribe para la licitación de un proyecto para remodelar una biblioteca. Él determina que tendrá una ganancia neta de 500 mil pesos si obtiene el contrato y una pérdida neta de 5600 pesos si su oferta falla. Si la probabilidad de tener el contrato es de $\frac{1}{4}$, calcule el pago esperado.

5.24 Una compañía de seguros realiza un análisis sobre el registro de demandas de 320 asegurados en un periodo de 6 años, es decir, efectúa una determinación empírica de la distribución de probabilidad de la variable X= número de demandas en cinco años.

x	0	1	2	3	4	5	6
$p(x)$	0.287	0.268	0.224	0.123	0.069	0.020	0.009

1. Calcule el valor medio del número de demandas

2. Calcule la desviación estándar.

5.25 El número de días X que le lleva a una empresa entregar un paquete de una ciudad V a otra ciudad M tiene la siguiente distribución

x	2	3	4	5	6
$p(x)$	0.55	0.25	0.1	0.05	0.05

1. Encuentre la media y la desviación estándar
2. ¿Cuál es la probabilidad de que un paquete tarde más de tres días en llegar de la cuidad V a la ciudad M?
3. Si dos paquetes son enviados de la ciudad V en diferentes meses, ¿cuál es la probabilidad de que ambos paquetes lleguen a la ciudad M en al menos 4 días después de que fueron enviados?

5.26 La administración de un laboratorio envía sus distribuidores a probar un medicamento. A partir de un periodo éste tiene efecto sobre el 30 % de los pacientes en los casos en que la ha prescrito. Si en el caso de un médico particular, administra la medicina en cuatro pacientes, ¿cuál es la probabilidad de que esta medicina tenga efecto en al menos tres de los pacientes?

5.27 En una bolsa se tiene 30 canicas, de las cuales 10 son negras y 20 blancas. Se considera sacar cinco canicas de la bolsa, se seleccionan éstas con reemplazo, es decir, se toma una y se regresa a la bolsa.

1. Construya la distribución de probabilidad para las canicas negras.
2. Obtenga la probabilidad de que $p(2) = P(X = 2)$.
3. Calcule la probabilidad de que al menos una canica es negra.

5.28 En una de las salas de un complejo de cines, en un día de aniversario la administración realiza una promoción. Si uno de los asistentes saca una bola blanca de una urna, no paga la entrada. La urna contiene 50 bolas, 45 rojas y 5 blancas. Hay 12 personas formadas en la fila, ¿cuál es la probabilidad de que tres personas pasen gratis?

5.29 En una industria un proceso produce artículos, de ellos se tiene que el 1 % es defectuoso. De un lote de producción, el administrador toma una muestra grande $n = 100$, calcule la probabilidad de que en la muestra no hay productos defectuosos.

5.30 La variable aleatoria X tiene una distribución binomial con parámetros n y p, tal que n es el número de ensayos y p la probabilidad de éxito. Encuentre las siguientes probabilidades:

1. $P(X \geq 2)$, y $P(X = 2)$ si $n = 4$ y $p = 0.23$
2. $P(8 \leq X \leq 12)$ y $P(X \leq 6)$ si $n = 14$ y $p = 0.30$
3. $P(X \leq 8)$ si $n = 16$ y $p = 0.25$
4. $P(8 \leq X \leq 10)$ y $P(X \leq 6)$ si $n = 10$ y $p = 0.70$

Distribución Poisson

Las probabilidades de éxito y de fracaso podemos relacionarlas con muchas actividades de la vida diaria, por ejemplo cuando nos referimos a un producto defectuoso o no defectuoso, si nos gustó un servicio o no. A los fabricantes de cualquier tipo de producto, en particular les resulta muy importante conocer la distribución de probabilidad de artículos defectuosos para resolver los problemas y para conocer el tiempo en que pueden garantizar sus productos.

5.31 Un fabricante de chips para tarjetas electrónicas que se usan en computadoras escoge 5 chips, donde la probabilidad de artículos defectuosos es de $p = 0.2$. En este caso es adecuado utilizar la distribución binomial para estudiar el tipo de defectos de chips en tarjetas electrónicas. Los parámetros correspondientes para esta distribución binomial son $n = 5$ y $p = 0.2$. A continuación se muestra la tabla para estas características y se redondea a 3 decimales.

x	0	1	2	3	4	5
$p(x)$	0.328	0.410	0.205	0.051	0.006	0.000
$P(\leq x)$	0.328	0.738	0.943	0.994	1	1

1. ¿Cuál es la probabilidad de que exactamente 2 chips estén defectuosos?
2. ¿Cuál es la probabilidad de que a lo más 2 chips estén defectuosos?
3. ¿Cuál es la probabilidad de que 3 o más chips estén defectuosos?
4. Interprete los resultados.

5.32 Las llamadas telefónicas al número de emergencia 066 se sabe que siguen una distribución Poisson con una media de dos llamadas por minuto. Calcule la probabilidad de que:

1. Se tienen cero llamadas en un minuto.
2. Hay al menos cinco llamadas en un minuto.
3. Hay al menos seis llamadas en una hora.

5.33 En una ciudad hay algunos tramos donde la velocidad está restringida a 80 km/h. Se sabe por los registros que 30 % de los automovilistas rebasan ese límite de velocidad. De esos, 15 automovilistas pasan inadvertidos por los policías o los radares de control. En este caso, la variable aleatoria es el número de autos que rebasa la velocidad permitida. Así se obtiene la binomial (15,0.3). Utilice el paquete estadístico en la opción correspondiente a la binomial.

1. ¿Cuál es la probabilidad de que 7 o más autos no respeten el límite de velocidad?
2. Encuentre la probabilidad de que entre 3 y 8 autos sobrepasen el límite de velocidad.

3. Comente con sus compañeros los resultados que obtuvo.

5.34 La probabilidad de éxito de una vacuna contra la gripe es 0.74. Calcule la probabilidad de que una vez que esta se administre a 20 personas:

1. Ninguno tenga gripe.
2. Todos contraigan la gripe.
3. Dos de ellos contraigan la gripe.

5.35 Una máquina produce suelas para zapato industrial y se sabe que produce un 7 por 1000 suelas defectuosas.

Hallar la probabilidad de que al examinar 50 suelas sólo haya una defectuosa.

5.36 El propósito de este ejercicio es que use el CalEst para calcular las siguientes probabilidades de un binomial. Si la variable X es binomial con $n = 10$ y $p = 0.3$.

a.- $P(X = 3)$, b.- $P(X \geq 4)$, c.- $P(X > 6)$, d.- $P(X < 7)$, e.- $P(X \leq 1)$, f.- $P(3 < X < 8)$, g.- $P(4 \leq X < 8)$, h.- $P(X > 12)$, i.- $P(X < 11)$.

5.37 Un equipo de fútbol ha ganado 40 % de 20 partidos después de haber metido primero un gol. ¿Cuál es la probabilidad de ganar al menos 6 juegos? ¿Cuál es la probabilidad de ganar más de 8 juegos?

5.38 La probabilidad de ganar una rifa es $p = 0.1$. Si se hacen tres intentos, ¿cuál es el valor esperado?

1. Construya la tabla de la distribución de probabilidad binomial para $X = 0, 1, 2$ y 3. Utilice la fórmula $\sum_{i=1} x_i p\left(x_i\right)$ para encontrar el valor esperado.
2. Use la fórmula del valor esperado para la distribución binomial.
3. Compare los resultados.
4. Encuentre la varianza y la desviación estándar.

5.39 De una población grande se selecciona una muestra de 1000 personas de las cuales el 20 % de ellas será auditada por la aduana. ¿Cuál es el valor esperado de las personas que serán auditadas?

5.40 Los ojos color café claro son una característica genética. Supongamos que en dicho genotipo de los futuros padres existe una probabilidad de ocurrencia del 75 %. Calcular los posibles eventos si se planea procrear seis hijos.

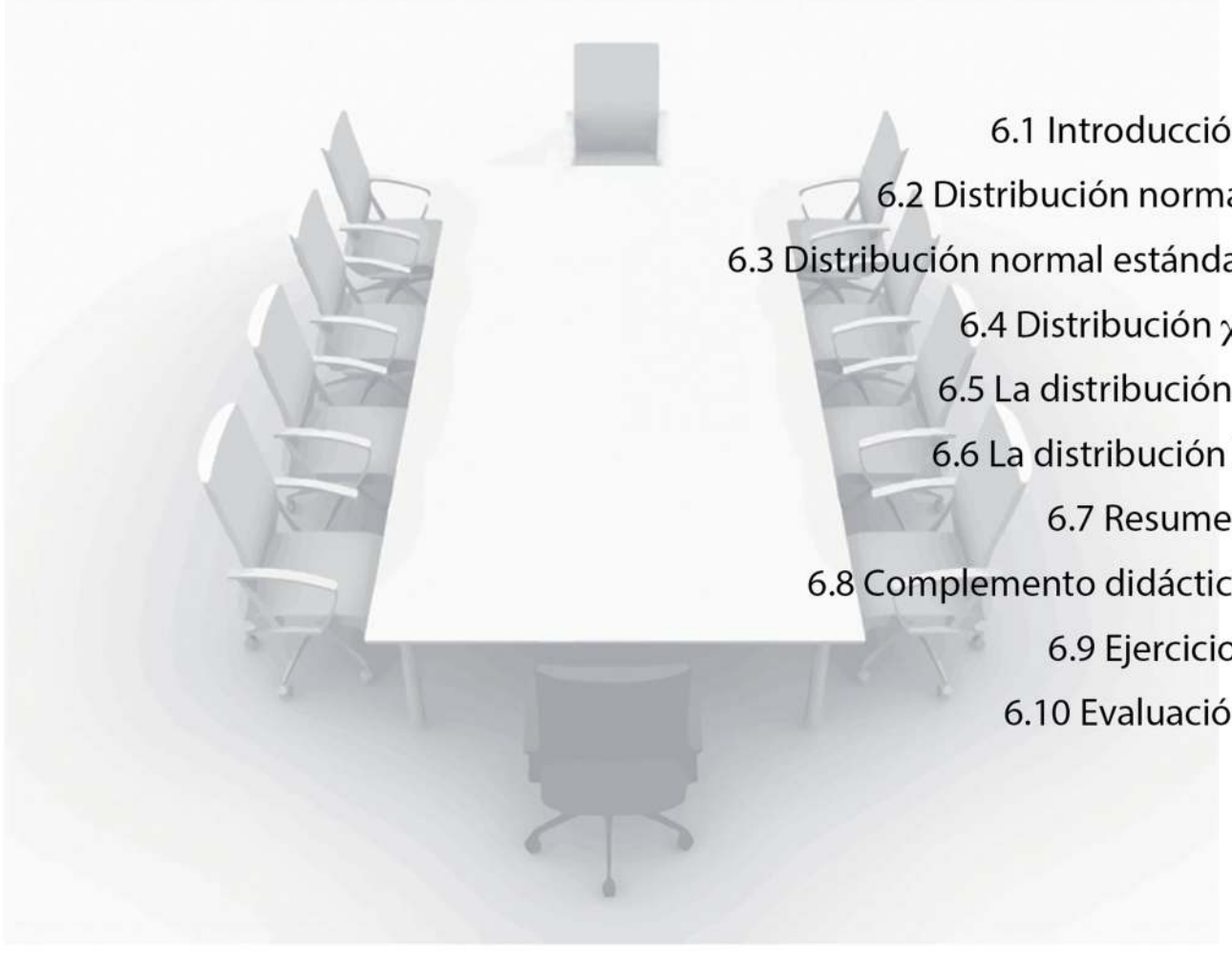
5.41 Observe numéricamente la aproximación a una binomial con parámetros $n = 100$ y $p = 0.04$ y una Poisson con parámetro $\lambda = 4$.

5.8 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.

Capítulo 6

Distribución de probabilidad: variables continuas



No es el conocimiento, sino el acto de aprendizaje, y no la posesión, sino el acto de llegar ahí, que concede el mayor disfrute.

Carl F. Gauss

Competencia general

Comprender y adquirir habilidad para la aplicación de la distribución de probabilidad que más se utiliza en estudios estadísticos: la distribución normal. Además examinar tres distribuciones básicas en la inferencia estadística: la Ji cuadrada, la $t - Student$, la F.

Competencias específicas

- Conocer y caracterizar el modelo de probabilidad normal, así como comprender que muchas variables que explican el comportamiento de fenómenos o procesos se pueden modelar mediante la distribución normal.
- Distinguir con claridad las funciones de densidad y de probabilidad de una normal, y extender estas ideas a las otras distribuciones.
- Aprender a calcular probabilidades con este modelo usando las tablas clásicas y mediante el uso de **CalEst**.
- Conocer y aplicar el método de transformación para calcular probabilidades de cualquier distribución normal con parámetros μ y σ, método que se denomina procedimiento de estandarización.
- Explicar cómo se puede verificar que un conjunto de datos sigue una distribución normal.
- Examinar mediante técnicas gráficas las distribuciones empírica y teórica.
- Identificar las condiciones para utilizar la distribución de probabilidad normal para aproximar la distribución binomial.
- Adquirir habilidad en el cálculo de probabilidades usando estas distribuciones para aplicarlas con facilidad en problemas de estimación y prueba de hipótesis: inferencia estadística.
- Comparar usando técnicas gráficas los valores de las distribuciones t-Student y normal.
- Emplear estas distribuciones para resolver una variedad de problemas relacionados con situaciones reales.

6.1 Introducción

El objetivo de este capítulo es presentar modelos de probabilidad que resumen la estructura de variación de una variable aleatoria continua. Por lo general, estos modelos se refieren como funciones de densidad, y existe una amplia variedad de ellos tal y como se muestra en la figura 6.1. La meta es encontrar la probabilidad, la cual está representada por el área bajo la densidad de la curva, donde la media corresponde al centro de gravedad de la curva. Primero se describirá la distribución normal, también referida como gausiana, en honor a Gauss. Ésta desempeña un papel relevante en la inferencia estadística, ya que la probabilidad de esta distribución permite evaluar la incertidumbre de las decisiones tomadas en función de la información de la muestra.

Usando esta distribución se calcularán las probabilidades y los percentiles, y se motivará la idea de trabajar con la normal estándar.

Posteriormente se presentan otras distribuciones conocidas como distribuciones muestrales, χ^2, la $t - Student$, la F. Éstas también son importantes en la inferencia estadística. La aplicación de estas variables vendrá en los capítulos siguientes, pero las exponemos aquí con el fin de tenerlas como referencia en el contexto de las distribuciones de probabilidad.

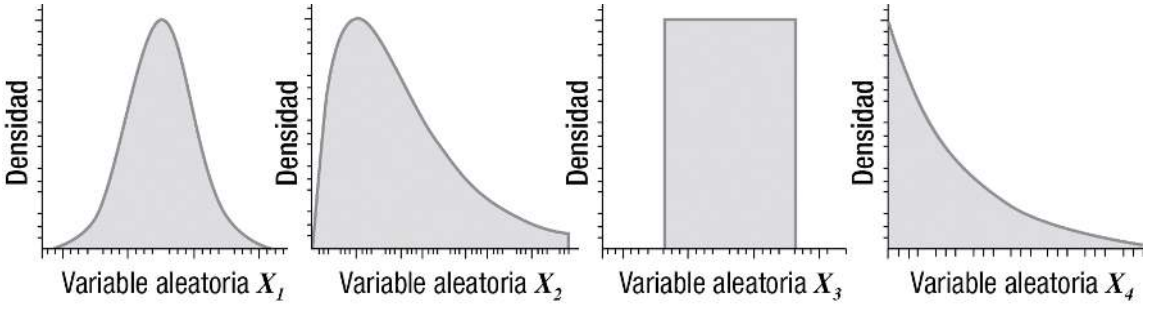


Figura 6.1 Modelos de probabilidad para una variable aleatoria con diferentes funciones de densidad.

Variables aleatorias continuas

Una característica de una variable aleatoria discreta es que sólo toma valores separados, distintos o contables. No todas las mediciones son de este tipo; por ejemplo, al medir el tiempo de llegada a la escuela, los valores pueden ser 40 o 41 minutos, o cualquier número entre 40 y 41 minutos, tal como 40.36 minutos. No existe separación ni valores distintos en este caso. Además, en este intervalo es posible un número infinito de números. De modo que los resultados no son contables como en el caso de la variable aleatoria discreta.

A la variable aleatoria como la del tiempo de llegada a la escuela se le llama variable aleatoria continua. Sus valores posibles forman un intervalo continuo y las probabilidades de las variables aleatorias continuas

se asocian sólo con intervalos de observaciones, no con valores individuales, como ocurre en el caso de las variables aleatorias discretas. Las típicas variables continuas miden alguna cantidad, como pueden ser el tiempo, el peso, la altura, el volumen, la presión sanguínea y la concentración de los químicos en la sangre con el fin obtener un seguro de vida, costos por mantenimiento en diferentes empresas, el ingreso promedio. Otros ejemplos de variables aleatorias continuas serían:

- El consumo de energía.
- El tiempo entre llamadas en un celular.
- El avalúo catastral.
- Ingreso mensual.
- Costos de mantenimiento.
- La inversión en seguros de vida.

En resumen, en las áreas de la administración y la economía muchas variables tienen una distribución aproximada a una normal. En el contexto de las distribuciones discretas, tales como la binomial o Poisson, cuando se tienen muchos datos se pueden aproximar a través de una normal.

6.2 Distribución normal

Características de la distribución normal

La distribución de probabilidad normal se aplica de manera frecuente para estudiar procesos cuya variable aleatoria es continua. La densidad de probabilidad de esta distribución se caracteriza por los parámetros $(\mu$ y $\sigma^2)$, es decir, su media (μ) y su varianza (σ^2); recuerde qué σ es la desviación estándar. Esta distribución es simétrica con respecto a la μ y tiene forma de campana, como se indica en la figura 6.2 la cual describe la forma de la normal. La expresión matemática que caracteriza la función de densidad de la normal está dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/(2\sigma^2)}. \tag{6.1}$$

donde $\pi = 3.1416$ y $e = 2.7183$

En resumen, la distribución de probabilidad normal es una curva simétrica en forma de campana. En estadística se le llama usualmente curva normal.

Distribución normal

La variable X tiene una distribución normal con media μ y desviación estándar σ se denota por:

$$X \sim N(\mu, \sigma^2)$$

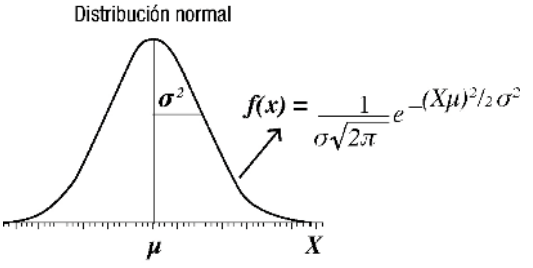


Figura 6.2 Distribución de probabilidad normal, con media μ y varianza σ^2.

La distribución normal es importante porque es un modelo apropiado para explicar la distribución de muchas medidas, tales como la estatura, el peso y las pruebas de aptitud, por mencionar algunos casos.

Campana de Gauss

La distribución normal se debe a Gauss, también conocida como "Campana de Gauss", la cual se usa ampliamente en el mundo de la estadística. Gauss, matemático, astrónomo y físico alemán, conocido como "el principe de las matemáticas", estudió en la Universidad de Götinga (1795-1798) y su tesis doctoral consistió en probar el teorema fundamental del álgebra. En 1823 se publicó su libro de estadística, en donde aparece el estudio de la normal.

El mundo de la información 1. Promedio de calificaciones

Los exámenes se aplican para saber cuál es el nivel de conocimientos de los estudiantes. Casi siempre se aplica un examen de ingreso a un nuevo nivel escolar, con el propósito de determinar quiénes tienen los conocimientos requeridos (aceptados) y quiénes carecen de esos conocimientos (rechazados). En nuestro caso, deseamos saber qué porcentaje de alumnos será rechazado si se aplica un examen de selección a 5 mil estudiantes.

Una función de densidad de probabilidad, como la de la figura 6.2, es una curva suave que representa la distribución de probabilidad de una variable aleatoria continua.

Preguntas sobre la naturaleza del problema

La variable de respuesta X es el promedio de la calificación que un estudiante obtiene al evaluarse en diferentes temas. Con esa información, se puede construir la distribución de la variable X. Se tiene establecido que los estudiantes que obtengan menos de 6 no ingresen al siguiente nivel. ¿Qué ocurre si ese porcentaje es alto y hay lugares disponibles? Los estudiantes con un promedio mayor que 9 se considerarán excelentes, pero ¿qué proporción de estudiantes están dentro del rango 6 y 9?

Por otro lado, si tienen un promedio entre 7.5 y 9 se considera que esos estudiantes les irá bien en el nuevo nivel escolar. Los que obtienen menos de 7.5 batallarán mucho o terminarán por abandonar la escuela. En la tabla 6.1 se describe el resumen de la información generada del examen.

Tabla 6.1 Resumen estadístico del examen

Media	7.510	Mínimo	5
Mediana	7.5	Máximo	10
Moda	7.3	Rango	6.9
C_1	6.9	Varianza	0.831
C_2	8.1	Desviación estándar	0.911

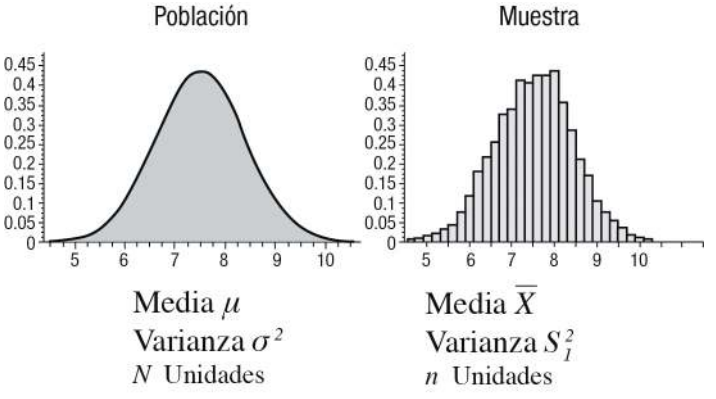


Figura 6.3 Modelo de probabilidad normal (Población) e histograma (muestra) que describe la distribución de las calificaciones.

Datos: Se obtuvieron 5000 datos, y a continuación se presenta el resumen estadístico. La descripción gráfica de estos datos se presenta mediante el histograma, figura 6.3. Debido a que el tamaño de la muestra es muy grande, $n = 5000$, podremos aproximar el histograma a un modelo teórico. Para este estudio, el modelo corresponde a una distribución normal.

Se sobrepone la curva normal al histograma (figura 6.4) donde se observa una buena aproximación del modelo normal (teórico) para modelar la distribución de las calificaciones. Siguiendo los datos de la tabla 6.1, se considera que la media muestral es $\overline{X} = 7.510$ y la desviación estándar $S = 0.911$. En este caso es necesario advertir que debido a la superposición de la curva normal y su buena aproximación a

la distribución de los datos, los parámetros y estadísticos pueden coincidir. Sin embargo, no se modela la distribución normal con los estadísticos $(\overline{X}, S)$, sino que se emplean los parámetros (μ y σ^2); para el presente estudio se propondrán $\mu = 7.5$ y $\sigma = 0.9$.

La superposición de la densidad normal es buena para usarla como un modelo adecuado para describir la distribución de las calificaciones.

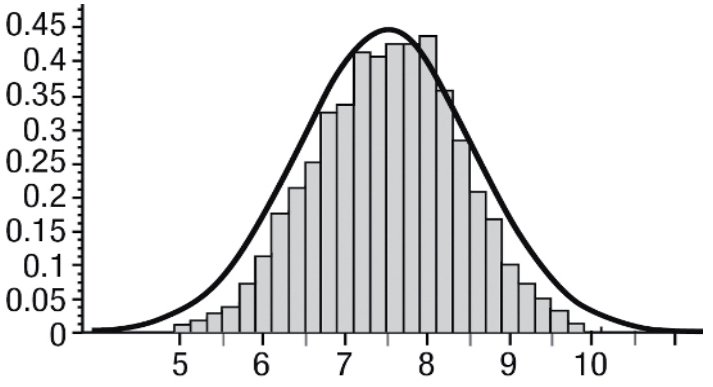


Figura 6.4 Superposición de la densidad normal para usarlo como un modelo adecuado para describir la distribución de las calificaciones.

Observaciones:

1. Del histograma hemos aprendido a calcular valores que cumplen o las relaciones que satisface la variable para explicar las respuestas de un problema planteado e interpretar la naturaleza de un estudio, usando las frecuencias o proporciones. Además se ha visto la utilidad del polígono de frecuencias acumulado para obtener conclusiones relevantes en cuanto a la naturaleza de las preguntas planteadas.

2. La normal es el modelo teórico que aproxima o explica la distribución de una variable aleatoria de una muestra, tal y como se indica en la figura 6.4. Una vez que se conocen los parámetros μ y σ^2, se puede dibujar su distribución y calcular sus valores para comprender los resultados de la investigación.

3. Una dificultad que surge para graficar una normal y obtener las probabilidades sobre las variables, es que para cada par de μ y σ^2 se tiene una distribución diferente. Ésta cambia de posición, originado por el valor de μ, y de amplitud ocasionado por el valor σ^2. Entonces es complicado obtener los valores que satisfagan las propiedades de cada normal y que no se pierda la idea original de la variable de estudio, es complicado. La figura 6.5 ilustra el material educativo contenido en **CalEst** que ayuda a resolver totalmente este problema; entre otras ventajas de este material destaca su su efecto visual, lo que facilita la comprensión del cálculo de probabilidades ya que ayuda a aprender a escribir de manera abstracta las expresiones de probabilidad, y permite tener claridad para distinguir entre los conceptos de función densidad, la curva de la normal, y la distribución de probabilidad de la normal, así como el polígono de frecuencias acumulado.

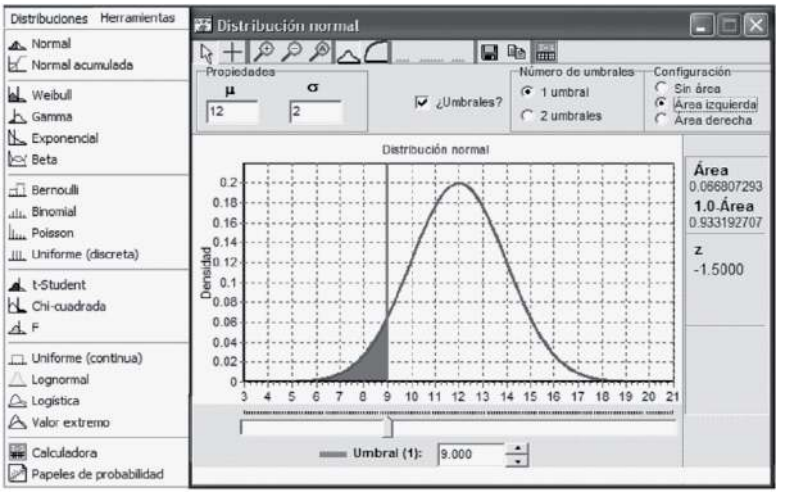


Figura 6.5 Descripción del CalEst para la presentación y cálculo de probabilidades usando la normal con media μ y desviación estándar σ.

Descripción de CalEst para la distribución normal

A continuación se presenta una guía que describe los puntos fundamentales de la distribución normal tal y como aparece en **CalEst**. Siguiendo estos pasos, usted se preparará en la operación de esta distribución que desempeña un papel relevante en el estudio de la estadística. Se espera que con la práctica que realice usando este material, tendrá suficiente habilidad para el cálculo de probabilidades bajo esta distribución. Además, madurará en sus ideas y experiencia para aplicar esta distribución a lo largo del capítulo, así como en el resto del libro y en los estudios que se proponga.

Como puede verse en la figura 6.5, en el bloque de distribuciones de **CalEst** aparecen las opciones para la normal, entre ellas la función densidad y acumulada. Con éstas se pueden calcular probabilidades o los valores de la variable X correspondientes a diferentes percentiles. Una vez que se está en el módulo de distribuciones, se acciona éste y aparece una lista de distribuciones; a partir de ahí seleccione la normal, enseguida verá esta distribución con media $\mu = 0$ y desviación estándar $\sigma = 1$.

Ahora, con el fin de trabajar en esta distribución y conocer su aplicación en el cálculo de probabilidades usando **CalEst**, se ha fijado una media $\mu = 12$ minutos y una desviación estándar $\sigma = 2$ minutos (para ver la gráfica en los cuadros para μ y σ indique los valores de 12 para μ, y en σ escriba 2, a continuación oprima la tecla enter); la variable X describe el tiempo de un tipo de servicio bancario. **Nota**: recurriendo a este mismo procedimiento, se pueden obtener los valores de las probabilidades para cualquier pareja de (μ, σ) en una distribución normal. Vea la figura 6.6.

1. Sin umbrales, en el cuadro sin la paloma ($\surd$) aparece la función Densidad. Para conocer el valor de la función para diferentes valores de la variable, use el signo más en azul, el cual aparece en la parte superior izquierda. Esto le permite visualizar el concepto de función, ya que para cada valor de la variable X tiene un valor de $f(x)$; en esta situación se está aplicando la expresión 6.1

2. Con umbrales, éstos se emplean para calcular probabilidades. Un umbral permite estimar el área a la derecha o a la izquierda; su valor aparece en un recuadro superior a la derecha. Ahí también se indica la diferencia de $1 - el\ área$, abajo aparece el valor de la variable Z que corresponde a la normal estándar que se indica más adelante. Observe el cursor en la figura 6.6, el cual crea una didámica ya que lo puede mover a la derecha o a la izquierda, y en cada caso tendrá una nueva área: *probabilidad a la derecha o a la izquierda*.

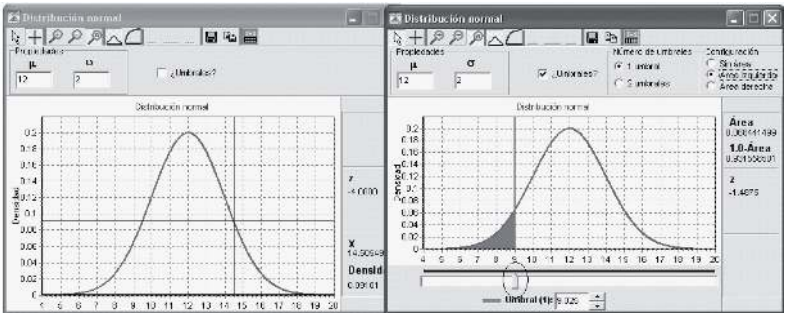


Figura 6.6 Características de la distribución normal para una media $\mu = 12$ y $\sigma = 2$.

3. Dos umbrales, colocados abajo, arriba, o uno arriba y otro abajo, permiten estimar la probabilidad entre dos valores de la variable. Muestra el valor de la diferencia de $1 - el\ área$, y los valores de la normal estándar respectivos a los valores de la variable X.

4. Para obtener percentiles, se mueven los umbrales considerando el valor de las áreas. En la gráfica, la precisión con los umbrales se da hasta milésimas.

5. Una mayor precisión en el cómputo de las probabilidades se obtiene usando un calculador que viene integrado, que se verá más adelante. Éste se activa accediendo al último cuadro que aparece en el segundo renglón a la derecha. Ahí se pide dar el valor de la media y desviación estándar, luego para obtener probabilidades se deben dar valores de la variable X. La otra opción es dar una probabilidad y saber a qué valores de la variable corresponden para una pareja (μ, σ).

6. Cuando escriba el valor de μ oprima enter para cambiar la media en la distribución, luego el valor de σ, y nuevamente oprima enter.

7. Aparecen dos figuras en azul que corresponden a la función densidad y acumulada de la normal, con éstas se puede obtener una u otra de manera alternativa.

8. Las lupas le permiten agrandar (+), empequeñecer (-) y actualizar (A) la figura.

9. Active el símbolo + para encontrar el valor de la función Densidad para un valor de X, para descativarlo use la flecha ubicada en la parte superior izquierda.

Características de la distribución de probabilidad normal

Como se describe en la figura 6.2 la media es el centro de la campana. Puesto que esta distribución es simétrica alrededor de μ, indica que 50 % de las observaciones estarán por debajo de la media y 50 %

serán mayores que la media. En este sentido la media es también la mediana de la distribución. En las figuras de abajo se señalan otras características relevantes de la distribución normal, las que se comentan a continuación.

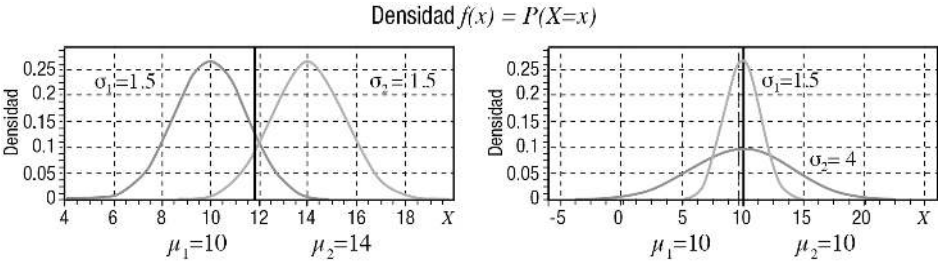


Figura 6.7 Distribución normal con diferentes medias y desviaciones estándar.

- **Diferentes medias y desviaciones estándar**. En la figura 6.7 se muestran cuatro densidades distintas para la distribución probabilidad normal. Claramente se percibe que si la desviación estándar (σ) crece, la densidad de la curva se hace más dispersa, gráfica derecha $\mu_1 = 10 \;\; \mu_2 = 10$ $\sigma_1 = 1.5 \;\; \sigma_2 = 4$. En el caso de diferentes valores de σ, se interpreta diciendo que hay mayor heterogeneidad en una población cuando la desviación estándar es mayor. La normal se mueve si media μ cambia, aquí se mantuvo fija la desviación estándar, $\mu_1 = 10 \;\; \mu_2 = 14 \;\; \sigma_1 = 1.5 \;\; \sigma_2 = 1.5$, gráfica a la izquierda. También se puede presentar el caso de diferentes medias y diferentes varianzas. Estos dos parámetros (μ, σ) cambian la apariencia de la normal.

- **Probabilidades y número de desviaciones estándar**. Aunque la curva de la normal es continua de manera infinita en ambas direcciones, la mayor parte de la distribución está dentro tres desviaciones estándar a cada lado de μ, ver figura 6.8. Para una distribución normal se tiene que:

  El 68.26 % de la distribución está entre $\mu - \sigma$ y $\mu + \sigma$ o, dicho de otra manera, la probabilidad del intervalo cuya longitud es de una desviación estándar a cada lado de la media es:

  $$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$

  El 95.44 % de la distribución está entre $\mu - 2\sigma$ y $\mu + 2\sigma$, de manera análoga, la probabilidad del intervalo cuya longitud es de dos desviaciones estándar a cada lado de la media es:

  $$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$

  El 99.74 % de la distribución está entre $\mu - 3\sigma$ y $\mu + 3\sigma$. Finalmente, la probabilidad del intervalo cuya longitud es de tres desviaciones estándar a cada lado de la media es:

  $$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

  En la figura 6.8 la parte sombreada corresponde a la distribución señalada o probabilidad indicada; estas tres relaciones se conocen como la *regla empírica*.

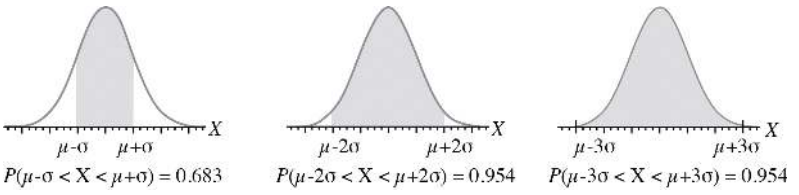


Figura 6.8 Probabilidades y áreas sombreadas a una, dos y tres desviaciones estándar de la media.

Regla empírica

En una distribución normal o aproximadamente simétrica, cerca del 68 % de las observaciones están a una desviación estándar de la media; alrededor del 95 % de las observaciones se hallan a dos desviaciones estándar de la media y por ahí del 99.7 %, en realidad todas las observaciones se encontrarán a tres desviaciones estándar de la media.

1. Reunimos en una sola gráfica esta representación para tener una visión integral del número de desviaciones estándar alrededor de la media (gráfica a la izquierda de la figura 6.9).

2. La distribución acumulada de la normal tiene una forma similar a la que aparece en la gráfica a la derecha, figura 6.9. A ésta se le agregó su relación con una desviación estándar. Esta descripción es importante para estimar el valor de la desviación estándar. Por ejemplo, la media corresponde a 0.5 de la distribución acumulada, eje vertical. Si sumamos a 0.5 la cantidad de 0.341[1] en el eje vertical tendremos el valor acumulado de 0.841. A partir de este valor se traza una línea horizontal a la curva. Luego, del punto de contacto con la curva se traza una línea perpendicular al eje horizontal; ese punto está a una desviación estándar, σ.

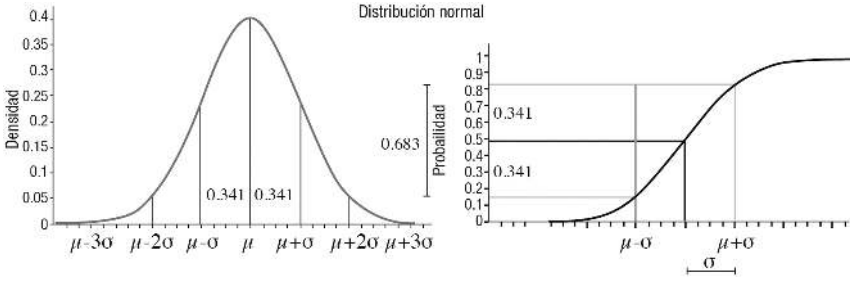


Figura 6.9 Izquierda, probabilidades a una, dos y tres desviaciones estándar de la media. Derecha, distribución de probabilidades acumuladas para la normal y la relación a una desviación estándar de la media.

[1] El punto 0.341 corresponde a la mitad de 0.683.

Ejemplo 6.1

La administración de una empresa tiene como estrategia realizar llamadas a sus clientes para conocer aspectos relevantes de su servicio. Entre los registros de la evalución de la atención, se tiene considerado el tiempo de la llamada. La *variable aleatoria tiempo de llamada* X, sigue una distribución normal con media $\mu = 215$ segundos, y una desviación estándar $\sigma = 35$ segundos Encuentre la probabilidad de que la llamada dure: 1.- De 180 segundos o menos. 2.- De 260 segundos o más. 3.- Entre 180 y 270 segundos inclusive.

Solución

1. La curva es la función de densidad, en esta etapa por la normal. El área bajo la curva limitada por el rango de la variable X, en este escenario entre $(-\infty, +\infty)$, infinito y más infinito, en símbolos $-\infty < X < +\infty$. Así el área comprende el 100 %, la proporción total es 1.0, el área está comprendida entre 0 y 1, es decir $0 \leq A(x) \leq 1$. Vea la figura 6.10.

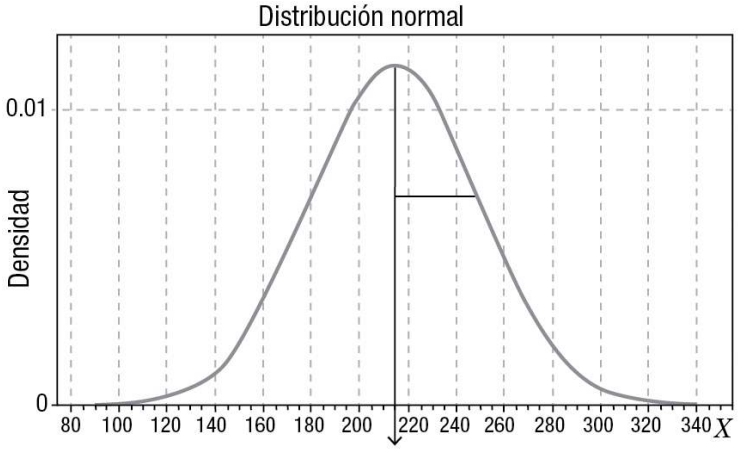


Figura 6.10 Distribución normal, con media $\mu = 215$ y desviación estándar $\sigma = 35$.

2. En la figura 6.11 describe un área limitada por la línea $X = 180$, un valor de la variable, la curva y el eje horizontal X que corresponde a la variable. Ésta representa una porción del área de total referida en el punto 1. Se hace un cálculo de esa área, en el **CalEst** muestra su valor, éste aparece señalado en la gráfica y es igual a 0.158. La explicación se da en el contexto de la figura 6.11. Esta área especifica la probabilidad de que el tiempo de llamada sea menor a 180 minutos, y se expresa mediante la siguiente fórmula: $P(X \leq 180) = 0.159$.

Probabilidad

Para que el usuario pueda conseguir este valor, se presenta una tabla, referida como tabla de la normal. Para alcanzar ésta, primero se transforma la variable original X en una nueva variable Z, la cual tiene una media $\mu = 0$ y varianza $\sigma^2 = 1$, desviación estándar $\sigma = 1$. Los detalles de este cambio se mostrarán en el siguiente apartado. Vale la pena comentar que este recurso de transformación evita que para cada par de media μ y varianza σ^2 se tenga una tabla, lo que resultaría laborioso. Claro que con el medio tecnológico de la actualidad ese cálculo se vuelve inmediato y prácticamente sencillo, tal y como se indica en el **CalEst**. El esfuerzo ahora consiste en asimilar los conceptos que nos dejan la distribución, su aplicación y la interpretación de los problemas.

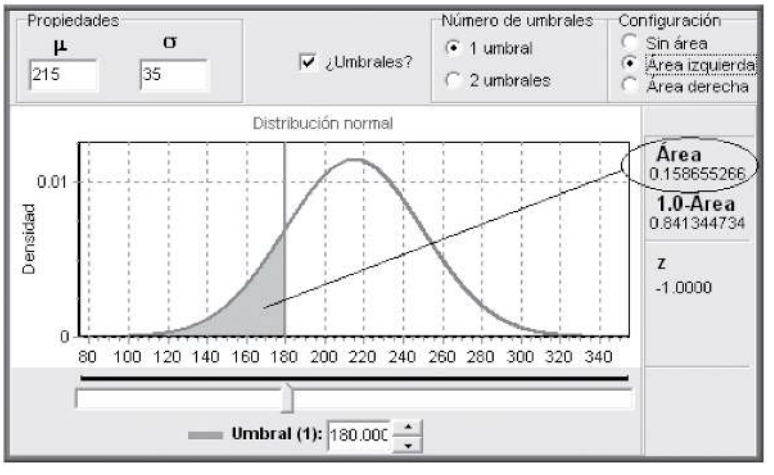


Figura 6.11 La probabilidad de que X sea menor o igual a 180.

3. Ahora, la probabilidad de que X sea mayor o igual a 260, $P(X \geq 260) = 0.099$. Este resultado se describe en la gráfica izquierda de la figura 6.12.

4. La gráfica a la derecha de la figura 6.12 muestra el cálculo de la probabilidad entre dos puntos x_1 y x_2, $P(180 \leq X \leq 270) = 0.783$, en general $P(x_1 \leq X \leq x_2)$. Un cálculo alternativo se obtiene mediante la siguiente expresión:

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$$

Para ejemplificar esta fórmula, primero se obtiene la $P(X \leq 270) = 0.942$, del inciso 1, $P(X \leq 180) = 0.159$, se sustituyen los valores y se obtiene que

$$P(x_1 \leq X \leq x_2) = P(180 \leq X \leq 270) = P(X \leq 270) - P(X \leq 180) = 0.942 - 0.159 = 0.783.$$

En resumen: la probabilidad representan una porción del área de la curva,

Caso 1. $P(X \leq x)$ figura 6.11. **Caso 2**. $P(X \geq x)$ gráfica izquierda, figura 6.12. **Caso 3.** $P(x_1 \leq X \leq x_2)$ gráfica derecha en la figura 6.12.

.

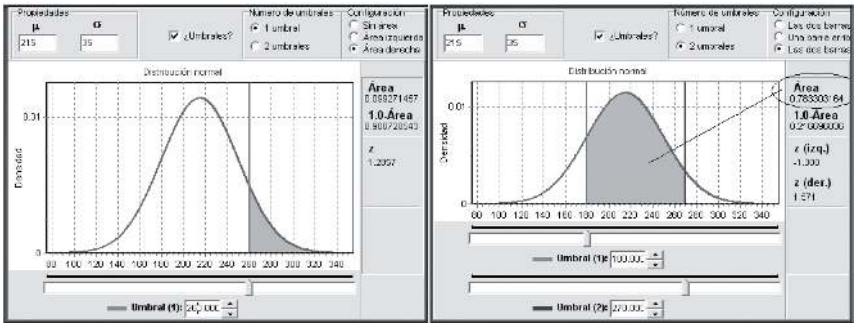


Figura 6.12 Izquierda: probabilidad a la derecha de X, $P(X \geq 260) = 0.099$. Derecha: probabilidad entre los puntos x_1 y x_2, $P(180 \leq X \leq 270) = 0.783$.

Ejemplo 6.2

Considerando el mundo de la información 1, se mostrará de manera intuitiva cómo calcular una probabilidad usando el histograma y relacionarlo con el caso 3.

1. Calcular la probabilidad de que un estudiante obtenga una calificación entre $x_1 = 7.0$ y $x_2 = 7.2$, es decir: $P(7.0 \leq X \leq 7.2)$.
2. Determinar la probabilidad de que un estudiante obtenga un promedio entre 7.5 y 9, esto es, calcular $P(7.5 \leq X \leq 9)$

Solución

1. El cómputo de $P(7.0 \leq X \leq 7.2)$ se hará con ayuda del histograma que se reproduce en la figura 6.13. Después, se formalizará el cálculo de probabilidad en una distribución normal. En principio la idea es calcular la probabilidad de que un alumno tenga una calificación entre 7 y 7.20, es decir:

$$P(7.0 \leq X \leq 7.2) = P(X \leq 7.2) - P(X \leq 7.0)$$

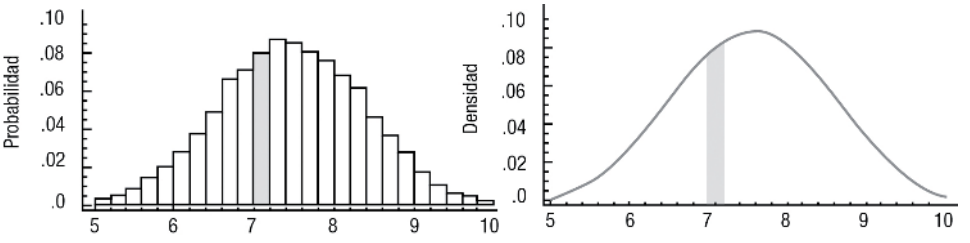


Figura 6.13 Probabilidad entre los números 7.0 y 7.2, $P(7.0 \leq X \leq 7.2) = 0.07954$.

El área correspondiente a la barra señalada del histograma es de 0.07954, lo que equivale a la probabilidad que se desea, es decir:

$$P(7.0 \leq X \leq 7.2) = 0.07954$$

La presentación gráfica de esta formulación se presenta en la figura 6.13. Una solución alternativa para calcular la probabilidad $P(7.0 \leq X \leq 7.2) = P(X \leq 7.2) - P(X \leq 7.0)$ es usar la gráfica de distribución de probabilidad acumulada. En la figura 6.14 se muestra el procedimiento para estimar esta probabilidad, donde el valor se indica de manera aproximada.

$$P(7.0 \leq X \leq 7.2) = P(X \leq 7.2) - P(X \leq 7.0)$$
$$P(X \leq 7.2) = 0.369$$
$$P(X \leq 7.0) = 0.289$$

En la gráfica se advierte que la probabilidad acumulada hasta 7.5 es de 0.369 y la probabilidad acumulada hasta 7.2 se aproxima a 0.289. La diferencia es 0.369-0.289=0.080, así $P(7.0 \leq X \leq 7.2) = P(X \leq 7.2) - P(X \leq 7.0) = 0.080$, que nos acerca al valor anterior, calculado mediante el histograma.

2. De manera gráfica, calcular la probabilidad $P(7.5 \leq X \leq 9)$ es una situación similar a la expuesta en la figura 6.13.

La probabilidad $P(7.5 \leq X \leq 9)$ puede obtenerse de manera aproximada usando la suma del área de los rectángulos, debida a la estrecha relación entre éstos y la gráfica de la normal; sin embargo, es preciso resaltar que este tipo de situaciones no se presentan siempre, tal situación se describe en la figura 6.15.

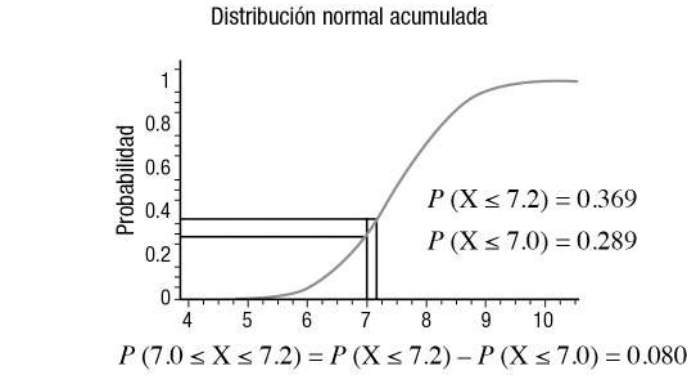


Figura 6.14 Cálculo de probabilidades usando la normal acumulada.

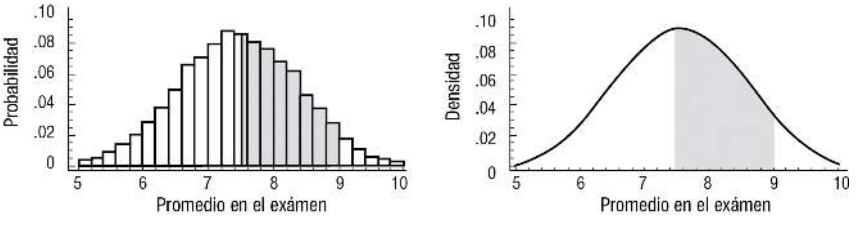


Figura 6.15 Probabilidad entre las calificaciones 7.5 y 9.

Podemos recurrir al uso de la gráfica acumulada como se ve en la figura 6.16. A continuación, se realizan los cálculos con apoyo a la salida que genera el programa.

$$P(7.5 \leq X \leq 9.0) = P(X \leq 9.0) - P(X \leq 7.5)$$
$$P(X \leq 9.0) = 0.952$$
$$P(X \leq 7.5) = 0.500$$

Por lo tanto: $P(7.5 \leq X \leq 9.0) = P(X \leq 9.0) - P(X \leq 7.5) = 0.952 - 0.5 = 0.452$. Se realiza la estimación aproximada aplicando las ideas expuestas en el inciso 1 del ejemplo 6.2 mediante la figura 6.13.

Comentarios:

1. Para la obtención de estos cálculos esencialmente se requiere de la ayuda del cómputo, tal y como se ha mostrado con las salidas del programa **CalEst**. Dada la media μ y la desviación estándar σ, se puede elaborar una gráfica como la descrita en la figura 6.16, donde el eje horizontal muestra los valores de la variable de interés en el problema de estudio, y el eje vertical la probabilidad que relaciona estos valores $P(X \leq x)$. Esto es de utilidad primordial porque se puede dar una probabilidad y encontrar el valor de la variable X, o dar el valor de la variable X y encontrar la probabilidad.

2. Las gráficas generan cálculos aproximados, sin embargo en el programa se tiene un calculador de

probabilidad que da valores con mayor precisión. Y realizar los cálculos como se indicó en el punto anterior. Ver figura 6.17.

3. En la práctica se recurre a un cambio de variable, para obtener las probabilidades de una distribución normal. Esto consiste en estandarizar la variable original X en una Z, ésta se denomina la variable normal estándar y se estudiará en el siguiente apartado. La ventaja de esta es que genera una distribución de probabilidad normal, por decirlo así única: estándar. Sin embargo para tener el valor de la variable original habrá que hacer una operación extra.

4. Podemos decir que al usar el CalEst, la normal estándar queda como un caso particular. Sin embargo se presentará esta última con detalle.

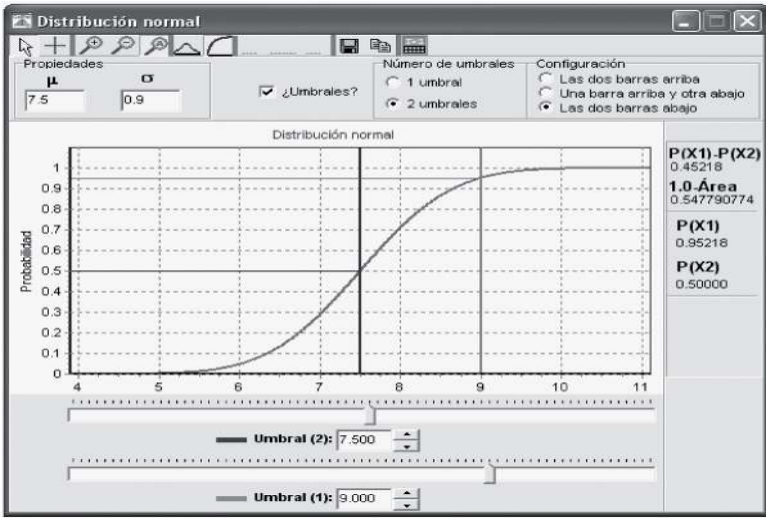


Figura 6.16 Salida del programa CalEst que muestra la distribución normal acumulada con los cálculos para el ejemplo.

Cálculo de la probabilidad con el CalEst

Una vez que usted está familiarizado con el cálculo de la probabilidad en la distribución normal, use el CalEst puede obtener estos valores, recuerde que equivale a calcular el área de bajo la curva para el o los valores de referencia. Lo importante en este punto es que puede obtener el valor de la probabilidad para cualquier pareja de parámetros de la normal, la media μ y la desviación estándar σ. También, dada la probabilidad, se puede saber a qué valor o valores corresponde. Procedimiento: apriete el cuadro que describe una calculadora. Aparecen las opciones: Directa o Inversa, figura 6.17.

1. Si es directa, el paso es indicar el valor de los parámetros μ y σ, y enseguida el valor o los valores de la variable. A continuación señalar si se trata de una probabilidad entre dos puntos, o ya sea a la izquierda o derecha del valor.

2. Si es Inversa, nuevamente indicar los valores de los parámetros μ y σ, así como el valor de la probabilidad, y a continuación indicar qué valor o valores de la variable se quieren encontrar.

.

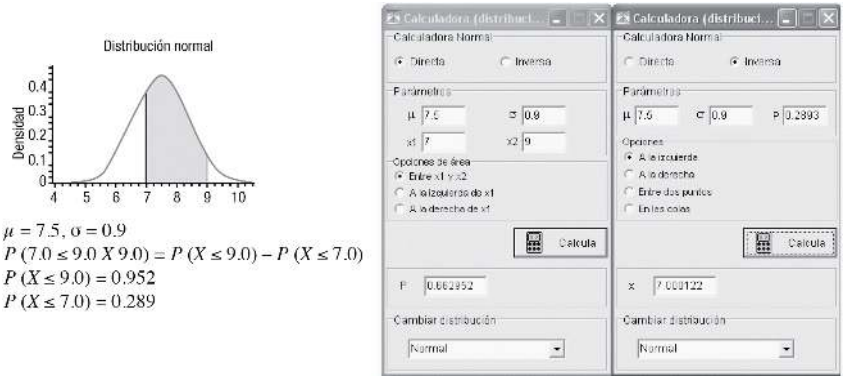


Figura 6.17 Operación de probabilidades usando el Calculador de Probabilidad en **CalEst**.

6.3 Distribución normal estándar

El cálculo de probabilidades usando la variable aleatoria real es complicado. Por ello, se recurre a un procedimiento matemático que mediante un cambio de variable, facilita el trabajo operativo. Dado que el cálculo de la densidad normal estándar es más sencillo, también puede construirse la distribución de probabilidad acumulada de la normal estándar, y en este caso el cálculo de probabilidades se hace más simple.

Usar tablas para encontrar probabilidades de una distribución normal con parámetros μ y σ puede ser una idea buena, pero existen limitaciones prácticas porque los valores μ y σ no están restringidos. Éstos dependen de la variable de interés, por lo que existe una infinidad de parejas de μ y σ que pueden ser de interés. No es factible tener una tabla para cada posible par.

Este problema se resuelve por el método que trasforma cada variable aleatoria X que se distribuye como una normal en una variable aleatoria normal estándar Z.

Procedimiento de estandarización de la distribución de probabilidad normal

Al valor de la variable aleatoria se le resta la media μ y se divide entre la desviación estándar σ. La

representación de esta operación se escribe por la expresión:

$$Z = \frac{X - \mu}{\sigma} \tag{6.2}$$

La variable aleatoria Z corresponde a una normal estándar. La representación gráfica de esta operación se describe en la figura 6.18.

La distribución normal estándar tiene una media de 0 y una desviación estándar de 1.

La distribución normal estándar (variable aleatoria Z) tiene una densidad, cuya forma es una campana con:

Media $\mu = 0$
Desviación estándar $\sigma = 1$

La distribución normal estándar se denota por $N(0, 1)$ Z se distibuye como una normal y se indica $Z \sim N(0, 1)$.

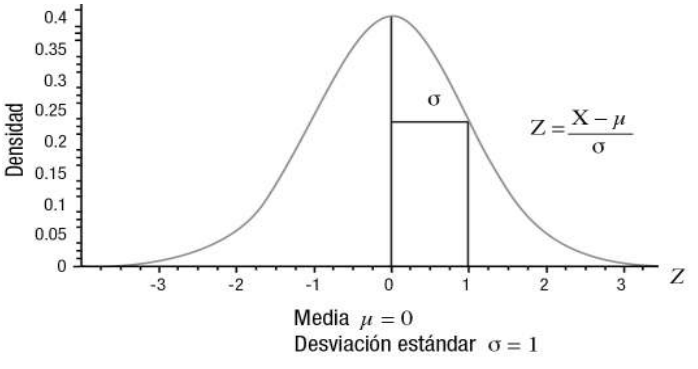


Figura 6.18 Distribución de probabilidad normal estándar.

Ejemplo 6.3

El objetivo de este ejemplo es calcular probabilidades de una distribución normal usando el procedimiento de estandarización. El valor de la variable X corresponde a las calificaciones, μ y σ son los parámetros que corresponden a la media y a la desviación estándar.

Solución (desde un punto de vista clásico)

La intención es calcular la probabilidad de las calificaciones entre 7.5 y 9, por lo que se sustituyen estos valores en la expresión Z y se obtienen los siguientes resultados:

$$Z = \frac{X - \mu}{\sigma},\ Z = \frac{7.5 - 7.501}{0.911} = 0 \text{ y } Z = \frac{9 - 7.501}{0.911} = 1.645$$

Estos valores se buscan en la tabla de valores de Z en una distribución normal estándar, la cual se

encuentra en el anexo Tablas. En la Tabla 6.1 se reproduce una porción de la tabla normal estándar. En ésta aparecen tres columnas de valores de Z con sus respectivas probabilidades acumuladas; vea los valores de ésta para el cálculo del valor de $Z = 1.645$. Por los cálculos realizados para Z, se pretende obtener la probabilidad acumulada de 0 y para 1.645, esto es: $P(Z \leq 0)$ y $P(Z \leq 1.645)$

Estos valores son:

$$P(Z \leq 0) = 0.5, \quad P(Z \leq 1.645) = 0.950$$

La diferencia entre estas probabilidades corresponde a:

$$P(0 \leq Z \leq 1.645) = P(Z \leq 1.645) - P(Z \leq 0) = 0.95 - 0.5 = 0.45$$

Así, para el problema de las calificaciones la expresión anterior equivale a:

$$P(7.5 \leq X \leq 9) = P(X \leq 7.5) - P(X \leq 9.0) = 0.45$$

Por último, se determina la probabilidad que se deseaba. En ese sentido, 45 % de los estudiantes que presentaron el examen alcanzarán una calificación promedio entre 7.5 y 9.

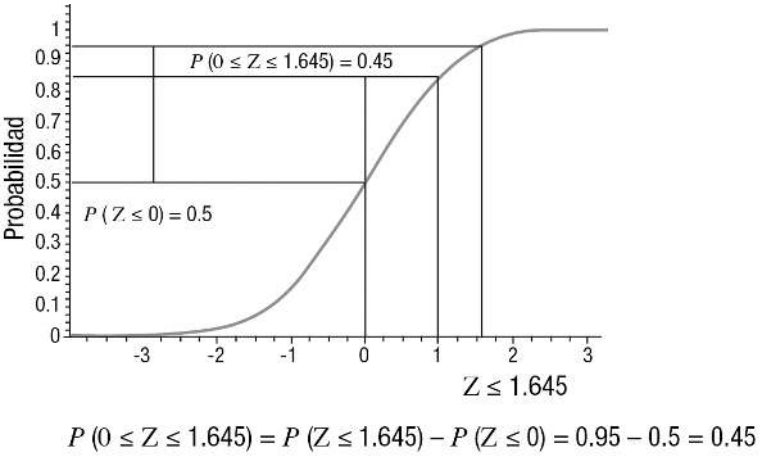


Figura 6.19 Distribución de probabilidad normal estándar acumulada.

En la figura 6.19 se describe la distribución normal estandarizada de manera acumulada, y se reproduce el valor de probabilidad acumulada a partir del valor de $Z = 1.645$.

La tabla normal estándar es una tabla de probabilidades para una variable aleatoria Z. A continuación se describe el procedimiento para usar la tabla de la normal estándar y así calcular probabilidades. La tabla de la normal estándar da el área a la izquierda de un valor específico de z:

$P(Z \leq z) =$ Área bajo la curva a la izquierda de z
$P(a \leq Z \leq b) =$ (Área a la izquierda de b)-(Área a la izquierda de a)
$P(0 \leq Z \leq 1) = 0.341$

Propiedades derivadas de la simetría de una curva normal estándar.

i. $P(Z \leq 0) = 0.5$
ii. $P(Z \leq -z) = 1 - P(Z \leq z) = P(Z \geq z)$

Observe que en la figura 6.20 se muestra la aplicación de estas propiedades de simetría con el valor de $z = 1.5$, o $z = -1.5$.

Tabla 6.2 Valores de Z y su probabilidad acumulada.

Z	$P(Z \leq z)$	Z	$P(Z \leq z)$	Z	$P(Z \leq z)$
-3	0.001350	-3.00	0.001350	0.500	0.691462
-2	0.022750	-2.50	0.006210	1.000	0.841345
-1	0.158655	-2.00	0.022750	1.040	0.850830
0	0.500000	-1.96	0.024998	1.200	0.884930
1	0.841345	-1.80	0.035930	1.400	0.919243
2	0.977250	-1.60	0.054799	1.600	0.945201
3	0.998650	-1.40	0.080757	1.645	0.950015
		-1.20	0.115070	1.800	0.964070
		-1.00	0.158655	2.000	0.977250
		-0.80	0.211855	2.200	0.986097
		-0.60	0.274253	2.400	0.991802
		-0.40	0.344578	2.600	0.995339
		-0.20	0.420740	2.800	0.997445
		0.00	0.500000	3.000	0.998650

Probabilidad de un intervalo

La probabilidad de un intervalo es la probabilidad de que una variable aleatoria tome un valor entre dos puntos dados de la variable X, es decir, $P(x_1 \leq X \leq x_2)$. Para la variable estandarizada Z, se tiene: $P(z_1 \leq Z \leq z_2)$.

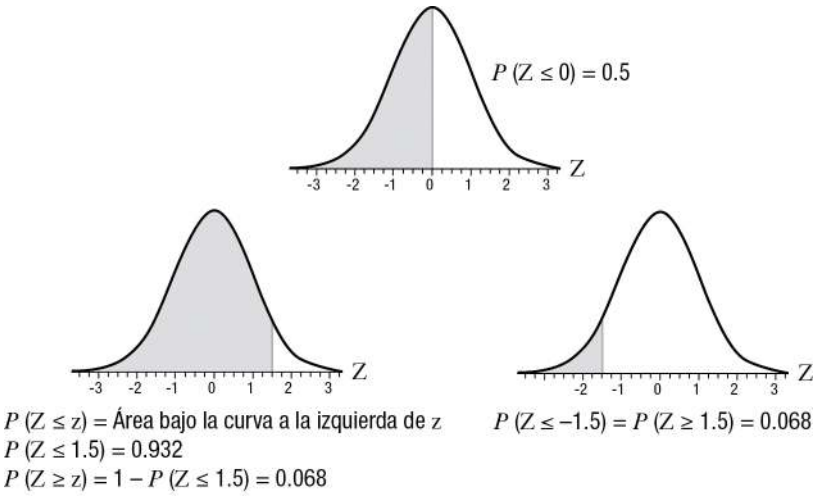


Figura 6.20 Propiedades de simetría de la curva normal.

Solución mediante el uso de CalEst

Con el **CalEst** aplique el procedimiento descrito para el uso del calculador. Entonces:

1. Si es directa, el paso es indicar el valor de los parámetros $\mu = 0$ y $\sigma = 1$, enseguida el valor o los valores de la variable. A continuación señalar si se trata de una probabilidad entre dos puntos, o ya sea a la izquierda, o derecha del valor.

2. Si es Inversa, nuevamente indicar los valores de los parámetros $\mu = 0$ y $\sigma = 1$, así como el valor de la probabilidad, y a continuación indicar qué valor o valores de la variable se quieren encontrar.

Ejemplo 6.4

Encontrar las probabilidades $P(Z \leq 1.35)$ y $P(Z > 1.35)$.

Solución

En la figura 6.21 se reproduce una proporción de la tabla de la normal que aparece en el anexo A. En la primera columna y en el renglón uno están dados los valores de la variable Z. Por ejemplo, para $z = 1.35$ se busca en la columna el número 1.3 y en el renglón el número 0.05, donde se interceptan estos valores

nos encontramos con el valor 0.9115. Este valor corresponde a la distribución de probabilidad acumulada para la normal en el punto $z = 1.35$, es decir, $P(Z \leq 1.35) = 0.9115$. En la gráfica se muestra el área relacionada a $P(Z > 1.35)$.

En consecuencia,

$$P(Z \leq 1.35) = 0.9115$$

Como se advierte $(Z > 1.35)$ es el complemento de $(Z \leq 1.35)$, por tanto,

$$P(Z > 1.35) = 1 - P(Z \leq 1.35) = 1 - 0.9115 = 0.0885$$

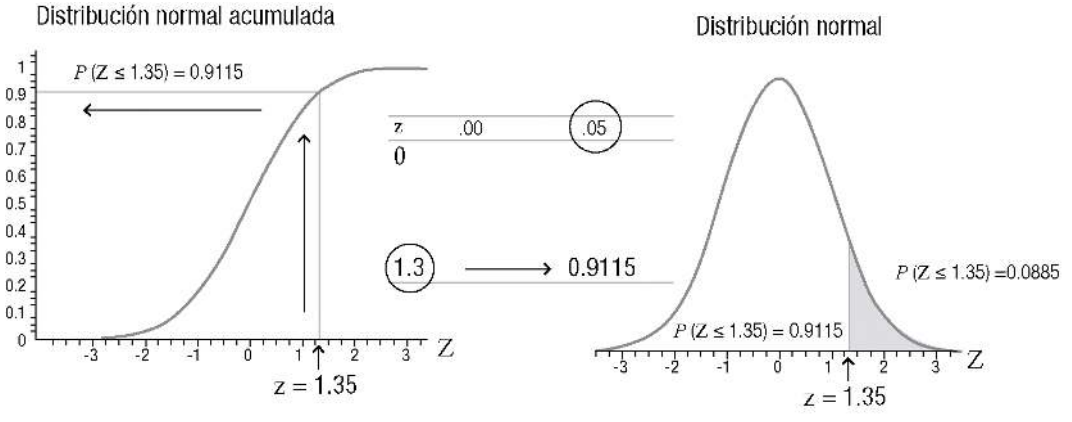


Figura 6.21 Cálculo de probabilidades en la normal estándar, $P(Z > 1.35)$.

Ejemplo 6.5

En la práctica es importante determinar los valores de las variables para una probabilidad dada. Encontrar el valor de z que indique que el 2.5 % de los productos están fuera de especificación, es decir, la z que cumpla $P(Z > z) = 0.025$ ver figura 6.22.

Solución

Usamos la propiedad de que:

$$P(Z \leq z) = 1 - P(Z > z) = 1 - 0.025 = 0.975$$

Ahora, localizamos el valor dentro de la tabla, y se observa el valor correspondiente de Z. En este caso $z = 1.96$.

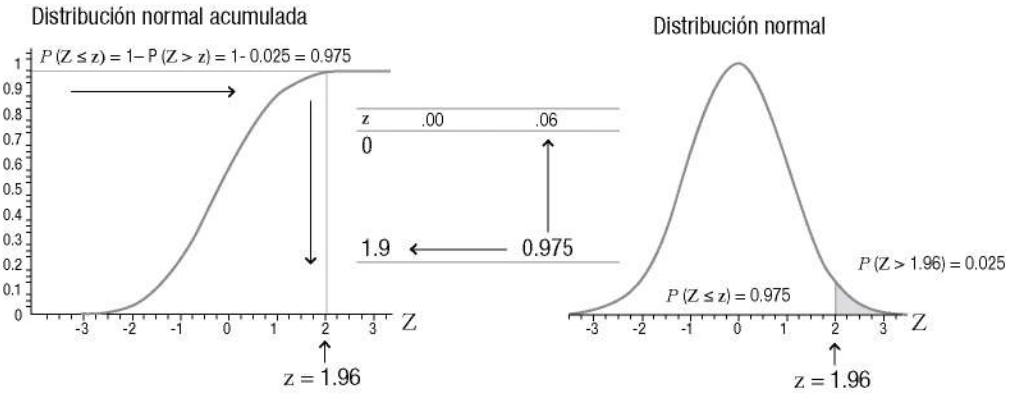


Figura 6.22 Cálculo de probabilidades en la normal estándar, $Z > 1.95$.

Ejemplo 6.6

En pruebas de laboratorio se estableció que el rendimiento en promedio de un coche de cuatro cilindros es de $\mu = 12.6$, número de kilómetros por litro (nkl), con una desviación estándar 0.8. Se desea encontrar la probabilidad de que el rendimiento sea menor que 12 nkl.

Solución

Si X es la variable aleatoria que representa nkl y es razonable suponer que tiene una distribución de probabilidad normal, entonces se busca $P(X < 12)$. Para calcular la probabilidad, relacionamos ésta con la normal estándar, de modo que:

$$Z = \frac{X - \mu}{\sigma} \leq \frac{12 - 12.6}{0.8} = -0.75$$

Entonces,

$$P(X \leq 12) = P(Z \leq -0.75) = 0.2266$$

Es una probabilidad relativamente pequeña para tener un rendimiento al menos de 12 kilómetros por litro, situación que ayuda una mejora económica.

Percentiles en una distribución normal

Para encontrar el p-enésimo percentil x_p en una distribución normal con media μ y desviación estándar σ mediante la normal estándar. Primero se busca el percentil de una normal estándar z_p y luego se aplica la siguiente relación:

$$x_p = \mu + \sigma z_p,$$

para encontrar el percentil para x_p. Esta expresión convierte el percentil de la normal estándar al percentil de una distribución normal con parámetros μ y σ. Esta fórmula es importante en la inferencia estadística.

Procedimiento para encontrar el percentil de una normal

El promedio de p-ésimo percentil, x_p, $0 < p < 1$, de una distribución normal con media μ y desviación estándar σ es:

1. Redondear p a milésimas y luego encontrar el valor de z_p en el renglón y la columna en la tabla de la normal que está en el anexo. Éste es el p-ésimo percentil de una normal estándar (ejemplo 6.3).
2. Convertir x_p aplicando la expresión:

$$x_p = \mu + \sigma z_p$$

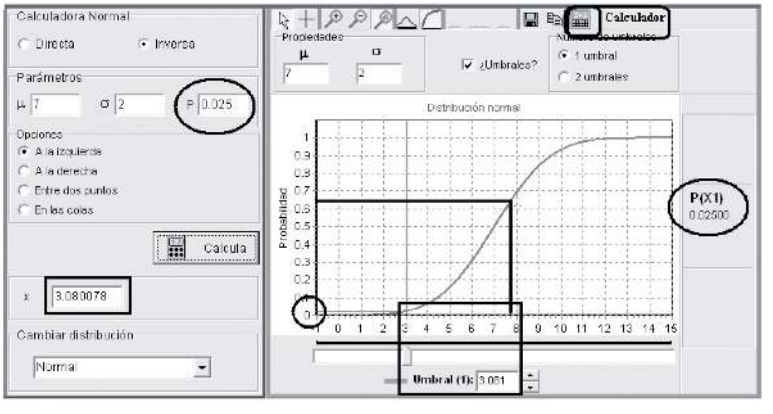


Figura 6.23 Descripción del calculador y la gráfica de percentiles.

Ejemplo 6.7

Encontrar el percentil 0.0249 aproximado a 0.025 de una normal media $\mu = 7$ y desviación estándar $\sigma = 2$.

Primero se averigua el valor de z que corresponde a ese percentil, esto es: $P(Z \geq z) = 0.025$. De la tabla 6.2, se obtiene que $z_{0.025} = -1.96$. Observe que la marca en el extremo superior derecho de la figura muestra donde está el calculador, note que los tres círculos indican la probabilidad, en el cuadro se muestra el valor de la variable con las características que se piden, la utilidad de la calculadora. Para determinar el valor de x_p, aplicamos la fórmula establecida, en la que:

$$x_p = \mu + \sigma z_p = 7 + 2 \times (-1.96) = 3.08$$

La gráfica a la derecha en la figura 6.23 muestra la distribución de probabilidad con los parámetros ($\mu = 7$, $\sigma = 2$). Observe que la gráfica de la distribución de probabilidad acumulada, también se emplea para calcular los percentiles. El procedimiento es identificar el percentil en el eje vertical y a partir de ese valor trazar una línea paralela a la horizontal hasta la gráfica y luego la perpendicular al eje horizontal, y el punto que intercepta es el que corresponde a ese percentil. En la gráfica, además del 0.025 percentil se trazó de manera aproximada el percentil 0.65 y corresponde a $x_p = 7.7$.
Para los cálculos con la normal estándar, use *el calculador* como se muestra en la figura 6.23 para una media $\mu = 0$, desviación estándar $\sigma = 1$ y una probabilidad (percentil) de $P = 0.025$, a continuación indicar el valor a la izquierda y enseguida aparece el valor de la variable Z.

Luca Pacioli

Uno de los primeros estadísticos fue Luca Pacioli, de nombre completo Fray Luca Bartolomeo de Pacioli o Luca di Borgo San Sepolcro, cuyo apellido también aparece escrito como Paccioli y Paciolo (Sansepolcro, 1445 - 1517), fue un fraile franciscano y matemático italiano, precursor del cálculo de probabilidades.

Ejemplo 6.8

El coordinador de la administración de una agencia de servicios lleva el registro del tiempo de atención a clientes. Bajo el supuesto de que esta variable tiene una distribución normal con $\mu = 17$ minutos y desviación estándar $\sigma = 3$ minutos. Calcule las siguientes probabilidades estudiar esta situación.
1. Determinar la probabilidad de que el tiempo de atención esté entre 13 y 22 minutos.
2. Más de 22 minutos.
3. Entre 15.5 y 18.5 minutos.

Solución

Se resolverá este problema usando el paquete estadístico. Nuevamente en la opción Probabilidad, seleccionamos la que se refiere a la Distribución de probabilidad normal, en ésta aparecerá una pantalla como la que se muestra en la figura 6.5. Para poder calcular probabilidades nos guiaremos por los umbrales. Hay tres elecciones posibles para operar con éstos. La elección depende de la comodidad del usuario, aquí hemos seleccionado la que está en medio. Luego en Propiedades aparecen los parámetros de la distribución normal. Ahí se han puesto los que indica el ejemplo. Cabe observar que para agrandar, achicar o mover la gráfica se usan los tres cuadros de escalas. Ahora procedemos al cálculo de las probabilidades mencionadas.

1. Para el primer punto se pide calcular la probabilidad $P(13 \leq X \leq 22)$, lo que se hace en este caso, es poner un umbral en 13 y el otro en 22 (figura 6.24). En el siguiente cuadro aparece la solución, es decir, los valores de la variable estandarizada y el área que representa la probabilidad que se pide encontrar. En símbolos esto es:

$$Z_1 = \frac{X-\mu}{\sigma} = \frac{13-17}{3} = -1.33, \quad Z_2 = \frac{X-\mu}{\sigma} = \frac{22-17}{3} = 1.67$$
$$P\left(\frac{13-17}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{22-17}{3}\right) = P(-1.33 \leq Z \leq 1.67) = 0.861$$
$$P(13 \leq X \leq 22) = 0.861$$

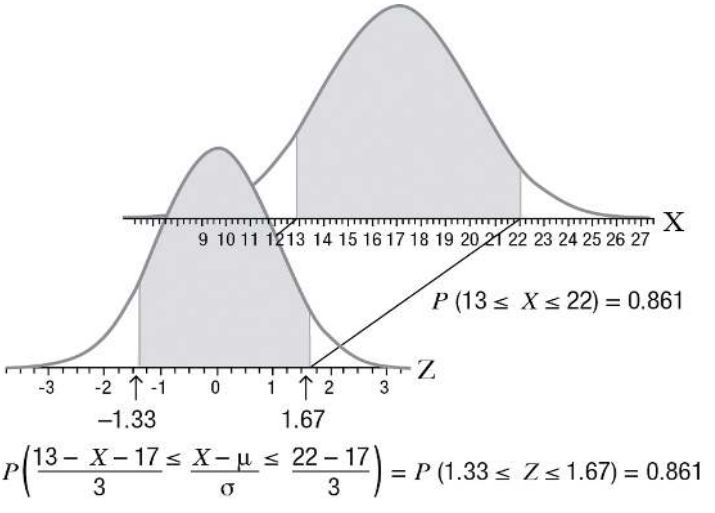


Figura 6.24 Cálculo de probabilidad en una normal con referencia a la norma estándar.

2. La probabilidad $P(X > 22)$ se obtiene ahora moviendo sólo un umbral, en este caso el de abajo que nos lleva a la derecha de la distribución gráfica a la izquierda en la figura 6.25. El otro se mantiene a la izquierda lo más posible, y la solución la puede verificar en su paquete. Ahí se observa lo que a continuación se expresa con símbolos.

$$Z = \frac{X-\mu}{\sigma} = \frac{22-17}{3} = 1.67$$
$$P\left(\frac{X-\mu}{\sigma} > \frac{22-17}{3}\right) = P(Z > 1.67) = 0.047$$
$$P(X > 22) = 0.047$$

3. De manera análoga se obtiene el cálculo de la tercera probabilidad, gráfica a la derecha en la figura 6.25, es decir:

$$P\left(\frac{15.5-17}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{18.5-17}{3}\right) = P(-0.5 \leq Z \leq 0.5) = 0.383$$
$$P(15.5 \leq X \leq 18.5) = 0.383$$

Solución utilizando la distribución normal estándar de CalEst

Mediante el calculador estadístico CalEst se pueden obtener probabilidades de la distribución normal estándar. Esto ofrece una muy buena alternativa al uso de las tradicionales tablas. Lo único que se debe realizar en esta situación es mover los umbrales, tal como se indicó anteriormente.

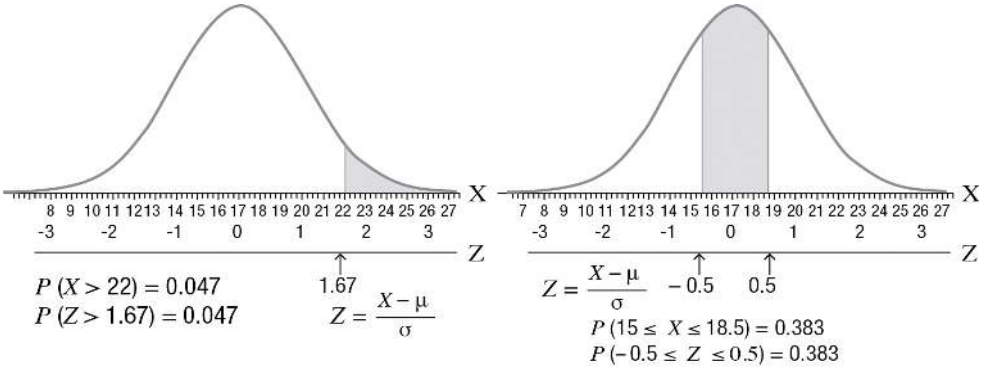


Figura 6.25 Cálculos de probabilidad en la distribución normal en las escalas X y Z.

Comentario: En la inferencia estadística por lo general se proponen uno o dos puntos de referencia para tomar una decisión con base en la información proporcionada por los datos.

El mundo de la información 2. Agencias de colocación

Hoy en día se han creado empresas que capacitan o captan a personas para que puedan conseguir empleo. Los gerentes de esas empresas tienen elaborada una batería de pruebas de aptitud, cuya finalidad es que sus clientes potenciales tengan la garantía de contratar personas con las habilidades que se requieren.

Preguntas sobre la naturaleza del problema

¿Cuál es la variable aleatoria? ¿Qué distribución tiene y cuáles son los valores de los parámetros? ¿Qué información requeriría un cliente? **Información:** La variable aleatoria es la calificación de la prueba y tiene una distribución normal con media $\mu = 75$ y varianza $\sigma^2 = 25$ ($\sigma = 5$) puntos. El interés está en que se necesita conocer el valor de los puntos en que la proporción de las personas que aplicaron la prueba es de 0.95, y con esa información saber si se cumple la expectativa de que el puntaje de aptitud esté entre 65 y 85 puntos. Se busca encontrar los puntos x_1 y x_2 en la expresión: $P(x_1 \leq X \leq x_2) = 1-\alpha = 0.95$. Utilizando el calculador como se explicó en el ejemplo 6.7, se tiene que: $P(65.2 \leq X \leq 84.8) = 1 - \alpha = 0.95$, es decir, los valores correspondientes son: $x_1 = 65.2$ y $x_2 = 84.8$. Interpretando esta información, tenemos que se cumplen las expectativas del cliente. Así el valor de $\alpha = 0.05$, considerando estos dos valores, se toma la información para los extremos izquierdo y derecho, de esa manera el valor de α se divide entre 2. *En resumen*, este es un punto clave para el trabajo futuro de los próximos capítulos, las expresiones siguientes describen esta proporción usando la variable original X: aptitud medida en puntos y la variable estandarizada Z; estas ideas se muestran en la figura 6.26.

$$P(X < 65.2) = \alpha/2 = 0.025 \quad P(65.2 \leq X \leq 84.8) = 1 - \alpha = 0.95 \quad P(X > 84.2) = \alpha/2 = 0.025$$
$$P(Z < -1.96) = \alpha/2 = 0.025 \quad P(-1.96 \leq Z \leq 1.96) = 1 - \alpha = 0.95 \quad P(Z > 1.96) = \alpha/2 = 0.025$$

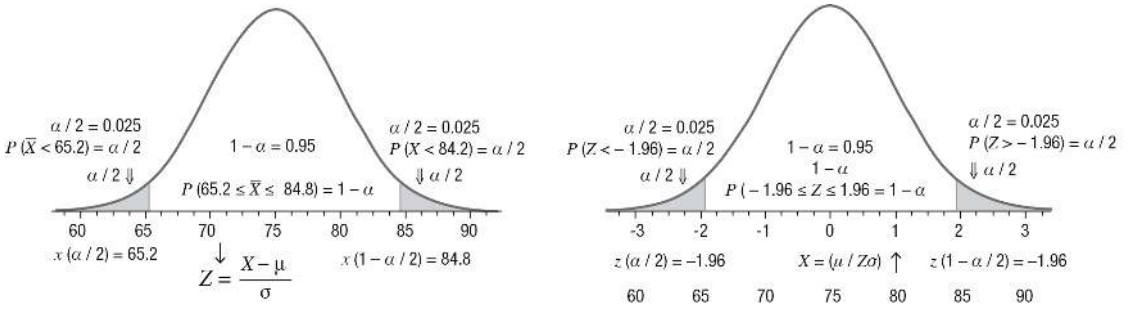


Figura 6.26 Obtención de dos valores para una probabilidad establecida.

Ejemplo 6.9

Distribución empírica y distribución teórica. La administración de una compañía registró el tiempo, en segundos, que dura una llamada con un cliente en el que se le ofrece un servicio. Se piensa que esta variable X (tiempo de llamada) presenta una distribución normal. En el registro se tomó una muestra aleatoria de 4000 personas. La distribución acumulada del tiempo de llamada se muestra con la línea poligonal en la figura 6.27. A esta línea poligonal se le conoce como distribución empírica y se compara con una distribución normal. Pero, como se advierte en este caso, las distribuciones son muy aproximadas.

De manera intuitiva se dice que el tiempo de espera tiene una distribución normal[2].

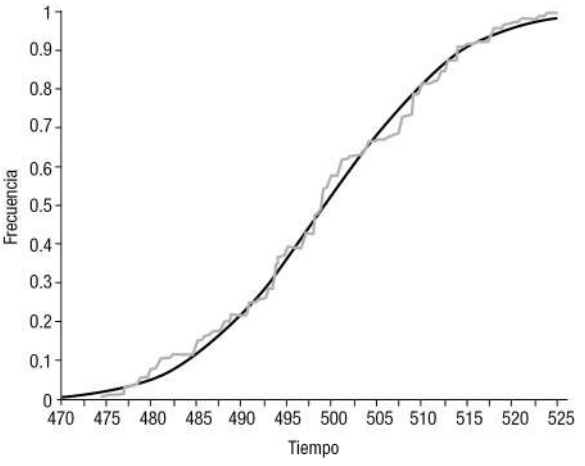


Figura 6.27 Relación entre una distribución empírica y la distribución de una normal (teórica).

6.4 Distribución χ^2

Nota. Las siguientes tres distribuciones: χ^2, $t - Student$, y F, se aplicarán en los siguientes capítulos, en una primera lectura de este capítulo las puede saltar y regresar cuando lo requiera.

Si $Z_1, Z_2, ...Z_n$ son variables normales estándar independientes, entonces se dice que la variable χ está definida por:

$$\chi = Z^2 + Z_2^2 + ... + Z_n^2$$

y tiene una distribución Ji (Chi) cuadrada con n grados de libertad. Se denota $\chi \sim \chi_n^2$ para indicar que χ tiene una distribución Ji cuadrada con n grados de libertad.

Casos de estudio

1. El ingreso mensual de una muestra de 20 administradores, recién egresados de una universidad, seleccionados aleatoriamente, tienen una media muestral de 8 mil pesos y una desviación estándar

[2]Existen procedimientos estadísticos con mayor precisión para verificar si una variable aleatoria tiene una distribución normal.

de 550 pesos.

2. Una compañía produce bolsas de cacahuate. El proceso genera miles de bolsas y cada una debe tener el mismo peso, de no ser así le ocasionará pérdidas económicas. Sin embargo existe una variación en el peso de cada bolsa, la cual debe ser mínima. Si la población de pesos tiene una distribución normal, en este caso se desea hacer inferencia estadística sobre la varianza o desviación estándar.

En ambos ejemplos se tiene interés en estudiar la varianza σ^2, de nueva cuenta se plantea la relación entre población y muestra (figura 6.28). En el procedimiento de seleccionar una muestra nos conduce a plantear una nueva variable tal y como se indica a continuación.

Definición:

Si la variable aleatoria X tiene una distribución normal entonces la variable aleatoria

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

tiene un distribución ji o chi cuadrada para cualquer muestra de tamaño $n > 1$.

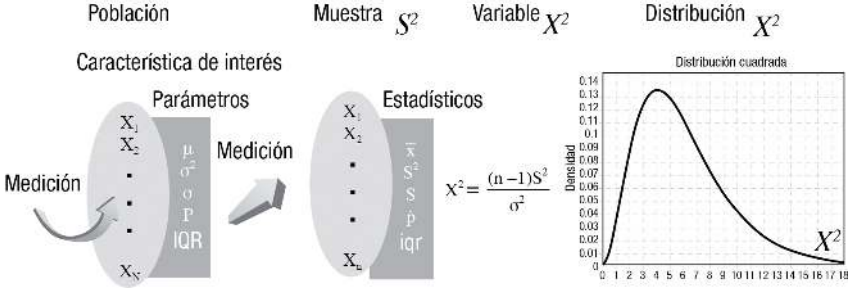


Figura 6.28 Construcción de la variable χ^2 considerando la relación población y muestra.

Propiedades de la distribución

1. Todos los valores de la distribución χ^2 son positivos.
2. La forma de la distribución depende de los grados de libertad, están dados por la muestra, es decir $gl = n - 1$.
3. El área bajo la curva es igual a 1.
4. Esta distribución es sesgada positivamente. Ver figura 6.29.

Esta distribución será de utilidad en el estudio de inferencia estadística sobre la varianza y en pruebas de bondad de ajuste que se estudiarán más adelante. Para llevar a cabo está este procedimiento, es

necesario determinar los valores críticos en la distribución χ^2 con $n-1$ grados de libertad, los cuales están relacionados con la probabilidad de esta distribución.

En esta dirección se especifica un valor α tal que: $(0 < \alpha < 1)$ y se calcula alguna de las siguientes tres probabilidades:

$$P(\chi^2 > \chi^2(n-1, 1-\alpha)) = \alpha$$
$$P(\chi^2(n-1, \alpha) < \chi^2) = 1-\alpha$$
$$P(\chi^2(\alpha/2, n-1) < \chi^2) = \alpha/2,\ P(\chi^2(1-\alpha/2, n-1) > \chi^2) = \alpha/2.$$

Estas probabilidades se obtienen directamente usando el **CalEst** y se tiene la ventaja de conocer la distribución, además de saber que se está calculando la probabilidad para diferentes valores de la variable aleatoria. Por lo general, la mayoría de libros la calculan mediante el empleo de tablas para la χ^2 y se obtiene mediante la expresión $P(\chi^2 > \chi^2(n-1, 1-\alpha)) = \alpha$.

Guía para encontrar los valores de la χ^2

1. Especifique el nivel de significancia α (probabilidad α).
2. Determine los grados de libertad $gl = n$.
3. Los valores de la distribución χ^2 se encuentran en la gráfica de la distribución Chi cuadrada en **CalEst**, o con la opción de la tabla que muestra el ambiente de la pantalla.
   a) Use el umbral para moverse a la derecha o izquierda según el valor de α.
   b) Use dos umbrales que correspondan a $\frac{1}{2}\alpha$ y 1-$\frac{\alpha}{2}$.

La distribución χ^2 se utiliza para hacer inferencia sobre la varianza (prueba de hipótesis e intervalos de confianza). Ver el capítulo de Prueba de Hipótesis para una población, así como la aplicación de la χ^2 en las pruebas de bondad de ajuste.

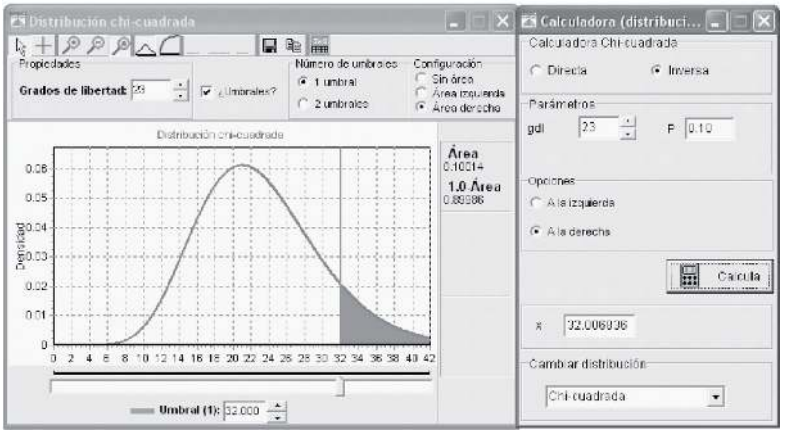


Figura 6.29 Descripción de la distribución usando **CalEst**, y los elementos para calcular las probabilidades.

Ejemplo 6.10

Encontrar el valor crítico χ^2_{cd} a la derecha cuando $n = 24$ y $\alpha = 0.10$.

Solución mediante el uso de CalEst

Los grados de libertad son $n - 1 = 24 - 1 = 23$. La gráfica de la figura 6.29, **CalEst**, muestra una χ^2 con 23 grados de libertad y un área sombreada (probabilidad) de $\alpha = 0.10$ en la parte derecha, por lo tanto el valor crítico es $\chi^2_{cd} = 32$.

Ejemplo 6.11

Se repite el ejemplo anterior con el fin de ilustrar la probabilidad en las colas de la distribución χ^2. Encontrar el valor crítico χ^2_{cd} a la derecha cuando $n = 24$ y $\alpha = 0.10$.

Solución

Los grados de libertad son $n - 1 = 24 - 1 = 23$. La gráfica de la figura 6.30 izquierda muestra una χ^2 con 23 grados de libertad y un área sombreada (probabilidad) de $\alpha = 0.10$ en la parte derecha. $\chi^2 = 32$ es decir: $P(\chi^2 > \chi^2(23, 0.90) = 32) = 0.10$, finalmente $P(\chi^2 \geq 32) = 0.10 \quad gl = 23$.

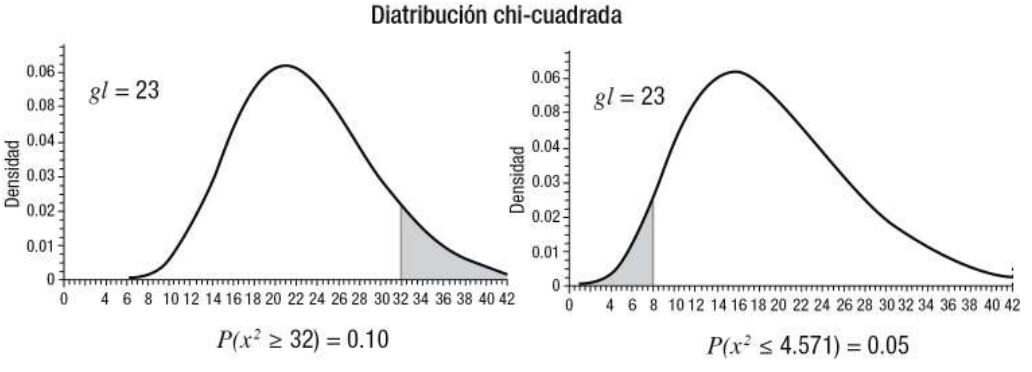


Figura 6.30 Probabilidades a la izquierda y derecha de la distribución.

Ejemplo 6.12

Encontrar el valor crítico χ^2_{ci} a la izquierda cuando $n = 12$ y $\alpha = 0.05$.

Solución

Los grados de libertad son $n = 12 - 1 = 11$. La gráfica de la figura 6.30 derecha muestra una χ^2 con 11 grados de libertad y el área sombreada a la izquierda de $\alpha = 0.05$, es decir: $P(\chi^2 \leq 4.571) = 0.05$ $gl = 11$. El área a la derecha de esta distribución es $1 - \alpha = 1 - 0.05 = 0.95$. $\chi^2 = 4.571$.

Ejemplo 6.13

Encontrar los valores críticos χ^2 a la derecha e izquierda de la distribución cuando $n = 13$ y $\alpha = 0.05$.

Solución

Los grados de libertad de la distribución son $gl = n-1 = 13-1 = 12$. La gráfica de la figura 6.31 describe la distribución de χ^2 con 12 grados de libertad y el área sombreada (probabilidad) de $\frac{1}{2}\alpha = 0.025$ en cada cola de la distribución. El área central es $1 - \alpha = 0.95$. Así, $\chi^2_I = 4.403$ $\chi^2_{cd} = 23.336$.

$$P(4.403 \leq \chi^2 \leq 23.336) = 0.95 \; gl = 12$$

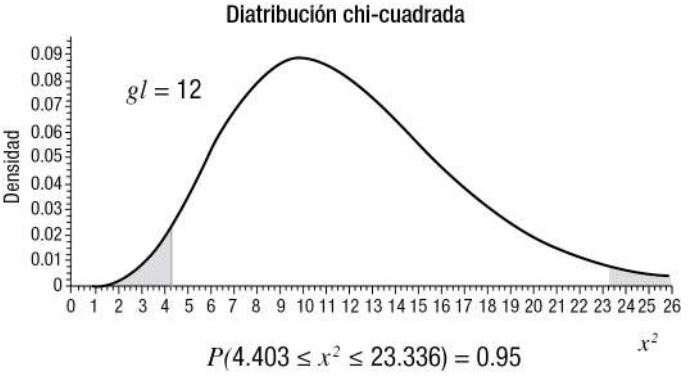


Figura 6.31 Cálculo de probabilidades para la distribución χ^2, entre dos valores de la variable aleatoria.

El intervalo de confianza de $(1-\alpha)\,\%$ **para** σ^2

Una aplicación de esta distribución es establecer un intervalo de confianza del $(1-\alpha)\,\%$ para σ^2:

$$\left(\frac{(n-1)S^2}{\chi^2(1-\alpha/2,n-1)},\frac{(n-1)S^2}{\chi^2(\alpha/2,n-1}\right)$$

y un intervalo de confianza del $(1-\alpha)\,\%$ para σ es:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2(1-\alpha/2,n-1)}},\sqrt{\frac{(n-1)S^2}{\chi^2(\alpha/2,n-1}}\right)$$

William Sealy Gosset (1876-1937)

Estadístico británico. Empleado por la firma cervecera Guinnes en Dublín, en 1906 fue enviado por la empresa a trabajar con K. Pearson en el University College de Londres, donde llevó a cabo sus principales contribuciones a la estadística, publicadas bajo el pseudónimo de Student. Estudió el problema de la estimación para muestras pequeñas, analizando la distribución del estadístico luego llamado t de Student.

6.5 La distribución t

En la vida real, para realizar estudios usando la distribución normal se requieren muestras suficientemente grandes ($n \geq 30$). Esta situación no es práctica, una alternativa para hacer inferencia sobre la media μ es usar la distribución t. Si Z y χ^2_{n-1} son variables aleatorias independientes, donde Z tiene una distribución normal estándar y χ^2 sigue una distribución Ji cuadrada con $n-1$ grados de libertad. Entonces se dice que la variable aleatoria está definida por:

$$T_{n-1}=\frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

Tiene una distribución $t-Student$ con $n-1$ grados de libertad, esta representación se puede obtener del grupo de distribuciones presentadas por **CalEst**. Nota cultural: La media y la varianza para ésta distribución son respectivamente:

Media	$\mu(T_{n-1}) = 0$
Varianza	$Var(T_{n-1}) = \frac{n-1}{n-2}$

Ejemplo 6.14

La forma de la distribución con 6 grados de libertad se muestra en la figura 6.32, en esta se muestra el cálculo de un valor crítico a la izquierda: esto es $P(T_{n-1} \leq -2.467) = 0.024$. Se muestra la probabilidad complementaria $P(T_{n-1} > -2.467) = 1 - 0.0243 \simeq 0.976$. Nota: se usa el símbolo $\simeq$ por errores de redondeo al considerar el umbral en milésimas.

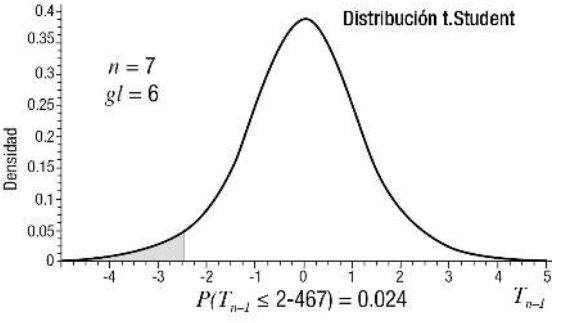


Figura 6.32 Descripción de la distribución $t - Student$ con 6 grados de libertad y probabilidad a la izquierda.

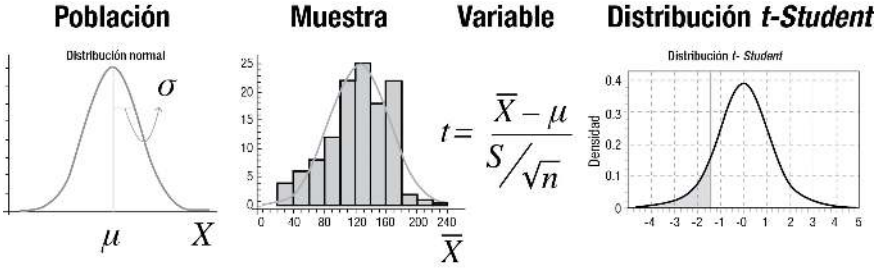


Figura 6.33 Descripción de la variable $t - Student$ en el contexto de población y muestra.

Distribución muestral

La distribución $t - Student$ desempeña un papel relevante en la inferencia estadística, de ahí la importancia de su estudio, y su aplicación se verá en los próximos capítulos. Así para el caso de muestras

pequeñas n, y dado que en muchas situaciones prácticas la desviación estándar σ de la población es desconocida, si la distribución de una variable aleatoria X es aproximadamente normal, entonces:

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

tiene una distribución $t - Student$. La figura 6.33 muestra la relación de la población en que la variable $t - Student$ permitirá concluir sobre la población usando la información de la muestra.

Los valores críticos de t se denotan por t_c, y se tiene $P(T_{n-1} > t_c) = \alpha$, donde α está entre 0 y 1. La distribución t tiene las siguientes propiedades:

1. La distribución t es de forma acampanada y simétrica alrededor de la media.
2. La distribución t es una familia de curvas, cada una es determinada por un parámetro llamado grados de libertad. Los grados de libertad son igual al tamaño de la muestra menos uno, $gl = n - 1$.
3. El área total bajo la curva es 1 o 100 %.
4. La media, la mediana y la moda de la distribución son igual a cero.

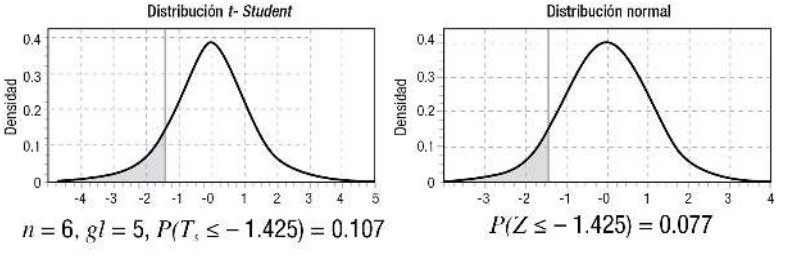


Figura 6.34 Relación entre las distribuciones $t - Student$ y la normal estándar, $gl = 5$.

5. A medida que el número de grados de libertad crece, la distribución se aproxima a una normal. Aunque es un resultado asintótico, después de los 29 gl, la distribución $t - Student$ se aproxima en centésimas a la normal estándar; la figura 6.34 describe el caso para $gl = 5$ y la figura 6.35 la situación $gl = 29$.

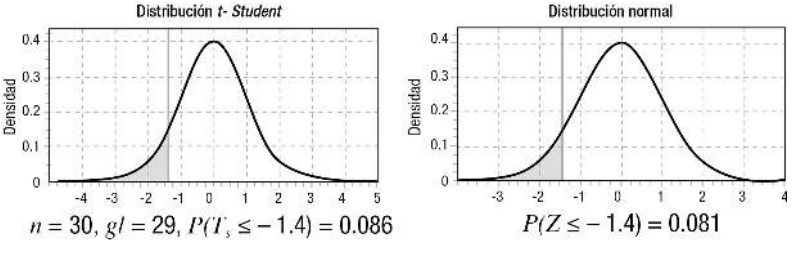


Figura 6.35 Relación entre las distribuciones $t - Student$ y la normal estándar, $gl = 29$.

Se observa que debido a la simetría:

$$\alpha = P(-T_{n-1} \geq t_c) = P(T_{n-1} \leq t_c) = 1 - P(T_{n-1} \geq -t_c)$$

Por lo que:

$$P(T_{n-1} \geq -t_c) = 1 - \alpha$$

Se llega a la conclusión de que:

$$-t_c(\alpha - 1, n - 1) = t(\alpha - 1, n - 1)$$

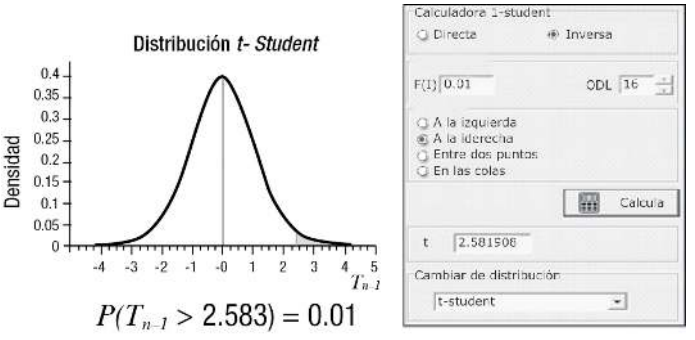


Figura 6.36 La distribución de probabilidad T para $\alpha = 0.01$ y $gl = 16$.

Ejemplo 6.15

Encontrar el valor crítico t_c en la cola derecha de la distribución $t - Student$ con $\alpha = 0.01$ y $n = 17$.

Solución

Los grados de libertad son $gl = n - 1 = 17 - 1 = 16$. Desde luego que existe una tabla que muestra algunos de los resultados principales de la distribución de probabilidad T; esta tabla se anexa al final del libro. En la figura 6.36 se muestra el bloque de distribución de **CalEst**: la distribución t de *Student*. El área sombreada corresponde al valor de $\alpha = 0.01$ y $t_0 = 2.583$ es el valor crítico. En símbolos:

$$P(T_{n-1} > 2.583) = 0.01$$

En la tabla de la derecha, aparece la alternativa para encontrar los valores t de la distribución para un valor de α dado. En este caso hay que usar la inversa. También se pueden encontrar probabilidades para diferentes valores de T.

Ejemplo 6.16

Usar la tabla de probabilidades de la distribución t con 13 gl para encontrar la probabilidad:

a. $P(T \leq -0.45)$, *b.* $P(T \geq 2.56)$, *c.*$P(-1.9 \leq T \leq 1.9)$.

Solución

a. $P(T \leq -0.45) = 0.330$, *b.* $P(T \geq 2.56) = 0.0118$, *c.*$P(-1.9 \leq T \leq 1.9) = 0.920$. Opcional: Mediante una gráfica en el CalEst visualice estos resultados.

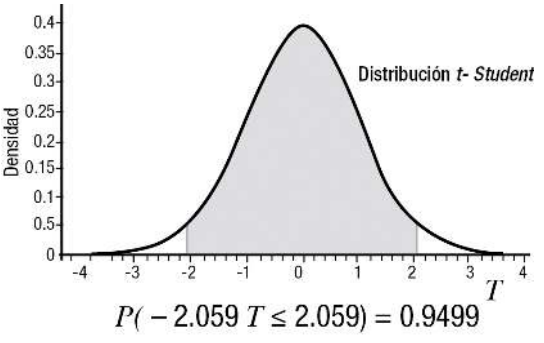


Figura 6.37 Cálculo de la probabilidad para distribución T considerando dos valores.

Ejemplo 6.17

Encontrar los valores críticos t_{ci} a la izquierda de la distribución, y t_{cd} a la derecha de la distribución para conocidas o referidas como colas derecha e izquierda, respectivamente, de la distribución $t - Student$, con $\alpha = 0.05$ y $n = 26$, notación:

$$t_{ci} = t(gl, \alpha/2) = t(25, 0.025) = -2.059$$
$$t_{cd} = t(gl, 1 - \alpha/2) = t(25, 0.975) = 2.059$$

.

Solución

Los grados de libertad son $n - 1 = 26 - 1 = 25$. En este caso se toma $\frac{1}{2}\alpha = \alpha/2$, para indicar la probabilidad de la cola izquierda y $1 - \alpha/2$ para referirse a la cola derecha de la distribución, es decir:

$\alpha/2 = 0.025$ y $1 - \alpha/2 = 0.975$. Los valores de $t_{ci} = -2.059$ y $t_{cd} = 2.059$. Ver figura 6.37.
$P(\ -2.059 \leq T \leq 2.059) = 0.9499$.

6.6 La distribución F

Si $\chi^2_{(n)}$ y $\chi^2_{(m)}$ son variables aleatorias Ji cuadradas con n y m grados de libertad, respectivamente, entonces se dice que la variable aleatoria $F(n, m)$ se define por:

$$F(n, m) = \frac{\chi^2_{(n)}/n}{\chi^2_{(m)}/m},$$

como una distribución F con n y m grados de libertad. La notación grados libertad en el numerador $gl_N = n$, y en el denominador $gl_D = m$.

Considerando el modelo de la obtención de una muestra a partir de la población, en esta situación se plantea la estrategia de comparar dos poblaciones. La variable aleatoria F permite inferir si las poblaciones son similares para alguna característica de estudio. En la figura 6.38 se presenta el proceso de obtención de la muestra de cada población; en la figura se observa que $n_1 = n$ y $n_2 = m$.

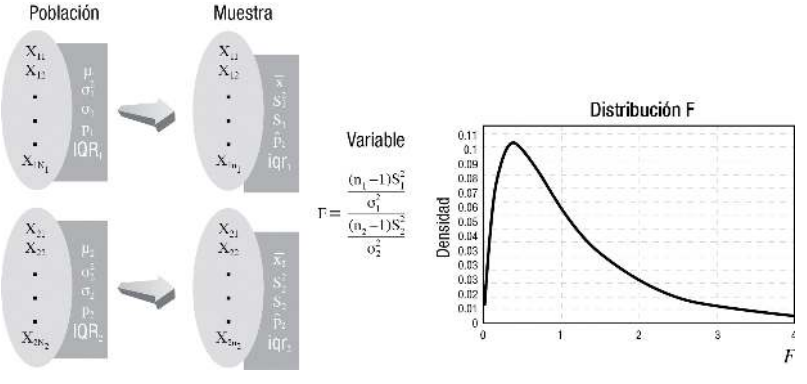


Figura 6.38 Construcción de la variable aleatoria F a partir de la relación Población y Muestra.

Ejemplo 6.18

En la figura 6.39 se presenta la distribución F con 4 y 7 grados de libertad, con un valor de $\alpha = 0.05$, es decir: $F(4, 7, 0.05) = 4.121$. También se ejemplifica el uso de la tabla que funciona como calculadora

de esta distribución. Es conveniente realizar varios ejercicios con base en esta distribución para obtener un mayor dominio sobre el cálculo de los valores de F dada una probabilidad, o calcular probabilidades a partir de un valor de F.

Al valor de α se le conoce como *nivel de significancia* y es la probabilidad de que $F(n, m, \alpha)$ sea mayor que F_{cd}; a este valor de F_{cd} se le conoce como *punto crítico* a la derecha de la distribución F. Es decir:

$$P(F(n, m) > F_D(4, 7, \alpha)) = \alpha$$

El valor de α está entre 0 y 1 $(0 < \alpha < 1)$, la distribución F cumple con la propiedad:

$$F(n, m, 1 - \alpha) = \frac{1}{F(m, n, \alpha)}$$

Otra propiedad de la distribución F en su relación con la distribución *t-Student,* es : $F(1, m, \alpha) = t^2(m, \alpha/2)$.

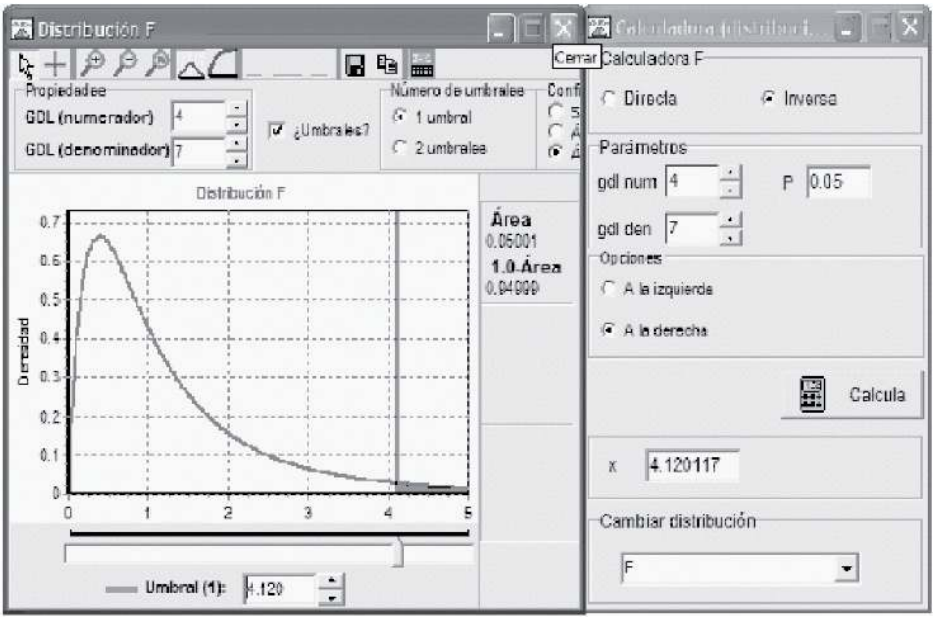


Figura 6.39 Probabilidad a la derecha de 4.12 en una distribución F.

Guía para encontrar los valores críticos para la distribución F

1. Especificar el nivel de significancia α.

2. Determinar los grados de libertad para el numerador $gl_{N.}$

3. Determinar los grados de libertad para el denominador gl_D.

4. Use la distribución F en **CalEst**.

   a) Valor de α cola derecha, describe la probabilidad a la derecha del punto $F_{cd}(n, m, \alpha)$, es decir: $P(F(n, m) > F_{cd}(n, m, \alpha)) = \alpha$

   b) Valor de α cola izquierda, se obtiene la probabilidad a la izquierda del punto $F_{ci}(n, m)$:

$$F_{ci}(n, m, 1 - \alpha) = \frac{1}{F_{cd}(m, n, \alpha)}$$

Ejemplo 6.19

Ilustración de propiedad $F_I(n, m, 1 - \alpha) = \frac{1}{F_D(m,n,\alpha)}$.

Caso 1. $gl_N = 3$ y $gl_D = 6$ y $\alpha = 0.05$ $F_D(3, 6, 0.0) = 4.791$; de manera inversa, dado el valor de F, se tiene el valor de la probabilidad $P(F > 4.757) = 0.05$.

Cálculos:

$$F_I(6, 3, 1 - 0.05) = \frac{1}{F(3, 6, 0.05)} \cong \frac{1}{4.791} \cong 0.209$$

Nota: Dada esta propiedad en los libros de estadística, generalmente aparecen los valores de la distribución F a la derecha.

Una ventaja al utilizar el *CalEst* es que usando los cursores puede obtener rápidamente los valores críticos de F a la izquierda de α y a la derecha de α.

Caso 2. $gl_N = 6$ $gl_D = 3$ y $\alpha = 0.05$, $F_D(6, 3, 05) = 8.940$

$$F_I(3, 6, 1 - 0.05) = \frac{1}{F(6, 3, 0.05)} = \frac{1}{8.940} \cong 0.113$$

En la figura 6.40 se presenta el caso bilateral de la distribución F.

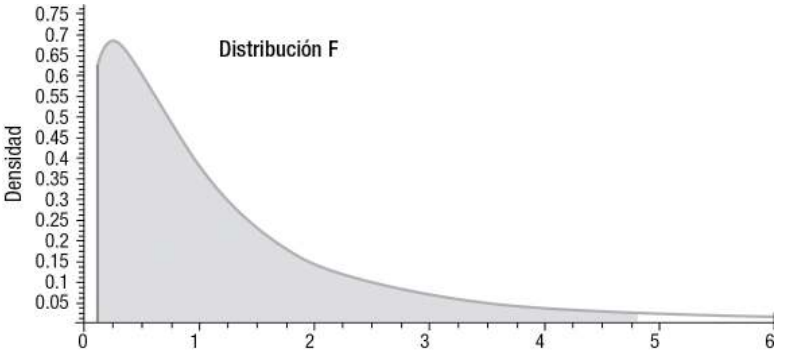


Figura 6.40 Valores de F para una probabilidad $1 - \alpha/2$ y $\alpha/2$.

6.7 Resumen

Distribución de probabilidad normal	Curva simétrica en forma de campana, con media μ y desviación estándar σ. De manera usual se le llama curva normal y se denota por N(μ, σ).
Distribución de probabilidad normal estándar (variable aleatoria Z)	Presenta una densidad, cuya forma es una campana con: Media $\mu = 0$ Desviación estándar $\mu = 1$, y se denota por N(0; 1).
Tabla normal estándar	Tabla de probabilidades para una variable aleatoria Z.

Procedimiento de estandarización de la distribución de probabilidad normal

Al valor de la variable aleatoria se le resta la media μ y se divide entre la desviación estándar σ. La representación de esta operación se escribe por la expresión:

$$Z = \frac{X - \mu}{\sigma}$$

Probabilidades y número de desviaciones estándar

Aunque la curva de la normal es continua de manera infinita en ambas direcciones, la mayor parte de la distribución está dentro de tres desviaciones estándar a cada lado de μ. Para una distribución normal se tiene que:

El 68.26 % de la distribución está entre $\mu - \sigma$ y $\mu + \sigma$, dicho de otra manera, la probabilidad del intervalo cuya longitud es de una desviación estándar a cada lado de la media es:

$$P(\mu - \sigma < X < \mu + \sigma) = 0.683$$

El 95.44 % de la distribución está entre $\mu - 2\sigma$ y $\mu + 2\sigma$, de manera análoga, la probabilidad del intervalo cuya longitud es de dos desviaciones estándar a cada lado de la media es:

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.954$$

El 99.74 % de la distribución está entre $\mu - 3\sigma$ y.$\mu + 3\sigma$. Finalmente, la probabilidad del intervalo cuya longitud es de tres desviaciones estándar a cada lado de la media es:

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

Percentiles en una distribución normal

Para encontrar el p-enésimo percentil x_p en una distribución normal con media μ y desviación estándar σ mediante la normal estándar. Primero se busca el percentil de una normal estándar z_p y luego se aplica la siguiente fórmula:

$$x_p = \mu + \sigma z_p,$$

para encontrar el percentil para x_p. Esta expresión convierte el percentil de la normal estándar al percentil de una distribución normal con parámetros μ y σ.

Cálculo de probabilidades en una distribución normal

$$\begin{array}{lll} P(X < 65.2) = \alpha/2 = 0.025 & P(65.2 \leq X \leq 84.8) = 1 - \alpha = 0.95 & P(X > 84.2) = \alpha/2 = 0.025 \\ P(Z < -1.96) = \alpha/2 = 0.025 & P(-1.96 \leq Z \leq 1.96) = 1 - \alpha = 0.95 & P(Z > 1.96) = \alpha/2 = 0.025 \end{array}$$

6.8 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

6.9 Ejercicios

Distribución normal

6.1 El tiempo de respuesta para un estímulo tiene una distribución normal con media $\mu = 46$, $\sigma = 4$. ¿Qué porcentaje de valores:

1. i.¿Son mayores de 46? ii. ¿Son mayores de 50?
2. i.¿Son mayores de 40? ii. ¿Son menores que 38?
3. i.¿Son menores que 49? ii. ¿Están entre 45 y 49?
4. i.¿Están entre 50 y 54? ii. ¿Es mayor que 56 y menor que 46?
5. i.¿Están dentro 1.5 desviaciones estándar de la media?
6. i.¿Está fuera de 2.3 desviaciones estándar de la media?

6.2 En un proceso de producción un producto tiene un encogimiento con una media de 80 mm y una desviación estándar de 12. Si un producto se selecciona al azar, encuentre cada una de las siguientes probabilidades:

1. i. $P(80 \leq X \leq 92)$, ii. $P(71 \leq X \leq 80)$, iii. $P(X \leq 92)$, iv. $P(X \geq 56)$, iv. $1 - P(65 \leq X \leq 98)$,
2. i. $1 - P(66 \leq X)$, ii. $1 - P(X \geq 104)$, iii. $P(53 \leq X \leq 59)$, iv. $P(X = 80)$.

6.3 Para un grupo de estudiantes de secundaria y preparatoria, el tiempo de solución del rompecabezas de la República Mexicana tiene una distribución normal con una media de 150 segundos y una desviación estándar de 10. Encuentre:

1. El primer, segundo y tercer cuartil.
2. El 65 percentil y el 35 percentil.
3. El rango del percentil del valor 165, es decir $P(X \leq 165)$.
4. El rango del percentil del valor 145.

6.4 Supongamos que X es una variable aleatoria con una distribución normal con media $\mu = 54$ y desviación estándar $\sigma = 6.8$. Calcule:

1. $P(40 \leq X \leq 56)$.

2. $P(X \geq 64)$.

3. w tal que $P(X \geq w) = 0.32$.

4. w tal que $P(X < w) = 0.063$.

6.5 El tiempo en que el cajero bancario tarda en atender a los clientes tiene una distribución normal con media $\mu = 10$ minutos y desviación estándar $\sigma = 2$ minutos.

1. ¿Qué porcentaje de clientes espera menos de 8 minutos?

2. ¿Cuál es la probabilidad de que el cajero tarde en atender a un cliente un tiempo mayor que 16 minutos?

6.6 En estudios realizados por una firma para medir el coeficiente intelectual de las personas que solicitan un empleo hay una distribución normal con media $\mu = 95$ y desviación estándar $\sigma = 13.5$.

1. ¿Cuál es la probabilidad de un coeficiente intelectual mayor a dos desviaciones estándar de la media?

2. ¿Qué porcentaje de las personas presenta un coeficiente intelectual superior a 120?

3. ¿Qué porcentaje de las personas muestra un coeficiente intelectual entre 80 y 120?

6.7 El número de calorías de una sopa en un menú se distribuye como una normal con media 200 y desviación estándar de 5. Encuentre la probabilidad de que la sopa contenga:

1. Más de 210 calorías.

2. Entre 190 y 200 calorías.

3. Encuentre el número de calorías que corresponde el 14 percentil.

6.8 En un restaurante sirven manzanas como postre. El peso de ellas muestra una distribución normal con media 60 gramos y una desviación estándar de 0.8 gramos. ¿Cuál es la probabilidad de que la manzana que se le sirve a la siguiente persona pese 50 gramos?

6.9 Las calificaciones de admisión presentan una distribución normal con una media de 500 y una desviación estándar de 100. Encuentre la probabilidad de que un estudiante:

1. Obtenga una calificación mayor que 650.

2. Obtenga una calificación menor que 375.

3. Obtenga una calificación entre 350 y 575.

4. Si la escuela admite sólo a los que tienen una calificación mayor a 670, ¿cuál es la proporción de los estudiantes que pueden ser admitidos?

5. ¿En qué límite se debe fijar la calificación si el 50 % de los estudiantes debe ser admitido?

6. ¿Cuál debe ser la calificación límite si el tope de admisión es del 15 % de estudiantes?

6.10 Supongamos que Z es una distribución normal estándar. Encuentre el percentil de la distribución en cada caso.

1. Debajo de $z = 2.0$.

2. Debajo de $z = 2.6$.

3. Debajo de $z = 1.36$.

4. Debajo de $z = -1.36$.

5. Entre $z = -1.42$ y $z = 1.25$.

6. Entre $z = -2.82$ y $z = -0.58$.

6.11 Supongamos que Z es una distribución normal estándar. Encuentre:

1. $P(Z < 1.64)$.

2. $P(Z \geq 1.96)$.

3. $P(-1.35 \leq Z \leq 1.35)$.

4. $P(1.22 \leq Z \leq 2.47)$.

5. El valor de z tal que 5 % del área esté debajo de éste.

6.12 Encuentre los percentiles 0.05, 0.01, 0.10, 0.025, 0.90, 0.95 y 0.68 en una distribución normal estándar.

6.13 El tiempo que dura un embarazo tiene una distribución normal con media $\mu = 266$ días y una desviación estándar de 16 días.

1. Usando esta gráfica encuentre la proporción de embarazos entre 285 y 305 días.

2. Una compañía de seguros tiene un plan familiar de salud. En una de sus cláusulas indica que no se cubren los costos de hospitalización si el nacimiento es en menos de 217 días después de la fecha de casamiento. Encuentre la probabilidad de que el nacimiento se dé en menos de 217 días.

3. Encontrar el 90 percentil.

4. Encontrar el 30 percentil.
5. Encontrar el rango percentil del valor de 275 días.
6. ¿Cuál es la probabilidad del que el embarazo termine en 280 días o más?
7. ¿Cuáles son los valores Z de la normal estándar que corresponden a los valores de X en el inciso 1?

6.14 Describa las características de una curva normal.

6.15 El sobrepeso es un factor de riesgo para la salud. El índice de masa corporal (IMC) se define como el peso entre la estatura al cuadrado. Se supone que este índice, en una población de adultos, se distribuye como una distribución de probabilidad normal con $\mu = 27.3$ y $\sigma = 4.1$.

1. Encuentre i) $P(X > 25)$ ii) $P(25 \leq X \leq 30)$ iii) $P(X > 30)$.

6.16 El periodo de gestación para un bebé con pocos días de nacido tiene una distribución normal con una $\mu = 266$ y $\sigma = 16$. Encuentra el valor de la variable Z (Z es la variable aleatoria que caracteriza a la normal estándar) para los siguientes valores: a.- 280 días b.- 250 días c.- 270 días.

6.17 Se calificó un examen, en el que la media fue de 500 y la desviación estándar de 100. Se sabe que 10,000 estudiantes realizaron el examen y que sus calificaciones tenían una distribución simétrica que se puede aproximar por una curva de densidad normal.

1. ¿Cuántos estudiantes obtuvieron una calificación entre 400 y 600?
2. ¿Cuántos estudiantes obtuvieron una calificación entre 300 y 700?
3. ¿Cuántos estudiantes obtuvieron una calificación entre 200 y 800?

6.18 X: Tiempo en que Pedro tarda en correr 5 kilómetros, medida en minutos. La media del grupo de corredores de la categoría de Pedro es 30 minutos, con una desviación estándar de 5 minutos. ¿Cuál es la probabilidad de que un corredor tarde menos de 19 minutos? ¿Cuál es la probabilidad de que un corredor tarde más de 38 minutos? Si compiten 1350 corredores ¿cuántos corren entre 24 y 36 minutos?

6.19 Las edades de defunción de hombres en México después de los 50 años sigue una distribución normal con media 70 años y una desviación estándar de 5.

1. ¿Cuál es la probabilidad de que una persona muera entre 60 y 65 años?
2. Encuentre la probabilidad de que una persona muera después de los 80 años.
3. Calcule la probabilidad de que una persona muera antes de los 60 años.

6.20 El número de documentos que revisan los administradores de una secretaria por hora tiene una distribución normal con media 9 y desviación estándar de 1.8. Encuentre las siguientes probabilidades: i.-1 $- P(6 \leq X)$, ii. $1 - P(X \geq 10)$, iii. $P(5 \leq X \leq 9)$, iv. $P(X > 8)$.

6.21 ¿Será cierto que las administraciones del sector público son eficientes? El tiempo de atención a clientes en oficinas de un gobierno local sigue una distribución normal, con media 17 minutos con una desviación estándar de 4.5 minutos. Qué porción de tiempo de espera:

1. Es mayor a 24 minutos.
2. Es cuando mucho 12 minutos.
3. Está entre 8 minutos y 24 minutos.

6.22 La administración de una empresa financiera ha llevado el registro del número de solicitudes de préstamo que reciben a la semana, ésta sigue una distribución normal con media 56.6 y una desviación estándar de 9.1 solicitudes. Establezca la probabilidad de que en una semana la empresa reciba:

1. Entre 58 y 67 solicitudes.
2. Más de 65 solicitudes.
3. Más de 60 solicitudes.

Distribución $t - Student$

6.23 Encontrar los valores críticos para la cola izquierda en cada inciso.

1. $t(0.0\ 5, 7)$, $t(0.0\ 1, 7)$, $t(0.005, 7)$, $t(0.10, 7)$.
2. $t(0.05, 12)$, $t(0.01, 12)$, $t(0.005, 12)$, $t(0.10, 12)$.
3. $t(0.05, 25)$, $t(0.01, 25)$, $t(0.005, 25)$, $t(0.10, 25)$.

6.24 Encontrar las siguientes probabilidades:

1. $P(T \geq 2.7)$, $P(T \leq -1.56)$, $P(-2 \leq T \leq 2)$ con $n = 18$.
2. $P(T \geq 3.5)$, $P(T \leq -0.58)$, $P(-2.5 \leq T \leq 2.5)$ con $n = 35$.
3. $P(T \geq 1.35)$, $P(T \leq -3.5)$, $P(-3 \leq T \leq 3)$ con $n = 6$.

Distribución χ^2

6.25 En cada uno de los siguientes casos $\chi^2(0.01, n-1)$, $\chi^2(0.025, n-1)$, $\chi^2(0.95, n-1)$ y $\chi^2(0.99, n-1)$. Encontrar estos valores de χ^2 con los siguientes grados de libertad (gl): *a.-* $gl = 9$, *b.-* $gl = 15$, *c.-* $gl = 25$.

6.26 Si los grados de libertad de la distribución son $gl = 2$, encuentre los valores de x en los siguientes casos: *a.* $P(\chi^2 \geq x) = 0.01$, *b.* $P(\chi^2 \geq x) = 0.05$, *c.* $P(\chi^2 \geq x) = 0.99$, *d.* $P(\chi^2 \geq x) = 0.01$, *e.* $P(\chi^2 \geq x) = 0.9$, *f.* $P(\chi^2 \geq x) = 0.5$.

6.27 Considerando $gl = 18$, calcule las siguientes probabilidades: *a.* $P(\chi^2 > 25.989)$ *b.* $P(\chi^2 \leq 7.015)$ *c.* $P(7.015 < \chi^2 < 9.390)$ *d.* $P(10.865 < \chi^2 < 28.869)$.

6.28 Encuentre el valor crítico de χ^2 para la cola izquierda cuando $n = 18$ y $\alpha = 0.01$.

6.29 Encuentre el valor crítico de χ^2 para la cola derecha cuando $n = 30$ y $\alpha = 0.05$.

6.30 Encuentre el valor crítico de χ^2 para dos colas cuando $n = 19$ y $\alpha = 0.05$. Es decir, encuentre el valor χ^2_D con $\frac{1}{2}\alpha$ y el valor χ^2_I con $1 - \frac{\alpha}{2}$.

6.31 Si X tiene una distribución χ^2 con $gl = 10$, encuentre la probabilidad $P(3.25 \leq X \leq 20.5)$.

6.32 Si X tiene una distribución χ^2 con $gl = 5$, determine las constantes c y d tal que $P(c < X < d) = 0.95$ y $P(X < c) = 0.025$.

Distribución F

6.33 Calcular los valores de F en la distribución F para los siguientes casos:

$F(0.025,\ 7, 5)$, $F(0.1,\ 3, 8)$, $F(0.05,\ 5, 7)$, $F(0.9,\ 5, 7)$, $F(0.01,\ 10, 12)$. Use las gráficas y tablas de la distribución F del *CalEst*.

6.34 Calcular el valor de F en la distribución F con $gl_N = gl_D = 24$ y $\alpha = 0.001$. Dadas estas condiciones, encontrar la probabilidad a la derecha si $F = 3.83$.

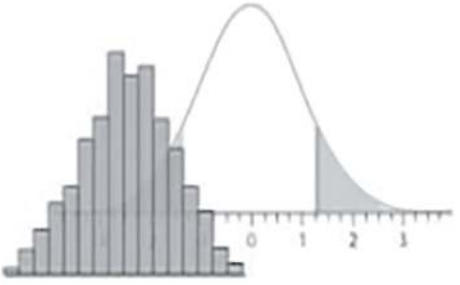
6.35 Encontrar el valor crítico derecho de F_D cuando los valores de α son: $\alpha = 0.05$, $\alpha = 0.025$, $\alpha = 0.01$ y $\alpha = 0.005$ respectivamente, donde los grados de libertad son: $gl_N = 6$ y $gl_D = 29$. Para estas condiciones encontrar el valor crítico a la siguiente F_I.

6.36 Encontrar las probabilidades a la derecha de $F = 5.40$ para los tres siguientes pares de grados de libertad: i. $gl_N = 4$ y $gl_D = 40$, ii. $gl_N = 6$ y $gl_D = 29$, iii. $gl_N = 10$ y $gl_D = 12$.

6.37 Si $F_0 = 4.83$ con $gl_N = 4$ y $gl_D = 8$, encontrar las probabilidades: $P(F(4,8) > 4.83)$, $P(\frac{1}{F(8,4)} < 0.97)$.

6.10 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.



Capítulo 7

Estimación por intervalos de confianza

Dime y lo olvido, enséñame y lo recuerdo, involucrame y lo aprendo.

Benjamín Franklin

Competencia general

Comprender la importancia de la inferencia estadística y el procedimiento de estimación por intervalo de confianza. Para diferentes parámetros de una población y relacionarlos en el contexto de problemas reales.

Competencias específicas

- Mediante un ejemplo conocer la relación entre parámetros y estadísticos.
- Observar la relación que existe entre la variable en una población y la variable en una muestra.
- Comprender las caracterícas de un parámetro y su estimación. Así como estudiar el procedimiento de estimación puntual y por intervalos de confianza del parámetro.
- Ilustrar el proceso de estimación estadística mediante ejemplos y definir los términos estadísticos que comprenden la estimación.
- Aprender a interpretar los intervalos de confianza para los parámetros de la media, proporción y varianza.
- Estudiar las distribuciones de probabilidad apropiadas para estimar los intervalos de confianza y calcular las probabilidades que se requieren en el proceso de estimación.
- Explicar los términos que comprenden los intervalos de confianza y observar cómo cambian la longitud de los intervalos en función de los niveles de confianza, varianza y tamaño de la muestra.
- Aplicar la fórmula del error de muestreo para calcular el tamaño de muestra.

7.1 Introducción

Existe una variedad de situaciones en la práctica que son de mucho interés para los estudiosos en problemas de administración y economía. Una serie de breves escenarios en los que se requiere información para su análisis e interpretación se indican a continuación. En algunos centros de estudio se tiene interés por conocer cuál es el promedio del salario básico de los trabajadores que cotizan en la seguridad social, considere como referencia un periodo, por ejemplo el último trimestre o un año. De este entorno, se desprenden varias reflexiones, conocer si el salario real se mantiene ante las situaciones de inflación. Observar el porcentaje que el salario básico acumula en un año, ya sea como pérdida o ganancia, si se compara con periodos pasados. Estimar el porcentaje de trabajadores ocupados que ganan entre uno y tres salarios mínimos diarios. Evaluar el total de personas que no reciben salario a pesar de estar consideradas como personas con ocupación. Así como, saber la variación de los ingresos mensuales que obtiene una familia de trabajadores. Estudiar el porcentaje real de pérdida del poder adquisitivo de los trabajadores formales e informales en los últimos cinco años a partir de un periodo de referencia.

En este capítulo se verá que la inferencia estadística es el proceso que permite obtener conclusiones sobre los parámetros con base en las propiedades de una muestra seleccionada de una población. En una etapa inicial, se describirá el estudio de la estimación puntual, y con base en los resultados metodológicos de esa explicación, se presentarán los conceptos para realizar la estimación puntual y por intervalo. Primero se mostrará las ideas del procedimiento de estimación puntual y enseguida el método de estimación referido como intervalos de confianza de un parámetro. En función del parámetro que se tenga interés en estimar, se habla de intervalos de confianza para una media, una proporción o una varianza. Una característica relevante de este proceso de estimación es que da un rango de valores y una probabilidad. En particular, el valor de la probabilidad indica un grado de certidumbre de que un intervalo contenga al parámetro. Nota: más adelante en nuevos capítulos, se verán distintas situaciones que requerirán el estudio de diferentes parámetros a los tratados en esta parte.

Motivación de la estimación estadística

Escenario 1

Un problema actual que de muchas maneras tiene un efecto económico importante es la obesidad. Esta es una de las enfermedades que afectan a los seres humanos debido, entre otras causas, a una dieta alimentaria inadecuada. En esta situación, las personas forman un conjunto grande de individuos al que se le denomina *población*. Una variable que ayuda a identificar el estado físico de una persona es el índice de masa corporal: IMC. El valor de esta variable se puede obtener conociendo el peso y la estatura del individuo:

$$IMC = peso/estatura^2$$

Para tomar decisiones adecuadas en la dieta alimenticia de la población, es necesario conocer si un

programa de alimentación saludable ayuda en *promedio* a reducir el IMC. Es evidente que estudiar a cada individuo de la *población* resulta costoso y complicado, por esto para tener una idea sobre el efecto del programa en el problema de la obesidad se recurre a examinar sólo una parte de la *población* mediante una *muestra* de personas obesas, con la idea de que el *promedio* de la reducción del IMC en esta *muestra* sea una buena estimación del *promedio* de reducción en toda la *población*. De esta manera se podría decir que el tratamiento para adelgazar fue en promedio efectivo.

Escenario 2

La contaminación, *población*, tiene efectos sobre la economía ya que genera costos para mantenerla a niveles bajos; también se destina dinero para curar enfermedades ocasionadas por partículas contaminantes. Los datos aleatorios en estos casos se obtienen a través de estaciones de monitoreo que permiten tomar *muestras* del aire o de agua para evaluar el nivel de contaminación.

Escenario 3

La vida de anaquel de una empresa que produce alimentos empacados es bajo, ya que tiene un lapso de empleo de 35 días. Esta situación genera una merma financiera por la cantidad de material elaborado que se desperdicia. Una estrategia para remediar esta situación es realizar un experimento agregando diferentes cantidades de un aditivo químico para alargar la duración de ese producto. Si el resultado de este proceso indica que la duración del alimento es ahora en promedio de 42 días, ¿se podrá afirmar que este aditivo mejora la vida de anaquel del producto? La meta de realizar un experimento de este tipo es determinar la influencia que uno o varios factores tienen sobre una característica de la población a la cual se denomina respuesta. De esta forma se obtiene conocimiento acerca del fenómeno o proceso en estudio. Nótese que los datos aleatorios proporcionados por las unidades de la población son producto de los resultados del experimento. En este caso la población es más conceptual que real.

Escenario 4

Varios fenómenos o estudios de la naturaleza (económicos, sociales, industriales) se pueden representar mediante un modelo. Entonces las observaciones de una muestra provienen de una población hipotética de valores simulados.

El procedimiento que permite, a través de los datos o de la información en una *muestra*, tener un conocimiento de la *población* se conoce como inferencia estadística. Este conocimiento se puede obtener mediante un intervalo de valores o considerando un solo valor. La inferencia estadística es una herramienta muy útil para resolver una gran cantidad de cuestiones que se presentan en la vida cotidiana, o en problemas sociales y biotecnología, en estudios económicos, en el desarrollo tecnológico y en la investigación científica en general.

En la figura 7.1 se ejemplifica una población de personas y en el círculo se describe la *muestra*. La esencia de la inferencia estadística es adquirir conocimiento sobre una población a través de la información proporcionada por la *muestra*, y así extraer conclusiones generales sobre la *población* objeto de estudio.

En la figura 7.2 el círculo grande representa la población objeto de estudio y el pequeño se refiere a la muestra que se toma de la población. En realidad, el círculo de la muestra debe estar dentro de la

población como se indica en la figura 7.1, pero mediante esta gráfica puede visualizarse la relación entre probabilidad e inferencia estadística. Por lo regular se dice que la muestra tiene información relevante de la población. Una muestra aleatoria es el proceso de seleccionar n unidades de una población que contiene N unidades, de manera que cada una de las unidades tiene la misma probabilidad $\frac{1}{N}$ de ocurrir. Cuando una muestra se selecciona de esta manera, se le denomina *muestreo aleatorio simple*.

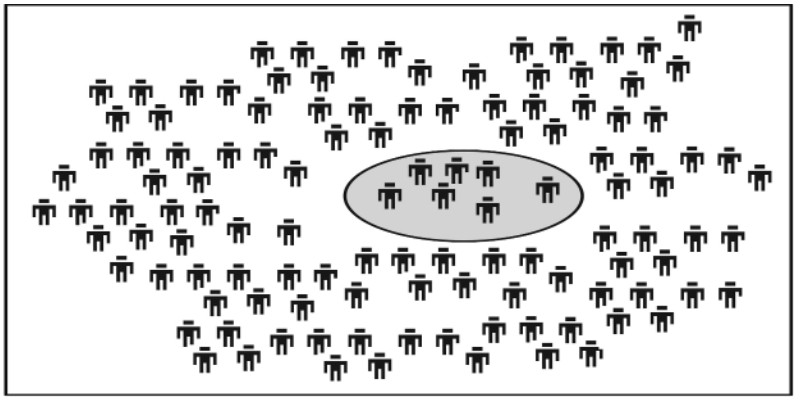


Figura 7.1 Idea de la relación entre *población* y *muestra*.

Tipos de muestreo

La probabilidad se aplica para seleccionar una muestra, se requiere que la población sea lo más homogénea posible. Ésta también desempeña un papel relevante en la inferencia estadística, ya que en términos probabilísticos se mide la estimación. Existen diferentes tipos de muestreo para elegir una muestra, los más utilizados son el aleatorio simple, el sistemático, por estrato y por conglomerados.

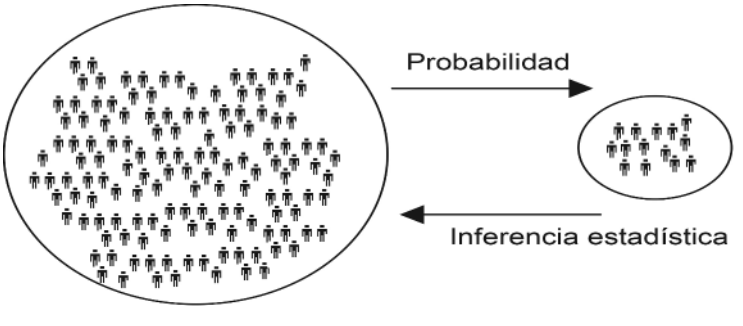


Figura 7.2 Descripción del procedimiento de inferencia estadística.

En resumen, necesitamos aprender a calcular la verosimilitud de una muestra particular seleccionada de una población. En ese sentido, lo que estamos haciendo es recorrer el camino de la población a la muestra, como la flecha de arriba en la figura 7.2. Sin embargo, nuestra meta final es ir de la muestra a la población, esto es, hacer lo que representa la flecha de abajo de la figura 7.2. Con ello, decimos que se usa la información de la muestra para elaborar afirmaciones en términos de probabilidad sobre el comportamiento de la población. Esta temática es un fuerte componente con temas relacionados con el conocimiento y la investigación. En este espacio se hace un breve resumen de elementos relacionados con esta área.

Poblaciones finitas

Las poblaciones que se pueden enumerar y así tener una lista de cada uno de sus elementos se denominan finitas. Si el interés de un investigador, una persona o una organización es tener la información de toda la población específica, entonces a ese proceso se le denomina censo.

7.2 Estimación estadística: puntual o por intervalo

La *estimación* es el proceso que permite inferir sobre los posibles valores de los parámetros que describen la población.

Proceso: Como es muy probable que se desconozcan los valores de los parámetros que representan a una población, se recurre a la información proporcionada en la muestra para tener una idea de los valores de los parámetros. Lo que puede resultar lógico es identificar un descriptor numérico para la muestra. Este estadístico, llamado *estimación puntual*, se puede usar para estimar la medida correspondiente a la población. Recordemos que un *estadístico* es la medida numérica que se calcula a partir de los datos observados en una muestra.

Una *estimación puntual* es un número calculado apartir de la muestra, y éste se usa para estimar un parámetro de la población. Un *estimador puntual* es una fórmula que se emplea para calcular la estimación puntual en un conjunto de datos.

Hay que tener presente que un *parámetro* es una medida numérica de la población. Los parámetros en realidad son desconocidos. Una meta en el procedimiento de estimación, es encontrar dos valores que comprendan una proporción fija de las medidas muestrales; éste procedimiento se llama *estimación por intervalo* .

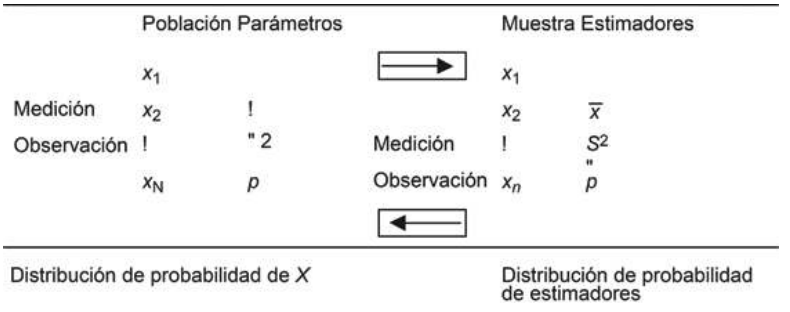


Figura 7.3 Caracterización estadística de la *población*: **parámetros**; y de la *muestra*: **estimadores**.

Las expresiones para la media y varianza poblacional, así como para la media y varianza muestral son:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}, \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}, \qquad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} \tag{7.1}$$

Estadístico

El estadístico es una medida que se obtiene a partir de una muestra, en este momento se están mencionando y trabajando con tres de ellos, a saber la media $\overline{X}$, la varianza S^2 y la proporción $\widehat{p}$.

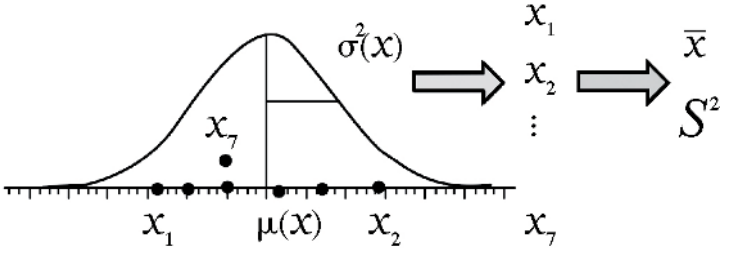


Figura 7.4 Distribución de la variable X característica de la población con parámetros $(\mu(X), \sigma^2)$, caso normal.

Para hacer inferencia estadística acerca de la media μ de la *población*, se debe comprender que la media muestral $\overline{X}$ es una variable aleatoria y tiene una distribución de probabilidad que permite evaluar el proceso de estimación. El enfoque de este capítulo es comprender cómo funciona la $\overline{X}$ para obtener

un conocimiento aproximado de la media poblacional. En relación con la figura 7.3 se han presentado tres parámetros de la *población* (μ, σ^2 y p) éstos son por lo general desconocidos. Los correspondientes estimadores, también conocidos como *estadísticos*, recogen la información de la *muestra* y son: ($\overline{X}, S^2$ y $\widehat{p}$). En capítulos posteriores se verán otros parámetros y sus correspondientes estimadores. Supóngase, por el momento, la variable X: estatura de las mujeres mayores de 18 años que pertenecen a un centro escolar y ésta tiene una distribución de probabilidad normal con media $\mu = 165$ cm y $\sigma = 5$ cm. En general X representará la variable aleatoria y ésta tiene un distribución de probabilidad, en particular en la figura 7.4 se ve que la distribución de la variable original tiene una distribución normal con media μ y varianza σ^2, es decir $X \sim N(\mu, \sigma^2)$. Cuando el interés consiste en estimar μ ¿cómo se sabrá si efectivamente la distribución de probabilidad de la variable X es normal o aproximadamente simétrica? Si no lo es, ¿afecta la estimación? ¿La estimación depende de σ^2? es decir, ¿cómo saber el valor de σ^2 si no conocemos el valor de μ? Con la finalidad de presentar el procedimiento de estimación, inicialmente se supone conocida o se aproxima por un tamaño suficientemente grande de la muestra.

Medición de una característica

Dada una característica de la población descrita por la variable aleatoria X, se escogen de manera aleatoria, digamos $n = 7$, elementos de la población y se mide esa característica de la muestra. Se calcula la media y la varianza.

El mundo de la información 1. Salario Profesional

Una cuestión de suma importancia para los profesionistas que acaban de egresar de la universidad y están en busca de empleo, es el salario.

Una universidad le encargó a una empresa que realiza estudios de mercado una encuesta para saber, entre otros factores, el salario percibido por las personas que terminaron hace tres años o menos la carrera de ingeniero industrial en diferentes universidades. En este caso la población de interés la forman todos los ingenieros industriales que egresaron de las universidades en México y tienen tres años o menos de haber egresado y están trabajando.

Preguntas sobre la naturaleza del problema

El salario establece el centro de las relaciones de intercambio entre las personas y las organizaciones. Todas las personas dentro de las organizaciones ofrecen su tiempo y su fuerza de trabajo a cambio de dinero. Esto representa el intercambio de una equivalencia entre derechos y obligaciones recíprocas entre el empleado y el empleador. Varias preguntas surgen para conocer los niveles de salarios: ¿qué tan competitivo es el salario de un ingeniero industrial recién egresado? ¿Cómo es este salario con respecto a otras profesiones? ¿Existen empresas que pagan mejor que otras?

Recopilación de datos: Con el propósito de plantear la idea establecida entre los datos de una población y el procedimiento para obtener una muestra y a partir de ella, realizar la inferencia estadística sobre la población, consideremos una empresa pequeña que constituye la población, en este caso de 50 ingenieros industriales y su salario dividido entre 60.57* para expresarlo en salarios mínimos. En la tabla 7.1 se identifica con ID a los 50 ingenieros para facilitar el procedimiento de extracción de la muestra y el número de salarios mínimos. (∗) Salario mínimo en la zona B en México, en el año 2012.
Recordemos que a partir de la información de la muestra pueden conocerse las características de la población. En este caso, se considera una población pequeña donde se puede calcular su media y su desviación estándar. La media y la desviación estándar para esta población pequeña $N = 50$ son:

$$\mu = 101.4 \qquad \text{y} \qquad \sigma^2 = 250.9$$

Tabla 7.1 Salarios de una población de 50 ingenieros.

ID	Salario	ID	Salario	ID	Salario	ID	Salario	ID	Salario
01	104	11	92	21	93	31	92	41	99
02	111	12	84	22	90	32	98	42	100
03	85	13	109	23	70	33	101	43	129
04	90	14	120	24	113	34	88	44	87
05	84	15	126	25	113	35	112	45	95
06	100	16	61	26	121	36	108	46	94
07	100	16	61	26	121	36	108	46	94
08	77	18	81	28	137	38	120	48	114
09	106	19	118	29	126	39	104	49	77
10	123	20	11	30	98	40	105	50	98

Objetivo: El objetivo es observar que cuando se toma una muestra, ésta nos aproxima al valor real de la media poblacional; en la práctica este valor es desconocido. En este caso se tiene una población pequeña, $n = 50$, pero real de la cual se puede calcular directamente la media μ, y la varianza σ^2. La meta de este planteamiento es que a partir de esta información se genere el procedimiento para seleccionar una muestra de la población y con esos datos calcular la media y varianza.

La media y la desviación estándar para la muestra son $\overline{X}$ y S, respectivamente. Así, para estimar la media de la población con la media de la muestra, se realiza lo siguiente:

$$\mu = \overline{X} - \text{un error} \qquad \text{o bien} \qquad \mu = \overline{X} + \text{un error} \tag{7.2}$$

Las preguntas centrales en este caso son: ¿qué tan pequeño es el error? ¿con qué confianza obtenemos el resultado? Con el propósito de motivar estas ideas se plantea una estrategia para seleccionar la muestra. Para el contexto de la estimación de la media, observe que el error realmente mide la discrepancia entre los valores del estadístico y el parámetro: $error = \overline{X} - \mu$.

Procedimiento: Se requiere de un procedimiento aleatorio para seleccionar una muestra, esto con el propósito de evitar alguna tendencia personal o criterio de la persona que toma la muestra. Básicamente se siguen tres formas de escoger una muestra. Una de ellas es considerar una urna o en su lugar una caja o bolsa, lo cual es laborioso por tener que hacer el material. La segunda es contar con una tabla de números aleatorios; esta es eficiente pero hay que ir observando los números. Mediante el uso de la tecnología se puede obtener una muestra de números aleatorios, para ello se requiere el empleo de un paquete estadístico, que resulta muy práctico. Además éste tiene la facilidad de seleccionar la muestra directa de valores de una lista finita de la población. Sin embargo, el procedimiento tiene la limitante de que el algoritmo depende del valor de inicio para generar la muestra. Lo que suele llamarse muestreo seudo aleatorio.

Atendiendo a la primera condición de selección de la muestra, se meten en un caja 50 papeles numerados del 1 al 50 y se selecciona una muestra de tamaño $n = 5$. Una vez que se saca un papel se observa el número, se anota el salario al que se refiere, luego se regresa el papel y se repite la regla, cuatro veces en este caso. Los cinco papeles que salieron para la muestra son: 18, 41, 40, 45, 22 y los salarios correspondientes son: 81, 99, 105, 95, 90. Por lo tanto, la media muestral es:

$$\overline{x} = \frac{81 + 99 + 105 + 95 + 90}{5} = \frac{470}{5} = 94$$

Este valor es una estimación puntual del valor de la media poblacional μ. Como se puede observar, existe una discrepancia con el valor real de la media de 7.4 (valor del error), utilizando la ecuación 7.2 se sigue:

$$\mu = \overline{x} + \text{error} = 94 + 7.4$$

Muestreo con reemplazo y sin reemplazo

Como se habrá observado, al extraer el papel que identifique a una persona para conocer su salario, éste puede regresarse a la caja (*con reemplazo*) o no regresarse (*sin reemplazo*). Reemplazar el papel en un millón de casos realmente no importa. De hecho, al reemplazar el primer papel antes de sacar el segundo, las observaciones en la primera y la segunda extracciones son totalmente independientes. No obstante, si no se reemplaza el papel, el segundo resultado de la extracción afectará ligeramente la segunda en poblaciones grandes. En cambio, en las poblaciones pequeñas el efecto sí es importante. Por otro lado, el muestreo con reemplazo genera observaciones independientes, es decir el resultado de una selección no se ve afectada por un selección previa.

Muestreo aleatorio simple: El desarrollo matemático es más sencillo si las observaciones son independientes. En este capítulo se va a suponer el muestreo aleatorio con reemplazo, al cual también suele llamarse *muestreo aleatorio simple.* Además se contemplará el tema de muestreo sin reemplazo y se señalarán las correcciones que se necesitan hacer en el procedimiento de estimación de acuerdo con la forma de selección de la muestra. Existen otros tipos de muestreo y se indicarán más adelante.

Ejemplo 7.1.1

Caso: $n = 5$. Siguiendo los datos de salario de *El mundo de la información 1*, este ejemplo tendrá soluciones para diferentes tamaños de muestra. 1.- Obtener 12 muestras aleatorias simples de tamaño $n = 5$, estimar la media en cada caso y la discrepancia con respecto a la media poblacional μ. 2.- Describir de manera gráfica los resultados del inciso anterior.

Solución. Selección de muestras de tamaño $n = 5$

Las siguientes muestras fueron seleccionadas mediante el muestreo aleatorio simple, ver la tabla 7.2. En los últimos dos renglones se han calculado $\overline{x}$ y $d = \overline{x} - \mu$, y para cada caso $d_i = \overline{x}_i - \mu$ con $i = 1, 2, \ldots, 12$

Tabla 7.2 Selección de 5 muestras de tamaño $n = 5$ y el cálculo de $\overline{x}$ y $\overline{x} - \mu$.

1	2	3	4	5	6	7	8	9	10	11	12
120	104	137	81	129	113	85	104	113	98	81	105
84	111	137	92	113	100	84	61	87	93	129	120
111	85	113	111	92	118	90	113	70	112	87	100
118	84	84	100	70	108	111	95	88	104	104	70
61	90	104	95	87	87	104	94	113	104	114	84
98.8	94.8	115	95.8	115	98.2	105.2	94.8	93.4	94.2	103	95.8
-2.6	-6.6	13.5	-5.6	-3.2	9.8	-6.6	-8.0	-7.2	0.8	1.6	-5.6

En la figura 7.5 se describen 12 muestras de tamaño $n = 5$, para cada una se ha calculado la media y se observa la discrepancia que existe con respecto a la media de la población. Se observa que con esta pequeña información, $\overline{x}$ se aproxima a μ, donde la media del penúltimo renglón es $\overline{\overline{x}} = 100.3$ y la media de la discrepancia $d_i = \overline{x}_i - \mu$ es -1.64, es en promedio pequeña. En la vida real sólo se escoge una muestra y la idea es que la discrepancia entre estas medias sea lo menor posible y de esa manera inferir sobre el parámetro.

Las preguntas que se plantean son ¿qué conocimiento sobre la población se debe tener, y qué propiedades estadísticas se requieren para alcanzar esa meta? En resumen, la población debe ser lo más homogénea posible, de no ser así se requiere otro método de muestreo. Por otro lado, desde un punto de vista técnico, si este proceso se repite varias veces, se generan diferentes valores para la media, así se puede considerar la variable aleatoria la media del salario $\overline{X}$. Ahora, se necesita elaborar el procedimiento estadístico para que a partir de las características de la muestra, se conozca la distribución de probabilidad de la variable $\overline{X}$

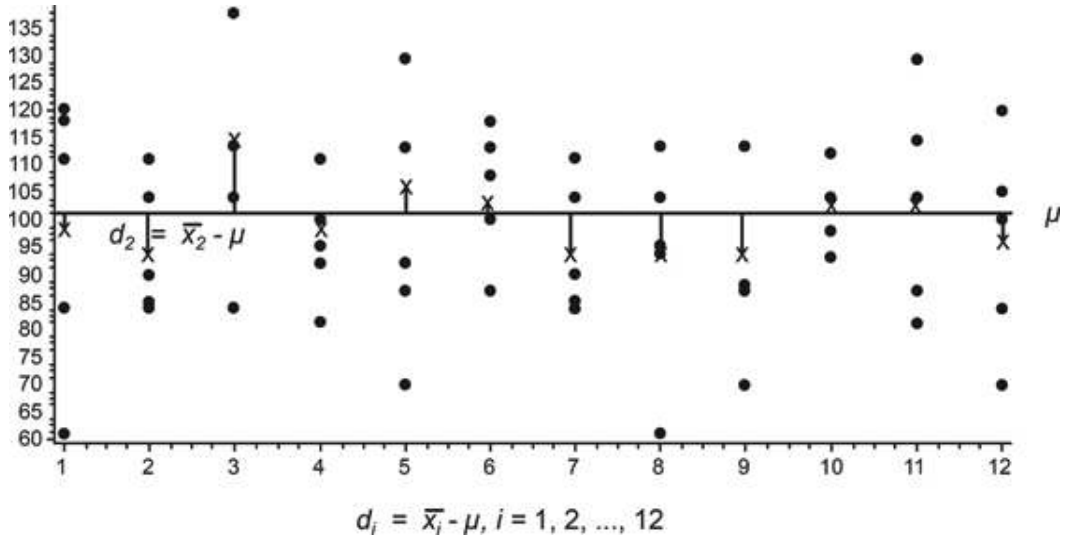


Figura 7.5 Describe 12 muestras de tamaño 5, la x representa la media en la muestra, la línea es la media de la población.

Ejemplo 7.1.2

Caso $n = 10$.

1.- Seleccionar de manera aleatoria 5 muestras de tamaño $n = 10$, siguiendo el mismo esquema del primer inciso, y calcular $\overline{x}$ y $\overline{x} - \mu$. 2.- Describir los procedimientos aleatorios para seleccionar una muestra. 3.- Evaluar el efecto del tamaño de la muestra, mostrar los resultados y explicar ésta situación en problemas reales. ¿Qué diferencias observa en $\overline{x} - \mu$ para los casos mostrados?

Solución. Selección de una muestra aleatoria de tamaño $n = 10$

Ahora se extrae de la caja con reemplazo 10 papeles y anotamos el número que corresponde al salario. Las muestras y los valores seleccionados de 12 muestras de tamaño $n = 10$ y el cálculo de $\overline{x}$ y $\overline{x} - \mu$, se presentan en la tabla 7.3. Se puede observar que también existe discrepancia entre las medias $d = \overline{x} - \mu$ en varios casos ésta tiene la apariencia de ser más pequeña que en caso de $n = 5$.

Desde luego hay mayor información, pero aun así sigue habiendo diferencia entre las medias. $d_i = \overline{x}_i - \mu$. Observe que la media del penúltimo renglón es $\overline{\overline{x}} = 101.725$, que se aproxima a la media $\mu = 101.4$, además la media de la discrepancia, $d_i = \overline{x}_i - \mu$ es -0.475 que en promedio resulta pequeña, lo que indica que al aumentar el tamaño de muestra la estimación es mejor.

Tabla 7.3 Selección de 5 muestras de tamaño $n = 10$ y el cálculo de $\overline{x}$ y $\overline{x} - \mu$

1	2	3	4	5	6	7	8	9	10	11	12
129	11	113	93	123	94	100	88	111	118	105	112
111	104	87	84	84	61	90	84	126	113	98	87
95	84	126	129	99	100	85	117	84	85	114	123
123	137	98	90	87	61	109	85	113	101	98	111
112	105	87	93	85	106	90	98	129	111	101	94
113	90	98	88	120	120	77	117	112	77	11	113
98	101	105	99	100	121	112	61	100	87	101	121
109	92	106	120	104	137	98	61	126	120	81	70
113	90	104	98	88	94	105	105	95	111	106	90
126	99	87	88	104	77	104	126	105	121	61	81
112.9	101.3	101.3	98.2	99.4	97.1	97.0	94.2	110.1	104.4	94.6	100.2
11.5	-0.1	-0.1	-3.2	-2.0	-4.3	-4.4	-7.2	8.7	3.0	-6.8	-1.2

Ejemplo 7.1.3

Diferentes procedimientos aleatorios para seleccionar una muestra.

Solución del procedimiento para la selección de una muestra aleatoria

Existen varios procedimientos para seleccionar una muestra; los más utilizados son la urna, la tabla de números aleatorio y mediante tecnología usando el material educativo CalEst.

Mediante una urna. Es un mecanismo adecuado para seleccionar una muestra, requiere la inversión para tener uno. En su defecto, se puede utilizar una caja con papeles numerados.

Utilizando una tabla de números aleatorios. En la tabla 7.4 se despliega una parte pequeña de números aleatorios dispuestos en filas y columnas. Una alternativa para seleccionar una muestra es escoger una columna y un renglón de manera aleatoria; por ejemplo, suponga que se tiene en la columna 6 y el renglón 7 el número 6250. Dado que la lista de salarios es de 50 números, en este caso se toman por pares, es decir el 62 y el 50; como el 62 no está en la lista se desecha y se toma el 50, así el salario respectivo es 98. Se sigue con el siguiente número, ya sea siempre de manera horizontal o vertical y se repite la regla. Existen tablas más grandes. Otra opción es usar el directorio telefónico en la sección blanca, se selecciona una hoja al azar. Para escoger un número de la lista de 1 a 50, se pueden considerar las últimas dos cifras del número del teléfono. Continuar con el número de teléfono que sigue, así hasta completar el tamaño de la muestra deseada.

Tabla 7.4 Números aleatorios.

	1	2	3	4	5	6	7	8	9	10	11	12
1	0690	6873	0178	5604	8337	6133	1385	0801	3767	1263	1064	4334
2	6792	2551	7005	8120	4329	5040	4237	4333	6846	2075	6301	1162
3	5987	0141	2634	2267	1956	0228	1314	4393	5444	8820	4357	6456
5	0465	0062	4160	2437	9790	7434	0587	2607	6363	5325	8185	3084
6	7096	6616	3118	1916	4463	3186	0305	5188	1686	0964	3193	4690
7	3157	5876	3498	2646	9368	6250	0470	5432	3111	7054	7882	9462
8	0045	6554	5076	4475	8604	2463	3349	1123	7606	5923	0750	2761
9	3364	7862	8652	0132	9559	1014	1946	1529	6474	7965	0092	1297
10	6926	3120	4627	2970	8423	6383	4336	5568	3637	6926	0971	2233

Solución mediante el uso de CalEst

(Apartado Herramientas opción Generador de números). Una alternativa: con el avance tecnológico, ahora se puede seleccionar una muestra mediante el software estadístico. En particular utilizando el **CalEst**, se elige el apartado de herramientas, ahí aparece la opción seleccionar datos aleatoriamente, tal y como aparece en la figura 7.6. En la hoja de captura aparecen los 50 datos de salarios, al emplear la opción indicada, el proceso para obtener una muestra es como sigue:

- **1)** Escoger la columna de datos de la cual se quiere tomar una muestra.
- **2)** Indicar el tamaño de la muestra.
- **3)** Indicar la columna donde desea que aparezcan los datos de la muestra.

Repitiendo este mecanismo se pueden generar varias muestras y luego continuar con el análisis estadístico. Esta situación ahorra tiempo y esfuerzo en la elección de la muestra. En este apartado de herramientas en **CalEst** también aparece la opción de generar números aleatorios.

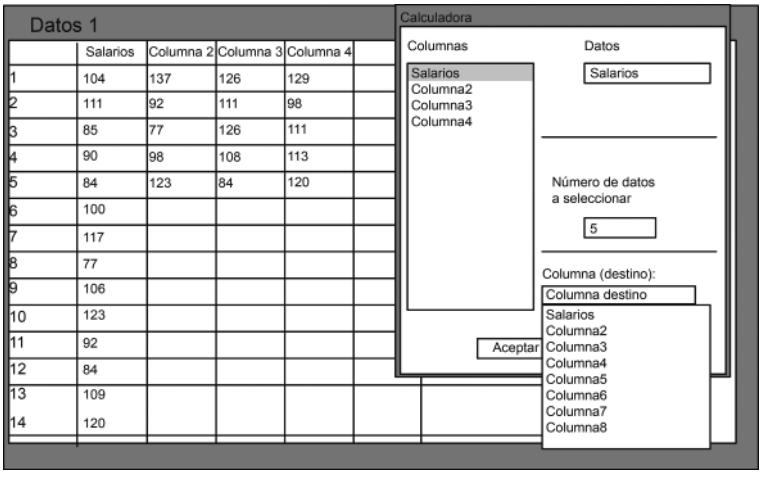


Figura 7.6 Procedimiento para elegir una muestra mediante el CalEst.

Ejemplo 7.1.4

Caso: evaluar el efecto del tamaño de la muestra, mostrar los resultados y explicar ésta situación en problemas reales. ¿Qué diferencias observa en $\overline{x} - \mu$ para los casos mostrados?

Solución efecto del tamaño de muestra

Observemos que la discrepancia entre la media muestral $\overline{x}$ y la media poblacional μ es menor cuando la *muestra es de mayor tamaño*. ¿Qué ocurriría si se aumentara el tamaño de la muestra? ¿Por qué? Una vez descubierto el procedimiento para seleccionar muestras aleatorias, en la siguiente tabla se presentan las medias $\overline{x}$ para muestras de tamaño $n = 15$ en el primer renglón, la media para las 12 columnas es $\overline{\overline{x}} = 101.658$, y $n = 30$ en el segundo, la media de las 12 columnas es $\overline{\overline{x}} = 100.5$.

1	2	3	4	5	6	7	8	9	10	11	12
97.9	105.1	102.9	103.3	104.9	103.3	99.5	99.0	103.7	101.9	97.9	100.5
103.6	96.5	101.5	102.1	96.6	98.9	102.9	99.5	99.4	103.5	103.2	98.3

Desde luego, cuando se tiene mayor información, n más grande, la discrepancia entre la media de la muestra y la media de la población se hace más pequeña. En la tabla 7.5 se observa una comparación entre las diferencia de las medias conforme aumenta el tamaño, n, de la muestra. Así en el primer renglón $n = 5$, luego $n = 10$, $n = 15$ y $n = 30$ respectivamente.

En la figura 7.7 se explica la distribución de la variable $\overline{X}$ para distintos valores de n. Se observa que la distribución $\overline{X}$ con $n = 30$ es más compacta alrededor de la media.

Para examinar la distribución de $\overline{X}$ con diferentes tamaños de muestra se usa la información del ejemplo anterior. En el primer caso, se obtienen 120 muestras de tamaño $n = 5$ y se calculan las medias. La distribución de esas 120 medias se observa en el primer histograma de la figura 7.7 donde se puede observar la variación de esta distribución.

Tabla 7.5 Presenta la comparación de la discrepancia al aumentar el tamaño de muestra

1	2	3	4	5	6	7	8	9	10	11	12
-2.6	-6.6	13.6	-5.6	-3.2	9.8	-6.6	-8.0	-7.2	0.8	1.6	-5.6
11.5	-0.1	-0.1	-3.2	-2.0	-4.3	-4.4	-7.2	8.7	3.0	-6.8	-1.2
-3.5	3.7	1.5	1.9	3.5	1.9	-1.9	-2.4	2.3	0.5	-3.5	0.9
2.2	-4.8	0.1	0.7	-4.7	-2.5	1.5	-1.9	-2.0	2	1.7	-1.3

En el segundo histograma se describe la distribución de $\overline{X}$ para 120 muestras de tamaño $n = 20$. Como se advierte en esta situación, existe menor variación con respecto a la distribución anterior. Finalmente, se aumentó el tamaño de la muestra a 30, y la distribución de $\overline{X}$ se presenta en el tercer histograma. Ahí se observa que los valores de la media muestral están más próximos al valor de la media. μ

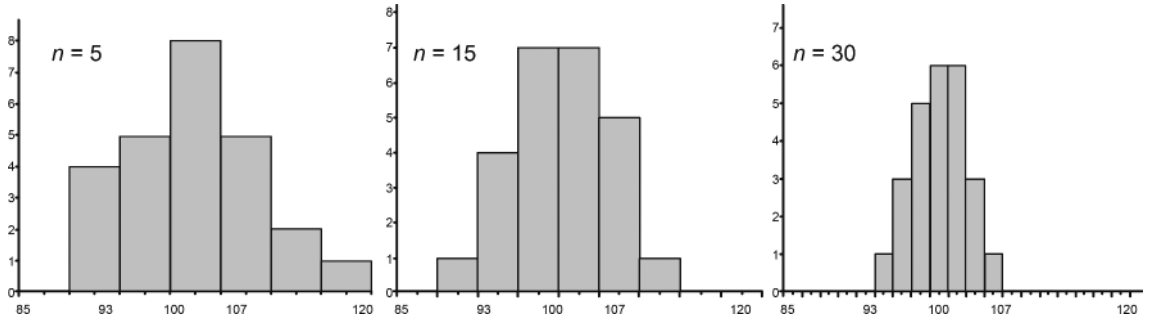
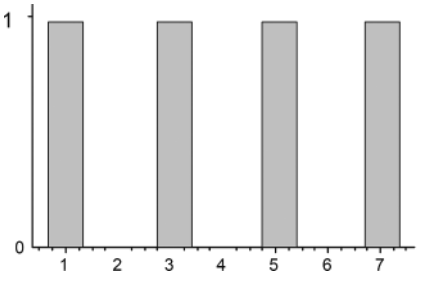


Figura 7.7 Distribución de las medias para 3 muestras de tamaño 5, 15 y 30, respectivamente, de la población del salario.

Es claro que se adquiere mayor conocimiento de lo que ocurre con los salarios de una población en la medida en que el tamaño de la muestra crezca, pero desde luego tendría que gastar más recursos y tiempo para obtener esa información.

7.3 Distribución de la media muestral

Con la finalidad de tener una idea del procedimiento de la relación de la población con una muestra, se plantea, de manera hipotética, una población pequeña con la distribución de la variable aleatoria y se indican los parámetros para la media y varianza. De ahí se toman todas las posibles muestras con un determinado tamaño n y se calcula la media y varianza, de tal manera que ahora se tiene una población de muestras. A partir de esta información se estudia la distribución de probabilidad de las variables media y varianza de las muestras. Inicialmente se presentará la teoría para la primera, para el caso de la variable varianza de la muestra se motivarán algunas ideas y la estimación de la varianza se explicará en otro apartado de este capítulo.



X: número de pólizas vendidas				
Días	x	$x - \mu$	$(x-\mu)^2$	$p(x)$
1	3	-1	1	$\frac{1}{4}$
2	5	-1	1	$\frac{1}{4}$
3	1	-3	9	$\frac{1}{4}$
4	7	3	9	$\frac{1}{4}$
Total	16	0	20	$\frac{1}{4}$

Figura 7.8 Distribución del número de pólizas vendidas, con media $\mu = 4$ y varianza $\sigma^2 = 5$.

Ejemplo 7.2.1

Un agente de seguros utiliza 4 días de la semana para salir a vender. Después de una publicidad realizada por la agencia, el número de pólizas vendidas fue de $x_1 = 3$, $x_2 = 5$, $x_3 = 1$, $x_4 = 7$. Esta pequeña referencia consiste en una población de cuatro elementos, $N = 4$. Se tiene que la media es $\mu = 4$ y la varianza $\sigma^2 = 5$, ver capítulo 3. La distribución de probabilidad de la variable X, número de pólizas vendidas caracteriza a esta población. Se observa en el eje horizontal en la gráfica de la figura 7.8, el número de pólizas vendidas en cuatro días, la frecuencia en el eje vertical y ésta es uniforme. En el cuadro, al lado de la gráfica se hace un resumen estadístico que permitirá calcular la media y varianza, respectivamente; en la última columna se indica la frecuencia relativa con $p(x)$.

Los cálculos para la media y la varianza son:

$$\begin{aligned} \mu &= \sum_{i=1}^{N} x_i p(x_i) = 3 \cdot \frac{1}{4} + 5 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 7 \cdot \frac{1}{4} = 4 \\ \sigma^2 &= \sum_{i=1}^{N} (x_i - \mu)^2 p(x_i) = (-1)^2 \cdot \frac{1}{4} + 1^2 \cdot \frac{1}{4} + (-3)^2 \cdot \frac{1}{4} + 3^2 \cdot \frac{1}{4} = 5 \end{aligned} \tag{7.3}$$

Si se toma una muestra de dos días de venta:

- **(1)** ¿Cuáles son todas la posibles muestras?
- **(2)** Estimar la media de cada muestra, la discrepancia con respecto a la media $\mu = 4$ y la varianza.
- **(3)** Obtener la media y la varianza de las medias de la muestra.
- **(4)** Calcule la suma del cuadrado de las discrepancias y divídalo entre el número de medias estimadas.
- **(5)** ¿Qué relación se observa con los valores de la media de la población y los resultados en el inciso (4)?
- **(6)** Bosquejar las gráficas de las distribuciones para la media y varianza.

Solución. Extracción de la muestra con reemplazo

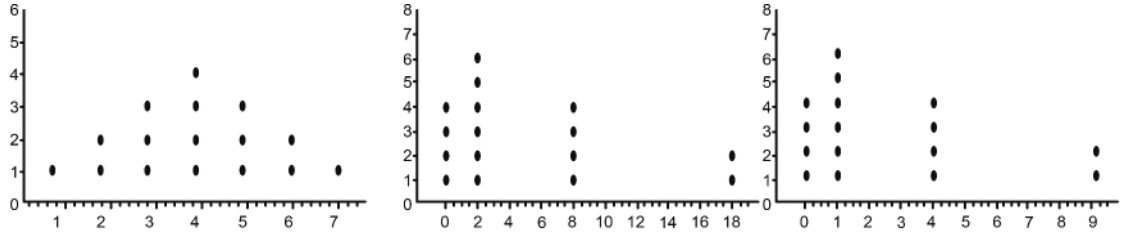
Se meten en una urna o bolsa cuatro canicas de diferentes colores para representar valores de las ventas de seguros y se extrae una canica, se observa el color y se anota el valor respectivo, se regresa ésta a la urna y se repite el proceso. Considere los siguientes colores: roja=3, verde=5, azul=1 y negra=7. Todos los posibles valores de una muestra de tamaño $n = 2$ así como la media, la varianza y la diferencia entre las medias al cuadrado se presentan en la tabla 7.6; esta información se sintetiza en las gráficas de la figura 7.9. Para fijar ideas, ahí se puede observar la simetría de $\overline{X}$, así como la asimetría de las distribuciones de S^2 y $(\overline{x} - \mu)^2$, las cuales se comentarán al final del capítulo.

Tabla 7.6 Cálculo de la media, varianza de las muestras de tamaño 2 de una Población venta de pólizas.

Ref	Muestra	$\overline{x}$	S^2	$\overline{x} - \mu$	$(\overline{x} - \mu)^2$	Ref	Muestra	$\overline{x}$	S^2	$\overline{x} - \mu$	$(\overline{x} - \mu)^2$
1	1 1	1	0	-3	9	9	5 1	3	8	-1	1
2	1 3	2	2	-2	4	10	5 3	4	2	0	1
3	1 5	3	8	-1	1	11	5 5	5	0	1	1
4	1 7	4	18	0	0	12	5 7	6	2	2	4
5	3 1	2	2	-2	4	13	7 1	4	18	0	0
6	3 3	3	0	-1	1	14	7 3	5	8	1	1
7	3 5	4	2	0	0	15	7 5	6	2	2	4
8	3 7	5	8	1	1	16	7 7	7	0	3	9
Suma		24	40	-8	20			40	40	8	20

Con los valores reportados en la tabla 7.6 se realizan los cálculos apropiados para la media y la varianza de la variable media muestral $\overline{X}$, éstos se expresan en la ecuación (7.4), nótese que $q(x_i)$ se refiere a la frecuencia de los valores en cada distribución. Para aplicar ésta expresión, note $x_i = \overline{x}_i$, es decir:

$(\overline{x}_i q(\overline{x}_i))$, $x_i = S_i^2$, es decir: $(S_i^2 q(S_i^2))$, y $x_i = (\overline{x}_i - \mu)^2$, es decir: $((\overline{x}_i - \mu)^2 q((\overline{x}_i - \mu)^2))$. La media $(\overline{X})$ se denota por $\mu_{\overline{X}}$ y la varianza $(\overline{X})$ por $\sigma^2_{\overline{X}}$. Advierta que $\mu_{\overline{X}} = \text{media}(\overline{X}) = 4$ coincide con la media μ, y la media de los valores de S^2 coincide con el parámetro σ^2. Finalmente $\sigma^2_{\overline{X}} = \text{varianza}(\overline{X}) = 2.5$, corresponde a $\frac{1}{2}$ de la varianza σ^2, registre que $n = 2$. ¿Estos resultados se pueden formalizar? ¿Cómo?



$\overline{X}$	1	2	3	4	5	6	7	Total
$q(X_i)$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	1
$\overline{X}q(X_i)$	$1\frac{1}{16}$	$2\frac{2}{16}$	$3\frac{3}{16}$	$4\frac{4}{16}$	$5\frac{3}{16}$	$6\frac{2}{16}$	$7\frac{1}{16}$	$\frac{64}{16}$

S^2	0	2	8	18	Total
$q(X_i)$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	1
$S^2q(X_i)$	$0\frac{4}{16}$	$2\frac{6}{16}$	$8\frac{4}{16}$	$18\frac{2}{16}$	$\frac{80}{16}$

$(x-\mu)^2$	0	1	4	9	Total
$q(X_i)$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	1
$(x-\mu)^2q(X_i)$	$0\frac{4}{16}$	$1\frac{6}{16}$	$4\frac{4}{16}$	$9\frac{2}{16}$	$\frac{40}{16}$

Figura 7.9 Distribución de las variables muestrales y el cálculo correspondiente de la media y la varianza.

$$\begin{aligned}
\mu_{\overline{X}} &= \sum_{i=1}^{16} \overline{x}_i q(x_i) = 1 \cdot \frac{1}{16} + 2 \cdot \frac{2}{16} + 3 \cdot \frac{3}{16} + 4 \cdot \frac{4}{16} + 5 \cdot \frac{3}{16} + 6 \cdot \frac{2}{16} + 7\frac{1}{16} = 4 \\
S^2 &= \sum_{i=1}^{16} S_i^2 q(x_i) = 0 \cdot \frac{4}{16} + 2 \cdot \frac{6}{16} + 8 \cdot \frac{4}{16} + 18 \cdot \frac{2}{16} = \frac{80}{16} = 5 \\
\sigma^2_{\overline{X}} &= \sum_{i=1}^{16} (\overline{x}_i - \mu)^2 q(x_i) = 0 \cdot \frac{4}{16} + 1 \cdot \frac{6}{16} + 4 \cdot \frac{4}{16} + 9 \cdot \frac{2}{16} = \frac{40}{16} = 2.5
\end{aligned} \tag{7.4}$$

Ejemplo 7.2.2

Analizar el caso de seleccionar una muestra sin reemplazo

Solución. Extracción de la muestra sin reemplazo

Con la finalidad de tener un panorama completo en la toma de una muestra, se ejemplifica el tema sin reemplazo. En el apartado de resultados teóricos se muestran los ajustes en la estimación de parámetros propios de este caso. En este esquema considere primero todas las posibles muestras que se obtienen sin reemplazo; del ejemplo se tiene una población $N = 4$ elementos, para seleccionar las muestras de tamaño n se utiliza la expresión ${}_N C_n^{(1)}$, en este ejemplo ${}_4C_2 = \frac{4!}{2!(4-2)!} = 6$. En la tabla 7.7 se presentan las muestras y los cálculos relacionados. La fórmula de combinación para obtener todas las muestras sin

reemplazo de tamaño n es ${}_NC_n = \frac{N!}{n!(N-n)!}$.

Los cálculos afines a la tabla 7.7, media$(\overline{X}) = \mu_{\overline{X}}$ son:

$$\begin{aligned}\mu_{\overline{X}} &= \sum_{i=1}^{6} \overline{x}_i q(x_i) = 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{2}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6} = \frac{24}{6} = 4 \\ \sigma^2_{\overline{X}} &= \sum_{i=1}^{6} (x_i - \mu)^2 q(x_i) = 0\cdot\frac{2}{6} + 1\cdot\frac{2}{6} + 4\cdot\frac{2}{6} = \frac{10}{6} = \frac{5}{3}\end{aligned} \tag{7.5}$$

donde $x_i = \overline{x}_i$, $x_i = S_i^2$ y $x_i = (\overline{x}_i - \mu)^2$, la media$(\overline{X})$ coincide con μ para el caso de la varianza se requiere un ajuste, que se verá en el siguiente apartado.

Tabla 7.7 Cálculo del muestreo sin reemplazo.

Muestra	$\overline{x}$	S^2	$(\overline{x} - \mu)^2$
(3,5)	4	2	0
(3,1)	2	2	4
(3,7)	5	8	1
(5,1)	3	8	1
(5,7)	6	2	4
(1,7)	4	18	0
	24	40	10

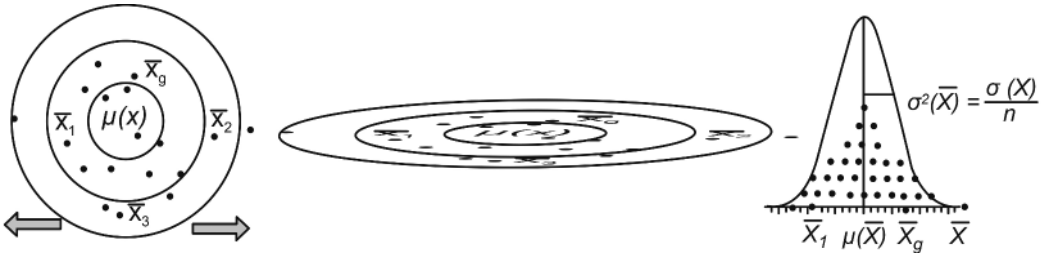


Figura 7.10 Una población de medias muestrales, cuya finalidad es estimar el parámetro $\mu(X)$, luego su distribución $(\overline{X})$ y sus características.

Resultado teórico de la variable aleatoria $\overline{X}$

El valor de la media muestral $\overline{X}$ varía de una muestra a otra como se puede observar en la figura 7.5. Con el fin de evaluar o tener una idea sobre las medias que se obtienen de las observaciones en una muestra, imagine que cada $\overline{x}$ es el tiro realizado por un tirador a un blanco, donde el blanco es el

parámetro. Si los valores de la media se colocan alrededor del parámetro, como se observa en la figura 7.10, éstos tienden a una distribución de probabilidad. Esta es normal si la distribución de la variable original X es normal, o si el tamaño de muestra es grande como se verá a continuación. Asimismo en la figura 7.10 se puede observar la cercanía al parámetro, situación que da lugar a *la precisión medida mediante la varianza. El tamaño de la muestra* también desempeña un papel importante en *la precisión de la estimación.* Llamamos a $\overline{X}$ un estimador puntual, pero éste también es una variable aleatoria. En unidades anteriores se explicó que una variable aleatoria tiene una media, una desviación estándar y una distribución de probabilidad, de modo que la variable aleatoria $\overline{X}$ tiene una media, una desviación estándar y una distribución de probabilidad.

Se hace notar que la media es ahora la media de todos los posibles valores de $\overline{X}$ y se denota por:

$$\mu(\overline{X}) = \mu_{\overline{X}}$$

Ésta corresponde al parámetro de la distribución de $\overline{X}$. La varianza que se relaciona a los valores de $\overline{X}$ se indica por:

$$\sigma^2(\overline{X}) = \sigma^2_{\overline{X}}$$

De manera análoga, esta es el parámetro para la desviación estándar de la distribución $\overline{X}$. Mediante métodos matemáticos se puede demostrar que dichos parámetros se relacionaban con los parámetros de la variable aleatoria X. Esta asociación se establece mediante las siguientes expresiones. La media de la distribución muestral $\overline{X}$ es:

$$\mu_{\overline{x}} = \mu \tag{7.6}$$

Ajuste de la varianza cuando el muestreo es con reemplazo: la varianza de la distribución muestral $\overline{X}$ es:

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n} \tag{7.7}$$

La desviación estándar de la distribución muestral $\overline{X}$ es:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.8}$$

Ésta última expresión revela que la desviación estándar disminuye en la medida que el tamaño de la muestra crece, véase la figura 7.11. La desviación estándar $\sigma_{\overline{x}}$ se conoce como error estándar. La expresión 7.7 es válida cuando el muestreo se realiza con reemplazo, así como cuando la población es muy grande. De tal manera que si el muestreo se lleva a cabo sin reemplazo existe un factor de corrección, así la varianza de la distribución muestral es:

$$\sigma^2_{\overline{x}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) \tag{7.9}$$

Por consiguiente, el error estándar es:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{7.10}$$

donde $\sqrt{\frac{N-n}{N-1}}$ se conoce como el factor de corrección.

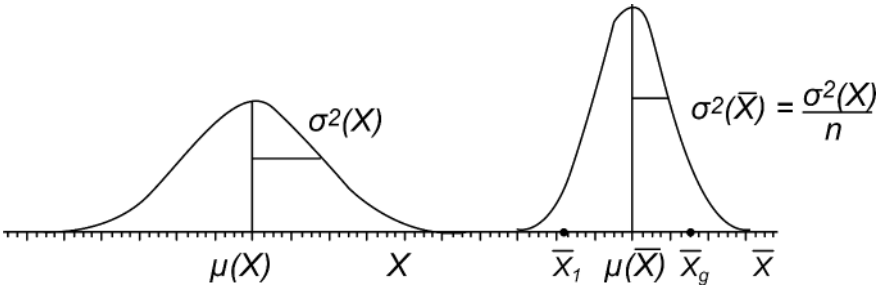


Figura 7.11 Descripción de la relación teórica entre las distribuciones de las variables X y $\overline{X}$.

Ejemplo 7.3

Con la finalidad de ejemplificar un caso sobre la relación entre la varianza de la muestra y el tamaño de la muestra, considérese la siguiente situación. Se ha seleccionado una muestra de tamaño $n = 10$ con $\sigma_{\overline{x}} = 9$, ¿cuántas observaciones más se necesita tomar para reducir $\sigma_{\overline{x}}$ a 4.5, a 3 o a 1?

Solución

La expresión $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ relaciona la desviación estándar de la media muestral $\overline{X}$, la desviación estándar de la variable X y el tamaño de muestra n. Con la información proporcionada por los datos se necesita conocer el valor de σ. Así, la expresión $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ se puede escribir como $\sigma = \sqrt{n}\sigma_{\overline{x}}$ o bien como $\sigma^2 = n\sigma^2_{\overline{x}}$. Sustituyendo los valores se tiene que $\sigma^2 = 10 \times 9^2 = 810$.

Para encontrar un valor del tamaño de muestra se requiere tener los valores de las desviaciones estándar σ y $\sigma_{\overline{x}}$, esto es, $n = \frac{\sigma^2}{\sigma^2_{\overline{x}}}$. Entonces, el tamaño de muestra si se desea reducir $\sigma_{\overline{x}}$ a 4.5 es:

$$n = \frac{\sigma^2}{\sigma^2_{\overline{x}}} = \frac{810}{(4.5)^2} = 40$$

Con las 10 observaciones que se tienen, entonces se necesitarán 30 observaciones más. Nótese que habrá un mejor conocimiento de los parámetros de la población si hay más observaciones en la muestra. Esta situación se ve reflejada si la desviación estándar $\sigma_{\overline{x}}$ disminuye; en tal caso, se dice que aumenta la precisión de la estimación.

Práctica mediante el uso de CalEst. La finalidad es ejemplificar la distribución normal y su relación entre la varianza σ^2 de la distribución de la variable X y la varianza $\sigma^2_{\overline{X}}$ distribución de la variable $\overline{X}$. Información: se tiene que $\sigma^2 = 729$ ($\sigma = 27$), $n = 9$, por lo tanto $\sigma^2_{\overline{X}} = 81$ ($\sigma_{\overline{x}} = 9$). La gráfica de la figura 7.12 extiende la idea de la gráfica de la figura 7.11.

La gráfica normal más extendida en la figura 7.12 muestra la normal con ($\mu = 60$, $\sigma = 27$) y la más delgada corresponde a una normal con ($\mu_{\overline{x}} = 60$, $\sigma_{\overline{x}} = 9$) distribución muestral. A partir de esta descripción se pueden realizar varios ejercicios para estudiar y comprender esta relación. Esta práctica será de mucha utilidad para comprender los conceptos que exponen en diferentes libros sobre inferencia estadística, y reproducir la descripción gráfica que presentan sobre la distribución normal, la normal estándar y la t-Student.

Resultado técnico: Si las mediciones de X vienen de una distribución normal, se sigue que la distribución muestral de $\overline{X}$ también es normal (véase la figura 7.12).

$$\overline{X} = \mu_{\overline{x}} + Z\sigma_{\overline{x}} = \mu_{\overline{x}} + Z\frac{\sigma}{\sqrt{n}}$$

En el proceso de estandarizar, ésta se puede escribir como una normal estándar, así:

$$Z = \frac{\overline{X} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

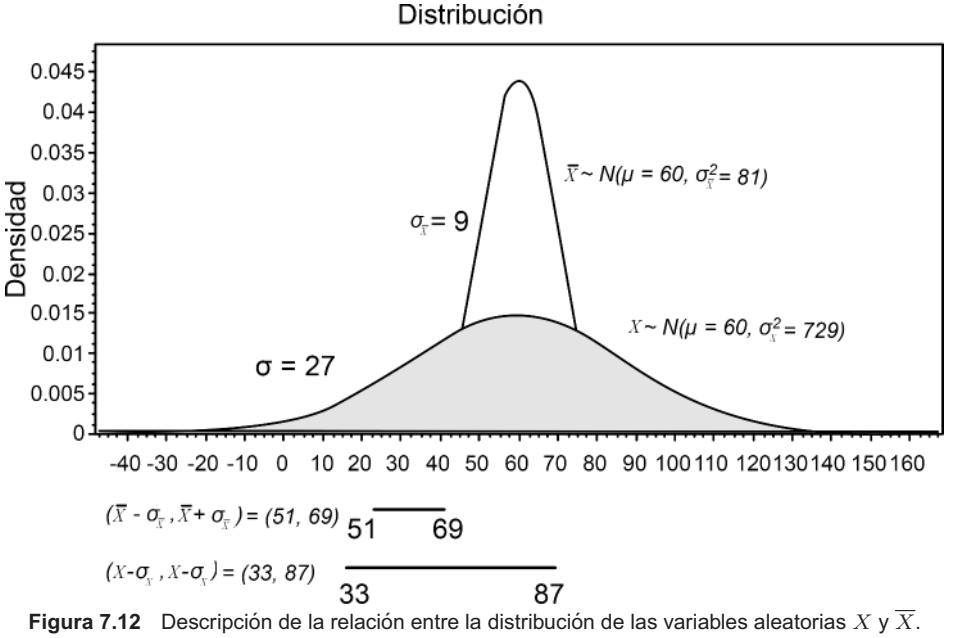


Figura 7.12 Descripción de la relación entre la distribución de las variables aleatorias X y $\overline{X}$.

Ejemplo 7.4

Las ventas X en miles de pesos por semana de una empresa que elabora alimentos tiene una distribución normal con parámetros $\mu = 750$ y $\sigma = 30$. Con el propósito de aumentar las ventas dicha empresa realiza una promoción; para evaluar la propaganda escogen al azar 25 tiendas. Si la promoción es exitosa esperan que en promedio las ventas sean mayor que 765 pesos. ¿Cuál es la probabilidad de que la media muestral sea mayor que 765? Interprete la respuesta de este cálculo.

Solución

La solución clásica es calcular la media y la varianza, en este caso las derivadas de la muestra, y estandarizar para consultar las tabla de la distribución normal estándar y conseguir el valor de la probabilidad. Considerando la información planteada, la distribución de la variable X es normal con parámetros $\mu = 750$ y $\sigma = 30$, entonces $\overline{X}$ tendrá una distribución normal con media $\mu = 750$ y desviación estándar $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{25}} = 6$.

Se requiere encontrar la probabilidad $P(\overline{X} \geq 765)$, donde $\overline{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$, desde un punto de vista clásico se recurre al proceso de estandarización para utilizar la tabla de la distribución normal estándar:

$$P(\overline{X} \geq 765) = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{765 - 750}{6}\right) = P(Z \geq 2.5) = 0.006$$

La figura 7.13 muestra la relación entre la distribución de probabilidad normal sobre la variable de interés y la estandarizada. Puesto que Z es una variable normal estándar, su probabilidad se obtiene usando las tablas de la normal estándar.

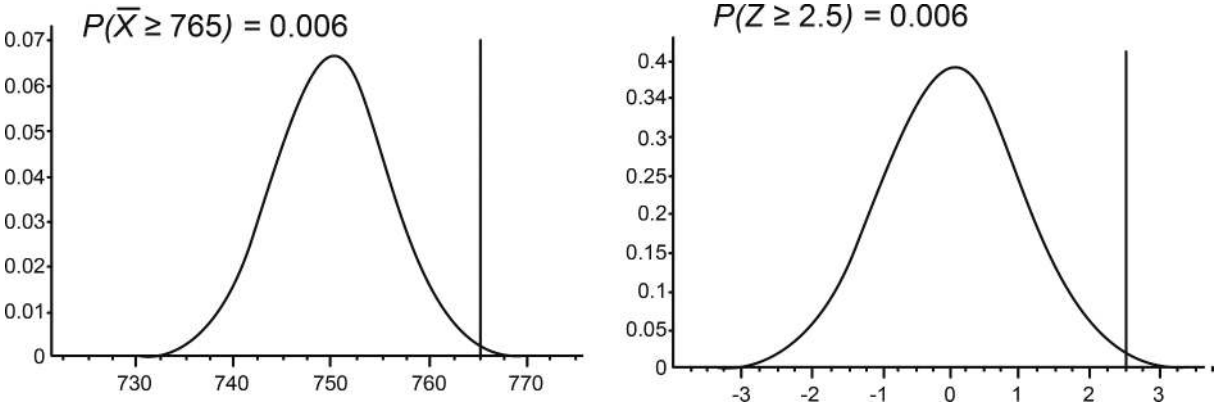


Figura 7.13 Cálculo considerando la distribución de probabilidad la normal original y la normal estándar.

Solución mediante el uso de CalEst

Vaya al módulo de distribuciones de **CalEst** y en éste use la normal con parámetros $\mu = 750$ y $\sigma_{\overline{X}} = 6$ (error estándar). Recuérdese que moviendo el cursor puede calcular cualquier probabilidad, como se ve en la figura 7.13, $P(\overline{X} \geq 765) = 0.006$. La alternativa a las tablas de la normal estándar es recurrir al calculador de la distribución normal, tal y como se vio en el capítulo 6, y se describe en la figura 7.14.

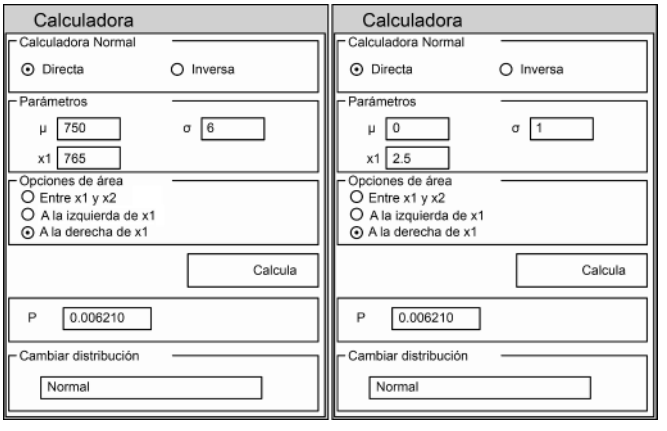


Figura 7.14 Cálculo de las probabilidades para una distribución normal, a la izquierda la original y a la derecha la estándar.

Teorema del límite central

Si se seleccionan muestras aleatorias de tamaño n de una población, sin importar la forma de su distribución, con media μ y desviación estándar σ, cuando n es suficientemente grande, la distribución de la variable $\overline{X}$ se aproxima a la distribución normal con:

$$\text{Media} \quad \mu_{\overline{X}}, \quad \text{tal que} \quad \mu_{\overline{X}} = \mu$$

$$\text{Desviación estándar} \quad \sigma_{\overline{X}} \quad \text{igual a} \quad \frac{\sigma}{\sqrt{n}}$$

7.4 Teorema del límite central

En función de los resultados observados en el ejemplo 7.3, se ve que cuando se aumenta el tamaño de muestra la distribución de $\overline{X}$ se aproxima a una distribución normal. Esta situación es importante y se enuncia en lo que se conoce como *teorema de límite central*.

Este resultado es muy apropiado, dado que especifica la distribución $\overline{X}$ para muestras grandes. Recurriendo a la simulación por computadora, en la figura 7.15 se muestran algunos casos.

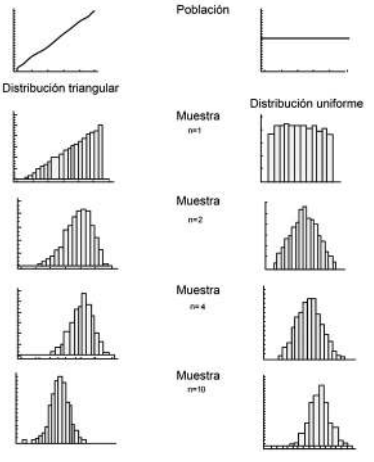


Figura 7.15 Ilustración de la distribución muestral para diferentes tamaños de muestra cuando la población original no tiene una distribución normal.

Por lo general, desde un punto de vista práctico es suficiente con que el tamaño de n sea de 20 o 30 para considerar la distribución de $\overline{X}$ como normal. En referencia a la figura 7.15, en las gráficas en la parte superior, se observa la forma de la distribución de la población, luego se examina con histogramas la distribución de la media muestral considerando diferentes tamaños de muestra. En el primer caso (gráfica superior izquierda) se presenta una distribución triangular. Si se selecciona una muestra de tamaño $n = 1$, la forma de la distribución muestral dada por el histograma es similar a la de la población. Si el tamaño de muestra n crece, la distribución de $\overline{X}$ se aproxima a una normal.

Una situación similar (gráfica superior derecha) ocurre cuando la distribución de la población es uniforme. Si la muestra es de tamaño $n = 1$, el histograma reproduce la distribución original. Si n crece, la distribución de $\overline{X}$ se aproxima a una normal.

Ejemplo 7.5

La administración de una empresa que manufactura aparatos eléctricos efectúa pruebas de habilidad y destreza durante el proceso de contratación de personal. La calificación de esas pruebas conforman una

población con una media $\mu = 112$ y desviación estándar $\sigma = 8$. a) ¿Cuál es la media $\mu_{\overline{x}}$ y la desviación estándar $\sigma_{\overline{x}}$ de una distribución muestral $\overline{X}$ cuya muestra es de tamaño $n = 16$? b) ¿Cuál es la probabilidad de que la calificación de la media muestral esté entre 109 y 115? c) Elaborar una tabla que muestre el efecto del tamaño de muestra considerando el inciso b e interprete.

Solución

En este caso se desea conocer los parámetros $\mu_{\overline{x}}$ y $\sigma_{\overline{x}}$ de la distribución muestral $\overline{X}$; entonces se recurre a los resultados del teorema del límite central. Para ello se utilizan las expresiones:

$$\mu_{\overline{x}} = \mu \quad \text{y} \quad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

Se tiene que $\mu_{\overline{x}} = \mu = 112$ y $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8}{\sqrt{16}} = 2$.

Observe que mediante la relación $\overline{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$, así para $\overline{X} = 109$ se tiene que $Z_i = Z(\alpha/2) = -1.5$ y con $\overline{X} = 115$ se obtiene $Z_d = Z(1 - \alpha/2) = 1.5$ entonces el resultado mediante la normal estándar es: $P(-1.5 \leq Z \leq 1.5) = 0.866$. Utilizando la alternativa del *CalEst* con parámetros $(\mu = 112, \sigma_{\overline{x}} = 2)$ se tiene que $P(109 \leq \overline{X} \leq 115) = 0.866$.

Tamaño de muestra	1	8	16	36	49
$\frac{\sigma}{\sqrt{n}}$	8	2.82	2	1.34	1.14
Proporción entre 109 y 115	0.29	0.71	0.87	0.97	0.99

Se subraya que al crecer el tamaño de muestra se tiene como resultado una menor variabilidad en la distribución de la media muestral. Por consiguiente, al tener una proporción mayor entre los valores de 109 y 115 se aproxima al valor real de la media este comprendida entre ellos.

Resumen del teorema del límite central

Se pueden combinar los tres puntos del teorema del límite central y se obtiene la figura 7.16, en la cual se describe la distribución muestral $\overline{X}$ cuando n es suficientemente grande. Bajo el supuesto de la distribución normal ($\mu = 112$, y $\sigma = 8$), se sabe que 68 % de los valores caen dentro de una desviación estándar de la media. Mientras que 95 % cae dentro de dos desviaciones estándar de la media y 99.7 % cae dentro de tres desviaciones estándar de la media. Del ejemplo 7.5 se puede precisar un intervalo que comprende una proporción fija de las medias muestrales. Así se puede determinar un intervalo que abarque el 95 % de las medias muestrales a partir de $n = 16$, esto es conocer los valores de $\overline{X}_i$ y $\overline{X}_d$ tal que la proporción sea 95, es decir $P(\overline{X}_i \leq \mu \leq \overline{X}_d) = 0.95$, los valores de la variable Z son:

$Z(\alpha/2) = Z(0.025) = -1.96$ y $Z(1-\alpha/2) = Z(0.975) = 1.96$, recuerde que: $\mu_{\overline{x}} = \mu$

$$\overline{X}_i = \mu - Z(\alpha/2)\frac{\sigma}{\sqrt{n}} = 112 - (1.96)\frac{8}{\sqrt{16}} = 108.08$$

$$\overline{X}_d = \mu + Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} = 112 + (1.96)\frac{8}{\sqrt{16}} = 115.92$$

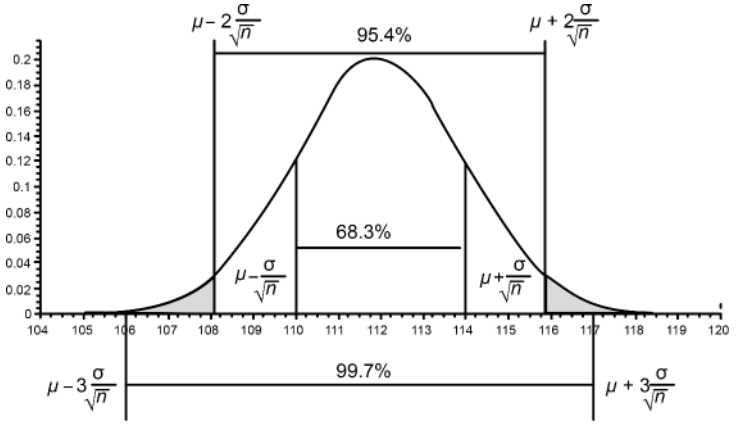


Figura 7.16 Ilustra la distribución muestral $\overline{X}$, a 1, 2 y 3 desviaciones estándar alrededor de $\mu_{\overline{x}} = \mu$.

De esta manera, el 95 % de todas las medias muestrales en referencia a $n = 16$ se encuentran entre 108.08 y 115.92.

Con respecto a la variable aleatoria $\overline{X}$, se puede advertir que el 68.3 % de las veces que observaremos una media muestral ésta caerá dentro de una desviación estándar alrededor de la media poblacional μ desconocida. De manera similar, 95.4 % de las veces observaremos una media muestral que cae dentro de dos desviaciones estándar de μ, y 99.7 % de las veces veremos una media muestral que cae dentro de tres desviaciones estándar de μ. Esta idea da lugar al concepto conocido como *intervalo de confianza* o una *estimación por intervalo,* que se verá más adelante.

Ejemplo 7.6

Se muestra el teorema con el lanzamiento de dados, esta actividad es ilustrativa ya que permite con la simulación de dados tener una idea conceptual del TLC al lanzar uno, dos, tres y más dados. Es decir se tienen muestras con $n = 1, 2, 3, \ldots, 10$. En este ejemplo se mostrará el caso para $n = 2$, se describe el espacio muestra y se toma la suma de los dados para construir la distribución de $\overline{X}$, se estima la media y la varianza de las medias de las muestras.

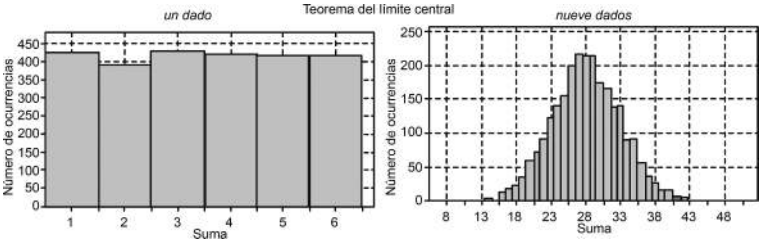


Figura 7.17 Ejemplifica el teorema de límite central, distribución uniforme $n = 1$, luego con $n = 9$.

Solución mediante el uso de CalEst

Mediante **CalEst** se ilustra la regularidad estadística (probabilidad frecuentista). Primer caso, se lanza un dado $n = 1$; el modelo para este experimento es una distribución uniforme. Para observar esta situación lance 1000 veces el dado de 100 en 100 ¿Qué observa? El número de lanzamientos del dado puede seguir creciendo. Ahora lance 5000 veces el dado de 100 en 100 vaya observando los cambios en la distribución.

Variando el número de dados lanzados, por ejemplo para $n = 2, n = 3, n = 4, n = 5$ y $n = 10$. Practicar el mismo procedimiento lanzando los dados 1000, 2000, 3000, 4000 y 5000, al realizar esta actividad varias veces. En cada caso la distribución tiende a hacerse simétrica. Los casos para $n = 1$ y $n = 9$ se presentan en la figura 7.17, la gráfica de la izquierda muestra el caso con un dado, distribución aproximadamente uniforme. Luego se muestra la situación con nueve dados, se observa en este caso una distribución cercana a una normal.

7.5 Intervalos de confianza para una media, proporción y varianza

Idea general de un Intervalo de confianza: media. En la figura 7.18 se muestran las ideas centrales en el proceso de estimación de un intervalo de confianza para la media μ. Considere la variable aleatoria: el gasto al semestre que realizan los estudiantes universitarios en libros. Se identifica como población a las personas que realizan estudios universitarios y la variable aleatoria tiene una distribución de probabilidad con una media y varianza. Como punto de partida, no se conocen ni la distribución ni los parámetros que la caracterizan. Así, la meta es plantear una estrategia para verificar cuál es la distribución de probabilidad que tiene la variable X y estimar los parámetros de ésta. Ante esta situación, el interés primario es construir un intervalo de confianza para la media; para explicar el procedimiento se supone que la distribución de probabilidad es *conocida* y es una *normal*; del mismo modo se supone que la varianza es *conocida*, en la figura 7.18, se pone con signo de interrogación para tenerla presente. En una próxima sección se relajará el supuesto de la varianza. En otro capítulo se presentará el procedimiento para verificar si se cumple el supuesto de normalidad.

Las líneas verticales entre la $\mu =?$ son los límites que establecen el nivel de confianza; como referencia se requerirán los valores de la variable aleatoria Z, normal estándar. Se observa la probabilidad entre esos dos valores, lo que se denomina el 95 % de confianza, gráfica derecha de la figura 7.19. A partir de la expresión:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

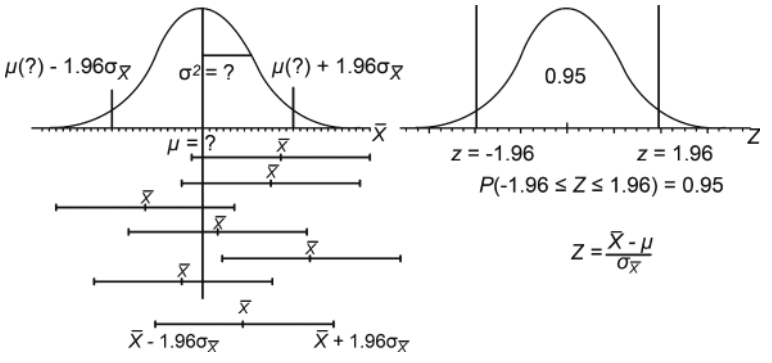


Figura 7.18 Presentación gráfica de la estimación de la media por intervalo, izquierda. Nivel de confianza, gráfica derecha.

Recuerde que $\overline{X}$ es un estimador puntual, en este caso tiene una distribución normal con media μ y varianza $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$, la relación entre las variables $\overline{X}$ y Z es:

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

El valor de Z se sustituye en $P(-1.96 \leq Z \leq 1.96) = 0.95$ y sigue que:

$$P(-1.96 \leq \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} \leq 1.96) = 0.95$$

realizando operaciones, se tiene la expresión equivalente:

$$P(-1.96\sigma_{\overline{X}} \leq \overline{X} - \mu \leq 1.96\sigma_{\overline{X}}) = 0.95$$

multiplicando por -1 cada término de la desigualdad y se obtiene la ecuación.
$P(-1.96\sigma_{\overline{X}} \leq \overline{\mu} - \overline{X} \leq 1.96\sigma_{\overline{X}}) = 0.95$, finalmente

$$P(\overline{X} - 1.96\sigma_{\overline{X}} \leq \mu \leq \overline{X} + 1.96\sigma_{\overline{X}}) = 0.95$$

La última ecuación indica que con una probabilidad del 0.95, μ estará entre los límites de $\overline{X} \mp 1.96\sigma_{\overline{X}}$. Una sencilla percepción de esta situación se ve reflejada en las 6 líneas abajo de la distribución, las cuales, en esta ilustración, 5 contienen a la media μ, y una no la contiene. La línea 7 en la figura 7.18 describe la idea general para un intervalo, los valores de la media $\overline{x}$ aparecen en el centro del intervalo.

La *distancia* para ambos lados del centro está indicada por la expresión

$$Z(\alpha/2)\sigma_{\overline{X}} = -Z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

donde $z(\alpha/2)$ es un valor de la variable aleatoria Z de una distribución de probabilidad normal estándar y sus valores escritos mediante una expresión de probabilidad $P(-1.96 \leq Z \leq 1.96) = 0.95$ muestran la confianza que se plantea para la estimación de la media con un nivel de confianza del 95 %, $\alpha = 0 - 05$. Así, la estimación de la media μ por intervalo, comprende los valores:

$$(\overline{x} - 1.96\sigma_{\overline{X}}, \overline{x} + 1.96\sigma_{\overline{X}})$$

y se le llama intervalo confianza del 95 %. En la figura 7.19 se muestra una idea general. Con el ánimo de integrar las ideas expuestas, se considera un ejemplo simplificado.

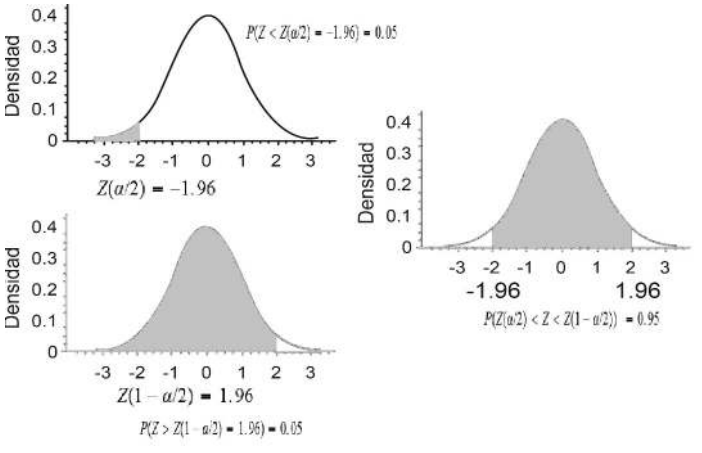


Figura 7.19 Distribución normal estándar para la construcción del intervalo de confianza.

Ejemplo 7.7

Para saber cuánto gana en el mercado laboral un ingeniero industrial recién egresado, se toma una muestra de tamaño $n = 30$ (se les pregunta su salario). Con la información proporcionada por la encuesta se obtiene una media de $\overline{x} = 6200$ del salario. La media muestral $\overline{x}$ es una estimación puntual confiable de μ, pero probablemente no esté exactamente sobre la μ. En lugar de esta idea, se puede especificar con una alta probabilidad: piense en un 0.95, que un rango en particular cubre la verdadera media.

Solución

A partir de los datos de la muestra con una desviación estándar de 279.45, así se tiene que:

$$Z\sigma_{\overline{X}} = -1.96\frac{279.45}{\sqrt{30}} \simeq -100, Z(1-\alpha/2)\sigma_{\overline{X}} = 1.96\frac{279.45}{\sqrt{30}} \simeq 100$$

de este modo (6200-100,6200+100), y se dice que el intervalo de 6100 a 6300 cubre la media μ con una probabilidad de 0.95.

Esto es un ejemplo de un *intervalo de confianza*. De manera intuitiva se repasan los elementos que integran a éste. Dicho intervalo comprende dos límites: uno inferior *LI*, izquierda, y otro superior *LS*, derecha. En el ejemplo citado, $LI = 6100$ y $LS = 6300$. Por otro lado, el intervalo de confianza tiene un valor de probabilidad, el cual suele conocerse como *nivel de confianza* y se denota por $1-\alpha$. Para el ejemplo será: $\alpha = 0.05$ y $1-\alpha = 0.95$. Expresado en términos de porcentaje se dice que hay un intervalo de 95 % de confianza. En general, un intervalo de confianza para la media poblacional se presenta en una proposición de probabilidad como se muestra en la siguiente fórmula:

$$P(LI \leq \mu \leq LS) = 1-\alpha$$

Se dice que el $(1-\alpha)\,\%$ de confianza la media se encuentra entre los límites LI o $LI = \overline{x}_i$ y LS o $LS = \overline{x}_d$, donde

$$LI = \overline{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}, \text{ y } LS = \overline{x} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

Intervalos de confianza para la media en muestras grandes

Intervalo de confianza cuando la población es normalmente distribuida y la desviación estándar es conocida. Con el fin de ilustrar el procedimiento para la estimación de la media μ mediante un intervalo de confianza, se propone que la muestra sea seleccionada de una población cuya distribución es normal y bajo el supuesto de que se conoce la desviación estándar. No obstante, en la práctica el valor de σ no se conoce, aunque más adelante se verá cómo obtener un intervalo de confianza sin estos supuestos.

Intervalo de confianza

El intervalo de confianza o estimación por intervalo es un rango de valores con una probabilidad asociada o un nivel de confianza. La probabilidad cuantifica la verosimilitud de que el intervalo contenga el parámetro poblacional.

Detalles técnicos: Por lo tanto, como ya se ha indicado, la discrepancia entre las medias μ y $\overline{X}$ da lugar a un margen de error conocido como error muestral, $\mu - \overline{X} = -ei$, o $\mu - \overline{X} = ed$. En realidad la media μ no se conoce, y en el proceso de selección de muestra $\overline{x}$ varía de muestra a muestra. Así, en el caso que se ha planteado, el valor del error e dependerá del nivel de significancia que se desee y de la distribución muestral. De esa manera se tendrá el máximo valor del error $ei = z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ que corresponderá a la distancia que se tiene a la izquierda, y $ed = z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$ corresponderá a la distancia que se tiene a la derecha de la media muestral. Finalmente se tienen los valores de los límites inferior LI y superior LS del intervalo de confianza:

$$LI = \overline{X} + ei, \qquad LS = \overline{X} + ed$$

Valor del error

Dado un nivel de significancia $(1-\alpha)$, el valor del error e corresponde a la mayor distancia posible que hay entre los valores del parámetro que se estima y la estimación puntual. En general, para cubrir $(1-\alpha)\,\%$ de los valores de una distribución normal se tiene que:

$$P(z(\alpha/2) \leq Z \leq z(1-\alpha/2)) = 1-\alpha$$

Considere la distribución muestral de $\overline{X}$, escrita como una normal estándar se tiene que:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Se sustituye ésta en la expresión anterior y se simplifica; entonces:

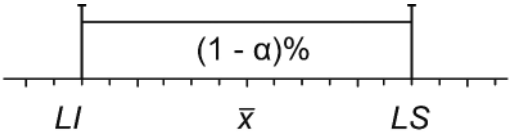
$$P(\overline{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Reuniendo esta información, se tiene que el error ei y ed distancias hacia ambos lados de la media $\overline{x}$, es:

$$ei = z(\alpha/2)\frac{\sigma}{\sqrt{n}} \text{ o } ed = z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$

Por consiguiente, la figura 7.20 describe esta información, los límites inferior y superior son:

$$LI = \overline{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}} \qquad LS = \overline{x} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$$



$$LI = \bar{x} + z(\alpha/2)\frac{\sigma}{\sqrt{n}} \qquad LS = \bar{x} + z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}}$$

Figura 7.20 Esquema general del intervalo de $(1-\alpha)\,\%$ confianza para la media, en que se indican los límites inferior y superior.

En general, se señala que con un $(1-\alpha)\,\%$ de confianza la media μ pertenece al intervalo (LI, LS); es decir $\mu \in (LI, LS)$.

Interpretación del intervalo de confianza

La interpretación del intervalo de confianza se relaciona con la oportunidad de que un intervalo contenga al parámetro, esto es, si toman 100 muestras diferentes de la misma población, por ejemplo para el caso de la media, se obtendrán 100 medias muestrales diferentes y, por consiguiente, 100 intervalos diferentes. Si cada uno de estos intervalos está dentro de 90 % de confianza, entonces teóricamente se esperaría que aproximadamente el 90 de los 100 contendrán a la media μ y 10 no.

Para motivar una idea de esta situación, en el siguiente problema se plantea un entorno que considera una realidad, con los datos de la población, se ha construido un escenario donde se han reproducido diferentes muestras con tamaño $n = 5$ y $n = 10$. La finalidad de esta simulación es que observe que al obtener diferentes muestras para cada uno de los tamaños, los intervalos de confianza contienen en su mayoría a la media.

El mundo de la información 2: Apoyo económico

Para una población de 100 estudiantes de un centro educativo, la administración del sector escolar desea conocer el gasto de transporte a la semana de estos alumnos con la intención de gestionar una mejor beca ante el municipio. A partir de la información generada de esta población, la media es $\mu = 110$ y la desviación estándar correspondiente es $\sigma = 21$.

Preguntas sobre la naturaleza del problema

¿Cómo puede el administrador usar esta información para gestionar el aumento de la beca para todos los estudiantes? ¿Cuál es la precisión de la información si aumenta el número de estudiantes que debe estar en la encuesta? ¿Qué información proporciona la media en cada muestra para tener una idea clara del gasto en transporte de los alumnos?

Se utilizó el siguiente procedimiento:

1. Se seleccionaron 20 muestras diferentes de tamaño $n = 5$ para los 100 datos de la población. En cada caso se calculó la media de la muestra con sus límites respectivos. En la figura 7.21, se describe una gráfica cuya línea central representa a la media de la población (parámetro) y en la que se

puede ver cuántos intervalos del 95 % de confianza contienen a la media, así como los que no la contienen.

2. Se repite el proceso del inciso anterior pero ahora la muestra tiene un tamaño $n = 10$. Nuevamente se observa la media de la población, y los intervalos que la contienen.

Interpretación: observe que por cuestiones de azar, en el primer caso sólo un intervalo de los 20 no contiene a la media, es decir 95 % sí contiene a la media. En el segundo caso, 16 intervalos contienen a la media, lo que equivale a un porcentaje de 80 %. También se puede notar que el ancho de los intervalos es más pequeño en la gráfica de la derecha, ya que $n = 10$.

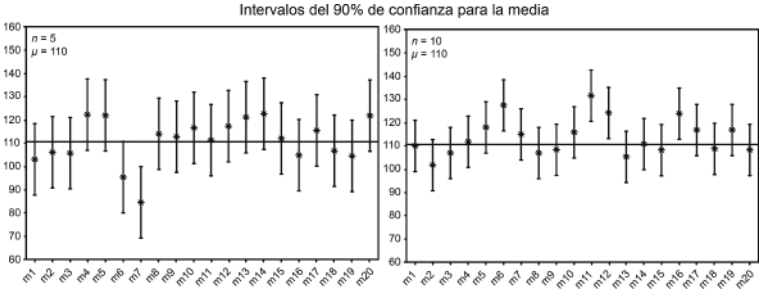


Figura 7.21 Se han construido 20 intervalos de confianza, para la primer gráfica con tamaño de muestra $n = 5$, y para la otra con $n = 10$.

Ejemplo 7.8

Con el fin de establecer lineamientos nutritivos para los comedores de un centro universitario, la administración de la universidad desea estimar el índice de masa corporal (IMC) para evaluar el estado de salud de su población. Se toma una muestra aleatoria de tamaño $n = 30$ con media $\overline{x} = 26.8$. Con referencia a estudios realizados con anterioridad, se sabe que la desviación estándar de la población es $\sigma = 2.8$. Se busca, entonces, encontrar un intervalo de confianza de 95 % para la media del IMC.

Solución

En resumen, la información proporcionada indica que $\overline{X} = 26.8$, $n = 30$ y $\sigma = 2.8$. Con el fin de estimar el intervalo se realizan los cálculos correspondientes para establecer la distancia de la media a los límites del intervalo y luego estimar los límites. Entonces, la desviación estándar de $\overline{X}$,

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2.8}{\sqrt{30}} = 0.51 \quad \text{por lo tanto} \quad e = 1.96\frac{\sigma}{\sqrt{30}} = (1.96)(0.51) = 1$$

por lo tanto los límites del intervalo de confianza son:

$$LI = \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} = 26.8 - 1 = 25.8, LS = \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} = 26.8 + 1 = 27.8$$

Ahora bien, se puede concluir, con 95 % de confianza, que la verdadera media μ del IMC está entre 25.8 y 27.8. En términos prácticos se tiene que, en promedio, las personas de la población considerada presentan sobrepeso, ya que cuando el IMC es superior a 25 hay sobrepeso.

Procedimiento para el cálculo de un intervalo considerando otros niveles de confianza: El proceso para deducir el valor correcto de Z en cualquier nivel de confianza es el que aparece a continuación.

1. Tomar el valor de α y dividirlo entre 2.
2. Mirar el valor de Z en la tabla de valores de una normal, ver en la tabla del capítulo 6 o en el ápendice (encontrar el más cercano a éste). Como alternativa para encontrar el valor de la variable Z, vea el material didáctico en **CalEst** (figura 7.22).
3. Leer el valor de Z correspondiente para obtener el valor de $z(\alpha/2)$ y así $z(1-\alpha/2)$.

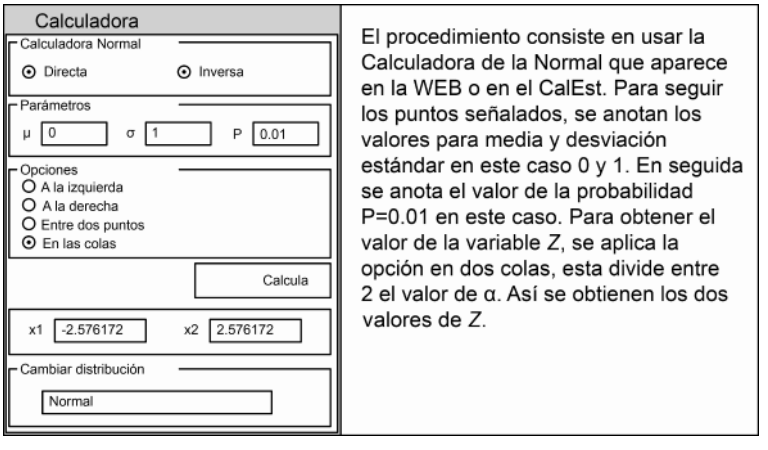


Figura 7.22 Procedimiento para el cálculo del intervalo de confianza, mediante **CalEst**.

Ejemplo 7.9

Con el objetivo de ilustrar el procedimiento, se recurre al contexto del ejemplo 7.8, donde $\overline{X} = 26.8$, $n = 30$ y $\sigma = 2.8$.. Calcule un intervalo de 99 % de confianza para la media.

Solución

En este caso, $1-\alpha = 0.99$, así $\alpha = 0.01$; por lo tanto, $\alpha/2 = 0.005$. Buscando en la tabla de una normal el valor de Z, se tiene que los valores correspondientes son: $z(\alpha/2) = -2.58$ y $z(1-\alpha/2) = 2.58$

Una alternativa para establecer el valor de la variable Z es utilizar la tabla del capítulo 6 o al final del libro, así como la calculadora que viene en el CalEst. Su aplicación se ha expuesto en el capítulo de distribuciones de probabilidad normal. Usando la calculadora, primero se da el valor de los parámetros de la normal estándar $\mu = 0$, y $\sigma = 1$, a continuación el valor de la probabilidad, para este ejemplo $p = 0.01$. El primer paso consiste dividir α entre 2, al considerar la opción en dos colas se hace esa operación. La salida indica los valores de la variable Z.

Los límites del intervalo quedarían como sigue:

$$\text{Límite inferior:} \qquad LI = \overline{X} + ei = \overline{X} - 2.58\frac{\sigma}{\sqrt{n}} = 26.8 - 1.32 = 25.48$$

$$\text{Límite superior:} \qquad LS = \overline{X} + ed = \overline{X} + 2.58\frac{\sigma}{\sqrt{n}} = 26.8 + 1.32 = 28.12$$

Con esta información se construye el intervalo, así μ está contenido en (25.48, 28.12) con un 99 % de confianza.

Ejemplo 7.10

En el contexto de la economía familiar, los profesores de una universidad tienen que comer fuera de su casa durante la semana. Se escogió un día de la semana y se tomó una muestra aleatoria, de profesores que usan el comedor, de tamaño $n = 20$, y se anotó el consumo en pesos. Un estudio previo por parte del administrador de la concesión, se sabe que la distribución de la variable consumo tiene una distribución aproximadamente simétrica y con varianza $\sigma^2 = 72.25$. Los datos se indican abajo:

48	43	54	49	51	47	41	24	52	46
63	51	53	59	44	55	39	62	49	62

Construir un intervalo de confianza de 95 % para μ.

Solución 1

Con los datos proporcionados por la muestra se calcula la media, es decir: $\overline{x} = 49.6$, a continuación se obtiene el valor de e en función del nivel de significancia, que para el ejemplo $(1-\alpha) = (1-0.05) = 0.95$ y así obtener los límites inferior y superior del intervalo de confianza. Operaciones: $ei = z(\alpha/2)\frac{\sigma}{\sqrt{n}} = -1.96\frac{8.5}{\sqrt{20}} = -3.727$, $ed = z(1-\alpha/2)\frac{\sigma}{\sqrt{n}} = 1.96\frac{8.5}{\sqrt{20}} = 3.727$. Entonces los límites son:

$$\text{Límite inferior:} \qquad LI = \overline{X} + ei = \overline{X} - 1.96\frac{\sigma}{\sqrt{n}} = 49.6 - 3.725 = 45.873$$

$$\text{Límite superior:} \qquad LS = \overline{X} + ed = \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} = 49.6 + 3.725 = 53.327$$

Con esta información se construye el intervalo, así μ está contenido en $(45.87, 53.33)$ con un $95\,\%$ de confianza.

Solución: utilizando el CalEst

Recurrimos al avance tecnológico de tal manera que proporcione los resultados que se han indicado en la solución operativa. Mediante el paquete estadístico que se ha venido empleando se generan las soluciones que se esperan. Como se sabe, inicialmente en el Calculador Estadístico se capturan los datos en una hoja y luego se marca la opción Inferencia, a continuación se escoge la alternativa intervalos de confianza.

En la figura 7.23, se muestra a la izquierda la hoja que aparece al aplicar la opción, se indica el Nivel de confianza, así como si se conoce o no la Desviación estándar, y finalmente se señala la columna en la que se tienen los datos, en este caso Consumo. Enseguida se tienen los cálculos y por consiguiente el intervalo de confianza, como se ve, la solución es idéntica a la clásica, sólo que ahora con apoyo de la calculadora. En el próximo apartado se explica la opción múltiples niveles que aparecen en la figura 7.23.

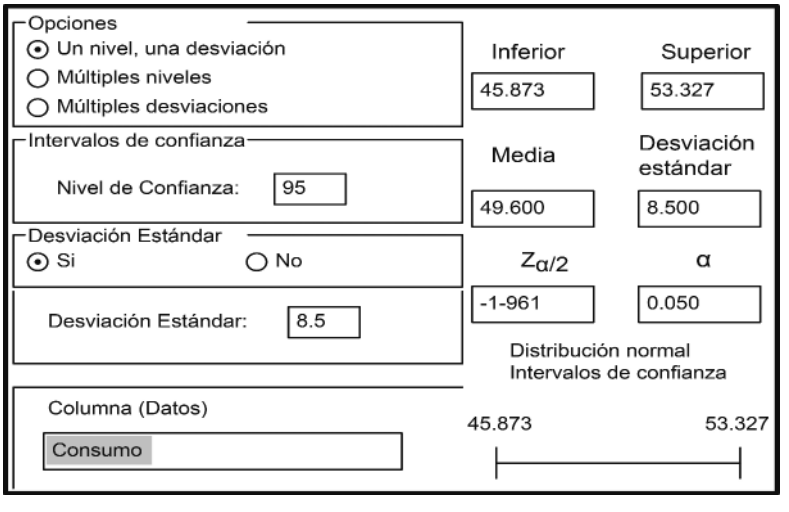


Figura 7.23 Intervalo de confianza para la media, con varianza conocida.

Observaciones generales sobre los intervalos de confianza para μ. Como se puede observar, el intervalo de confianza del ejemplo 7.8 es menos amplio que el del ejemplo 7.9, pese a que en ambos casos

el tamaño de la muestra es el mismo. Mediante el siguiente ejemplo se muestra dicha situación para tres niveles de confianza.

Ejemplo 7.11

Volviendo a los datos del ejemplo 7.10, estime el intervalo de confianza para la media μ considerando tres niveles de confianza a la vez, es decir 90 %, 95 % y 99 % a la vez. En la figura 7.24 se exponen las tres situaciones.

Solución

Claramente se ve la diferencia en la longitud de los intervalos. Los resultados del ejemplo 7.10 indican que al aumentar el nivel de confianza, el margen de error ($ei = z(\alpha/2)\frac{\sigma}{\sqrt{n}}, ed = z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$) aumenta, pero disminuye la precisión en la estimación del parámetro. Por esa razón, es recomendable que se especifique el valor deseado y la precisión (el margen de error) requeridos, antes de iniciar un estudio. A partir de estos datos se podrá determinar el tamaño de muestra.

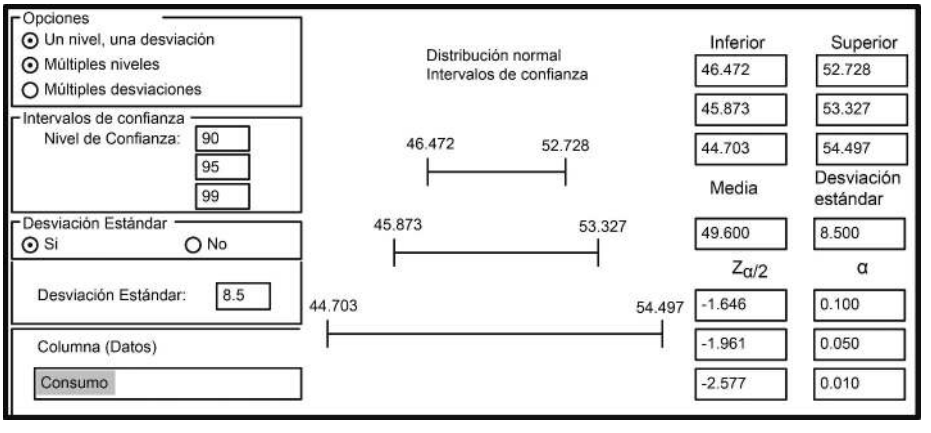


Figura 7.24 Estimación del intervalo de confianza para tres niveles diferentes.

Margen de error permisible

$ei = z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ entonces $ei\sqrt{n} = z(\alpha/2)\sigma$

$\sqrt{n} = \frac{z(\alpha/2)\sigma}{ei}$ finalmente $n = \left(\frac{z(\alpha/2)\sigma}{ei}\right)^2$

Determinar el tamaño de muestra n : El tamaño de muestra se determina despejando n de la fórmula del margen de error, esto es:

$$n = \left(\frac{z(\alpha/2)\sigma}{ei} \right)^2$$

Como se sabe, $z(\alpha/2)$ es el valor de la distribución normal que corresponde al nivel de confianza deseado, σ es la desviación estándar y ei el margen de error permisible ($ei = -ed$).

Ejemplo 7.12

La administración de un banco lleva el registro mensual del pago mínimo que deben realizar a las tarjetas de crédito un grupo determinado de clientes, establecido por el reporte de ingresos mensuales. Un estudio previo reportó una desviación estándar $\sigma = 120$ pesos. ¿Cuántas muestras debe tomar la administración para tener una certeza del 90 % de que el error en estimación no exceda de 8 pesos?

Solución

La información de la que se dispone es $\sigma = 120$, $1 - \alpha = 0.90$ y $\alpha/2 = 0.05$. El punto en el límite superior a 0.05 de una distribución normal estándar es $Z(0.05) = 1.645$, y el error (ei) tolerable es $ei = 8$. Con esta información se puede aplicar la fórmula para calcular el tamaño de muestra n, así

$$n = \left(\frac{z(\alpha/2)\sigma}{ei} \right)^2 = \left(\frac{(1.645)(120)}{8} \right)^2 = 24.68 \cong 25$$

Por lo tanto, el tamaño de muestra es $n = 25$.

Intervalo de confianza cuando la población no es normal y la desviación estándar es desconocida

Hemos visto que para una media calculada de una muestra aleatoria, la variabilidad de muestra a muestra en la media muestral está dada por la variabilidad de observaciones individuales dividida entre la raíz cuadrada del tamaño de la muestra, esto es, $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. No podemos usar esta expresión en muchas aplicaciones prácticas, pues ésta depende de la desviación estándar de la población σ, la cual es desconocida. Una solución alterna es usar la desviación estándar S de la muestra y remplazar el verdadero valor de σ desconocido por el de S. Ello permite que haya un estimador de $\sigma_{\overline{x}}$, al cual se le conoce como *error estándar*.

$$\text{El error estándar de la media muestral} = \frac{\text{desviación estándar muestral}}{\sqrt{\text{tamaño de la muestra}}} = \frac{S}{\sqrt{n}}$$

Una versión más amplia del teorema de límite central dice que no sólo para tamaños de muestras

suficientemente grandes es $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ aproximadamente una normal estándar, sino también $(\overline{X}-\mu)/(S/\sqrt{n})$ si se distribuye aproximadamente como una normal estándar.

Ejemplo 7.13

Para prevenir los gastos que repercuten en el ingreso familiar debido a malos hábitos en la alimentación, la dirección de una escuela tiene la norma de realizar un examen médico a los alumnos de recién ingreso. Entre los resultados del examen se encuentra el nivel de glucosa en sangre. Para tener una idea del promedio del nivel de glucosa se toma una muestra de 50 análisis efectuados a los estudiantes. Los valores fueron:

78	89	74	87	80	81	84	83	86	98	72	66	68	73	82
89	96	87	81	77	87	83	112	77	79	116	74	88	90	75
95	81	77	92	74	83	94	89	75	77	102	81	88	99	75
78	93	78	107	100										

A partir de estos números encontrar un intervalo de 90 % de confianza.

Solución

Esta muestra nos arroja los siguientes resultados: $n=50$, $\overline{X}=85$, $S=10.831$, $S/\sqrt{n}=1.531$. Aplicando los cuatro pasos para obtener los intervalos de confianza según el nivel deseado, se observa que $(1-\alpha)=0.90$, entonces $\alpha=0.10$.

En resumen, $\alpha/2=0.10/2=0.05$, $Z(\alpha/2)=-1.645$, $Z(1-\alpha/2)=1.645$. Con esta información podemos obtener los límites inferior y superior para el intervalo de confianza de 90 %.

$$\text{Límite inferior}: \qquad LI=\overline{X}-1.645\frac{S}{\sqrt{n}}=85-2.519=82.481$$

$$\text{Límite superior}: LS=\overline{X}+1.645\frac{S}{\sqrt{n}}=85+2.159=87.519$$

Es decir, el intervalo para la media μ de 90 % de confianza es:

$$82.481\leq\mu\leq 87.519$$

En conclusión, se dice que la población escolar tendrá en promedio un nivel de glucosa que está dentro de los estándares adecuados para la salud, ya que las asociaciones médicas recomiendan como meta específica un nivel de glucosa de 80 a 120 mg/dl antes de alimentos.

Intervalos de confianza para la media en muestras pequeñas: $t - Student$

Intervalo de confianza usando la $t - Student$. En las secciones anteriores se ha calculado el intervalo de confianza para estimar el parámetro μ, para ello se ha supuesto que la distribución de la variable aleatoria X es normal, y que la varianza σ^2 es conocida, y en el caso de que no sea así, entonces se propone tomar una muestra de tamaño $n \geq 30$. Tales supuestos han permitido construir la metodología para la estimación de un parámetro y estas ideas se extienden a muchos otros estudios e investigaciones que un administrador o un economista tenga interés en llevar a cabo. En problemas reales difícilmente se conocerá σ^2, y en la práctica tomar muestras de tamaño grande consume tiempo y es costoso. Tal situación se remedia considerando la distribución t-Student, explicada en el capítulo 6, en lugar de la distribución normal estándar. En resumen, considere que la variable aleatoria X tiene una distribución de probabilidad aproximadamente normal, entonces:

$$t = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

tiene una distribución $t - Student$ con $n - 1$ grados de libertad. En forma abreviada, distribución t y $gl = n - 1$.

Dado que el interés es estimar por intervalo de confianza el parámetro μ ahora usando la distribución t, en la tabla 7.8 se muestra un extracto de ésta y en el ejemplo 7.14 se ilustra su aplicación. Ahí se puede obtener el valor para t yendo a las probabilidades que se muestran en la tabla 7.8. También se puede obtener el valor de t partiendo de una probabilidad.

Ejemplo 7.14

Un valor de $\alpha = 0.10$ permitiría obtener los valores en la distribución t que se generan en el proceso de estimación de un intervalo del $(1 - \alpha)\,\% = 90\,\%$ de confianza. Obtenga estos valores usando la tabla 7.8, primero determine el valor superior a la cola derecha de la distribución $t(0.05)$ de la $t - Student$ con 7 grados de libertad con $\alpha/2 = 0.05$. Luego encuentre el valor inferior a la cola izquierda de la distribución $t(0.05)$ con $\alpha/2 = 0.05$.

Solución

Con $gl = 7$, el valor superior de la distribución t se encuentra en el renglón 4 y la columna 4 de la tabla 7.8 y éste es 1.895, es decir $P(t \geq t(7, 0.05) = 1.895) = 0.05$. De manera análoga, y puesto que la curva está centrada en 0, el valor para la cola izquierda es -1.895, o sea, $P(t \leq t(7, 0.05) = -1.895) = 0.05$. Integrando estos dos valores, en la figura 7.25 se muestra una distribución t con 7 gl, con un nivel de significancia del 90 %, es decir:

$$P(-1.895 \leq t \leq 1.895) = 0.90$$

Tabla 7.8 Puntos porcentuales de la distribución $t - Student$
Áreas de la cola superior α; $t(\alpha)$

	α	0.25	0.1	0.05	0.025	0.01	0.005
gl							
5		0.727	1.476	2.015	2.571	3.365	4.032
7		0.711	1.415	1.895	2.365	2.998	3.499
15		0.691	1.341	1.753	2.131	2.602	2.947
22		0.686	1.321	1.717	2.074	2.508	2.819
23		0.685	1.319	1.714	2.069	2.500	2.807
24		0.685	1.318	1.711	2.064	2.492	2.797
25		0.684	1.316	1.708	2.060	2.485	2.787

Solución mediante el uso de CalEst

En el apartado de distribuciones en el calculador estadístico, CalEst, aparece la distribución de probabilidad $t-Student$, y al utilizarla se pueden generar diferentes distribuciones de la t dependiendo de los grados de libertad, en particular para este ejemplo. La solución se describe en la figura 7.25 donde se muestra una distribución t con 7 gl, con un nivel de significancia del 90 %. A la derecha de la gráfica se observa el procedimiento para obtener los valores de la distribución t usando el calculador de probabilidad.

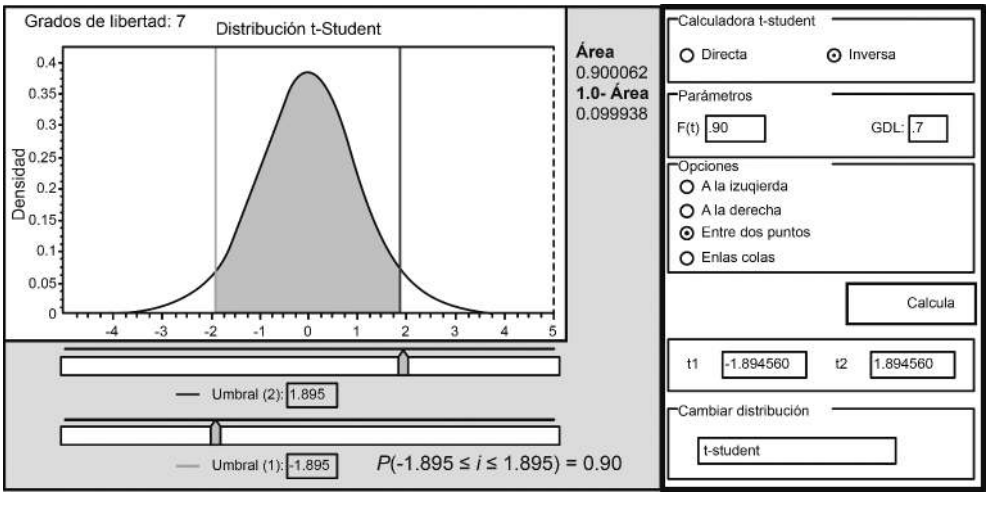


Figura 7.25 Valores de t entre los valores que generan una probabilidad de 0.90 en la distribución $t - Student$.

Ejemplo 7.15

En el desarrollo de nuevos productos o en el buen funcionamiento de los proceso en la industria, surgen problemas económico administrativos. En este caso una empresa que genera un producto químico requiere que la media del nivel de pH en el agua debe estar en 6.8. Si esto no ocurre, se tiene que invertir para modificar el proceso. Para ello se construye un intervalo y luego se observa si 6.8 está contenido en el intervalo. El técnico responsable de este proceso toma 19 muestras de agua y mide el pH de cada una de ellas. Los datos son:

6.7,	7.1,	6.8,	6.9,	6.5,	6.7,	6.6,	6.5,	6.5,	6.2
6.3,	6.6,	7.0,	6.7,	6.9,	6.5,	6.6,	6.9,	6.9	

Solución

La media y desviación estándar para los 19 datos son: $\overline{x} = 6.679$ y $s = 0.237$, respectivamente. El valor de t en la distribución t de Student para el 95 % de confianza implica tomar los valores de t correspondientes a las probabilidades menor a 0.025 y mayor a 0.975, es decir, $t_i = t(18, 0.025) = -2.1004$ o $t_d = t(18, 0.975) = 2.1004$.

Intervalo de confianza: Se completa el análisis estadístico estimando el intervalo del 95 % de confianza para la media.

$$\left(\overline{x} + t(n-1, \alpha/2)\frac{S}{\sqrt{n}}, \overline{x} + t(n-1, 1-\alpha/2)\frac{S}{\sqrt{n}}\right)$$

$$(6.679 - 2.1004(0.0544), 6.679 + 2.1004(0.0544) = (6.565, 6.793)$$

En este caso el intervalo de confianza no contiene al 6.8 y eso indica que la empresa debe invertir para mejorar el proceso.

Intervalos de confianza bajo diferentes supuestos

Una vez que se ha planteado la estimación por intervalo de confianza, se observa que existen diferentes condiciones en el proceso de estimación. Como se ha señalado, dependen del conocimiento sobre la distribución de probabilidad de la variable aleatoria original y la varianza. En ese sentido, se presenta un esquema, figura 7.26 y un cuadro, tabla 7.9, que resumen la descripción de los intervalos de confianza que se han usado para estimar el parámetro μ, considerando diferentes supuestos.

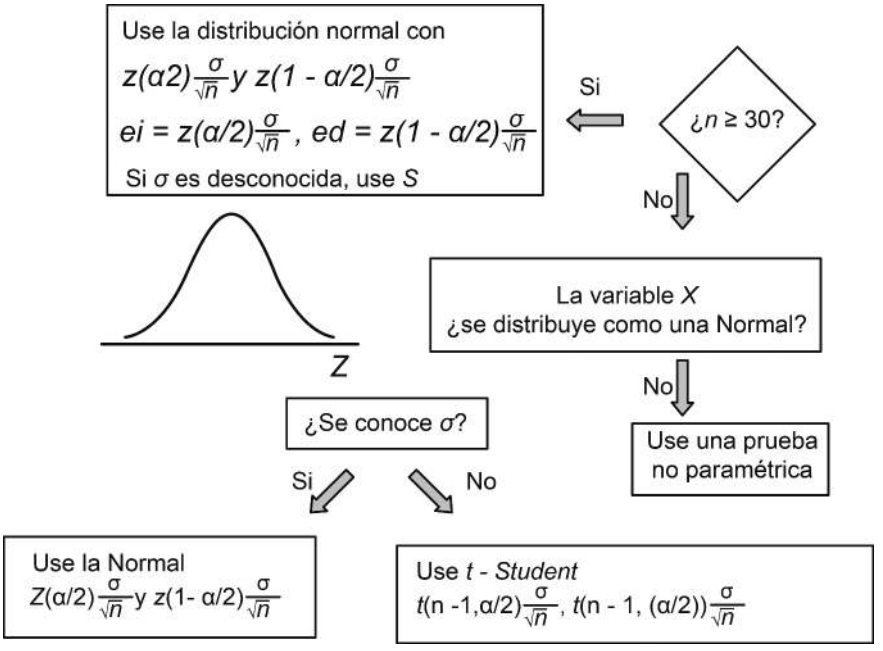


Figura 7.26 Esquema que resume la estimación por intervalo de confianza para el parámetro μ.

Tabla 7.9 Intervalos de confianza para μ.

Población	Desviación estándar	Tamaño de muestra	Intervalo de confianza para μ
Normal	Conocida	$n \leq 1$	$LI = \overline{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ $LS = \overline{X} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$
Normal	Desconocida	$n > 30$	$LI = \overline{X} + z(\alpha/2)\frac{S}{\sqrt{n}}$ $LS = \overline{X} + z(1-\alpha/2)\frac{S}{\sqrt{n}}$
Normal	Desconocida	$n \leq 30$	$LI = \overline{X} + t(gl = n-1, \alpha/2)\frac{S}{\sqrt{n}}$ $LS = \overline{X} + t(gl = n-1, (1-\alpha/2))\frac{S}{\sqrt{n}}$
No Normal	Conocida	$n \geq 30$	$LI = \overline{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}$ $LS = \overline{X} + z(1-\alpha/2)\frac{\sigma}{\sqrt{n}}$

Intervalos de Confianza para una proporción p

Como se ha mencionado anteriormente, la proporción es otro parámetro importante en diferentes problemas de la vida real. Existen situaciones prácticas en administración o economía social, familiar o empresarial en que se utilizan proporciones; por ejemplo: el porcentaje de personas que no pagan impuestos, el porcentaje de obreros a quienes no se les paga horas extras de trabajo, porcentaje de hogares que no cuentan con agua potable, o calles no pavimentadas, porcentaje de alumnos que abandonan la secundaria, etcétera.

Observación: por lo general los parámetros se simbolizan con letras griegas; en particular aquí se ha empleado la letra p para referirse al parámetro para la proporción, porque es frecuente representarlo de esa forma. Algunos escritores utilizan la letra griega π para señalar una proporción.

Ejemplo 7.16

Para estrategias de ventas, la administración de una corporación tiene interés en conocer el porcentaje de clientes que pagan en efectivo y aquellos que utilizan crédito. En este planteamiento la respuesta que se tiene es binaria. Describir la proporción de los que pagan en efectivo.

Solución

Se ha visto que a partir de la variable aleatoria X se puede construir la proporción:

$$X = \begin{cases} 1 & \text{si el cliente paga en efectivo} \\ 0 & \text{si el ciente paga a crédito} \end{cases}$$

Entonces una proporción se define como:

$$p = \frac{X}{N}$$

donde X son los clientes que pagan en efectivo de un total de N clientes en la población.

Estimador de la proporción. Un estimador para p está dado por $\widehat{p}$. Éste último se calcula a partir de los x elementos de la población que reúnen la característica: clientes que pagan en efectivo tomados de una muestra de tamaño n seleccionada de la población, esto es:

$$\widehat{p} = \frac{x}{n}$$

Así, $\widehat{p}$ es la *proporción muestral.* El error estándar del estimador $\widehat{p}$ es:

$$\sqrt{\frac{p(1-p)}{n}}$$

La proporción en una población se obtiene mediante $p = \frac{X}{N}$. La variable aleatoria X sigue una distribución binomial con parámetros (n, p). *Se requieren de las condiciones* $np \geq 5$ y $n(1-p) \geq 5$ *para una distribución binomial se aproxime a una normal*. Entonces la distribución muestral para $\widehat{p}$ es aproximadamente normal con parámetros:

$$\mu_{\widehat{p}} = \mu(\widehat{p}) = p \quad \text{y} \quad \sigma_{\widehat{p}} = \sigma(\widehat{p}) = \sqrt{p(1-p)/n}$$

Observación: Para calcular el error estándar de $\widehat{p}$, debemos conocer el valor de p. Sin embargo, ése es precisamente el valor que se intenta estimar. Dado que lo desconocemos, tiene sentido usar $\widehat{p}$ como un estimador de p en la fórmula para el error estándar.

Procedimiento para encontrar el intervalo de confianza de *p*

Con la experiencia adquirida en el cálculo del intervalo de confianza para el parámetro μ recuerde que se requiere calcular $e = z(\alpha/2)$(error estándar). A partir de ahí desarrollar las fórmulas que permiten encontrar los límites inferior y superior de un intervalo del $100(1-\alpha)\,\%$ de confianza para p, es necesario conocer la distribución muestral de $\widehat{p}$.

El cálculo de e es:

$$ei = Z(\alpha/2)\sqrt{\frac{p(1-p)}{n}} \cong Z(\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$ed = Z(1-\alpha/2)\sqrt{\frac{p(1-p)}{n}} \cong Z(1-\alpha/2)\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

Por lo tanto los límites del intervalo se expresan por:

Límite inferior	$LI = \widehat{p} - ei$
Límite superior	$LS = \widehat{p} + ed$

Ejemplo 7.17

Entre varias pruebas que realiza una empresa para contratar a su personal está la memoria a corto plazo. La administración contrata a un sicólogo para que realice las pruebas de selección a una muestra de 200 personas; para ello le ayudan varios de sus asistentes. En particular, esta prueba consiste en mostrar una tarjeta con 16 palabras a cada una de las personas por 30 segundos, a continuación se les distrae por un minuto platicando con los entrevistados. Finalmente se le pide a la persona que diga las palabras que recuerda, para ello se da un minuto. El investigador plantea que 23 % de las personas recuerdan 8 o

más palabras, construya un intervalo del 95 % de confianza para evaluar esta tesis. La información que recogieron de las $n = 200$ entrevistas es que $\widehat{p} = 0.27$ recuerdan 8 o más palabras.

Solución

La parte operativa inicia verificando las condiciones que garanticen usar una distribución normal, es decir: $200(0.23) = 46 \geq 5$ y $200(0.77) = 154 \geq 5$.

Los valores correspondientes para la variable normal estándar son $Z(\alpha/2) = -1.96$ y $Z(1-\alpha/2) = 1.96$. El intervalo de confianza del 95 % para este ejemplo se calcula de acuerdo con el valor de Z.

$$\left(\widehat{p} + Z(\alpha/2)\sqrt{\widehat{p}(1-\widehat{p}/n)}, \widehat{p} + Z(1-\alpha/2)\sqrt{\widehat{p}(1-\widehat{p})/n)}\right)$$

$$\left(0.27 - 1.96(\sqrt{0.27(0.73)/200)}, 0.27 + 1.96(\sqrt{0.27(0.73/200)}\right) = (0.208, 0.332)$$

Lo que indica que aproximadamente entre el 21 % y 33 % recuerdan 8 o más palabras.

Inferencia sobre la varianza σ^2 y σ

La varianza en diferentes procesos tanto administrativos como económicos debe ser conocida, ya que cuando ocurren situaciones con una variabilidad grande, las conclusiones sobre éstas crean incertidumbre. Por ejemplo, una compañía produce bolsas de cacahuate. El proceso genera miles de bolsas y cada una debe tener el mismo peso. Sin embargo, existe una variación en el peso de cada bolsa, la cual debe ser baja, si no fuera así la compañía se enfrenta a pérdidas económicas, dado que una alta varianza ocasiona bolsas con pesos bajos, lo cual puede repercutir en una multa y desprestigio, así como dar cacahuates de más. Si la población de pesos tiene una distribución normal, en este caso se desea hacer inferencia estadística sobre la varianza o desviación estándar.

Estimador puntual

El estimador puntual para σ^2 es S^2 y el estimador puntual para σ es S, además S^2 es un estimador insesgado para σ^2.

En situaciones similares que generan proyectos administrativos e influyen en la economía de una empresa, una excesiva variación en las dimensiones de un producto contribuye a tener una calidad pobre, ya que la expectativa del cliente es tener un producto uniforme, por ello es importante minimizar la variabilidad. Otras variables económicas que generan una importante inseguridad son los precios de algunos produc-

tos, como el precio de alguna medicina la cual tiene una varianza grande, ya que el importe cambia dependiendo de los establecimientos donde se compre. Situaciones de desigualdad social son generadas por las diferencias de salarios o grados de educación; en tales casos la varianza es grande.

En resumen, la varianza desempeña un papel importante para explicar el desempeño de los procesos o fenómenos, ya que una dispersión grande de los datos en torno a la media genera una gran variabilidad de las características de estudio. Finalmente, un aspecto relevante en muchas situaciones es hacer inferencia estadística sobre la varianza o la desviación estándar de un proceso con la finalidad de poder reducir ésta.

Nota: Esta distribución se planteó en el capítulo 6, sin embargo en la siguiente sección se toman las características relevantes de la distribución para construir los intervalos de confianza.

Intervalos de confianza para σ^2

Es necesario usar la distribución conocida como Ji cuadrada (Chi cuadrada) χ^2 en la construcción de intervalos de confianza para la varianza y la desviación estándar. Así:

Si una variable aleatoria X tiene una distribución normal, entonces la distribución:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

es una distribución Ji cuadrada para muestras de tamaño $n > 1$. Esta distribución tiene cuatro propiedades:

1. Todo los valores χ^2 son mayores o iguales a cero.
2. La distribución Ji cuadrada es una familia de curvas, cada una determinada por los grados de libertad.
3. El área bajo la curva de una distribución Ji cuadrada es igual a 1.
4. La distribución Ji cuadrada es sesgada positivamente.

Caracterización de la distribución probabilidad χ^2

Si $Z_1, Z_2, ...Z_n$ son variables normales estándar independientes, entonces se dice que la variable χ está definida por:

$$\chi = Z^2 + Z_2^2 + ... + Z_n^2$$

y tiene una distribución Ji (Chi) cuadrada con n grados de libertad. Se denota $\chi \sim \chi_n^2$ para indicar que χ tiene una distribución Ji cuadrada con n grados de libertad. Véase capítulo 6.

Ideas para la estimación por intervalo usando la χ^2

Para llevar a cabo esta inferencia, es necesario determinar los valores críticos en la distribución χ^2 con $n-1$ grados de libertad.

En esta dirección se especifica un valor α tal que $(0 < \alpha < 1)$ y se calcula alguna de las siguientes probabilidades: a la derecha $P(\chi^2 > \chi^2(n-1,\alpha)) = \alpha$, a la izquierda $P(\chi^2(n-1,1-\alpha) < \chi^2) = 1-\alpha$, similarmente para $\alpha/2$

$$P(\chi^2(n-1,\alpha/2) > \chi^2) = \alpha/2, \quad P(\chi^2(n-1,1-\alpha/2) < \chi^2) = 1-\alpha/2$$

Solución operativa clásica

Para calcular estas probabilidades asociadas a la construcción de intervalos de confianza de la χ^2, véase la tabla 7.10.

Tabla 7.10 Puntos porcentuales de la distribución X^2
Áreas de la cola superior α; $X^2(\alpha)$

	α	0.99	0.975	0.95	0.90	0.10	0.05	0.025
gl								
1		——	0.001	0.004	0.016	2.706	3.841	5.024
2		0.020	0.051	0.103	0.211	4.605	5.991	7.378
3		0.115	0.216	0.352	0.584	6.251	7815	9.348
8		1.646	2.180	2.733	3.490	13.362	15.507	17.535
11		3.053	3.816	4.575	5.578	17.275	19.675	21.920
19		7.633	8.907	10.117	11.651	27.204	30.144	32.852
23		10.196	11.689	13.091	14.848	32.007	35.172	38.076

Solución mediante el uso de CalEst

En forma similar a otras distribuciones de probabilidad, el cálculo de éstas se obtiene utilizando **CalEst**. Con esta representación se tiene la ventaja de conocer la distribución, además de saber que se está calculando la probabilidad para diferentes valores de la variable aleatoria, que es difícil visualizarlo desde la tabla.

Guía para encontrar los valores de la χ^2

1. Especifique el nivel de significancia α (probabilidad α).
2. Determine los grados de libertad $gl = n$.
3. Los valores de la distribución χ^2 se encuentran en la gráfica de la distribución Chi cuadrada.

   a) Use el umbral para moverse a la derecha o izquierda según el valor de α.

b) Use dos umbrales que correspondan a $\frac{1}{2}\alpha$ y 1-$\frac{\alpha}{2}$.

La distribución χ^2 se utiliza para hacer inferencia sobre la varianza mediante intervalos de confianza, en la tabla 7.10 se presenta una síntesis de la tabla χ^2. La idea es que comprenda y adquiera habilidad para el cálculo de probabilidades de la χ^2 empleando esta tabla. También se extiende la práctica con el uso de la tecnología.

Ejemplo 7.18

Encontrar los valores de $\chi^2(\alpha/2)$ a la derecha (χ^2_d) y a la izquierda (χ^2_i) en una distribución de probabilidad χ^2, cuando $n = 24$ y $\alpha/2 = 0.05$. Así se puede calcular la probabilidad entre esos límites, tal que $P(\chi^2_i \leq \chi^2 \leq \chi^2_d) = 1 - \alpha = 0.90$.

Solución operativa clásica

Puede usar la tabla de la distribución de probabilidad χ^2 para obtener el valor. Una síntesis de ésta se presenta en la tabla 7.10, los valores que proporciona esta tabla corresponden al área derecha, es decir $P(\chi^2 \geq \chi^2_{tabla}) = p$. Dado que los grados de libertad en este caso son $gl = 23$, entonces, se ve en el renglón 9 y la columna 8 y se observa que $P(\chi^2 \geq 35.172) = 0.05 \;\; gl = 23$. Para el caso de la izquierda $P(\chi^2 \geq 13.091) = 0.95$ $gl = 23$. Aplicando la regla $P(\chi^2 < 13.091) + P(\chi^2 \geq 13.091) = 1$, se tiene que $P(\chi^2 < 13.091) = 1 - P(\chi^2 \geq 13.091) = 0.05$. En consecuencia, 90 % del área bajo la función densidad de la χ^2 se encuentra entre los valores 13.091 y 35.172, lo cual se expresa por:

$$P(13.091 \leq \chi^2 \leq 35.172) = 0.90$$

Esta fórmula se expresa en forma general como:

$$P(\chi^2_i \leq \chi^2 \leq \chi^2_d) = 1 - \alpha$$

Solución mediante el uso de CalEst

Una de las posibles ventajas al utilizar la tecnología es la comprensión de los conceptos de la función densidad y distribución de probabilidad de la χ^2 mediante un enfoque visual. De la colección de distribuciones presentadas por el **CalEst**, se encuentra esta distribución. Al escoger ésta aparece el contorno de la densidad para un grado de libertad.

En primera instancia se selecciona el grado de libertad requerido $n - 1$, y luego, similar a los casos de otras distribuciones se usan los umbrales para obtener el área y en consecuencia la probabilidad que se requieren. Inicialmente, usando la caluladora de probabilidades en el CalEst se mostrará el proceso para obtener los dos valores de la χ^2, ver figura 7.28. A continuación mediante los dos umbrales se determinan

los valores de la izquierda y derecha de la distribución χ^2, ya que estos son los que se utilizarán para calcular el intervalo de confianza para la varianza y desviación estándar.

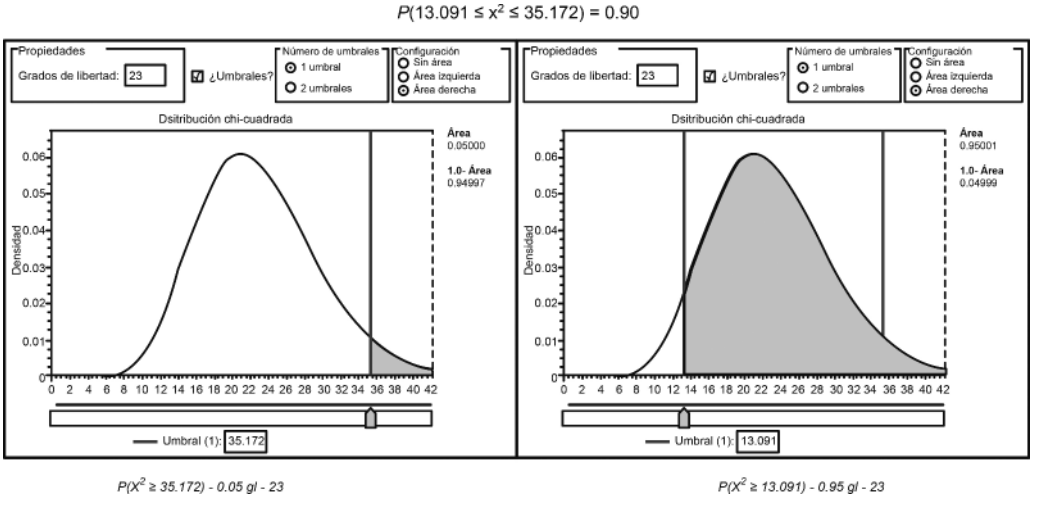


Figura 7.27 Cálculo de los valores de la distribución χ^2 para la construcción del intervalo de confianza para la varianza.

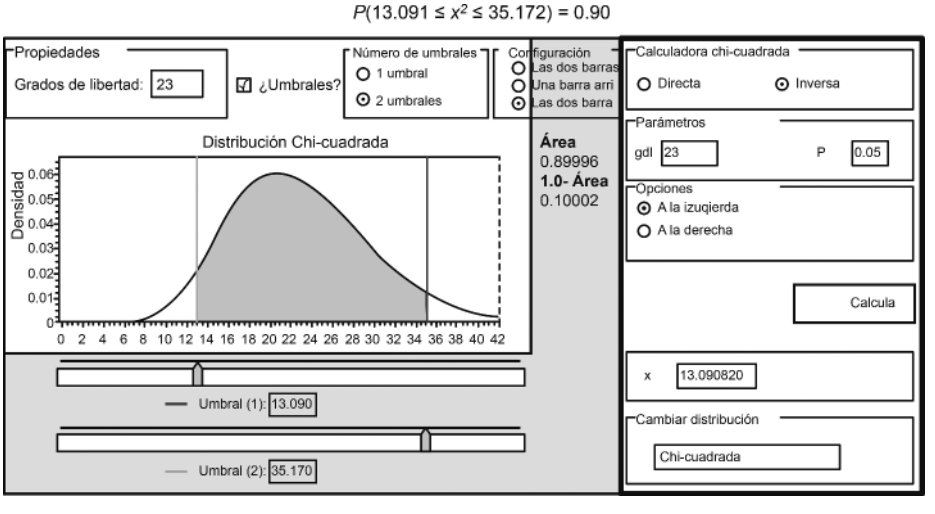


Figura 7.28 Cálculo de dos valores extremos en la distribución χ^2.

Enfoque visual para mostrar el uso de la tabla universal de la χ^2

Los grados de libertad son $n-1=24-1=23$. La gráfica de la figura 7.27 izquierda muestra una χ^2 con 23 grados de libertad y un área sombreada (probabilidad) de $\alpha=0.05$ en la parte derecha. $\chi^2=32.172$ es decir $P(\chi^2 \geq 32.172)=0.05 \;\; gl=23$. La gráfica de la derecha muestra el caso $P(\chi^2 \geq 13.091)=0.95$ $gl=23$. Esta gráfica permite reproducir el resultado del ejemplo y se muestra en la figura 7.27.

Enfoque visual cálculo directo

La figura 7.28 muestra el proceso del cálculo de probabilidades usando ambos umbrales, por lo tanto directamente se tiene $P(13.091 \leq \chi^2 \leq 35.172)=0.90$. La figura 7.28 también muestra el calculador de probabilidades para la χ^2.

Intervalo de confianza de $(1-\alpha)\,\%$ para σ^2

Una vez que se ha desarrollado el proceso operativo para el cálculo de los valores extremos χ_i^2 y χ_d^2 de la distribución χ^2, se procede a la construcción del intervalo de confianza a partir de la expresión:

$$P(\chi_i^2 \leq \chi^2 \leq \chi_d^2)=1-\alpha$$

Se sustituye en la expresión anterior $\chi^2=\frac{(n-1)S^2}{\sigma^2}$,

$$P(\chi_i^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_d^2)=1-\alpha$$

Invirtiendo cada término dentro de la expresión, se tiene:

$$P(\frac{1}{\chi_d^2} \leq \frac{\sigma^2}{(n-1)S^2} \leq \frac{1}{\chi_i^2})=1-\alpha$$

Finalmente:

$$P(\frac{(n-1)S^2 1}{\chi_d^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_i^2})=1-\alpha$$

De esa manera, el intervalo de confianza del $(1-\alpha)\,\%$ contendrá a la varianza σ^2 entre los valores:

$$\left(\frac{(n-1)S^2}{\chi_d^2}, \frac{(n-1)S^2}{\chi_i^2}\right) \tag{7.11}$$

Intervalo de confianza del $(1-\alpha)\,\%$ para σ

A partir de la fórmula (7.11), obteniendo la raíz cuadrada de cada término se generan los valores del intervalo de confianza para la desviación estándar, es decir:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi_d^2}}, \sqrt{\frac{(n-1)S^2}{\chi_i^2}}\right) \tag{7.12}$$

Ejemplo 7.19

En el contexto de la economía familiar está el precio de las medicinas, se selecciona una muestra de 9 farmacias de una ciudad mediana. Para un medicamento en específico se estudia la varianza del precio. La varianza de la muestra es 374.07. Suponga que la variable aleatoria: precio de este medicamento se distribuye como una normal. Calcule un intervalo de 90 % de confianza para la varianza y desviación estándar.

Solución

Revisando la tabla 7.10, dado que el número de grados de libertad es 8, $n-1=9-1=8$ y que el valor de $\alpha=0.10$ $(\alpha/2=0.05)$, observe que $\chi^2_d(0.05)=15.508$ y $\chi^2_i(0.95)=2.732$. Se sustituyen los datos en la expresión (7.11), se obtiene el intervalo del 90 % de confianza para la varianza:

$$\left(\frac{(n-1)S^2}{\chi^2_d},\frac{(n-1)S^2}{\chi^2_i}\right)=\left(\frac{(9-1)\times 374.07}{15.508},\frac{(9-1)\times 374.07}{2.732}\right)=(192.97,839.15)$$

El intervalo del 90 % de confianza para la desviación estándar queda establecido aplicando la ecuación (7.12):

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_d}},\sqrt{\frac{(n-1)S^2}{\chi^2_i}}\right)=\left(\sqrt{\frac{(9-1)\times 374.07}{15.508}},\sqrt{\frac{(9-1)\times 374.07}{2.732}}\right)=(13.89,28.97)$$

7.6 Resumen

Estimación puntual	Número calculado a partir de los datos obtenidos en una muestra. Se usa para estimar un parámetro de la población
Estimador puntual	Fórmula o regla que se usa para calcular la estimación puntual para un conjunto de datos.
Teorema del límite central	Este teorema establece que en una muestra aleatoria de una población con media μ y desviación estándar σ, cuando n es bastante grande, la distribución de $\overline{X}$ es aproximadamente normal con media $\mu_{\overline{x}}$ igual a μ y desviación estándar $\sigma_{\overline{x}}$ igual a $\frac{\sigma}{\sqrt{n}}$.

Intervalo de confianza	También conocido como estimación por intervalo, es un rango de valores con una probabilidad asociada o nivel de confianza $1-\alpha$. La probabilidad cuantifica la oportunidad de que el intervalo contenga el verdadero valor del parámetro poblacional.
Error estándar	Desviación estándar de la distribución muestral de un estimador puntual. Éste mide qué tanto varía de una muestra a otra el estimador puntual.
Distribución muestral	La distribución de un estimador puntual se llama distribución muestral.

Resumen del método estadístico en el proceso de estimación para una media

Descripción	Fórmula
Proceso de Estadarización	$\overline{X} = \mu_{\overline{X}} + Z\sigma_{\overline{x}} = \mu + Z\frac{\sigma}{\sqrt{n}}$ $Z = \frac{\overline{X}-\mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$
Media de $\overline{X}$	$\mu_{\overline{x}} = \mu$
Desviación estándar	$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$

Resumen del método estadístico en el proceso de estimación para una media

Descripción	Fórmula
Intervalo de confianza para μ cuando σ es conocida	Límite inferior $LI = \overline{X} + ei$ Límite superior $LS = \overline{X} + ed$ Donde $ei = \frac{Z(\alpha/2)\sigma}{\sqrt{n}}, ed = \frac{Z(1-\alpha/2)\sigma}{\sqrt{n}}$ Ancho del intervalo *ancho* $= ei + ed$

Descripción	Fórmula
Desviación estándar	Límite inferior $LI = \overline{X} - e$ Límite superior $LS = \overline{X} + e$ Donde $ei = \frac{t(n-1,\alpha/2)\sigma}{\sqrt{n}}$ $ed = \frac{t(n-1,(1-\alpha/2))\sigma}{\sqrt{n}}$ Ancho del intervalo *ancho* $= ei + ed$
Proceso de estandarización	$\overline{X} = \mu + t(n-1,\alpha/2)\frac{S}{\sqrt{n}}$ $\overline{X} = \mu + t(n-1,(1-\alpha/2))\frac{S}{\sqrt{n}}$ $t = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$
Tamaño de n, para estimar μ	$n = \frac{Z^2(\alpha/2)\sigma^2}{e^2}$

Resumen del método estadístico en el proceso de estimación para una proporción

Descripción	Expresiones
Intervalo de confianza para p	Límite inferior $LI = \widehat{p} + ei$ Límite superior $LS = \widehat{p} + ed$ Donde $ei = Z(\alpha/2)\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}$ $ed = Z(1-\alpha/2)\frac{\sqrt{\widehat{p}(1-\widehat{p})}}{\sqrt{n}}$ Ancho del intervalo *ancho* $= ei + ed$
Tamaño de n para estimar p	$n = \frac{Z^2(\alpha/2)p(1-p)}{ei^2}$

Descripción	Expresiones
Intervalo de confianza para σ^2	Límite inferior $LI = \frac{(n-1)S^2}{\chi^2_d}$ Límite superior $LS = \frac{(n-1)S^2}{\chi^2_i}$

Descripción	Expresiones
Intervalo de confianza para σ^2	Límite inferior $LI = \sqrt{\frac{(n-1)S^2}{\chi^2_d}}$ Límite superior $LS = \sqrt{\frac{(n-1)S^2}{\chi^2_i}}$

7.7 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

7.8 Ejercicios

Parámetro y estimación

7.1 Para medir la estatura de los estudiantes de su centro escolar, explique la forma en que se seleccionaría una muestra de 10 alumnos:

(a) De su grupo

(b) De su escuela

(c) De manera incorrecta, y decir por qué.

7.2 En una clínica de medicina familiar desean comparar la eficacia de una técnica naturista contra un método tradicional, esto es, con base en medicamentos, para tratar la gripe. Los médicos encargados del estudio dan el seguimiento a dos poblaciones pequeñas de 50 individuos. Ellos registran el número de días que las personas tardan en recuperarse de la gripe.

Método naturista	Método tradicional
1 3 4 3 5	1 4 5 7 8
3 4 6 6 3	3 9 8 8 8
6 4 3 4 6	4 9 9 1 6
7 5 4 5 5	4 1 3 9 7
4 4 5 5 6	8 2 3 1 9
3 4 4 7 4	1 7 5 1 1
6 3 5 5 5	1 6 8 2 9
5 4 4 4 5	4 1 1 1 3
4 5 4 3 4	4 2 4 9 4
5 4 2 4 4	1 3 8 1 1

(a) Asigne un número para identificar a cada uno de los individuos; primero a los que siguieron el método naturista y luego a los del otro método.

(b) Calcule la media μ para el número de días en los que se restablecen las 50 personas que se trataron con el método naturista y en seguida los que practicaron el método tradicional.

(c) Para cada método calcule la desviación estándar.

(d) Introduzca 50 papeles en una caja y en cada papel anote el número del individuo. Seleccione una muestra de 5 individuos de la población que siguió el método naturista. Anote en la tabla que sigue el número de días en que la persona se alivió, calcule la media y la discrepancia $\overline{X} - \mu$. Repita la operación para el método tradicional.

	Método naturista	Método tradicional
Selección 1		
Selección 2		
Selección 3		
Selección 4		
Selección 5		
$\overline{X}$		
$\overline{X} - \mu$		

(e) Seleccione una muestra de 10 individuos de la población que siguió el método naturista. Anote en la siguiente tabla el número de días en que la persona se alivió. Repite la operación para el método tradicional y calcula la media y la discrepancia .

	Método naturista	Método tradicional
Selección 1		
Selección 2		
Selección 3		
Selección 4		
Selección 5		
Selección 6		
Selección 7		
Selección 8		
Selección 9		
Selección 10		
$\overline{x}$		
$\overline{x} - \mu$		

(f) Compare los resultados de las muestras de tamaño 5 con las muestras de tamaño 10.

7.3 Mediante el uso del paquete estadístico (módulo Extras, opción Generador de números) o de la tabla de números aleatorios, repita el ejercicio anterior para seleccionar la muestra.

7.4 A continuación se exhibe una población de 50 ingenieros industriales que llevan trabajando 6 años. Calcule la media y la desviación estándar.

9806	8023	8383	8410	5395	5431	6257	8948	8293	7353
9179	9931	8595	8594	8318	8668	8130	7877	7458	7536
5823	6518	7720	7306	8821	6719	9755	6180	8523	7989
7519	5641	7988	7533	7363	10154	9858	8639	7587	10772
6671	7397	8180	7725	8224	10204	9063	6066	7784	8968

(a) Usando la tabla de números aleatorios, tome una muestra de 10 ingenieros industriales, calcule la media y la diferencia.

(b) Usando la tabla de números aleatorios, tome una muestra de 15 ingenieros industriales, calcule la media y la diferencia.

(c) Al observar los resultados de los incisos anteriores, ¿qué puede concluir?

(d) ¿Qué diferencia encuentra entre el inciso **(a)** de este problema y los datos del problema 1.1? ¿Cree que esa diferencia influya en la precisión de los resultados?

Distribución de la media muestral

7.5 Se han usado para $\overline{x}$ y $\overline{X}$ para referirse a la media muestral. ¿Para qué propósitos se usa la primera y para qué propósitos la segunda?

7.6 ¿Qué se entiende por distribución muestral de $\overline{X}$?

7.7 ¿Cómo es la media de la población de la media muestral $\overline{X}$ relacionado con la media de la población para las observaciones individuales X?

7.8 ¿Cómo es la desviación estándar de la población de la media $\overline{X}$ muestral relacionado con la desviación estándar de la población para las observaciones individuales X?

7.9 ¿Qué ocurre con la distribución muestral si crece el tamaño de una muestra?

7.10 Si seleccionamos muestras de una distribución normal, ¿qué se puede decir sobre la distribución de $\overline{X}$?

7.11 Para simplificar la relación entre la información de una población y una muestral, se considera una población pequeña de tamaño $N = 4$, ésta se refiere a una lista de cuatro personas con un crédito bancario de 10000, 20000, 30000, 40000, para facilitar la parte operativa, considere los valores de 1, 2, 3, 4 por 10 mil.

1. Calcule la media y la varianza para esta población.
2. Seleccione todas las muestras aleatorias de tamaño $n = 2$ con reemplazo, construya en una tabla los resultados, ahí mismo calcule las medias, las varianzas y las discrepancias de la media muestral y de la media de la población.
3. Describa la distribución de cada una de las medidas del inciso anterior y calcule la media y la varianza de las medias muestrales. Compare los resultados con los del inciso a.
4. Elija las muestras aleatorias de tamaño $n = 2$, sin reemplazo. Estime la media y la varianza de las medias muestrales. Contraste con las soluciones del inciso a.

7.12 La población de las ventas semanales de una empresa que labora pizzas, en miles de pesos, es 190, 240, 160, 210, y 300. Calcule y bosqueje la distribución muestral para las medias $\overline{X}$ y $S^2_{\overline{X}}$ varianzas muestrales considerando los siguientes casos:

1. Tome muestras de tamaño $n = 2$ con reemplazo y calcule la media de las medias y varianza muestrales.
2. Como en el inciso anterior con muestras sin reemplazo.
3. ¿Cuál es el valor del error estándar en a y b? Discuta los resultados.
4. ¿Cuáles son los errores estándar, si $n = 3$? ¿Qué diferencias observa? Explique.

5. Explique con sus palabras los conceptos de distribución muestral, la media de las medias, la media de las varianzas y el error estándar de la distribución muestral.

7.13 Se plantea una población pequeña con la información que se genera mediante la administración de una empresa que ofrece servicios. La variable es el tiempo de atención a un cliente. Los datos reportados son: 9, 6, 6, 5, 4, 3, 2.

1. Estime la media y varianza de ésta población.
2. Use la fórmula de combinaciones ${}_NC_n$ y construya una tabla que describa las muestras de tamaño, 1, 2, 3, 4, 5, 6 y 7 con las posibles muestras.
3. Prepare una tabla que represente las muestras de diferente tamaño, en cada caso obtenga las medias y calcule la media de las medias. ¿Describa lo que observa?
4. Construya una tabla que represente la distribución de las medias en cada muestra, y haga una gráfica para cada caso. Interprete sus resultados y elabore un escrito de sus conclusiones.

7.14 A continuación se dan los datos de la estatura de 100 jóvenes, cuya media es $\mu = 170$ y la desviación estándar es $\sigma = 7.5$. Usando la tabla de números aleatorios o el paquete estadístico, tome una muestra aleatoria de tamaño $n = 15$ y calcule la media muestral. Junte esta información con 20 compañeros y dibuje un diagrama de puntos para los 20 valores de $\overline{X}$.

181	173	175	184	174	165	160	177	176	178
164	180	162	160	154	185	166	166	177	182
173	163	177	172	166	169	158	177	173	172
164	162	184	170	183	165	183	158	170	179
174	160	160	173	169	160	164	168	157	173
163	165	172	163	171	183	177	157	172	183
164	172	171	184	170	177	172	167	170	170
171	164	161	168	163	179	174	172	175	164
157	166	158	167	172	190	164	173	169	157
178	165	167	172	176	162	174	159	176	169

7.15 Si se ha seleccionado una muestra de tamaño $n = 10$, con $\sigma_{\overline{X}} = 9$. ¿Cuántas observaciones más se necesitan tomar para reducir $\sigma_{\overline{X}}$ a 3 o a 1?

7.16 Si X tiene una distribución normal con $\mu = 20$ y $\sigma = 4$, calcule la probabilidad de que

1. $X > 21$,
2. $\overline{X} > 21$ si $\overline{X}$ se obtuvo de una muestra aleatoria de tamaño $n = 20$.

7.17 La variable X mide la habilidad motríz de los jóvenes para efectuar una actividad. Ésta tiene una distribución normal con parámetros $\mu = 110$ y $\sigma = 20$. Se desarrolla una nueva tecnología para aumentar la habilidad. Se toma una muestra de 36 jóvenes con una media $\overline{X} = 120$, calcule $P(\overline{X} \geq 120)$, ¿se podría concluir que es efectiva la nueva tecnología?

7.18 Una empresa mediana tiene 320 empleados, se tiene registrado que el número de horas que trabajan a la semana tiene una media $\mu = 37$ con una desviación estándar $\sigma = 6$.

1. ¿Cuál es la probabilidad de que un empleado seleccionado al azar, trabaje entre ± 2 horas a la semana en torno a la media?
2. Si se toma una muestra de $n = 16$ empleados, ¿cuál es la probabilidad de que trabaje entre ± 2 horas a la semana en torno a la media? En este caso considere el caso de muestreo sin reemplazo.

Teorema del límite central

7.19 Considere una población pequeña que se refiere al número de libros que compran al semestre siete estudiantes del primer año de administración, $N = 7$,. La distribución se indica a continuación:

X	1	2	3	4
$p(X)$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{2}{7}$	$\frac{3}{7}$

1. Determine el valor μ y σ^2. Para los siguientes incisos, asuma que para calcular la media $\overline{x}$ seleccione una muestra de tamaño $n = 49$.
2. ¿Cuál es el valor más pequeño y más grande para $\overline{x}$?
3. En referencia al TLC ¿Cuáles son los valores para $\mu_{\overline{x}}$ y $\sigma_{\overline{x}}$? ¿Cuál es la distribución de probabilidad de $\overline{x}$? Argumente su respuesta.

7.20 La administración de una línea de autobuses tiene registrado el tiempo, en segundos, de salida con una media $\mu = 400$ y una desviación estándar $\sigma = 120$. Suponga que la distribución de la media muestral es generada por muestras de tamaño $n = 100$. a. Encuentre el valor de $\mu_{\overline{x}}$. b. Encuentre $\sigma_{\overline{x}}$. c. Encuentre $P(360 \leq \overline{X} \leq 420)$. d. Encuentre $P(\overline{X} > 420)$. e. Si un valor de X es seleccionado, encontrar $P(360 \leq X \leq 420)$. Compare con el inciso c. f. Si un valor de X es seleccionado, encontrar $P(X > 420)$. Compare con el inciso d.

7.21 Repita el ejercicio 2 si $n = 400$. ¿Qué observa con respecto a las probabilidades en los incisos c y d?

7.22 Una empresa tiene el registro de los mensajes telefónicos para sus clientes, las duración de éstos tiene media $\mu = 220$ y varianza $\sigma^2 = 144$ segundos. Suponga que la distribución de la media muestral es generada por muestras de tamaño $n = 36$. a. Encuentre los valores de $\mu_{\overline{x}}$ y $\sigma_{\overline{x}}$. b. Obtenga $P(218 \leq \overline{X} \leq 222)$. c. Obtenga $P(\overline{X} > 223)$. d. Obtenga $P(221 < \overline{X} < 227)$.

7.23 Se considera el caso en que la variable original X de una población, es la variable aleatoria de los números (1, 2, 3, 4, 5, 6) y tiene una distribución de probabilidad uniforme, con media $\mu = 3.5$ y $\sigma^2 = 2.08$. De la población, se toma una muestra aleatoria, en esta ocasión la muestra corresponde a los números (2,3), así que la media de ésta muestra es $\overline{x} = 2.5$. Se observa que todas las posibles muestras corresponden al lanzar dos ($n = 2$) dados. Formalizando, sea la variable aleatoria X: el número de puntos o el número que se observa al lanzar un dado.

1. El modelo para este experimento, corresponde a una distribución de probabilidad uniforme, verifique que la media y la varianza de ésta distribución son: $\mu = 3.5$ y $\sigma^2 = 2.92$ respectivamente.

2. Se toma una muestra de tamaño $n = 2$, es decir, se lanza el dado dos veces, todas las posibles muestras de tamaño 2 al tirar los dados se tienen en la Tabla E1. A partir de esa, describa en tablas las distribución de la media y varianza de la muestra respectivamente, haga las gráficas en cada caso.

3. Verifique que: $\sum_{i=1}^{36}(x_i - \mu)^2 p(x_i) = 1.46$, sugerencia: elabore la tabla de distribución de la discrepancia $d_i = x_i - \mu$ al cuadrado.

Tabla E1. Descripción de todas las muestras de tamaño 2, con el cálculo de las medias y varianzas.

x_1	x_2	$\overline{x}$	S^2	x_1	x_2	$\overline{x}$	S^2	x_1	x_2	$\overline{x}$	S^2
1	1	1	0	3	1	2.0	2	5	1	3	8.0
1	2	1.5	0	3	2	2.5	2	5	2	3.5	4.5
1	3	2	2.0	3	3	3	0	5	3	4	2.0
1	4	2.5	4.5	3	4	3.5	0	5	4	4.5	0.5
1	5	3	8.0	3	5	4	0	5	5	5	0
1	6	3.5	12.5	3	6	4.5	4.5	5	6	5.5	0.5
2	1	1.5	0.5	4	1	2.5	4.5	6	1	3.5	12.5
2	2	2	0	4	2	3	0	6	7	4	8
2	3	2.5	0.5	4	3	3.5	0.5	6	3	4.5	4.5
2	4	3	2.0	4	4	4	0	6	4	5	2.0
2	5	3.5	4.5	4	5	4.5	0.5	6	5	5.5	0.5
2	6	4	8	4	6	5	2	6	6	6	0.0

4. En un sentido amplio, el proceso de lanzar los dos dados y observar el resultado se puede considerar como una muestra con reemplazo. Es decir, si se lanza un dado se ve uno de los 6 números, ahora si pone como regla que al lanzar el segundo dado no se permite el número que cayó en el primer dado, en tal caso, se vuelve a lanzar el dado hasta que el número sea diferente del primero. En esta condición, se especifica que el muestreo se lleva a cabo sin reemplazo. Construya las tablas similares para las distribuciones de la media, varianza y la discrepancia $d_i = x_i - \mu$ al cuadrado considerando esta situación.

7.24 En el contexto de aprendizaje, la siguiente práctica permite hacer la operación mental de retención al aplicar procedimientos y el hacer cálculos. En los siguientes dos puntos se propone construir la distribución de la media y varianza muestral con $n = 9$. Esto es, se lanza un dado nueve veces y se repite el proceso 25 veces.

1. Lanzar un dado 9 veces, registrar el número x que muestra el dado al caer. Calcule la media.
2. Repita este procedimiento 24 veces. Haga el histograma para las 25 medias. Calcule la media y la varianza de estas 25 medias. ¿Qué tanto se aproximan estos valores a los parámetros μ y σ^2? Utilice los resultados del teorema de límite central. A continuación sólo se presentan los dos primeros renglones de la tabla, en un hoja complete las 24 veces.

Lanzamiento	**Muestras de tamaño** $n = 9$									**Medidas**
1	6	3	2	2	5	1	4	4	1	3.11
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										

Ejercicios 1.4

7.25 La desviación estándar de una población es 9. ¿Cuál es la desviación estándar de $\overline{X}$ para una muestra aleatoria de tamaños a) $n = 25$, b) $n = 100$, c) $n = 400$?

7.26 La media de una población normal es igual a 25, con una desviación estándar de 4. Para una muestra aleatoria de tamaño $n = 8$, se necesita determinar a) la media de $\overline{X}$, b) la desviación estándar de $\overline{X}$ y c) la distribución muestral de $\overline{X}$.

7.27 Una compañía que produce baterías para automóvil otorga una garantía de 48 meses. Se sabe que la vida de una batería tiene una desviación estándar de 2 meses. Ocho amigos y yo hemos comprado estas baterías y los datos registrados del periodo de vida, en meses, fueron:

47.38	47.49	46.55	46.38	47.87	46.09	45.57	48.99	51.18

1. Calcule la media muestral de vida en estas baterías.
2. Calcule el valor de Z.
3. Con la información de estos datos, ¿se tiene razón en sospechar de la garantía de esta compañía?
4. ¿Esta media está dentro de una desviación estándar de los datos?

7.28 Un joven de 20 años de edad, que gusta del atletismo, ha empezado a practicar para participar en un medio maratón. Lleva el registro de su tiempo en 35 días corriendo 12 kilómetros; su promedio de carrera es $\overline{X} = 47.2$. El promedio poblacional de un corredor no tan bueno en la categoría de 18 a 25 es de 48.2, con una desviación estándar $\sigma = 1.6$.

1. Calcule el valor de Z.
2. ¿El tiempo de carrera de este corredor es bueno o es bastante rápido?

7.29 Las calificaciones en el ámbito nacional en la materia de historia universal tienen una media de 75 puntos y una desviación estándar de 5 puntos. En una escuela seleccionan una muestra de 50 estudiantes y les aplican la prueba; el promedio resultante fue de 80 puntos. Con base en estos datos, ¿se puede considerar que el nivel académico en historia universal de la escuela es bueno?

7.30 La media poblacional del tiempo de cocción, en una olla de barro, de una variedad de frijol es de 80 minutos con una desviación estándar de 8.

1. Si se selecciona una muestra aleatoria de tamaño $n = 64$, ¿cuál es la probabilidad de que la media muestral esté entre 78 y 82?
2. Con una muestra aleatoria de tamaño $n = 100$, ¿cuál es la probabilidad de que la media muestral esté entre 78 y 82?

7.31 La distribución del ingreso de las personas que trabajan en una maquiladora en una región central es $\mu = 13\ 000$ y $\sigma = 500$.

1. ¿Cuál es la distribución para $\overline{X}$ con base en una muestra aleatoria de 100 personas?

2. Evaluar la $P(\overline{X} > 13\ 500)$.

7.32 Para establecer un programa de becas, la dirección de una secretaría encargada del estudio ha registrado que el gasto mensual de los estudiantes de licenciatura en una comunidad tiene una media $\mu = 1500$ y desviación estándar $\sigma = 150$. Para establecer el monto de la beca el encargado del proyecto debe determinar que la probabilidad de la media en una muestra de tamaño $n = 30$: a. Esté entre 1250 y 1500. b.- Sea mayor que 1800. c.- Sea menor que 1800. d.- Sea mayor de 1250 y e.- Esté entre 1250 y 1750. Interprete los resultados que le de lugar a tomar una decisión.

7.33 Suponga que el peso X de un adulto hombre se distribuye como una normal con media $\mu = 77$ kg., y $\sigma = 9$ kg. Es decir $X \sim N(77, 9^2)$. Si tomamos una muestra de 16 adultos, ¿cuál es la probabilidad de que la media muestral $\overline{X}$ caiga entre 73 y 82 kgs?

7.34 El consumo bimestral de energía eléctrica en los hogares de una zona residencial es $\mu = 243$, con una desviación estándar de $\sigma = 62$ Kwh. Se ha seleccionado una muestra de tamaño $n = 16$ hogares. ¿Cuál es la probabilidad de que la media muestral del consumo de energía a.- sea mayor 220 Kwh? b.- esté entre 210 y 280 Kwh? c.- sea mayor a 260? Con esta información ¿será importante establecer un programa de ahorro en energía eléctrica? Si la distribución de la variable consumo de energía no tiene una distribución normal, ¿Ese tamaño es suficiente?

7.35 En una escuela el 60 % de los estudiantes no recibe atención médica, se selecciona una muestra de 150 estudiantes. Si la variable aleatoria X : no recibe atención médica, X tiene una distribución binomial, estime la probabilidad de que: a. X esté entre 82 y 101. b. X mayor que 97.

7.36 Una encuesta realizada en una ciudad, a gran escala, revela que el 30 % de la población adulta consume regularmente bebidas alcohólicas durante la comida. Considerando esta proporción, ¿cuál es la probabilidad de que de 1000 entrevistados el número de consumidores de bebidas alcohólicas sea: a. Menor a 280, b. 316 o más?

7.37 Una empresa fabrica aparatos eléctricos, la producción la almacenan en lotes de 250 piezas. Tienen una tasa de falla del 8 %. a.- Se sabe que sus clientes cambiarán de proveedor cuando su producto tienen más de 10 % de defectos. b.- Entre el 8 y 10 % de defectos los clientes le harán una nueva evaluación. c.- Si están entre 4 % y 8 % los clientes seguirán. d.- Menos del 4 % los clientes les harán un mayor pedido. ¿Cuál es la mejor decisión que tomen los clientes?

Intervalos de confianza para una media

7.38 Para estimar la media μ con la media muestral $\overline{X}$, encuentre (i) El error estándar de $\overline{X}$ y (ii) con un error marginal de $100(1-\alpha)$ % en cada uno de los siguientes casos:

1. $n = 144$, $\sigma = 24$, 1-$\alpha = 0.95$.
2. $n = 68$, $\sigma = 9.2$, $1 - \alpha = 0.99$
3. $n = 256$, $\sigma = 54$, $1 - \alpha = 0.90$

7.39 Determine la estimación puntual de la media poblacional μ y su margen de error con $100(1-\alpha)\,\%$ en cada uno de los siguientes casos:

1. $n = 150$, $\overline{X} = 86.2$, $S = 9.2$, $1 - \alpha = 0.975$.
2. $n = 225$, $\overline{X} = 925$, $S = 95$, $1 - \alpha = 0.90$.
3. $n = 144$, $\overline{X} = 0.786$, $S = 0.096$, $1 - \alpha = 0.95$.

7.40 Se pescan 58 truchas de un lago, con un promedio de 1920 gramos y una desviación estándar de 678 gramos. De estos datos:

1. Estime la media de las truchas de este lago.
2. Calcule el margen de error de 90 %.
3. Determine un intervalo de confianza de 99 %

7.41 Se toman 25 muestras diferentes del gasto que realizan los estudiantes en la compra de útiles escolares al iniciar un semestre. Suponga que la media poblacional real es $\mu = 900$ y la desviación estándar $\sigma = 17.5$. En la gráfica se muestran los 25 intervalos, estime cuántos no contienen a la media, expréselo en porcentaje e interprete.

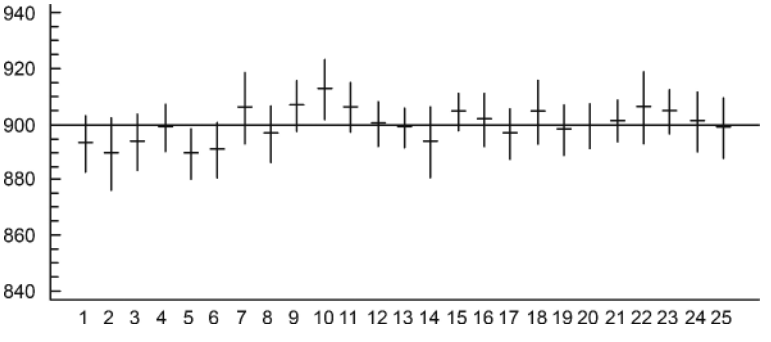


Figura 7.31 Ejercicio 4

7.42 Se da una muestra de 25 observaciones del tiempo de vida de unos focos. Esta variable viene de una población normal para la que se desconoce μ y la desviación estándar es $\sigma = 180$, la media muestral que se encontró es de 1200 horas. Construya un intervalo de confianza de 95 % para μ .

7.43 Se sabe que una variable aleatoria X tiene una distribución normal con media μ desconocida pero $\sigma^2 = 36$. Obtenga un tamaño de muestra n tal que la media esté en el intervalo $(\overline{x} - 1.5, \overline{x} + 1.5)$ con un nivel de confianza del 95 %.

7.44 Un sociólogo está interesado en estimar el ingreso semanal de los meseros en una ciudad grande. Entrevista a una muestra aleatoria de 75 meseros. La media y la desviación estándar son 3350 pesos y 368, respectivamente. Obtenga un intervalo de confianza de:

1. 90 % para el ingreso semanal de los meseros.
2. 95 % para el ingreso semanal de los meseros.

7.45 La elasticidad de un plástico que se utilizará para envolver alimentos se distribuye como una normal con media μ y varianza 110. Se mide la elasticidad de una muestra de 12 plásticos: 60, 53, 49, 72, 66, 57, 50, 82, 71, 62, 87, 48.

- Obtenga un intervalo de confianza de:

1. 90 % para la elasticidad.
2. 95 % para la elasticidad.

- Utilize el paquete estadístico para

1. 90 % para la elasticidad.
2. 95 % para la elasticidad.

7.46 Práctica para explorar las ideas conceptuales sobre los intervalos de confianza para μ. Sobre una población grande de estudiantes del nivel bachillerato, se seleccionó una muestra y se les preguntó sobre el número de horas que practican deporte a la semana. Se plantean cuatro situaciones diferentes:

1. Se seleccionó una muestra de 2500 estudiantes. La media muestral es $\overline{X} = 12.5$. La desviación estándar, σ, conocida es de 1.05 horas. Con base en los datos señalados:

   a) Encuentre un intervalo del 90 % de confianza para μ.
   b) Encuentre un intervalo del 92 % de confianza para μ.
   c) Encuentre un intervalo del 94 % de confianza para μ.
   d) Encuentre un intervalo del 96 % de confianza para μ.
   e) Encuentre un intervalo del 98 % de confianza para μ.
   f) Comente qué sucede al tamaño del intervalo de confianza en la medida en que el nivel de confianza se incrementa.

2. Se seleccionó una muestra de 2500 estudiantes. La media muestral es $\overline{X} = 10.5$ horas. La desviación estándar, σ, conocida es de 1.05 horas. Con base en los datos señalados:

   a) Encuentre un intervalo del 92 % de confianza para μ.

   b) Encuentre un intervalo del 94 % de confianza para μ.

   c) Encuentre un intervalo del 96 % de confianza para μ.

   d) Encuentre un intervalo del 98 % de confianza para μ.

   e) Compare los intervalos de confianza con los encontrados en el inciso a. Comente qué sucede al tamaño del intervalo de confianza en la medida en que el valor de la media muestral $\overline{X}$ cambió.

3. Se seleccionó una muestra de 2500 estudiantes. La media muestral es $\overline{X} = 12.5$ horas. La desviación estándar, σ, conocida es pero ahora tiene un valor de 2.05 horas. Con base en los datos señalados

   a) Encuentre un intervalo del 90 % de confianza para μ.

   b) Encuentre un intervalo del 92 % de confianza para μ.

   c) Encuentre un intervalo del 94 % de confianza para μ.

   d) Encuentre un intervalo del 96 % de confianza para μ.

   e) Encuentre un intervalo del 98 % de confianza para μ.

   f) Compare los intervalos de confianza con los encontrados en los incisos a y c. Comente qué sucede al tamaño del intervalo de confianza en la medida en que el valor de la desviación estándar de la población σ cambió.

4. Se seleccionó una muestra de 2000 estudiantes. La media muestral es $\overline{X} = 12.5$ horas. La desviación estándar, σ, conocida es de 1.05 horas. Con base en los datos señalados

   a) Encuentre un intervalo del 90 % de confianza para μ.

   b) Encuentre un intervalo del 92 % de confianza para μ.

   c) Encuentre un intervalo del 94 % de confianza para μ.

   d) Encuentre un intervalo del 96 % de confianza para μ.

   e) Encuentre un intervalo del 98 % de confianza para μ.

   f) Compare los intervalos de confianza con los encontrados en los incisos a y d. Comente qué sucede al tamaño del intervalo de confianza en la medida en que el valor del tamaño de muestra n cambió.

7.47 La administración de una cadena de supermercados está interesada en mejorar su servicio. Unos de los parámetros que midieron es el tiempo de espera en la caja registradora. El estudio se realizó en 5 tiendas midiendo 100 personas en cada tienda. Se obtuvieron los siguientes tiempos promedio de espera

(en minutos): 4.5, 6.3, 8.1, 5.1 y 7.2. La desviación estándar para cada tienda fue (en minutos): 2.5, 1.4, 5.3, 1.9 y 0.5, respectivamente. Obtenga el intervalo de confianza para cada tienda con un nivel del 95 %.

7.48 Para evaluar la necesidad de contribuir a la economía familiar un municipio pequeño quiere saber el estado de salud de los estudiantes. Una de las pruebas es medir el nivel de glucosa en la sangre. Seleccione al azar una escuela primaria, y de esa toma una muestra aleatoria a 50 estudiantes dando una media de 112mg/ml con una desviación estándar de 18mg/ml.

1. Obtenga el intervalo de confianza al 90 %
2. ¿Cuál es el error que se comete con la estimación anterior?

7.49 Un especialista de una dependencia de un gobierno estatal, realiza un estudio sobre el precio de unos artículos. Toma una muestra sobre el precio de un mismo producto a 20 comercios diferentes elegidos al azar en una ciudad. Suponiendo que los precios siguen una distribución normal con $\mu = 98.45$ pesos y $\sigma = 12.31$ pesos.

1. Obtenga el intervalo de confianza de 90 %.
2. Basados en los resultados del muestreo, un especialista establece que los intervalos de confianza deben estar entre 66.75 y 130.15. ¿Qué nivel de confianza está asociado a este intervalo?

7.50 Por estrategias de mercado, la administración de un supermercado desea conocer el tiempo que las personas pasa en la tienda. Considere que se toma una muestra aleatoria de 25 personas la media de permanencia en el supermercado es de 35 minutos. Suponga que la variable tiempo de permanencia, sigue una distribución normal con una desviación estándar de 8 minutos. Encuentre el intervalo de confianza de 95 % para la media de la población.

7.51 Considerando un ejemplo de economía empresarial. En una fábrica se empacan bolsas de café. Se toma una muestra de 16 bolsas y se encuentra que la media del contenido de las bolsas es de 500 gramos. El contenido de las bolsas sigue una distribución normal con una desviación estándar de 10.7 gramos. Encuentre e interprete el intervalo de confianza de 99 % para la media del contenido de las bolsas de café.

7.52 En un estudio de economía familiar. Se realiza una encuesta en 85 casas para conocer el nivel de endeudamiento en tarjetas de crédito. Considere que la variable aleatoria monto de la deuda por persona siguen una distribución normal con $\overline{x} = 6,710.21$ en pesos y $S = 765.35$. Obtenga los intervalos de confianza para la media μ 90 % y 97 %.

Intervalos de confianza para la media en muestras pequeñas y múltiples valores de S

7.53 Múltiples desviaciones permite considerar otros intervalos en términos a la desviación estándar. Estas observaciones están relacionadas con las expresiones $z(\alpha/2, gl = n-1)\frac{\sigma}{\sqrt{n}}$,y $t(\alpha/2, gl = n-1)\frac{S}{\sqrt{n}}$. Compruebe que la longitud de los intervalos también dependen del tamaño de muestra n; a medida que n crece se cuenta con mayor información y la precisión de los intervalos es mejor. Verifique estas

diferentes situaciones usando el ejercicio que se plantea en el inciso a, tambien use el programa CalEst para responder a sus preguntas.

1. Para evaluar el contenido de zinc en cajas de cereal, la secretaría de salud toma una muestra a 46 cajas de cereal y encuentre que la media del contenido de zinc es de 1.9 miligramos por gramo de cereal. Considere las siguientes tres desviaciones estándar : 0.4, 0.48 y 0.52. Encuentre los intervalos de confianza de 95 % para la concentración media de zinc.

7.54 La secretaría del medio ambiente de una gran ciudad desea determinar si el promedio de la cuenta de bacterias por unidad volumen de agua en una presa está dentro del nivel de seguridad de 200. Un ingeniero bioquímico de esa secretaría toma una muestra de 10 unidades de volumen de agua y cuenta el número de bacterias:

211	212	178	173	196
182	217	193	199	186

1. Encuentre un intervalo de confianza de 90 % para la media de una población normal.
2. Encuentre un intervalo de confianza de 99 % para la media de una población normal.

7.55 Un tipo de aislante de cable eléctrico se sometió a prueba para ver en cuál nivel de voltaje se presentaba alguna falla. Se realizaron 12 pruebas y los voltajes registrados son:

52	64	38	66	52	68
60	44	48	46	70	62

Encuentre un intervalo de confianza de 95 % para la media de una población normal.

7.56 Con el fin de mejorar el aislante se elaboró un nuevo cable eléctrico. Los voltajes de falla para 14 pruebas son:

36	44	41	53	38	36	34
54	52	37	44	51	35	44

1. Encuentre un intervalo de confianza de 95 % para la media de una población normal.
2. Se podría concluir que este nuevo cable es mejor, igual o peor que el del ejercicio anterior, ¿por qué?

7.57 En el contexto de la economía familiar, se seleccionó una muestra de 30 parejas que se encontraban en una explanada de las llamadas comidas rápidas, y se les preguntó el gasto que habían realizado ese día en comida. Los datos redondeados a pesos son:

90	110	85	135	140	112	108	143	162	138
149	99	160	148	152	106	117	97	146	91
135	106	110	122	152	136	126	108	135	58

1. Verifique si los datos siguen una distribución normal, use el papel de probabilidad normal.
2. Encuentre un intervalo de confianza del 95 % y 97 % de confianza para la media, varianza y desviación estándar del gasto en comidas que realizan las familias.
3. Con un 99 % de confianza, ¿es el gasto al menos de 105 pesos?

7.58 La secretaria de salud de un estado quiere evaluar el apoyo económico que ha otorgado a las mujeres embarazadas en una zona rural. El peso de 23 niños recién nacidos en la clínica de esa zona rural tiene el siguiente registro: un media de $\overline{x} = 3072.6g$ y un desviación estándar de $S = 152g$. Encuentre los intervalos de confianza de 88 %, 95 % y 97 % para la media de una distribución de probabilidad normal. Use el apoyo tecnológico.

7.59 En un proceso químico es importante que una solución que se va a usar como reactivo tenga un pH de 8.2 exactamente, ya que de no ser así le ocasionaría una pérdida económica a la compañia. Después de 10 mediciones realizadas, se obtuvieron los siguientes valores:

8.18	8.16	8.17	8.22	8.19
8.17	8.15	8.21	8.16	8.18

1. Obtenga un intervalo del 90 % confianza, ¿este intervalo contiene al valor 8.2?
2. Obtenga un intervalo del 88 % confianza, ¿este intervalo contiene al valor 8.2?

7.60 Determinar el tamaño de muestra para estimar el verdadero valor de μ entre ± 0.2 con un nivel del 90 % de confianza. Considere que la variable de interés X tiene una distribución normal con varianza $\sigma^2 = 0.36$.

7.61 El coordinador del departamento de administración de una universidad, tiene que presentar en una reunión de trabajo, el desempeño de los aspirantes en un examen de admisión a la maestría en administración de empresas. Éste incluía 15 preguntas. Tomó una muestra de 25 solicitudes y encontró que la media de la calificación de los solicitantes fue de 9.6 con una desviación estándar de 1.5. Encuentre los intervalos de confianza de 95 % y 99 % para la calificación.

7.62 Una oficina de recursos humanos realiza pruebas de actitud a todos los solicitantes. De una muestra de 9 solicitudes, el promedio del puntaje fue de 234.8 con una desviación estándar. Considere que los puntajes siguen una distribución normal. Encuentre los intervalos de confianza de 80 % y 90 %.

7.63 Una fábrica produce azulejos para baños. Con el fin de un estudio económico, la administración considera un lote de producción. Se toman 24 azulejos y se pesan. La variable peso sigue una distribución normal con una desviación estándar de 54 gramos. La media del peso de los azulejos es de 1846 gramos.

1. Encuentre el intervalo de confianza de 99 % para el lote de producción.
2. Si el tamaño de la muestra es de 15 azulejos, ¿qué ocurre con el intervalo de confianza de 99 %?

3. Si el tamaño de la muestra sigue siendo de 24 azulejos pero la desviación estándar es de 62 gramos. ¿Cómo cambia el intervalo de confianza de 99 %?

Intervalos de confianza para una Proporción

7.64 Se toma una muestra de 124 estudiantes para conocer su opinión sobre el control de la natalidad. Del total, 86 se manifiestan a favor. Calcular un intervalo de confianza del 95 % para la proporción de la población.

7.65 La secretaría de salud desea saber el porcentaje de niños que han cubierto la serie de vacunas básicas en una zona rural. Se busca un nivel de confianza del 90 %, y una precisión . Obtener el tamaño de muestra. Como no hay una idea sobre la proporción de la población, como estrategia inicial se supone que $p = 0.20$.

7.66 Un biólogo está probando una variedad de semillas de girasol. Pone a germinar 120 semillas de las cuales sólo 106 germinaron. Calcule un intervalo de confianza del 95 % para este resultado.

7.67 Se sabe que el 25 % de personas de una comunidad consumen pescado. Durante dos meses se lleva a cabo un programa para aumentar su consumo. Al finalizar ese periodo se levanta una encuesta a 84 personas y se obtuvo un.

1. Obtener un intervalo del 90 % de confianza para la proporción poblacional p.

2. Obtener un intervalo del 95 % de confianza para la proporción poblacional p.

3. ¿Se puede concluir que el programa fue eficiente para aumentar el consumo?

Práctica para la χ^2 e intervalos de confianza para una varianza

7.68 En cada uno de los siguientes casos $\chi^2(0.01, n-1)$, $\chi^2(0.025, n-1)$, $\chi^2(0.95, n-1)$ y $\chi^2(0.99, n-1)$. Encontrar estos valores de χ^2 con los siguientes grados de libertad (gl): *a.-* $gl = 9$, *b.-* $gl = 15$, *c.-* $gl = 25$.

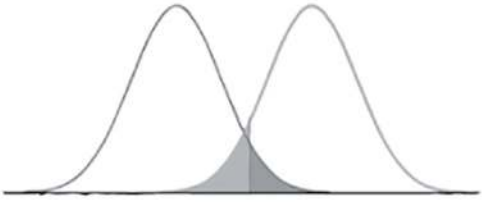
7.69 Si los grados de libertad de la distribución son $gl = 2$, encuentre los valores de x en los siguientes casos: *a.* $P(\chi^2 \geq x) = 0.01$, *b.* $P(\chi^2 \geq x) = 0.05$, *c.* $P(\chi^2 \geq x) = 0.99$, *d.* $P(\chi^2 \geq x) = 0.01$, *e.* $P(\chi^2 \geq x) = 0.9$, *f.* $P(\chi^2 \geq x) = 0.5$.

7.70 Encontrar el valor crítico χ^2_i a la izquierda cuando $n = 12$ y $\alpha = 0.05$.

7.71 Encontrar los valores críticos χ^2 a la derecha e izquierda de la distribución cuando $n = 13$ y $\alpha = 0.05$.

7.9 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.



Capítulo 8

Prueba de hipótesis sobre un parámetro

Es mucho mejor una respuesta aproximada a la pregunta correcta, la cual es comúnmente vaga, que la respuesta correcta a la pregunta errónea, la cual puede hacerse de una forma precisa. La combinación de ciertos datos y un deseo doloroso de una respuesta, no aseguran que una respuesta razonable se pueda extraer del cuerpo de datos.

J.W. Tukey.

Competencia general

Comprender los conceptos esenciales para plantear hipótesis estadísticas, tal que, se adquieran habilidad para probarlas e interpretarlas.

Competencias específicas

- Identificar los elementos básicos de estadística y probabilidad en el planteamiento de una prueba de hipótesis estadística.
- Describir diferentes situaciones o problemas donde se puedan plantear hipótesis estadísticas.
- Redactar con la mayor claridad posible los conceptos estadísticos básicos en el procedimiento de prueba de hipótesis.
- Adquirir habilidad para conocer el vocabulario que enmarca el planteamiento de hipótesis estadística, así como el procedimiento de prueba.
- Calcular las probabilidades que caracterizan el nivel de significancia, la región de rechazo y los tipos de error.
- Comprender la manera en que se construye la estrategia para realizar una prueba de hipótesis estadística.
- Explicar lo que se comprende por estadístico de prueba.
- Reconocer diferentes escenarios para utilizar el estadístico de prueba apropiado.
- Aplicar los conceptos y el procedimiento para realizar ejercicios en diferentes situaciones.
- Establecer la relación entre la estimación de intervalos de confianza y pruebas de hipótesis.
- Leer un artículo de divulgación en temas de administración o economía para identificar los conceptos estadísticos en prueba de hipótesis.

8.1 Introducción

En el contexto de la administración existe una variedad de situaciones en las que se plantean soluciones que tienen que ver con la captación de información, con la cual se examina si los proyectos son exitosos. La administración de una empresa de autobuses ha recibido a través de encuestas sobre sus servicios, la información de que se plantea una corrida, por parte de los clientes, a una hora diferente a la propuesta. Planea hacer el cambio, con el supuesto de que beneficiará a más del 70 % de sus clientes. En este caso la hipótesis se formaliza H: el cambio beneficia a más del 70 % de clientes. La variable aleatoria es el cambio beneficia: X, con valores sí, no. La información se obtendrá de una muestra de un grupo de clientes seleccionados al azar.

El efecto económico aparece en muchas fábricas, por ello resulta importante el control estadístico del proceso, en general referido como control de calidad. En un proceso se toman pequeñas muestras cada determinado tiempo, una hora, para saber si el producto en cuestión cumple con las especificaciones de calidad. Así en un proceso de pintado de un aparato, la media del grosor debe ser igual a 5 mm; si esto se cumple, la señal del sistema de control es que la producción no pare. En caso contrario se detiene el proceso, entonces la hipótesis a estudiar es H: la media del grosor de pintura es diferente a 5 mm. Los valores de la variable aleatoria X, grosor de pintura, sus valores se miden en milímetros. La medición del grosor de pintura a una muestra de aparatos proporcionará la información para decidir si se rechaza la hipótesis.

En el ambiente de la economía familiar, en México un investigador afirma que los empleados que trabajan en las agencias de limpieza ganan a la semana en promedio menos de dos salarios mínimos. En ese sentido, la media del salario a la semana es de 775 pesos. La hipótesis se escribe como, H: la media de salario a la semana de los trabajadores de limpieza es menor a 775 pesos. La variable aleatoria X es el salario a la semana en pesos. Seleccionar una muestra de empleados de limpieza proporcionará la información para la prueba. Note que, el salario mínimo es una referencia temporal.

8.2 Planteamiento y conceptos básicos de una hipótesis estadística

El mundo de la información 1. Economía familiar

Los trabajadores que prestan sus servicios en las llamadas plazas de venta de saldos (outlet), tienen que trasportarse todos los días, lo que implica un gasto importante, e influye en el gasto familiar. Como una prestación los dueños de cada local se organizan a través de la administración de la plaza para ofrecerles el alquiler diario de autobuses.

Preguntas sobre la naturaleza del estudio

¿Cuánto gastan, en pesos, los empleados a la semana en el traslado? ¿Ayudará esta estrategia ahorrar tiempo de traslado y así mejorar el rendimiento de los empleados? ¿Cómo se puede medir el rendimiento

en el desempeño de las actividades que realizan los empleados? ¿Cómo se organiza la información para evaluar esta situación? ¿Qué información se requiere? ¿Cuáles son las variables aleatorias involucradas en este estudio?

Estrategias estadísticas sobre el estudio. El propósito inicial es hacer algunas afirmaciones sobre el planteamiento del estudio. Así por ejemplo se dice que: los empleados gastan en promedio al menos 192 pesos a la semana, lo que repercute en su ingreso, ¿vale la pena considerarlo? El tiempo que realizan en trasladarse de su casa a la plaza es en promedio mayor a 50 minutos, con el servicio de trasporte, ¿se beneficia al empleado en el ahorro de tiempo? Con el fin de tener un conocimiento sobre esta situación se recurre a tomar una muestra de la población de empleados en la plaza que usen el transporte urbano. Las variables aleatorias implícitas en el estudio son X_1: gasto en el transporte, X_2: tiempo de traslado. Para evaluar el desempeño se utiliza un cuestionario del que al final se obtendrá como reporte una calificación. Como parte de un proceso de abstracción se plantea la afirmación en términos de un parámetro, en este caso la media, como una hipótesis.

Hipótesis: la media del gasto de transporte realizado por los empleados es mayor a 192 pesos.

$$H : \mu > 192$$

De manera análoga:

Hipótesis: el tiempo medio de traslado realizado por los empleados es mayor a 50 minutos.

$$H : \mu > 50$$

Estas suposiciones se identifican como hipótesis alternativas. Con el fin de verificar estas afirmaciones sobre la media μ, se contrastan con la negación a la que se denomina hipótesis nula.

En símbolos, éstas se expresan por: $H_0 : \mu \leq 192$, gastan al menos 192 pesos y $H_0 : \mu \leq 50$, respectivamente.

Prueba de hipótesis

Una *hipótesis* es una afirmación que expresa el valor del parámetro de una población, por ejemplo la media poblacional μ. En la prueba de hipótesis, la idea es dar el beneficio de la duda a la hipótesis nula, si el valor del parámetro es razonable. La hipótesis nula se rechaza sólo si los datos de la muestra indican que el valor del parámetro es no razonable, esto se verá con más detalle en el análisis de datos. La *inferencia estadística* es un procedimiento cuyo objetivo es generar una conclusión sobre una población, mediante la información que proporciona una muestra seleccionada de datos. Una rama importante de la inferencia es la *prueba de hipótesis*. Ésta consiste en un procedimiento para seleccionar entre dos hipótesis, conocidas como *hipótesis nula* e *hipótesis alternativa.*

Selección de la hipótesis: Cuando en la práctica se desea hacer un estudio del desempeño de un proceso, como evaluar si nuevas estrategias de cambio son efectivas o la explicación de un problema, tal situación involucrará una hipótesis. Ésta se propondrá como una afirmación, un jucio o una idea. Por lo general, le corresponderá indentificar la afirmación como hipótesis alternativa y la negación como nula.

Hipótesis nula. Es una proposición que indica que no hay diferencia (no hay efecto, no hay cambio). Ésta se plantea usualmente en términos del parámetro (medida de la población) contiene el signo igual, y se denota por H_0.

Hipótesis alternativa. Es una afirmación que indica la verdad del parámetro en lugar de la hipótesis nula. Usualmente se expresa con los símbolos $<$, $>$ o $\neq$. Ésta se denota por H_1.

Papel de la distribución normal en la prueba de hipótesis

Como preliminares se considera el cálculo de probabilidades de la distribución normal que son importantes en el proceso de prueba de hipótesis estadísticas. Se examina un caso específico, para ello considere la variable X que sigue una de probabilidad normal con los siguientes valores: $\mu = 36$ y desviación estándar $\sigma = 10$.

El cálculo de la probabilidad toma como indicador un umbral x_c, éste sirve como referencia para decidir si una de las hipótesis es sustentada por la información de la variable X tomada de una muestra aleatoria. Para estos valores de la normal, se plantea obtener la probabilidad en alguna de las siguientes tres situaciones, menor o igual a $x_c = 20$, mayor o igual a $x_c = 52$, o entre $x_c = 20$ y $x_c = 52$, es decir:

$$i.\ P(X \leq 20), \quad ii.\ P(X \geq 52), \quad iii.\ P(20 \leq X \leq 52).$$

En la figura 8.1 se muestra el cálculo de estas tres probabilidades en particular, observe la gráfica a la derecha de esta figura. Ahí se muestra la probabilidad a la izquierda de 20, y a la derecha de 52, este cálculo, también se le conoce por la probabilidad en las colas, izquierda y derecha. Se destaca que estas ideas conceptualizan el uso de la probabilidad en las pruebas de hipótesis y en el cálculo de los intervalos de confiaza.

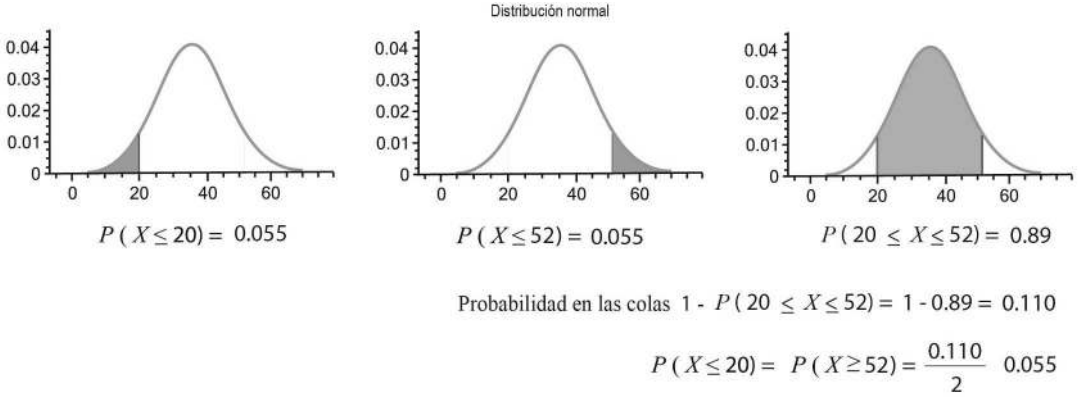


Figura 8.1 Cálculo de la probabilidad entre dos valores, umbrales, de la variable X.

Este procedimiento se extiende a procesos en los que la variable aleatoria es un estadístico, por ejemplo la media muestral, para la normal, o se tienen estadísticos con otro tipo de distribución tal como la t de *Student*, o la χ^2. Una vez familiarizado con el cálculo de probabilidades, visto en el capítulo 6, se aplica éste a la metodología de prueba de hipótesis.

Motivación de una prueba de hipótesis

El propósito principal de las pruebas de hipótesis es hacer posible una elección adecuada entre dos hipótesis, que por lo general pueden referirse a valores de parámetros. Para construir el mecanismo de elección de una de las hipótesis, se supone que la distribución de la variable aleatoria de interés sigue un cierto tipo conocido; por lo general se plantea la distribución de probabilidad normal. La metodología que se presenta en este capítulo se refiere a la prueba de hipótesis paramétrica, en particular para la media μ, la proporción p y la varianza σ^2.

Estudiar a la población, en principio, a través de estos tres parámetros, fundamenta nuestras investigaciones acerca de los problemas relacionados con la administración y economía en particular. Las preguntas surgen de las unidades que componen a la población que se desea observar o experimentar. Para alcanzar esta meta recurrimos a la información que proporciona la muestra. En la tabla 8.1 se exponen las afirmaciones hipotéticas que se hacen sobre la población y los elementos de la muestra para verificarlas. Por ejemplo, en la venta de tortillas durante un mes del año, para determinada ciudad, se puede afirmar sobre la media del precio de la tortilla considerando los diferentes lugares de venta. Además puede ser de interés corroborar la proporción de expendios que respetan el precio oficial, así como verificar la homogeneidad en el precio, que se estudia a través de la varianza.

Tabla 8.1 Descripción de la inferencia estadística mediante el planteamiento de una hipótesis.

Población:	**Muestra**
Hipótesis: Una afirmación sobre el parámetro de la población	**Estadísticos**
$(\mu,\ \rho,\ \sigma^2)$	$(\overline{X},\ \hat{\rho}\ S^2)$
Afirmaciones	**Cálculos**
$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$	$\overline{x}_m$
$H_0 : p = p_0$ vs $H_1 : p \neq p_0$	$\widehat{p}_m$
$H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$	S_m^2

Con la finalidad de motivar el contexto de cómo nace el planteamiento de una hipótesis estadística, así como la idea general del método para construir la prueba de hipótesis, se recurre a la simulación de un sistema de seguridad que consiste en la lectura del iris. Desde luego el procedimiento de prueba genera una serie de conceptos estadísticos y de probabilidad que es necesario comprender y dominar para adquirir un amplio conocimiento sobre la prueba. El aprendizaje adquirido en este capítulo se puede aplicar a diferentes escenarios tanto de administración como en economía.

Propuesta visual de aprendizaje de los conceptos de prueba de hipotesis

Mediante la lectura del iris se describirá la estructura del material didáctico que aquí se expone. En esta situación, la referencia del iris se genera usando la distancia de Hamming (Daugman, 2003) y tiene valores entre 0 y 100 (%). La idea consiste en simular un sistema de seguridad que lee el iris a un usuario para que pueda acceder a un lugar. Con la indicación proporcionada por una lectura, el sistema identifica al usuario y como resultado lo declara impostor o auténtico. En general la información se obtiene leyendo el iris de una persona. La variable aleatoria X es la lectura del iris, y se supone que X tiene una distribución de probabilidad normal con media μ y desviación estándar σ.

Prueba de hipótesis mediante la lectura del iris: impostor o auténtico. Con la finalidad de presentar el mecanismo de la prueba de hipótesis estadística se ha elaborado un módulo que simula un sistema de seguridad. Aquí se hace la descripción de manera gráfica y operativa para construir el procedimiento de la prueba de hipótesis. Para utilizarlo recurra al apartado de solución usando la tecnología descrito más adelante.

La acción consiste en la lectura del iris tal y como se muestra en la figura 8.2, mediante ésta un sistema identificará si un usuario es impostor o auténtico. Así, la información se obtiene leyendo el iris de una persona. La variable aleatoria X es la lectura del iris, X tiene una distribución de probabilidad normal con media $\mu = 36$ y desviación estándar $\sigma = 10$ tal como se muestra en la gráfica a la izquierda en la figura 8.2 y describe la lectura de un impostor. Para el caso de un auténtico, la variable X tiene una distribución de probabilidad normal con media $\mu = 68$ y desviación estándar $\sigma = 10$, como se ve en la gráfica a la derecha en la figura 8.2. Los valores para los parámetros de la normal y los indicados en el eje horizontal corresponden a las lecturas expresadas en porcentaje. La línea negra representa la lectura del iris de una persona, que corresponde a la información; en este caso se observa que está ubicada a la derecha, lo cual indica que el usuario es auténtico y por lo tanto el sistema acertó.

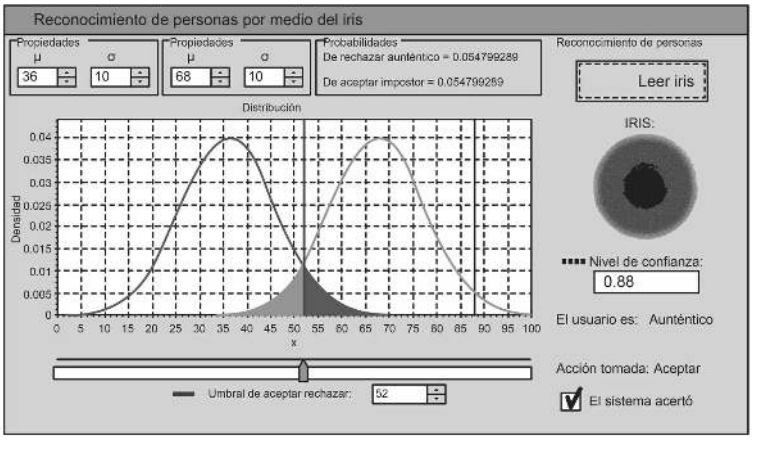


Figura 8.2 Descripción del simulador que lee el iris de una persona para clasificarlo como auténtico o impostor.

Observe que existe una línea entre las dos distribuciones, esta genera un área a la derecha y otra a la izquierda. Asocie esta indicación con las probabilidades previamente calculadas. Más adelante se verá

qué papel desempeñan éstas en el contexto de una prueba de hipótesis.

Todo contraste va a plantear la comparación entre dos hipótesis en el contexto de la información generada por la muestra. Se plantean las hipótesis conocidas como la nula y la alternativa

$$\begin{aligned} H_0 &: \text{El sistema reconoce a un usuario impostor,} \\ H_1 &: \text{El sistema reconoce a un usuario auténtico.} \end{aligned} \tag{8.1}$$

Observación

H_0 representa la hipótesis nula. H_1 representa la hipótesis alternativa.

Los métodos de prueba de hipótesis permitirán hacer una elección adecuada entre estas hipótesis, con un determinado nivel de error; por ejemplo, el sistema puede no reconocer a un usuario impostor cuando realmente es un impostor. Dos errores pueden cometerse al juzgar una hipótesis y se refieren como error tipo I y error tipo II. El error tipo I tiene lugar cuando se rechaza la hipótesis nula siendo verdadera. El error tipo II se comete al no rechazar una hipótesis nula que es falsa. Estos errores se expresan en términos de probabilidad, y es aquí donde surge la necesidad del cálculo de probabilidades.

Descripción del error tipo I es la probabilidad de rechazar H_0 cuando es verdadera. Esto se expresa como:

$$\alpha = P(\text{rechazar } H_0 \mid H_0 \text{ es verdadera}). \tag{8.2}$$

De este planteamiento, un punto que se debe tener presente en la estructura de la prueba de hipótesis, es que para realizarla, se supone que la hipótesis nula es *verdadera* (cierta).

Observe en la figura 8.2 que el cálculo de la probabilidad dada por la expresión (8.2), es:

$$\alpha = P(X \geq 52 \mid H_0 \text{ es verdadera}) = 0.055 \tag{8.3}$$

La interpretación del error tipo I para el caso del iris es, el sistema no rechaza, acepta al impostor, cuando realmente el sistema debe rechazar al impostor.

Descripción del error tipo II , es la probabilidad de no rechazar cuando es falsa. Esto se expresa como:

$$\beta = P(\text{no rechazar } H_0 \mid H_0 \text{ es falsa}) \tag{8.4}$$

Análogamente, la probabilidad de la expresión (8.4) es:

$$\beta = P(X < 52 \mid H_0 \text{ es falsa}) = 0.055 \tag{8.5}$$

La interpretación del error tipo II para el caso del iris es, el sistema rechaza a un auténtico, cuando realmente el sistema debe reconocer a un auténtico. En resumen:

Error tipo I: Probabilidad de aceptar un impostor,

Error tipo II: Probabilidad de rechazar un auténtico.

Importancia del umbral

En referencia a la figura 8.2, el umbral $\overline{x}_c$ llamado punto crítico, separa a las distribuciones, en esta situación, con una probabilidad igual para ambos lados, no siempre es el caso. La probabilidad del error tipo I, por lo general se fija de manera convencional, los valores que frecuentemente se consideran son 0.1, 0.05, 0.01, cabe resaltar que estos valores ayudarán en la decisión que aporten los datos para rechazar la hipótesis nula. Observe que si el valor del estadístico cae a la izquierda del umbral, en relación a la figura 8.2, entonces esta información apoya a la hipótesis nula, en caso contrario, si el valor del estadístico cae a la derecha, los datos no dan evidencia para apoyar la hipótesis nula y por tanto se rechaza.

Nuevamente observe la figura 8.2, al oprimir donde dice: leer iris, a partir de ahí aparece un valor de la variable de respuesta, referido como nivel de confianza, en este caso 0.88 y representa el 88 %. En la parte gráfica aparece una línea que señala ese valor 0.88, y en la parte inferior del lado derecho se describe la situación generada por la información, el usuario es Auténtico, Acción tomada: Aceptar, la relación de esta acción con el procedimiento de prueba de hipótesis caracteriza rechazar la hipótesis Nula. La interpretación de la acción tomada por la encuesta (datos de la muestra), *el sistema acertó*. Esta se asocia a una de las cuatro decisiones mencionadas en la tabla 8.2.

Tabla 8.2. Prueba de hipótesis y lectura del iris.

Se tiene cuatro decisiones (juicios) a considerar con la información proporcionada por el valor de la variable.
1) La hipótesis nula, verdadera, el sistema reconoce a un usuario impostor y no lo admite (no rechaza H_0), acción tomada es correcta.
2) La hipótesis nula, verdadera, el sistema reconoce a un usuario impostor y lo admite (rechaza H_0), acción tomada es incorrecta y genera el error tipo I.
3) La hipótesis nula, no es verdadera, el sistema reconoce a un usuario auténtico y lo admite (acepta), acción tomada es correcta.
4) La hipótesis nula, falsa, el sistema reconoce a un usuario como un auténtico y no lo admite cuando realmente es un auténtico, acción tomada es incorrecta y genera el error tipo II

Resumen del planteamiento de hipótesis y los tipos de error

Con la información proporcionada por la tabla 8.2 y la figura 8.2 se puede resumir como se muestra en la tabla 8.3. De la gráfica, el área a la derecha del punto crítico bajo la hipótesis nula indica que se rechaza H_0 cuando en realidad H_0 es verdadera. El otro, el área a la izquierda del punto crítico considerando la hipótesis alternativa, indica que no se rechaza H_0; en realidad, H_0 es falsa, así:

Tabla 8.3 Esquema de las cuatro decisiones en la prueba de hipótesis.

	Hipótesis nula	
Decisión	H_0 es verdadera	H_0 es falsa
No rechazar H_0	Decisión correcta	β : Error tipo II
Rechazar H_0	α : Error tipo I	Decisión correcta

Recopilación de los términos tratados en el iris

El proceso para verificar la recomendación del sistema de seguridad considerando n personas se plantea mediante dos hipótesis, el procedimiento de la prueba de hipótesis consiste en seleccionar una de ellas, la elección se puede sintetizar como sigue:

1. En términos de probabilidad se ha planteado como referencia el valor de α, la probabilidad del error tipo I. Es decir, $\alpha = P(\overline{X} \leq \overline{x}_{ci})$ o $\alpha = P(\overline{X} \geq \overline{x}_{cd})$. Donde $\overline{x}_{ci}$ o $\overline{x}_{cd}$ expresan el valor crítico a la izquierda o a la derecha. Nota: el valor de α también se refiere como *nivel de significancia.*

2. Se obtiene la información y con ella se calcula la media de la muestra $\overline{x}_m$: entonces, si la media $\overline{x}_m$ es menor que el *valor crítico* $\overline{x}_{ci}$ o mayor que el *valor crítico* $\overline{x}_{cd}$, se deduce que no hay evidencia para apoyar a la hipótesis nula. En caso contrario los datos generados por la información no dan elementos para rechazar la hipótesis nula. Así se puede concluir en el contexto del problema.

3. Se ha calculado la probabilidad de la variable $\overline{X}$ con respecto a la media muestral, es decir $valor - p = P(\overline{X} \geq \overline{x}_m)$. El $valor - p$, se denomina *nivel de significancia descriptivo*, es la probabilidad de tener valores de la muestra, al menos tan grande como los que se obtuvieron dado que la hipótesis nula es verdadera. Fisher vio el $valor - p$ como un índice que mide la fuerza de la evidencia contra la hipótesis nula, Sterne and Smith (2001).

Analogía entre el sistema de seguridad y prueba de hipótesis

A partir de este sistema de simulación "lectura del iris" descrito en la figura 8.2, se puede utilizar para explicar las probabilidades de ambos errores o sus complementos. Si se repiten muchas veces las lecturas y se calculan las proporciones en los errores, éstas estiman de manera aproximada los valores de las probabilidades. También se pueden generar diferentes escenarios sobre el sistema de seguridad al cambiar los valores de la media y desviación estándar. Esta actividad permitirá simular distintos sistemas, lo que hace posible comprender los papeles que desempeñan cada uno de los elementos que intervienen en la prueba de hipótesis. Nota: en el apartado complemento didáctico se planea una actividad usando la opción iris. Una propuesta que puede ser de interés, es que planee varias actividades utilizando esta alternativa didáctica.

A continuación, en tres tablas se presenta la analogía del sistema de seguridad planteado mediante el iris y lo que corresponde a una prueba de hipótesis estadística.

Tabla 8.4 Analogía entre la prueba de hipótesis y sistema de seguridad.

Sistema de seguridad	Prueba de Hipótesis
1. Lectura del iris	1. Experimento
2. Usuario	2. Efecto
3. Suposición: El usuario es impostor	3. Hipótesis nula H_0: el efecto es cero
4. Aseveración: El usuario es auténtico	4. Hipótesis alternativa H_1: el efecto es real
5. Censor de lectura	5. Experimentador
6. Tarea del censor de lectura: A: Demostrar que la suposición es no verdadera (que el usuario es auténtico) y que la aseveración es verdadera.	6. Tarea del experimentador: A: Demostrar que la hipótesis nula puede no ser verdadera (que el efecto no es cero) y que la hipótesis alternativa puede ser verdadera (se rechaza H_0), el efecto puede ser real.

Tabla 8.5 Analogía entre la prueba de hipótesis y sistema de seguridad

Sistema de seguridad	Prueba de hipótesis
7. Naturaleza de la prueba	7. Naturaleza del experimento
A: Censor débil:	A: Alta confianza/baja potencia de la prueba.
• Alta confianza de que cuando el usuario es impostor realmente es así.	• Alta confianza $(1-\alpha)$ de que cuando el efecto encontrado sea real, éste realmente es así.
• Alto riesgo de admitir a un auténtico que es impostor.	• Alto riesgo (β) de demostrar que un efecto real sea cero.
• Admitir a un impostor como impostor.	• Concluir que un efecto cero es cero.
• Una prueba pobre para admitir que un verdadero auténtico es auténtico.	• Un experimento pobre para verificar que un efecto real verdadero es real.

Tabla 8.6 Analogía entre la prueba de hipótesis y sistema de seguridad

Sistema de seguridad	Prueba de hipótesis
B: Censor estable:	B: Baja confianza /alta potencia de la prueba
• Baja confianza de que cuando el usuario es impostor realmente es así.	• Baja confianza $(1-\alpha)$ de que cuando el efecto encontrado sea real, éste realmente es así.
• Bajo riesgo de admitir a un auténtico que es impostor.	• Bajo riesgo (β) de demostrar que un efecto real sea cero.
• Admitir a un impostor como impostor.	• Concluir que un efecto cero es cero.
• Una prueba buena para admitir que un verdadero auténtico es auténtico.	• Un experimento bueno para verificar que un efecto verdadero es real.

La prueba de hipótesis usando muestras aleatorias, lectura del iris

Mediante la lectura del iris se pueden tomar muestras aleatorias para verificar si un sistema reconoce a un auténtico, tal como se planteó en la expresión (8.1). Para ello, suponga que la lectura del iris es mayor a 36. La hipótesis correspondiente al impostor es igualdad o menor a 36, y en principio se considera la igual, en ese sentido el iris de un impostor se distribuye como una normal con media $\mu = 36$. Planteamiento de las hipótesis, se quiere autenticidad de un usuario, así como hipótesis alternativa el usuario es un auténtico y la hipótesis nula, el usuario es un impostor (no es un auténtico).

Para generar la idea de la prueba, se fijará que la distribución del auténtico es una distribución de probabilidad normal con media $\mu = 65$ donde la desviación estándar σ se supone conocida con un valor de 30 ($\sigma = 30$). La figura 8.3 reproduce el planteamiento de ambas hipótesis. Observe a partir de la figura y del planteamiento de la hipótesis alternativa que el valor crítico está a la derecha de la distribución normal que representa a la hipótesis nula, por ello la probabilidad del nivel de significancia se calcula a la derecha, es decir $\alpha = P(\overline{X} \geq \overline{x}_{cd})$ o $\alpha = P(Z \geq z(1-\alpha))$.

El proceso para obtener la información, consiste en elegir una muestra al azar de una población de posibles usuarios. Se propone una muestra de $n = 9$ usuarios. Las lecturas del iris (multiplicadas por 100) son: 24, 32, 78, 76, 73, 69, 86, 54, 46. Por lo tanto, la media de la muestra es: $\overline{x}_m = 59.8$. A continuación se describirá la secuencia de paso para mostrar el procedimiento para realizar la prueba de hipótesis.

Procedimiento para realizar la prueba de hipótesis

1. Plantear las hipótesis:

   a) Una hipótesis nula, $H_0 : \mu = 36$ (Usuario impostor)

   b) Establecer la hipótesis alternativa, $H_1 : \mu > 36$ (Usuario auténtico)

2. Proponer un nivel de significancia, valor de α (establecido de acuerdo al valor de la probabilidad del error tipo I que se desee tener), en este caso $\alpha = 0.05$. Encontrar un valor de referencia (punto crítico) a partir de una distribución de probabilidad normal que corresponde al *estadístico de prueba* $\overline{X}$. Calcular el valor crítico para $\overline{X}$ correspondiente al nivel de significancia propuesto, es decir obtener: $\overline{x}_c$

$$\overline{x}_c = \mu + z(1-\alpha)\frac{\sigma}{\sqrt{n}} = 36 + 1.645\frac{30}{\sqrt{9}} = 52.45$$

3. Comparar el valor del estadístico calculado $\overline{x}_m$ con el punto crítico $\overline{x}_c$. Para fijar ideas considere el caso de la media, si el estadístico $\overline{x}_m$ resulta mayor que el punto crítico $\overline{x}_c$, entonces se rechaza la hipótesis nula. En este caso, $\overline{x}_m = 59.8 > \overline{x}_c = 52.45$, observe la figura 8.3.

4. Dar una conclusión e interpretación al problema estudiando, para este ejemplo se rechaza la hipótesis nula.

Interpretación: El valor del estadístico cae a la derecha del valor de referencia (umbral en la figura 8.3), es decir $\overline{x}_c = 52.45 < \overline{x}_m = 59.8$ o acorde con la regla de decisión $\overline{x}_m = 59.8 > \overline{x}_c = 52.45$, por lo

tanto se rechaza la hipótesis nula y se concluye que en promedio el sistema de seguridad identifica a los usuarios auténticos.

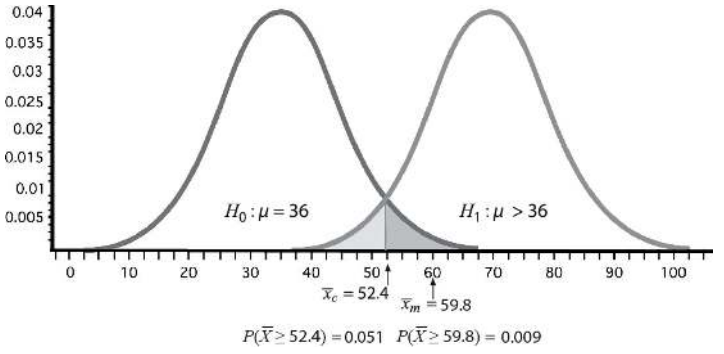


Figura 8.3 Descripción de la metodología para realizar una prueba de hipótesis estadística.

Comentario: Se puede decir que este procedimiento es directo, ya que considera la parte real del problema que se estudia; en resumen se establece el punto crítico con la referencia de la probabilidad, error tipo I. Luego, con información de la muestra, se observa si la media está antes o después del valor del punto crítico. Puede observar que este procedimiento depende de la ayuda tecnológica y gráfica de la distribución normal en el CalEst.

Prueba de hipótesis usando el estadístico Z

Dada la opción de usar la tabla de la distribución de probabilidad normal estándar, implica que la distribución normal original, que representa la parte real del proceso, se estandarice. Se repiten los pasos anteriores pero ahora mediante el proceso de estandarización. Los presentamos de manera sintética.

1. Plantear las hipótesis:

   a) Hipótesis nula, $H_0 : \mu = 36$ (Usuario impostor)

   b) Hipótesis alternativa, $H_1 : \mu > 36$ (Usuario auténtico)

2. Proponer un nivel de significancia, valor de α, en este caso $\alpha = 0.05$. Encontrar un valor de referencia (punto crítico) a partir de una distribución de probabilidad; la normal corresponde al estadístico (estimación muestral estándar) Z. Calcular el valor crítico para Z correspondiente al nivel de significancia propuesto, es decir obtener: z_c

$$z_c = z(1 - \alpha) = 1.645$$

3. Comparar el valor del estadístico calculado z_m con el punto crítico z_c.

$$\text{Proceso de estandarización: } z_m = \frac{\overline{x}_m - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{59.8 - 36}{10} = 2.38$$

Si el estadístico z_m resulta mayor que el punto crítico z_c, entonces se rechaza la hipótesis nula. En este caso, $z_m = 2.38 > z_c = 1.645$.

4. Dar una conclusión e interpretación al problema estudiado. En esta situación, la decisión es rechazar H_o ya que $z_m = 2.38$ está a la derecha del valor crítico, es decir $z_m = 2.38 > z_c = 1.645$.

Se copia la figura 8.3 en la figura 8.4; observe ahora los valores de la variable Z, el umbral o valor crítico es $z_c = 1.64$, y el valor $z_m = 2.38$ corresponde al valor del estadístico $\overline{x} = 59.8$.

Comentario: Este es el procedimiento clásico, que recurre a la distribución normal estándar; por uso y costumbre se dice que busque el valor crítico en las tablas de la normal.

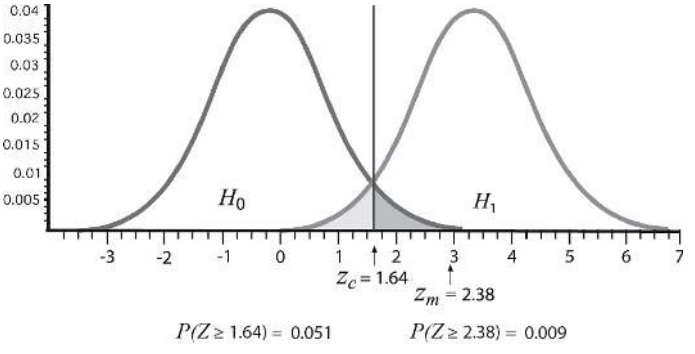


Figura 8.4 Proceso de prueba de hipótesis estandarizado.

Procedimiento de la prueba de hipótesis utilizando el $valor - p$

El $valor - p$, es un procedimiento alternativo para verificar si los datos apoyan la hipótesis nula. En este caso, el criterio para decidir sobre la hipótesis nula es:

rechazar H_0 si el $valor - p$ es menor que α,
no rechazar H_0 si el $valor - p$ es mayor e igual que α

Entonces el $valor - p$ es:

$$valor - p = P(\overline{X} \geq \overline{x}_m), \text{ o}$$
$$valor - p = P(Z \geq z_m)$$

En el ejemplo, se tiene que: $valor - p = P(\overline{X} \geq \overline{x}_m) = valor - p = P(\overline{X} \geq 59.8) = 0.009$, o $valor - p = P(Z \geq 2.38) = 0.009$, note que esta probabilidad es menor que la de $\alpha = 0.055$. Por lo tanto, se rechaza la hipótesis nula. Nota: Es muy conveniente calcular el $valor - p$ con el objetivo de tener presente la fuerza de la evidencia contra la hipótesis nula.

Comentario 1. Es común que los paquetes estadísticos reporten el $valor - p$. Esta aparente simplicidad provoca en los usuarios de los paquetes mucha desinformación de los conceptos estadísticos. Lo

recomendable es estudiar con detalle la relación de las probabilidades con el procedimiento de prueba de hipótesis.

Comentario 2. Por lo general, no existe un acuerdo para determinar qué tan pequeño debe ser el valor de α con el propósito de rechazar H_0 para establecer una fuerte evidencia a favor de la hipótesis H_1. Puesto que se puede dar el caso de que un investigador opte un valor de $\alpha = 0.05$, mientras un segundo investigador considere un nivel de $\alpha = 0.01$ existe la posibilidad de que en el primer caso se rechace H_0 y no así en el segundo, con base en los mismos datos. En ese sentido, será adecuado presentar los resultados en términos del $valor - p$, el cual se puede interpretar como el menor valor de α con el que se puede rechazar H_0, en función del valor del estadístico de prueba.

Comentario 3. Cuando se toma la decisión de rechazar H_0 en términos del valor de α, se dice que el resultado es estadísticamente significativo. Una discusión interesante respecto de este tema, que es importante considerar, aparece en el libro de Prieto y Herranz (2005).

Comentario 4. En el resumen se presentará de manera gráfica la relación entre estos tres procedimientos de la prueba de hipótesis, usando como referencia los datos de este apartado.

Resumen de los conceptos de la prueba de hipótesis

Hipótesis es una afirmación sobre la característica, parámetro, de una población.
Prueba de hipótesis es un procedimiento que permite verificar la afirmación sobre la población, a través de la información de la muestra y como referencia la probabilidad.
Error tipo I. Rechazar una hipótesis nula verdadera. La probabilidad de cometer el error tipo I es igual al *nivel de significancia* α, valor de referencia para la *prueba de hipótesis*.
Error tipo II. Probabilidad de no rechazar una hipótesis nula que es falsa.
Valor crítico o punto crítico es el umbral, calculado a partir del nivel de significancia. El valor es la referencia para verificar la evidencia de H_0.
El estadístico de prueba es una medida de discrepancia entre los que expresan los datos y lo que se esperaría para evaluar si H_0 es verdadera.
El $valor - p$ **es la probabilidad**, calculada bajo el supuesto de que la hipótesis nula es verdadera, y es el área de cola de la distribución que va más allá del valor del estadístico de prueba.

El planteamiento general de las hipótesis para la μ de una población es:

La hipótesis nula:

$$H_0 : \mu = \mu_0$$

La hipótesis alternativa es alguna de las siguientes opciones:

1. $H_1 : \mu > \mu_0$
2. $H_1 : \mu < \mu_0$
3. $H_1 : \mu \neq \mu_0$

A las dos primeras se les conoce como *pruebas de un lado* y a la tercera como *prueba bilateral*. Con la finalidad de ejemplificar las tres situaciones de la prueba de hipótesis, la gráfica a la derecha en la figura 8.5 señala la descripción de la prueba de hipótesis 1, y la gráfica de izquierda caracteriza la segunda prueba, a éstas se les denomina pruebas de una cola, observe la probabilidad del error tipo I en esta figura.

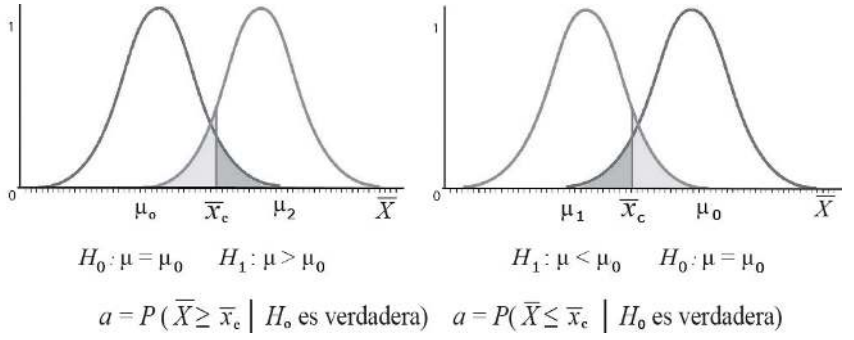


Figura 8.5 Descripción de las pruebas de hipótesis cuando la alternativa es mayor que μ_0 o menor que μ_0.

Asimismo, la figura 8.6 describe la prueba de hipótesis para ambos lados; es importante observar que la probabilidad del error tipo I se divide entre dos. En la práctica, cuando se establece el nivel de significancia se toma $\alpha/2$. En la gráfica es representado por α_1 y α_2.

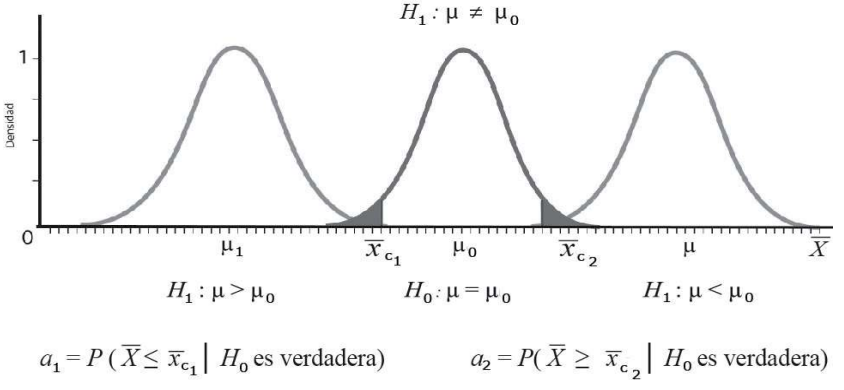


Figura 8.6 Descripción de la prueba de hipótesis cuando la alternativa es diferente a μ_0.

Ejemplo 8.1

En una compañía, la administración ha invertido en capacitación para sus ingenieros de proceso con el propósito de aumentar la productividad. La variable X, el número de toneladas producidas se distribuye como una normal con media μ y varianza $\sigma^2 = 4$, $X \sim N(\mu, \sigma^2)$. La meta es superar la media actual de

8 toneladas y proponen como alternativa llegar a una media de 9 toneladas. La observación consistirá en tomar una muestra de la producción. Con base en la muestra intentar decidir si la capacitación fue eficiente. Escogen una muestra de tamaño $n = 9$ y el valor de media $\overline{x}_m = 9.29$

Nota: Cuando se plantea realizar una mejora de un proceso, proyecto o actividad, es común que la afirmación se exprese diciendo que la acción tomada originará un cambio favorable, el cual se medirá con la información en la muestra. Para el ejemplo, se dirá que la capacitación provocará un incremento en la producción. Así la administración pudo haber sostenido que el entrenamiento es exitoso si la media de la producción es mayor a 8. Esto da lugar a tener varias hipótesis alternativas; aquí se ha escogido la media ($\mu = 9$) con el fin de ilustrar varios conceptos en el procedimiento de la prueba de hipótesis e interpretación de la conclusión, en este caso para saber si la capacitación influye en el rendimiento.

Solución operativa clásica

1. Planteamiento de las hipótesis:

$$H_0 : \mu = 8,$$
$$H_1 : \mu = 9 \ (\mu > 8).$$

2. A continuación, se fija el nivel de significancia, en este caso se propone $\alpha = 0.05$. En el supuesto de que la *hipótesis nula es verdadera* ($\mu = 8$), el siguiente paso es calcular el punto crítico. Dado que se tiene la información del tamaño de muestra $n = 9$, la desviación estándar $\sigma = 2$ y el nivel de significancia $\alpha = 0.05$, entonces $z(1 - \alpha) = z(0.95) = 1.645$. De la expresión indicada en la lectura del iris se tiene que:

$$\overline{x}_c = \mu_0 + z(1 - \alpha)\frac{\sigma}{\sqrt{n}} = 8 + 1.645\frac{2}{3} = 9.097 \simeq 9.1$$

3. Este procedimiento permite conluir que H_0 se rechaza si el valor observado de la muestra $\overline{x}_m$ es mayor que el valor del punto crítico $\overline{x}_c$. En este ejemplo, $\overline{x}_m = 9.29$, se cumple que: $\overline{x}_m = 9.29 > \overline{x}_c = 9.1$, por lo tanto se rechaza la hipótesis nula.

4. El plan de capacitación resultó adecuado.

Observación 1. Con base en la información de la muestra, se puede formalizar la estrategia para saber si los datos dan evidencia en decidir si se rechaza o no se rechaza H_0.

$$\text{Se rechaza } H_0 \text{ si } \overline{x}_m > \overline{x}_c, \text{ con } \alpha = \alpha_0.$$

Observación 2: Utilizando la expresión (8.2) se verifica que:

$$\alpha = P(\overline{X} \geq \overline{x}_c \mid H_0 \ \mu = \mu_0 = 8) = P(\overline{X} \geq 9.1 \mid \mu = 8) = 0.05 \tag{8.6}$$

Este valor que indica la ecuación (8.6), se puede verificar en el **CalEst**. Por otro lado, siguiendo el tradicional proceso de estandarización se tiene que:

$$\alpha = P\left(\frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{9.1 - 8}{\frac{2}{3}} \mid \mu = 8\right) = P(Z \geq 1.645 \mid \mu = 8) = 0.05$$

Interpretación sobre el nivel de significancia α y el valor de β

En la tabla 8.2 se han presentado dos decisiones que dan lugar a dos errores, α: error tipo I rechaza H_0 cuando ésta es verdadera ($H_0 : \mu = \mu_0$), y β : error tipo II no rechazar H_0 cuando ésta es falsa ($\mu > \mu_0$ o $\mu < \mu_0$). Ambos dependen de una probabilidad, se parte fijando el valor de significancia (tamaño de la prueba) α ¿qué ocurre con el valor de β? ¿De qué depende su valor? ¿Es importante considerarlo?

Resumen del ejemplo 8.1, se ha realizado la prueba de $H_0 : \mu = 8$ contra la alternativa $H_1 : \mu = 9$ con un nivel de significancia $\alpha = 0.05$. La conclusion que se sigue de la prueba es rechazar H_0 si el valor observado $\overline{x}_m$ de la muestra es mayor que 9.1 ($\overline{x}_m = 9.21 > 9.1$).

Es conveniente, tener en mente que la prueba ha generado una región crítica, debido al valor de $\alpha = 0.05$, para cualquier hipótesis alternativa $\mu = \mu_1$, en particular $\mu_1 > \mu_0$. Esto da la idea de que el error del tipo II debe tomar un valor, en particular para la cola derecha, en lugar del valor $\alpha = 0.05$.

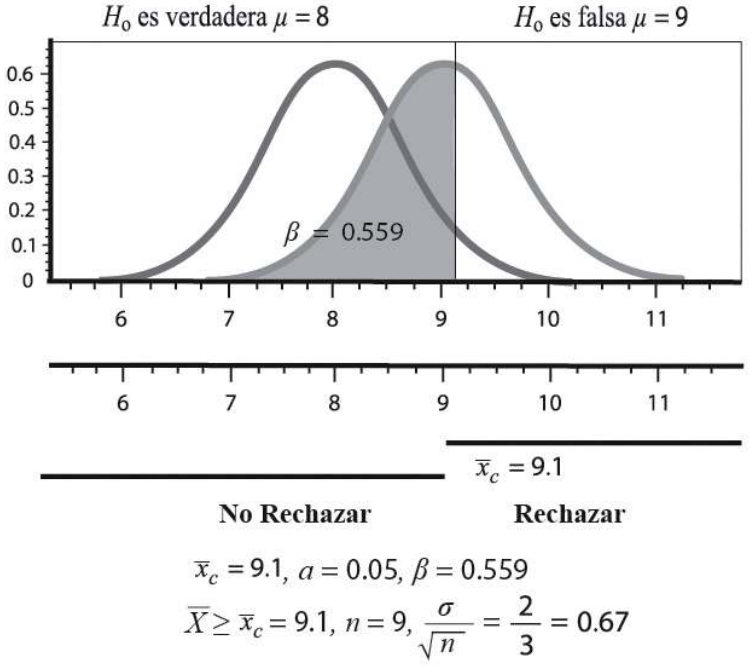


Figura 8.7 Descripción de las probabilidades de los errores tipo I y tipo II, con n = 9.

El cálculo de la probabilidad del error tipo II es:

$$\beta = P(\overline{X} < \overline{x}_c = \mu_0 + z\frac{\sigma}{\sqrt{n}} = 9.1 \mid H_0 \text{ es falsa}) = 0.559 \tag{8.7}$$

Usando la transformación estándar:

$$\beta = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{9.1 - 9}{0.667} \mid \mu = 9\right) = P(Z < 0.150 \mid \mu = 9) = 0.559$$

Para fijar ideas, se describe los cálculos de ambas probabilidades error tipo I y tipo II, y la interpretación en la figura 8.7. En este ejemplo, para un $\overline{x}$ relativamente mayor que 9.1, se debería tener la seguridad con una confianza $(1 - \alpha = 0.95)$ de que se está en lo correcto cuando se rechaza H_0. Es claro, que se puede decidir no rechazar H_0 cuando en realidad se debería rechazar. De aquí surge la idea del motivo de asociar la hipótesis alternativa con lo que se espera establecer. Al tomar la decisión, surge la interpretación del resultado y está en relación con la probabilidad de los errores. Si se rechaza H_0 y se toma la hipótesis alternativa, entonces se tiene una probabilidad de error α que será incorrecta. Si no se rechaza H_0, entonces H_0 puede ser verdadera o tener un error del tipo II, en el ejemplo, $\beta = 0.559$. No se podría sentir seguridad en recomendar H_0 justo cuando no se tiene suficiente evidencia de rechazar esta. Por ello es preferible decir que no se rechaza H_0 en vez de que se acepta H_0, simplemente no se tiene la suficiente evidencia estadística para declarar H_0 falsa en un nivel de significancia $\alpha = 0.05$. Esto no sería tan grave, si el error tipo II es pequeño.

Un tamaño de muestra más grande genera más información, vea el caso en que $n = 49$, se observa cual es el impacto sobre el valor de β, se complementa la información en la figura 8.8. La selección del tamaño de muestra a permitido especificar ambos valores de α y β, esto permite concluir que el *tamaño de muestra es esencial*. Para calcular la probabilidad del error tipo II, primero se obtiene el valor de punto crítico (nueva referencia al cambiar el tamaño de muestra). Se conserva el valor de $\alpha = 0.05$, así:

$$\overline{x}_c = \mu_0 + z(1 - \alpha)\frac{\sigma}{\sqrt{n}} = 8 + 1.645\frac{2}{7} = 8.477$$

Para este caso, el valor de β es:

$$\beta = P\left(\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{8.477 - 9}{0.29} \mid \mu = 9\right) = P(Z < -1.803 \mid \mu = 9) = 0.036$$

Interpretación de los errores tipo I y tipo II en la práctica

En la tabla 8.7 se presenta una interpretación práctica de los riesgos de los valores de (α) y (β), y describe las formas que comúnmente se refiere a estos errores.

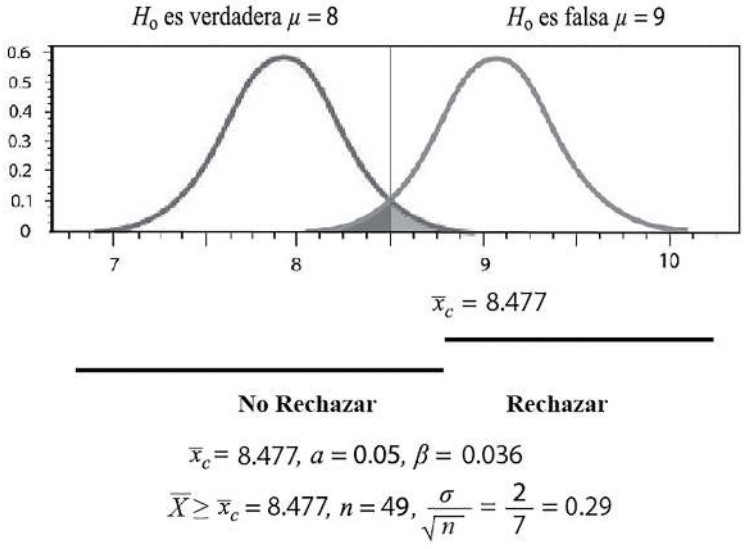


Figura 8.8 Descripción de las probabilidades de los errores tipo I y tipo II, con n = 49.

Tabla 8.7 Evaluación e interpretación de los errores tipo I y tipo II

Riesgo de α	Riesgo de β
• Riesgo de rechazar una hipótesis nula verdadera.	• Riesgo de no rechazar una hipótesis nula falsa.
• Riesgo de detectar una diferencia irreal.	• Riesgo de no detectar una diferencia real.
• Error tipo I.	• Error tipo II.
• Riesgo de decir que un proceso no funciona bien cuando no es así.	• Riesgo de decir que un proceso funciona mal cuando no es así.
• $(1-\alpha)$ probabilidad de aceptar una hipótesis nula verdadera.	• $(1-\beta)$ probabilidad de rechazar una hipótesis nula falsa.
• $(1-\alpha)$ probabilidad de llamar bueno a un producto bueno.	• $(1-\beta)$ probabilidad de llamar malo a un producto malo.
• $(1-\alpha)$ confianza de la prueba.	• $(1-\beta)$ potencia de la prueba.

Resumen: prueba de H_0 usando el estadístico $\overline{X}$

1. Planteamiento de las hipótesis:

   a) Hipótesis nula $H_0 : \mu = \mu_0$

   b) Hipótesis alternativa $H_1 : \mu > \mu_0$

2. Considerando un valor para α (probabilidad del error tipo I). Así:

$$\overline{x}_c = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}}$$

3. Se usó el estadístico de prueba $\overline{X}$ y su valor para la muestra tomada es $\overline{x}_m$, se comparan $\overline{x}_m$ y $\overline{x}_c$ Se rechaza H_0 si $\overline{x}_m > \overline{x}_c$.

4. Se concluye de acuerdo a lo que se desea verificar.

Prueba de H_0 usando el estadístico estandarizado Z

El procedimiento equivalente es usar como estadístico de prueba la variable Z de una normal estándar y su valor punto crítico es $z_c = z(1-\alpha)$

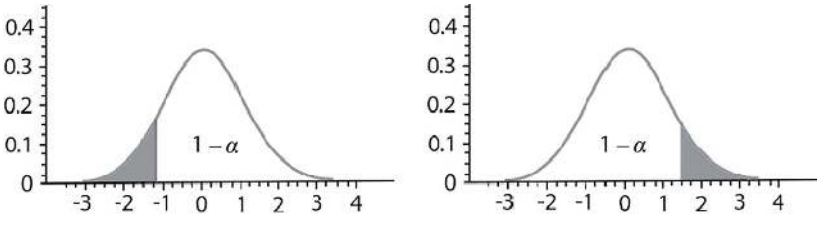
Se rechaza H_0 si $z_m > z(1-\alpha) = z_c$. (Donde $z_m = \frac{\overline{x}_m - \mu_0}{\frac{\sigma}{\sqrt{n}}}$)

Prueba de H_0 usando el $valor - p$

Se rechaza H_0 si $valor - p = P(\overline{X} \geq \overline{x}_m) < \alpha$ caso variable original, o $valor - p = P(Z \geq z_m) < \alpha$ para la variable estandarizada.

Observación: así para un valor $\overline{x}_m$ de $\overline{X}$ y un valor z_m de Z se sigue que:

$$\overline{x}_m = \mu_0 + z_m\frac{\sigma}{\sqrt{n}} > \overline{x}_c = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}}$$



α	$z(\alpha)$	$1-\alpha$	$z(1-\alpha)$
0.1	-1.282	0.90	1.282
0.05	-1.645	0.95	1.645
0.025	-1.96	0.975	1.96
0.01	-2.326	0.99	2.326
0.005	-2.576	0.995	2.576

Figura 8.9 Valores que usualmente se proponen para α, y los correspondientes a z

Solución mediante el uso de CalEst

Aquí se reproduce la solución dada para el ejemplo 8.1, para ello se retoman las opciones visuales del Calculador Estadístico, así se ha elaborado un mecanismo que simula la metodología de prueba de hipótesis. La idea de este proyecto es permitir que los usuarios desarrollen habilidades que les brinden la facilidad de aprender estos conceptos de prueba de hipótesis.

La figura 8.10 transcribe los resultados explicados en el ejemplo 8.1, los cuales se van a exponer aquí mediante una serie de pasos de la función del simulador.

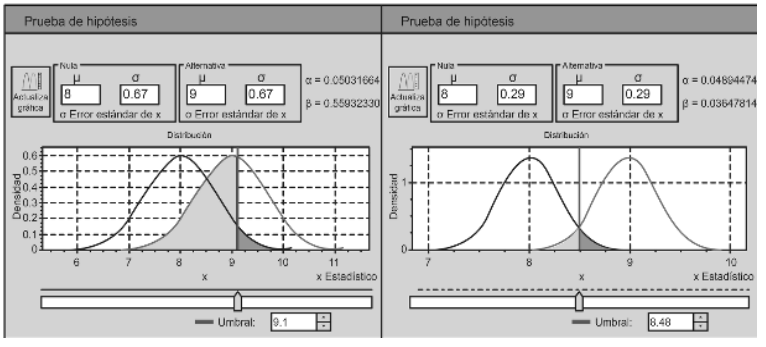


Figura 8.10 Opción visual que simula la prueba de hipótesis de la media.

Para tener el simulador, recurra a la opción didáctica en el CalEst: motivación para prueba de hipótesis indicada con el símbolo $\mathcal{H}_0$. Se sugiere utilizar esta opción para aplicarla a diferentes escenarios y con ello asimilar los conceptos de prueba de hipótesis.

1. Seleccione la opción $\mathcal{H}_0$, el título es Prueba de hipótesis. Ahí aparecerán dos distribuciones de probabilidad normal, una representa la hipótesis nula con media $\mu = 0$ y el error estándar $\sigma = 1$, y la otra representa una de las posibles hipótesis alternativas, con media $\mu = 1$ y el error estándar $\sigma = 1$. Nota: como en este proceso de simulación se especifica la hipótesis alternativa, ésta es referida como hipótesis simple.

2. Plantee una prueba escribiendo los valores para los parámetros que caracterizan a las hipótesis nula y la alternativa, respectivamente.

   a) Para el ejemplo se escribe las medias y el error estándar para cada hipótesis, así las medias son $\mu_0 = 8$ y $\mu_1 = 9$ y el error estándar $\sigma = 0.67$.

   b) Luego se coloca el umbral donde se obtenga el valor del nivel de significancia, es decir $\alpha = 0.05$. En este caso en $\overline{x}_m > \overline{x}_c = 9.1$. Esta situación se observa en la gráfica a la izquierda de la figura 8.10. La discusión se realizó en el contexto de la solución operativa clásica. Observe que en el extremo superior izquierdo aparecen el cálculo de las probabilidades de los errores tipo I (α) y tipo II (β).

3. Considere diferentes situaciones aumentando el tamaño de la muestra que influye en el error estándar y comente los resultados. Así como al considerar otras hipótesis alternativas, esto da lugar a plantearse varias preguntas ¿habrá la mejor hipótesis alternativa? ¿Existe algún criterio para saber cuándo establecer una alternativa? ¿Se puede calcular en términos de probabilidad?

En la solución del ejemplo 8.1, se examinó el caso de un tamaño de muestra diferente observe la gráfica derecha de la figura 8.10.

8.3 Prueba de hipótesis para una media: muestras grandes

Por el trabajo presentado hasta aquí, se está en la posibilidad de presentar una serie de ejemplos sobre pruebas de hipótesis, inicialmente, para la media. La finalidad es ilustrar situaciones sobre posibles aplicaciones y que completen conceptos a cerca de la prueba de hipótesis. El camino es considerar primero muestras de tamaño grande donde se emplea la distribución normal.

Ejemplo 8.2

Un administrador lleva una serie de modificaciones en los procesos de gestión con el fin de reducir el tiempo en que los clientes realizan un trámite, ya que actualmente en su dependencia el tiempo para realizar dicho trámite administrativo tiene una media de 17 días con una desviación estándar de 3 ($\sigma = 3$). Toma una muestra de tamaño $n = 50$ para verificar si sus cambios resultan efectivos, con $\overline{x}_m = 15.9$. La variable aleatoria X : el tiempo que toma realizar el trámite, ésta tiene una distribución normal con media μ y varianza σ^2.

Solución operativa clásica

El primer paso es interpretar lo que el administrador se propone con las reformas, esto es, reducir el actual tiempo de trámite, es decir $\mu < 17$. A continuación presenta el planteamiento hipotético y los pasos para efectuar la prueba de hipótesis, esto es:

1. Hipótesis:

$$H_0 : \mu = 17$$
$$H_1 : \mu < 17$$

2. Para un nivel de significancia $\alpha = 0.025$ ($z(\alpha) = z(0.025) = -1.96$). Entonces, en el supuesto de que la hipótesis nula H_0 es verdadera, el valor del punto crítico es:

$$\overline{x}_c = \mu + z(\alpha)\frac{\sigma}{\sqrt{n}} = 17 - 1.96\frac{3}{\sqrt{50}} = 16.17$$

Este valor corresponte al punto de referencia para rechazar o no rechazar la hipótesis nula H_0.

3. A partir de la información de la muestra, se calcula la media $\overline{x}_m$ y se rechaza la hipótesis nula H_0,

$$\text{si } \overline{x}_m < \overline{x}_c$$

Puesto que $15.9 < 16.17$, se rechaza H_0.

4. Se reduce el proceso de gestión a menos de 17.

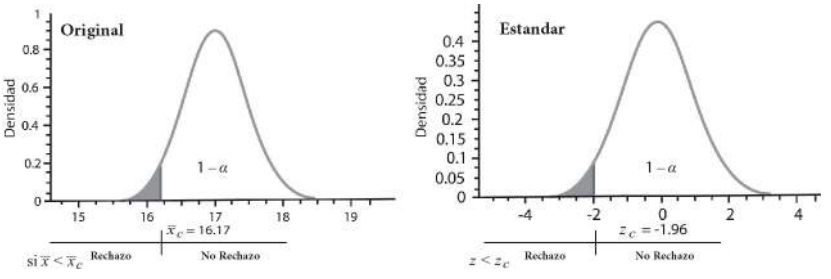


Figura 8.11 Descripción gráfica de la prueba de hipótesis en el ejemplo 8.2.

En la figura 8.11 se describe este proceso de la prueba de hipótesis para el ejemplo, a la izquierda está el planteamiento para los valores originales de la variable X y a la derecha está el proceso estandarizado, que equivale decir rechazar la hipótesis nula H_0, si $z_m < z_c = -1.96$. Donde:

$$z_m = \frac{15.9 - 17}{3/\sqrt{50}} = \frac{-1.1}{0.424} = -2.594$$

Ya que $z_m = -2.594 < z_c = -1.96$ se concluye de manera similar. En un ejemplo posterior se mostrarán los cuatro pasos del procedimiento de prueba usando el estadístico Z.

Note en las gráficas de la figura 8.11 que el punto crítico (umbral) divide a la región en dos partes; a éstas se les conoce como *región de rechazo y no rechazo.*

Potencia de la prueba

Una vez que se tienen las probabilidades de los errores tipo I y II, el complemento de la probabilidad de β se le refiere como *potencia de la prueba*, esto es $p(\mu) = 1 - \beta$.

> La potencia $p(\mu)$ de la prueba de H_0 es la probabilidad de rechazar H_0 cuando el verdadero valor del parámetro es μ.

La potencia de la prueba para la alternativa $H_1 : \mu < \mu_0$ es la probabilidad que se obtiene mediante la siguiente expresión.

$$p(\mu) = 1 - \beta = P\left(\overline{X} \leq \mu_0 + z(\alpha)\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1\right) \tag{8.8}$$

Aplicando la fórmula (8.8) al ejemplo, se puede calcular la potencia de la prueba para la $\mu = 15.6$, $\overline{x}_c = \mu_0 + z(\alpha)\frac{\sigma}{\sqrt{n}} = 17 - 1.96(0.424) \doteq 16.17$

$$p(15.6) = 1 - \beta = P\left(\overline{X} \leq 16.17 \mid \mu = \mu_1\right) = 0.91 \tag{8.9}$$

En la figura 8.12 la potencia expresada para la variable de la normal estándar Z, $P(Z \leq 1.341) = 0.91$

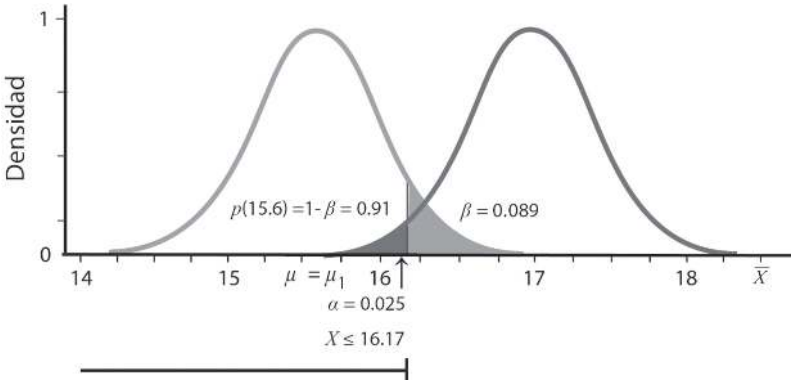


Figura 8.12 Presentación visual del cálculo de la potencia de la prueba μ = 15.6.

Tamaño de muestra

En esta parte se presenta el procedimiento para calcular el *tamaño de muestra* en la prueba de hipótesis. La importancia de ésta, entre otros puntos, es que indicará qué tamaño se requiere para alcanzar una determinada potencia de la prueba.

Así bajo H_0 se tiene que $P(\overline{X} \geq \overline{x}_c) = \alpha$, $(P(Z \geq z(\alpha)) = \alpha)$ para una región crítica de tamaño α. La probabilidad del error tipo II para la alternativa μ_1 escrita con el punto $\overline{x}_c = \mu_0 + z(\alpha)\frac{\sigma}{\sqrt{n}}$, se expresa por:

$$\beta = P(\overline{X} \geq \overline{x}_c = \mu_0 + z(\alpha)\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1)$$

También se puede expresar por:

$$\beta = P(\overline{X} \geq \overline{x}_c = \mu_1 + z(1-\beta)\frac{\sigma}{\sqrt{n}} \mid \mu = \mu_1)$$

Observando que $z(\alpha) = -z(1-\alpha)$, por lo tanto $\mu_0 - z(1-\alpha)\frac{\sigma}{\sqrt{n}} = \mu_1 + z(1-\beta)\frac{\sigma}{\sqrt{n}}$, de aquí sigue que:

$$n = \frac{(z(1-\alpha) + z(1-\beta))^2\sigma^2}{(\mu_0 - \mu_1)^2}$$

Para el caso en que $\alpha = 0.025$, $\beta = 0.05$, considerando una diferencia de 1 entre las medias $d = \mu_0 - \mu_1 = -1$, $\sigma = 3$. Sustituyendo se obtiene:

$$n = \frac{(1.96 + 1.645)^2 3^2}{(-1)^2} \doteq 117$$

Caso muestras grandes

Ejemplo 8.3

Caso prueba bilateral o de dos colas. Los procesos industriales tienen una componente administrativa y económica muy importante. Los procesos de llenado son muy comunes y resultan un buen ejemplo para mostrar la hipótesis alternativa de diferencia, porque si una empresa da más producto pierde en su utilidad, o puede ser multada por una autoridad si da menos. Por ejemplo, en un proceso de llenado, el peso de bolsas de uvas pasas no debe ser ni más ni menos de 336 gramos. Para verificar que el proceso cumple con esta especificación, el responsable del proceso toma una muestra de 40 bolsas de un lote de producción. La media del peso de estas bolsas fue de $\overline{x}_m = 329.91$-valor en la muestra-, con una desviación estándar de $s = 11$. Nota: no se proporciona el valor de σ, pero dado que el valor de la muestra es grande, se usará el valor de s como un estimado; así $\sigma = 11$. La variable aleatoria X es el peso de llenado de las bolsas, Planteadas estas condiciones por el resultados del TLC visto en el capítulo 7, entonces $\overline{X}$ se toma como distribución de probabilidad normal.

Solución operativa clásica

Por pasos:

1. Las hipótesis que se plantean para describir esta situación son:

$$H_0 : \mu = 336$$
$$H_1 : \mu \neq 336$$

   Observe que en este caso el que la hipótesis alternativa sea $H_1 : \mu \neq 336$, da la posibilidad de que $\mu < 336$, o que $\mu > 336$. Como se ha dicho con anterioridad, la prueba de hipótesis es de las *llamadas bilaterales* o de dos colas.

2. Para verificar, si los datos apoyan la hipótesis nula con un nivel de significancia de $\alpha = 0.05(5\,\%)$. En este caso, al tener una hipótesis bilateral, el valor de α se parte en dos, una para la cola izquierda

y la otra para la cola derecha, por consiguiente el nivel de significancias es $\frac{\alpha}{2}$. Así $z_{ci} = z(\alpha/2) = z(0.025) = -1.96$, y $z_{cd} = z(1-\alpha/2) = z(0.975) = 1.96$.

3. Se tienen dos estadísticos de prueba, uno para la alternativa $\mu < 336$ y otro para la alternativa $\mu > 336$. De esa manera se tienen dos puntos críticos, estos son:

$$\overline{x}_{ci} = \mu_0 + z(\alpha)\frac{\sigma}{\sqrt{n}}, \text{ el otro es } \overline{x}_{cd} = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}} \tag{8.10}$$

Con la información los valores de estos son:

$$\overline{x}_{ci} \triangleq 332.59 \text{ y } \overline{x}_{cd} \triangleq 339.41$$

El procedimiento indica que se debe comparar el valor del estadístico $\overline{x}_m = 329.91$, que en este caso resultó el más pequeño, con el valor crítico $\overline{x}_{ci} = 332.6$. Se sigue que $\overline{x}_m < \overline{x}_{ci}$, por lo que se rechaza la hipótesis nula. Nota: Puede ocurrir que $\overline{x}_m$ sea el valor más grande a la derecha, entonces se rechaza H_0, si $\overline{x}_m > \overline{x}_{cd}$

4. La muestra indica que el llenado de bolsas de pasas tiene un peso menor al establecido.

Estos cuatro puntos se muestran con la gráfica de la figura 8.13.

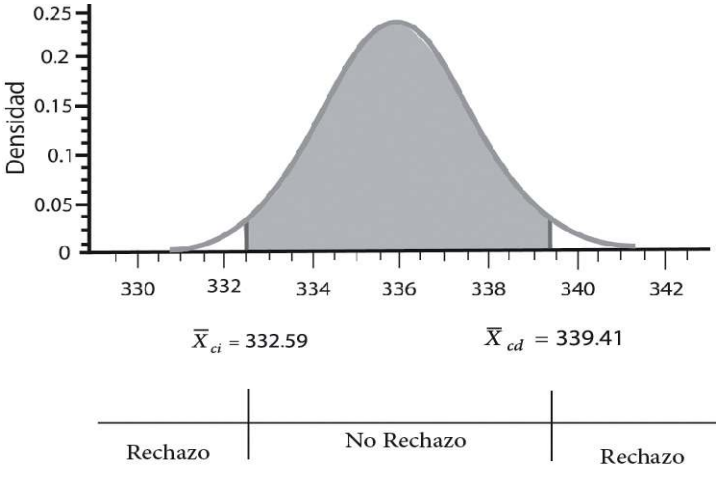


Figura 8.13 Descripción de la prueba de hipótesis cuando la alternativa es $H_1 : \mu \neq \mu_0$, ambos lados.

Prueba aplicando la distribución normal estándar

El procedimiento usando la normal estándar es el siguiente.

1. Las hipótesis que se plantean para describir esta situación son:

$$H_0 : \mu = 336$$
$$H_1 : \mu \neq 336$$

2. Reproduciendo el punto 2 del ejemplo 3, se escriben los valores críticos usando la distribución normal estándar con $\alpha = 0.05$, se tiene que $z_{ci} = z(\alpha/2) = z(0.025) = -1.96$, y $z_{cd} = z(1-\alpha/2) = z(0.975) = 1.96$

3. Bajo el supuesto de que la hipótesis nula es verdadera, se realiza el cálculo del estadístico de prueba considerando la normal estándar:

$$z_m = \frac{\overline{x}_m - \mu}{\sigma/\sqrt{n}} = \frac{329.91 - 336}{11/\sqrt{40}} = -3.5$$

   Se compara este valor con uno de los valores de la distribución estándar para $\alpha = 0.05$. Se ve que se satisface $z_m = -3.5 < -1.96$ por lo tanto se rechaza la hipótesis nula.

4. La conclusión es la misma con los datos originales.

Se puede reproducir en **CalEst** la información que se ha usado para realizar la prueba de hipótesis, esto es, aprovechando el efecto visual y la facilidad de cálculo de la distribución normal original y estándar estudiadas en el capítulo 6. Con la finalidad de completar esta consideración, en la figura 8.14 se muestra la gráfica generada por **CalEst**, tal y como aparece en la calculadora. En la opción de distribuciones se emplea la normal; cuando se abre esta elección, la distribución normal estándar es la que aparece inicialmente.

Entre los valores de los *umbrales* (puntos críticos) se tiene la región de no rechazo. Luego, a la izquierda del umbral 1, y a la derecha del umbral 2 se tiene la región de rechazo. Observe que los valores de los umbrales corresponden a los valores de $z_{ci} = -1.96$ y $z_{cd} = 1.96$. El área sombreada corresponde al nivel de significancia, $\alpha/2 = 0.025$ a la izquierda, y $\alpha/2 = 0.025$ a la derecha. Se añade en la figura 8.14 la tabla denominada Calculadora que da con mayor precisión el valor de los valores de una variable que tenga una distribución de probabilidad normal; además ésta es de utilidad para otras aplicaciones y ejercicios.

Dado que $z_m < z_{ci}$, se rechaza H_0. Conclusión: esta situación indica que las bolsas de uvas pasas dan menos del peso que se espera. Por lo tanto habrá que estudiar el proceso para lograr que las bolsas den el peso.

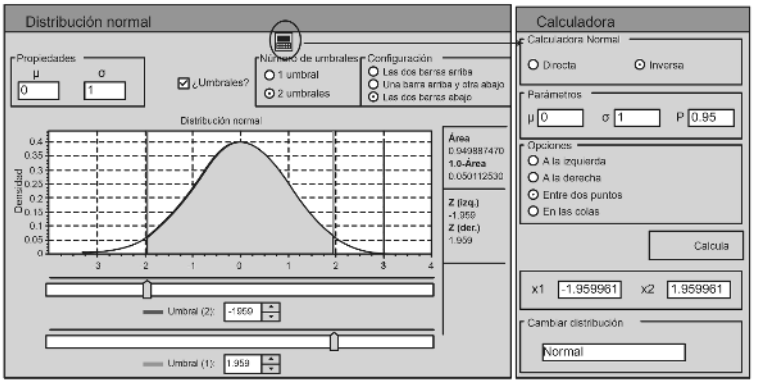


Figura 8.14 Descripción de los valores críticos en una prueba de hipótesis bilateral.

Relación entre el intervalo de confianza y la prueba de hipótesis bilateral

Nuevamente recurriendo a la expresión (8.10), y con la información propuesta del ejemplo, se puede construir el intervalo de confianza de $(1-\alpha)\,\%$. En este ejemplo se establece el de $95\,\%$ de confianza, esto es:

$$(\overline{x}_m + z(\alpha/2)\tfrac{\sigma}{\sqrt{n}}, \overline{x}_m + z(1-\alpha/2)\tfrac{\sigma}{\sqrt{n}})$$

$$(329.91 - 1.96(1.74), 329.91 + 1.96(1.74))$$

Se reporta que el intervalo de confianza de $95\,\%$ para la media μ es:

$$(326.4996, 333.3204)$$

Interpretación: se observa que este intervalo no contiene a la media $\mu = 336$; se concluye diciendo que se rechaza la hipótesis nula, lo que confirma lo dicho por la prueba de hipótesis.

Solución mediante el uso de CalEst

Se ha expuesto la metodología de la prueba de hipótesis para una media, por el momento para muestras grandes. Se han ilustrado los conceptos estadísticos y de probabilidad básicos en este procedimiento, así como, la parte operativa para efectuar la prueba. El detalle operativo se puede transcribir usando el CalEst y obtener la ganancia visual del proceso de un prueba, que ilustra con claridad los errores tipo I y tipo II mediante la generación de gráficas.

En este caso hay que recurrir a la opción: Inferencia, luego Prueba de hipótesis μ_1, continuar para el caso una media y finalmente el estadístico Z. La indicación para acceder a esta opción se muestra en el cuadro como se presenta en la figura 8.15. Una vez que se han seleccionado las opciones se tiene un cuadro en el que hay que escribir la información proporcionada por los datos muestrales. Una vez completa la información, se oprime el botón Aceptar y se despliega la información desarrollada en el ejemplo 8.3.

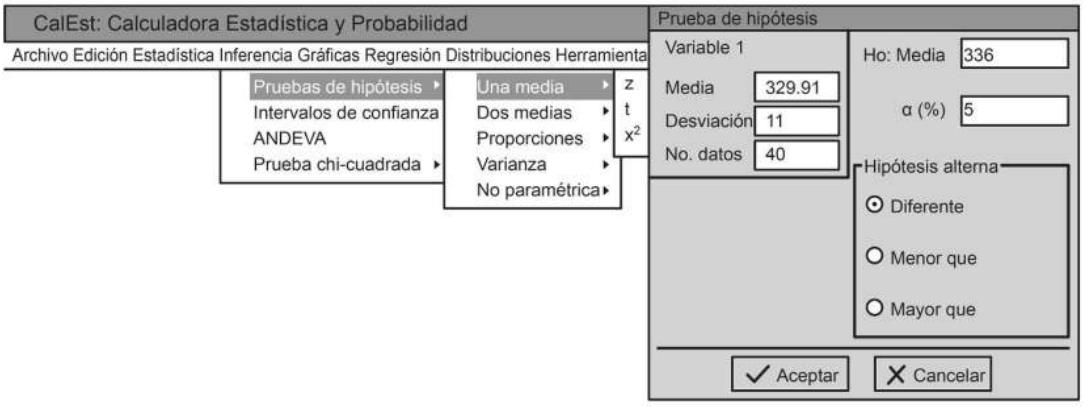


Figura 8.15 Descripción del procedimiento para realizar la prueba de hipótesis para una media usando **CalEst**.

Finalmente el reporte se obtiene mediante la salida de los resultados mostrados por **CalEst**, como se distingue en la figura 8.16. Se puede observar que en este informe se extraen los elementos que son básicos para la prueba de hipótesis y su conclusión. Así, se generan: el estadístico de prueba calculado y los valores críticos; con esta descripción se puede concluir si se rechaza o no se rechaza la hipótesis nula. En la figura 8.16 se destaca la parte visual del material didáctico, tal y como se describe en la figura 8.13. A la derecha se ilustra la gráfica de la distribución normal; sin necesidad de pasar por el proceso de estandarizar en la normal, se ven los valores críticos. Así se tiene que la media $\overline{x}_m = 329.91$ es menor que el valor crítico $\overline{x}_c = 332.6$ y sigue la conclusión ya anotada. La figura 8.14, describe el procedimiento de la prueba de hipótesis considerando la normal estándar.

Prueba de la hipótesis H_0 mediante la comparación de probabilidades

Se puede observar que en el cuarto renglón de resultados mostrados en la figura 8.16, aparece un valor de p ($valor - p$) éste corresponde a la probabilidad que deja a la izquierda el estadístico calculado, recuerde que, es el denominado *nivel de significancia descriptivo*, $valor - p$. En símbolos: $P(Z \leq -3.5) = 0.00046$. Esta probabilidad se compara con el *nivel de significancia* α, ésta es otra alternativa para decidir sobre la hipótesis nula. Por lo tanto:

$$\text{si } valor - p < \alpha, \text{ se rechaza } H_0$$

Note que en el caso de las pruebas bilaterales se tiene $\frac{\alpha}{2}$, en ese sentido se tiene el valor correspondiente de p para los dos lados.

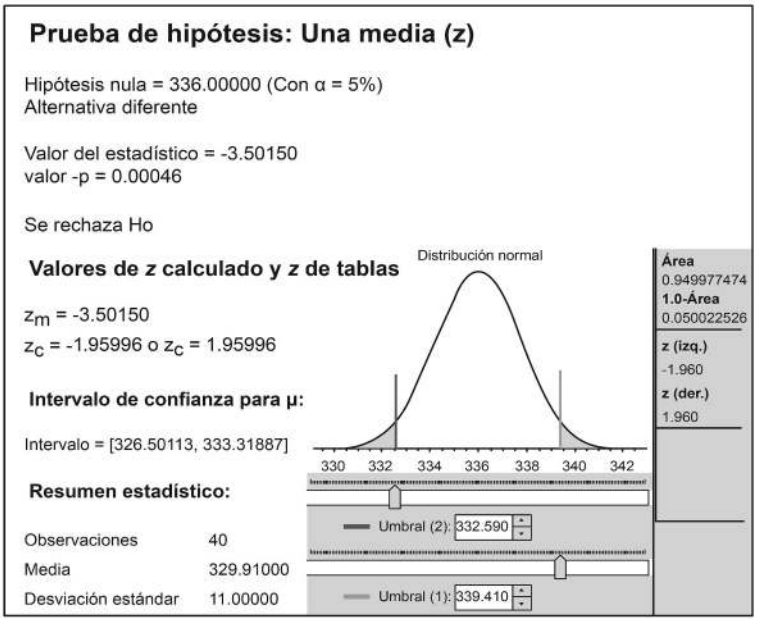


Figura 8.16 Reporte de la salida de la prueba de hipótesis para una media, generada por **CalEst.**

Ejemplo 8.4

Con la finalidad de tener un mayor número de utilidades en un proceso de curtiduría, la administración plantea que el índice de elongación de una piel, debe estar arriba de 64 unidades. En una muestra de $n = 45$ pieles, la media del índice de elongación es de $\overline{x}_m = 65.5$ y desviación estándar de $S = 3.5$. Considerando que el tamaño de muestra es suficientemente grande, en este caso se usará σ por S. El nivel de significancia que se propone es $\alpha = 0.05$

Solución

Siguiendo las etapas propuestas para realizar la prueba de hipótesis:

1. Las hipótesis correspondientes en este ejemplo son:

$$H_0 : \mu = 64$$
$$H_1 : \mu > 64$$

2. El nivel de significancia es $\alpha = 0.05$, de esta manera la región de rechazo se establece considerando el punto crítico:

$$\overline{x}_c = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}} = 64 + 1.645\frac{3.5}{\sqrt{45}} \doteq 64.858$$

3. A continuación, se compara el valor del estadístico $\overline{x}_m = 65.5$ con el valor crítico $\overline{x}_c \doteq 64.858$, se observa que:

$$\overline{x}_m = 65.5 > \overline{x}_c \doteq 64.858$$

Verifique con la gráfica de la figura 8.17.

4. En efecto, los datos dan evidencia de que se incrementan las ganancias de la curtiduría, ya que se rechaza H_0. La elongación de la piel supera las 64 unidades.

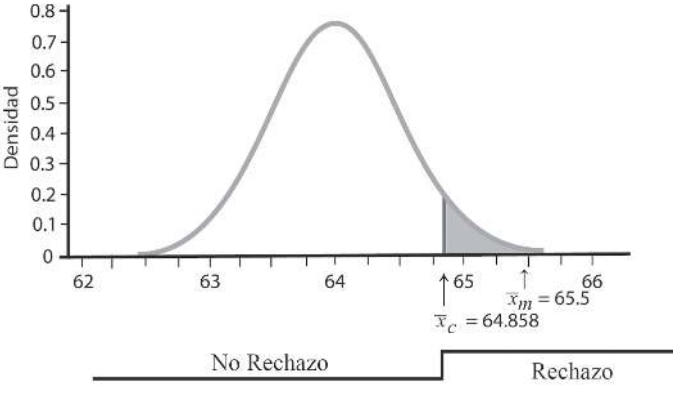


Figura 8.17 Prueba de hipótesis para H_0: $\mu = 64$ vs $H1 : \mu > 64$.

Solución con la variable Z mediante el uso de CalEst

Pruebe la hipótesis siguiendo el proceso usando la normal estándar; esto se transcribe en la figura 8.18. Observe que $z_m = 3.067 > z_c = 1.645$. Los datos no dan evidencia para apoyar a la hipótesis nula, por lo tanto se rechaza y efectivamente el índice de elongación está por arriba de 64 unidades.

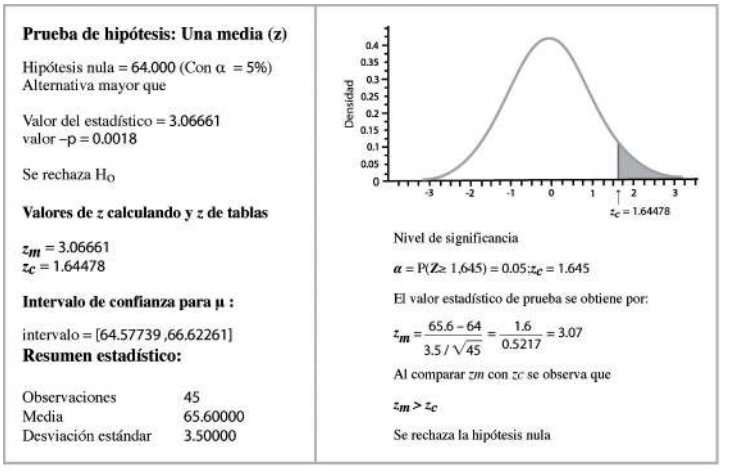


Figura 8.18 Descripción de la prueba de hipótesis $H_0 : \mu = 64$ vs $H_1 : \mu > 64$ usando la normal estándar.

Ilustración de la prueba de hipótesis mediante la comparación de probabilidades. Se interpreta el resultado considerando el nivel de significancia descriptivo $valor - p$, observe que el valor crítico $z_c = 3.07$, se calcula la probabilidad de que este valor sea mayor que 3.07. Así:

$$valor - p = P(Z \geq 3.07) = 0.00107$$

Entonces se confirma la decisión planteada, se rechaza H_0 porque $p < \alpha = 0.05$. La información proporcionada por esta muestra permite concluir que las pieles tienen una elongación mayor a 64.

Intervalo de confianza del $95\,\%$ para la media μ

Considerando los conceptos planteados en el capítulo 7, alternativamente se puede estimar el intervalo de confianza $(1 - \alpha)\,\%$, la expresión es:

$$(\overline{x}_m + z(\alpha/2)\frac{\sigma}{\sqrt{n}}, \overline{x}_m + z(1 - \alpha/2)\frac{\sigma}{\sqrt{n}})$$

Si $\alpha = 0.05$, entonces el intervalo de confianza del $95\,\%$ de confianza es:

$$(65.6 - 1.96(.5217), 65.6 + 1.96(.527)) = (64.577, 66.623)$$

Conclusión, con un $95\,\%$ de confianza la media estará contenida entre los valores de 64.58 y 66.23. Observe que este intervalo no contiene a la media $\mu = 64$ que caracteriza la hipótesis nula.

.

8.4 Prueba de hipótesis para una media: muestras pequeñas

Se han expuesto los conceptos básicos de la prueba de hipótesis, en este apartado se extienden estas ideas al caso de muestras pequeñas. En esa dirección la distribución que se emplea es la t de Student, que se estudió en el capítulo 6 y que se recomienda repasar para recordar las propiedades de la t. La aplicación de esta distribución a la prueba de hipótesis es importante, ya que se puede decir que es más práctica en el estudio de problemas reales, pues no requiere el conocimiento de la varianza σ^2. Se apoya esencialmente en la información que proporciona la muestra de la población. El supuesto relevante en su aplicación es que la variable aleatoria de interés debe seguir una distribución normal o aproximadamente simétrica, conjetura que habrá de confirmarse, también, tomando en cuenta la información de la muestra.

Recuerde que un punto central en la prueba de hipótesis es tener el punto crítico que será la referencia para contrastarlo con el estadístico de prueba. Se da una guía para encontrar los valores críticos en una distribución $t-$Student.

1. Identificar el nivel de significancia α
2. Identificar los grados de libertad, $gl : n-1$
3. Usar la distribución t de Student. Esta se aplica considerando las siguientes situaciones de la prueba de hipótesis:

   a) Al lado izquierdo, α. El valor crítico es: $t_{ci} = t(n-1, \alpha)$.

   b) Al lado derecho, $1-\alpha$. El valor crítico es: $t_{cd} = t(n-1, 1-\alpha)$.

   c) Bilateral (en este caso se ve señalado el lado izquierdo, use $\alpha/2$, el valor crítico es: $t_{ci} = t(n-1, \alpha/2)$ o lado derecho con $1-\alpha/2$, el valor crítico es: $t_{cd} = t(n-1, (1-\alpha/2))$.

Resumen de la prueba de hipótesis usando la t-Student. Una vez que se han comprendido los principios que justifican la prueba de hipótesis, en las siguientes tres figuras, (8.19, 8.20 y 8.21) se hace un resumen de la prueba de hipótesis para cada una de las situaciones indicadas. Estas figuras representan la distribución $t-Student$, con $gl = 9$, $\alpha = 0.05$, así para el lado izquierdo: $t_{ci} = t(9, 0.05) = -1.83426$, el lado derecho: $t_{cd} = t(9, 0.95) = 1.83426$, y el caso bilateral: $t_{ci} = t(9, 0.025) = -2.264$, $t_{cd} = t(9, 0.975) = 2.264$. La relación entre la media de la muestra y la distribución $t - Student$ es:

$$t_m = \frac{\overline{x}_m - \mu_0}{\frac{S}{\sqrt{n}}}$$

En la opción distribuciones del CalEst seleccione la distribución $t-$Student.

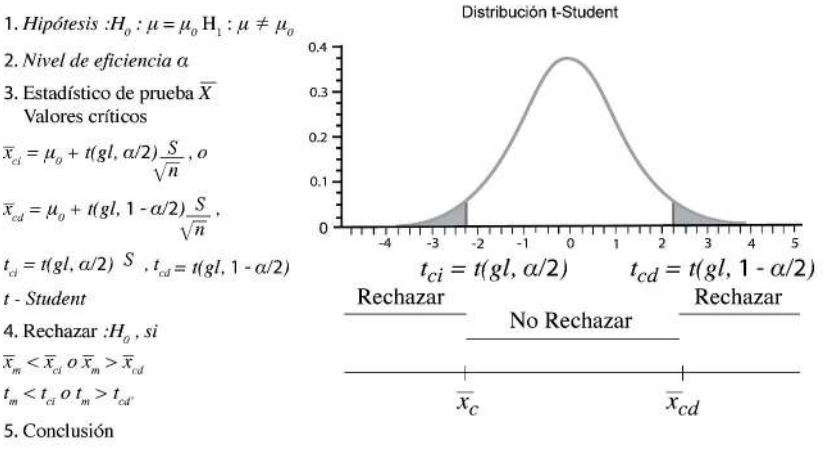


Figura 8.19 Prueba de hipótesis para muestras pequeñas cuando la alternativa es menor.

Planteamiento hipotético, muestras pequeñas

En esta parte se harán, nuevamente, los tres planteamientos de hipótesis para la media. Sin embargo, el procedimiento de la prueba de estas hipótesis se basa en el estadístico $t - Student$.

1. Prueba de hipótesis cuando $H_0 : \mu = \mu_0$ vs $H_1 : \mu < \mu_0$. El resumen de la descripción de esta prueba de hipótesis se presenta en la figura 8.19.

2. Prueba de hipótesis cuando $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$. El resumen de la descripción de esta prueba de hipótesis se presenta en la figura 8.20.

3. Prueba de hipótesis cuando $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$.

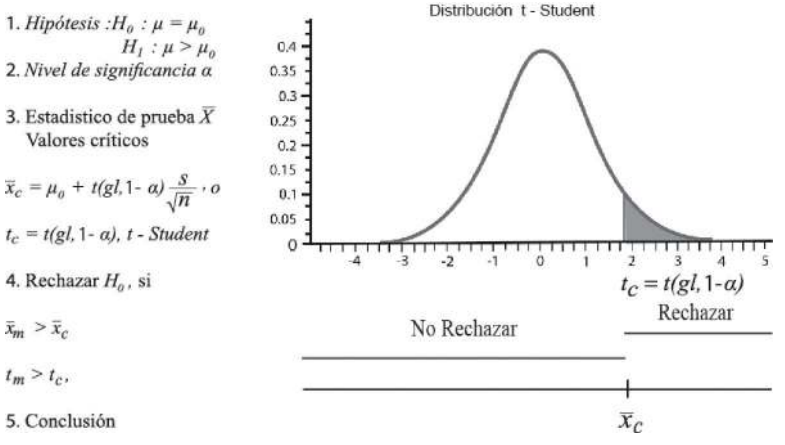


Figura 8.20 Prueba de hipótesis cuando H1: $\mu > \mu_0$

El resumen y la descripción de la prueba de hipótesis bilateral, se presenta en la figura 8.21.

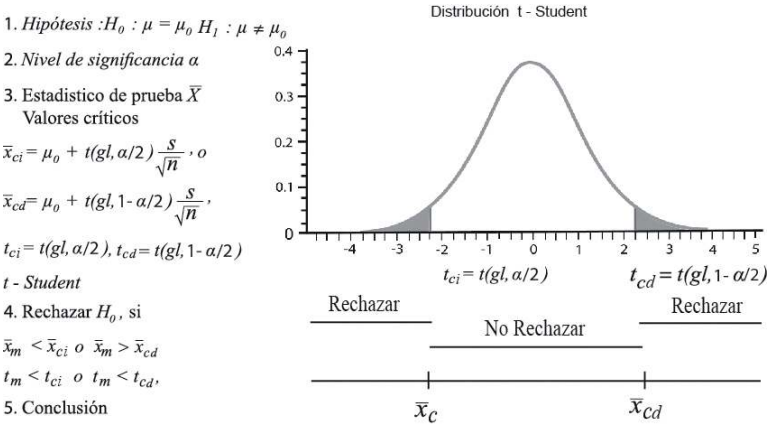


Figura 8.21 Prueba de hipótesis cuando $H_1 : \mu \neq \mu_0$.

Procedimiento de la prueba de hipótesis: regla de decisión

La idea principal de la prueba de hipótesis es comparar dos números, el valor del estadístico de prueba con el valor de una distribución de probabilidad (éste se obtiene a partir de un valor de significancia -probabilidad- establecido por α), considere como punto crítico $\overline{x}_c$:

$$\overline{x}_c = \mu + t(n-1, \alpha)\frac{S}{\sqrt{n}}$$

Se compara el valor del estadístico $\overline{x}_m$ con el punto crítico $\overline{x}_c$, regla de decisión para las tres hipótesis:

1. Si la hipótesis alternativa es $H_1 : \mu < \mu_o$ se rechaza la hipótesis si $\overline{x}_m$ es menor que $\overline{x}_c$, donde $\overline{x}_c = \mu + t(n-1, \alpha)\frac{S}{\sqrt{n}}$.

2. Si la hipótesis alternativa es $H_1 : \mu > \mu_o$ se rechaza la hipótesis si $\overline{x}_m$ es mayor que $\overline{x}_c$, donde $\overline{x}_c = \mu + t(n-1, (1-\alpha))\frac{S}{\sqrt{n}}$.

3. Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$, se rechaza la hipótesis si $\overline{x}_m$ es menor que $\overline{x}_{ci} = \mu + t(n-1, \alpha/2)\frac{S}{\sqrt{n}}$ o mayor que $\overline{x}_{cd} = \mu + t(n-1, (1-\alpha/2))\frac{S}{\sqrt{n}}$.

Considere el proceso de estandarización:

$$t_m = \frac{\overline{x}_m - \mu}{\frac{S}{\sqrt{n}}}$$

Se compara con un valor de la distribución t_c de Student $t(n-1, \alpha)$, donde α es una probabilidad y $n-1$ son los grados de libertad.

1. Si la hipótesis alternativa es $H_1 : \mu < \mu_o$ se rechaza la hipótesis si t_m es menor que $t_c = t(n-1, \alpha)$.

2. Si la hipótesis alternativa es $H_1 : \mu > \mu_o$ se rechaza la hipótesis si t_m es mayor que $t_c = t(n - 1, (1 - \alpha))$.

3. Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$, se rechaza la hipótesis si t_m es menor que $t_c = t(n-1, \alpha/2)$ o mayor que $t_c = t(n - 1, (1 - \alpha/2))$.

- Considere el nivel de significancia descriptivo, el $valor - p$, una alternativa es comparar la probabilidad p que deja el estadístico de prueba (a la derecha o izquierda) con el valor de significancia -probabilidad- α . Análogamente:

1. Si la hipótesis alternativa es $H_1 : \mu < \mu_o$ se rechaza la hipótesis si $valor - p$ es menor que α.

2. Si la hipótesis alternativa es $H_1 : \mu > \mu_o$ se rechaza la hipótesis si $valor - p$ es menor que α.

3. Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$ se rechaza la hipótesis si $2(valor - p)$ es menor que α.

Ejemplo 8.5

En la dulcería de un complejo de cines el tiempo de atención al cliente tarda mucho. Debido a las quejas que recibe la administración han decidido bajar el tiempo a menos de 180 segundos (3 minutos). Después de haber implementado una serie de estrategias para mejorar el servicio, se toma una muestra de 16 clientes y se mide el tiempo que llevó, sólo en ser atendidos, no el tiempo de espera en la fila. Los datos se dan a continuación. Con un nivel de significancia de $\alpha = 0.05$, ¿resultó efectivo el cambio impulsado por la administración?

197	194	176	174	217	186	221	188
211	196	167	238	179	212	191	233

Solución

1. El planteamiento de las hipótesis son:

$$H_0 : \mu = 180$$
$$H_1 : \mu < 180$$

2. El nivel de significancia propuesto para realizar esta prueba es $\alpha = 0.05$, de la tabla de la distribución $t - Student$, el valor del punto crítico correspondiente con 15 grados de libertad es: $t(n - 1, \alpha) = t(15, 0.05) = -1.752$, así $t_c = -1.752$.

3. La información derivada de la muestra indica que $\overline{x}_m = 201.875$, $S = 19.926$, $n = 16$, entonces el punto crítico es:

$$\overline{x}_c = \mu_0 + t(n-1,\alpha)\frac{S}{\sqrt{n}} = 180 - 1.752\frac{19.926}{4} = 171.272$$

Comparando este valor con la media de la muestra, se observa que $\overline{x}_m = 201.875 > 171.272$.

4. Los datos no dan evidencia para rechazar la hipótesis nula. Se concluye que la estrategia puesta en marcha por la administración no resulta eficiente, ya que no disminuyó el tiempo de atención.

La prueba usando el estadístico t_m

Considerando el punto 3, con el estadístico t_m se tiene que su valor es

$$t_m = \frac{\overline{x}_m - \mu}{\frac{S}{\sqrt{n}}} = \frac{201.875 - 180}{\frac{19.926}{\sqrt{16}}} = 4.391$$

Se comparan los valores de t_c y t_m, dado que $t_m = 4.391 > t_c = -1.752$, entonces no se rechaza la hipótesis nula. Se verifica la conclusión anterior.

Nivel de significancia descriptivo

Alternativamente, como el $valor - p = P(t_m \leq -0.44671) = 0.3307$ es mayor que $\alpha = 0.05$ (nivel de significancia) no se rechaza H_0 . Lo que indica que la atención al cliente en ese complejo de cines no está por debajo de 180.

Intervalo de confianza

Con el fin de completar la inferencia de los datos se construye un intervalo del $(1-\alpha)\,\%$ de confianza para una media con respecto al caso de muestras pequeñas.

$$\left(\overline{x}_m + t(n-1,\alpha/2)\frac{S}{\sqrt{n}}, \overline{x}_m + t(n-1,1-\alpha/2)\frac{S}{\sqrt{n}}\right)$$

Para construir un intervalo de $95\,\%$ de confianza, se recurre a la distribución t y se tiene que el valor de $t(n-1,1-\alpha/2) = t(15,0.975) = 2.134$.

$$(201.875 - 2.134(4.982), 201.875 + 2.134(4.982))$$

$$(191.245, 212.503)$$

Note que los valores del intervalo están por arriba de 180, desde luego la estrategia realizada por la administración no tuvo los efectos esperados.

Ejemplo 8.6

Una empresa en economía está interesada en realizar un estudio para estimar el nivel de deuda de los usuarios de tarjetas de crédito, debido al alto interés que cobran las entidades bancarias. El investigador considera que si el pago mínimo por tarjeta de crédito al mes rebasa los mil pesos, significa que las personas tienen una deuda importante acumulada con el banco que expide la tarjeta. Selecciona una muestra aleatoria de 18 clientes de una institución bancaria, así $n = 18$; los datos redondeados a pesos se presentan a continuación. Como se desea tener un error del tipo I pequeño, se toma un nivel de significancia de $\alpha = 0.025$.

1237	1162	943	1093	1109	1143	1204	1198	944
951	1370	1163	1005	929	1028	1151	1132	1207

Solución mediante el uso de CalEst

Esta solución se presenta mediante el empleo de **CalEst**, el que se ha utilizado para realizar los cálculos en este trabajo. En la figura 8.22 se muestra la opción en inferencia estadística para la prueba t. Se muestran las posibilidades para realizar la prueba. No obstante se pueden seguir los pasos indicados en el procedimiento de una prueba de hipótesis. La salida de este apoyo tecnológico facilita el cálculo y proporciona una ayuda visual para comprender la prueba. Sin embargo, éste no sustituye el conocimiento ni la creatividad que se requiere en el planteamiento de un problema, ni el dominio y compresión de la metodología de prueba de hipótesis.

Hipótesis

$$H_0 : \mu = 1000, \quad H_1 : \mu > 1000$$

Procedimiento: en resumen, se rechaza H_0 puesto que $t_m > t_c$, o el $valor - p = 0.00067 < \alpha$.

Conclusión: Al rechazar la hipótesis nula, los datos dan evidencia de la significancia estadística para considerar que el pago mínimo al mes de un tarjeta habiente es superior a 1000 pesos.

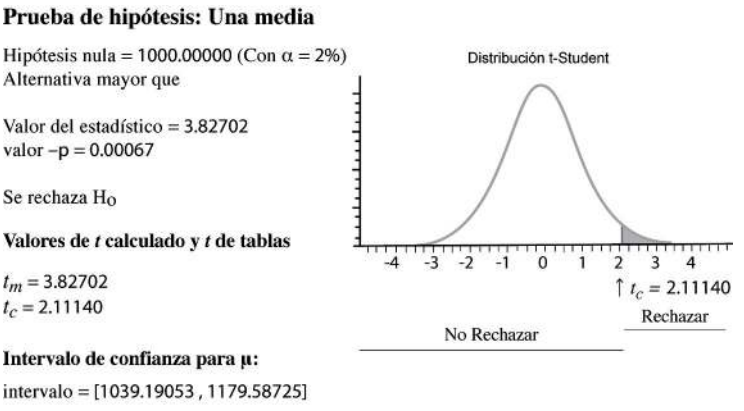


Figura 8.22 Reporte de la salida de la prueba de hipótesis usando **CalEst**.

8.5 Prueba de hipótesis para una proporción

Las pruebas de hipótesis sobre proporciones se emplean en muchas actividades de estudio, prácticas e investigación relacionadas con la administración y la economía. Para fijar ideas, aquí se presentan unos ejemplos.

1. En referencia a la economía, el hábito de fumar impacta en el gasto diario y ocasiona un egreso, ya que a la larga tiene efectos en la salud. Además acarrea otros problemas, entre ellos las limitantes que ponen los seguros a los fumadores. En el contexto de un estudio se piensa que entre el grupo de fumadores, la media del número de cigarros que fuma una persona al día es de 7, lo que representa un aproximado diario de 14 pesos. Así, se considera que:

   - El porcentaje de personas que fuman menos de 7 cigarros al día es de 44 %.

2. La administración de muchas empresas siempre tiene el interés de evaluar el impacto que poseen sus productos en la preferencia de las personas. En este sentido, se desea conocer si las campañas publicitarias tienen un efecto favorable entre los consumidores. Para ello se lanza un producto después de una campaña y se presume que la campaña fue exitosa si:

   - El porcentaje de personas que aprueban el producto es mayor al 25 %.

Los elementos que intervienen en un ejemplo como éstos son:

*É*xito=E, la persona seleccionada está a favor de la propuesta.

Falla=F, la persona seleccionada no está a favor de la propuesta.

$P(E) = p$, la proporción de los participantes a favor de la propuesta.

$P(F) = q = 1 - p$, la proporción de los participantes en contra de la propuesta.

N : Tamaño de la población.

n : número de pruebas, es decir, número de participantes en la muestra.

X : número de elementos favorables en la población.

x : número de elementos favorables en la muestra.

$p = \frac{X}{N}$ es estimado por $\widehat{p} = \frac{x}{n}$

La proporción en una población se obtiene mediante $p = \frac{X}{N}$. La variable aleatoria X sigue una distribución binomial con parámetros (n, p). *Se requieren de las condiciones* $np \geq 5$ y $n(1-p) \geq 5$ *para una distribución binomial se aproxime a una normal.* Entonces la distribución muestral para $\widehat{p}$ es normal con parámetros:

$$\mu_{\widehat{p}} = \mu(\widehat{p}) = p \ \text{ y } \ \sigma_{\widehat{p}} = \sigma(\widehat{p}) = \sqrt{p(1-p)/n}$$

La expresión estandarizada que depende de la muestra es:

$$z_m = \frac{\widehat{p}_m - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{\widehat{p}_m - p}{\sqrt{p(1-p)/n}}$$

Los intervalos de confianza $(1-\alpha)\,\%$ se calculan de acuerdo con el valor de z, normal estándar.

$$\left(\widehat{p}_m + z(\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p})/n}), \widehat{p}_m + z(1-\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p}/n})\right)$$

Con esta síntesis de conceptos se puede exponer las expresiones para realizar una prueba de hipótesis para una proporción; éstas se plantean como sigue:

$$H_0 : p = p_0$$
$$H_1 : p < p_0, \text{ o } H_1 : p > p_0, \text{ o } H_1 : p \neq p_0,$$

Metodología para realizar la prueba $H_0 : p = p_0$ **vs** $H_1 : p > p_0$

1. Identificar las hipótesis nula y alternativa.

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

2. Especificar el nivel de significancia, dar el valor de α.

3. Determinar los puntos críticos,

   a) Para esta prueba se tiene que:

   $$\widehat{p}_c = p_0 + z(1-\alpha)\sigma(\widehat{p}) = p_0 + z(1-\alpha)\sqrt{p(1-p)/n}$$

   Luego se contrasta $\widehat{p}_m$ con $\widehat{p}_c$, es decir ¿$\widehat{p}_m > \widehat{p}_c$?

   b) Para el estadístico Z, entonces se comparan el punto crítico $z_c = z(1-\alpha/2)$ con z_m indicada por la expresión $z_m = \frac{\widehat{p}_m - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{\widehat{p}_m - p}{\sqrt{p(1-p)/n}}$. De manera similar, como en el punto 3a se tiene que:

   $$¿z_m > z_c?$$

   c) Mediante el $valor - p$

   $$¿valor - p < \alpha?$$

   donde el $valor - p = P(\widehat{p} > \widehat{p}_m)$ para el estadístico $(\widehat{p})$. Así el $valor - p = P(Z > z_m)$ para el estadístico z.

4. Si la respuesta es afirmativa en alguna de las relaciones anteriores se rechaza la hipótesis nula. Y se interpreta para el estudio en cuestión.

.

Ejemplo 8.7

En el contexto de la economía en la salud, un estudio indica que menos de 20 % de los adultos en una población son alérgicos a un medicamento; en una muestra aleatoria de 100 adultos de esa población el 15 % indicaron que son alérgicos. Con $\alpha = 0.01$ hay evidencia para apoyar lo que señaló el estudio.

Solución

La información generada por la muestra es: $\widehat{p}_m = 0.15$, $n = 100$, $z(\alpha = 0.01) = -2.326$

1. Las hipótesis se plantean como sigue:

   $$H_0 : p = 0.2$$
   $$H_1 : p < 0.2$$

   Se cumplen las restricciones $100(0.2) = 20 \geq 5$ y $100(0.8) = 80 \geq 5$. En la figura 8.23 se presenta la distribución normal con media $\mu_{\widehat{p}} = \mu(\widehat{p}) = p$ y desviación estándar $\sigma_{\widehat{p}} = \sigma(\widehat{p}) = \sqrt{p(1-p)/n}$. Ahí se describe la información esencial del ejemplo para alcanzar las conclusiones.

2. El nivel de significancia se estableció por $\alpha = 0.01$, el valor crítico de z para el α establecido es $z_c = z(\alpha) = z(0.01) = -2.326$.

3. Ahora se calcula el punto crítico

$$\widehat{p}_c = p_0 + z(\alpha)\sqrt{p(1-p)/n} = 0.2 - 2.326(\sqrt{(0.2)(0.8)/100}) = 0.107$$

4. Puesto que $\widehat{p}_c = 0.107 < \widehat{p}_m = 0.15$ ($\widehat{p}_m > \widehat{p}_c$), no se rechaza la hipótesis nula. Así que los datos dan evidencia que menos de 20 % de los adultos son alérgicos a ese medicamento.

.

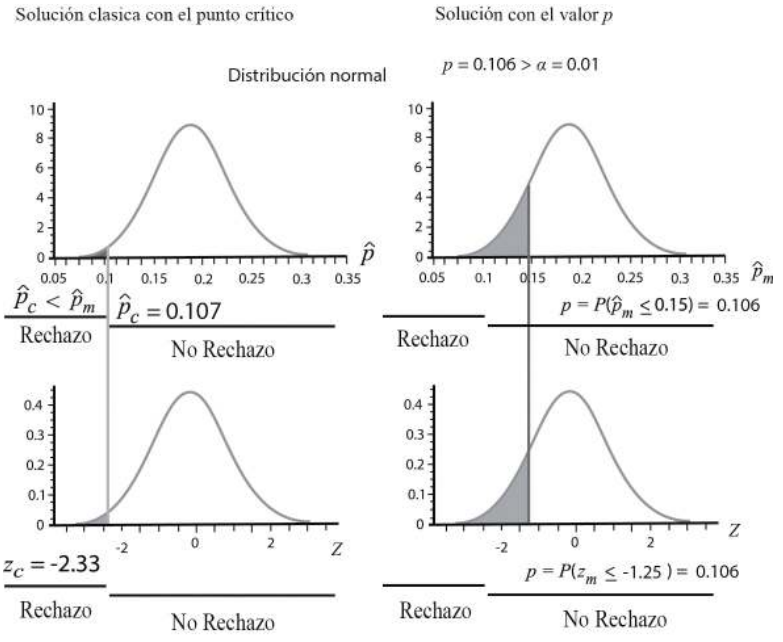


Figura 8.23 Prueba de hipótesis para una proporción para el caso $H_0 : p = 0 : 2$ vs $H_0 : p < 0 : 2$

Soluciones complementarias

1. Mediante el estadístico estandarizado Z se tiene que:

$$z_m = \frac{\widehat{p}_m - p}{\sqrt{p(1-p)/n}} = \frac{0.15 - 0.20}{\sqrt{(0.2)(0.8)/100}} = -1.25$$

dado que $z_m = -1.25 > z_c = -2.33$, se concluye que no se rechaza la hipótesis nula.

2. Utilizando el valor de significancia descriptivo, $valor - p$. La probabilidad $valor - p = P(\widehat{p}_m \leq 0.15) = 0.106$. La probabilidad de que z sea menor que -1.25, este es $p - valor = P(z_m \leq$

$-1.25) = 0.106$, y representa el área que deja a la izquierda el valor del estadístico, es decir $valor - p = P(z < -1.25)$. Así se obtiene el $valor - p = 0.106 > \alpha = 0.01$.

3. Finalmente, el intervalo de confianza del $(1-\alpha)\,\%$ para una proporción p se calcula de acuerdo con el valor de z. La expresión para el intervalos es:

$$\left(\widehat{p}_m + z(\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p})/n}), \widehat{p}_m + z(1-\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p})/n})\right)$$

$$(0.058, 0.242)$$

Lo que indica que con un 99 % de confianza la población que es alérgica a los medicamentos está entre el 6 % y 24 %.

Ejemplo 8.8

Un administrador requiere para sus procesos una serie de habilidades de sus empleados. Para un proceso de admisión contrata a un sicólogo, quien junto con sus ayudantes aplican una batería de 16 pruebas con la finalidad de determinar las destrezas de los aspirantes. Se tomó una muestra de 200 de ellos; las personas que superen 8 o más pruebas son fuertes candidatos para que sean contratados. Por experiencia, el sicólogo considera que más de 23 % cumplen esas expectativas. En este caso, el 27 % ($\widehat{p}_m = 0.27$) de los aspirantes pasaron 8 o más de las pruebas. Con esta información, ¿se verifica el supuesto del sicólogo?

Solución mediante el uso de CalEst

La solución de este ejemplo se describirá mediante el uso del **CalEst**, véase la figura 8.24, aunque primero se hace un resumen. Recuerde que éste contribuye en la parte visual para comprender el problema y da la facilidad de operación pero no sustituye la creatividad para abordar diferentes problemas. Para realizar esta prueba usando el material educativo, se va a la opción *Inferencia,* ahí se escoge $\mathcal{H}_0$ prueba de hipótesis, luego selecciona una proporción.

Aparece un cuadro que habrá que completar para hacer la prueba, el cual se aprecia en la figura 8.24. Una vez escrita la información, se tiene el reporte que aparece a la izquierda. La gráfica de la normal completa la idea de la prueba. Enseguida se hace un resumen de la prueba.

La parte operativa inicia verificando las condiciones que garanticen usar una distribución normal, es decir: $200(0.23) = 46 \geq 5$ y $200(0.77) = 154 \geq 5$. Así, se puede recurrir a la gráfica de la distribución normal para llevar a cabo el procedimiento de prueba de hipótesis.

1. Las hipótesis para este estudio se plantean por:

$$H_0 : p = 0.23$$
$$H_1 : p \neq 0.23$$

2. El nivel de significancia que se propone para realizar esta prueba es de $\alpha = 0.05$; para este valor de α se puede verificar, en la tabla de la normal estándar, que los valores de z son: $z_{ci} = z(0.025) = -1.96$ y $z_{cd} = z(0.975) = 1.96$.

3. Los valores críticos para la prueba son:

$$\widehat{p}_{ci} = p_0 + z(\alpha/2) * \sqrt{\widehat{p}(1-\widehat{p}/n)} = 0.23 - 1.96 * \sqrt{0.27(0.73/200)} = 0.167$$
$$\widehat{p}_{cd} = p_0 + z(1-\alpha/2) * \sqrt{\widehat{p}(1-\widehat{p})/n)} = 0.23 + 1.96 * \sqrt{0.27(0.73/200)} = 0.293$$

4. No se rechaza la hipótesis nula ya que $\widehat{p}_m = 0.27 < \widehat{p}_{cd} = 0.293$. Se concluye que los datos dan evidencia para no rechazar la hipótesis nula y por lo tanto la conjetura del sicólogo no se cumple en este caso.

Análogamente para el valor de z: El cálculo del estadístico de prueba mediante la normal estándar es como sigue:

$$z_m = \frac{\widehat{p}_m - p}{\sqrt{p(1-p)/n}} = \frac{0.27 - 0.23}{\sqrt{(0.23)(0.77)/200}} = 1.34$$

Como este valor es menor a 1.96, $z_m < z_{cd}$, no se rechaza la hipótesis nula.

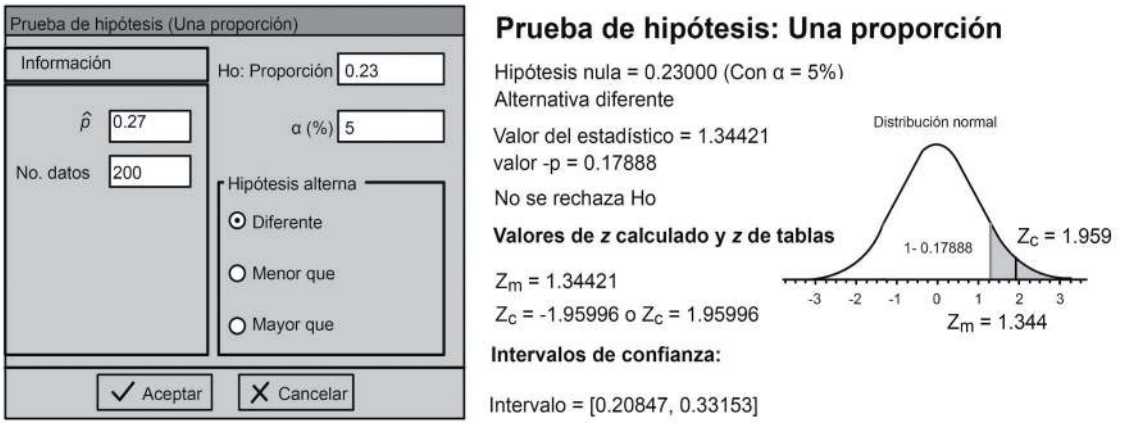


Figura 8.24 Reporte de la salida al realizar la prueba de hipótesis de una proporción.

Considerando el $valor - p$**:** También, se concluye no rechazar H_0 usando el valor p, ya que el valor de la probabilidad correspondiente a la distribución del estadístico $\widehat{p}$, es $valor - p = P(\widehat{p} \geq p_m) = 0.177$, para la variable z_m el valor calculado es $(valor - p = P(Z \geq z_m) = 0.177)$, en ambos casos $p > \alpha$.

Intervalo de confianza: El intervalo de confianza de 95 % para este ejemplo se calcula de acuerdo con el valor de Z.

$$\left(\widehat{p} + z(\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p}/n)}, \widehat{p} + z(1-\alpha/2)(\sqrt{\widehat{p}(1-\widehat{p})/n})\right)$$

$$\left(0.27 - 1.96(\sqrt{0.27(0.73)/200)}, 0.27 + 1.96(\sqrt{0.27(0.73/200)}\right) = (0.208, 0.332)$$

Lo que indica que entre 21 % y 33 % superan 8 o más de las pruebas. Nota: observe con mucho cuidado que el valor estimado por la muestra es de $\widehat{p}_m = 0.27$, y el margen en que puede estar el verdadero valor de p es amplio.

8.6 Prueba de hipótesis sobre una varianza σ^2 y σ

Motivación. La varianza desempeña un papel importante para explicar el desempeño de los procesos o fenómenos, ya que una dispersión grande de los datos en torno a la media genera una gran variabilidad de las características de estudio. Por ejemplo, una excesiva variación en las dimensiones de un producto contribuye a tener una calidad pobre, siendo que la expectativa del cliente es tener un producto uniforme, por ello es importante minimizar la variabilidad. Un aspecto relevante en muchas situaciones es conocer la varianza o la desviación estándar de un proceso con la finalidad de poder reducir ésta. Un resultado importante de la teoría estadística es el siguiente:

Estimación de σ^2

El estimador puntual para σ^2 es S^2 y el estimador puntual para σ es S, además S^2 es un estimador insesgado para σ^2.

Con la finalidad de realizar la prueba de hipótesis sobre la varianza σ^2, es necesario aplicar la distribución

conocida como Ji cuadrada (Chi cuadrada) χ^2. Si una variable aleatoria X tiene una distribución normal, entonces la distribución:

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

es una distribución Ji cuadrada para muestras de tamaño $n > 1$. El estudio de esta distribución se describe en el capítulo 6 y se utilizó en el capítulo 7 en la sección intervalos de confianza sobre σ^2.

Formulación de la prueba de hipótesis para σ^2

Para investigar la posible diferencia significativa que existe entre la varianza de una población σ^2 (o desviación estándar σ) y un valor de una varianza σ_0^2 preseleccionada (σ_0), el procedimiento se plantea como sigue:

> No existe diferencia significativa entre la varianza de una población σ^2 y un valor preseleccionado para la varianza σ_0^2.

La hipótesis nula fija en la igualdad, se expresa por:

$$H_0 : \sigma^2 = \sigma_0^2$$

Las posibles tres hipótesis alternativas son:

$$1.- H_1 : \sigma^2 < \sigma_0^2 \qquad 2.- H_1 : \sigma^2 > \sigma_0^2 \qquad 3.- H_1 : \sigma^2 \neq \sigma_0^2$$

El caso 1, se ilustra en el ejemplo 8.9. El procedimiento general de la prueba para los tres planteamientos se ilustran en el resumen.

Ejemplo 8.9

En los proyectos de mejora, la administración de una empresa que opera cines decidió aplicar un programa de capacitación para reducir a 2.9 minutos la desviación estándar del tiempo de servicio en sus dulcerías, desde que una persona está formada hasta que es atendida. Una muestra aleatoria de 23 atenciones a clientes tiene una desviación estándar 2.1 minutos. Con $\alpha = 0.01$ (10 %), ¿existe evidencia para sostener que el programa de mejora es exitoso, reduciendo la varianza? Resumen de la información muestral: $n = 23$ y $s = 2.1$ minutos.

Solución operativa clásica

1. Las hipótesis son:

$$H_0 : \sigma^2 = (2.9)^2$$
$$H_1 : \sigma^2 < (2.9)^2$$

2. $\alpha = 0.1$ es el nivel de significancia propuesto. Entonces el valor crítico $\chi^2_c = \chi^2(gl, \alpha) = \chi^2(22, 0.1) = 14.042$.

3. El valor del estadístico dada la información es:

$$\chi^2_m = \frac{(n-1)S^2_m}{\sigma^2} = \frac{(23-1)*4.41}{8.41} = 11.53$$

4. Se tiene que $\chi^2_m = 11.53 < \chi^2_c = 14.042$. Dada esta relación no se rechaza la hipótesis nula H_0 y se concluye que existe una reducción de la variabilidad en la atención al cliente. ¿Se puede conseguir una mayor mejora en el servicio?

Nota: Observe que en el punto 2 de la solución, el valor crítico para la χ^2_c representa el área izquierda en la figura 8.25, es decir $\chi^2(22, 0.1)$. Esto lo hemos presentado así para conservar nuestra propuesta de los puntos críticos, área a la izquierda (α o $\alpha/2$) y el área a la derecha ($1-\alpha$, o $1-\alpha/2$.) La mayoría de libros lo describen como el área a la izquierda ($1-\alpha$, o $1-\alpha/2$) y el área a la derecha (α o $\alpha/2$). En realidad los valores son equivalentes consulte el resumen de este capítulo prueba de hipótesis para σ^2.

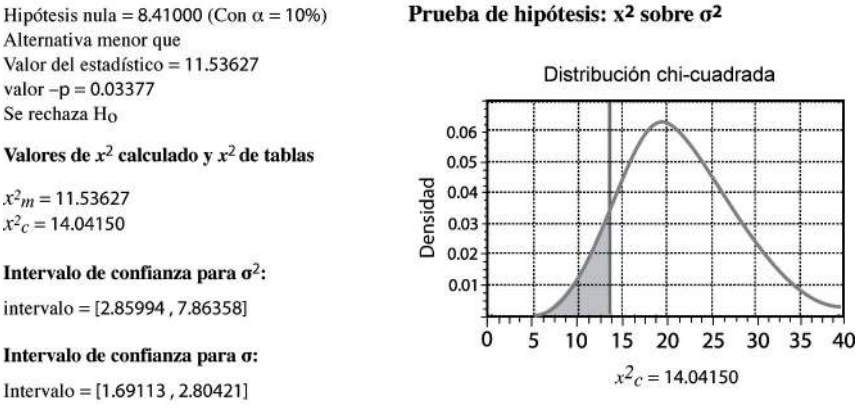


Figura 8.25 Resumen de la salida de la prueba de hiptesis usando **CalEst**.

Solución mediante el uso de CalEst

En esta parte se muestra el procedimiento para hacer una prueba de hipótesis utilizando el programa **CalEst**. En la opción inferencia se selecciona la prueba de hipótesis y ahí se escoge la varianza, se completa el cuadro con la información de la muestra y del planteamiento hipotético, tal y como se puntualiza en la figura 8.26.

Prueba de hipótesis: x^2 sobre σ^2

Hipótesis nula = 8.41000 (Con α = 10%)
Alternativa menor que

Valor del estadistico = 11.53627
valor -p=0.03377

Se rechaza H_0

Valores de x^2 calculado y x^2de tablas

x^2_m = 11.53627
x^2_c = 14.04150

Intervalo de confianza para σ^2:

intervalo = [2.85994 , 7.86358]

Intervalo de confianza para σ:

Intervalo = [1.69113 , 2.80421]

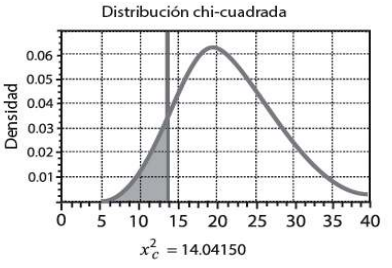


Figura 8.26 En inferencia: Pruebas de hipótesis: Una media χ^2. Luego llene la hoja.

Intervalo del (1-α) % confianza para σ^2

$$\left(\frac{(n-1)S^2}{\chi^2_{(22,1-\alpha/2)}}, \frac{(n-1)S^2}{\chi^2_{(22,\alpha/2)}}\right)$$

El resultado al sustituir los valores en la expresión anterior proporciona el intervalo del 90 % de confianza para σ^2 :

$$(\frac{97.02}{33.924}, \frac{97.02}{12.338}) = (2.8599, 7.8635)$$

Intervalo del 90 % confianza para σ es: $(1.6927, 2.804)$. La desviación estándar en la atención a clientes estará con un 90 % de confianza entre 1.7 minutos y 2.8 minutos en atención al cliente. El reporte de la prueba y el cálculo de los intervalos se muestra en la figura 8.25.

8.7 Resumen

Planteamiento de una hipótesis:

Hipótesis nula. Es una proposición que indica que no hay diferencia (no hay efecto, no hay cambio). Ésta se plantea usualmente en términos del parámetro (medida de la población) y contiene el signo igual, y se denota con H_0.

Hipótesis alternativa. Es una afirmación que indica la verdad del parámetro en lugar de la hipótesis nula. Usualmente se expresa con los símbolos $<$, $>$ o $\neq$. Ésta se denota con H_1. **Decisiones en la prueba de hipótesis.**

	Hipótesis nula	
Decisión	H_0 es verdadera	H_0 es falsa (alternativa)
No rechazar H_0	Decisión correcta	β : Error tipo II
Rechazar H_0	α : Error tipo I	Decisión correcta

Procedimiento para realizar una prueba de hipótesis

En resumen, se plantea el procedimiento para realizar una prueba de hipótesis siguendo una serie de etapas:

1. Plantear las hipótesis:

   a) Hipótesis alternativa: H_1.

   b) Establecer la hipótesis nula, H_0.

2. Proponer un nivel de significancia, valor de α. Encontrar un valor de referencia (punto crítico) a partir de una distribución de probabilidad (por ejemplo la normal, la normal estándar, la $t-Student$, la $Ji-cuadrada$). Calcular los valores críticos para las variables $(\overline{X}, \widehat{p}, S^2)$ correspondiente al nivel de significancia propuesto, es decir obtener: $(\overline{x}_c, \widehat{p}_c, S_c^2)$.

3. Comparar el valor del estadístico calculado $(\overline{x}_m, \widehat{p}_m, S_m^2)$ con el punto crítico $(\overline{x}_c, \widehat{p}_c, S_c^2)$, o comparando la probabilidad del estadístico con el valor de α. Para fijar ideas considere el caso de la media, si el estadístico $\overline{x}_m$ resulta menor o mayor que el punto crítico $\overline{x}_c$, entonces se rechaza la hipótesis nula. Nota: El procedimiento es similar para la proporción o varianza.

4. Dar una conclusión e interpretación al problema que se estudia.

Planteamiento general de las hipótesis para la media de una población:

La hipótesis nula:

$$H_0 : \mu = \mu_0$$

La hipótesis alternativa es alguna de las siguientes tres opciones:

$$1.\text{-}\ H_1 : \mu > \mu_0, 2.\text{-}\ H_1 : \mu < \mu_0, 3.\text{-}\ H_1 : \mu \neq \mu_0$$

Prueba de H_0 usando el estadístico $\overline{X}$. (Caso 1)

Considerando un valor para α (probabilidad del error tipo I) en la prueba de $H_0 : \mu = \mu_0$ contra la alternativa $H_1 : \mu > \mu_0$, se usó el estadístico de prueba $\overline{X}$ un valor $\overline{x}_m$ de éste se compara con el punto

crítico $\overline{x}_c$ para rechazar o no rechazar H_0. El procedimiento equivalente es usar como estadístico de prueba la variable Z de una normal estándar con su valor punto crítico $z(1-\alpha)$. Así:

$$\overline{x}_c = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}}$$

Observe que:

$$\overline{X} = \mu_0 + Z\frac{\sigma}{\sqrt{n}}$$

Así para un valor $\overline{x}_m$ de $\overline{X}$ y un valor z_m de Z se sigue que:

$$\overline{x}_m = \mu_0 + z_m\frac{\sigma}{\sqrt{n}} > \overline{x}_c = \mu_0 + z(1-\alpha)\frac{\sigma}{\sqrt{n}}$$

Se rechaza H_0 si $\overline{x}_m > \overline{x}_c$

Prueba de H_0 usando el estadístico estandarizado Z

Se rechaza H_0 si $z_m > z(1-\alpha) = z_c$. (Donde $z_m = \frac{\overline{x}_m - \mu_0}{\frac{\sigma}{\sqrt{n}}}$)

Prueba de la hipótesis H_0 mediante la comparación de probabilidades

A partir del *nivel de significancia descriptivo, valor* $-$ *p*. Esta probabilidad se compara con el *nivel de significancia* α, la cual es otra alternativa para decidir sobre la hipótesis nula. Por lo tanto:

$$\text{si } valor - p < \alpha, \text{ se rechaza } H_0$$

Comparación de los criterios de prueba de hipótesis

Para el caso: lectura del iris, observe la relación entre los estadísticos que se obtienen a partir de los datos de la muestra: x_m, z_m y el nivel de significancia descriptivo $valor - p$, así como los valores de referencia, valores críticos, x_c, z_c y α. Se muestra la dependencia de estos valores mediante la figura 8.28, la gráfica a la izquierda muestra los valores críticos y la otra gráfica a los estadísticos de muestra. Entonces en el método para la prueba de hipótesis, viendo las gráficas se compara los cuadros, los exágonos o los óvalos. El primero muestra el plan natural con los datos originales de la muestra, el segundo es el procedimiento, limitado, a la normal estándar y finalmente la regla que compara las probabilidades.

Prueba de hipótesis cuando $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ muestras pequeñas usando la distribución $t - Student$

La idea principal de la prueba de hipótesis es comparar dos números, el valor del estadístico de prueba con el valor de una distribución de probabilidad (éste se obtiene a partir de un valor de significancia, probabilidad, establecido por α); considere:

$$\overline{x}_c = \mu + t(n-1,\alpha)\frac{S}{\sqrt{n}}$$

Se compara el valor del estadístico $\overline{x}_m$ con el punto crítico $\overline{x}_c$.

- Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$, se rechaza la hipótesis si $\overline{x}_m$ es menor que $\overline{x}_{ci} =$

$\mu + t(n-1, \alpha/2)\frac{S}{\sqrt{n}}$ o mayor que $\overline{x}_{cd} = \mu + t(n-1, (1-\alpha/2))\frac{S}{\sqrt{n}}$.

Considere:

$$t_m = \frac{\overline{x}_m - \mu}{\frac{S}{\sqrt{n}}}$$

Se compara con un valor de la distribución t_c de Student $t(n-1, \alpha)$, donde α es una probabilidad, y $n-1$ son los grados de libertad.

- Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$, se rechaza la hipótesis si t_m es menor que $t_{ci} = t(n-1, \alpha/2)$ o mayor que $t_{cd} = t(n-1, (1-\alpha/2))$.

  Una alternativa es comparar la probabilidad p que deja el estadístico de prueba (a la derecha o izquierda) con el valor de significancia, probabilidad, α. Análogamente:

- Si la hipótesis alternativa es $H_1 : \mu \neq \mu_o$ se rechaza la hipótesis si $2(valor - p)$ es menor que α.

La prueba $H_0 : p = p_0$ vs $H_1 : p > p_0$

1. Especificar el nivel de significancia, dar el valor de α.
2. Determinar los puntos críticos,

   a) Para esta prueba se tiene que:

   $$\widehat{p}_c = p_0 + z(1-\alpha)\sigma(\widehat{p}) = p_0 + z(1-\alpha)\sqrt{p(1-p)/n}$$

   Luego se contrasta $\widehat{p}_m$ con $\widehat{p}_c$, es decir se rechaza la hipótesis nula si $\widehat{p}_m > \widehat{p}_c$

   b) Para el estadístico Z, entonces se comparan el punto crítico $z_c = z(1-\alpha/2)$ con z_m indicada por la expresión $z_m = \frac{\widehat{p}_m - \mu_{\widehat{p}}}{\sigma_{\widehat{p}}} = \frac{\widehat{p}_m - p}{\sqrt{p(1-p)/n}}$. Se rechaza la hipótesis nula si $z_m > z_c$.

   c) Mediante el $valor - p$ se rechaza la hipótesis nula si $valor - p < \alpha$, donde el $valor - p = P(\widehat{p} > \widehat{p}_m)$ para el estadístico $(\widehat{p})$. Así el $valor - p = P(Z > z_m)$ para el estadístico z.

Prueba de hipótesis para la varianza σ^2

La hipótesis nula expresada por la igualdad:

$$H_0 : \sigma^2 = \sigma_0^2$$

Tres casos de hipótesis alternativas:

$$\text{1.- } H_1 : \sigma^2 < \sigma_0^2 \qquad \text{2.- } H_1 : \sigma^2 > \sigma_0^2 \qquad \text{3.- } H_1 : \sigma^2 \neq \sigma_0^2$$

Valores Críticos:

Caso 1: $\chi^2_{c*} = \chi^2(gl, \alpha)$, Caso 2: $\chi^2_{c\#} = \chi^2(gl, 1-\alpha)$, Caso 3: $\chi^2_{ci} = \chi^2(gl, \alpha/2)$, y $\chi^2_c = \chi^2(gl, 1-\alpha/2)$.

El estadístico de prueba:

$$\chi^2_m = \frac{(n-1)S^2}{\sigma^2}$$

Se rechaza H_0, si:

$$\text{Caso 1. } \chi^2_m < \chi^2_{c*}, \quad \text{Caso 2. } \chi^2_m > \chi^2_{c\#}, \text{ Caso 3. } \chi^2_m < \chi^2_{ci} \text{ o } \chi^2_m > \chi^2_{cd}$$

8.8 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

R. A. Fisher

Incluso los científicos necesitan sus héroes y R.A. Fisher fue sin duda el héroe de la estadística del siglo XX. Sus ideas transformaron la disciplina de tal forma que hasta un César o un Alejandro hubieran envidiado. B. Efron.

8.9 Ejercicios

Planteamiento y conceptos básicos de una hipótesis estadística

8.1 Describa el procedimiento para realizar una prueba de hipótesis y coméntelo.

8.2 En el contexto de la prueba de hipótesis:

1. Explique y comente por qué la regla de decisión tiene que emplearse en indicar si la hipótesis nula debería rechazarse. ¿Cuál es el papel que desempeña la probabilidad en esta decisión?
2. ¿Por qué se acepta la hipótesis nula como cierta?
3. ¿Cómo determina los valores críticos para llevar a cabo la prueba de hipótesis, los reales y los estandarizados? Utilice las gráfica de la normal para bosquejar las ideas.
4. ¿Cómo influye el tamaño de la muestra en la prueba de hipótesis? ¿Afecta los valores críticos?
5. ¿Cómo influye el nivel de significancia en los valores críticos?
6. Explique la diferencia entre los errores tipo I y tipo II. Escriba la relevancia de cada uno de ellos en la prueba de hipótesis.

8.3 Se sabe que la cantidad promedio de nicotina que contienen los cigarros es de $\mu = 1.6$ y se sabe que tiene una desviación estándar de $\sigma = 0.8$. Una compañía tiene el reto de reducir esos niveles de nicotina cambiando su proceso, en lo que se refiere al quemado de la hoja. Una vez realizadas las correcciones en el proceso se toma una muestra de $n = 20$ cigarros y el promedio de nicotina en esos cigarros es $\overline{x}_m = 1.56$.

1. ¿Cómo puede esa compañía saber que su proceso es efectivo para reducir el nivel de nicotina?
2. ¿Qué hipótesis se puede plantear la empresa para evaluar su proceso?
3. ¿Cómo se puede traducir esa información en hipótesis estadísticas?
4. ¿Cuál es la población objeto de estudio? ¿Cuál es la muestra y cómo se obtienen ésta?
5. Escriba qué parámetros dan información sobre el problema.
6. Indique cuáles son los estadísticos que contribuyen a la información del problema.
7. Bosqueje la distribución del estadístico e indique los parámetros de esta distribución, es decir la media y la varianza.
8. Describa el estadístico de prueba cuando se conoce la desviación estándar e indique cuál es su distribución de probabilidad.
9. Describa el estadístico de prueba cuando no se conoce la desviación estándar e indique cuál es su distribución de probabilidad.
10. ¿Cómo podría plantear la estrategia para conocer si el proceso de la empresa fue eficiente en la reducción de los niveles de nicotina?

8.4 Se sabe que los niveles altos de colesterol son un problema para la salud de los individuos. Se proponen dos tratamientos, A y B, para reducir los niveles de colesterol. Se presume que el tratamiento A es mejor que el tratamiento B.

1. ¿Qué hipótesis se puede plantear en esta situación?
2. ¿Cómo se puede traducir esta situación en hipótesis estadísticas?
3. ¿Cuál es la población objeto de estudio? ¿Qué procedimiento seguiría para realizar este estudio?
4. ¿Cuáles son las muestras para el estudio y cómo se obtiene éstas?
5. Escriba qué parámetros dan información sobre el problema.
6. Indique cuáles son los estadísticos que contribuyen a la información del problema.
7. Bosqueje la distribución del estadístico e indique los parámetros de esta distribución es decir la media y la varianza.
8. ¿Qué supuesto se debe hacer sobre las varianzas? ¿Cómo puede verificar que se cumple este supuesto?
9. Describa el estadístico de prueba cuando se conoce la desviación estándar e indique cuál es su distribución de probabilidad.
10. Describa el estadístico de prueba cuando no se conoce la desviación estándar e indique cuál es su distribución de probabilidad.
11. ¿Cómo podría plantear la estrategia para conocer si el tratamiento A es mejor que el tratamiento B?

8.5 Con la finalidad de tomar diferentes medidas correctivas en el aprovechamiento escolar, la administración de un centro escolar tiene interés en conocer el rendimiento de sus alumnos. Para ello la escuela aplica una prueba de 100 preguntas para determinar el conocimiento general de sus estudiantes; se tomó una muestra de 45 de ellos al finalizar el ciclo escolar de primaria. El director de la escuela considera que se tiene un nivel bajo si la prueba arroja una media menor a 70 puntos. El resumen de la siguiente información es: la media $\overline{x}_m = 68.14$, la desviación estándar $\sigma = 10.34$ y el tamaño de muestra $n = 45$.El director quiere ser conservador ante la posibilidad de cometer el error tipo I y propone el nivel de confianza $\alpha = 0.02$.

1. Plantee la hipótesis.
2. Realice la prueba.

Prueba de hipótesis para una media: muestras grandes

8.6 Se sabe que una estrategia de venta que siguen muchas tiendas comerciales grandes es que los clientes compren a crédito. Una de estas tiendas tiene 10 mil clientes que han usado la cuenta y por consiguiente tienen un adeudo. El administrador del departamento de ventas estima que la media de la cantidad de adeudo es superior a los 3 mil pesos. Selecciona una muestra de 36 clientes y la media es 2,700 pesos con una desviación estándar de 520 pesos. Verifique la hipótesis con un alfa de 0.05.

Prueba de hipótesis para una media: muestras pequeñas

8.7 Conteste a las siguientes cuestiones:

1. ¿Qué nos indica la prueba $t - Student$?
2. Defina y explique el $valor - p$
3. ¿Describa una situación en la que es mejor usar una prueba de dos colas a una de una cola?
4. ¿Qué valores de t_m corresponden a valores grandes o pequeños del $valor - p$?

8.8 En un proceso de admisión la administración de una compañía contrata a un sicólogo. Éste aplica una prueba de memoria en la que el tiempo de respuesta debe ser 3 minutos. Se aplica la prueba a 10 empleados y se mide el tiempo de respuesta una vez que se cumplieron los tres minutos. Con un nivel de significancia 0.05, los datos ¿apoyan la hipótesis nula de que el tiempo después de los tres minutos es cero? Los dato reportados son 1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4

8.9 Una empresa que genera un producto químico requiere que la media del nivel de pH en el agua debe estar en 6.8. Si eso no ocurre se genera un costo por retrabajo, lo que representa una pérdida económica para la empresa. La administración de la empresa le solicita un reporte al técnico responsable de este proceso. En el contexto de estudio se toman 19 muestras de agua y mide el pH de cada una de ellas. Los datos son:

6.7	7.1	6.8	6.9	6.5	6.7	6.6	6.5	6.5	6.2
6.3	6.6	7.0	6.7	6.9	6.5	6.6	6.9	6.9	

8.10 En el Banco Central están debatiendo sobre vender o no parte de las reservas que tienen en dólares con el objetivo de evitar la devaluación de la moneda nacional. Un grupo está a favor, mientras que otro se opone totalmente a dicha medida.

1. Formule la prueba de hipótesis nula y alternativa desde el punto de vista de los que quieren vender parte de las reservas.
2. Formule la prueba de hipótesis nula y alternativa desde la perspectiva de los que se oponen.

8.11 El gerente de una agencia de seguros está diseñando un nuevo plan de incentivos para los vendedores con el objetivo de aumentar el volumen de ventas. Para comparar si el nuevo esquema de incentivos es efectivo, el gerente prueba el nuevo esquema por un mes con un grupo de vendedores.

1. Desarrolle la prueba de hipótesis nula y alternativa que ayude al gerente a tomar una decisión.
2. Comente en la conclusión cuándo H_0 no puede ser rechazada.
3. Comente en la conclusión cuándo H_0 puede ser rechazada.

8.12 El responsable de calidad de una fábrica empacadora de azúcar quiere revisar que los nuevos procedimientos han reducido el número de bolsas de azúcar que se rellenen por debajo de la cantidad prometida. Exprese la hipótesis nula y alternativa e indique los resultados que proporcionarían una evidencia sólida.

8.13 El administrador de un hotel en Cancún ha concluido que los huéspedes en promedio gastan por un fin de semana la cantidad de $4,500 o menos. El contador del hotel ha notado un incremento en lo que los huéspedes gastan en los últimos meses. El contador va a tomar la muestra de un fin de semana para probar lo que dice el administrador del hotel.

1. ¿Cuál planteamiento de prueba de hipotesis debe ser utilizada? Explique el porqué.

$$\begin{array}{lll} H_0 : \mu \geq 4500 & H_0 : \mu \leq 4500 & H_0 : \mu = 4500 \\ H_1 : \mu < 4500 & H_1 : \mu > 4500 & H_0 : \mu \neq 4500 \end{array}$$

2. Comente en la conclusión cuándo H_0 no puede ser rechazada.

3. Comente en la conclusión cuándo H_0 puede ser rechazada.

8.14 En una fábrica de llantas siguiendo el proceso actual se fabrican en promedio 90 neumáticos por hora con una desviación estándar de 7. Una agencia consultora ha propuesto un nuevo proceso. El gerente de producción dice que el proceso será cambiado si hay suficiente evidencia que el nuevo proceso incrementará la producción promedio. Se realizan pruebas con el nuevo procedimiento por 32 horas. Durante este periodo, se produjeron 93 llantas en promedio por hora. ¿Qué le recomendaría usted al gerente? ¿Quedarse con el viejo proceso o cambiar al nuevo? ¿Por qué?

8.15 Un fabricante de aceite de oliva asegura que sus botellas contienen en promedio 375 ml. Se sabe que el contenido sigue una distribución normal con una desviación estándar de 25 ml. Se toma una muestra de 20 botellas. El contenido promedio de esta muestra es de 371. Realice una prueba de hipótesis de que el contenido es al menos el prometido (375 ml). Realice la prueba con 10 % de nivel de significancia.

8.16 En una tienda revisan los productos que sus proveedores les entregan. Uno de los productos son baterías. El productor asegura que las baterías tienen un tiempo de vida de 60 horas. De pruebas anteriores, se sabe que el tiempo de vida tiene una distribución normal y una desviación estándar de 4 horas. De la entrega de hoy, el inspector tomó 8 baterías y midió su tiempo de vida. En promedio estas 8 baterías funcionaron 58.5 horas. Prueba con un 10 % nivel de significancia la prueba de hipótesis de que el cargamento durará al menos las 60 horas prometidas por el fabricante.

8.17 Una compañía telefónica mide el número de mensajes que sus usuarios envían por día. En promedio se envían 9.6 mensajes por día por usuario. Un investigador de mercado cree que los usuarios con puestos ejecutivos envían más mensajes que el resto de los usuarios. Un grupo de ejecutivos será usado para el estudio. Se plantea la prueba de hipótesis de la siguiente manera:

$$H_0 : \mu \leq 9.6$$
$$H_1 : \mu > 9.6$$

1. ¿Cuál es el error Tipo I en esta situación? ¿Cuáles son las consecuencias de cometer este error?
2. ¿Cuál es el error Tipo II en esta situación? ¿Cuáles son las consecuencias de cometer este error?

8.18 La etiqueta de un jugo de naranja de 1 litro, asegura tener en promedio 900 mg de calcio. Responda las siguientes preguntas para una prueba de hipótesis que podría ser utilizada para probar la aseveración en la etiqueta.

1. Desarrolle la prueba de hipotesis pertinente para el problema ¿Cuál es el error Tipo I en esta situación? ¿Cuáles son las consecuencias de cometer este error?
2. ¿Cuál es el error Tipo II en esta situación? ¿Cuáles son las consecuencias de cometer este error?

8.19 En una compañía de automóviles, un grupo de investigación concluyó que el uso del nuevo sistema de control electrónico representa una mejora de 5 km por litro en la eficiencia del consumo de combustible de los que lo usan. Se evaluó el sistema en 100 autos tomados al azar, obteniéndose una mejora promedio de 4.3 km por litro con una desviación estándar de 2.7 km/l. Realice una prueba de hipótesis de que la media de la mejora de la población es al menos de 5 km/l. Obtenga el valor-p, e interprete sus resultados.

8.20 Un hospital realizó una encuesta sobre la satisfacción del servicio dado. Los entrevistados tenían que calificar el servicio de una escala de 0 (muy malo) a 5 (muy bueno). De los 172 pacientes entrevistados, la calificación promedio fue de 3.9 con una desviación estándar de 0.8. Realice una prueba con nivel de significancia de 1 %. Considere que la prueba nula es que la calificación es a lo mucho 3.0 y la prueba alternativa es mayor que 3.0.

8.21 Una compañía decide anunciar sus productos en internet esperando subir sus ventas en un 20 %. De una muestra de 30 tiendas, se observó que las ventas aumentaron en promedio 17.3 % con una desviación estándar de 4.5. Realice una prueba de hipótesis de que el aumento de las ventas promedio en todas las tiendas fue de al menos 20 %. Use un nivel de significancia de 5 %.

8.22 Un proceso de envasado de enjuague para el cabello llena las botellas con 600 gramos de producto. Se tomó una muestra obteniéndose las siguientes mediciones (en gramos):

608	598	597	589	601	599	575	588	595	591

Realice con un nivel de significancia del 10 % una prueba de hipótesis de que el contenido de las botellas de enjuague tienen al menos la cantidad prometida.

Pruebas de hipótesis para una proporción

8.23 Con el fin de conocer la memoria a corto plazo un sicólogo realiza una prueba con 200 personas, para ello le ayudan varios de sus asistentes. La prueba consiste en mostrar una tarjeta con 16 palabras a cada una de las personas por 30 segundos, a continuación se les distrae por un minuto platicando con los entrevistados. Finalmente se le pide a la persona que diga las palabras que recuerda, para ello se da un minuto. El investigador plantea que 23 % de las personas recuerdan 8 o más palabras. La información que recogieron de las $n = 200$ entrevistas es que $\widehat{p} = 0.27$ recuerdan 8 o más palabras. Verifique que se cumple la siguiente hipótesis.

$$\begin{aligned} H_0 &: p = 0.23 \\ H_1 &= p \neq 0.23 \end{aligned}$$

8.24 Un médico supone que más de 55 % de las personas que viven en una zona cercana a una refinería tienen algún problema de salud relacionado con el aparato respiratorio. Para confirmar su supuesto realizé con un equipo de médicos y varios laboratorios 425 pruebas en una localidad y encontró que 255 padecían un mal respiratorio.

8.25 Para las siguientes situaciones, formule las hipótesis nula y alternativa

1. Menos de 40 % de los estudiantes de la carrera de administración leen tres libros de cultura general durante el semestre escolar.

2. Más de 65 % de los estudiantes del primer semestre de la carrera de economía sólo estudian con las notas del profesor.

3. Una mayoría de personas no están de acuerdo con el horario de verano.

4. A más de la mitad de los estudiantes les gusta ir al restaurant de la universidad.

5. Al menos el 50 % de los profesores utilizan los nuevos recursos de la tecnología para enseñar a sus alumnos.

6. No más de un tercio de personas votarán por el nuevo candidato.

8.26 El banco ha tratado de incrementar la proporción del número de créditos otorgados por dicha institución bancaria a los clientes con la categoría de preferentes. De los 160 créditos otorgados este mes, 64 están identificados como clientes preferentes. ¿El banco habrá tenido éxito por atraer más clientes preferentes? Use $\alpha = 0.025$; obtenga el $valor - p$.

8.27 La administración de un gobierno considera que a lo más el 15 % de los establecimientos de la localidad no cumple con los precios oficiales. De una muestra de 200 tiendas se encontró que 17 exceden los precios oficiales. ¿Existe suficiente evidencia para rechazar lo dicho por la administración?

8.28 Considerando una zona comercial en una ciudad grande, más del 70 % de los empleados toma más de 2 horas, todos los días, para llegar a su centro de trabajo. De una muestra de 300, 224 indicaron que sí emplean ese tiempo para trasladarse. Exprese las hipótesis y realice la prueba con $\alpha = 0.05$. ¿Cuál es el $valor - p$?

8.29 Encuentre el $valor - p$ para la siguiente hipótesis:

1. $H_0 : p = 0.5$ vs $H_1 : \mu \neq 0.5$, $z_m = -2.26$.
2. $H_0 : p = 0.2$ vs $H_1 : \mu > 0.2$, $z_m = 1.46$.

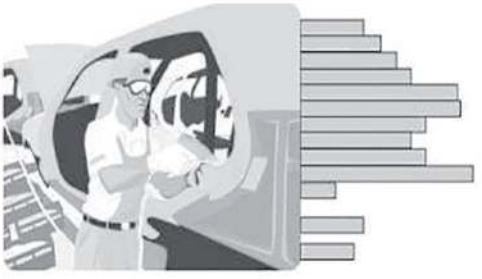
Prueba de hipótesis sobre la varianza

8.30 Un sistema de riego no da uniformidad a la distribución del agua si la varianza es mayor a 0.25 unidades $(\text{cm/hr})^2$. Para probar la uniformidad de la aplicación del agua para un nuevo sistema se midió la cantidad de agua después de 1 hora en 41 lugares seleccionados aleatoriamente. La media y varianza registrados fueron: 0.85 y 0.27, respectivamente.

8.31 Un fabricante de hilo industrial señala que la tensión de su producto al final de la línea de producción tiene una varianza diferente a 15.9 unidades. Un auditor selecciona una muestra aleatoria de 15 carretes al final de la línea, los cuales muestran una varianza de 21.8 unidades. En el supuesto de que la población tiene una distribución normal, verifique que estos datos muestran suficiente evidencia para rechazar la hipótesis nula; considere un $\alpha = 0.05$ para el nivel de significancia.

8.10 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.



Capítulo 9

Inferencia estadística para dos poblaciones

La búsqueda de la verdad debe ser el objeto de nuestra actividad; es el único fin digno de ella. En primer lugar, sin duda debemos esforzarnos por aliviar los sufrimientos humanos; pero ¿por qué? No sufrir es un ideal negativo que sería alcanzado con más seguridad por el aniquilamiento del mundo. Si queremos librar, cada vez más, al hombre de sus preocupaciones materiales, es para que pueda emplear su libertad reconquistada en el estudio y la contemplación de la verdad.

Jules Henri Poincaré

Competencia general

Extender los conocimientos de inferencia estadística a dos poblaciones para construir intervalos de confianza y realizar prueba de hipótesis sobre diferencia de parámetros.

Competencias específicas

- Conocer situaciones donde es necesario estimar la diferencia de parámetros, para los casos de dos medias, dos proporciones y la razón de varianzas.
- Identificar los conceptos principales -variable aleatoria, distribución de probabilidad y parámetros- que caracterizan a dos poblaciones.
- Reproducir las ideas principales de los métodos de estimación para la diferencia de parámetros.
- Identificar los supuestos básicos en la diferencia de dos poblaciones para realizar la prueba de hipótesis, tales como varianzas conocidas iguales o diferentes en muestras pequeñas.
- Describir la diferencia entre la selección de muestras independientes y pareadas.
- Explicar el procedimiento para hacer inferencia estadística en dos poblaciones, considerando las diferentes condiciones estadísticas.
- Establecer la relación entre la estimación de intervalos de confianza y pruebas de hipótesis para dos poblaciones.
- Leer un artículo de divulgación en temas de administración o economía para identificar los conceptos estadísticos en inferencia estadística para dos poblaciones.

9.1 Introducción

Planteamiento estadístico para comparar dos poblaciones

En el trabajo observacional o experimental es frecuente que surja la necesidad de comparar dos poblaciones. En cuanto a los temas de interés en administración y economía:

1. Se puede examinar el gasto que los estudiantes destinan en material educativo o libros al semestre, ya que de alguna manera puede repercutir en la economía familiar. La comparación se sugiere entre los alumnos de dos centros educativos de nivel superior. La variable aleatoria es el gasto en pesos que realizan los estudiantes. En este caso los centros educativos ejemplificarán *dos poblaciones*, así la variable de cada una de éstas tiene una distribución de probabilidad, que se supone normal, con su respectiva media y varianza. Este planteamiento da lugar a estimar los parámetros de las poblaciones y evaluar la diferencia de medias para saber si el gasto es diferente. También, considerar la razón entre las varianzas para saber si el gasto en libros es *homogéneo* entre los dos centros educativos. Las muestras seleccionadas en esta parte se les llaman *muestras independientes*.

2. El rendimiento laboral es un asunto de interés para muchos comercios y compañías, en tal situación la administración desempeña un papel importante en alcanzar tal meta. En una empresa, el gerente de capacitación diseña un programa para mejorar la productividad. La evaluación del rendimiento laboral se logra mediante la aplicación de una serie de cuestionarios relacionados con el desempeño; al final de éstos se obtiene un índice que califica el rendimiento. En este caso, la variable es la calificación que establece el índice. El procedimiento del gerente para realizar esta actividad es como sigue: Paso 1, toma una muestra de la población de trabajadores, y evalúa el rendimiento. Paso 2, aplica el programa de capacitación a los trabajadores seleccionados en la muestra, a éstos les aplica la prueba de rendimiento y los califica. Preste atención a lo siguiente: la población descrita en el paso 1 se denominará *población 1*, que representa a los trabajadores antes de aplicar el programa. En el paso 2 se tendrá la *población 2* y corresponde a los que recibieron la capacitación. Ahora, observe que la muestra 1 es la que se refiere a los trabajadores antes del entrenamiento, la muestra 2 toma a los mismos, pero a los que se les aplicó el programa. Estas muestras se conocen con el nombre *muestras pareadas*.

Una de las metas de este capítulo es aprender a diferenciar esta situación y aplicar la metodología apropiada para realizar la inferencia estadística ante estas situaciones. Otros ejemplos que pueden ser de interés en el estudio de dos poblaciones son:

- Existen empresas que realizan su manufactura en zonas urbanas y zonas rurales; se puede preguntar: ¿el salario que perciben los empleados de ambas zonas es similar?
- Una actividad de compra moderna se practica por internet. ¿El porcentaje de mujeres y hombres que compran por internet es igual?

- El registro ambiental que se lleva a cabo sobre la actividad industrial tiene un impacto económico. En éste se realiza una serie de estudios que generan información sobre emisiones atmosféricas entre otras. ¿El índice de contaminación reportado por dos estaciones, la primera en una zona industrial y otra en una zona residencial, es el mismo?

9.2 Parámetros y estimación

El modelo descrito sobre el quehacer de la estadística en la comparación de dos poblaciones se propone en el siguiente problema.

El mundo de la información 1. Consumo de enrgía eléctrica

El consumo de energía eléctrica, medida en Kwh, en los hogares forma parte de la economía familiar. En el contexto de administrar de manera adecuada el gasto, los expertos recomiendan estar al pendiente de este consumo. Por otro lado, poseer equipos electrodomésticos da lugar a tener un mayor desarrollo social y ese hecho impacta en los índices con los que los gobiernos miden el progreso. En esa dirección, un administrador tiene interés en conocer si existe un mayor consumo de energía al bimestre en las casas que tienen un horno de microondas, que las que no lo tienen. En este estudio se consideran casas de una zona con un nivel social similar.

Preguntas sobre la naturaleza del problema

¿Cómo se caracterizan las poblaciones para este estudio? ¿Cuánta energía consume un hogar con horno de microondas? ¿Cuánta energía consumen los hogares sin este aparato eléctrico? ¿Qué estrategia se sigue para tomar una muestra de cada población? ¿Cuáles son las variables del estudio? ¿Tienen éstas una distribución normal? ¿Qué tamaño de muestra se requiere de cada población para el estudio? ¿Existe diferencia significativa en el consumo de energía en estas dos situaciones? Contemplando otras ideas, ¿las personas que viven en esa zona, están informados sobre la alternativa de la energía solar? ¿Tienen acceso a ella?

Estrategias estadísticas sobre el estudio: Con la finalidad de tener la información lo más homogénea posible se toma una zona parecida en cuanto a factores económicos. Se toma una muestra de personas y se les pregunta si poseen horno de microondas. La respuesta es: sí o no. Luego, se les pregunta el consumo de energía en su casa medida en Kwh. Se puede clasificar a las personas en dos categorías que se denominarán poblaciones, una de ellas será la de hogares con horno de microondas y la otra las que no cuentan con ese aparato. Con esta información, ¿se tienen los elementos para decidir si hay diferencia en el consumo de energía entre estas dos poblaciones? Para responder a esta pregunta primero se hará una descripción de este entorno.

La idea es tener una región de estudio lo más homogénea posible, en ese sentido considere un centro de estudios superiores, y de ahí seleccionar estudiantes y preguntarles. Se pensaría que los estudiantes

que van en la misma escuela, por lo general son afines socialmente.

Escenario estadístico de dos poblaciones: Descripción de las poblaciones

- Población 1: {casas *con* horno de microondas}
- Población 2: {casas *sin* horno de microondas}

Se cree que el consumo de energía de las casas que tienen horno de microondas es mayor que las que no tienen ese aparato electrodoméstico, tome como referencia el consumo señalado en el último bimestre. La variable aleatoria es el consumo de energía:

X_1: Consumo de energía en casas *con* horno de microondas
X_2: Consumo de energía en casas *sin* horno de microondas

A continuación, se presentarán las ideas básicas para llevar a cabo la verificación de esa conjetura.

Escenario estadístico de dos poblaciones en general: Determinación de las poblaciones:

- Población 1: $\{u_{11}, u_{12}, u_{13}, ..., u_{1N_1}\}$
- Población 2: $\{u_{21}, u_{22}, u_{23}, ..., u_{2N_2}\}$

La población 1 tiene N_1 unidades y la población tiene N_2 unidades, a las que se les medirá una característica específica para comparar las poblaciones una vez que se ha planteado alguna conjetura sobre ellas. Variable aleatoria X_1 para la población 1, y variable aleatoria X_2 para la población 2 (véase la figura 9.1)

Distribución de probabilidad de las variables:

X_1: Tiene una distribución de probabilidad normal con media μ_1 y varianza σ_1^2, $X_1 \sim N(\mu_1, \sigma_1^2)$
X_2: Tiene una distribución de probabilidad normal con media μ_2 y varianza σ_2^2, $X_2 \sim N(\mu_2, \sigma_2^2)$

Las preguntas: ¿existe diferencia entre las poblaciones?

$$\mu_1 - \mu_2 =? \qquad \frac{\sigma_1^2}{\sigma_2^2} =? \qquad p_1 - p_2 =?$$

¿Cómo se compara? ¿Cómo se estima?

Selección aleatoria de las muestras:

Muestra 1: $\{u_{11}, u_{12}, u_{13}, ..., u_{1n_1}\}$	Muestra 2: $\{u_{21}, u_{22}, u_{23}, ..., u_{2n_2}\}$

donde n_1 es bastante menor que N_1 y n_2 es bastante menor que N_2.

Observación: En principio, la meta principal es estudiar la distribución de probabilidad con respecto a la diferencia de las medias muestrales $\overline{X}_1 - \overline{X}_2$, considerando los casos en muestras grandes y muestras pequeñas, así como la distribución para la diferencia de proporciones $\widehat{p}_1 - \widehat{p}_2$, y finalmente la distribución sobre la razón de varianzas $\frac{S_1^2}{S_2^2}$. Estos temas se presentarán en las siguientes secciones.

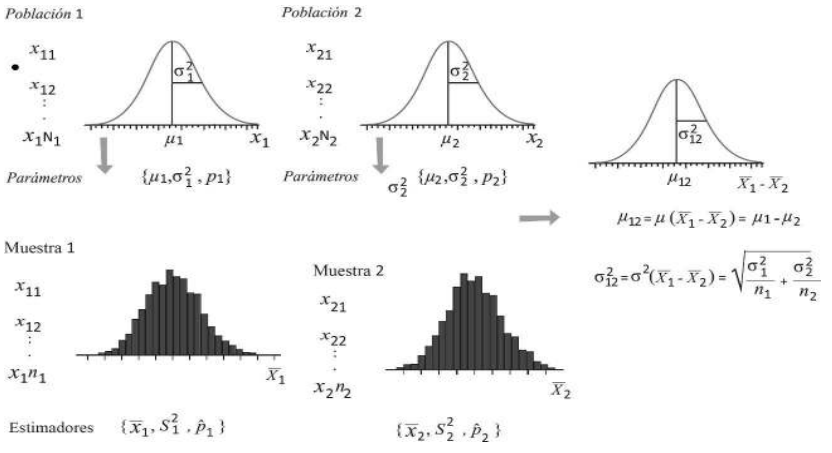


Figura 9.1 Panorama para hacer inferencia de dos poblaciones y la distribución muestral de $\overline{X}_1 - \overline{X}_2$.

Con la información expuesta en la figura 9.1, se retoman las ideas de los capítulos 7 y 8, y se extienden a la comparación de dos poblaciones. Así, la estimación puntual de la diferencia entre $\mu_1 - \mu_2$ se expresa por la diferencia entre las dos medias muestrales, $\overline{X}_1 - \overline{X}_2$. La estimación por intervalo y la prueba de hipótesis considerando diferentes supuestos, son similares a los tratados anteriormente. Las variantes de estos conceptos y el desarrollo de los procedimientos para construir el intervalo de confianza y hacer prueba de hipótesis en relación a la diferencia de las medias se explicarán en la primera parte de este capítulo.

En síntesis, para llevar a cabo inferencia estadística entre dos poblaciones considerando muestras independientes, se tiene que:

- Las muestras deben ser seleccionadas de manera aleatoria.
- Las muestras deben ser independientes. Dos muestras son independientes si la muestra seleccionada de una población no está relacionada con la muestra de la otra población.
- Cada población debe tener una distribución normal.

El procedimiento mediante la tecnología

Nota: En el proceso de prueba de hipótesis para dos poblaciones o la estimación por intervalo de confianza, se tiene la opción de usar la tecnología. Los conceptos de error tipo I y error tipo II desempeñan el mismo

papel que ya se discutió en el capítulo anterior. Ahora se puede emplear el calculador estadístico para facilitar la parte operativa, pero es importante tener en mente los fundamentos teóricos para realizar pruebas de hipótesis en la comparación de dos poblaciones. El ambiente de **CalEst** para este tema se presenta en la figura 9.2.

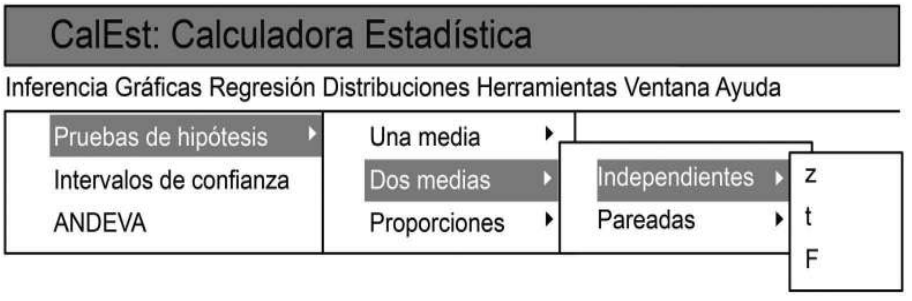


Figura 9.2 Opción para realizar los cálculos para hacer inferencias sobre dos poblaciones.

Se realizarán pruebas de hipótesis para muestras independientes tal y como indican en la figura 9.2, ésta puede ocurrir cuando las muestras seleccionadas de dos poblaciones, son grandes y en este caso se usa la distribución normal, o la normal estándar z, o pequeñas y en tal caso se usa la distribución $t-$Student. Para realizar las prueba de hipótesis sobre dos poblaciones se parte del supuesto de que las varianzas de la población son iguales, para verificarlo se realiza una prueba de hipótesis sobre las varianzas y en ese caso se usa la prueba F.

9.3 Prueba de hipótesis para la diferencia de medias: muestras independientes

Descripción de las hipótesis en referencia a dos poblaciones.

1. La hipótesis nula H_0 es la hipótesis estadística que usualmente indica que no hay diferencia entre los parámetros de dos poblaciones.
2. La hipótesis alternativa es la hipótesis estadística que indica la relación entre las dos poblaciones y se plantea como menor, mayor o diferente.

Los tres casos son:

$$\begin{array}{lll} H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 & H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 & H_1 : \mu_1 > \mu_2 & H_1 : \mu_1 < \mu_2 \end{array} \tag{9.1}$$

Una notación equivalente para este planteamiento se tiene para la hipótesis nula: $H_0 : \mu_1 - \mu_2 = 0$ y la hipótesis alternativa: $H_1 : \mu_1 - \mu_2 \neq 0$, $H_1 : \mu_1 - \mu_2 > 0$ y $H_1 : \mu_1 - \mu_2 < 0$

La prueba de hipótesis para la diferencia de medias: muestras grandes

Con el fin de verificar las pruebas planteadas en 9.1, se debe considerar las siguientes condiciones:

- Las muestras deben ser seleccionadas de manera aleatoria.
- Las muestras deben ser independientes. Dos muestras son independientes si la muestra seleccionada de una población no está relacionada con la muestra de la otra población.
- El tamaño de muestra seleccionada de cada población debe ser mayor de 30, o si no cada población debe tener una distribución normal con varianzas conocidas.

Si se cumplen estas condiciones, entonces la distribución muestral de la diferencia de medias $\overline{X}_1 - \overline{X}_2$ es una distribución normal con:

$$\begin{aligned} &\text{media} && \mu_{dm} = media(\overline{X}_1 - \overline{X}_2) = \mu_{\overline{x}_1} - \mu_{\overline{x}_2} = \mu_1 - \mu_2 \\ &\text{y error estándar} && \sigma_{dm} = \sigma(\overline{X}_1 - \overline{X}_2) = \sqrt{\sigma^2_{\overline{x}_1} + \sigma^2_{\overline{x}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

Procedimiento para realizar la prueba de hipótesis $H_0 : \mu_1 = \mu_2$.

1. Plantear las hipótesis (caso: $H_1 : \mu_1 \neq \mu_2$)

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 \\ H_1 &: \mu_1 \neq \mu_2 \end{aligned}$$

2. Determinar el nivel de significancia α. En este planteamiento, se tiene una prueba bilateral, entonces los valores de Z en la normal estándar son: $z_{ci} = z(\alpha/2)$ y $z_{cd} = z(1 - \alpha/2)$.

3. A partir de la información del inciso anterior se calcula la referencia para la prueba, en esta situación los estadísticos son:

$$\begin{aligned} (\overline{x}_1 - \overline{x}_2)_{ci} &= \mu_1 - \mu_2 + z(\alpha/2)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad \text{o,} \\ (\overline{x}_1 - \overline{x}_2)_{cd} &= \mu_1 - \mu_2 + z(1 - \alpha/2)\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned} \tag{9.2}$$

   Estos se comparan con la información de la muestra $(\overline{x}_1 - \overline{x}_2)_m$, es decir, si $(\overline{x}_1 - \overline{x}_2)_m < (\overline{x}_1 - \overline{x}_2)_{ci}$ o $(\overline{x}_1 - \overline{x}_2)_m > (\overline{x}_1 - \overline{x}_2)_{cd}$ se rechaza la hipótesis nula. Nota: el subíndice m se emplea aquí con el fin de hacer énfasis de la diferencia de las medias en la muestra y resaltar la comparación, pero es suficiente $\overline{x}_1 - \overline{x}_2$.

4. Se concluye en función del estudio planteado.

Pruebas de hipótesis utilizando la normal estándar

Los puntos 1 y 2 del procedimiento anterior se repiten, sólo es suficiente con calcular el estadístico z_m para llevar a cabo la comparación con los puntos críticos: z_{ci} y z_{cd}, y dicidir si los datos apoyan o no la hipótesis nula. Este es el método que se usa con frecuencia, ya que sólo se requiere la distribución normal estándar.

Dadas estas características de la distribución muestral $(\overline{x}_1 - \overline{x}_2)_m$, el estadístico de prueba estandarizado toma la forma:

$$z_m = \frac{(diferencia\ observada)_m - (diferencia\ hipótesis)}{Error\ estándar}$$

La expresión del estadístico de prueba para comparar dos medias es: $\overline{x}_1 - \overline{x}_2$ y en forma estandarizado se indica por:

$$z_m = \frac{(\overline{x}_1 - \overline{x}_2)_m - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{9.3}$$

Entonces, si $\overline{z}_m < \overline{z}_{ci}$ o $\overline{z}_m > \overline{z}_{cd}$ se rechaza la hipótesis nula. En resumen, éste es el patrón que se sigue para realizar la prueba de hipótesis en la comparación de dos medias usando la distribución Z (muestras grandes).

Prueba de hipótesis mediante la comparación de probabilidades

De manera análoga a lo planteado en el capítulo 8, también se puede justificar si los datos apoyan a la hipótesis nula, a través del *nivel de significancia descriptivo.* Considerando la diferencia de las medias muestrales, este se obtiene mediante:

$$\begin{aligned} valor - p &= P((\overline{X}_1 - \overline{X}_2) > (\overline{x}_1 - \overline{x}_2)_m) \qquad \text{o,} \\ valor - p &= P((\overline{X}_1 - \overline{X}_2) < (\overline{x}_1 - \overline{x}_2)_m) \end{aligned}$$

En el proceso estandarizado se tiene que:

$$\begin{aligned} valor - p &= P(Z > z_m) \qquad \text{o,} \\ valor - p &= P(Z < z_m) \end{aligned} \tag{9.4}$$

Luego se compara este valor con el nivel de significancia α, esto es:

$$\text{Si, } valor - p > \alpha,$$

no se rechaza la hipótesis nula,

$$\text{Si, } valor - p < \alpha,$$

se rechaza la hipótesis nula.

Intervalo $(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$

Con la información generada por los datos se puede construir el intervalo $(1-\alpha)\,\%$ de confianza para la diferencia de medias $\mu_1 - \mu_2$, cuya interpretación indicará explicaciones importantes del objeto de estudio. Éste se escribe de la siguiente manera:

$$(\overline{x}_1 - \overline{x}_2)_m + z_{ci}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < (\mu_1 - \mu_2) < (\overline{x}_1 - \overline{x}_2)_m + z_{cd}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \tag{9.5}$$

donde $z_{ci} = z(\alpha/2)$ y $z_{cd} = z(1-\alpha/2)$.

Ejemplo 9.1

La administración de una empresa de reparto necesita comprar una línea de autos compactos. Consulta por internet dos marcas diferentes de automóviles, que se denotarán como A y B, con dos años de uso. Seleccionan una muestra de 36 autos para la marca A y de 42 para la B. La marca que en promedio resulte más barata será la que adquieran. μ_1 será la media de la marca A y μ_2 la media de la otra marca. Verificar que hay diferencia en el precio promedio de las marcas de los coches compactos con un nivel de significancia de $\alpha = 0.05$. Estime el intervalo de confianza de 95 %. Como se sabe que ya se tiene habilidad en el cálculo de los estadísticos, la media y la varianza en cada muestra, en miles de pesos, se presentan en la tabla del ejemplo.

Resultados al evaluar el precio de dos marcas de autos compactos

Marca	Número de autos	Media muestral	Desviación estándar muestral
A	$n_1 = 36$	$\overline{x}_1 = 84.6$	$S_1^2 = 73.96$
B	$n_2 = 42$	$\overline{x}_2 = 81.1$	$S_2^2 = 38.44$

Solución clásica operativa

Una vez planteadas las hipótesis, habrá que calcular las expresiones, 9.2, para el caso original. La estandarizada mediante las ecuaciones 9.3, o la del $valor - p$ usando las fórmulas 9.4. Este es el trabajo que se realizará a continuación. Observe que no se conocen las varianzas originales σ_1^2, y σ_2^2, en esta situación se usan las varianzas de la muestra.

1. Hipótesis

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

2. Si $\alpha = 0.05$, se tiene que $z_{ci} = z(\alpha/2) = z(0.025) = -1.96$ y $z_{cd} = z(1-\alpha/2) = z(0.975) = 1.96$.

3. Como se indicó, se sustituyen los valores generados por la muestra en la expresión 9.2, y en el supuesto de que H_0 es cierta $\mu_1 - \mu_2 = 0$, se sigue que:

$$(\overline{x}_1 - \overline{x}_2)_{ci} = \mu_1 - \mu_2 - 1.96\sqrt{\frac{73.96}{36} + \frac{38.44}{42}} = 0 - 3.377 = -3.377 \text{ o,}$$
$$(\overline{x}_1 - \overline{x}_2)_{cd} = \mu_1 - \mu_2 + 1.96\sqrt{\frac{73.96}{36} + \frac{38.44}{42}} = 0 + 3.377 = 3.377$$

   El valor $(\overline{x}_1 - \overline{x}_2)_m = 84.6 - 81.1 = 3.5$, dado que $(\overline{x}_1 - \overline{x}_2)_m = 3.5 > (\overline{x}_1 - \overline{x}_2)_{cd} = 3.377$. Se rechaza la hipótesis nula.

4. La administración preferirá comprar los autos de la marca B. En el contexto de la solución, vale notar que la varianza del precio de los coches de la marca B es más pequeña, hay mayor homogeneidad en este precio, ¿será significativa la comparación entre estas varianzas?

Solución considerando el valor de z

Aquí, únicamente se reproducirá el cálculo del valor de z_m; sustituyendo los valores en la ecuación 9.3 se sigue:

$$z_m = \frac{3.5 - 0}{1.723} = 2.031$$

Como se puede ver, $z_m = 2.03 > z_{cd} = 1.96$, se tiene la misma conclusión anterior.

Solución considerando el $valor- \ p$

La conclusión es similar ya que $valor- \ p$ para una distribución normal con media cero y desviación estándar 1.723 $(N(0, 1.723))$, $valor- \ p = P((\overline{X}_1 - \overline{X}_2) > 3.5) = 0.021$, o para $(N(0,1))$ $valor- \ p = P(Z > 2.031) = 0.021$. Dado que $valor- \ p < \alpha/2 = 0.025$ se rechaza H_0. Verifique estos valores consultando el capítulo 6, para una distribución normal $N(0, 1.723)$ y $N(0,1)$

Intervalo $95\,\%$ de confianza para $\mu_1 - \mu_2$

Con la información generada por los datos se puede construir el intervalo $95\,\%$ de confianza para la diferencia de medias $\mu_1 - \mu_2$, cuya interpretación indicará explicaciones importantes del objeto de estudio. Se sustituyen los valores del ejemplo en la ecuación 9.5:

$$\begin{aligned} 3.5 - 1.96(1.723) &< (\mu_1 - \mu_2) < 3.5 + 1.96(1.723), \\ 0.123 &< (\mu_1 - \mu_2) < 6.877 \end{aligned}$$

Observe que este intervalo no contiene al cero, otra interpretación que indica que los datos no dan evidencia para apoyar la hipótesis nula.

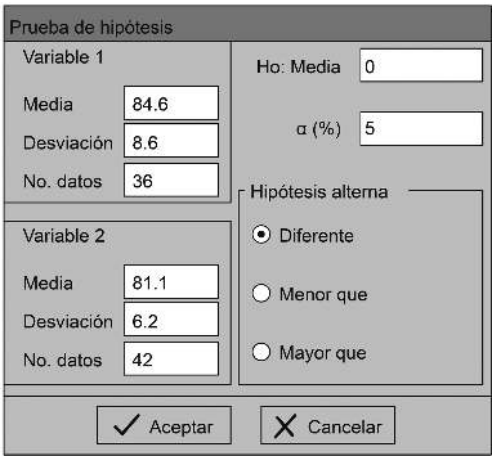


Figura 9.3 Hoja para completar la información y llevar a cabo la prueba de hipótesis.

Solución mediante el uso de CalEst

El procedimiento seguido para hacer la prueba de hipótesis se puede reproducir mediante la ayuda de **CalEst**, lo cual desde luego facilita de manera importante los cálculos, además de que se tiene el apoyo visual de la distribución para identificar las regiones establecidas por los puntos críticos. En la figura 9.2 se ha indicado la descripción del acceso para la prueba de comparación de medias para muestras independientes.

Una vez que aplica esta opción, se tiene la imagen que reporta la figura 9.3, la cual aparece con la información del ejemplo 1. Una vez que se da Aceptar, se tiene un resumen similar al señalado en la solución operativa (ver la figura 9.4), ésta se completa con las gráficas de la distribución normal original y la estándar, y ambas muestran el $valor - p$.

A nivel de conclusión, se observa que una vez que se ha comprendido el proceso de prueba de hipótesis la parte operativa es cómoda usando la tecnología. Nota: para practicar y repasar los conceptos de los errores tipo I y tipo II, se recomienda utilizar la opción didáctica $\mathcal{H}_0$ en **CalEst**.

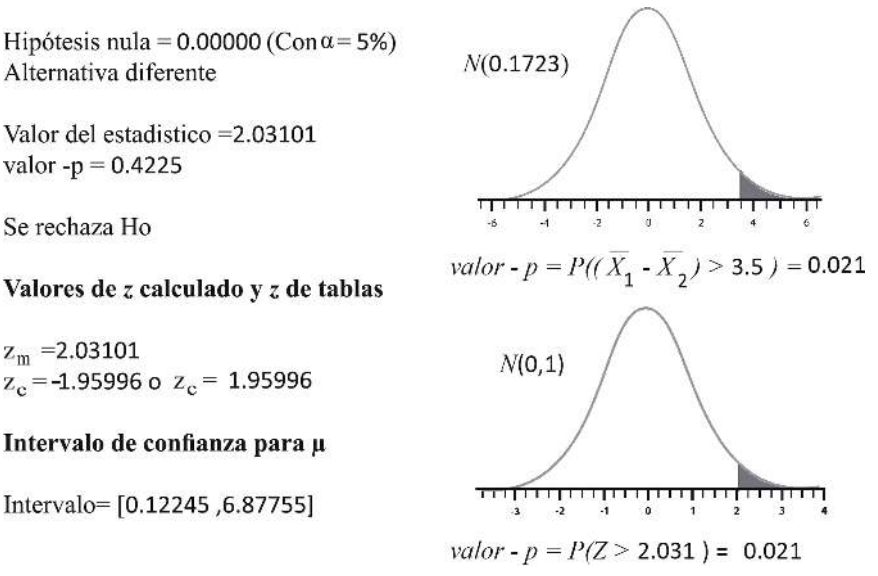


Figura 9.4 Reporte del **CalEst** en la prueba de hipótesis $\mu_1 \neq \mu_2$.

La prueba $t - Student$ para la diferencia de medias: muestras pequeñas

Este es un caso muy práctico para verificar una prueba de hipótesis, ya que se requiere menos información. En esta situación, lo relevante es verificar que los datos vienen de poblaciones que tienen una distribución normal y varianzas desconocidas e iguales. Ante estas observaciones, se usa la prueba $t - Student$ para probar la diferencia de las medias μ_1 y μ_2 entre dos poblaciones cuando una muestra es seleccionada aleatoriamente de manera independiente de cada población. Las expresiones para realizar la prueba son similares al proceso del estadístico z, y se considera la situación equivalente para la distribución $t - Student$. La metodología se enumera a continuación:

Procedimiento para realizar la prueba de hipótesis $H_0 : \mu_1 = \mu_2$, para la $t - Student$.

1. Plantear las hipótesis (caso: $H_1 : \mu_1 \neq \mu_2$)

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

2. Determinar el nivel de significancia α. En este planteamiento se tiene una prueba bilateral, entonces los valores de $t - Student$ son: $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1 - \alpha/2)$. Donde $gl = n_1 + n_2 - 2$

3. A partir de la información del inciso anterior se calcula la referencia para la prueba, en esta situación los puntos críticos o de referencia para verificar la hipótesis nula son:

$$\begin{aligned}(\overline{x}_1-\overline{x}_2)_{ci} &= \mu_1-\mu_2+t(gl,\alpha/2)ES \qquad \text{o,}\\ (\overline{x}_1-\overline{x}_2)_{cd} &= \mu_1-\mu_2+t(gl,1-\alpha/2)ES\end{aligned} \tag{9.6}$$

Donde:

$$S_p=\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$$

La desviación estándar S_p se conoce como *ponderada* y se obtiene en el supuesto de que $\sigma_1^2=\sigma_2^2$. Este último supuesto se verifica mediante una prueba de hipótesis que se verá más adelante. En el contexto inferencia estadística, S_p^2 es la estimación de la varianza para la distribución de la variable aleatoria $\overline{X}_1-\overline{X}_2$ (diferencia muestral de medias). Así S_p es el error estándar de esta distribución.

$$ES=Error\ estándar=S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \tag{9.7}$$

En esta prueba, los puntos críticos se comparan con la información de la muestra $(\overline{x}_1-\overline{x}_2)_m$, es decir, si $(\overline{x}_1-\overline{x}_2)_m<(\overline{x}_1-\overline{x}_2)_{ci}$ o $(\overline{x}_1-\overline{x}_2)_m>(\overline{x}_1-\overline{x}_2)_{cd}$ se rechaza la hipótesis nula. Nuevamente, se indica que el subíndice m se emplea aquí para hacer énfasis en la diferencia de las medias en la muestra y resaltar la comparación, pero es suficiente $\overline{x}_1-\overline{x}_2$. Para el caso de las hipótesis $H_1:\mu_1-\mu_2<0$, o $H_1:\mu_1-\mu_2>0$, de una cola, se muestra en los ejemplos.

4. Se concluye en función del estudio planteado.

Prueba de hipótesis utilizando la $t-Student$

Así como en el caso de la prueba de hipótesis para muestras grandes, los puntos 1 y 2 del procedimiento anterior se repiten, sólo es suficiente con calcular el estadístico t_m para llevar a cabo la comparación con los puntos críticos: t_{ci} y t_{cd}, y decidir si los datos apoyan o no a la hipótesis nula. Este es el método que usa con frecuencia, ya que sólo se requiere la distribución $t-Student$. En la práctica del quehacer estadístico, se tiene una síntesis de esta distribución presentada en una tabla, la cual contiene algunos valores críticos.

Dadas estas características de la distribución muestral $(\overline{x}_1-\overline{x}_2)_m$, el estadístico de prueba estandarizado toma la forma:

$$t_m=\frac{(diferencia\ observada)_m-(diferencia\ hipótesis)}{Error\ estándar}$$

La expresión del estadístico de prueba para comparar dos medias es: $\overline{x}_1-\overline{x}_2$ y en forma estandarizado se indica por:

$$t_m = \frac{(\overline{x}_1 - \overline{x}_2)_m - (\mu_1 - \mu_2)}{S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{9.8}$$

En resumen, el patrón que se sigue para realizar la prueba de hipótesis en la comparación de dos medias usando la distribución $t-Student$ para la situación planteada, la hipótesis bilateral se rechaza si $t_m < t_{ci}$ o si $t_m > t_{cd}$. En el caso de la hipótesis de una cola, la decisión es rechazar la hipótesis H_0, si $t_m < t_c$ o si $t_m > t_c$.

Prueba de hipótesis mediante la comparación de probabilidades

De manera análoga a lo planteado en el capítulo 8, también se puede justificar si los datos apoyan a la hipótesis nula, a través del *nivel de significancia descriptivo.* Considerando las medias muestrales, éste se obtiene mediante:

$$\begin{aligned} valor - p &= P((\overline{X}_1 - \overline{X}_2) > (\overline{x}_1 - \overline{x}_2)_m) \qquad \text{o,} \\ valor - p &= P((\overline{X}_1 - \overline{X}_2) < (\overline{x}_1 - \overline{x}_2)_m) \end{aligned}$$

En el proceso estandarizado es:

$$\begin{aligned} valor - p &= P(t > t_m) \qquad \text{o,} \\ valor - p &= P(t < t_m) \end{aligned} \tag{9.9}$$

Luego se compara este valor con el nivel de significancia α, esto es:

$$\text{Si, } valor - p > \alpha,$$

no se rechaza la hipótesis nula,

$$\text{si, } valor - p < \alpha,$$

se rechaza la hipótesis nula.

Intervalo $(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$

Con la información generada por los datos se puede construir el intervalo $(1-\alpha)\,\%$ de confianza para la diferencia de medias $\mu_1 - \mu_2$, cuya interpretación aportará explicaciones importantes del objeto de estudio, y se escribe de la siguiente manera:

$$(\overline{x}_1 - \overline{x}_2)_m + t_{ci}ES < (\mu_1 - \mu_2) < (\overline{x}_1 - \overline{x}_2)_m + t_{cd}ES \tag{9.10}$$

donde $ES = S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$, y $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1-\alpha/2)$.

Inferencia si $\sigma_1^2 \neq \sigma_2^2$

Si no existe evidencia que confirme el supuesto sobre la igualdad de las varianzas, es decir $\sigma_1^2 \neq \sigma_2^2$, se requiere hacer un ajuste a los grados de libertad de la distribución $t - Student$ para aplicar los procedimientos propuestos de inferencia estadística. A nivel de breviario cultural, no se ha podido determinar una distribución que ajuste de manera exacta este proceso. Se ha propuesto hacer una precisión en los grados de libertad para usar la distribución $t - Student$ como una aproximación. Cuando las varianzas de dos poblaciones son diferentes, $\sigma_1^2 \neq \sigma_2^2$, los grados de libertad adecuados son:

$$gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{9.11}$$

Ejemplo 9.2

Desde el punto de vista de la administración estratégica, las empresas están trabajando continuamente en la corrección de sus procesos para alcanzar mejoras y ser competitivos. La finalidad es ganar mercados, pero también lograr ahorros. En un proceso de moldeo por inyección una compañía tiene problemas asociados al encogimiento de las piezas producidas después del curado. Este encogimiento contribuye a una creciente variabilidad del producto. El objetivo del grupo de trabajo es determinar mediante una estrategia experimental los factores que contribuyen a disminuir el encogimiento y reducir la variabilidad del producto. En este ejemplo se tiene la relación del encogimiento de una muestra de 24 piezas del proceso antes de llevar a cabo los cambios, y a cada una de ellas se le midió el encogimiento. Después de realizar las innovaciones señaladas por los resultados del experimento, se tomó una muestra, también de 24 piezas, y se observó el encogimiento. Se plantea verificar mediante una prueba de hipótesis, con $\alpha = 0.01$, si realmente los cambios indican una diferencia en el encogimiento. Estimar mediante un intervalo de confianza de 99 %. Se cuenta con la referencia para evaluar el impacto económico de la mejora, la cual se presentará en la interpretación. Una compilación estadística de la información arrojada por las mediciones en el proceso original y después del cambio se registra en la tabla de abajo.

Resumen estadístico del moldeo de inyección

Proceso	Número de unidades	Media muestral	Desviación estándar muestral
Original	$n_1 = 24$	$\overline{x}_1 = 0.147$	$S_1^2 = 5.085x10^{-4}$
Mejora	$n_2 = 24$	$\overline{x}_2 = 0.099$	$S_2^2 = 2.288x10^{-4}$

Planteamiento de las hipótesis:

$$H_0 : \mu_1 = \mu_2, \text{ no hay diferencia.}$$
$$H_1 : \mu_1 \neq \mu_2, \text{ existe diferencia en el encogimiento.}$$

Solución operativa clásica

Se hacen las operaciones empleando las fórmulas 9.6, 9.8, 9.9 y 9.10. Luego se presentan en una tabla los pasos de la prueba de hipótesis (ver figura 9.5). Cálculos básicos:

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{(23)5.085x10^{-4} + (23)2.288x10^{-4}}{46}} \doteq 0.0192$$

$$ES = Error\ estándar = S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 0.0192\sqrt{\left(\frac{1}{24}+\frac{1}{24}\right)} \doteq 5.543x10^{-3}$$

Estimación de los valores críticos estandarizados:

$$t_{ci} = t(gl, \alpha/2) = t(46, 0.025) = -2.013 \text{ y } t_{cd} = t(46, 0.975) = 2.013.$$

Estimación de los valores críticos:

$$\begin{aligned}(\overline{x}_1 - \overline{x}_2)_{ci} &= \mu_1 - \mu_2 + t(gl, \alpha/2)ES = 0 - 2.013(5.543x10^{-3}) = -0.011 \text{ o,}\\ (\overline{x}_1 - \overline{x}_2)_{cd} &= \mu_1 - \mu_2 + t(gl, 1-\alpha/2)ES = 0 + 2.013(5.543x10^{-3}) = 0.011\end{aligned}$$

Cálculo del estadístico:

$$(\overline{x}_1 - \overline{x}_2)_m = 0.147 - 0.099 = 0.048$$

El estadístico estandarizado:

$$t_m = \frac{(\overline{x}_1 - \overline{x}_2)_m - (\mu_1 - \mu_2)}{S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \doteq \frac{0.048 - 0}{5.543x10^{-3}} \doteq 8.72$$

Nota: el símbolo $\doteq$ es una representación debido al redondeo a milésimas. Valor del nivel de significancia descriptivo:

$$valor - p = P(t > t_m) = P(t > 8.72) = 2.621x10^{-11}$$

Procedimiento para obtener el intervalo de confianza:

$$0.048 - 0.011 < (\mu_1 - \mu_2) < 0.048 + 0.011$$

El intervalo de confianza:

$$0.037 < (\mu_1 - \mu_2) < 0.059$$

El impacto económico de la mejora implicó la siguiente ganancia, el costo por unidad en el proceso original era de 3.88 pesos, con la mejora el costo de la unidad es de 1.50 pesos. Así que por unidad hay una diferencia de $3.88 - 1.50 = 1.38$ pesos de ahorro. En cuanto a la producción masiva se obtiene un ahorro sustancial.

Resumen de la prueba de hipótesis estadistica
1. $H_0 : \mu_1 = \mu_2$ vs $H1 : \mu_1 \neq \mu_2$
2. Determinar el punto crítico calculado $t_{ci} = -2.013$, $t_{cd} = 2.013$
$(\overline{x_1} - \overline{x_2})_m = 0.48 > (\overline{x_1} - \overline{x_2})_{cd} = 0.11$
3. $t_m = 8.72 > t_{cd} = 2.013$
$valor\text{-}p = 2.621 \text{x} 10^{-11} << \alpha/2 = 0.025$
4. Se rechaza H_0 lo que indica que la mejora fue exitosa
5. El intervalo del 95% de confianza $0.037 < (\mu_1 - \mu_2) < 0.059$

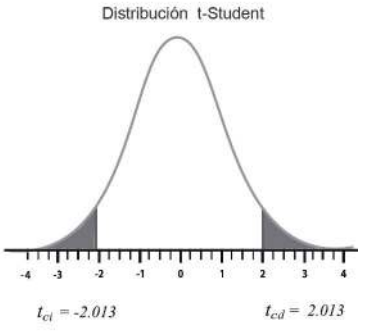


Figura 9.5 Metodología de la prueba de hipótesis usando la $t - Student$.

Ejemplo 9.3

En una investigación, la administración del deporte de alto rendimiento tiene programas de entrenamiento e inicialmente se desea comparar el desempeño atlético de dos grupos de jóvenes. El primero es un grupo control y el otro es un grupo entrenado. La capacidad de recuperación indica si el efecto del entrenamiento es adecuado. La evaluación física después de completar un programa para ambos grupos se muestra en la tabla:

Control	83	91	97	97	108	111	111	117	117	125	125	140				
Entrenado	64	83	83	85	91	97	97	97	103	108	111	111	117	117	125	125

Se supone que el grupo entrenado tiene un mejor rendimiento que el grupo control.

$$H : \text{El grupo entrenado } \mu_2 \text{ tiene un mejor rendimiento que el grupo control } \mu_1.$$

Eso indica que la media del grupo entrenado debe ser menor que la del grupo control, ya que se espera que su recuperación sea más rápida. Para confirmar el supuesto, se plantean las siguientes hipótesis:

$$H_0 : \mu_1 = \mu_2 \ (\mu_1 - \mu_2 = 0)$$
$$H_1 : \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0)$$

la que se desea verificar con un nivel de significancia del 0.10. Complete el ejemplo y estime el intervalo de 90 % confianza para la diferencia entre estos grupos.

Solución mediante el uso de CalEst

Una vez que se ha adquirido una maduración y habilidad en la prueba de hipótesis para la diferencia de medias, en esta ocasión se desea aprovechar el uso de la tecnología para realizar la inferencia estadística sobre algún estudio que se esté realizando, debido a que la tecnología facilita la parte opertiva. Para ello se mostrará el procedimiento desde la captura hasta la salida de la información. Usando el reporte elaborado por este material de cómputo, se puede reproducir la metodología de prueba de hipótesis planteada y aplicada en los subapartados anteriores.

1. Se capturan los datos en **CalEst**: ir primero a la opción Archivo, crear una hoja de cálculo y capturar los datos. El nombre de la columna se escribe sobre la columna o usando el ratón al marcar el extremo izquierdo del rectángulo. La descripción de este procedimiento se muestra en la figura 9.6.

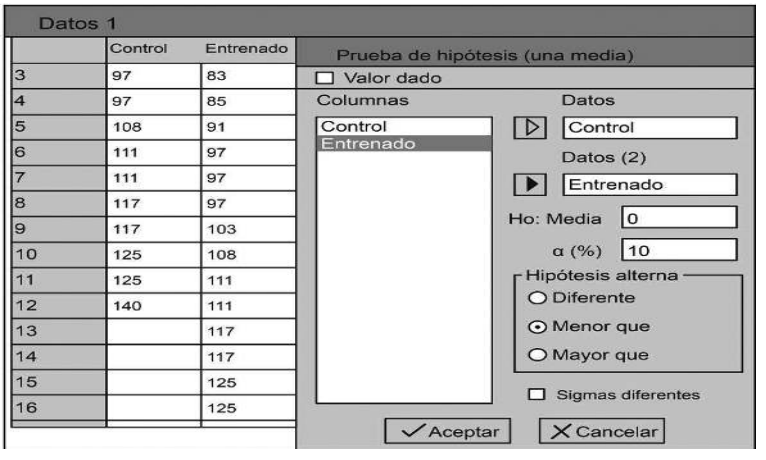


Figura 9.6 Hojas de captura y referencia para la diferencia de medias.

2. Aplicando la alternativa que aparece en la figura 9.2, se realiza la prueba t. La descripción de los resultados que aparecen en la figura 9.7, es la hoja de salida calculada por la parte operativa del calculador estadístico. Con ese reporte se hace la interpretación.

3. Se rechaza la hipótesis nula, lo que indica que la recuperación del grupo entrenado es más rápido. Esta conclusión se sigue de las siguientes observaciones:

   a) Se observa el valor crítico de referencia $t_c = 1.315$ y el de la muestra es $t_m = 1.571$, $t_m > t_c$.

   b) Un cálculo extra indica que $(\overline{x}_1 - \overline{x}_2)_c = \mu_1 - \mu_2 + t(26, 90)ES = 0 + 1.315(10.631) = 8.199$. El ES calculado mediante la ecuación 9.7. Del resumen estadístico se tiene que $(\overline{x}_1 - \overline{x}_2)_m = 9.792$, comparando con el valor crítico $(\overline{x}_1 - \overline{x}_2)_c$, se ve que $(\overline{x}_1 - \overline{x}_2)_m > (\overline{x}_1 - \overline{x}_2)_c$.

   c) El $valor - p = 0.064 > \alpha = 0.1$

4. Finalmente, usando la expresión 9.10, el intervalo del 90 % de confianza es: $-0.83 < (\mu_1 - \mu_2) < 20.42$. Indica la diferencia de las medias en la capacidad de recuperación entre estos dos grupos.

.

Prueba de hipótesis para dos medias independientes

Hipótesis nula = 0.00000 (Con α = 10)
Alternativa mayor que

Valor del estadístico = 1.5709435753924
valor –p = 0.0641433645318017

Se rechaza H_0

Valores de *t* calculado y *t* de tablas

$t_m = 1.57094$
$t_c = 1.31544$

Intervalo de confianza para de:

intervalo = [-0.839. 20.432]

Resumen estadístico

	Muestra 1	Muestra 2
Observaciones	12	16
Media	110.66667	100.87500
Mediana	111.00000	100.00000
Varianza	235.33333	289.18333
Desviación estándar	153.34058	17.00539
Error estándar	4.42844	4.25135
Rango	51.00000	61.00000
Máximo	140.00000	125.00000
Mínimo	189.00000	64.00000

Figura 9.7 Hoja de salida de los cálculos para efectuar la prueba t.

9.4 Prueba de hipótesis para la diferencia de medias: muestras pareadas

El mundo de la información 2. Antes y después

En diferentes actividades que realizamos de manera cotidiana, surge el interés por conocer si una idea o una propuesta darán mejores resultados. Estos planteamientos van desde temas sencillos hasta complejos, a continuación se plantean algunos.

Con el propósito de ahorrar energía se propone cambiar los focos tradicionales por focos llamados ahorradores. En consecuencia se espera economizar el consumo de energía en Kw/hora. ¿Cómo realizar el estudio para ver si se logra tal fin? Por ejemplo, seleccionar una muestra de 13 casas al azar de una colonia, luego comparar el consumo de energía un bimestre antes, es decir con los focos tradicionales y otro después ya con los focos ahorradores. ¿Cuál es la hipótesis?

En la administración de la planeación educativa se sugiere que si a los profesores de educación primaria (vale para cualquier nivel) se les da una capacitación y actualización en varios temas, esto repercutirá en

el aprendizaje y aprovechamiento de los estudiantes. Se toma una muestra, digamos 15 estudiantes, y se les evalúa. Posterior a la preparación del profesor, se toma otra muestra de 15 estudiantes de los mismos profesores y se compara. La descripción de cómo se selecciona la muestra contempla tener el antes y después, mientras que la homogeneidad de la unidad viene del profesor. Ante este planteamiento, ¿cuál es la hipótesis?

En el estudio de muestras pareadas, considere la siguiente situación: la administración de una compañía que fabrica llantas quiere comparar su llanta XP-500 con la similar XR de la competencia. La evaluación se hará mediante el desgaste de la llanta. La verificación se efectúa recorriendo una distancia de mil kilómetros en una de las carreteras que la compañia utiliza para tal fin. Los procedimientos propuestos por el administrador son:

1. Selecciona 14 automóviles de la misma marca y modelo. A 7 de esos les ponen la llanta P y a los otros la llanta R.

2. Selecciona 7 automóviles al azar, y selecciona también al azar 2 de las 4 posiciones y en esas coloca la llanta P, mientras que en las dos restantes va la llanta R.

Así compara el desgaste de las llantas, considerando el antes y después. ¿Qué procedimiento es mejor? ¿Cuál es la hipótesis de investigación? ¿Cuál es la hipótesis nula? ¿Cuál es la hipótesis alternativa? Se pueden plantear proyectos para el cambio de valores, de apreciaciones, de conducta, de rendimiento, así como comparar planes de seguros o económicos, entre muchos otros. Para completar la idea, un administrador tiene la intención de mejorar el ambiente laboral. Para tener una idea de esa situación, aplica un cuestionario a los empleados sobre este punto, el antes. Luego realiza una serie de actividades tendientes a mejorar el ambiente mediante videos, programas de capacitación, entrevistas. Vuelve a aplicar el cuestionario, el después, y hace la comparación. ¿Cuál es la hipótesis?

Observación: las unidades experimentales o de observación deben ser organizadas lo más parecido posible.

Planteamiento del problema muestras pareadas

Una categoría de estudio que da lugar a verificar la diferencia de medias entre dos poblaciones consiste en seleccionar muestras de cada una de ellas, donde las unidades de observación o experimentales sean las mismas o lo más homogéneo posible. La idea general del proceso radica en aplicar a una muestra aleatoria cierta acción y luego a la misma muestra o a una lo más similar posible otra acción. A continuación evaluar mediante una prueba de hipótesis si existe diferencia entre estas acciones, es decir la que se aplica antes y la que se realiza después. A este tipo de procedimiento se le denomina muestras pareadas, ya que corresponde a los mismos individuos o que éstos sean lo más semejante posible.

La población estará descrita por diferencias que corresponderían a las características antes y después de la acción, eso es:

$$\textit{Población de diferencias: } \{d_1, d_2, d_3, \ldots, d_N\}$$

La población de diferencias tiene N unidades, donde $d_1 = x_{a1} - x_{d1}, d_2 = x_{a2} - x_{d2}, ..., d_n = x_{an} - x_{dn}$.

La variable aleatoria $d = X_a - X_d$ representa la diferencia de las cualidades antes y después de la acción. Distribución de probabilidad de la variable:

$d = X_a - X_d$ tiene una distribución de probabilidad normal con media $\mu_d = \mu_1 - \mu_2$
y varianza $\sigma_d^2 = \frac{1}{N}\sum_{i=1}^{N}(d_i - \mu_d)^2$, así $d \sim N(\mu_d, \sigma_d^2)$

La conjetura que se plantea es establecer si existe un cambio en la diferencia de poblaciones al aplicar las acciones antes y después. Esto da lugar a plantear la hipótesis siguiente:

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

La hipótesis nula indica que no existe efecto al aplicar las acciones. ¿Cuál es el mecanismo para verificar esta prueba?

La técnica para comparar dos poblaciones mediante muestras pareadas al realizar un experimento es:

1. Seleccionar la muestra de manera aleatoria.

Muestra antes: $\{u_{a1}, u_{a2}, u_{a3}, ..., u_{an}\}$	Muestra después: $\{u_{d1}, u_{d2}, u_{d3}, ..., u_{dn}\}$

2. Las muestras son pareadas (*dependientes*), cada una de las unidades en la primera muestra deben ser la misma en la segunda muestra (*par*), o lo más homogénea posible.
3. Ambas poblaciones tienen una distribución normal.
4. x_{ai} y x_{di} caracterizan las mediciones de las i unidades en cada una de las muestras. Así los datos que se obtienen estan representados por:

$$d_i = x_{ai} - x_{di}$$

En resumen, las estadísticas son:

$$\text{tamaño de la muestra: } n \quad total = \sum_{i=1}^{n} d_i \quad \overline{d} = \frac{\sum_{i=1}^{n} d_i}{n} \quad S_d^2 = \frac{\sum_{i=1}^{n}\left(d_i - \overline{d}\right)^2}{n-1} \tag{9.12}$$

En síntesis, la forma en que se lleve a cabo este experimento da lugar a muestras dependientes, generalmente denominadas *muestras pareadas.* La estrategia consiste en experimentar sobre la misma unidad experimental aplicando un tratamiento después. En caso de que las unidades no puedan ser las mismas se buscará que los pares sean lo más homogéneo posible.

Para verificar la hipótesis se recurre a la metodología planteada en el capítulo 8. Para este nuevo escenario, lo que se requiere es identificar el punto de referencia, es decir: valor del punto crítico. Como las muestras por lo general son pequeñas se usará la distribución $t - Student$.

Síntesis de la selección y análisis de muestras pareadas

Pa	Pd	Parámetros	Antes	Después	Diferencias
x_{a1}	x_{d1}		x_{a1}	x_{d1}	$d_1 = x_{a1} - x_{d1}$
x_{a2}	x_{d2}		x_{a2}	x_{d2}	$d_2 = x_{a2} - x_{d2}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	
x_{aN_1}	x_{dN_2}	$\mu_d = \mu_1 - \mu_2$ $\sigma_d^2 = \frac{1}{N}\sum(d_i - \mu_d)^2$	x_{an}	x_{dn} $S_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \overline{d})^2$	$d_n = x_{an} - x_{dn}$ $\overline{d} = \frac{1}{n}\sum d_i$
Distribución de probabilidad de X_a					
Distribución de probabilidad de X_d				Distribución de probabilidad del estimador $\overline{x}_a - \overline{x}_d$	
Pa: Población 1 Pd: Población 2					

Procedimiento para verificar una prueba en muestras pareadas

1. Plantear las hipótesis (caso: $H_1 : \mu_d \neq 0$)

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

2. Determinar el nivel de significancia α. En este planteamiento se tiene una prueba bilateral, entonces los valores de $t - Student$ en la normal estándar son: $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1 - \alpha/2)$. Dadas estas condiciones se sigue que la distribución de la diferencia de medias $\overline{d}$ tiene una distribución t de Student con $n - 1$ grados de libertad.

3. A partir de la información del inciso anterior se calculan los valores críticos que son los puntos de referencia para la prueba, en esta situación los valores del estadístico son:

$$\begin{aligned} \overline{d}_{ci} &= \mu_d + t(gl, \alpha/2)\frac{S_d}{\sqrt{n}} \text{ o,} \\ \overline{d}_{cd} &= \mu_d + t(gl, (1 - \alpha/2))\frac{S_d}{\sqrt{n}} \end{aligned} \tag{9.13}$$

   Estos se comparan con la información de la muestra $\overline{d}_m$, es decir, si $\overline{d}_m < \overline{d}_{ci}$ o $\overline{d}_m > \overline{d}_{cd}$ se rechaza la hipótesis nula.

4. Se concluye en función del estudio planteado.

Para el caso estandarizado: Dadas estas condiciones, se sigue que la distribución de la diferencia de medias $\overline{d}$ tiene una distribución t de Student con $n - 1$ grados de libertad. Así, el estadístico es:

$$t_m = \frac{\overline{d}_m - \mu_d}{\frac{S_d}{\sqrt{n}}}, \quad gl = n - 1 \tag{9.14}$$

Finalmente para concluir, se comparan t_m y t_{ci} o con t_{cd}.

Ejemplo 9.4

Para incrementar las ventas, la administración de una empresa que vende alimentos, decidió hacer una propaganda, para la cual hacen una inversión. Para dar seguimiento a su propuesta, tomaron una muestra de 10 de sus diferentes locales sobre los cuales se hará el estudio. Observaron el monto de las ventas, indicadas por miles de pesos a la semana, antes y después de la promoción. Los resultados de antes y después se anotan en la siguiente tabla:

Antes	273	246	233	286	280	286	289	246	280	283
Después	242	245	242	279	282	231	308	279	291	269
Diferencia	31	1	-9	7	-2	55	-19	-33	-11	14

Mediante una prueba de hipótesis se puede verificar si el gasto en promoverse les permitió aumentar el volumen de ventas.

Solución

Planteamiento de las hipótesis:

$$H_0 : \mu_d = 0, \text{ no resultó efectiva la publicidad.}$$
$$H_1 : \mu_d < 0, \text{ resultó efectiva la publicidad.}$$

Operaciones. Los cálculos estadísticos para la diferencia son:

$$n = 10, \quad \sum_{i=1}^{10} d_i = 34, \quad \overline{d}_m = 3.4, \quad S_d = 25.325$$

Se ha propuesto un nivel de significancia de $\alpha = 0.05$, considerando que el tamaño de muestra es pequeño se emplea la distribución t $-Student$ con $gl = n - 1$ grados de libertad. Determinar el punto crítico calculando $t_c = t(9, 0.05) = -1.834$. Atendiendo que la alternativa es $H_1 : \mu_d < 0$, el valor de referencia para la diferencia es:

$$\overline{d}_c = \mu_d + t(gl, \alpha/2)\frac{S_d}{\sqrt{n}} = 0 - 1.834(\frac{25.325}{\sqrt{10}}) = -14.688.$$

El estadístico estandarizado es:

$$t_m = \frac{\overline{d}_m - \mu_d}{\frac{S_d}{\sqrt{n}}} = \frac{3.4 - 0}{8.009} = 0.424.$$

así $t_m > t_c = -1.834$.

Conclusión. Con esta información no se rechaza la hipótesis nula, dado que:

1. $\overline{d}_m = 3.4 > \overline{d}_c = -14.688$, con los datos de la muestra, o

2. $t_m = 0.424 > t_c = -1.834$, en el proceso de estandarización, o

3. $valor - p = 0.662 > \alpha = 0.05$, usando el nivel de significancia descriptivo $valor - p$.

Estas dos últimas relaciones se recogen en la figura 9.8. Por lo tanto, la inversión en la promoción para incrementar las ventas no resultó relevante.

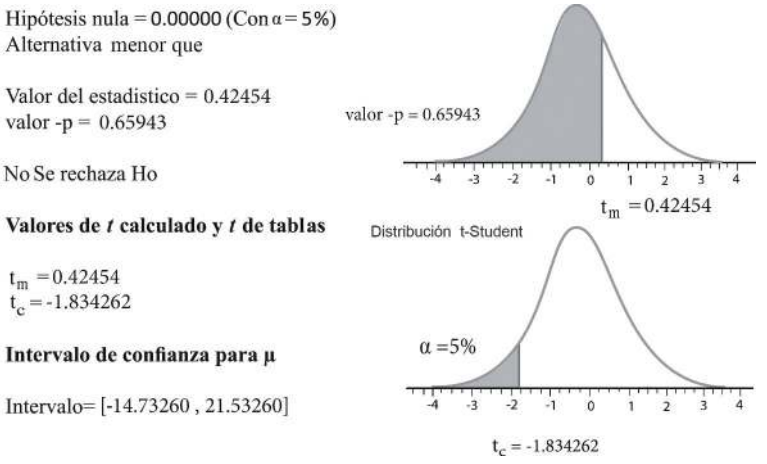
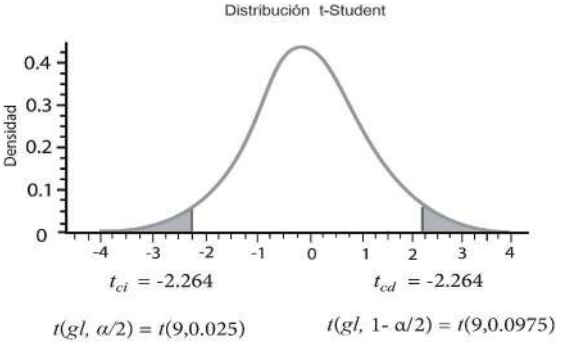


Figura 9.8 Reporte e interpretación de la prueba de hipótesis para muestras pareadas.

Finalmente el intervalo de confianza: consultando los valores de la distribución $t - Student$, para un nivel de 95 %, se tiene que $t_{ci} = t(gl, \alpha/2) = t(9, 0.025) = -2.264$ y $t_{cd} = t(gl, 1 - \alpha/2) = t(9, 0.975) = 2.264$.



Distribución t-student para el intervalo de confianza.

$$LI = (\overline{d})_m + t_{ci}\frac{S_d}{\sqrt{n}}, LD = (\overline{d})_m + t_{cd}\frac{S_d}{\sqrt{n}}$$
$$LI \leq \mu_d \leq LD$$
$$3.4 - 2.264(8.009) \leq \mu_d \leq 3.4 + 2.264(8.009)$$
$$-14.733 \leq \mu_d \leq 21.532$$

Eso quiere decir que con una confianza de 95 % las ventas oscilarán entre una mejora de aproximadamente 15 mil pesos y una disminución de 21.5 mil pesos. En términos prácticos, indica que la publicidad no produjo aumento en las ventas.

Resumen para estimar intervalos $(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$

Para las diferencia de medias a continuación se describen en la siguiente tabla los cálculos para su estimación en intervalos del $(1-\alpha)\,\%$ de confianza, considerando los diferentes tamaños de muestra.

Intervalos $(1-\alpha)\,\%$ de confianza para $\mu_1 - \mu_2$

Muestras Independientes	Muestras Grandes	$LI = (\overline{x}_1 - \overline{x}_2)_m + z_{ci}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $LD = (\overline{x}_1 - \overline{x}_2)_m + z_{cd}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ $z_{ci} = z(\alpha/2)$ y $z_{cd} = z(1-\alpha/2)$ Utilice S_1^2 y S_2^2, si σ_1^2 y σ_2^2 son desconocidos
Muestras Independientes	Muestras pequeñas	$LI = (\overline{x}_1 - \overline{x}_2)_m + t_{ci}S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $LD = (\overline{x}_1 - \overline{x}_2)_m + t_{cd}S_p\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1-\alpha/2)$ $gl = n_1 + n_2 - 2$, si $\sigma_1^2 = \sigma_2^2$ $gl'(1.11)$, si $\sigma_1^2 \neq \sigma_2^2$

Muestras Pareadas	$LI = (\overline{d})_m + t_{ci}\frac{S_d}{\sqrt{n}}$ $LD = (\overline{d})_m + t_{cd}\frac{S_d}{\sqrt{n}}$ $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1-\alpha/2)$ gl=$n-1$

9.5 Prueba de hipótesis para la diferencia de proporciones

Introducción. En el tema de la comparación de dos proporciones, se indicará una serie de casos que resultan relevantes en estudios de la vida cotidiana. Aquí se hará una mención de ellos resaltando las preguntas sobre la naturaleza del problema, sobre todo de actividades que están involucradas con la administración y la economía.

- El ausentismo es una situación en la que se involucra una serie de cambios en la administración ya sea como prevención o estrategias de hacer frente al problema ya que indirectamente afecta en la economía de las empresas. ¿El porcentaje de ausentismo, al semestre, entre dos turnos laborales es similar? Este porcentaje, ¿es mayor entre trabajadores menores o mayores de 35 años?
- En estrategias de mercado, los administradores están al pendiente de ir a la vanguardia en la forma de ganar más clientes. Por ello resulta importante estudiar el porcentaje entre hombre y mujeres que compran por internet. En este contexto, también existe interés en conocer si hay alguna diferencia entre el porcentaje cuando los compradores son profesionistas o no profesionistas.
- El impacto económico que genera un laboratorio al producir un nuevo medicamento, ¿existe una diferencia significativa en los porcentajes de personas que se mejoran usando el nuevo remedio al anterior? En esa misma dirección se puede evaluar: ¿hay diferencia significativa entre el porcentaje de efectividad que producen dos medicinas diferentes en el tratamiento de una enfermedad?

Prueba de hipótesis para dos proporciones

Con la finalidad de abordar estudios como los mencionados, es importante observar que de inicio se tienen dos poblaciones. El planteamiento general es: considere dos poblaciones, la primera con N_1 unidades, la segunda con N_2 unidades, es decir:

$$\text{Población 1: } \{u_{11}, u_{12}, u_{13}, ..., u_{1N_1}\} \text{ y} \quad \text{Población 2: } \{u_{21}, u_{22}, u_{23}, ..., u_{2N_2}\}$$

A las cuales se les mide una característica, que es la misma para cada población. Para una oportunidad las variables W_1 y W_2 se describen por:

$$W_1 = \begin{cases} 1 & \text{si se cumple con la característica en la población 1 (éxito)} \\ 0 & \text{no se cumple con la característica en la población 1 (fracaso)} \end{cases}$$

$$W_2 = \begin{cases} 1 & \text{si se cumple con la característica en la población 1 (éxito)} \\ 0 & \text{no se cumple con la característica en la población 1 (fracaso)} \end{cases}$$

Ahora considere las variables: X_1 el número de éxitos en un total de N_1 oportunidades, y X_2 el número de éxitos en un total de N_2 oportunidades. Entonces las proporciones de ambas poblaciones son:

$$p_1 = \frac{X_1}{N_1},\ p_2 = \frac{X_2}{N_2}$$

A partir de este planteamiento se puede preguntar: ¿existe diferencia entre las proporciones de esas poblaciones? ¿Cómo se compara? ¿Cómo se estima la diferencia?

Para el caso del muestreo se hace una selección aleatoria de unidades de tamaño n_1 y n_2 en cada población respectivamente. Las muestras son:

Muestra 1: $\{u_{11}, u_{12}, u_{13}, ..., u_{1n_1}\}$	Muestra 2: $\{u_{21}, u_{22}, u_{23}, ..., u_{2n_2}\}$

donde n_1 es una muestra de N_1 y n_2 es una muestra de N_2; las proporciones estimadas son:

$$\widehat{p}_1 = \frac{X_1}{n_1},\ \widehat{p}_2 = \frac{X_2}{n_2}$$

Donde X_1 el número de éxitos en un total de n_1 casos, y X_2 el número de éxitos en un total de n_2 casos. Las condiciones que se deben seguir para realizar la comparación entre proporciones son:

1. Seleccionar las muestras de manera aleatoria.
2. Las muestras deben ser independientes.
3. Las muestras deben ser suficientemente grandes para usar, como una aproximación, la distribución normal, así se debe cumplir:

$$n_1 p_1 \geq 5,\ n_1(1-p_1) \geq 5,\ n_2 p_2 \geq 5 \quad y \quad n_2(1-p_2) \geq 5.$$

Dadas estas condiciones se tiene que la distribución muestral de $\widehat{p}_1 - \widehat{p}_2$, diferencia entre proporciones, es una *normal con media:*

$$\mu(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2,$$

y un *error estándar:*

$$\sigma(\widehat{p}_1 - \widehat{p}_2) = \sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

donde:

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2},\ x_1 = n_1\widehat{p}_1,\ x_2 = n_2\widehat{p}_2.$$

Intervalo del $(1-\alpha)\,\%$ **confianza para** $\mu(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$.

Observe que el estadístico $\widehat{p}_1 - \widehat{p}_2$ es un estimador de $\mu(\widehat{p}_1 - \widehat{p}_2) = p_1 - p_2$. La distribución de probabilidad de $\widehat{p}_1 - \widehat{p}_2$ es una normal con media $\mu(\widehat{p}_1 - \widehat{p}_2)$ y varianza $\sigma^2(\widehat{p}_1 - \widehat{p}_2)$. Como se ha explicado en los capítulos 7 y 8, con esta información se puede construir el intervalo de confianza para la media $\mu(\widehat{p}_1 - \widehat{p}_2)$, entonces el *intervalo del* $(1-\alpha)\,\%$ *confianza para la diferencia de proporciones* $p_1 - p_2$ es:

$$\begin{aligned} &LI = (\widehat{p}_1 - \widehat{p}_2)_m + z(\alpha/2)\sqrt{\overline{p}(1-\overline{p})\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)}, \\ &LD = (\widehat{p}_1 - \widehat{p}_2)_m + z(1-\alpha/2)\sqrt{\overline{p}(1-\overline{p})\left(\tfrac{1}{n_1} + \tfrac{1}{n_2}\right)} \\ &LI \leq p_1 - p_2 \leq LD \end{aligned} \tag{9.15}$$

Elementos para la prueba de hipótesis

Entonces para probar la hipótesis de diferencia entre dos proporciones p_1 y p_2 cuando las muestras se extraen de manera aleatoria de dos poblaciones, el *estadístico de prueba* es:

$$(\widehat{p}_1 - \widehat{p}_2)_m \tag{9.16}$$

los puntos críticos para comparar el estadístico y realizar la prueba de hipótesis $H_0 : p_1 = p_2$ vs $H_1 : p_1 \neq p_2$, son:

$$\begin{aligned} (\widehat{p}_1 - \widehat{p}_2)_{ci} &= (p_1 - p_2)_{H_0} + z_{ci}(\alpha/2)\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ (\widehat{p}_1 - \widehat{p}_2)_{cd} &= (p_1 - p_2)_{H_0} + z_{cd}(1-\alpha/2)\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \end{aligned} \tag{9.17}$$

En $(p_1 - p_2)_{H_0}$ indica que se está bajo el supuesto de que la hipótesis nula es verdadera.

El *estadístico prueba en forma estandarizado:*

$$z_m = \frac{(\widehat{p}_1 - \widehat{p}_2)_m - (p_1 - p_2)_{H_0}}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{9.18}$$

Metodología para realizar la prueba $H_0 : p_1 = p_2$ **vs** $H_1 : p_1 \neq p_2$

1. Identificar las hipótesis nula y alternativa:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 \neq p_2$$

2. Especificar el nivel de significancia, dar el valor de α.
3. Determinar los puntos críticos,
   a) Usando la ecuaciones 9.17, luego contrastar con el estadístico 9.16, es decir:

$$¿(\widehat{p}_1 - \widehat{p}_2)_m < (\widehat{p}_1 - \widehat{p}_2)_{ci}?$$
$$¿(\widehat{p}_1 - \widehat{p}_2)_m > (\widehat{p}_1 - \widehat{p}_2)_{cd}?$$

   b) Si se usa el estadístico Z, entonces se comparan los puntos críticos $z_{ci} = z(\alpha/2)$ y $z_{cd} = z(1-\alpha/2)$ con z_m indicada por la expresión 9.18. De manera similar, como en el punto 3 se tiene que:

$$¿z_m < z_{ci}?$$
$$¿z_m > z_{cd}?$$

   c) Mediante el $valor - p$

$$¿valor - p < \alpha/2?$$
$$¿valor - p < (1-\alpha/2)?$$

   donde el $valor - p = P((\widehat{p}_1 - \widehat{p}_2) < (\widehat{p}_1 - \widehat{p}_2)_m)$ o $valor - p = P((\widehat{p}_1 - \widehat{p}_2) > (\widehat{p}_1 - \widehat{p}_2)_m)$ para el estadístico $(\widehat{p}_1 - \widehat{p}_2)$. Así el $valor - p = P(Z < z_m)$ o $valor - p = P(Z > z_m)$ para el estadístico z.

4. Si la respuesta es afirmativa en alguna de las relaciones anteriores se rechaza la hipótesis nula. Y se interpreta para el estudio en cuestión.

Los casos:

$$H_0 : p_1 = p_2$$
$$H_1 : H_1 : p_1 > p_2, \text{ o } \ H_1 : p_1 < p_2$$

se muestran en los siguientes dos ejemplos.

Ejemplo 9.5

La administración de una empresa que se dedica a estudios de mercado hace un análisis sobre la proporción de la compra de productos por internet. Se tiene la creencia de que los hombres compran más que las mujeres. Se obtuvo una muestra de 310 personas, de ellas 150 son hombres y 54 de éstos indicaron que compraban por internet, mientras que de las 160 mujeres fueron 40 las que se dedicaban a ese fin. Se propone un nivel de significancia $\alpha = 0.05$

Solución al ejemplo 5

Los datos generados en este estudio son:

Muestra	Proporciones	Tamaño	
Hombres	$\hat{p}_1 = 0.36$	$n_1 = 150$	$x_1 = 54$
Mujeres	$\hat{p}_2 = 0.25$	$n_2 = 160$	$x_2 = 40$

Se cumplen las condiciones:

$$\begin{aligned} n_1 p_1 &= 150(0.36) \geq 5,\ n_1(1-p_1) = 150(0.64) \geq 5, \\ n_2 p_2 &= 160(0.25) \geq 5\ y\ n_2(1-p_2) = 160(0.75) \geq 5. \end{aligned}$$

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{54 + 40}{150 + 160} = \frac{94}{310} = 0.30$$

$$(\widehat{p}_1 - \widehat{p}_2)_m = (0.36 - 0.25) = 0.11$$

Las operaciones para $\alpha = 0.05$, por lo que $z_c(0.95) = 1.645$

$$\begin{aligned} (\widehat{p}_1 - \widehat{p}_2)_c &= (p_1 - p_2)_{H_0} + z_c(0.95)\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \\ (\widehat{p}_1 - \widehat{p}_2)_c &= 0 + 1.645\sqrt{0.30(0.70)\left(\frac{1}{150} + \frac{1}{160}\right)} = (1.645)(0.053) = 0.087 \end{aligned} \tag{9.19}$$

El estadístico prueba en forma estandarizado:

$$z_m = \frac{(\widehat{p}_1 - \widehat{p}_2)_m - (p_1 - p_2)}{\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.11 - 0}{0.053} = 2.075 \tag{9.20}$$

$$valor - p = P((\widehat{p}_1 - \widehat{p}_2) > (\widehat{p}_1 - \widehat{p}_2)_m) = P((\widehat{p}_1 - \widehat{p}_2) > 0.11) = 0.019$$

$$valor - p = P(Z > z_m) = P(Z > 2.075) = 0.019$$

1. Las hipótesis:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 > p_2$$

2. El nivel de significancia $\alpha = 0.05$, $P(Z \geq 1.645) = 0.05$ (véase la gráfica superior en la figura 9.9).

3. Resumen de los cálculos de los puntos críticos y estadísticos.

   a) $(\widehat{p}_1 - \widehat{p}_2)_c = 0.087$ y $z_c = 1.645$

   b) $(\widehat{p}_1 - \widehat{p}_2)_m = 0.11$ y $z_m = 2.075$

   c) $valor - p = P((\widehat{p}_1 - \widehat{p}_2) > 0.11) = 0.019$ gráfica de en medio, y $valor - p = P(Z > 2.075) = 0.019$ gráfica inferior en la figura 9.9.

4. Comparando, véase la figura 9.9:

   a) $(\widehat{p}_1 - \widehat{p}_2)_m = 0.11 > (\widehat{p}_1 - \widehat{p}_2)_c = 0.087$, gráfica de en medio.

   b) $z_m = 2.075 > z_c = 1.645$, gráfica inferior.

   c) $valor - p = 0.019 < 0.05$

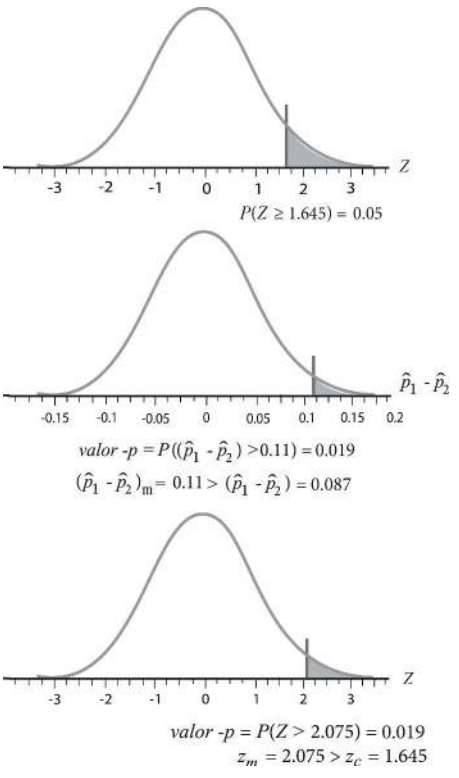


Figura 9.9 Describe el vínculo entre los puntos críticos y el $valor - p$.

Conclusión. De los incisos del punto 4, confirman que se rechaza la hipótesis nula. Por lo tanto la proporción de hombres que compran por internet es mayor que la de las mujeres.

En la figura 9.9 se muestran los resultados de este ejemplo presentando la relación entre los puntos críticos y el $valor - p$, indicando también la referencia del nivel de significancia.

Ejemplo 9.6

Por entrevistas realizadas a algunos de sus trabajadores, la administración de una organización supone que los del primer turno tienen menor preferencia que los del segundo turno en laborar jornadas de 10 horas diarias en lugar de 8, para completar su semana de 48 horas. Para verificar esa posibilidad selecciona una muestra de 210 trabadores de los cuales 120 son del primer turno y 54 están de acuerdo por trabajar 10 horas diarias, y de los 90 del segundo turno, 63 indicaron que están de acuerdo con 10. Plantee las hipótesis correspondientes. Para verificar la conjetura use un nivel significancia del 5 %. Los

datos generados en este estudio son:

Muestra	Proporciones	Tamaño	
Primer Turno	$\widehat{p}_1 = 0.45$	$n_1 = 120$	$x_1 = 54$
Segundo turno	$\widehat{p}_2 = 0.70$	$n_2 = 90$	$x_2 = 63$

Se cumplen las condiciones:

$$\begin{aligned} n_1 p_1 &= 120(0.45) \geq 5,\ n_1(1-p_1) = 120(0.55) \geq 5, \\ n_2 p_2 &= 90(0.70) \geq 5\ y\ n_2(1-p_2) = 90(0.30) \geq 5. \end{aligned}$$

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{54 + 63}{120 + 90} = \frac{117}{210} = 0.56$$

$$(\widehat{p}_1 - \widehat{p}_2)_m = (0.45 - 0.70) = -0.25$$

Solución

El argumento de la administración plantea que la proporción de trabajadores del primer turno que pretenden laborar 10 horas diarias en lugar de ir 6 días a la semana, es menor que la proporción de trabajadores del segundo turno. Así la hipótesis que se plantea es:

$$H_0 : p_1 = p_2$$
$$H_1 : p_1 < p_2$$

Con la finalidad de probar la hipótesis, en esta solución sólo se usará un criterio. Así se rechaza la hipótesis nula, si:

$$(\widehat{p}_1 - \widehat{p}_2)_m < (\widehat{p}_1 - \widehat{p}_2)_c \tag{9.21}$$

y en caso contrario, no se rechaza, esto es:

$$(\widehat{p}_1 - \widehat{p}_2)_m > (\widehat{p}_1 - \widehat{p}_2)_c \tag{9.22}$$

Ya que el nivel de significancia $\alpha = 0.05$, $z_c(\alpha) = z_c(0.05) = -1.645$

$$(\widehat{p}_1 - \widehat{p}_2)_c = (p_1 - p_2)_{H_0} + z_c(0.05)\sqrt{\overline{p}(1-\overline{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = (-1.645)(0.068) = -0.114$$

Se concluye rechazar H_0 ya que $(\widehat{p}_1 - \widehat{p}_2)_m = -0.25 < (\widehat{p}_1 - \widehat{p}_2)_c = -0.114$. Se resuelve de manera similar para H_0 usando el estadístico estandarizado, puesto que $z_m = -3.644 < z_c - 1.645$, se tiene el mismo

resultado para el nivel de significancia descriptivo, observe el $valor - p$. El $valor - p = P((\widehat{p}_1 - \widehat{p}_2)_m) \leq -0.25) = 0.00013$

Ejemplo 9.7

Muchos estudiantes tienen que trabajar para poder sostener a su familia. Se toma una muestra de los que asisten a la escuela en el turno matutino y se les pregunta si trabajan, y de igual forma a los del turno vespertino. La información recopilada es la siguiente: del turno matutino de 110 entrevistados 52 trabajan, y 57 de 100 del otro turno. La administración de la escuela indica que existe diferencia en la proporción entre los alumnos que trabajan en ambos turnos. Los datos generados en este estudio son:

Muestra	Proporciones	Tamaño	
Primer Turno	$\widehat{p}_1 = 0.473$	$n_1 = 110$	$x_1 = 52$
Segundo turno	$\widehat{p}_2 = 0.57$	$n_2 = 100$	$x_2 = 57$

Se cumplen las condiciones:

$$\begin{aligned} n_1 p_1 &= 110(0.473) \geq 5,\ n_1(1-p_1) = 110(0.527) \geq 5, \\ n_2 p_2 &= 100(0.57) \geq 5\ y\ n_2(1-p_2) = 100(0.43) \geq 5. \end{aligned}$$

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{52 + 57}{110 + 100} = \frac{109}{210} = 0.52$$

$$(\widehat{p}_1 - \widehat{p}_2)_m = (0.473 - 0.57) = -0.097$$

Solución mediante el uso de CalEst

La solución de este ejemplo se hará a partir del uso de **CalEst**; la opción es ir a inferencia estadística y tal como se describe en la figura 9.2 se accede a la prueba de hipótesis de proporciones y ahí se selecciona la de dos. En la figura 9.10 se presenta la hoja de captura, en la cual se escriben los datos del ejemplo o problema a estudiar, a continuación se escribe el valor de la hipótesis nula, el nivel de significancia y la hipótesis alternativa que corresponda.

Una vez que se ha completado el cuadro, se aplica la opción aceptar, en seguida aparecen los resultados que se indican a la derecha de la figura referida. En esta versión de **CalEst**, aparecen los criterios del $valor - p$ y del estadístico estandarizado z_m, con ese reporte se pueden obtener las conclusiones. Ya que se proporciona el valor crítico, no se rechaza H_0 si z_m está entre -2.576 y 2.576, es decir, $-2.576 \leq z_m \leq 2.576$, y se rechaza en caso contrario. En este ejemplo se satisface la primera situación y por lo tanto no se rechaza la hipótesis nula. El $valor - p > \alpha$ confirma la conclusión. Note que también se

comparte el intervalo de confianza, en este ejemplo del 99 %; la expresión de referencia para este cálculo se aporta en 9.15. Así:

$$-0.275 \leq p_1 - p_2 \leq 0.081$$

Observe que este intervalo contiene al cero, es una información complementaria que constata la conclusión ya anunciada.

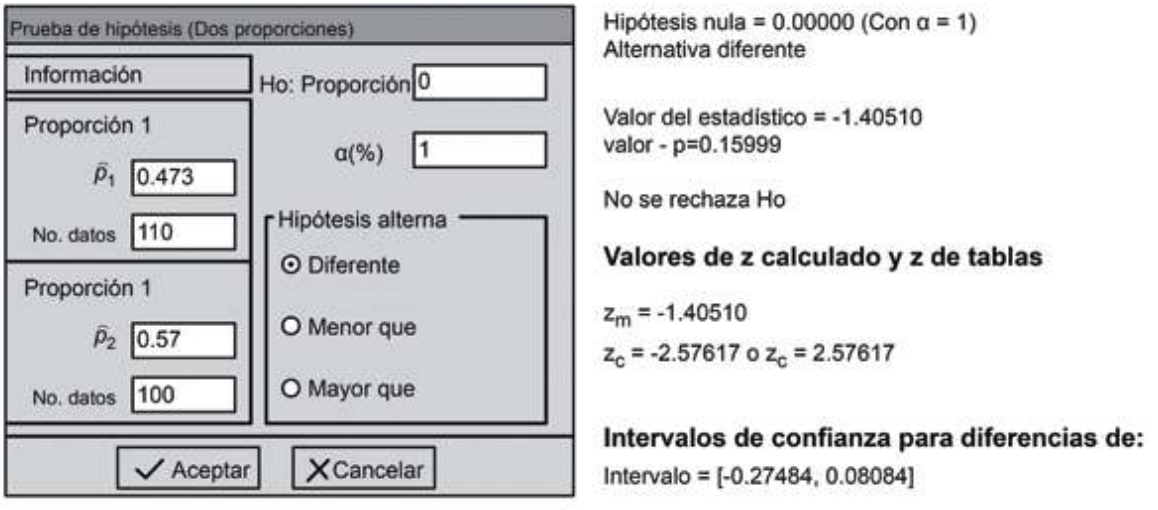


Figura 9.10 Indica cómo se llena la hoja para realizar la prueba de hipótesis de dos proporciones y los resultados que genera.

9.6 Prueba de hipótesis para la razón de varianzas

Prueba F, comparación de varianzas

Uno de los supuesto fundamentales para realizar la prueba de comparación de medias es la igualdad de varianzas; se interpreta diciendo "homogeneidad de poblaciones, o igualdad de varianzas". Así, la estimación por intervalos de confianza y la comparación de varianzas mediante una prueba de hipótesis son procedimientos importantes en estadística para verificar la homogeneidad entre dos poblaciones o tratamientos.

Considere la descripción expuesta en la figura 9.1, ahí se propone el estudio de dos poblaciones. Se propone la distribución de probabilidad de las variables X_1 y X_2 :

X_1: Tiene una distribución de probabilidad normal con media μ_1 y varianza σ_1^2, $X_1 \sim N(\mu_1, \sigma_1^2)$
X_2: Tiene una distribución de probabilidad normal con media μ_2 y varianza σ_2^2, $X_2 \sim N(\mu_2, \sigma_2^2)$

Las preguntas: ¿existe diferencia entre las varianzas de las poblaciones?

$$\frac{\sigma_1^2}{\sigma_2^2} = ?$$

¿Cómo se estiman y Cómo se comparan? Las varianzas de la muestra son:

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2}{n_1 - 1} \text{ y}$$

$$S_2^2 = \frac{\sum_{i=1}^{n_2}(x_{2i} - \overline{x}_2)^2}{n_2 - 1}$$

Donde n_1 es el tamaño de muestra de la población 1, y n_2 es el tamaño de muestra de la población 2.

Estimación

S_1^2 es un estimador insesgado de σ_1^2, $E(S_1^2) = \sigma_1^2$
S_2^2 es un estimador insesgado de σ_2^2, $E(S_2^2) = \sigma_2^2$

En el supuesto de que $\frac{\sigma_1^2}{\sigma_2^2} = 1$, se construye el siguiente estadístico:

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2}{S_2^2} \tag{9.23}$$

Entonces la expresión 9.23 sigue una distribución de probabilidad F con $gl_n = n_1 - 1$ grados de libertad en el numerador y $gl_d = n_2 - 1$ grados de libertad en el denominador, la expresión que caracteriza esta situación y permite obtener un intervalo del (1-α) % de confianza es:

$$P\left(F(gl_n, gl_d, \alpha/2) \leq \frac{S_1^2}{S_2^2} \leq F(gl_n, gl_d, 1-\alpha/2)\right) = 1 - \alpha$$

Por ejemplo, suponga que $gl_n = n_1 - 1 = 16$, $gl_d = n_2 - 1 = 20$ y $\alpha = 0.10$, la figura 9.11, usando el **CalEst**, muestra este caso.

$$F(16, 20, 0.05) = 0.439 \text{ y } F(16, 20, 0.95) = 2.184$$
$$P\left(0.439 \leq \frac{S_1^2}{S_2^2} \leq 2.184\right) = 0.90$$

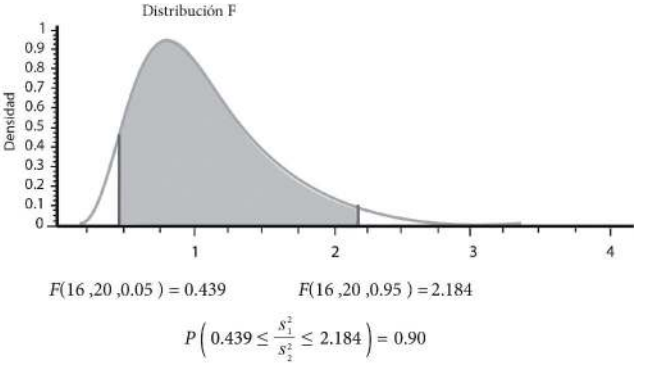


Figura 9.11 Probabilidad de la distribución F.

Propiedad de la distribución F

Propiedad de la distribución F

Si X tiene una distribución de probabilidad $F(gl_n, gl_d)$ entonces la variable $Y = \dfrac{1}{X}$ tiene una distribución de probabilidad $F(gl_d, gl_n)$.

Con el propósito de estimar un intervalo de confianza para la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$, es importante estudiar la distribución de probabilidad del recíproco de una variable X, donde X tiene una distribución de probabilidad F, con $gl_n = n_1 - 1$ grados de libertad en el numerador y $gl_d = n_2 - 1$ grados de libertad en el denominador. Observe que se intercambian los grados de libertad, del numerador $n_1 - 1$ y del denominador $n_2 - 1$. A continuación se presentará la relación entre estas distribuciones para el cálculo de probabilidades en procedimientos bilaterales.

Para fijar ideas, considere a F_I la cola izquierda de la distribución F y asociada con α o $\alpha/2$, y F_D la cola derecha y relacionada con $(1-\alpha)$ o $(1-\alpha/2)$. Entonces el recíproco de esta distribución se obtiene intercambiando los grados de libertad y usando la cola contraria de la distribución. Esto es:

$$F_I = F(16, 20, 0.05) = \frac{1}{F(20, 16, 0.95)} = \frac{1}{2.276} = \frac{1}{F_D^*} = 0.439,$$

note que se han permutando los grados de libertad, del numerador $n_1 - 1$ y del denominador $n_2 - 1$. En general la propiedad se puede expresar por:

Encontrar el valor del intervalo en la cola izquierda se obtiene mediante la expresión:

$$F_I = F(gl_n, gl_d, \alpha) = \frac{1}{F_D^*(gl_d, gl_n, 1-\alpha)}$$

donde F_D^* es la distribución F, donde los grados de libertad en el numerador son $gl_n = n_2 - 1$ y grados de libertad en el denominador son $gl_d = n_1 - 1$

La relación para la cola derecha de la distribución F

$$F_D = F(16, 20, 0.95) = \frac{1}{F(20, 16, 0.05)} = \frac{1}{0.459} = \frac{1}{F_I^*} = 2.184,$$

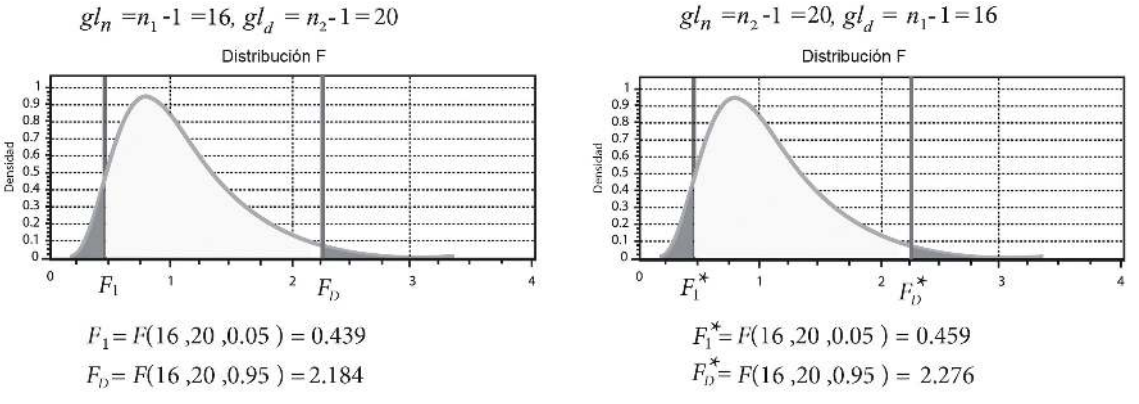


Figura 9.12 Distribuciones de la F, que muestra la propiedad de ésta.

En la figura 9.12 se muestran las gráficas de la distribución de probabilidad F, a la derecha con los grados de libertad $gl_n = n_1 - 1 = 16$, $gl_d = n_2 - 1 = 20$ y $\alpha/2 = 0.05$, y a la izquierda $gl_n = n_2 - 1 = 20$, $gl_d = n_1 - 1 = 16$ y $\alpha/ = 0.05$. Las relaciones:

$$F_I^* = F(20, 16, 0.05) = \frac{1}{F_D} = \frac{1}{F(16, 20, 0.95)} = \frac{1}{2.184} = 0.459,$$

$$F_D^* = F(20, 16, 0.95) = \frac{1}{F_I} = \frac{1}{F(16, 20, 0.05)} = \frac{1}{0.439} = 2.276,$$

Observación, los límites F_I^* y F_D^* son los que corresponderán a los valores del intervalo de confianza para la razón de varianzas. En resumen:

$$
\begin{aligned}
F_I^* &= F(gl_d, gl_n, 0.05) = \frac{1}{F_D} = \frac{1}{F(gl_n, gl_d, 0.95)} \\
F_D^* &= F(gl_d, gl_n, 0.95) = \frac{1}{F_I} = \frac{1}{F(gl_n, gl_d, 0.05)}
\end{aligned}
\tag{9.24}
$$

Nota: Utilice la distribución de probabilidad F en **CalEst** para obtener los valores correspondientes a las relaciones indicadas; en el capítulo 6 se ha indicado que para acceder a ella tiene que ir al módulo de distribuciones y de ahí a la calculadora de la distribución F. Para estos datos ver la figura 9.13. Le recordamos que los valores que genera la gráfica son aproximados a milésimas y la calculadora hasta diezmilésimos.

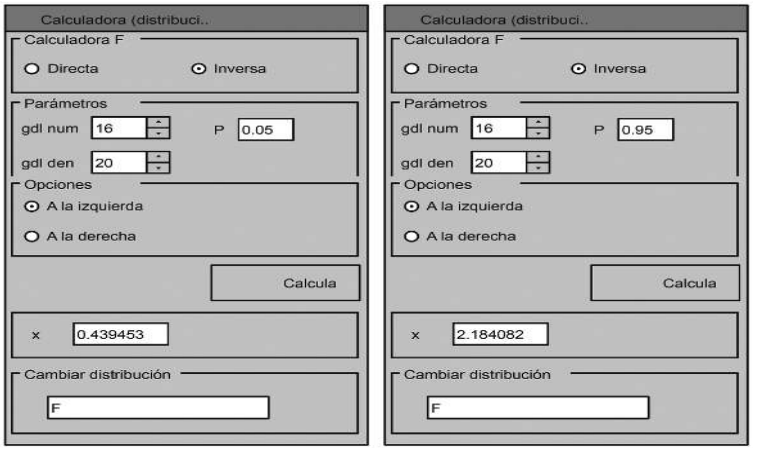


Figura 9.13 Cálculo de la distribución de probabilidad F, $gl_n = n_1 - 1 = 16$, $gl_d = n_2 - 1 = 20$ y $\alpha/2 = 0.05$.

Metodología para la inferencia en la igualdad de varianzas

Siguiendo un mecanismo similar al de las pruebas discutidas en este capítulo, se sigue que para la igualdad de varianza $\sigma_1^2 = \sigma_2^2$, también referido como razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2} = 1$

1. Planteamiento de las hipótesis:

$$
\begin{aligned}
H_0 &: \sigma_1^2 = \sigma_2^2 \qquad H_1 : \sigma_1^2 \neq \sigma_2^2 \text{ o,} \\
H_0 &: \frac{\sigma_1^2}{\sigma_2^2} = 1 \qquad H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1
\end{aligned}
$$

2. Dado el nivel de significancia en esta prueba la distribución de probabilidad que se emplea es la F, estudiada en el capítulo 6. Entonces los valores críticos son $F_{ci} = F(gl_n, gl_d, \alpha/2)$ y $F_{cd} = F(gl_n, gl_d, (1 - \alpha/2))$, donde $gl_n = n_1 - 1$, $gl_d = n_2 - 1$ y α.

3. Se calcula el valor del estadístico para los datos generados en el estudio, es decir:

$$F_m = \frac{S_1^2}{S_2^2}$$

No se rechaza la hipótesis nula si se cumple que:

$$F_{ci} \leq F_m \leq F_{cd}$$

En caso contrario, se rechaza la hipótesis.

4. Se concluye en el contexto del problema que se está estudiando.

Ejemplo 9.8

En el marco de la economía familiar, en dos tiendas de autoservicio el precio de la canasta básica varía día con día y una persona tiene que decidir en cuál de las dos tiendas comprar para no desequilibrar su presupuesto. El precio en el riesgo de compra está asociado con la desviación estándar del precio diario de la canasta. Se tienen seleccionadas muestras aleatorias de ambas tiendas; la tienda A: 31 días con una desviación estándar de 5.7; la tienda B: 30 días con una desviación estándar de 3.5. Con un nivel de significancia $\alpha = 0.05$, ¿se puede concluir que en alguna tienda hay mayor riesgo de comprar y desequilibrar el presupuesto? Resumen de la información que proporciona la muestra:

$$\begin{array}{llll} \text{Muestra 1} & n_1 = 31 & S_1^2 = (5.7)^2 = 32.49 \\ \text{Muestra 2} & n_2 = 30 & S_2^2 = (3.5)^2 = 12.25 \end{array}$$

Operaciones

$$F_m = \frac{32.49}{12.25} = 2.652$$

Cálculo del valor crítico en función del nivel de significancia $F_{ci} = F(30, 29, 0.025) = 0.48$, y $F_{cd} = F(30, 29, 0.975) = 2.091$ puesto que 2.652 no está entre los valores 0.48 y 2.1 se rechaza H_0.

Alternativamente se prueba la hipótesis utilizando el nivel de significancia descriptivo p, $2p = P(F_m > 2.652) = 2(0.00514) = 0.01028$, puede observar que $p < \alpha = 0.025$, por lo tanto se concluye rechazar H_0. La tienda A tiene mayor varianza en sus precios.

Intervalo del (1-α) % confianza para la razón de varianzas : $\frac{\sigma_1^2}{\sigma_2^2}$

Observe la expresión 9.23, para poder determinar el *intervalo de confianza para la razón de varianzas* , se requiere de algunas operaciones que nos conducen a la aplicación de la propiedad del recíproco de la distribución F; así, el intervalo es:

$$\left(\frac{S_1^2}{S_2}F_I^*, \frac{S_1^2}{S_2^2}F_D^*\right)$$

Aplicando las fórmulas 9.24 se expresa el intervalo como en la literatura estadística:

$$\left(\frac{S_1^2}{S_2}\frac{1}{F_D}, \frac{S_1^2}{S_2^2}F(gl_d, gl_n, 1-\alpha/2)\right)$$

Determinación de los valores para la distribución F, para los límites a la izquierda $F_I^* = \frac{1}{F(29,30,0.975)} = 0.478$, y derecha $F_D^* = F(30-1, 31-1, 0.975) = 2.083$, donde gl_n, gl_d son los grados de libertad para el numerador y denominador, respectivamente. Así, el intervalo de 95 % de confianza para la razón de varianza es:

$$((2.652)(0.478), (2.652)(2.083)) = (1.268, 5.523)$$

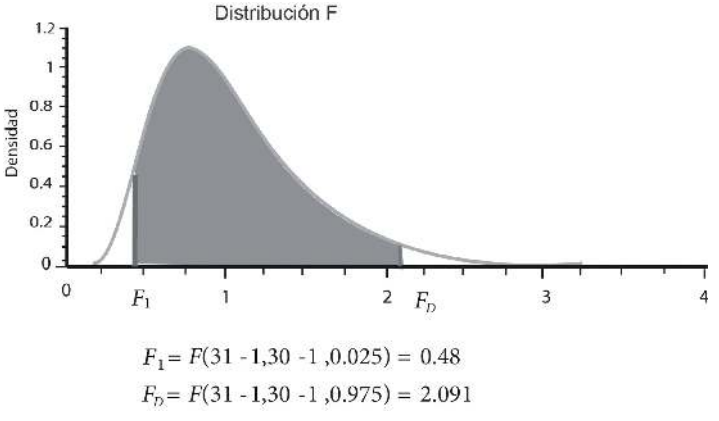


Figura 9.14 Valores de la distribución F para construir el intervalo de 95 % de confianza.

Una interpretación de interés es la relación entre este intervalo y la prueba de hipótesis, observe que el intervalo no contiene al 1. Así, para el caso de la hipótesis alternativa $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$, el intervalo del $(1-\alpha)\,\%$ confianza es una opción para probar esta hipótesis. En consecuencia, si el 1 no está contenido en el intervalo, se rechaza la hipótesis nula, tal y como ocurre en este ejemplo.

La descripción gráfica de los resultados de la prueba se muestra en la figura 9.14; la facilidad de la gráfica de la distribución de probabilidad F es relevante para comprender con mayor claridad un intervalo de confianza. Análogamente puede usar la descripción de esta distribución para explicar la prueba de hipótesis, para ello conviene que haga dos gráficas, una de ellas que señale el nivel de significancia y así ver el valor del punto crítico, la otra que indique el valor del estadístico calculado.

Ejemplo 9.9

Un profesor ha diseñado una estrategia para mejorar la velocidad de lectura; en este sentido intenta reducir la varianza en el tiempo de lectura. Primero hace una prueba para conocer la varianza del tiempo de lectura que actualmente tienen sus alumnos y selecciona una muestra aleatoria de 10 alumnos. La prueba de velocidad de lectura arrojó una varianza de 144 minutos^2. En su planteamiento consideró una muestra de 21 alumnos; la varianza del tiempo de lectura fue de 100 minutos^2. Proponiendo un nivel de significancia $\alpha = 0.10$, ¿existe suficiente evidencia para respaldar la estrategia del profesor?

Resumen de la información que proporciona la muestra:

$$\begin{array}{lll} \text{Muestra 1} & n_1 = 10 & S_1^2 = 144 \\ \text{Muestra 2} & n_2 = 21 & S_2^2 = 100 \end{array}$$

Solución mediante el uso de CalEst

Este ejemplo será resuelto con apoyo de la tecnología, en este caso se trata de hacer inferencia sobre la varianza. Después del planteamieto estadístico del problema se comentan los resultados apoyados en los resultados que proporciona el paquete estadístico. La descripción operativa que genera el CalEst sobre la prueba de hipótesis sobre la razón de varianzas se describe en la figura 9.16.

1. La hipótesis de trabajo se expresa por:

   H_t : La varianza del tiempo de lectura actual es mayor a la varianza del tiempo bajo su esquema.

   Las hipótesis estadísticas son:

$$\begin{aligned} H_0 &: \sigma_1^2 = \sigma_2^2 \\ H_1 &: \sigma_1^2 > \sigma_2^2 \end{aligned}$$

2. El nivel de significancia $\alpha = 0.10$, proporciona el valor de referencia, $F_c = F(10-1, 21-1, 0.90) = 1.96$, verifique este valor usando la distribución F (use la tabla de esta distribución, o el calculador en **CalEst**, como se muestra en la figura 9.15).

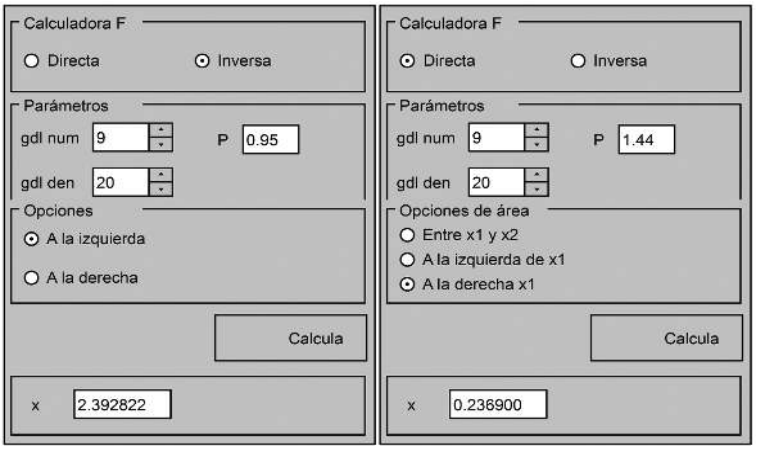


Figura 9.15 Calculador de la distribución F, activos en el **CalEst**.

3. Resumen de los cálculos de los puntos crítcos y estadísticos de prueba

   a) Puntos críticos $F_c = 1.96$, y estadístico de prueba $F_m = \frac{144}{100} = 1.44$

   b) $valor - p = P(F \geq 1.44) = 0.237$, observe que $0.237 = p > \alpha = 0.1$

4. Puesto que $F_m < F_c$ concluye que no se rechaza H_0. Similarmente si usa la relación $p > \alpha$.

Los datos no dan evidencia para concluir que la varianza en el tiempo de la lectura es mayor en el grupo 1, el método del profesor no fue eficiente.

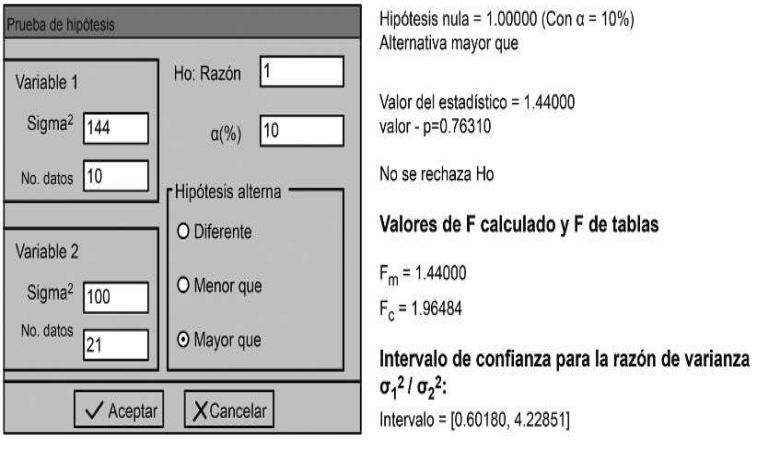


Figura 9.16 Hoja de captura en el **CalEst** y reporte de la prueba F.

9.7 Prueba de hipótesis para más de dos poblaciones

Análisis de varianza: ideas generales. En el capítulo 6 y en los apartados previos se plantearon pruebas de hipótesis para una y dos poblaciones, respectivamente. En esta parte se desarrolla un procedimiento para comparar más de dos poblaciones a la vez; este procedimiento se conoce como análisis de varianza (abreviada ANDEVA). Los datos se pueden obtener ya sea mediante observación de algún hecho, a través de una estrategia experimental. Esta última desempeña un papel relevante para hacer la comparación entre más de dos poblaciones, y se le llama *diseño completamente al azar*

El mundo de la información 3: preferencia por un producto

En el modelo para mejorar la calidad en el servicio, está el apartado del consumo. Para mantenerse a un nivel competitivo, muchas empresas procuran no tener mermas económicas y buscan aumentar su ganancia. Por ello las administraciones planean estrategias para mantenerse, además de ganar nuevos mercados. Mediante cuestionarios tienen información de la preferencia que los clientes tienen sobre sus productos, así como exploración del mercado. ¿Qué tipo de sondeo utilizan para adquirir conocimiento sobre la venta de diferente tipo de producto? ¿Qué observaciones necesitan realizar para compararse con otros competidores? Con ese conocimiento tienen que ajustar sus procesos, lo cual los lleva a aplicar técnicas de experimentación. En esa dirección, aplican métodos de administración estratégica para ser expertos en el mercado y recurren al desarrollo tecnológico generado por profesionistas altamente capacitados. Considere un ejemplo sencillo y simplificado en el desarrollo de nuevos productos: un ingeniero que trabaja en biotecnología de alimentos quiere conocer la preferencia que tienen las personas por cuatro tipos de helados elaborados en condiciones diferentes. En este caso las hipótesis son:

- Hipótesis nula: H_0: las personas prefieren por igual los cuatro helados.
- Hipótesis alternativa: H_1: existe preferencia por alguno de los helados.

¿Qué plan debe seguir para obtener la información? ¿Cómo debe medir?

En este caso, como se trata del desarrollo de un nuevo producto que busca desde un punto económico ganar mercados, se aplica la estrategia experimental. Esa técnica le ayudará a tener una idea sobre la naturaleza del diseño de experimentos. A continuación se expondrán algunas nociones elementales sobre este tema.

Preliminares: La estrategia experimental para este estudio es como sigue: suponga que hay 12 personas y que cada una de tres personas prueban cada helado (h_1, h_2, h_3 y h_4 representan los cuatro helados). Las doce personas se seleccionan aleatoriamente y se les da a probar el helado correspondiente, es decir:

h_1	h_1	h_1	h_2	h_2	h_2	h_3	h_3	h_3	h_4	h_4	h_4
8	3	10	12	5	6	11	9	4	7	2	1

La medición se realiza en una escala hedónica con valores que van de 1 a 10, donde 1 es el disgusto de

las personas por el helado y 10 es la aceptación total. Se realizaron 10 preguntas.La estructura de esta estrategia experimental se presenta en la tabla 10.1, con los resultados reportados por las personas.

Tabla 10.1 Resultados experimentales

	Helados			
	1	2	3	4
	74	46	80	85
	78	56	70	82
	73	49	77	89
n_j	3	3	3	3
$\overline{x}_j$	75	50.33	75.67	85.33
S_j^2	7	26.33	26.33	12.33

La idea principal en este caso es probar si la variación entre grupos (entre helados) es similar a la variación dentro de grupos (cada helado).

Hipótesis. En resumen, el planteamiento general de este esquema es: para el ejemplo se tienen 4 poblaciones de interés; el procedimiento del análisis de varianza se plantea probar la hipótesis:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ &H_1 : \text{algún par es diferente } \mu_1 \neq \mu_2, \text{ ó } \mu_1 \neq \mu_3 \text{ ó } \mu_1 \neq \mu_4 \\ &\qquad\qquad\qquad\qquad\quad \mu_2 \neq \mu_3, \text{ ó } \mu_2 \neq \mu_4 \text{ ó } \mu_3 \neq \mu_4 \\ &\qquad\qquad\qquad\qquad\quad \mu_3 \neq \mu_4 \end{aligned} \tag{9.25}$$

Este esquema se conoce como *diseño completamente al azar* La meta de este procedimiento es construir el estadístico de prueba para verificar esta hipótesis.

Términos básicos en el diseño. Primero se describe lo que se entiende por factor.

> Un **factor** es una variable que se usa para diferenciar un grupo de una población a otra. Esta es una variable que puede estar relacionada con la variable de interés. Un **nivel** es uno de los posibles valores que el factor puede tomar.

En el ejemplo el factor es el tipo de helado, el nivel es cada uno de los helados y caso se tienen cuatro niveles. A continuación se formaliza lo que se llama *variable de respuesta*, en este caso la evaluación de la preferencia del helado.

> La **variable de respuesta** es una variable cuantitativa, que son las mediciones sobre las unidades que estarán sujetas a observación o experimentación.

La descripción de la observación o experimento:

Un **diseño completamente al azar** es una estructura que permite observar o experimentar en los diferentes niveles de un factor, y los objetos que se observan o miden son asignados aleatoriamente a uno de los niveles del factor.

En la terminología de esta estructura cada nivel del factor, se define como *tratamiento*. En el ejemplo, el tratamiento es el tipo de helado que probarán las personas. **Un tratamiento es una característica o nivel particular del factor.** También en cada uno de los tratamientos hay tres observaciones, es decir $n_1 = n_2 = n_3 = n_4 = 3$ en esta situación el diseño es balanceado porque el tamaño de muestra es igual en cada caso.

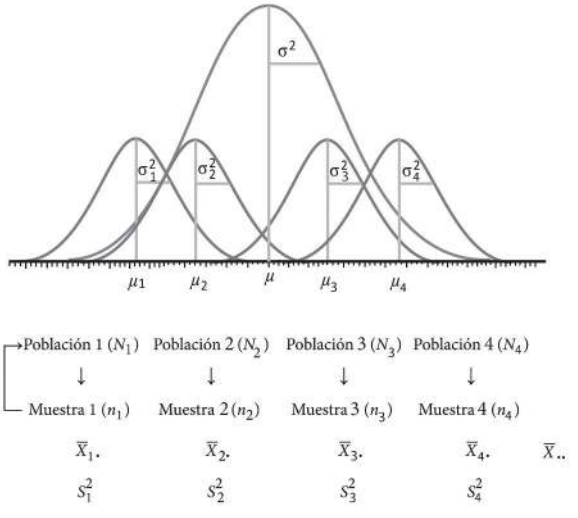


Figura 9.17 Caracterización de cuatro poblaciones con referencia a una población general.

Análisis de los datos

El procedimiento estadístico para analizar los datos del experimento se llama análisis de la varianza (ANDEVA).

Análisis de la varianza es una técnica que se utiliza para analizar la varianza en los datos para determinar cuándo más de dos poblaciones tienen medias iguales.

Para la aplicación del análisis de la varianza se parte de tres supuestos (una idea aproximada de esto se exhiben en la figura 9.17.

1. Las poblaciones objeto de estudio tienen una distribución de probabilidad normal.

2. Las varianzas de estas poblaciones son iguales.

3. Las muestras se seleccionan de manera independiente.

Planteamiento estadístico

Considerando la situación de 4 poblaciones tal y como se muestra en la figura 9.17, la finalidad es verificar si efectivamente los datos vienen de cuatro poblaciones diferentes o corresponden a una. Si se presenta esta última situación se dice que no hay efecto de tratamiento. ¿Cómo se verifica? Esto se realiza a través de una prueba de hipótesis, similar a la propuesta por la expresión. Para realizar la prueba, recuerde que se necesita construir un estadístico y éste surge de la información que proporciona la muestra tomada de una población. A nivel general, para cuatro poblaciones la estructura de las muestras aleatorias se recogen en la tabla 9.2. Observe la siguiente composición de las muestras:

1. Cada muestra tiene sus estadísticas que son la media y la varianza. ¿Qué tanta varianza existe entre las muestras? ¿Cómo se puede estimar?

2. Dentro de cada muestra existe una varianza y en función del tamaño de la muestra la idea es ponderar éstas. ¿Qué efecto tiene esta varianza ponderada en el análisis de los datos?

3. ¿Cómo se pueden usar las varianza del punto 1 y del punto 2 en el análisis estadístico?

Tabla 10.2 Estructura de las muestras aleatorias seleccionadas de las poblaciones.

Muestra 1	Muestra 2	Muestra 3	Muestra 4	Total
x_{11}	x_{21}	x_{31}	x_{41}	
x_{12}	x_{22}	x_{32}	x_{42}	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_{1n_1}	x_{2n_2}	x_{3n_1}	x_{4n_1}	
$\overline{x}_{1\cdot}$	$\overline{x}_{2\cdot}$	$\overline{x}_{3\cdot}$	$\overline{x}_{4\cdot}$	$\overline{x}_{\cdot\cdot}$
S_1^2	S_2^2	S_3^2	S_4^2	

Para aprender cómo se efectúa el análisis estadístico siguiendo los tres puntos citados, se propone el siguiente ejemplo.

Ejemplo 9.10

Por la actividad que realiza en las ventas de un servicio, un agente de seguros tiene interés en evaluar cuatro instituciones bancarias sobre el tiempo que les toma a las personas en efectuar un trámite. Para ello toma como referencia a los que no son clientes de la institución. Selecciona las cuatro instituciones de un centro comercial, luego de manera aleatoria selecciona a cinco clientes de cada institución. Toma el tiempo en que son atendidos. Los resultados se muestran en la tabla 9.3. ¿Existe diferencia entre las

instituciones en el tiempo de atención? Para contestar la pregunta utiliza un nivel de significancia de 10 %

	Banco 1	Banco 2	Banco 3	Banco 4	Total
	12	13	24	29	
	16	10	17	31	
	17	14	23	26	
	14	18	20	28	
	15	12	18	22	
$\overline{x}_{i\cdot}$	$\overline{x}_{1\cdot} = 14.8$	$\overline{x}_{2\cdot} = 13.4$	$\overline{x}_{3\cdot} = 20.4$	$\overline{x}_{4\cdot} = 27.2$	$\overline{x}_{\cdot\cdot} = 18.95$
S_i^2	$S_1^2 = 3.7$	$S_2^2 = 8.8$	$S_3^2 = 9.3$	$S_4^2 = 11.7$	
$(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})$	-4.15	-5.55	1.45	8.25	

donde $i = 1, 2, 3, 4$ bancos. La figura 9.18 describe la relación gráfica de la información en cada muestra.

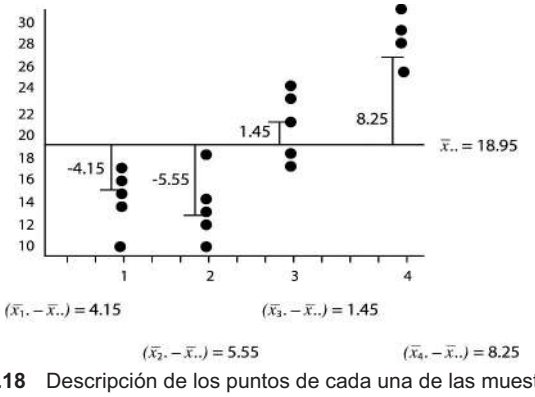


Figura 9.18 Descripción de los puntos de cada una de las muestras.

Solución

El punto de partida es plantear las hipótesis:

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \\ &H_1 : \text{Algún par de medias es diferente } \mu_1 \neq \mu_2, \text{ ó } \mu_1 \neq \mu_3 \text{ ó } \mu_1 \neq \mu_4 \\ &\qquad\qquad \mu_2 \neq \mu_3, \text{ ó } \mu_2 \neq \mu_4 \text{ ó } \mu_3 \neq \mu_4 \end{aligned} \tag{9.26}$$

Metodología para construir el estadístico de prueba

1. Cálculos para obtener la varianza entre bancos (grupos: muestras)

						Total
Medias		$\overline{x}_{1\cdot}= 14.8$	$\overline{x}_{2\cdot}= 13.4$	$\overline{x}_{3\cdot}= 20.4$	$\overline{x}_{4\cdot}= 27.2$	$\overline{x}_{\cdot\cdot}= 18.95$
Diferencias	$(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})$	-4.15	-5.55	1.45	8.25	
Diferencias por el tamaño de muestra	$n_i\,(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$	$5(17.22)$	$5\,(30.80)$	$5\,(2.10)$	$5\,(68.06)$	590.95

donde n_i describe la suma de cada una de las unidades en cada muestra, 590.95, es la suma de los cuatro términos de $n_i\,(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2$, en símbolos:

$$\sum_{i=1}^{4} n_i\,(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2 = 590.95$$

La varianza, que en el contexto del análisis de esta estructura, se denomina cuadrado medio entre grupos, y se escribe:

$$CM_{entre} = \frac{\sum_{i=1}^{4} n_i\,(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2}{4-1} = \frac{590.95}{3} = 196.98 \tag{9.27}$$

2. Estimación de la varianza (ponderada) dentro de cada muestra. Cada columna de la tabla siguiente muestra la discrepancia al cuadrado que hay en cada muestra. Se suman estos cuadrados y se divide entre el tamaño de muestra en cada tratamiento n_i, así se obtiene la varianza, en este caso de cada banco.

   La varianza ponderada toma en cuenta el tamaño de la muestra, esta se escribe como:

$$S^2_{ponderada} = \frac{(5-1)S_1^2 + (5-1)S_2^2 + (5-1)S_3^2 + (5-1)S_4^2}{(5-1) + (5-1) + (5-1) + (5-1)}$$

$$S^2_{ponderada} = \frac{4(3.7 + 8.8 + 9.3 + 11.7)}{16} = \frac{134}{16} = 8.375 \tag{9.28}$$

   En el ambiente de la estructura que se está considerando, esta varianza se le denomina cuadrado medio dentro de grupos y la expresión es:

$$CM_{dentro} = S^2_{ponderada}$$

Banco 1	Banco 2	Banco 3	Banco 4
$(x_{1j}-\overline{x}_{1\cdot})^2$	$(x_{2j}-\overline{x}_{2\cdot})^2$	$(x_{3j}-\overline{x}_{3\cdot})^2$	$(x_{4j}-\overline{x}_{4\cdot})^2$
$j=1,...,5$	$j=1,...,5$	$j=1,...,5$	$j=1,...,5$
$(12-14.8)^2$	$(13-13.4)^2$	$(24-20.4)^2$	$(29-27.2)^2$
$(16-14.8)^2$	$(10-13.4)^2$	$(17-20.4)^2$	$(31-27.2)^2$
$(17-14.8)^2$	$(14-13.4)^2$	$(23-20.4)^2$	$(26-27.2)^2$
$(14-14.8)^2$	$(18-13.4)^2$	$(20-20.4)^2$	$(28-27.2)^2$
$(15-14.8)^2$	$(12-13.4)^2$	$(18-20.4)^2$	$(22-27.2)^2$
$Sumar\,(x_{1j}-\overline{x}_{1\cdot})^2$ $\sum_{j=1}^{5}(x_{1j}-\overline{x}_{1\cdot})^2=14.8$	$Sumar\,(x_{2j}-\overline{x}_{2\cdot})^2$ $\sum_{j=1}^{5}(x_{2j}-\overline{x}_{2\cdot})^2=35.2$	$Sumar\,(x_{3j}-\overline{x}_{3\cdot})^2$ $\sum_{j=1}^{5}(x_{3j}-\overline{x}_{3\cdot})^2=37.2$	$Sumar\,(x_{4j}-\overline{x}_{4\cdot})^2$ $\sum_{j=1}^{5}(x_{4j}-\overline{x}_{4\cdot})^2=46.8$
$S_1^2=\frac{14.8}{5-1}=3.7$	$S_2^2=\frac{35.2}{5-1}=8.8$	$S_3^2=\frac{37.2}{5-1}=9.3$	$S_4^2=\frac{46.8}{5-1}=11.7$

3. En el punto 1 y 2 se han construido dos varianzas, una entre grupos y la otra dentro de grupos; la razón de estas varianzas proporcionan el estadístico que se requiere para la prueba de hipótesis. Esto se expresa por

$$\text{razón de varianzas} = RV = \frac{CM_{entre}}{CM_{dentro}}$$

De acuerdo con el resultado indicado por la expresión 9.23, esta RV tiene una distribución de probabilidad F, observe con cuidado, donde los grados de libertad en el numerador es el número de grupos menos uno, es decir: $gl_n = 4-1 = 3$, los grados de libertad para el denominador son: $(5-1)+(5-1)+(5-1)+(5-1) = 20-4$ que corresponde al total de observaciones menos el número de grupos, así $gl_d = 20-4 = 16$. En referencia a lo presentado en el apartado anterior, el estadístico F_m es RV, por lo tanto

$$F_m = \frac{CM_{entre}}{CM_{dentro}} \tag{9.29}$$

El valor calculado es:

$$F_m = \frac{CM_{entre}}{CM_{dentro}} = \frac{196.98}{8.375} = 23.52$$

Metodología para la prueba de hipótesis

Ahora, con los elementos que permitieron construir el valor del estadístico de prueba F_m, expresión 9.29, se está en la posibilidad de realizar la prueba sobre las hipótesis planteadas en 9.26. Reproduciendo el procedimiento de prueba de hipótesis discutidos en las unidades anteriores, se sigue que:

1. Planteamiento de hipótesis, escritas en 9.26.

2. Dado el nivel de significancia $\alpha = 0.10$, el valor crítico F_c, está dado por la distribución de probabilidad F. A partir de la tabla de la distribución o del calculador, sigue que: $F_c = F(gl_n, gl_d, 1-\alpha) = F(3, 16, 0.90) = 2.462$, figura 9.19.

3. Comparamos el estadístico con el valor crítico y se observa que: $F_m > F_c = 2.462$. Además, el $valor - p = P(F \geq F_m) = P(F \geq 23.52) = 0.000$, bastante más pequeño que el valor de α.

4. Por lo tanto se rechaza la hipótesis nula. Eso quiere decir que al menos un par de los grupos tiene medias diferentes. Hay bancos en que en promedio los clientes tardan más tiempo en realizar los trámites.

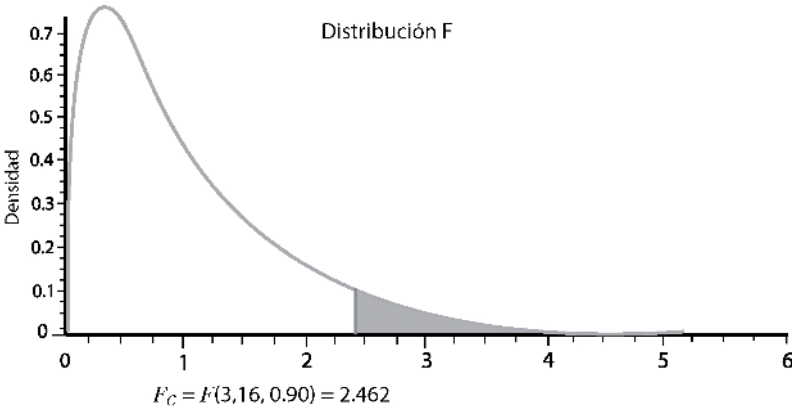


Figura 9.19 Valor del punto crítico en la distribución F.

Tabla del análisis de la varianza (andeva). Es frecuente representar en una tabla los cálculos descritos para la varianza tanto entre como dentro. En ésta se describen por columna: las fuentes de variación, grados de libertad de la varianza, la suma de cuadrados, el cuadrado medio y el estadístico $F - m$, o razón de varianzas.

Tabla 9.2 Análisis de la varianza

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	Fm
Entre tratamientos	$4 - 1 = 3$	$SC_{trat} = 590.95$	$CM_{trat} = \frac{590.95}{3}$	$\frac{CM_{trat}}{CM_{dentro}} = 23.52$
Dentro tratamientos	$20-4 = 16$	$SC_{dentro} = 134$	$CM_{dentro} = \frac{134}{16}$	
Total	$20 - 1 = 19$	$SC_{total} = 724.95$		

Solución mediante el uso de CalEst

En el material del cálculo estadístico CalEst, en la opción inferencia está la alternativa para realizar el andeva; una vez capturados los datos en la hoja en cada columna aparece el tratamiento, se aplica la alternativa y se genera la información requerida para completar la prueba según se describió en la metodología.

En la figura 9.20 se reproduce la salida; en este reporte se muestra el resumen estadístico, así como la tabla del andeva indicando el $valor - p$, que le permitirá concluir. Nota: el nivel de significancia que se empleó en el CalEst es $\alpha = 0.05$. También aparecen los intervalos de confianza para cada grupo.

Análisis de varianza
Media total
18.95000
Media tratamientos
14.80000
13.40000
20.40000
27.20000

Análisis de varianza

gl	suma de c	cuadrado medio	razón de varianza	valor-p
3	590.95000	196.98333	23.52040	0.00000
16	134.00000	8.37500		
19	724.95000			

Desviación estándar ponderada
2.89396

Límites de intervalos de confianza de (1–α)

12.05595,	17.54405
10.65595,	16.14405
17.65595,	23.14405
24.45595,	29.94405

Figura 9.20 Reporte del análisis de la varianza para los datos del ejemplo 10.

Comparaciones múltiples. Una vez que se rechazó la hipótesis nula, se sabe que al menos un par de medias es diferente. ¿Cuál de ese par es? En esta situación se realizan las pruebas de hipótesis para cada par de medias que se describen a continuación.

$$
\begin{aligned}
&H_0\text{: } \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2,\\
&H_0\text{: } \mu_1 = \mu_3 \text{ vs } H_1 : \mu_1 \neq \mu_3,\\
&H_0\text{: } \mu_1 = \mu_4 \text{ vs } H_1 : \mu_1 \neq \mu_4,\\
&H_0\text{: } \mu_2 = \mu_3 \text{ vs } H_1 : \mu_2 \neq \mu_3,\\
&H_0\text{: } \mu_2 = \mu_4 \text{ vs } H_1 : \mu_2 \neq \mu_4\\
&H_0\text{: } \mu_3 = \mu_4 \text{ vs } H_1 : \mu_3 \neq \mu_4
\end{aligned}
$$

Para probar estas hipótesis, aquí se recurre a utilizar la fórmula expresada en la prueba de muestras independientes descrita por la ecuación 9.10 y ésta es:

$$(\overline{x}_1 - \overline{x}_2)_m + t_{ci}ES < (\mu_1 - \mu_2) < (\overline{x}_1 - \overline{x}_2)_m + t_{cd}ES$$

De tal forma, si el intervalo contiene al cero no se rechaza la hipótesis de nula y se dice que el par de

medias es igual. Considere varios puntos antes de hacer la verificación en el ejemplo. El primero de ellos es el error estándar, éste se escribe por:

$$ES = \sqrt{CM_{dentro}}\left(\sqrt{\frac{1}{n_i}+\frac{1}{n_j}}\right)$$

Para el ejemplo $ES = \sqrt{CM_{dentro}}\left(\sqrt{(1/5)+(1/5)}\right) = \sqrt{8.375}(0.632) = 2.894(0.632) = 1.830$. El segundo corresponde al cálculo de los grados de libertad para la distribución $t - Student$ son gl_d (gl dentro de grupos). En este caso $t_{ci} = t(16, 0.025) = -2.12$ y $t_{cd} = t(16, 0.975) = 2.12$. Nota: en esta exposición se ha decidido hacer estas pruebas de comparación múltiple usando el estadístico de prueba t, porque cumple con las metas planteadas, sin embargo, en la literatura, principalmente de diseño de experimentos, existen varias pruebas para este objetivo.

Comentario: por último, se puede preguntar ¿por qué no hacer estas pruebas desde el inicio en lugar del análisis de varianza? La cuestión es que si se hacen las comparaciones de manera independiente al inicio, la probabilidad del error tipo I crece. De esta forma, para cualquier comparación entre las medias, se podría decir que son diferentes cuando no lo son. Similarmente ocurriría con el error tipo II, si se decidiera que las medias no son diferentes cuando realmente lo son.

Solución de comparaciones múltiple para el ejemplo 10

Se ejemplifica el primer caso: la diferencia de medias para los dos primeros bancos son: $\overline{x}_{1\cdot} = 14.8 - \overline{x}_{2\cdot} = 13.4 = 1.4$, se sustituyen los valores en el intervalo y se obtiene que:

$$1.4 - 2.12(1.83) < (\mu_1 - \mu_2) < 1.4 + 2.12(1.83)$$

$$-2.480 < (\mu_1 - \mu_2) < 5.280$$

El intervalo contiene al cero, por lo tanto no se rechaza la hipótesis nula en ésta comparación, por lo tanto las medias de los bancos 1 y 2 son iguales. Entonces el tiempo que tardan en promedio los clientes en estas dos instituciones es el mismo. Extendiendo el reporte generado por el CalEst en el andeva, se observa en la figura 9.21, los intervalos de confianza para las diferencias de medias. Se recomienda hacer el resto de las comparaciones y verificarlas con lo observado en la figura 9.21.

Observe que los intervalos de las pruebas 2 a la 6 no contienen al cero, por lo tanto se rechaza la hipótesis nula, y todas esas diferencias de medias son distintas. Salvo en los dos primeros bancos donde el tiempo de atención es el mismo en los demás es mucho más tardado en promedio.

H_0: $\mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$,	1.40000	–2.48067	5.28067
H_0: $\mu_1 = \mu_3$ vs $H_1 : \mu_1 \neq \mu_3$,	–5.60000	–9.48067	–1.71933
H_0: $\mu_1 = \mu_4$ vs $H_1 : \mu_1 \neq \mu_4$,	–12.40000	–16.28067	–8.51933
H_0: $\mu_2 = \mu_3$ vs $H_1 : \mu_2 \neq \mu_3$,	–7.00000	–10.88067	–3.11933
H_0: $\mu_2 = \mu_4$ vs $H_1 : \mu_2 \neq \mu_4$,	–18.80000	–17.68067	–9.91933
H_0: $\mu_3 = \mu_4$ vs $H_1 : \mu_3 \neq \mu_4$,	–6.80000	–10.68067	–2.91933

Figura 9.21 Diferencia de medias con sus respectivos intervalos.

Componentes de la variación total

En resumen, el planteamiento general del diseño completamente al azar es: supóngase que hay k poblaciones de interés; el procedimiento del análisis de varianza se plantea probar la hipótesis:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$
$$H_1 : \text{algún par de medias es diferente } \mu_i \neq \mu_{i'}$$
$$\text{donde } i \neq i',\ i = 1, ..., k$$

En este diseño se tienen dos componentes de variación: el que se refiere a la variabilidad que se atribuye entre los promedios de las k poblaciones que se están comparando, y la variabilidad interna dentro de cada una de las k poblaciones; a esta última se le conoce como *error experimental*. En la tabla 9.3 se distinguen k muestras o grupos, en el ámbito experimental: tratamiento, cada una con diferente tamaño, es decir: $n_1, n_2, \ldots, n_k$. Comentario: nuevamente, se llama la atención en ver que en estadística se construyen las estructuras de variación indicando qué tan alejados están los datos de la media (referencia). Primero, se observa qué discrepancia hay de la media de cada muestra con respecto a la llamada gran media. Esta última es la media de todas las observaciones y se denota por $\overline{x}_{...}$. Que la suma de los cuadrados de estas discrepancias dividido por el número de muestras menos uno, da lugar a una varianza. Luego, se percibe la diferencia de dentro de cada muestra y que la suma de los cuadrados de estas discrepancias dividida entre el tamaño de la muestra menos uno, genera la varianza correspondiente. Esta idea esquemática proporcionada por la tabla 9.3 permitirá tener una mayor claridad en la construcción del estadístico para probar las hipótesis para más de dos poblaciones.

Es importante destacar que hay dos índices i y j, los cuales desempeñan un papel central para identificar número de muestras, el primero, $i = 1, 2, \ldots, k$, el segundo, el tamaño de muestra en cada grupo, así $j = 1, 2, \ldots, n_i$.

Tabla 9.3 Ilustración del cálculo de los cuadrados medios para el análisis de la varianza.

	Muestra 1	**Muestra 2**	$\cdots$	**Muestra K**	**Total**
CM_{dentro} Varianza dentro	$(x_{11}-\overline{x}_{1\cdot})^2$	$(x_{21}-\overline{x}_{2\cdot})^2$	$\cdots$	$(x_{k1}-\overline{x}_{k\cdot})^2$	
	$(x_{12}-\overline{x}_{1\cdot})^2$	$(x_{22}-\overline{x}_{2\cdot})^2$	$\cdots$	$(x_{k2}-\overline{x}_{k\cdot})^2$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$(x_{1n1}-\overline{x}_{1\cdot})^2$	$(x_{2n2}-\overline{x}_{2\cdot})^2$	$\ldots$	$(x_{kn_k}-\overline{x}_{k\cdot})^2$	
	S_1^2	S_2^2	$\ldots$	S_k^2	
	$\overline{x}_{1\cdot}$	$\overline{x}_{2\cdot}$	$\cdots$	$\overline{x}_{k\cdot}$	$\overline{x}_{\cdot\cdot}$
	$(\overline{x}_{1\cdot}-\overline{x}_{\cdot\cdot})^2$	$(\overline{x}_{2\cdot}-\overline{x}_{\cdot\cdot})^2$	$\cdots$	$(\overline{x}_{k\cdot}-\overline{x}_{\cdot\cdot})^2$	
CM_{entre} Varianza entre	$\dfrac{n_1(\overline{x}_{1\cdot}-\overline{x}_{\cdot\cdot})^2+n_2(\overline{x}_{2\cdot}-\overline{x}_{\cdot\cdot})^2+\cdots+n_k(\overline{x}_{k\cdot}-\overline{x}_{\cdot\cdot})^2}{N-k}$				

La finalidad de la prueba de hipótesis es comparar estos dos componentes de variabilidad; si éstos resultan ser iguales se concluye que la variabilidad entre promedios de la población no se considera significativa, es decir, no se rechaza la hipótesis nula; $H_0 : \mu_1 = \mu_2 = ... = \mu_k$.

La variación entre poblaciones mide qué tan diferente son los tratamientos de la gran media. Ésta se llama el cuadrado medio de tratamientos y se denota por CM_{trat}.

El cuadrado medio de tratamientos CM_{trat} se obtiene dividiendo la suma de cuadrados entre tratamientos, entre el número de tratamientos menos 1, esto es:

$$CM_{trat} = \frac{SC_{trat}}{k-1}$$

La fórmula para la suma de cuadrados entre tratamientos es:

$$SC_{trat} = \sum_{i=1}^{k} n_i \left(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot}\right)^2,$$

donde $\overline{x}_{i\cdot}$ es el promedio de los k tratamientos.

La variación dentro de tratamientos toma en cuenta la varianza y el tamaño de muestra en cada tratamiento; a éste se le conoce como el cuadrado medio dentro de tratamientos y se denota por:

$$CM_{dentro} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + ... + (n_k-1)S_k^2}{(n_1-1)+(n_2-1)+...+(n_k-1)} \tag{9.30}$$

donde $(n_i - 1)S_i^2 = \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i.})^2$, $i = 1, 2, ..., k$

Las componentes de la suma de cuadrados entre y dentro de tratamientos son parte de un gran total; éste se denomina la suma de cuadros total. Así la fuente de variación está compuesta por estos tres elementos. En resumen: la suma de cuadrados total (SCT=SC_{total}) es la medida de la variación en el conjunto de datos del experimento. Ésta se escribe por:

$$SC_{total} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{..})^2$$

donde k es el número de tratamientos y n_i el tamaño de muestra en cada tratamiento. La expresión general es:

$$SC_{total} = SC_{trat} + SC_{dentro}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{..})^2 = \sum_{i=1}^{k} n_i (x_{i.} - \overline{x}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{i.})^2$$

Observación. El numerador de la expresión 9.30 se le conoce como la suma de cuadrados dentro de tratamientos. El denominador son los grados de libertad asociados a esta fuente de variación, los cuales son: $(n_1 - 1) + (n_2 - 1) + (n_k - 1) = n_1 + n_2 + ... + n_k - k = N - k$.

La variable que resulta de dividir el cuadrado medio entre y el cuadrado medio dentro de tratamientos tiene una distribución F con $gl_n = k - 1$ grados de libertad en el numerador y $gl_d = N - k$ grados de libertad en el denominador; esto es $F(k - 1, N - k, 1 - \alpha)$.

Los resultados de este procedimiento se recogen en una tabla que se conoce como *análisis de varianza* En la tabla 9.4 se describen las expresiones para realizar los cálculos realizados.

Procedimiento de la prueba de hipótesis

1. Plantear las hipótesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, H_1 : al menos una media de una población es diferente de otra.

2. Para seleccionar el nivel de significancia se calculan los grados de libertad para el numerador gl_n y denominador gl_d, se calcula el valor crítico referencia $F_c = F(gl_n, gl_d, 1 - \alpha)$.

3. Calcular el valor $F_m = \frac{CM_{trat}}{CM_{dentro}}$.

4. Comparar F_m con F_c si $F_m > F_c$ se rechaza H_0 y se concluye a favor de la hipótesis alterna.

5. Interpretar en el contexto del problema.

Tabla 9.4 Análisis de la varianza.

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	Fm
Entre tratamientos	$k-1$	SC_{trat}	$CM_{trat} = \frac{SC_{trat}}{k-1}$	$\frac{CM_{trat}}{CM_{dentro}}$
Dentro tratamientos	$N-k$	SC_{dentro}	$CM_{dentro} = \frac{SC_{dentro}}{N-k}$	
Total	$N-1$	SC_{total}		

Ejemplo 9.11

Debido a que el tema de salud es un factor económico que se debe considerar, en la actualidad la licenciatura en nutrición se imparte en muchas universidades. Por otro lado, en muchas ciudades existen consultorios en esta disciplina. Un nutriólogo está probando cuatro tratamientos (A, B, C y D) para reducir el peso. Cuenta con 18 personas para realizar el experimento y aplica de manera aleatoria uno de los tratamientos a cada persona; después de un periodo de prueba anota el número de kilogramos que redujeron las personas. En la tabla del ejemplo se anota el registro (6.5 indica que una persona redujo 6 kilogramos con 500 gramos). ¿Cuál es la hipótesis que se plantea verificar el nutriólogo? ¿Cuáles son los resultados principales para construir el estadístico de prueba F? Resultados:

	Tratamiento			
	A	B	C	D
	6.5	6.0	5.5	4.6
	6.5	5.8	5.2	4.5
	5.8	5.8	4.9	4.5
	5.7	5.5	4.8	4.8
	5.6	5.1		
n_j	5	5	4	4
$\overline{x}_j$	6.02	5.64	5.10	4.60
S^2	0.197	0.123	0.100	0.020

Solución

1. El planteamiento de las hipótesis es:

$$H_1 : \mu_A = \mu_B = \mu_C = \mu_D$$

(Los cuatro tratamientos producen en promedio la misma reducción de peso).

$$H_2 : \text{al menos un par de tratamientos es diferente}$$

2. Considere un nivel de significancia $\alpha = 0.05$, los grados de libertad correspondientes para el numerador y denominador son en este caso: $gl_n = 4-1 = 3$, $gl_d = 18-4 = 14$. Así $F_c = F(3, 14, 0.95) = 3.344$, figura 9.22.

3. Procedimiento operativo para calcular el estadístico F_m. Con la información reportada en la tabla del ejemplo se pueden construir los cuadrados medios entre y dentro de tratamientos, es decir:

$$CM_{trat} = \frac{5(6.02-5.394)^2 + 5(5.64-5.394)^2 + 4(5.10-5.394)^2 + 4(4.60-5.394)^2}{4-1} = 1.709$$

$$CM_{dentro} = \frac{(5-1)(0.197) + (5-1)(0.123) + (4-1)(0.100) + (4-1)(0.020)}{(5-1)+(5-1)+(4-1)+(4-1)} = 0.117$$

Así $F_m = \frac{1.709}{0.117} = 14.596$.

4. Puesto que $F_m = 14.596 > F_c = 3.344$ se rechaza la hipótesis nula, lo que indica que alguno de los tratamientos es efectivo para reducir de peso.

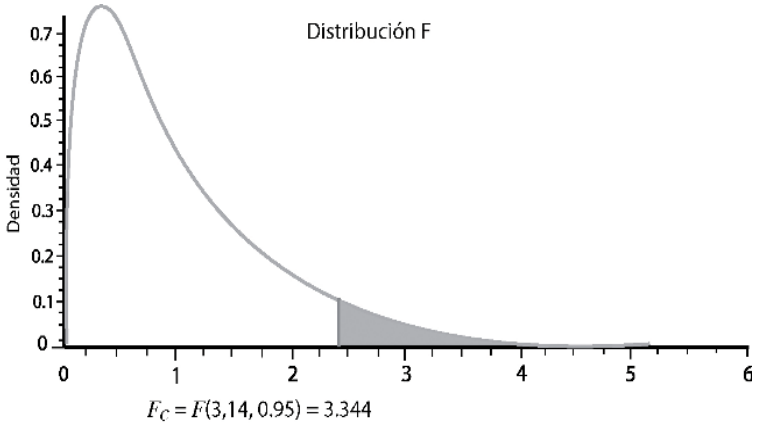


Figura 9.22 Valor crítico en la distribución F.

9.8 Resumen

Resumen y guía para realizar la prueba

Resumen y guía para realizar la prueba diferencia de medias-muestras independientes	
Explicación	Ecuaciones
1. Identificar las hipótesis nula y alternativa	$H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 < \mu_2$, $\mu_1 > \mu_2$, o $\mu_1 \neq \mu_2$
2. Especificar el nivel de significancia	Dar el valor de α o $\alpha/2$
Determinar el punto crítico (Identificando la distribución)	Usar $(\overline{x}_1 - \overline{x}_2)_c$, z_c, la t_c, $t - Student$,
3. Calcular el estadístico de prueba	Estimar la expresión $(\overline{x}_1 - \overline{x}_2)_m$ Aplicar la expresión 9.3 : z_m Calcular t_m
4. Tomar la decisión estadística: rechazar o no rechazar Interpretar los resultados	Comparar $(\overline{x}_1 - \overline{x}_2)_m$ con $(\overline{x}_1 - \overline{x}_2)_c$ Estándar Comparar z z_m con z_c Estándar Comparar t t_m con t_c Donde c es ci o cd

Resumen y guía para realizar la prueba diferencia de medias-muestras pareadas	
Explicación	Ecuaciones
1. Identificar las hipótesis nula y alternativa	$H_0 : \mu_d = 0$ $H_1 : \mu_d < 0$, $\mu_d > 0$, o $\mu_d \neq 0$
2. Especificar el nivel de significancia	Dar el valor de α o $\alpha/2$
Determinar el punto crítico (Identificando la distribución)	Usar $(\overline{x}_d)_c$, la $t - Student$, t_c
3. Calcular el estadístico de prueba	Estimar la expresión $(\overline{x}_d)_m$ Aplicar la expresión 9.14 Para calcular t_m
4. Tomar la decisión estadística: rechazar o no rechazar Interpretar los resultados	Comparar $(\overline{x}_d)_m$ con $(\overline{x}_d)_c$ Estándar comparar t_m con t_c Donde c es ci o cd

Resumen y guía para realizar la prueba diferencia de proporciones	
Explicación	Ecuaciones
1. Identificar las hipótesis nula y alternativa	$H_0 : p_1 = p_2$ $H_1 : p_1 < p_2$, $p_1 > p_2$, o $p_1 \neq p_2$
2. Especificar el nivel de significancia	Dar el valor de α
Determinar el punto crítico (Identificando la distribución)	Usar $(\widehat{p}_1 - \widehat{p}_2)_c$, z_c,
3. Calcular el estadístico de prueba	Estimar la expresión $(\widehat{p}_1 - \widehat{p}_2)_m$ Aplicar la expresión 1.3: z_m
4. Tomar la decisión estadística: rechazar o no rechazar	Comparar $(\overline{x}_d)_m$ con $(\overline{x}_d)_c$ Comparar $(\widehat{p}_1 - \widehat{p}_2)_m$ con $(\widehat{p}_1 - \widehat{p}_2)_c$ Comparar z_m con z_c
Interpretar los resultados	

Resumen y guía para realizar la prueba razón de varianzas	
Explicación	Ecuaciones
1. Identificar las hipótesis nula y alternativa	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_1 : \frac{\sigma_1^2}{\sigma_2^2} < 1$, $\frac{\sigma_1^2}{\sigma_2^2} > 1$, o $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$
2. Especificar el nivel de significancia	Dar el valor de α
Determinar el punto crítico (Identificando la distribución)	Usar $\left(\frac{(n_1-1)S_1^2}{\sigma_1^2} / \frac{(n_2-1)S_2^2}{\sigma_2^2}\right)_c$, F_c,
3. Calcular el estadístico de prueba	Estimar la expresión $\left(\frac{(n_1-1)S_1^2}{\sigma_1^2} / \frac{(n_2-1)S_2^2}{\sigma_2^2}\right)_m$ Aplicar la expresión 1.3: F_m
4. Tomar la decisión estadística: rechazar o no rechazar	Comparar $\left(\frac{(n_1-1)S_1^2}{\sigma_1^2} / \frac{(n_2-1)S_2^2}{\sigma_2^2}\right)_m$ con $\left(\frac{(n_1-1)S_1^2}{\sigma_1^2} / \frac{(n_2-1)S_2^2}{\sigma_2^2}\right)_c$ Comparar F_m con F_c
Interpretar los resultados	

Resumen del cálculo de los cuadrados medios para el análisis de la varianza.

	Muestra 1	Muestra 2	$\cdots$	Muestra K	Total
CM_{dentro} Varianza dentro	$(x_{11}-\overline{x}_{1\cdot})^2$	$(x_{21}-\overline{x}_{2\cdot})^2$	$\cdots$	$(x_{k1}-\overline{x}_{k\cdot})^2$	
	$(x_{12}-\overline{x}_{1\cdot})^2$	$(x_{22}-\overline{x}_{2\cdot})^2$	$\cdots$	$(x_{k2}-\overline{x_{k\cdot}})$	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$(x_{1n1}-\overline{x_{1\cdot}})^2$	$(x_{2n2}-\overline{x_{2\cdot}})^2$	$\cdots$	$(x_{kn_k}-\overline{x}_{k\cdot})^2$	
	S_1^2	S_2^2	$\cdots$	S_k^2	
	$\overline{x}_{1\cdot}$	$\overline{x}_{2\cdot}$	$\cdots$	$\overline{x}_{k\cdot}$	$\overline{x}_{\cdot\cdot}$
	$(\overline{x}_{1\cdot}-\overline{x}_{\cdot\cdot})^2$	$(\overline{x}_{2\cdot}-\overline{x}_{\cdot\cdot})^2$	$\cdots$	$(\overline{x}_{k\cdot}-\overline{x}_{\cdot\cdot})^2$	
CM_{entre} Varianza entre	$\dfrac{n_1(\overline{x}_{1\cdot}-\overline{x}_{\cdot\cdot})^2+n_2(\overline{x}_{2\cdot}-\overline{x}_{\cdot\cdot})^2+\cdots+n_k(\overline{x}_{k\cdot}-\overline{x}_{\cdot\cdot})^2}{N-k}$				

9.9 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

9.10 Ejercicios

Prueba de hipótesis para la diferencia de medias: muestras independientes.

9.1 Uno de los responsables de la administración en un distrito escolar desea comparar el rendimiento de los estudiantes de preparatoria abierta y el sistema escolar en los exámenes de matemáticas. En el

primer sistema se usa para preparar a los estudiantes un tipo de material que se denominará A, en el escolar usan un material tipo B. Si μ_1 es la media de la calificación que obtendrán los estudiantes en el sistema abierto y μ_2 la media para los estudiantes del sistema escolar, ¿serán las medias diferentes? En la tabla siguiente se reportan los resultados que se obtuvieron al evaluar una muestra aleatoria de cada sistema.

Material	Número de estudiantes	Media muestral	Desviación estándar muestral
A	$n_1 = 35$	$\overline{x}_1 = 78$	$S_1 = 8$
B	$n_2 = 40$	$\overline{x}_2 = 75$	$S_2 = 6$

1. Plantee las hipótesis y realice la prueba de hipótesis usando como estadísticos la diferencia muestral y el valor de Z, distribución normal estándar. Considere para este caso $\alpha = 0.1$.
2. Calcule el $valor - p$ y compare con el nivel de significancia α, obtenga su conclusión y arguméntela.
3. Estime el intervalo de 90 % de confianza e interprete los valores.

9.2 Subraye el inciso que responde correctamente a la siguiente pregunta: si un intervalo de confianza del 99 % para $\mu_1 - \mu_2$ está dada por $4.3 < (\mu_1 - \mu_2) < 8.1$, ¿cuáles de las siguientes conclusiones se pueden obtener basándose en este intervalo?

1. Aceptar $H_0 : \mu_1 = \mu_2$ con $\alpha = 0.05$ si su alternativa es $H_a : \mu_1 \neq \mu_2$
2. Rechazar $H_0 : \mu_1 = \mu_2$ con $\alpha = 0.01$ si su alternativa es $H_a : \mu_1 \neq \mu_2$
3. Rechazar $H_0 : \mu_1 = \mu_2$ con $\alpha = 0.01$ si su alternativa es $H_a : \mu_1 < \mu_2$
4. Aceptar $H_0 : \mu_1 = \mu_2$ con $\alpha = 0.01$ si su alternativa es $H_a : \mu_1 \neq \mu_2$
5. Aceptar $H_0 : \mu_1 = \mu_2 + 3$ con $\alpha = 0.01$ si su alternativa es $H_a : \mu_1 \neq \mu_2 + 3$

9.3 El director de cierta escuela primaria piensa que 6 de 10 padres de familia tiene problemas económicos para surtir la lista de útiles escolares que se pide al inicio de curso. Un profesor de la misma escuela dice no estar de acuerdo con dicha información, y para fundamentar su conjetura pregunta a 35 padres de familia si tuvieron o no problemas económicos para surtir sus listas, encontrando una respuesta afirmativa en 23 de ellos. Utilizando una $\alpha = 10$ pruebe la hipótesis del profesor.

9.4 Como parte de un proyecto de investigación, un psicólogo seleccionó una muestra aleatoria de 12 muchachas y otra de 9 muchachos. Luego, le pidió a cada individuo que dibujara una figura masculina. El tiempo promedio que utilizaron la mujeres fue de 8 minutos con una varianza de 18 minutos. Para los hombres, el tiempo medio fue de 13 minutos con una varianza de 22.5. ¿Indican estos datos que los hombres en promedio gastan más tiempo en dibujar una figura masculina que las mujeres? Considere un

nivel de significancia de 0.05.

9.5 Dos métodos de memorización se usan para determinar cuál produce mejor retención. Se seleccionan nueve parejas de estudiantes y se incluyen en el estudio. Las parejas se forman de acuerdo a su I.Q., y sus hábitos de estudio, los métodos son asignados aleatoriamente a cada elemento de la pareja. Las calificaciones obtenidas son las siguientes:

	1	2	3	4	5	6	7	8	9
Método A	90	86	72	65	44	52	46	38	43
Método B	85	87	70	62	44	53	42	35	46

Determine si hay una diferencia significativa en la efectividad de los dos métodos. Aplicando la prueba t de Student.

9.6 En el estudio de la característica de calidad de un producto en agronomía económica, se tienen dos procedimientos asegurados por reguladores de precisión (1:BA/2,4-D 2:K/2,4-D); después de cinco semanas se tiene la información del promedio y desviación estándar como sigue:

Regulador 1	Regulador 2
$\overline{x} = 0.0976$	$\overline{y} = 0.0789$
$S_1 = 0.117$	$S_2 = 0.382$

1. ¿En qué regulador hubo mayor variación? ¿Por qué?
2. ¿Existe diferencia entre los reguladores?
3. Estime el intervalo de 90 % de confianza?

9.7 Una agencia de venta de casas, tiene el registro del tiempo que tardan en vender casas en las zonas sur y norte. La administración piensa que en la zona sur se venden más rápido. Seleccionan una muestra de 11 casas en la zona sur y 14 casas en la otra zona. Los datos son:

Zona sur	102	78	71	81	129	67	78	110	131	136	92
Zona norte	164	114	49	87	180	153	110	134	62	102	146
	168	90	132								

Verifique la consideración de la administración; use $\alpha = 0.025$

9.8 Se conoce que existen lineamientos para establecer el salario mínimo. La gerencia de una secretaría considera que no hay diferencia entre los salarios de los obreros de empresas transnacionales con los de empresas nacionales. El resumen de la información del salario, en pesos, a la semana de obreros especializados se presenta en la siguiente tabla:

Transnacionales	Nacionales
$n_1 = 22$	$n_2 = 18$
$\overline{x}_1 = 2190$	$\overline{x}_2 = 1938$
$S_1^2 = 1150.25$	$S_2^2 = 1088.75$

1. Escriba el procedimiento para probar la hipótesis, con $\alpha = 0.05$, use el estadístico $(\overline{X}_1 - \overline{X}_2)$.
2. Similar al inciso a, pero use el estadístico $t - Student$.
3. ¿Cuál es el $valor - p$, el nivel de significancia descriptivo?
4. Estime el intervalo de 95 % de confianza.

9.9 Se supone que el tiempo de permanencia en un hospital para un mismo tratamiento no debe ser diferente. Una compañía de seguros realiza un estudio considerando dos médicos diferentes. La información de 20 pacientes para el mismo tratamiento, 10 con el médico 1 y los otros 10 con el médico 2, es la siguiente:

Médico 1	4	5	6	2	6	4	3	4	4	5
Médico 2	3	6	6	4	3	3	4	2	4	4

1. Plantee sus hipótesis considerando el supuesto.
2. Pruebe la hipótesis; use $\alpha = 0.02$

9.10 Una empresa citadina planea trasladar su empresa a zonas rurales, el motivo es la reducción del salario y prestaciones a los trabajadores. Un asesor de una empresa independiente desarrolla el proyecto. Éste lo realiza con apoyo económico gubernamental por aquello de generar empleos.

1. ¿Cuál es la suposición que considera la administración de la empresa? Escriba en términos de hipótesis.
2. Después de tres meses, mientras cierra la empresa la planta en la ciudad, se toma una muestra de 10 obreros en la zona rural y 10 obreros de la zona urbana, en cada caso se les pregunta el salario semanal. Con la información que se recoge en la tabla siguiente plantee las hipótesis estadísticas y pruébelas usando un nivel de significancia de 0.01.
3. ¿Cuál es el efecto económico de la empresa?

Rural	1024 920	780 780	815 850	715	1100	760	1300
Urbana	1100 1530	1350 1700	1520 1800	1460	1650	1150	1900

9.11 Las plantas A y B de una empresa elaboran harinas. Una característica de calidad importante es la humedad; éstas tienen por valor objetivo $70 \pm \Delta$. La administración de la empresa quiere saber si existe diferencia en las desviaciones al valor objetivo en estas dos plantas. Los valores son:

Planta											$\overline{y}$	S^2	gl
A	14	-1	13	22	26	4	-17	8	12	-2			
B	23	13	-11	-9	1	-7	2	-9	-4	-10			

1. Diga cómo realizaría el estudio.
2. Plantee la hipótesis que se prueba.
3. Realice los cálculos para obtener el estadístico de prueba.
4. Pruebe la hipótesis.
5. Obtenga sus conclusiones.
6. ¿Qué impacto tiene esta desviación en los costos?

9.12 En un estudio se encontró que el sueldo semanal, en miles de pesos, en una muestra aleatoria de mujeres y de hombres con cierta ocupación fue:

Mujeres	5.2	5.6	4.5	5.4
Hombres	6.0	5.7	6.2	5.2

Considere que esta variable se distribuye normalmente, investigue si hay evidencia de discriminación sexual (en cualquier sentido) en los sueldos pagados en esta ocupación (use $\alpha = 0.5$). Conteste lo que se pide en los incisos de las preguntas siguientes:

1. Plantee sus hipótesis:

H (Del estudio)
H_0:
H_1:

2. Plantee su región de rechazo.
3. Si el valor del estadístico calculado es: $\frac{-0.58}{.46\sqrt{.5}} = -1.78$

   a) ¿Qué decisión estadística tomaría?

   b) ¿Cómo interpreta en términos del problema su conclusión estadística?

Prueba de hipótesis para la diferencia de medias: muestras pareadas

9.13 Calcule el $valor - p$ para cada una de las siguientes pruebas de hipótesis:

$H_1 : \mu_d < 0$	$n = 15$	$t_1 = -1.86$
$H_1 : \mu_d \neq 0$	$n = 9$	$t_1 = 2.23$
$H_1 : \mu_d > 0$	$n = 24$	$t_1 = 2.76$
$H_1 : \mu_d > 0.54$	$n = 16$	$t_1 = 2.86$

9.14 Use el estadístico $\overline{d}$ para establecer el criterio para probar la hipótesis y aplíquelo a los siguientes casos:

$H_1 : \mu_d < 0$	$n = 15$	$\alpha = 0.01$
$H_1 : \mu_d \neq 0$	$n = 12$	$\alpha = 0.10$
$H_1 : \mu_d > 0$	$n = 9$	$\alpha = 0.05$
$H_1 : \mu_d > 0.54$	$n = 16$	$\alpha = 0.025$

9.15 Siguiendo el procedimiento 2 planteado en muestras pareadas, la administración de una empresa siguió el estudio del desgaste de su llanta P en comparación con la de la competencia, llanta R. Los datos reportados son:

Carro	1	2	3	4	5	6	7
Llanta P	120	84	95	48	116	90	112
Llanta R	135	87	107	46	125	102	113

1. ¿Cuáles son los valores de las diferencias?
2. Estime la media y varianza de las diferencias.
3. Plantee las hipótesis.
4. Realice la prueba.

9.16 En cuanto a los salarios que perciben los recién egresados de la carrera de ingeniería industrial, se piensa que los egresados de una escuela pública tienen un sueldo menor respecto a los que se gradúan de centros de educación superior privados. Se consulta a doce ingenieros y la quincena en miles de pesos es la que se describe en la tabla de abajo:

Pública	3.9	4.3	5.2	1.4	1.8	5.2	2.7	3.4	3.9	8.0	4.5	1.3
Privada	5.7	6.7	8.3	6.6	5.6	4.8	8.8	9.4	7.5	6.3	4.7	7.8

1. ¿Cómo justificaría la comparación de muestras pareadas?

2. Plantee sus hipótesis y explíquelas.

3. Use la prueba $t - Student$, con $\alpha = 0.1$.

4. Interprete sus resultados.

9.17 Por ser un problema de salud y que tiene un impacto económico, en la actualidad es frecuente encontrar tratamientos para adelgazar pero que mantienen el nivel de nutrición en buenos niveles. Diez personas se someten a un programa nuevo coordinado por una investigadora en líneas naturales para bajar de peso. Los datos en la tabla indican el peso, en kg, antes y un mes después del tratamiento.

Antes	98	92	80	62	92	78	89	81	71	80
Después	93	87	73	59	91	65	80	78	65	73

1. ¿Cómo plantearía sus hipótesis para este proyecto?

2. Elabore un diagrama de caja para cada una de las situaciones del antes y después.

3. Verifique si el programa fue eficiente; use un nivel de significancia de 0.01.

4. Construya e interprete el intervalo de confianza de 95 % para la diferencia de pesos, con este tipo de muestras.

9.18 La gerencia de una empresa está preparando un grupo de personas para que desarrolle habilidades para la administración. Con ese objetivo contrata una agencia para que capacite al personal. Selecciona una muestra de 15 personas y les hace una evaluación inicial, luego del programa de capacitación evalúa de nuevo. El cuestionario aplicado consta de varios reactivos divididos en 6 rubros; el porcentaje total es de 100, los resultados de la puntuación es:

Antes	89	74	75	61	89	63	79	73	80	65	73	81
Después	92	83	83	56	91	58	83	80	95	80	74	83

1. Describa las hipótesis a probar y explíquelas.

2. Pruebe las hipótesis usando un $\alpha = 0.01$.

3. Calcule un intervalo de 95 % de confianza para la diferencia de medias. Con esta información ¿qué le diría a la agencia que capacitó?

9.19 En dos hospitales llevan a cabo el mismo tratamiento médico con una estancia en el hospital de 2 días. La agencia que vendió el seguro de gastos médicos mayores supone que el costo en ambos centros es el mismo. La administradora analiza 12 facturas del costo, en pesos, de cada hospital y la información es:

Hospital 1	5100	5820	4710	5900	6270	5820
	5900	6100	3780	3810	4390	4100
Hospital 2	5600	6990	4930	5690	6920	4530
	5800	3820	5290	5220	5970	6390

1. Haga un diagrama de caja para cada hospital y comente ambas gráficas.
2. Verifique el supuesto de la agencia de seguros, $\alpha = 0.05$

9.20 Economía familiar: se realiza la comparación en el precio de 10 medicinas seleccionadas al azar, de la sección de farmacias en dos cadenas grandes de distribución. Se puede considerar que una de las compañías tiene mayor precio que la otra. Los resultados de 10 medicinas en pesos son

Medicina	Empresa 1	Empresa 2
M1	390	420
M2	900	850
M3	590	600
M4	320	270
M5	275	275
M6	580	480
M7	720	720
M8	800	790
M9	440	440
M10	380	330

1. Calcule la diferencia entre los precios en cada una de las medicinas. Con la simple apreciación de estas diferencias, hay motivo para pensar que una empresa es más cara.
2. Obtenga la media y la desviación estándar de estas diferencias.
3. Plantee las hipótesis para esta situación.
4. En el supuesto de que los precios siguen una distribución normal, pruebe la hipótesis usando un nivel de significancia de 0.05.

Prueba de hipótesis para la diferencia de proporciones

9.21 Verifique que la diferencia de proporciones en un estudio para adelgazar no difiere. En un método de 75 personas adelgazaron 72 % y en el otro método de 80 personas bajaron de peso 66 %.

1. Plantee sus hipótesis

2. Realice la prueba con $\alpha = 0.01$.

9.22 Una empresa grande presenta un problema de ausentismo laboral. La administración está trabajando para resolver la situación, ya que en el último mes, en el turno matutino 31 de 125 personas no se presentaron y en el turno vespertino 42 de 120 no llegaron a laborar. ¿Es diferente el porcentaje en ambos turnos?

9.23 Un sociólogo afirma que el porcentaje de profesionistas que desea emigrar a provincia es mayor que el porcentaje de obreros que también les gustaría emigrar. Para corroborar lo anterior, escoge una muestra aleatoria de profesionistas y otra de obreros, registrando la siguiente información:

	n	desean emigrar
Profesionista	90	37
Obreros	110	38

(use $\alpha = 0.5$). Para este problema, construya la:

1. Hipótesis de investigación
2. Hipótesis nula
3. Hipótesis alterna
4. Región de rechazo de H_0

9.24 En una universidad ubicada en el sur de una ciudad grande se tomó una muestra de 1000 alumnos becados y de éstos al final del semestre aprobaron el 40 %, y en otra, situada en el norte, de los 900 becados aprobó 47 %. ¿Existe diferencia en el rendimiento de los estudiantes de estas universidades?

9.25 Se piensa que las empresas prefieren contratar egresados de centros escolares privados que los no privados, para puestos que consideran básicos. Se realizó un estudio en una zona industrial mediante pruebas para conseguir el trabajo; de una encuesta aplicada se tienen los siguientes resultados:

	Aceptados	Rechazados	Total
No privadas	32	27	59
Privadas	196	51	247

1. Planteé las hipótesis correspondientes.
2. Realice la prueba con $\alpha = 0.025$. Use los tres estadísticos.
3. Interprete sus resultados.

Prueba de hipótesis para la razón de varianzas

9.26 Utilizando la calculadora de distribuciones o las tablas, encuentre los valores de las siguientes distribuciones F:

1. $F(5, 10, 0.05) =?$ Es decir, con 5 y 10 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.05$.
2. $F(10, 18, 0.01) =?$ Es decir, con 10 y 18 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.01$.
3. $F(15, 10, 0.75) =?$ Es decir, con 15 y 10 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.05$.
4. $F(10, 20, 0.975) =?$ Es decir, con 10 y 20 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.05$.
5. $F(18, 10, 0.01) =?$ Es decir, con 18 y 10 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.05$.
6. $F(18, 20, 0.95) =$ Es decir, con 18 y 20 grados de libertad en el numerador y denominador, respectivamente, y $\alpha = 0.05$.

9.27 Se toman 20 muestras de la población 1 que corresponde a un grupo de hombres que ingresó al trabajo recientemente, y se obtiene una varianza muestral igual a 6.5 y una muestra de 25 a la población 2 que corresponde a un grupo de hombres que cuentan con una capacitación en administración, con una varianza muestral de 4.2. La variable es el tiempo de espera en un servicio que ofrece una empresa grande.

1. Pruebe la siguiente hipótesis con un nivel de significancia de 0.05 y dé su conclusión.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 > \sigma_2^2$$

2. Con los datos del problema anterior pruebe la siguiente hipótesis, con un nivel de significancia de 0.01:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

3. Si los datos cambian a $n_1 = 25$, $S_1^2 = 4.1$, $n_2 = 20$ y $S_2^2 = 7.8$, ¿cuál es su conclusión?

9.28 Una compañía de publicidad quiere saber qué género, hombres o mujeres, pasan más tiempo en las redes sociales. El estudio concluyó que las mujeres pasan más tiempo en promedio. Suponga que se realiza un estudio de seguimiento y se les pregunta a 25 hombres y a 30 mujeres el tiempo que pasan en las redes

sociales. La desviación estándar del tiempo que dedican a sitios de redes sociales fue de 29 y 47 minutos, respectivamente. Con un nivel de significancia de 0.01, determine si la muestra apoya la conclusión de que las mujeres tienen una varianza mayor del tiempo que pasan en las redes sociales que los hombres.

9.29 Estudios han demostrado que estudiantes que realizan sus actividades con distracciones tienen en promedio calificaciones más bajas que aquellos que no tienen distracciones. Se revisan las calificaciones de 16 estudiantes en ambas condiciones. La desviación estándar con distracciones fue de 2.3, mientras que sin distracciones fue de 1.6. Con un nivel de significancia de 0.05, ¿considera usted que los datos justifican la afirmación de que la varianza de las calificaciones con distracciones es mayor que la varianza sin distracciones?

9.30 Durante los últimos doce meses, la desviación estándar del costo de 10 materias primas para fabricar el producto A fue de 5.23 %, mientras que la desviación estándar del costo de 8 materias primas necesarias en la elaboración del producto B fue de 3.98 %. Realice una prueba para varianzas iguales con $\alpha = 0.05$ y exponga sus conclusiones sobre la variación del costo de las materias primas de ambos productos.

9.31 La administración de una empresa realiza un experimento para evaluar la habilidad que desarrollan dos grupos de trabajadores. Suponen que existe heterogeneidad entre los grupos y plantean la siguiente hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

La información que tienen del experimento es: $n_1 = 9$, $S_1^2 = 5.2$, $n_2 = 7$ y $S_2^2 = 4.8$.

1. Verifique el supuesto de la administración, con $\alpha = 0.05$.

2. Construya el intervalo de 95 % de confianza para la razón de varianzas.

9.32 La varianza en la venta de autos compactos usados de una marca es mayor a la de otra marca.

Prueba de hipótesis para más de dos poblaciones

9.33 En una investigación en el área de psicología realizada a cuatro grupos de trabajadores en una empresa, se obtuvo el siguiente resumen estadístico en un experimento de diseño completamente al azar.

		Grupo		
	I	II	III	IV
	16	24	16	25
	7	6	15	19
	19	15	18	16
	24	25	19	17
	31	32	6	42
		24	13	45
		29	18	
$\overline{x}_j$	19.40	22.14	15.00	27.33

De estos datos se presenta un resumen de la tabla del análisis de la varianza, ANDEVA.

Tabla de ANDEVA

Fuente de variación	gl	Suma de cuadrados	Cuadrados medios	F
Entre grupos	3	513.97		
Dentro de grupos	21	1749.36		
Total	24	2263.33		

Con esta información conteste lo siguiente:

1. Plantee la hipótesis nula.
2. ¿Cuál es su decisión estadística?
3. Realice la prueba de comparaciones múltiples y obtenga sus conclusiones.

9.34 El tiempo de atención al cliente en tres sectores diferentes en una empresa que presta servicios, se describe en la siguiente tabla. Si existe una diferencia el administrador de la empresa tiene que realizar un trabajo de capacitación para mejorar los tiempos.

Sectores		
A	B	C
286	324	322
262	339	256
318	259	278
278	319	295
359	339	284

1. Verifique que al menos un par de medias de los sectores es diferente, use 0.05.

2. Construya un intervalo de 99 % de confianza para la diferencia de medias en los pares de sectores A-B, A-C, B-C, interprete los resultados.

9.35 En una investigación se asignan a los voluntarios una de cuatro terapias diferentes. Debido a las dificultades para grabar el número de palabras emitidas por los participantes, éstos se asignaron en diferentes cantidades a cada terapia. ¿Existe diferencia entre las diferentes terapias?

9.36 En un proceso industrial, la administración de una empresa decide realizar una prueba de tensión en un plástico. Del proceso se toman aleatoriamente 12 muestras, posteriormente se envían cuatro muestras aleatorias a tres laboratorios diferentes. La prueba de tensión realizada por los tres laboratorios se describe en la siguiente tabla. El interés en este experimento es conocer si existe una diferencia en las mediciones realizadas en cada laboratorio.

lab	A	B	C	
	45	54	49	
	47	55	51	
	50	50	50	
	46	53	48	
total	188	212	198	598
$\overline{y}_{i\bullet}$	47	53	49.5	$\overline{y}_{\bullet\bullet} = 49.83$
S_i^2	4.67	4.67	1.67	
L_i	44.83	50.83	47.33	
L_s	49.17	55.17	51.67	

Planteamiento estadístico. El objetivo es conocer si existe diferencia entre los laboratorios para medir la tensión del plástico.

1. Diga cómo realizaría el estudio.

2. Plantee la hipótesis que se prueba.

3. Realice los cálculos para obtener la razón de varianza.

4. Pruebe la hipótesis.

5. Obtenga los intervalos de confianza para el laboratorio 1.

6. Construya la tabla del ANDEVA.

7. Si se rechaza la hipótesis nula realice la prueba de comparaciones múltiples.

9.37 Una empresa desea evaluar el impacto económico que produce la venta de un producto que se presenta en tres colores. El número de unidades vendidas durante un mes en una muestra de sus tiendas se describen en la tabla de abajo:

Efecto color	Azul	Verde	Naranja
	31	35	28
	29	32	34
	30	37	29
	36	34	33
	32	32	31

1. Exponga las hipótesis a probar, argumente su propuesta.
2. Pruebe la hipótesis, con $\alpha = 0.05$, si existe diferencia haga las comparaciones múltiples.
3. Estime el intervalo del 95 % de cada color, ¿qué observa?

9.11 Evaluación

En esta evaluación se plantea una serie de preguntas en las que se pide que el lector seleccione la respuesta correcta, y para hacerlo es necesario que el alumno aplique los conceptos expuestos.

Capítulo 10

Modelación Estadística

El compromiso con la acción significa desplegar nuestra capacidad de ser más grandes que la situación. Sin ese compromiso, permanecemos en un estado de análisis, paralizados, ahogando nuestro talento como eternos aprendices en un mundo de maestros.

William C. Miller.

Competencia general

Proporcionar una guía para conocer y comprender los elementos estadísticos más importantes en la elaboración de un modelo de regresión. Generalmente este tipo de modelo permite explicar y estudiar diferentes situaciones que se presentan en distintas áreas del conocimiento. La idea principal es seleccionar una muestra de individuos u objetos a los cuales se les medirán dos características, de esa manera el interés es evaluar cuándo las variables se relacionan, y establecer qué tan estrecha es esa relación. Indicar cómo la variable de respuesta se puede predecir a partir de la variable independiente, y luego estimar la predicción.

Competencias específicas

- Aplicar técnicas de graficación para datos bivariados y aprender a interpretarlas.
- Aprender a reproducir e interpretar la distribución de los datos para cada una de las variables en un diagrama de dispersión.
- Comprender cómo se puede detectar mediante una gráfica y de manera intuitiva la relación entre dos variables.
- Aprender a calcular el coeficiente de correlación para obtener una medida de asociación entre variables, y saber interpretar este coeficiente.
- Encontrar la naturaleza de la relación entre las variables X y Y de un conjunto de datos y usar esta relación para predecir la variable de respuesta Y de la variable de entrada X. Posteriormente construir el intervalo de confianza para la predicción y el modelo.
- Aprender a representar y estudiar en una gráfica la relación entre variables de tipo cualitativo y cuantitativo.
- Identificar si en la relación entre variables existe alguna tendencia que permita explicar con mayor precisión las preguntas planteadas en un problema.

Continúa

Competencias específicas. Continuación

- Realizar las pruebas de hipótesis y construir los intervalos de confianza para los parámetros de correlación y del modelo de regresión.
- Construir e interpretar el análisis de la varianza para el modelo de regresión.
- Explicar las características del modelo usando el coeficiente de determinación.
- Adquirir habilidad para evaluar el modelo mediante el análisis de residuales.
- Construir el modelo de regresión múltiple y hacer inferencia estadística sobre los parámetros de ese modelo.

10.1 Introducción

Una variable de respuesta se define como la característica de interés que se observa en los elementos o sujetos en una población. Hasta ahora se han analizado datos considerando una sola variable, salvo algunos ejemplos del capítulo 2. Sin embargo, al aplicar una encuesta o realizar un estudio casi siempre se presenta más de una variable para distinguir a los individuos de una población.

En particular, este capítulo se centrará en el análisis de los datos que contienen dos variables, a los que también se conoce como *bivariados*, esto es, cuando un par de mediciones se registran para cada elemento de la muestra estudiada. En esa dirección, se estudiará la relación entre dos variables. Cada variable puede ser cuantitativa o cualitativa, y existen tres posibles combinaciones entre ambas variables: (1) las dos son cualitativas, o (2) cuantitativas, o (3) una es cuantitativa y la otra cualitativa. Como en el caso de una variable, se describirán los datos mediante el uso de técnicas gráficas para mostrar las características entre cualquier relación, y se verá el *coeficiente de correlación* que indica el grado de asociación entre dos variables de tipo cuantitativo. A continuación se expone una serie de situaciones en las que intervienen dos variables y su posible relación.

1. Como se sabe, en muchos países se fija año con año un aumento al salario mínimo. La idea, en la práctica, de este aspecto es mejorar el poder adquisitivo de los trabajadores. Una variable cuantitativa discreta son los años, en este caso se toma como referencia o se parte de un año en específico, por ejemplo los últimos 12 años, los valores son 2001, 2002,. . . ,2012. La otra variable es cuantitativa y está dada en alguna moneda, por ejemplo en pesos mexicanos. ¿Qué tanto crece el salario mínimo cada año? ¿Se puede modelar esta relación?

2. Siguiendo la dirección del ejemplo anterior, se puede ver el efecto que tiene el precio de algún producto, por ejemplo la tortilla, en el salario mínimo. La nueva variable es el precio del producto, la cual es cuantitativa. Se evalúa la relación que pueda existir entre estas variables para observar si efectivamente el salario mínimo es competitivo o genera una merma para el empleado. Al aumentar el precio de la tortilla, ¿cómo afecta en el salario mínimo?

3. En un proyecto de administración, la inversión en la promoción de un producto ¿tienen efecto en el aumento de las ventas? Variables: inversión y ventas están dadas en una moneda, y ambas son cuantitativas.

4. Se desea ver si existe una relación entre la severidad de una lesión y los días que un paciente pasa en el hospital. Esta información será de utilidad para la administración, ya que le permitirá planear la disponibilidad de camas en el hospital. La variable "severidad de la lesión" se puede medir por apreciación del médico, puede ser cualitativa y tener varias categorías, las cuales pueden ir desde leve hasta muy severa. Una alternativa interesante es utilizar un método para cuantificar esta variable, tales como la escala Linker, que se vio en el capítulo 1, o mediante el empleo de técnicas de la teoría de conjuntos difusos.

5. En la temática de economía agrícola, es importante tener conocimiento de la relación entre el nivel de fertilizante aplicado (variable cuantitativa medida en gramos) aplicado y la producción de una cosecha, variable cuantitativa medida en toneladas. Esa información ayudará a que los agricultores tengan mejores criterios para planear sus siembras.

Como complemento cultural, se describirá la relación de más de dos variables, para ello se usarán técnicas gráficas y se presentarán ejemplos para casos donde existan tres o cuatro variables.

10.2 Análisis descriptivo de datos bivariados

El mundo de la información 1. Índice de ozono: Relación entre dos variables cuantitativas

Sin duda, el impacto de la contaminación tiene efectos administrativos y económicos. En particular, para medir la contaminación ambiental se emplea el índice de ozono. Con el propósito de tener una mejor información sobre los niveles de ozono en muchas ciudades, en el ámbito mundial se colocan estaciones para registrar la contaminación. Por lo general, las lecturas en estas estaciones se reportan cada hora. Los datos de estas lecturas permitirán a las autoridades encargadas de cuidar el medio ambiente, tomar acciones pertinentes en los casos extremos. Al final del día reporta el valor máximo de ozono, entre otras mediciones.

Preguntas sobre la naturaleza del problema

Se pueden formular algunas preguntas sobre las lecturas reportadas en las estaciones: ¿el nivel de ozono

es idéntico en todas las estaciones?, ¿existen lugares con mayores niveles de ozono?, ¿al aumentar el ozono en alguna estación aumenta en otra? Para responder a estas preguntas se observarán las lecturas en dos estaciones colocadas en lugares opuestos.

Datos del índice de ozono: Los datos registrados por las lecturas del ozono en dos estaciones son:

Día	10	11	12	13	14	15	16	17	18
NO	100	107	163	140	97	102	146	109	117
SE	95	70	91	164	110	113	134	103	113

Antes de realizar el análisis de esta información, primero se indica lo que se comprenderá por un diagrama de dispersión.

Diagrama de dispersión

El *diagrama de dispersión* es una técnica apropiada para describir datos *bivariados,* pues permite visualizar la posible relación entre dos variables cuantitativas. Estos diagramas representan una nube de puntos y cada uno es una pareja de los valores de cada variable. Con el fin de tener una evaluación numérica de la relación entre dos variables, en estadística se usa el término correlación, la cual permite medir qué tan sólida es la relación entre esas dos variables.

Día	NO	SE
10	100	95
11	107	70
12	163	91
13	140	164
14	97	110
15	102	113
16	146	134
17	109	103
18	117	113

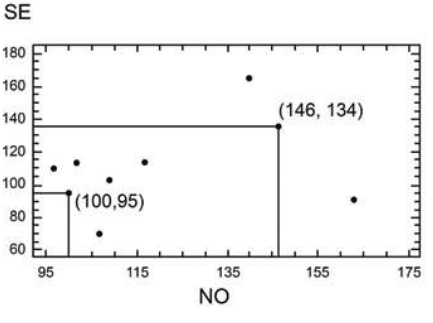


Figura 10.1 Diagrama de dispersión para el registro del ozono durante 9 días del mes de abril de 2005 en dos estaciones de monitoreo.

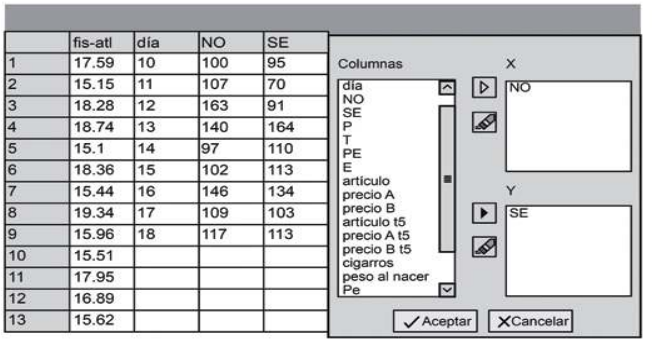
Análisis de la relación de los datos: En la figura 10.1 se muestra una tabla de datos y el diagrama de dispersión correspondiente a la relación del ozono (SO3) entre las estaciones noroeste (NO) y sureste (SE) de una gran ciudad del 10 al 18 de abril de 2005. El reporte indica el máximo de ozono que se registró esos días en cada estación. Por ejemplo, el máximo en esas dos estaciones durante el 10 de abril

está dado por la pareja de observaciones 100 y 95. Ese par de datos se pasa a la gráfica y es el punto (100,95)[1]. Así, cada par de observaciones se representa en un plano cartesiano. Para mayor comprensión de esto se describen los dos datos de las estaciones del día 16 de abril.

A partir de la figura 10.1 se puede interpretar que no existe una estrecha relación entre las variables, es decir, que al aumentar o disminuir el ozono en la estación noroeste, éste no aumenta o disminuye en la estación del sureste. Se puede concluir que no hay una tendencia marcada del nivel de ozono en alguna de la estaciones, ya que como se puede apreciar en la figura 10.1 en algunos días en la estación NO el nivel de ozono es mayor que el nivel de ozono de la estación SE.

Presentación gráfica usando la tecnología

En esta parte se presenta la aplicación del paquete estadístico CalEst con la finalidad de elaborar un diagrama de dispersión. En la opción de gráficas, aparece la alternativa de gráficas de dispersión. Antes de accionarla se necesita tener una hoja de datos en activo, donde se ha capturado la información que se va a graficar, también puede ser una hoja de datos nueva. En la figura 10.2 se ve la hoja de entrada con una serie de variables; en particular, en las columnas de la 2 a la 4 se tiene la información relacionada con el ozono. Una vez que se ha señalado la opción gráfica, aparece en la pantalla un cuadro donde se deben especificar las variables que se usarán para construir el diagrama de dispersión. Luego se da aceptar y se tiene el diagrama de dispersión.



	fis-atl	día	NO	SE
1	17.59	10	100	95
2	15.15	11	107	70
3	18.28	12	163	91
4	18.74	13	140	164
5	15.1	14	97	110
6	18.36	15	102	113
7	15.44	16	146	134
8	19.34	17	109	103
9	15.96	18	117	113
10	15.51			
11	17.95			
12	16.89			
13	15.62			

Figura 10.2 Presentación de la hoja que aparece en la opción datos bivariados para generar el diagrama de dispersión.

Una vez dado aceptar se obtiene el diagrama de dispersión tal y como se describe en la figura 10.3.

Construcción de un diagrama de dispersión

El procedimiento para trazar un diagrama de dispersión descrito por pasos es:

Paso 1. En una hoja o tabla de registro se anotan los valores observados en las dos variables cuantitativas continuas cuya relación será estudiada.

[1] Al punto (100,95) también se le denomina par ordenado, es decir, observación en el eje horizontal, observación en el eje vertical.

Paso 2. Se trazan los ejes horizontal (X) y vertical (Y). En el primer cuadrante de un plano cartesiano, la parte superior derecha de los ejes horizontal y vertical, respectivamente, quedará la figura. Conviene encontrar los valores mínimo y máximo de X y Y, porque ello permitirá marcar los valores de las variables sobre los ejes, pero hay que procurar que los ejes sean de igual magnitud entre las longitudes de los puntos sobre ellos.

Paso 3. Se grafica cada una de las parejas de la tabla de registro en el diagrama de dispersión. Si en la tabla están registradas parejas iguales que hacen que en la gráfica queden en el mismo punto, se traza un círculo sobre el punto para indicar que es un punto repetido. Si es el caso de que más puntos se repitan, entonces se trazan sucesivamente círculos concéntricos.

Paso 4. Poner etiquetas en el diagrama de dispersión para mayor claridad, y por último escribir el nombre de las variables en los ejes.

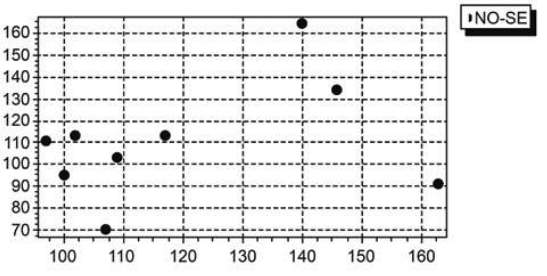


Figura 10.3 Diagrama de dispersión generado por el CalEst.

Ejemplo 10.1

Con el propósito de estudiar la relación entre el peso (en kilogramos) y tiempo (segundos) en una prueba de atletismo de 100 metros, un profesor de secundaria cronometró a 25 estudiantes no entrenados para conocer su rendimiento físico. Además, tomó el peso de cada participante. En este ejemplo, el peso y la velocidad registrados equivalen a dos variables cuantitativas continuas, P: peso, T: tiempo. Los datos registrados son:

P	T	P	T	P	T	P	T	P	T
53	16.97	53	20.46	51	19.81	38	12.05	48	14.77
41	14.88	52	17.93	58	22.41	70	23.49	60	19.2
59	19.35	60	16.74	62	20.31	54	15.47	67	21.76
57	18.93	58	16.27	57	19.07	52	18.57	55	18.32
45	16.56	51	18.48	55	19.79	61	22.98	47	16.34

Solución

En el diagrama de dispersión que se presenta en la figura 10.4 se puede observar una tendencia a incrementarse el tiempo en la medida en que el peso es mayor. Sin embargo, también se da el caso de que algunos estudiantes corren en menos tiempo aunque pesan más.

En la figura 10.4 se observa que los puntos encerrados en la elipse de abajo corresponden a los estudiantes que son rápidos, y los encerrados en la elipse de arriba, a los más lentos. Los puntos que no están encerrados muestran la tendencia de que a mayor peso más lentitud.

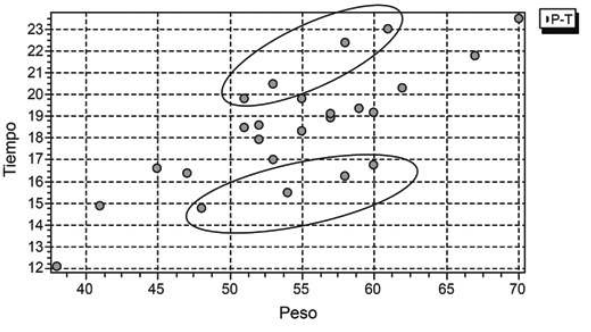


Figura 10.4 Relación entre peso y tiempo registrados en una prueba de 100 metros por 25 estudiantes.

Variable continua

Una variable continua es una variable numérica que puede tomar un número infinito de valores entre dos números.

El mundo de la información 2. Relación entre estatura - peso, y su distribución

Las administraciones escolares han dado impulso a nuevas carreras, entre ellas la licenciatura en nutrición. Un nutriólogo está realizando un estudio para su tesis de licenciatura. Necesita conocer si existe relación entre la estatura y el peso de un grupo de jóvenes.

Preguntas sobre la naturaleza del problema

Es importante identificar si a una mayor estatura se incrementa el peso. ¿Existirán jóvenes que tengan un peso alto siendo bajos en estatura? ¿Qué variable tiene mayor dispersión? Al conocer la distribución de los datos, ¿se puede decir que son simétricos para cada variable? **Datos:** Se toma una muestra al azar

de 45 de ellos. En la figura 10.5 se muestra la descripción de los datos observados.

P	T	P	T	P	T	P	T	P	T
47	143	47	148	54	155	47	141	66	157
46	141	58	155	62	163	51	152	59	154
56	158	60	153	42	133	54	149	52	144
51	149	59	148	61	168	58	155	49	147
57	156	68	166	58	146	48	137	53	151
59	159	63	153	66	162	53	145	60	148
63	162	55	158	52	148	46	144	48	143
54	157	68	172	60	147	55	151	60	154

Análisis de la información proporcionada por los datos. En el cuadrante se observa el diagrama de dispersión para ambas variables figura 10.5. Al eje X se le asignó la estatura y al Y el peso. Después se construyó el diagrama de caja para la estatura, el cual se muestra en la parte superior del eje horizontal. La distribución de los datos referentes a la estatura sigue una distribución más o menos simétrica, de modo que se puede decir que la media y la mediana coinciden. También se dibujó el diagrama de caja para el peso, pero en esta situación la distribución, como en el anterior, está ligeramente sesgada, donde la media es un poco menor que la mediana.

Con el propósito de elaborar los diagramas de caja, en la tabla 10.1, se presenta un resumen estadístico de cada una de las variables.

La información proporcionada por las medianas de estas distribuciones permitirá que se evalúe la existencia de una estrecha relación entre estas variables.

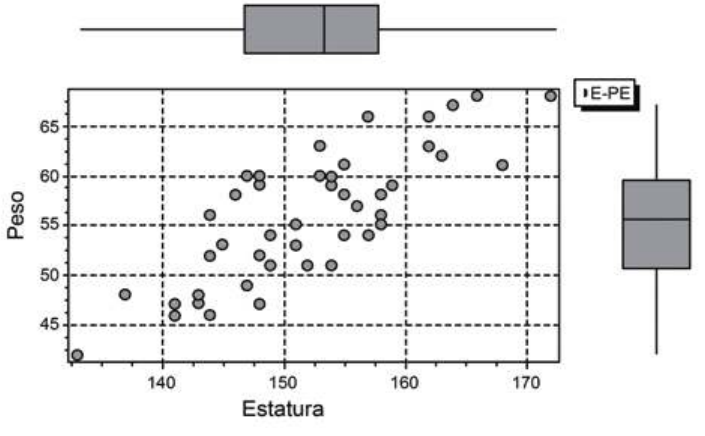


Figura 10.5 Relación entre estatura y peso de 45 jóvenes y la distribución de cada variable.

Tabla 10.1 Resumen estadístico de cada una de las variables.

Media	55.96	Estatura
Mediana	56	152.16
Mínimo	42	133
Máximo	68	172
Primer cuartil	51	147
Segundo cuartil	60	157.16
Desviación S	6.49	8.24

Prueba de correlación. Un objetivo importante en un estudio es observar si existe alguna correlación entre las variables de interés, tal como se mostró en la figura 10.5. La idea de buscar esta relación es tratar de explicar una posible causa-efecto entre las variables. Para averiguar más sobre ello, se puede utilizar la denominada *prueba de las medianas* . Este procedimiento estadístico es sencillo y práctico.

Ejemplo 10.2

Se consideran los datos estudiados en el problema 2 para mostrar la prueba de correlación usando las medianas. A continuación se efectúa un bosquejo intuitivo para observar si hay correlación entre dos variables. Esto se realiza a partir de la construcción de las medianas para la distribución de cada variable. En la figura 10.6 se puede observar que, a partir de las medianas, se ha dividido la nube de puntos en cuatro partes. Estas cuatro divisiones se marcan como i, ii, iii y iv.

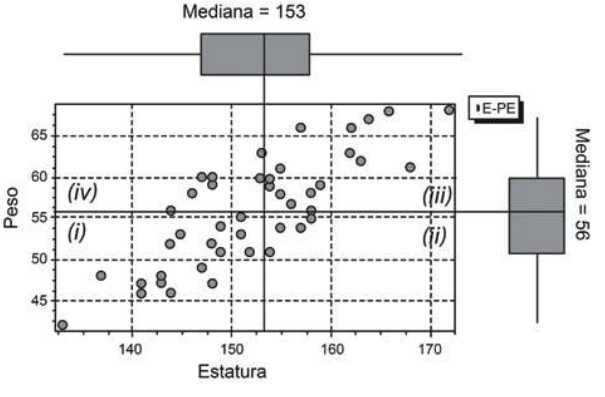


Figura 10.6 Descripción de la prueba de correlación entre la estatura y peso aplicando la razón de las medianas.

En cada cuadrante se pueden contar los puntos. De manera intuitiva, podría pensarse que hay una correlación positiva si la suma de puntos en (ii) y (iv) es menor que la suma de puntos en (i) y (iv). En este caso, se observa claramente que sí existe una correlación positiva entre la estatura y el peso en esta muestra de estudiantes, puesto que:

1. La suma de los puntos en $(i) + (iii) = 17 + 16 = 33$
2. La suma de los puntos en $(ii) + (iv) = 4 + 7 = 11$
3. Finalmente, $(i) + (iii) = 33 > (ii) + (iv) = 11$

El último resultado indica que hay una relación positiva entre las variables de estatura y de peso.

Formalización estadística de la prueba de correlación usando las medianas

La formalización es una descripción del *método de las medianas*. Para hacer una verificación formal sobre la prueba recurrimos a la tabla A1, al final de este capítulo. Esta tabla contiene un límite inferior *(LI)* y otro superior *(LS)*, una N que hace referencia al número de observaciones y un porcentaje de 5 %.

El propósito principal es contar los números que están en cada cuadrante generados por las medias y se comparan con estos límites. Para consolidar esta idea, consideremos el valor de N para el ejemplo, es decir, N=45. En la tabla se observa que el límite inferior es 15 y el superior es 30. Ahora si la suma de $(ii) + (iv) < 15$, decimos que hay una correlación positiva, o de manera análoga, si $(i) + (iii) > 30$. En efecto, se cuentan los puntos y se observa que se cumple este hecho. El procedimiento en forma de algoritmo es como sigue:

1. Encontrar las medianas de los ejes horizontal y vertical, dibujar una línea paralela a la horizontal y otra paralela a la vertical de modo que atraviesen las medianas.
2. Marcar los cuadrantes formados por las medianas; éstos generan cuatro áreas (i, ii, iii, iv), donde el área (i) corresponde al cuadrante derecho superior, y se identifican las áreas restantes siguiendo el sentido contrario a las manecillas del reloj. Luego se cuentan los puntos en cada área.
3. Contar el número de puntos para $(ii), (iv), (i)$ y (iii). Si existen puntos sobre los nuevos ejes, éstos también se cuentan. Supongamos que hay M puntos sobre los ejes, éstos se restan al tamaño de la muestra, es decir *N1=N-M*, este valor *N1* es el nuevo valor de N y es el que se toma como referencia en la tabla A1.
4. Se compara el número total de puntos en (ii) y (iv) con el valor límite inferior, si $(ii)+(iv) <$ límite inferior decimos que existe una correlación positiva. En esa misma dirección, también podemos comparar $(i)+(iii)$ con el límite superior, y obtenemos la misma conclusión (que hay una correlación positiva) si $(i) + (iii) >$ límite superior.
5. Análogamente, para establecer la existencia de la correlación negativa, contamos el número de puntos en las áreas (i y iii) y (ii y iv), respectivamente. Pero ahora la suma de puntos en (ii y iv) debe ser mayor a la suma de puntos (i y iii). Se puede usar como referencia la tabla A1 para llegar a una conclusión.

Coeficiente de correlación

En esta parte se expone el tema del coeficiente que indica la relación entre variables y la inferencia estadística sobre el parámetro que representa al coeficiente de correlación. El *coeficiente de correlación de la muestra*, que se denota con r, se calcula para establecer la relación entre dos variables X y Y en términos de un valor numérico. El valor de este coeficiente r indicará el grado de asociación entre las variables.

Cálculo del coeficiente de correlación r. A continuación se expone el procedimiento operativo para obtener el *valor de* r.

1. Se calcula la suma siguiente en términos de la variable X.

$$S_{xx} = \sum_{i=1}^{n} (X_i - \overline{X})^2 \tag{10.1}$$

2. Se calcula la suma siguiente en términos de la variable Y.

$$S_{yy} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \tag{10.2}$$

3. Se calcula la siguiente expresión:

$$S_{xy} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y}) \tag{10.3}$$

4. Se obtiene el coeficiente de correlación r, mediante la fórmula:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{10.4}$$

Complemento técnico. Las medidas numéricas calculadas a partir de la información de una muestra corresponden a los estadísticos. En ese sentido, r es el *coeficiente de correlación muestral* . El coeficiente de correlación que se refiere a la población se denota por la letra griega ro, esto es, ρ. Así, ρ es el *parámetro* que permite evaluar qué tan estrecha es la relación entre dos variables en una población. El *coeficiente de correlación* r, también se conoce como *coeficiente de correlación de Pearson*, se denota con r y está dado por:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \tag{10.5}$$

El coeficiente de correlación de la población se denota por la letra griega ρ.

Inferencia estadística sobre el parámetro ρ

En el estudio de la correlación entre las variables X y Y se plantea hacer *inferencia estadística*, , en esta situación específica, para el parámetro ρ, véase la figura 10.7. Como hemos visto, este procedimiento se realiza mediante una prueba de hipótesis o mediante intervalo de confianza. Para el primer caso, recordemos el procedimiento propuesto en el capítulo 9, siguiendo la regla estándar propuesta:

1. Plantear las hipótesis, es decir:

$$\begin{aligned} H_0 &: \rho = 0, \\ H_1 &: \rho \neq 0. \end{aligned}$$

2. Proponer el nivel de significancia α para la prueba, α es la probabilidad de rechazar la hipótesis nula si esta es verdadera. Por lo general, para α se consideran los siguientes valores 0.01, 0.05, 0.1, entre otros. Se obtienen los valores de la distribución t en relación con el nivel de significancia, esto es: $t_{ci} = t(gl, \alpha/2)$ y $t_{cd} = t(gl, 1-\alpha/2)$. Donde t tiene una distribución t de *Student* con $n-2$ grados de libertad.

3. En el supuesto de que la hipótesis nula es verdadera, se construye el punto crítico, y ya que en el planteamiento de esta hipótesis se tiene la diferencia, se requieren dos puntos críticos, uno a la izquierda y otro a la derecha es decir:

$$r_{ci} = \rho_0 + t(gl, \alpha/2)S(r), \text{ y } \ r_{cd} = \rho_0 + t(gl, 1-\alpha/2)S(r)$$

   El error estándar $S(r)$ del coeficiente de correlación r se plantea mediante la siguiente expresión:

El error estándar del coeficiente de correlación r	$S(r) = \sqrt{\frac{1-r^2}{n-2}}$

4. A continuación se realiza el estudio midiendo, observando o experimentando las variables de interés, con esa información se calcula el coeficiente de correlación r, lo denotamos por r_m. Se compara este valor con los puntos críticos, y se rechaza la hipótesis nula si:

$$r_m < r_{ci}, \text{ o } r_m > r_{cd}$$

   en caso contrario, es decir $r_{ci} < r_m < r_{cd}$, no se rechaza H_0.

Procedimientos alternativos para realizar la prueba

1. Se estandariza la variable aleatoria para verificar la prueba de hipótesis usando el estadístico t, esto es:

Prueba $t - Student$ para el coeficiente de correlación ρ	$t_m = \frac{r_m - \rho}{S(r)}$	(10.6)

Finalmente, se compara t_m con t_{ci} y t_{cd}. Se rechaza la hipótesis nula si $t_m < t_{ci}$, o $t_m > t_{cd}$. Si t_m está entre los valores críticos no se rechaza H_0.

2. Prueba de hipótesis mediante la comparación de probabilidades, se obtiene el $valor - p = P(r \leq r_m)$, o $valor - p = P(t \leq t_m)$ si $H_1 : \rho < 0$. Tener presente que el $valor - p = P(r \geq r_m)$, o $valor - p = P(t \geq t_m)$ si $H_1 : \rho > 0$. Al comparar este valor con el nivel de significancia, considerar el $valor - p$ para pruebas bilaterales dado que $\alpha/2$.

El intervalo del $(1 - \alpha)\,\%$ de confianza de ρ es:

$$L_i \leq \rho \leq L_f$$

donde $L_i = r + t(gl, \alpha/2)S(r)$, y $L_d = r + t(gl, 1 - \alpha/2)S(r)$, $S(r)$ es el error estándar .

En resumen: a partir de los datos de la muestra se calcula el estadístico r. En esta circunstancia, el estadístico r es una variable aleatoria que tiene una distribución de probabilidad con media $\mu(r)$ y varianza $\sigma^2(r)$.

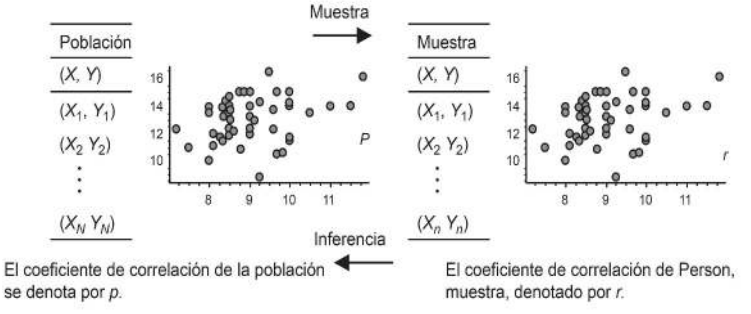


Figura 10.7 Relación entre el parámetro de correlación ρ y el estadístico r.

Ejemplo 10.3

La administración de una compañía con el fin de estudiar la productividad, evalúa la relación entre las variables, el número de horas trabajadas y la producción en toneladas. Si la eficiencia es adecuada, ellos esperarían que existe una relación positiva entre las variables, es decir que a mayor número de horas en el trabajo mayor producción. Verificar mediante una prueba de hipótesis que en efecto se cumple esta relación; use un nivel de significancia de $\alpha = 0.05$. Los datos son:

Observación	1	2	3	4	5	6	7	8
Horas	228	184	205	235	175	194	190	218
Producción	25	16	19	25	15	18	20	22

Solución

Usando la expresión 10.5, se tiene que $r_m = 0.957$.

1. El planteamiento hipotético para verificar si existe relación:

$$H_0 : \rho = 0, H_1 : \rho > 0.$$

2. El nivel de significancia es $\alpha = 0.05$, el valor del estadístico r que se deriva de este nivel corresponde al punto crítico, esto es $r_c = \rho + t(n-2, 1-\alpha)S(r)$, donde $S(r) = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-(0.957)^2}{8-2}} = 0.118$ y $t_c = t(n-2, 1-\alpha)$. Entonces $r_c = 0 + t(6, 0.95)S(r) = 1.945(0.165) = 0.230$
3. Se rechaza la hipótesis nula si $r_m > r_c$. Por lo tanto se rechaza la hipótesis nula.
4. En efecto existe una relación positiva entre las variables.

Si se usa el valor del estadístico $t-Student$, se obtiene el procedimiento alternativo en forma estándar. Así, se tiene que $t_m > t_c$, se rechaza la hiótesis nula. Donde $t_m = \frac{r_m - \rho}{S(r)} = \frac{0.957}{0.118} = 8.110$, consulte la expresión 10.6.

Hay que tener presente que la notación r_m y t_m es para enfatizar que son los valores que se obtienen a partir de la muestra para el estudio en curso.

Por otro lado, los puntos críticos para verificar la hipótesis $H_0 : \rho = 0$, $H_1 : \rho \neq 0$, son: $r_{c_i} = \rho + t(n-2, \alpha/2)S(r)$ y $r_{cd} = \rho + t(n-2, (1-\alpha/2))S(r)$. Los puntos críticos en forma estándar son: $t_{ci} = t(n-2, \alpha/2)$, $t_{cd} = t(n-2, (1-\alpha/2))$

El mundo de la información 3. Sistemas de lectura

En la actualidad existen sistemas para aumentar la velocidad de lectura de los individuos. La administración de una empresa en la que sus empleados tienen que estar leyendo constantemente reportes, contrató a una empresa que se dedica a vender uno de estos sistemas. Se seleccionó a uno de los trabajadores al azar y durante 8 semanas se observó el número de palabras que puede leer en un minuto.

Preguntas sobre la naturaleza del problema

¿Cómo se podría saber si realmente este sistema es eficiente? Al pasar cada semana ¿habrá un mayor número de palabras?, ¿en qué semana el salto de número de palabras que leyó el trabajador fue mayor?

Para responder algunas de estas preguntas, primero es necesario conocer el coeficiente de correlación. Si aumenta la velocidad de lectura del trabajador al transcurrir las semanas, se tiene un coeficiente de correlación y r debe ser positivo.

Datos: Se llevó a cabo el programa y cada semana se realizó una evaluación para determinar el número de palabras que lee por minuto el trabajador. Los datos se presentan a continuación:

Semanas	2	3	4	6	7	8
Velocidad	49	86	109	165	173	192

Análisis de la información proporcionada por los datos. Tenga en cuenta que para obtener el valor del coeficiente de correlación r, es necesario aplicar la expresión 10.5, para ello es necesario calcular las expresiones que aparecen en el numerador y denominador, es decir S_{xy}, S_x y S_y. Para facilitar el cálculo de estos tres valores se desglosan las expresiones indicadas en los pasos 1 a 3 en el procedimiento que ilustra el cálculo de r. En la tabla 10.2 se ilustra la parte operativa del procedimiento para calcular r.

Tabla 10.2 Operaciones para el cálculo del coeficiente de correlación.

Datos en x	$(X_i-\overline{X})$	$(X_i-\overline{X})^2$	**Datos en** Y	$(Y_i-\overline{Y})$	$(Y_i-\overline{Y})^2$	$(X_i-\overline{X})(Y_i-\overline{Y})$
2	-3	9	49	-80	6400	240
3	-2	4	86	-43	1849	86
4	-1	1	109	-20	400	20
6	1	1	165	36	1296	36
7	2	4	173	44	1936	88
8	3	9	192	63	3969	189
30		28	774		15850	659

Sustitución de valores en la fórmula para obtener r.

$$\begin{aligned} S_x &= 28, \quad S_y = 15\ 850, \quad S_xS_y = 659 \\ r &= \frac{659}{\sqrt{28}\sqrt{15\ 850}} = \frac{659}{666.246} = 0.989 \end{aligned}$$

Si el valor de r es cercano a 1, entonces se dice que existe una fuerte relación positiva entre las variables X y Y. Entonces, se concluye que existe una relación positiva fuerte entre el número de semanas y la velocidad de lectura. De modo que se interpreta diciendo que el método de lectura rápida resulta efectivo al pasar las semanas. Finalmente, se completa esta noción observando el diagrama de dispersión, figura 10.8. Ahí se observa que aumenta la velocidad de lectura con el paso de las semanas.

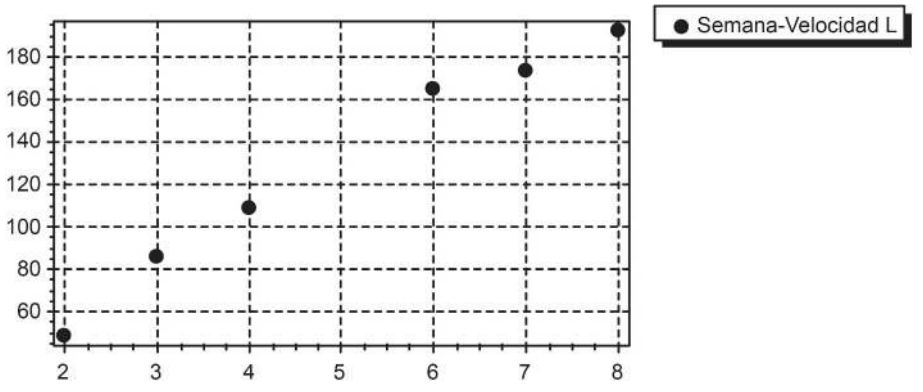


Figura 10.8 Se observa la correlación entre el número de semanas y la velocidad de lectura.

Ejemplo 10.4

Se realiza un experimento para estudiar la relación entre la dosis de un estimulante y el tiempo que un individuo tarda en reaccionar a él. Los datos registrados son:

Dosis (miligramos)	X	1	3	4	7	9	12	13	14
Tiempo de reacción (segundos)	Y	3.5	2.4	2.1	1.3	1.2	2.2	2.6	4.2

Calcular el coeficiente de correlación y presentar un diagrama de dispersión.

Solución

La finalidad del ejemplo es presentar un procedimiento que simplifique la parte operativa en el cálculo del coeficiente de correlación. Para alcanzar esta meta se proponen las siguientes expresiones alternativas para S_x, S_y y S_{xy}, considere las expresiones 10.1, 10.2 y10.3:

$$S_{xx} = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}, S_{yy} = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}, S_{xy} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

Con los cálculos que se muestran en la tabla 10.3 y usando estas expresiones, se obtiene el coeficiente de correlación r. Por lo tanto:

$$\begin{aligned} S_{xx} &= 665 - \frac{(63)^2}{8} = 168.875, \quad S_{yy} = 54.75 - \frac{(19.5)^2}{8} = 7.218, \\ S_{xy} &= 158 - \frac{(63)(19.5)}{8} = 4.438 \end{aligned}$$

El coeficiente de correlación es:

$$r = \frac{4.438}{\sqrt{168.875}\sqrt{7.218}} = 0.127$$

Este valor indica que no existe una correlación lineal entre las variables X: dosis de un estimulante y Y: tiempo de reacción. En la figura 10.9 se ve que el tiempo de reacción disminuye desde la primera dosis hasta una dosis de 9 miligramos; luego crece en la medida en que la dosis aumenta. En resumen, la relación entre estas dos variables tiende a ser en forma cuadrática.

Tabla 10.3. Cálculos alternativos para r.

I	X	Y	X^2	Y^2	XY
1	1	3.5	1	12.25	3.5
2	3	2.4	9	5.76	7.2
3	4	2.1	16	4.41	8.4
4	7	1.3	49	1.69	9.1
5	9	1.2	81	1.44	10.8
6	12	2.2	144	4.84	26.4
7	13	2.6	169	6.76	33.8
8	14	4.2	196	17.64	58.8
Total	63	19.5	665	54.75	158
	$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} Y_i$	$\sum_{i=1}^{n} X_1^2$	$\sum_{i=1}^{n} Y_1^2$	$\sum_{i=1}^{n} X_i Y_i$

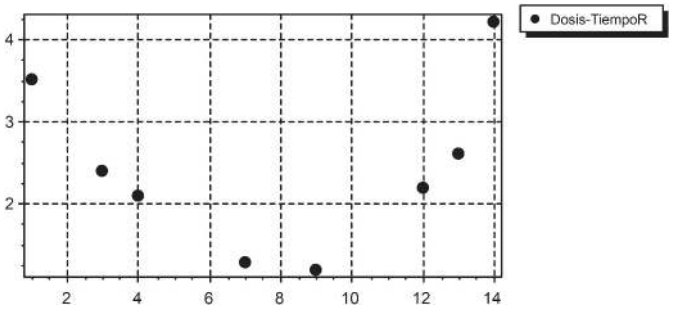


Figura 10.9 Diagrama de dispersión que describe la correlación entre las variables dosis y tiempo de reacción.

Ejemplo 10.5

Un instituto encargado de la protección al consumidor tiene información sobre los precios mínimos y máximos de derivados de aceites y grasas comestibles, los cuales se tomaron de una muestra de diferentes tiendas[2] que existen en una ciudad. En la tabla 10.4 se presentan los precios: los de A son los precios mínimos de cada uno de los artículos y se colocaron en el eje horizontal; los de B son los precios máximos de esos mismos artículos y se colocaron en el eje vertical. ¿Cómo se relacionan los precios de A con los de B?

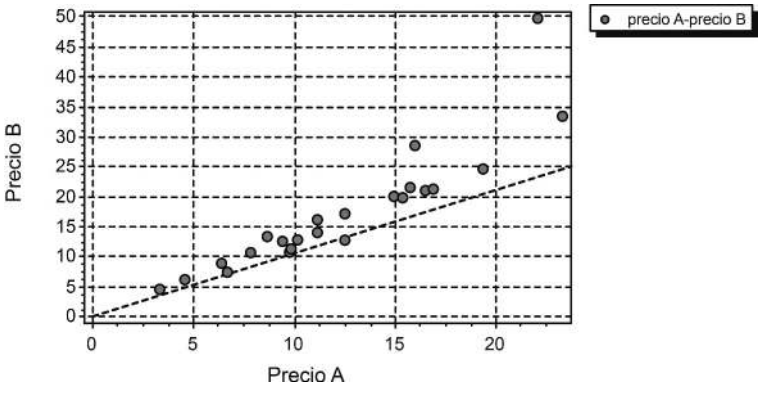


Figura 10.10 Indica la relación de precios para los derivados de aceites y grasas comestibles

Solución

En el diagrama de dispersión que aparece en la figura 10.10 se observa una correlación entre ambos precios. Mientras en unas tiendas es más barato un artículo, en otras su precio es más caro. En la práctica, esto nos indica que siempre es importante saber dónde es más barato comprar para ahorrar dinero. En ese sentido, el resumen de cálculos para obtener el coeficiente de correlación muestral es:

$$\sum_{i=1}^{23} X_i = 284.46, \sum_{i=1}^{23} Y_i = 395.71, \sum_{i=1}^{n} X_i^2 = 4132.056,$$

$$\sum_{i=1}^{23} Y_i^2 = 9043.935, \sum_{i=1}^{23} X_i Y_i = 5958.296$$

[2]En este ejemplo no es de interés identificar esas tiendas.

Por lo tanto:

$$S_{xx} = 4132.056 - \frac{(284.46)^2}{23} = 613.9, \qquad S_{yy} = 9043.935 - \frac{(395.71)^2}{23} = 2235.830,$$
$$S_{xy} = 5958.296 - \frac{(284.46)(395.71)}{23} = 1064.223$$

Tabla 10.4 Descripción entre los precios para diferentes artículos derivados de aceites y grasas.

Artículo	Precio A	Precio B	Artículo	Precio A	Precio B
1	22.13	49.6	13	10.20	12.6
2	11.13	16	14	15.73	21.44
3	6.70	7.29	15	16.90	21.3
4	12.48	12.72	16	15.35	19.8
5	16.00	28.5	17	9.45	12.44
6	9.83	10.5	18	23.3	33.4
7	12.5	17.14	19	6.44	8.80
8	11.16	14	20	3.30	4.3
9	4.6	6.05	21	7.85	10.47
10	8.7	13.1	22	16.55	21.03
11	9.90	11	23	14.91	19.78
12	19.35	24.45			

El coeficiente de correlación es:

$$r = \frac{1064.223}{\sqrt{613.9}\sqrt{2235.830}} = 0.908$$

Así, $r = 0.908$. La línea que se trazó a 45 grados del origen del cuadrante de precios, permite observar de manera clara que el mismo artículo es más caro en las tiendas que tienen el precio B. Aparte de esta interpretación, es conveniente resaltar que el punto 1 tiene la pareja de valores (22.9, 49.6), en éste se observa que la diferencia, más grande entre los precios A y B. Cuando un punto aparece, como en esta ocasión, se llama discrepante o aberrante, y es frecuente reconocerlo por la palabra outlier en inglés.

Ejemplo 10.6

Continuando con el ejemplo anterior, ahora se toman los precios de los 23 artículos derivados de aceites y grasas en dos tiendas de autoservicio, y se anota el precio al que los vende cada una de estas tiendas. En la tabla 10.5 se muestra la información recabada:

Tabla 10.5 Descripción entre los precios en la tienda A y la tienda B.

Artículo	Tienda A	Tienda B	Artículo	Tienda A	Tienda B
1	24.13	27.6	13	10.2	12.6
2	12.13	15	14	17.73	19.44
3	7.2	6.79	15	17.9	20.3
4	12.68	12.52	16	19.35	5.8
5	19	25.5	17	9.45	12.44
6	10.23	10.1	18	23.3	33.4
7	15.5	14.14	19	6.94	8.3
8	13.16	12	20	3.8	3.8
9	5.6	5.05	21	9.35	8.97
10	11.7	10.1	22	18.55	19.03
11	10.9	10	23	18.91	15.78
12	21.35	22.45			

Solución

En la figura 10.11 se muestra el diagrama de dispersión para ilustrar la relación entre los precios en ambas tiendas.

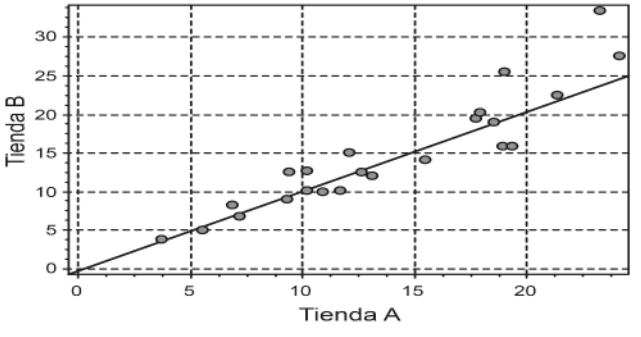


Figura 10.11 Relación de precios para los derivados de aceites y grasas comestibles en las tiendas A y B.

Como se observa, prácticamente las tiendas tienen diferentes precios, pero mientras que la tienda A vende más caros algunos artículos, otros están a un precio menor. En resumen, los precios llegan a compensarse. El coeficiente de correlación es $r = 0.924$, lo cual indica que hay una fuerte relación positiva entre los precios. Los cálculos para obtener r son:

$$\sum_{i=1}^{23} X_i = 319.06, \sum_{i=1}^{23} Y_i = 341.113, \sum_{i=1}^{23} X_i^2 = 5161.122, \sum_{i=1}^{23} Y_i = 6260.987, \sum_{i=1}^{23} X_i Y_i = 5600.637.$$

Por lo tanto:

$$S_{xx} = 5161.122 - \frac{(319.06)^2}{23} = 735.066, \quad S_{yy} = 6260.987 - \frac{(341.113)^2}{23} = 1201.94,$$
$$S_{xy} = 5600.637 - \frac{(319.06)(341.113)}{23} = 868.658$$

El coeficiente de correlación es:

$$r = \frac{868.658}{\left(\sqrt{735.066}\right)\left(\sqrt{1201.94}\right)} = 0.924$$

Cabe destacar que en el diagrama de la figura 10.11 hay tres artículos que en la tienda B son un poco más caros que en la tienda A, cuando el precio rebasa los 20 pesos. ¿Esta tendencia se da en otros artículos? De ser así, la tienda B resultaría más cara cuando un artículo cueste más de 20 pesos.

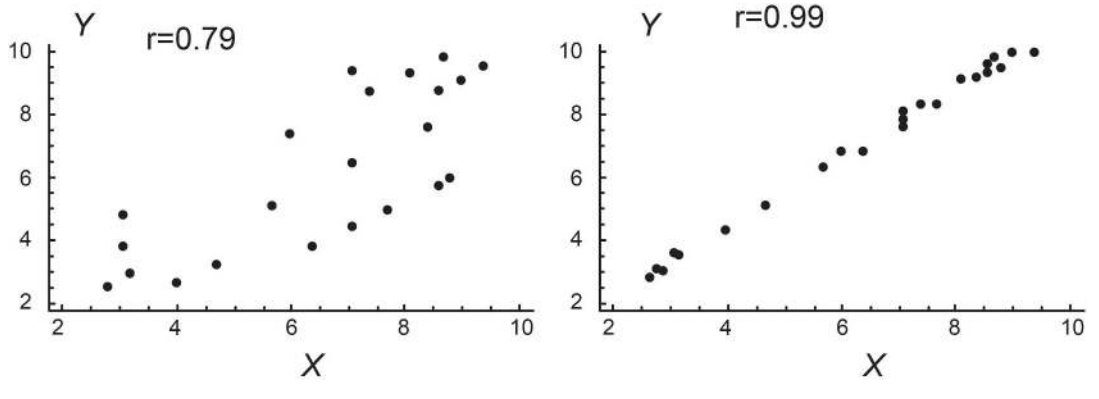


Figura 10.12 Par de diagramas de dispersión que muestran una correlación positiva, muy alta en la segunda gráfica.

Relación positiva o negativa entre variables. Existen otras situaciones que podrían caracterizar a los diagramas de dispersión y su relación con los coeficientes de correlación muestral r. En la figura 10.12 se ilustran dos diagramas de dispersión con un coeficiente de correlación $r = 0.79$ y otro con $r = 0.99$. En la figura 10.13 aparece otro par de gráficas con correlación negativa, con $r = -0.13$ y $r = -0.99$, respectivamente. Como se observa hay diferencias en la *tendencia de asociación* entre las variables X y Y.

Características importantes del coeficiente de correlación

1. El valor de r está entre -1 y 1, es decir, satisface la relación: $-1 \leq r \leq 1$.
2. La magnitud de r indica qué tan estrecha es la relación entre las variables y el signo indica la dirección. Entonces, si:
   - $r > 0$ si la nube de puntos de los valores (X, Y) es una banda que va desde la parte inferior izquierda a la parte superior derecha.
   - $r < 0$ si la nube de puntos de los valores (X, Y) es una banda que va desde la parte superior izquierda a la parte inferior derecha.
   - $r = 1$ si la nube de puntos de los valores (X, Y) es una línea recta, y se dice que es una relación positiva perfecta.
   - $r = -1$ si la nube de puntos de los valores (X, Y) es una línea recta, y se dice que es una relación negativa perfecta.
3. Un valor r cercano a cero indica que la relación lineal es muy débil.

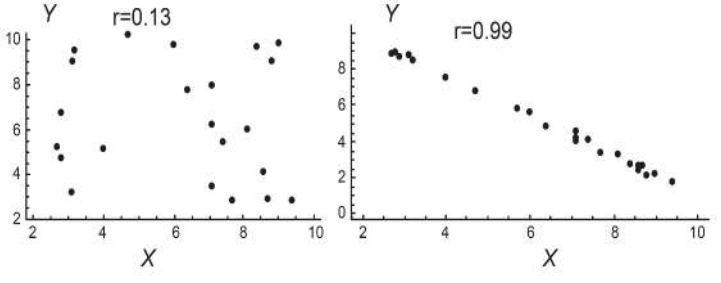


Figura 10.13 Este par presenta dos casos de correlación negativa, donde se destaca que el primero es cercano a cero.

Diagramas para describir la relación entre otro tipo de variables

Representación gráfica entre variables cualitativa y cuantitativa: En el capítulo 1 se clasificaron las variables por tipo, ya sea cualitativo o cuantitativo. Con la finalidad de completar el análisis de la información entre este tipo de variables, en este pequeño apartado se presenta su descripción gráfica.

El mundo de la información 4. Promedio final de licenciatura

Al finalizar sus estudios en la carrera de ingeniería industrial, en un tecnológico regional otorgan un reconocimiento a los tres mejores promedios. En la figura 10.14 se muestra la descripción de esos promedios considerando el periodo que abarca de 1996 a 2004.

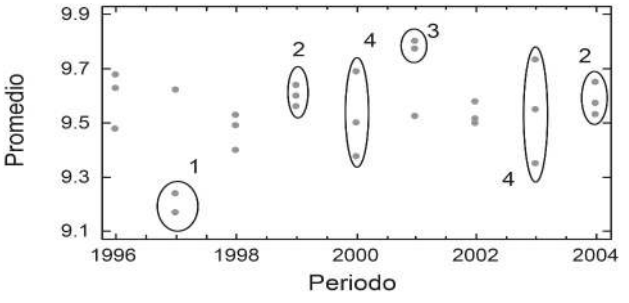


Figura 10.14 Descripción entre las variables periodo y promedio al finalizar la carrera de ingeniería industrial.

Preguntas sobre la naturaleza del problema

¿Los mejores promedios en las calificaciones se mantienen año con año? ¿Existe una homogeneidad entre los promedios de calificaciones dentro de cada periodo y entre periodo? ¿Se nota una diferencia por género en los mejores promedios?

Datos: Los datos obtenidos para cada periodo y por género se muestran a continuación:

P	G	C	P	G	C	P	G	C
1996	m	9.48	1999	m	9.6	2002	h	9.58
1996	m	9.68	1999	h	9.64	2002	m	9.51
1996	h	9.63	1999	m	9.56	2002	m	9.5
1997	m	9.62	2000	m	9.5	2003	h	9.55
1997	h	9.24	2000	m	9.38	2003	m	9.35
1997	m	9.17	2000	h	9.69	2003	m	9.73
1998	m	9.53	2001	h	9.77	2004	m	9.57
1998	m	9.4	2001	h	9.8	2004	m	9.53
1998	m	9.49	2001	m	9.52	2004	m	9.65
P: periodo, G: Género, C: calificación promedio								

Análisis de la información proporcionada por los datos: En este evento, la *variable cuantitativa* es el promedio en las calificaciones y la *variable cualitativa* es el periodo. La relación entre estas dos

variables se muestra en el diagrama de puntos de la figura 10.15. En ésta se han encerrado varios grupos de puntos a fin de visualizar el desempeño académico de esas generaciones.

En el círculo marcado con el número 1 se observa que en el año 1997 se obtuvieron los promedios más bajos del periodo. El número 3 señala que en el año 2001 se alcanzaron los promedios más altos. Los círculos con el número 2 muestran que los promedios están concentrados (muy pegados uno del otro) y de manera contraria en 2000 y 2003, mientras que en los marcados con el número 4 existe la mayor dispersión de los promedios.

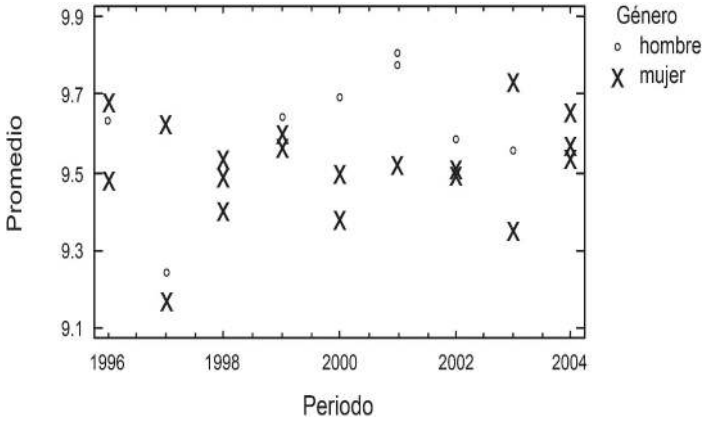


Figura 10.15 Se incorpora la variable género para obtener mayor información del estudio.

Variable cuantitativa y cualitativa

Para explorar la relación entre una variable cuantitativa y una cualitativa, los valores de la variable cuantitativa se grafican para cada grupo o categoría de la otra variable cualitativa usando la misma escala.

Una variable más: Aprovechando las facilidades de las técnicas modernas de graficación, es posible incorporar una nueva variable a la relación que ya se estableció. La variable que se integra es de tipo cualitativo. En nuestro caso será la variable género. Esta nueva variable se identificará en el diagrama de dispersión, donde se pondrán x para reconocer a las mujeres y círculos pequeños para señalar a los hombres.

En esta situación se tendrán tres variables y sus valores quedan descritos en la figura 10.15. A partir de ésta puede concluirse que las mujeres destacan entre los mejores promedios, ya que obtuvieron el primer lugar en más ocasiones y en la mayoría de las generaciones aparecen más.

El mundo de la información 5. Promedio generacional de una preparatoria

En este problema se estudia la relación de variables y distribución de una de ellas. Par ello se integrarán conceptos vistos en el capítulo 2, como lo es la distribución, además de conceptos del capítulo 3, como lo son el resumen estadístico y el diagrama de caja. Conocer cómo se desempeña una generación en un ciclo escolar es importante para la escuela porque le permite efectuar ajustes cuando los considere importantes. Es común usar un promedio de la calificación para estudiar el rendimiento escolar, lo que representaría la información. ¿Qué haría la dirección de una escuela para alcanzar el conocimiento del desempeño escolar de una generación? ¿Podría la dirección comparar el rendimiento de cada generación?

Preguntas sobre la naturaleza del problema

Como se sabe, el número de estudiantes por generación en una escuela preparatoria es alto. En el capítulo 2 se vio cómo estudiar la distribución para este tipo de datos, pero si se tienen varios conjuntos, por ejemplo diferentes generaciones, surge la pregunta: ¿cómo representar para diferentes generaciones la distribución? Una vez recabada la información, ¿cómo estudiar el desempeño escolar para cada generación?, ¿qué características se muestran en la distribución de los datos, promedio de calificaciones, en cada generación?

Datos: En esta ocasión los datos son el promedio de calificaciones que tuvieron los estudiantes en cada generación. Como sabemos, este volumen de información es muy alto, no se puede reproducir. Estos datos vendrán representados por el diagrama de caja como se muestra en el análisis de datos de este problema. **Análisis de la información proporcionada por los datos:** Mediante el diagrama de caja, la distribución del promedio de calificaciones y su relación con las generaciones (variable cualitativa), podemos visualizar la relación que existe en diferentes generaciones de una escuela de bachillerato. En la figura 10.16 se puede ver cuál fue la media y la mediana, así como la dispersión que hubo en cada generación. Así, por ejemplo, se observa que las generaciones 99 y 01 tuvieron en general menor desempeño que las otras.

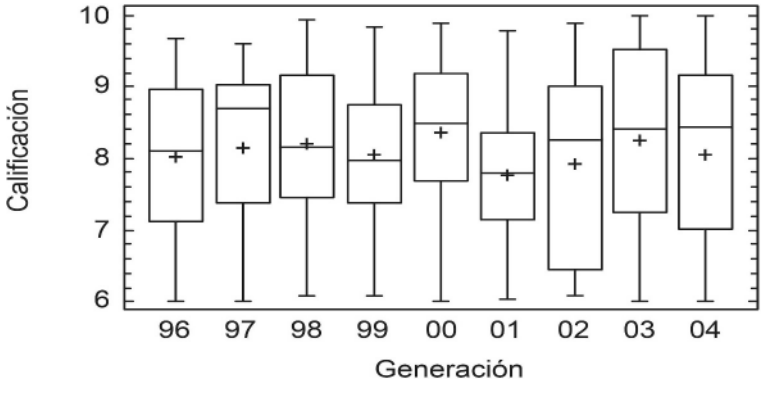


Figura 10.16 Relación entre las variables generación y calificaciones. Se destaca la distribución para cada generación. El signo + indica el promedio de calificaciones.

Aplicación a los datos de la contaminación de la Ciudad de México

Un ejemplo real sobre el análisis descriptivo de la información puede encontrarse en la siguiente dirección

www.calidadaire.df.gob.mx/calidadaire/productos/infocalidadaire/imecaanterior.php

En esa publicación se presenta un reporte estadístico interesante sobre análisis descriptivo de la información de la contaminación de la Ciudad de México. Son de particular importancia las páginas 22, 24, 28, 30, pues en ellas se presentan estudios donde se aplicaron técnicas como la mostrada en la figura 10.16.

El mundo de la información 6. Registro de datos olímpicos

Antecedentes: Los Juegos Olímpicos en una nueva etapa iniciaron en el año de 1896 y continúan hasta la fecha. Desde 1900, se cuenta con el registro del tiempo en segundos de los atletas que llegaron en primer lugar en la carrera varonil de los 1500 metros planos. Un reto en cada uno de los Juegos Olímpicos es mejorar las marcas anteriores. Para alcanzar esta meta los atletas siguen diferentes métodos de entrenamiento y estrategias de competencia. ¿Han mejorado las marcas desde la primera olimpiada a la fecha?

Preguntas sobre la naturaleza del problema

En esta circunstancia no se podrá dar una respuesta a preguntas que se puedan plantear porque no se conoce con detalle muchos problemas, pero se pueden citar algunas: ¿qué tan efectivo resultó el programa de entrenamiento seguido para mejorar una marca?, ¿afecta el tipo de zapato usado? ¿esa evolución se da por igual en la rama varonil y femenil?

Datos: La Tabla 10.6 muestra los datos que se han medido en esa competencia durante los Juegos Olímpicos de 1900 a 2012.

Tabla 10.6 Datos registrados en la especialidad de los 1500 metros en Juegos Olímpicos.

Año	1900	1904	1908	1912	1920	1924	1928	1932	1936
Tiempo	246.0	245.4	243.4	236.8	241.9	233.6	233.2	231.2	22.7
Año	1948	1952	1956	1960	1964	1968	1972	1976	1980
Tiempo	229.8	225.2	221.2	215.6	218.1	214.9	216.3	219.2	218.4
Año	1984	1988	1992	1996	2000	2004	2008	2012	
Tiempo	212.5	216.0	220.1	215.8	212.1	214.2	213.1	214.1	

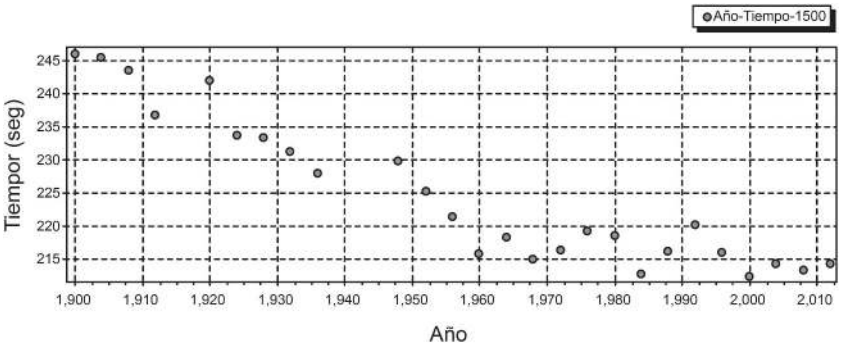


Figura 10.17 Descripción de cómo el tiempo en una competencia de los 1500 metros en Juegos Olímpicos ha mostrado una tendencia a la baja.

Análisis de la información proporcionada por los datos: Para saber cuál ha sido el avance de los tiempos durante los juegos se recomienda trazar un diagrama de dispersión (puntos) y describirlos mediante la aplicación de la opción diagrama de dispersión que viene en el paquete estadístico en el módulo de Estadística. En la figura 10.17 se describe la tendencia de la relación entre el transcurso de los años y el tiempo realizado por los participantes y, como se puede observar, se han ido mejorando los tiempos. Será interesante que expresen sus opiniones ante estos resultados olímpicos.

Relación entre dos variables cualitativas

Antecedentes: En el capítulo 2 se vió el caso de una variable cualitativa, se construyó la tabla de frecuencia para contar el número de observaciones que caían en cada nivel del grupo o categoría de la variable, y se describieron los resultados en un diagrama de barras. En el caso que se desee estudiar la relación entre dos variables de tipo categórico, se cuenta el número de observaciones que caen en cada combinación de los niveles de las dos variables. Este procedimiento permite construir lo que se conoce como tablas de contingencia. Se usa la palabra contingencia porque existe una relación entre dos variables, pues se puede decir que los valores que toma una variable son contingentes (dependen) de los valores de la otra variable. Para ilustrar la relación de este tipo de variables recurrimos a un ejemplo.

El mundo de la información 7. Costumbre de leer el horóscopo

Muchos individuos creen que los astros pueden influir en el destino de los seres humanos, por ello muchas personas acostumbran consultar su horóscopo. El interés es conocer si las personas acostumbran a leer el horóscopo, y además saber con qué frecuencia lo leen. Este tipo de problemas caen en el campo que se conoce como estudios de opinión.

Preguntas sobre la naturaleza del problema

¿Las personas se desempeñan en función de lo que indica su horóscopo? ¿Realmente las personas creen en la influencia de los astros?

Datos: Para averiguar si existe una relación del género (hombre o mujer) con el hábito de leer el horóscopo, se entrevistó a 246 personas y se les preguntó: "¿cada cuándo acostumbra usted leer su horóscopo?" En la tabla siguiente se muestran los valores de la variable "Costumbre de leer el horóscopo", que tiene cinco categorías, a saber: 1) Diariamente; 2) Regularmente; 3) Ocasionalmente; 4) Nunca; 5) No contesta. La otra variable, "Género, se divide en Mujer y Hombre.

Costumbre de leer el horóscopo.						
Género	1	2	3	4	5	Total
Mujer	6	40	72	36	6	160
Hombre	4	20	31	18	13	86
Total	10	60	103	54	19	246

Tabla de contingencia

Una tabla de contingencia es una tabla con renglones que representan los posibles valores de una variable y las columnas representan los posibles valores para la segunda variable. Las casillas (celdas) en la tabla representan el número de veces que ocurre cada par de valores.

Análisis de la información proporcionada por los datos: Un primer paso para descubrir la relación entre las variables "Costumbre de leer el horóscopo" y "Género" es cruzar la tabla para estas dos variables (número de casillas) y construir la tabla contando las veces que cada par de valores ocurre en cada casilla.

Los totales de cada columna aparecen en la parte inferior de la tabla. Así, el número de personas que leen regularmente el horóscopo son 60. De manera análoga, se obtiene el total por renglón. Por ejemplo, para el primero es 160, lo que equivale al número de mujeres que participaron en la encuesta. El gran total equivale a sumar el renglón de totales o la columna de totales, que en este caso es 246.

¿Qué puede decirnos la tabla acerca de la relación entre variables? Con un poco de experiencia, a partir de la tabla pueden construirse patrones. Sin embargo, resulta más apropiado calcular y graficar las proporciones para obtener un bosquejo de la relación. Para llevar a cabo este procedimiento es necesario conocer cómo se calculan las proporciones en una tabla de contingencias.

Primero, se calculan las proporciones considerando el gran total, que en el Problema 8 es el total de encuestados. Este equivale a dividir cada casilla entre 246. En la siguiente tabla se muestra la proporción de toda la tabla. Por ejemplo, $\frac{40}{246} = 0.163$ indica que 16.3 % son mujeres que leen de manera regular su horóscopo, comparado con 8.1 % de los hombres que también lo leen de manera regular.

Tabla 10.7 Costumbre de leer el horóscopo.

Género	1	2	3	4	5
Mujer	0.60	0.67	0.70	0.67	0.32
Hombre	0.40	0.33	0.30	0.33	0.68
Total	1.00	1.00	1.00	1.00	1.00

Con la información proporcionada por esta última tabla, se puede obtener la gráfica para la costumbre de leer el horóscopo para cada uno de los sexos. Ésta se muestra en la figura 10.18 y a partir de ella pueden obtenerse varias conclusiones. Por ejemplo, puede decirse que las mujeres están más interesadas en leer sus horóscopos, mientras que los hombres mostraron poco interés por responder a la pregunta.

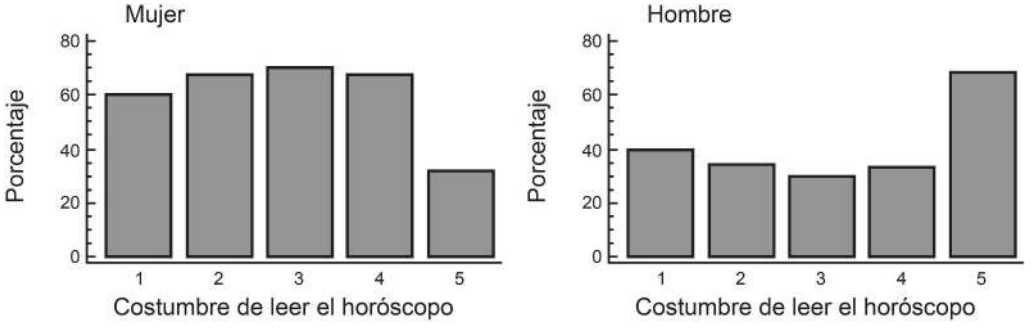


Figura 10.18 Gráfica para describir la relación que se establece entre dos variables cualitativas.

10.3 Modelo de regresión lineal

El mundo de la información

Cuando se estudia la relación entre dos variables existe el interés por predecir el valor de una de ellas a partir del valor de la otra. Con esta relación se pueden construir modelos que permitan aproximarse al conocimiento de los problemas en la realidad. En la figura 10.19, se describe el desempeño del salario por hora en manufacturas en México y China desde el transcurso de los años, 2002 al 2011. Primero, se puede evaluar el desempeño del salario en México al pasar los años, con respecto a las variables X: años y la variable Y: salario. De manera similar, se construye un modelo para China. Un punto de interés a destacar en la relación descrita en la gráfica, es que se puede comparar los sueldos de ambos países. Así una interpretación es que en China crecieron y en México prácticamente quedaron estancados. ¿Cómo construir un modelo a partir de esta relación?

Un ejemplo físico de la vida cotidiana, por ejemplo, predecir la distancia en que frena un automóvil

cuando éste se mueve a diferentes velocidades iníciales. En este contexto, la variable velocidad inicial del automóvil se conoce como variable de entrada, y la variable distancia de frenado es la respuesta. Es común denotar a la variable de entrada con X y la de respuesta con Y.

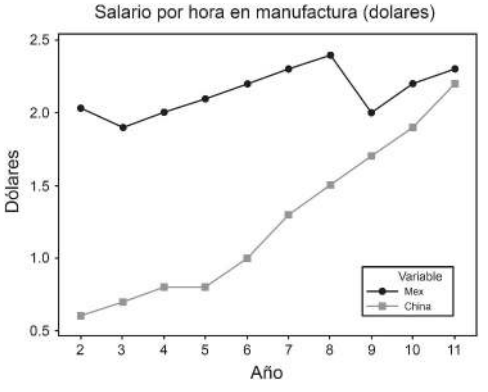


Figura 10.19 Fuente: Secretaria de Hacienda y Crédito Público. Tomada de la Jornada 23 septiembre de 2012.

El mundo de la información 8. Dificultad de audición: Modelo de regresión lineal

La administración de una compañía que fabrica aparatos para la sordera tiene interés en conocer cómo disminuye la capacidad de audición con la edad. Se realiza un estudio para conocer ese efecto. En tal observación, se desea determinar cómo el nivel de audición, al incrementar un nivel de sonido (decibeles) agradable al oído, depende de la edad.

Preguntas sobre la naturaleza del problema

¿Se puede establecer una relación entre las variables edad de una persona y el nivel de sonido? ¿Se puede predecir el nivel de sonido en función de la edad de una persona? ¿Cuál es el nivel del sonido cuando la edad se incrementa en una unidad?

Datos: Los datos que se obtuvieron en una muestra de ocho personas de diferente edad fueron:

Edad (años)	X	15	25	35	45	55	65	75	85
Nivel de sonido	X	56	57	64	64	68	74	78	85
(decibeles)									

En el contexto de este problema la variable edad X es la variable de entrada y el nivel de sonido la respuesta Y. Figura 10.20. A la izquierda el diagrama de dispersión para las variables edad y nivel de sonido y a la derecha se muestra la línea recta que mejor describe a los datos.

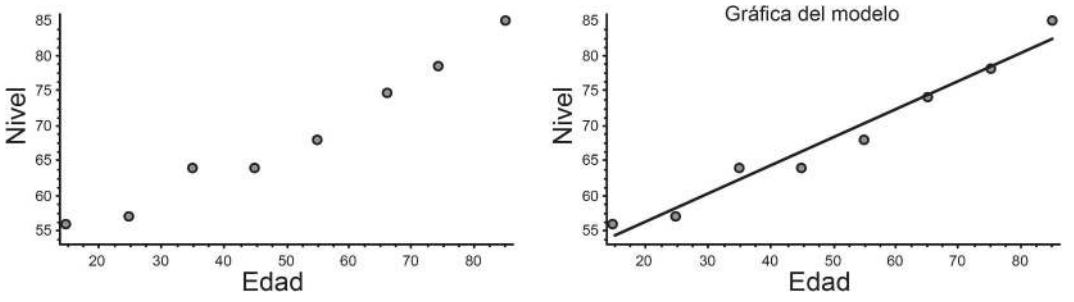


Figura 10.20 A la izquierda el diagrama de dispersión para las variables edad y nivel de sonido y a la derecha se muestra la línea recta que mejor describe a los datos.

Caracterización de la pobación y de la muestra en el modelo

Ahora en el contexto de dos variables (X, Y), X denominadas variable explicativa o de entrada y Y de respuesta. En un primer caso, la relación entre estas variables se puede modelar mediante una recta, referido como *la línea de regresión ,* tambien conocido como *modelo de regresión*. La caracterización estadística, necesita describir la población para el caso de estos datos bivariados. En la tabla 10.8 se describe la población, el modelo que describe la relación entre las variables y se señalan los parámetros, ahí también, se representa la muestra con sus respectivos estimadores. Éstos últimos son variables aleatorias y desempeñan un papel importante para estimar los parámetros del modelo, así como, proporcionar los elementos estadísticos relevantes para hacer inferencia sobre el modelo.

Tabla 10.8. Descripción de la población y muestra en la estimación del modelo de regresión.

	Población		Muestra
(X, Y)	Modelo	(X, Y)	Modelo
(X_1, Y_1)	$Y = \beta_0 + \beta_1 X + \varepsilon$	(X_1, Y_1)	$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$
(X_2, Y_2)	Parámetros	(X_2, Y_2)	Estimadores
$\vdots$	β_0 y β_1	$\vdots$	$\widehat{\beta}_0$ y $\widehat{\beta}_1$
(X_N, Y_N)		(X_n, Y_n)	

Determinación del modelo: Se supone que la respuesta (Y) es una variable aleatoria que se relaciona con la variable (X) por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, i = 1, ..., n \quad \text{donde:} \tag{10.7}$$

1. Y_i denota la i-ésima respuesta correspondiente a un valor de X_i.

2. Para cada valor de la variable X, los valores de la variable Y tienen una distribución de probabilidad normal con media $\mu = \beta_0 + \beta_1 X$, sobre la línea de regresión, con la misma varianza σ^2, homogeneidad en la varianzas, gráfica a la izquierda en la figura 10.21. Nota: este punto establece las condiciones estadísticas para realizar la inferencia estadística.

3. ε_i (epsilon) $\varepsilon_{1,\ldots,}\varepsilon_n$ son variables aleatorias no observables y suponemos que se distribuyen como una distribución de probabilidad normal con media cero y varianza σ^2.

4. Los parámetros β_0(Beta cero) es el punto en el cual la recta (10.7) intercepta a Y. β_1(Beta uno), pendiente de la recta 10.7, indica la cantidad que crece (decrece) la respuesta Y por cada unidad que aumenta la variable X. Figura 10.22.

Descripción e interpretación del modelo:

1. Identifique los valores de los parámetros $\beta_0, \beta_1, \sigma^2$ en el siguiente modelo estadístico:

$$Y = 8 - 6X + \varepsilon$$

   ε es una variable normal con media cero y desviación estándar 4. Solución: $\beta_0 = 8, \beta_1 = -6, \sigma^2 = 16$

2. Del modelo de regresión lineal determine la media y la desviación estándar de Y, para: $X = 1$, cuando $\beta_0 = 2$, $\beta_1 = 4$ y $\sigma = 3$. Solución $Y = 2 + 4X + \varepsilon$, media$(2 + 4X + \varepsilon) = 2 + 4x$, var$(Y)$ =var$(2 + 4X + \varepsilon)$ =var(ε)= 81.

3. Haga la gráfica de la línea de regresión para la media de la línea de regresión lineal.

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \text{con } \beta_0 = 7 \text{ y } \beta_1 = 2.$$

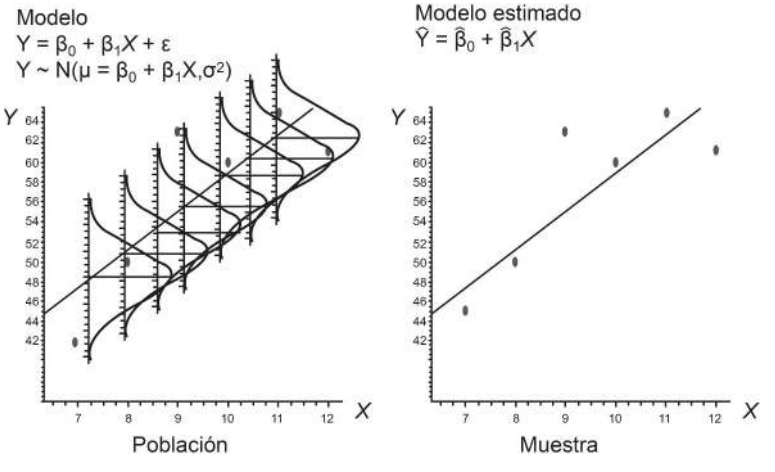
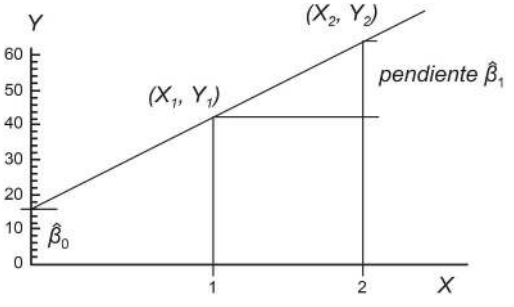


Figura 10.21 Para cada valor de la variable X, los valores de la variable Y tienen una distribución normal con media μ sobre la recta de regresión con la misma varianza σ^2.

Características de la línea. La línea recta que aparece en la gráfica que está a la derecha de la Figura 10.21 considera una aproximación a los puntos que corresponden a los datos bivariados de la muestra. En principio, como los puntos están muy próximos a la recta se sospecha de una posible relación lineal entre las variables. Ante esa realidad, se presenta la característica general de un modelo que describa esa *relación lineal*, denominado *modelo de regresión lineal:* $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X$, donde $\widehat{\beta}_0$ y $\widehat{\beta}_1$ son los estimadores de los parámetros β_0 y β_1. Observe que, si $\widehat{\beta}_1$ se aproxima a cero, entonces, la recta es horizontal, ésto indica que no hay ningún cambio en la respuesta Y, y por lo tanto, se dice que no existe una relación lineal entre las variables X y Y. A continuación, en los siguientes apartados se indica el procedimiento para obtener la ecuación de este modelo, y la inferencia estadística entorno a él.



La línea recta está determinada por dos constantes denotadas por $\hat{\beta}_0$, $\hat{\beta}_1$ como se muestra en la figura. La primera corresponde a la altura que está arriba del origen en el plano cartesiano (primer cuadrante) denominada *intercepto* $\hat{\beta}_0$. La segunda, conocida como *pendiente* $\hat{\beta}_1$, indica el incremento de la variable *Y* cuando la variable *X* aumenta una unidad

Figura 10.22 Línea recta que modela un conjunto de datos e indica el cambio del valor de la variable Y al aumentar X en una unidad.

La técnica de mínimos cuadrados y línea recta $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$

Una vez expuesto el escenario estadístico que caracteriza tanto a la población como a la muestra en el caso de datos bivariados, se observa que en este nuevo planteamiento aparecen los parámetros que describe el modelo, los cuales son: β_0 y β_1. Siguiendo el contexto de *inferencia estadística* que se ha ido construyendo a lo largo del libro, surgen varias preguntas, tales como:

1. ¿Cómo se estiman los parámetros del modelo?

2. ¿Cómo se usa la información de la muestra para estimar los parámetros?

Supóngase tentativamente que el modelo 10.7 es correcto. El problema de la estimación de los parámetros β_0 y β_1 se ve como la línea recta que mejor se ajuste al diagrama de dispersión. El procedimiento estadístico que permite determinar la recta que mejor se ajuste se conoce por el *método de mínimos cuadrados.* Los estimadores de los parámetros β_0, β_1 se denotan por $\widehat{\beta}_0$ y $\widehat{\beta}_1$ respectivamente, vease la figura 10.23.

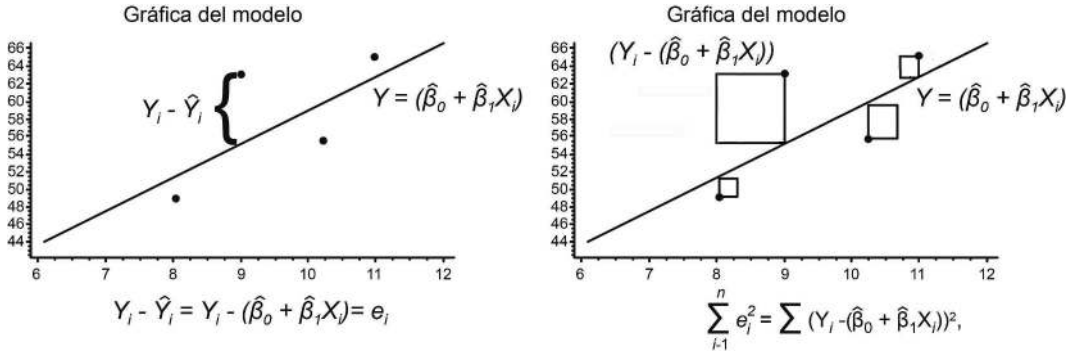


Figura 10.23 Línea recta $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ que modela a un conjunto de datos bivariados y la técnica de mínimos cuadrados.

Mínimos cuadrados

La técnica de mínimos cuadrados encuentra la ecuación de la línea que minimiza la suma de cuadrados de las desviaciones entre los datos y la línea (gráfica a la derecha de la figura 10.23).

Dibujar una línea recta para ajustar un conjunto de puntos, plantea satisfacer algún criterio. Primero, se quiere que la línea que se dibuja sea única. Segundo, se desea que la línea esté lo más cercana posible a todos los puntos. En la figura 10.23 izquierda se ven los puntos y la línea trazada para representar la relación entre variables. La distancia de cada punto a la línea: $Y_i - \widehat{Y}_i$ es una desviación; adelantando la idea que se puntualizará después, ese punto no está explicado por el modelo.

Considere que una línea arbitraria $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ se dibuja en un diagrama de dispersión (figura 10.23). En el valor X_i de la variable independiente, se tiene el valor observado Y_i y el correspondiente a la recta $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ (este valor $\widehat{Y}_i$: se le denomina valor predicho); la discrepancia entre los valores observados y predichos es:

$$Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) = e_i$$

y representa la distancia vertical del punto Y_i a la línea; a los e_i se les llama *residuales*.

La discrepancia de todos los puntos está representada por la siguiente expresión:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i))^2, \tag{10.8}$$

Así, D es una medida de la discrepancia de los puntos observados Y_i de la línea $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$. La

magnitud de D depende de la línea que se dibuje, es decir, dependen de $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Un buen ajuste hace la D tan pequeña como sea posible. El *método de mínimos cuadrados* permite calcular los valores estimados de los parámetros, y su objetivo es minimizar la siguiente discrepancia:

$$D = \sum_{i=1}^{n} (\text{respuesta observada-modelo})^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2. \tag{10.9}$$

Desviación

El valor $\widehat{Y}$ (Y gorro) es el valor que se predice (predicho) de Y para un valor seleccionado de X.
La distancia entre los valores predichos de Y, $\widehat{Y}_i$, y el valor observado de Y es una desviación y se expresa por $Y_i - \widehat{Y}_i$, a esta discrepancia se le llama residual.

Los valores que se obtienen de optimizar la expresión anterior se denominan *estimadores de mínimos cuadrados* de los parámetros del modelo, estos se representan por las cantidades $\widehat{\beta}_0$ y $\widehat{\beta}_1$. Mediante este procedimiento obtenemos el modelo que mejor se ajusta a los datos, el cual se representa por $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$.

La ecuación que representa la relación lineal entre las variables X y Y está dada por la línea recta

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X \tag{10.10}$$

y se le conoce como la *línea de regresión de mínimos cuadrados*, donde la *pendiente* $\widehat{\beta}_1$ se obtiene por la expresión:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \tag{10.11}$$

El intercepto es:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{10.12}$$

A los *estimadores del modelo de regresión* $\widehat{\beta}_0$ y $\widehat{\beta}_1$ también se les conoce como *coeficientes de regresión*

Comentario: El proceso mediante el cual se obtiene la expresión 10.11 es cálculo diferencial, que por lo general se ve en la mayoría de las licenciaturas. Derivar la expresión 10.9 es un buen ejercicio de aplicación de ese tema. Una buena observación de la fórmula que estima el parámetro del modelo de regresión, es decir, la pendiente de la línea, nos permite ver que las expresiones que la componen,

muestran una analogía con la derivación de la media y varianza estudiadas en el capítulo 3. Esto es interesante por lo que allá se consideraba para una variable, aquí esa expresión integra a ambas variables X y Y.

Así, el numerador es la suma del producto de las discrepancias de las variables con respecto a sus medias. El denominador es la suma de la discrepancia, de la variable X con respecto a la media, elevada al cuadrado. La razón de estas cantidades proporciona el estimador clave del parámetro del modelo de regresión. Nuevamente volvemos a hacer énfasis en identificar estos sencillos detalles que permiten generar la parte conceptual de los métodos estadísticos, que a partir de este punto el trabajo se torna simplemente en operativo. No olvide que la aplicación de éste concepto ayuda a resolver problemas planteados en la práctica y ayuda a entender el mundo real.

También es importante advertir que se ha producido un nuevo estimador, lo que abre el camino para aplicar el procedimiento de inferencia estadística, discutido en capítulos previos. Ahora el que entra en acción es el estadístico $\widehat{\beta}_1$. A continuación se preparará el escenario para construir la *metodología de prueba de hipótesis* y la obtención de los *intervalos de confianza* sobre los parámetros $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

Notación: La expresión 10.11 se puede simplificar realizando el producto, en el numerador, y desarrollando el cuadrado en el denominador. Este recurso operativo tiene la finalidad de facilitar los cálculos en la estimación de los parámetros, el análisis y en la evaluación del modelo. En ese sentido se plantean las siguientes expresiones.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \bar{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{10.13}$$

$$S_{xy} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n} X_iY_i - n\overline{XY} \tag{10.14}$$

$$S_{xx} = \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2, \tag{10.15}$$

$$S_{yy} = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \tag{10.16}$$

Las siguientes fórmulas son de utilidad porque permiten calcular los *estimadores de mínimos cuadrados* ; a continuación se muestran las expresiones:

El estimador de mínimos cuadrados de β_0 *es* $\widehat{\beta}_0$: $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1\overline{X}$

El estimador de mínimos cuadrados de β_1 *es* $\widehat{\beta}_1$: $\widehat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

La recta de regresión (estimada) o ajustada es: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$

Se puede observar que las expresiones S_{xx}, S_{xy} son las mismas que aparecen en el cálculo del coeficiente de correlación r. Éstas se pueden aplicar a los cálculos, así $\widehat{\beta}_1$ se rescribe como:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \tag{10.17}$$

La estimación del intercepto es:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{10.18}$$

Observación: una vez estimados por mínimos cuadrados los parámetros del modelo de regresión, queda por conocer cuál es la distribución de los estimadores, dado que es una condición importante para llevar a cabo la inferencia estadística sobre los parámetros, principalmente con respecto a la pendiente β_1 y el modelo, actividad que se realizará en el apartado de inferencia sobre el modelo de regresión.

Respuestas al mundo de la información 8: La línea de regresión de mínimos cuadrados para el problema dificultad de audición se obtiene mediante los cálculos ilustrados en la tabla 10.9

Utilizando los resultados del último renglón de la tabla 10.9 se obtienen los valores de las expresiones para calcular las medias ($\overline{X} = 50$, y $\overline{Y} = 68.25$) y la de las cantidades: S_{xx}, S_{yy} y S_{xy}:

$$S_{xx}= 24200 - 8\,(50)^2 = 4200, \quad S_y= 37986 - 8(68.25)^2= 721.5, \quad S_{xy}= 29010 - 8(50)(68.25) = 1710$$

La pendiente se obtiene por:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1710}{4200} = 0.407$$

El intercepto es:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = \frac{546}{8} - 0.407\frac{400}{8} = 47.9$$

Con los cálculos descritos en la tabla 10.9, también se puede obtener el coeficiente de correlación r:

$$r = \frac{1710}{\sqrt{4200}\sqrt{721.5}} = 0.982$$

Interpretación de los resultados: La línea de regresión de mínimos cuadrados es:

$$\widehat{Y} = 47.9 + 0.407X$$

El valor de la pendiente, $\widehat{\beta}_1$, indica que al aumentar la edad en un año, el nivel de sonido se incrementa en 0.407 decibeles. Si se desea predecir el nivel del sonido a la edad de 50 años, se sustituye $X = 50$ en $\widehat{Y} = 47.9 + 0.407X$ y se obtiene el resultado:

$$\widehat{Y} = 47.9 + 0.407(50) = 68.25$$

Es decir, si se tiene una edad de 50 el nivel del sonido correspondiente es de 68.25 decibeles. Finalmente el coeficiente de correlación resalta la fuerte relación que existe entre las variables X y Y.

Tabla 10.9 Cálculos alternativos para $\widehat{\beta}_1$.

i	X	Y	X^2	Y^2	XY
1	15	56	225	3136	840
2	25	57	625	3249	1425
3	35	64	1225	4096	2240
4	45	64	2025	4096	2880
5	55	68	3025	4624	3740
6	65	74	4225	5476	4810
7	75	78	5625	6087	5850
8	85	85	7225	7225	7225
Total	400	546	24200	37986	29010
	$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} Y_i$	$\sum_{i=1}^{n} X_i^2$	$\sum_{i=1}^{n} Y_i^2$	$\sum_{i=1}^{n} X_i Y_i$

El mundo de la información 9. Aplicación del modelo

Una aplicación interesante es establecer la relación que existe entre datos en el tiempo, la variable independiente X representa el periodo, ya sea en años, meses y otros. Entonces, la idea es desarrollar el modelo de regresión que define la relación entre las variables (X, Y) en el tiempo. El poder adquisitivo es una cuestión importante en el contexto de la economía. Por lo general, una referencia económica es el salario mínimo, éste se establece cada año y es una referencia para determinar los salarios. En este ejemplo se desea conocer si existe una relación lineal del salario mínimo a través de los años, y si es así qué interpretaciones se pueden obtener a partir del modelo de regresión. La información parte del salario mínimo otorgado en México para el área geográfica denominada "B" que va desde los años 2001 al 2012.

Preguntas sobre la naturaleza del problema

¿El precio de la tortilla se mantiene acorde a esa relación? ¿Cuál es el aumento de los productos de la canasta básica? ¿Se mantiene, competitivo el salario de los trabajadores? La varible X indica los años y Y el salario. Nota: con el fin de simplificar los cálculos, se han numerado los años del 1 al 12.

X	1	2	3	4	5	6	7	8	9	10	11	12
Y	37.9	40.0	41.8	43.7	45.3	47.2	49.0	50.9	53.3	55.8	58.1	60.5

Proporcione el modelo de regresión e interprete el parámetro β_1, haga una gráfica de dispersión de estos datos.

Solución

El resumen estadístico para obtener los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de los parámetros, aplicando las expresiones 10.12 y 10.11 es:

$$\overline{X} = 6.5,\ \overline{Y} = 48.66,\ S_{xx} = 143.0,\ S_{xy} \triangleq 287.67.$$

Así, sustituyendo en las fórmulas (10.17) se tiene que:

$$\widehat{\beta}_1 = \frac{287.67}{143} = 2.012, \quad \widehat{\beta}_0 = 48.66 - (2.012)(6.5) = 35.582$$

La gráfica del modelo se describe en la figura 10.24 y éste es:

$$\widehat{Y} = 35.582 + 2.012X$$

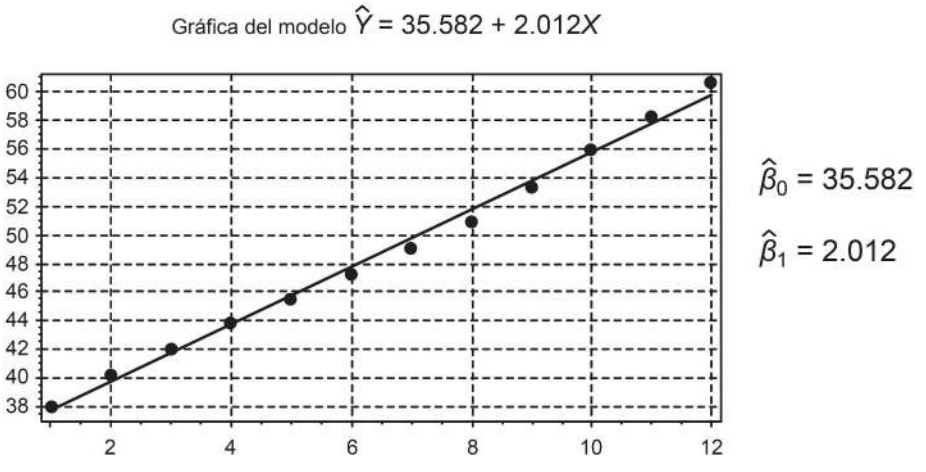


Figura 10.24 Diagrama de dispersión y modelo de regresión para el ejemplo 7.

Esto significa que cada año que pasa, el salario aumenta en ¡sólo dos pesos! Cabrían muchas preguntas, ante esta situación, con el fin de evaluar el poder adquisitivo de los trabajadores.

10.4 Inferencia estadística sobre los parámetros del modelo

A continuación se plantearán las hipótesis relevantes sobre el modelo y la metodología para verificarlas. Con los conceptos y resultados del procedimiento se continuará con la inferencia de los parámetros, la interpretación y evaluación del modelo.

1. Con la información de la muestra ¿se puede decir si los datos dan evidencia para concluir que existe una relación lineal?

2. ¿Cuáles son las hipótesis que se plantean para saber si el modelo explica los datos proporcionados por la muestra?

Planteamiento de la hipótesis sobre el modelo: $Y = \beta_0 + \beta_1 X + \varepsilon$

En un problema de análisis de regresión es de interés determinar cuándo la respuesta varía o no con la magnitud de la variable de entrada X. Considere el modelo de regresión lineal. Respuesta esperada $\mu = E(Y) = \beta_0 + \beta_1 X$. En ésta no hay cambio en X si y sólo si $\beta_1 = 0$. Con objeto de verificar si los datos apoyan o no este hecho, en términos estadísticos se plantea la siguiente hipótesis :

$$\begin{aligned} H_0 &: \text{No existe relación lineal entre } X \text{ y } Y \\ H_1 &: \text{Sí existe relación lineal entre } X \text{ y } Y \end{aligned} \tag{10.19}$$

Planteamiento de la hipótesis sobre la pendiente β_1

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{10.20}$$

Ambas hipótesis son equivalentes; si en la primera se rechaza la hipótesis nula, se concluye que sí existe relación lineal entre X y Y, tal resultado indicaría que $\beta_1 \neq 0$. En el primer caso (10.19) se recurre al análisis de varianza, y por lo tanto se verifica aplicando el estadístico de prueba F, la razón entre varianzas, tal y como se vió en el apartado 10.6. La segunda (10.20) se verifica empleando el estadístico de prueba $t - Student$.

Para probar la hipótesis (10.19), el planteamiento permitirá considerar la existencia de dos fuentes de varianza: una explicada por el modelo VAR 1, y otra no explicada por el modelo VAR 2; recuerde que la razón de dos varianzas sigue una distribución F con gl_{num} y gl_{dem}. A partir de un ejemplo se mostrará la metodología de la prueba de hipótesis mediante el análisis de la varianza.

Observación: conviene indicar que se realiza primero la prueba de hipótesis (10.19) considerando la motivación del análisis de la varianza efectuado en el apartado 10.6, dado que de este análisis se requiere el resultado del cuadrado medio del error (CM_{error}) para realizar la prueba de hipótesis (10.20). Comentamos, también que, en los cursos de estadística usualmente se presenta la segunda, empleando como supuesto el conocimiento del error estándar, derivado del CM_{error}.

Ejemplo 10.7

En la administración de una empresa, como parte estratégica se trabaja para incrementar la productividad de uno de sus procesos, con el objetivo de obtener una mayor utilidad y satisfacer los requisitos de uno de sus clientes. En una parte principal del proceso, el cambio de la temperatura influye en las impurezas, y al controlar éstas se alcanza la meta propuesta. La variable X es el aumento de la temperatura (en grados C) y la variable Y es la productividad, medida en porcentaje. Nota: con el fin de motivar la metodología para la construcción del modelo y su evaluación, en este ejemplo sólo se han tomado 7 valores. Los valores descritos para este estudio son:

Temperatura	X	6	7	8	9	10	11	12
Productividad	Y	44	42	50	63	60	65	61

Solución

Resumen de cálculos para estimar, por mínimos cuadrados, los parámetros del modelo de regresión:

$$\overline{X} = 9,\ \overline{Y} = 55,\ S_{xx} = 28 \text{ y } S_{xy} = 107$$

Aplicando las expresiones (10.11) y (10.12) se tiene que:

$$\begin{aligned} \widehat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{107}{28} = 3.821 \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} = 55 - 3.821(9) = 20.607 \end{aligned}$$

Entonces el modelo:

$$\widehat{Y} = 20.607 - 3.821X$$

Planteamiento de la hipótesis

$$H_0 : \text{No existe relación lineal entre } X \text{ y } Y$$
$$H_1 : \text{Sí existe relación lineal entre } X \text{ y } Y$$

Motivación de la metodología para probar esta hipótesis. Parte operativa En la siguiente tabla se presenta una serie de cálculos para realizar la prueba de hipótesis sobre la relación lineal, conocida por la tabla del análisis de la varianza, en siglas se indica por andeva.

Tabla 10.10 Descripción de los cálculos que explican la fuente de variación.

X	Y	$\hat{Y}$	$Y-\bar{Y}$	$\hat{Y}-\bar{Y}$	$Y-\hat{Y}$	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	$(Y-\hat{Y})^2$
6	44	43.54	-11	-11.46	0.46	121	131.423	0.212
7	42	47.36	-13	-7.46	-5.35	169	58.416	28.730
8	50	51.18	-5	-3.82	-1.17	25	14.600	1.392
9	63	55	8	0	8	64	0	64
10	60	58.82	5	3.82	1.18	25	14.600	1.392
11	65	62.64	10	7.64	2.36	100	58.416	5.570
12	61	66.46	6	4.46	-5.45	36	131.423	29.812
						540	408.878	130.84

Los resultados del último renglón de la tabla 10.10 representan la suma de cuadrados que corresponden a tres *fuentes de variación en el modelo de regresión* la *variación total* (SCT), la cual es la suma de las últimas dos columnas que se refieren a la *variación explicada por la regresión* (SCR) y la *variación que permanece sin explicar debido al error*. Estos se explican de manera gráfica en las figuras 10.25, 10.26 y 10.27.

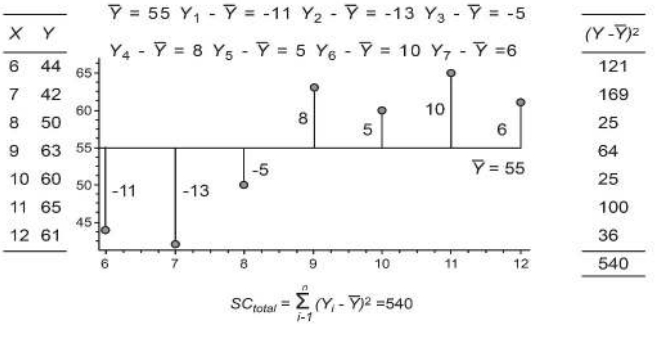


Figura 10.25 Describe la variación total, la que corresponde a las observaciones y a la media.

Los resultados se describen en las siguientes expresiones:

$$SC_{total} = \sum_{i=1}^{7} (Y_i - \overline{Y})^2 = 540$$

$$SC_{regresión} = \sum_{i=1}^{7} (\widehat{Y}_i - \overline{Y})^2 \stackrel{\circ}{=} 408.878$$

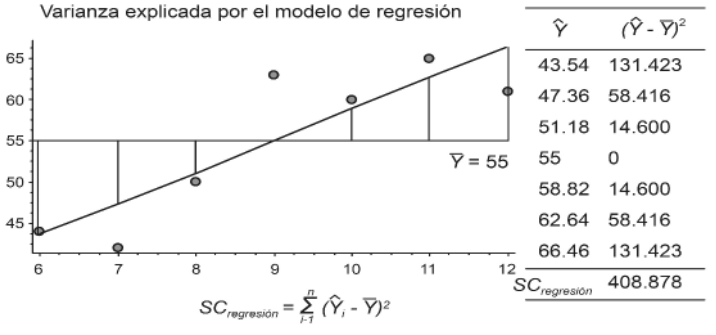
$$SC_{error} = \sum_{i=1}^{7} (Y_i - \widehat{Y}_i)^2 \stackrel{\circ}{=} 130.84$$

Nota: el símbolo ($\stackrel{\circ}{=}$) se refiere al cálculo redondeado. Como una medida de la adecuación del modelo lineal se examina qué tanto de la variación en la respuesta se explica por el modelo ajustado:

$$Y_i = \{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\} + (Y_i - \{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\}),$$

donde Y_i es el valor observado, $(\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$ explica la relación lineal y $(Y_i - \{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\})$ explica el residual. En una situación ideal donde todos los puntos están en la línea, los residuales son cero, de esta manera los valores de Y se explican por la dependencia lineal en X.

Para evaluar qué tanto los datos se ajustan al modelo se debe considerar la $SC_{regresión}$, de tal manera que si existe un buen ajuste, la $SC_{regresión}$ contribuye fuertemente a la SC_{total}. Nuevamente la situación ideal: todos los puntos sobre la línea recta nos indica que SC_{error} es cero, así SC_{total} queda totalmente explicada por los valores de X.



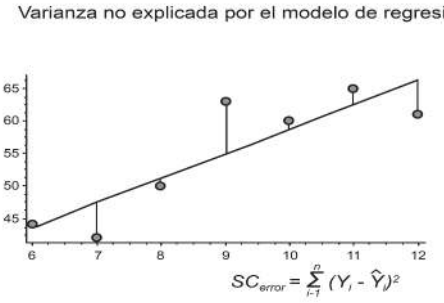
$\widehat{Y}$	$(\widehat{Y} - \overline{Y})^2$
43.54	131.423
47.36	58.416
51.18	14.600
55	0
58.82	14.600
62.64	58.416
66.46	131.423
$SC_{regresión}$	408.878

Figura 10.26 La fuente de variación relativa a la regresión: la varianza explicada por la regresión.

Construcción del estadístico de prueba F. Finalmente las *varianzas están compuestas por los grados de libertad de la regresión y del error y la razón* se indica por:

$$RV = Razón\ de\ varianzas = \left(\frac{SC_{regresión}}{gl_{regresión}}\right) \Big/ \left(\frac{SC_{error}}{gl_{error}}\right) \tag{10.21}$$

Donde los grados de libertad se obtienen de la siguiente manera, del total son el número de observaciones menos uno, $gl_{total} = n - 1$, los de la regresión son el número de parámetros menos uno, $gl_{regresión} = 2-1 = 1$, finalmente los del error son la diferencia en los anteriores, es decir $gl_{error} = gl_{total} - gl_{regresión} = (n-1) - 1 = n - 2$. Continuando con el ejemplo se tiene que:



Y	$\hat{Y}$	$(\hat{Y} - \overline{Y})^2$
44	43.54	0.212
42	47.36	28.730
50	51.18	1.392
63	55	64
60	58.82	1.392
65	62.64	5.570
61	66.46	29.812
SC_{error}		130.84

Figura 10.27 Variación no explicada debido al error.

RV es una variable aleatoria que tiene una distribución F con 1 y $n-2$ grados de libertad respectivamente, para decidir si los datos confirman la hipótesis nula se compara RV con el valor $F_c = F(1, n-2, \alpha) = F(1, 5, 0.05) = 6.607$, donde α es el nivel de significancia. Puesto que $RV > 6.607$ se rechaza la hipótesis nula, es decir, si existe una relación lineal entre las variables. Para obtener el valor de F_c, se recurre a las tablas de la distribución F en el apéndice, se puede disponer de la distribución mediante el empleo del paquete estadístico CalEst. El resumen de estos datos se recogen en la tabla del *análisis de la varianza* andeva-Tabla 10.11, siguiente:

Tabla 10.11. Análisis de la Varianza.

Fuentes de Variación	Suma de Cuadrados	GL	Cuadrado Medio	Razón de varianza
Modelo (R)	408.878	1	408.878	$RV = 15.625$
Residual (E)	130.84	5	26.168	
Total (T)	540	7		

Análisis de la varianza

La *variabilidad total* de los Y_i se expresa por:

$$SC_{total} = S_{yy} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2, \tag{10.22}$$

Se denomina la *suma de cuadrados total* y se denota por SC_{total}. Para medir la discrepancia de la linealidad, se utiliza la *suma de cuadrados de los residuales*:

$$SC_{error} = \sum_{i=1}^{n}(Y_i - \{\widehat{\beta}_0 + \widehat{\beta}_1 X_i\})^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}, \tag{10.23}$$

la SC_{error} es parte de la variabilidad de la *suma de cuadrados total*, para calcularla use las expresiones 10.14, 10.15 y 10.16. Nota: esta última expresión se obtiene sustituyendo los estimadores de mínimos cuadrados y haciendo un poco de álgebra. Para obtener la otra suma de cuadrados, se calcula la diferencia entre: S_{yy} y SC_{error}:

$$S_{yy} - SC_{error} = S_{yy} - \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{S_{xy}^2}{S_{xx}}, \tag{10.24}$$

Representa la variabilidad de los datos al modelo y se denomina *suma de cuadrados de la regresión;* se denota por $SC_{regresión}$. Así la variación total o SC_{total} queda expresada por dos componentes: $SC_{regresión}$ (*explicada por el modelo*) y SC_{error} (residual: *no explicada por el modelo*). A continuación se describe la variabilidad de las diferentes componentes del modelo.

$$SC_{total} = S_{yy}(total\ de\ Y)$$

explicada por la relación lineal:

$$SC_{regresión} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \frac{S_{xy}^2}{S_{xx}} \tag{10.25}$$

no explicada por el modelo residual:

$$SC_{error} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = SC_{total} - SC_{regresión}$$

Para resumir esta situación generalmente se presenta una tabla que se llama *análisis de la varianz*a (ANDEVA), Tabla 10.12 .

Tabla 10.12. Análisis de la Varianza.

Fuentes de Variación	Suma de Cuadrados	GL	Cuadrado Medio	Razón de varianza
Modelo (R)	$SC_{regresión}$	1	$\frac{SC_{regresión}}{1}$8	$RV = \frac{SCR/1}{SCE/(n-2)}$
Residual (E)	SC_{error}	$n-2$	$\frac{SC_{error}}{n-2}$	
Total (T)	SC_{total}	$n-1$		

Metodología para la prueba de hipótesis sobre el modelo usando el andeva

Así, la tabla 10.12 resume el procedimiento para probar la hipótesis :

$$H_0 : \text{No existe relación lineal entre } X \text{ y } Y$$
$$H_1 : \text{Sí existe relación lineal entre } X \text{ y } Y :$$

La *razón de varianzas RV*, indicada por la expresión 10.21, es *una variable aleatoria* que tiene una *distribución F* con 1 y $n-2$ grados de libertad, respectivamente; para decidir si los datos confirman la hipótesis nula se compara RV con el valor $F_c = F(1, n-2, \alpha)$.

$$Si\ RV > F_c = F(1, n-2, \alpha),\ \text{se rechaza la hipótesis } H_0$$

es decir, no existe relación lineal entre X y Y. Observe el siguiente procedimiento para obtener el valor de RV

$$CM_{regresión} = \left(\frac{SC_{regresión}}{gl_{regresión}}\right) \quad \text{y} \quad CM_{error} = \left(\frac{SC_{error}}{gl_{error}}\right)$$

Rescribimos la expresión 10.21 en términos de los *cuadrados medios*:

$$RV = Razón\ de\ varianzas = \frac{CM_{regresión}}{CM_{error}} \tag{10.26}$$

Nota: use la *distribución de probabilidad F* y el calculador de ésta en el paquete estadístico para verificar estos resultados estadísticos y conclusiones. Observe que en el caso del análisis de la regresión simple la prueba de hipótesis mediante las pruebas F y t son equivalentes. A continuación se verá la prueba usando la distribución t. Vea la propiedad de la distribución F con respecto a la t que se mostró en el capítulo 6. En referencia a la notación que se ha venido usando RV es $RV_m = F_m$ en términos de la muestra.

Coeficiente de determinación

En el análisis del modelo de regresión, una meta es evaluar el grado de asociación entre las variables X y Y. Existen dos medidas descriptivas que con frecuencia se emplean en la práctica para evaluar esta relación. Con el fin de fijar ideas, la SC_{total} mide la variación de las observaciones en Y_i cuando la variable X no se toma en cuenta, como se vió anteriormente. Sin embargo, en el caso del modelo, como ya se indicó, SC_{error} mide la variación en la variable, respuesta, Y_i cuando se utiliza la variable independiente X, también llamada predictora. Medir el efecto de X en reducir la variación en Y, consiste en expresar la reducción en la variación $SC_{regresión}$ como una proporción de la varianza total. Recuerde que la reducción se expresa por $SC_{regresión} = SC_{total} - SC_{error}$.

$$R^2 = \frac{SC_{regresión}}{SC_{total}}$$

Así, R^2 es un índice para evaluar el porcentaje de los datos que se explican mediante el modelo se llama *coeficiente de determinación*. Sustituyendo $SC_{regresión}$ en la expresión anterior y simplificando queda otra expresión para R^2, esto es,

$$R^2 = \frac{SC_{regresión}}{SC_{total}} = \frac{SC_{total} - SC_{error}}{SC_{total}} = 1 - \frac{SC_{error}}{SC_{total}}$$

Una expresión operativa para calcular el coeficiente se plantea en la siguiente ecuación,

$$\frac{SC_{regresión}}{SC_{total}} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Puesto que $0 \leq SC_{regresión} \leq SC_{total}$, entonces

$$0 \leq R^2 \leq 1.$$

Se interpretará R^2 como la reducción proporcional de la variación total asociada con el uso de la variable independiente, predictora, X. Por lo tanto, un valor grande de R^2, indica que la variación total de Y es reducida debido a la introducción de la variable independiente X. Observe los siguientes puntos:

1. Si todos los puntos caen sobre la línea recta, significa que la $SC_{error} = 0$, de esa manera $R^2 = 1$.

2. Cuando el modelo ajustado es una línea horizontal, en esa situación, se tiene que $\widehat{\beta}_1 = 0$ y observe que $\widehat{Y} \equiv \overline{Y}$, de ahí se sigue que $SC_{error} = SC_{total}$

3. Interpretación, si R^2 no es 0 o 1, sin embargo, estará entre estos valores. Por lo tanto, si R^2 es cercano a 1, se dice que existe una relación entre las variables X y Y.

4. La raíz cuadrada de R^2 proporciona el *coeficiente de correlación* , este corresponde al que se discutió en los apartados anteriores en este capítulo, así:

$$r = \pm\sqrt{\frac{S_{xy}^2}{S_{xy}S_{yy}}} = \pm\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{10.27}$$

Coeficiente de determinación para ejemplo 7

La recta que se ajustó :$\widehat{Y} = 55 - 3.821X$. ¿Qué tanto de la variabilidad es explicada por el modelo de regresión? Usando el resumen de los cálculos realizados en este ejemplo se puede estimar el valor del coeficiente de determinación R^2; aquí tomamos los datos de la tabla 10.11, donde $SC_{regresión} = 408.878$ y $SC_{total} = 540$, sustituyendo en la fórmula se obtiene:

$$R^2 = \frac{408.878}{540} = 0.76$$

El valor se interpreta diciendo que el 76 % de la variabilidad en Y se explica por la regresión lineal y se concluye que el modelo es satisfactorio. ¿Qué se puede concluir si R^2 es pequeño? ¿Qué procedimientos se deben confirmar para tal caso?

10.4.1 Inferencia con respecto a la pendiente β_1 y β_0

1. ¿Cuál es el procedimiento estadístico para verificar que no hay cambio en X si y sólo si $\beta_1 = 0$ aplicando el estadístico de prueba $t - Student$? Planteamiento de la hipótesis sobre la pendiente β_1

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{10.28}$$

2. ¿Cuál es el estadístico apropiado, en el ambiente de datos bivariados, para construir el intervalo de confianza para β_0 y β_1? Intervalo de $(1-\alpha)\,\%$ de confianza para β_0 y β_1

$$L_{0i} \leq \beta_0 \leq L_{0d}$$

$$L_{1i} \leq \beta_1 \leq L_{1d}$$

Para contestar ambas preguntas es necesario construir el estadístico de prueba, y en esa dirección se requiere estudiar las variables aleatorias $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que generaron los estimadores de mínimos cuadrados. ¿Cuál es la distribución de probabilidad, la media y varianza de estas variables?

Propiedades estadísticas de los estimadores

Las respuestas a estas preguntas se basan en las propiedades que tienen los estimadores de mínimos cuadrados, entre ellas se sabe que éstos tienen una distribución de probabilidad normal, con las siguientes características: media $\mu_0 = \beta_0$ y varianza $\sigma^2(\widehat{\beta}_0)$ para la ordenada al origen y media $\mu = \beta_1$ y varianza $\sigma^2(\widehat{\beta}_1)$ para la pendiente. Las expresiones para la varianza de los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ se representan mediante las siguientes expresiones:

$$\widehat{\sigma}^2(\widehat{\beta}_0) = \widehat{\sigma}^2(\frac{1}{n} + \frac{X^2}{S_{xx}}), \text{ y } \widehat{\sigma}^2(\widehat{\beta}_1) = \frac{\widehat{\sigma}^2}{S_{xx}}$$

Donde el estimador de la varianza σ^2 es $\widehat{\sigma}^2 = \frac{SC_{error}}{n-2}$, es decir el cuadrado medio del error es el estimador de la varianza σ^2. Nota: recuerde que SC_{error} se obtiene sustituyendo las expresiones 10.14, 10.15 y 10.16. En consecuencia, el estimador de la desviación estándar $\widehat{\sigma}$ es la raíz cuadrada de $\frac{SC_{error}}{n-2}$. En resumen, un resultado importante:

Varianza estimada	$\widehat{\sigma}^2 = \frac{SC_{error}}{n-2}$
Error estándar	$\widehat{\sigma} = \sqrt{\frac{SC_{error}}{n-2}}$

De estas expresiones se deriva el error estándar para ambos estadísticos:

Error estándar para el estadístico $\widehat{\beta}_1$	$ES(\widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{1}{S_{xx}}}$
Error estándar para el estadístico $\widehat{\beta}_0$	$ES(\widehat{\beta}_0) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{X^2}{S_{xx}}}$

Metodología para probar la hipótesis: $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$

Ahora queremos responder a la pregunta con el fin de probar la hipótesis planteada en la expresión 10.28. Recuerde que para ello se requiere proponer un estadístico de prueba. En el procedimiento establecido se han planteado tres esquemas relacionados entre sí, tal y como se describió en el capítulo 9. Aquí ese papel lo desempeñan los estadísticos β_1 y $t - Student$, además del $valor - p$. Así que con las propiedades del estadístico β_1 expuestas en el punto anterior y la figura 10.28, a continuación se determinan los puntos críticos a partir del nivel de significancia α.

1. Dado que la hipótesis alternativa tiene el signo $\neq$, indica que tiene dos valores: uno para la izquierda generado para $\alpha/2$, el valor crítico es $t_{ci} = t(gl, \alpha/2)$. El otro, a la derecha con $1 - \alpha/2$ y valor crítico es $t_{cd} = t(gl, (1 - \alpha/2))$. Se usa la distribución $t - Student$ ya que el tamaño de muestras es pequeño.

2. Los estadísticos de prueba derivados para esta situación son:

Izquierda	$\widehat{\beta}_{1ci} = \beta_1 + t(gl, \alpha/2)ES(\widehat{\beta}_1)$
Derecha	$\widehat{\beta}_{1cd} = \beta_1 + t(gl, (1 - \alpha/2))ES(\widehat{\beta}_1)$

   No se rechaza la hipótesis nula 10.28, si el valor $\widehat{\beta}_{1m}$ estimado de la muestra está entre estos valores, es decir $\widehat{\beta}_{1ci} \leq \widehat{\beta}_{1m} \leq \widehat{\beta}_{1cd}$, en caso contrario se rechaza.

3. Con esta información se concluye que el modelo es adecuado para explicar la relación entre variables.

Procedimiento alternativo usando el estadadístico estandarizado

La prueba anterior es similar con el estadístico $t-Student$, y en atención a la información que proporciona la muestra, éste tiene la forma:

$$t_m = \frac{(\widehat{\beta}_{1m} - \beta_1)}{\widehat{\sigma}(\widehat{\beta}_1)}$$

Se *distribuye* como una $t - Student$ con $n - 2$ gl. Se realiza la prueba de hipótesis planteada comparando este valor con los puntos críticos t_{ci} y t_{cd}. Entonces, no se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ si t_m está entre estos puntos, y de nuevo, se rechaza en caso contrario. Observe la figura 10.28.

Procedimiento alternativo con el $valor - p$

Finalmente se calcula la probabilidad para obtener el $valor-p$ y comparar éste con el nivel de significancia $\alpha/2$, para la prueba bilateral, $H_1 : \beta_1 \neq 0$, o con α para la prueba mayor $H_1 : \beta_1 > 0$ o menor $H_1 : \beta_1 < 0$.

Estas probabilidades, cuando se aplica la distribución $t - Student$, se obtienen mediante la expresión:

$$valor - p = P(t \geq t_m), \text{ o}$$
$$valor - p = P(t \leq t_m)$$

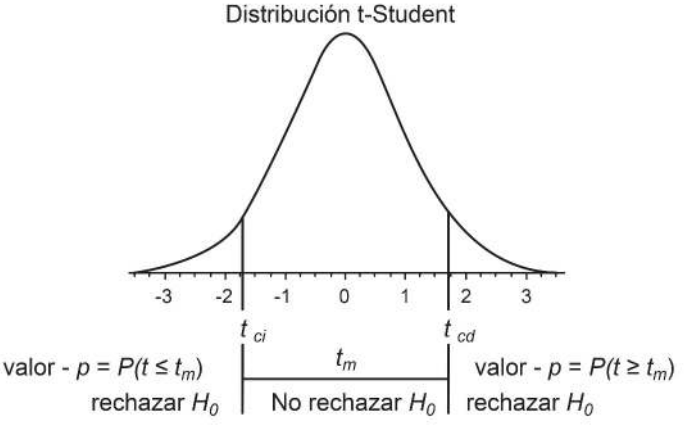


Figura 10.28 Procedimiento de la prueba de hipótesis $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

Comentario: aunque en este trabajo no se expondrá la prueba de hipótesis para el intercepto β_0, ésta sigue un procedimiento similar al del parámetro de la pendiente, por ejemplo usando la distribución $t - Student$

$$t_m = \frac{(\widehat{\beta}_{0_m} - \beta_0)}{\widehat{\sigma}(\widehat{\beta}_0)}$$

Intervalos de confianza para los parámetros del modelo de regresión

1. Intervalo de confianza para β_1. Se puede encontrar un intervalo de confianza para el parámetro β_1 usando la distribución t. Un intervalo de confianza para β_1 con un nivel de $100(1-\alpha)\,\%$ se obtiene mediante la expresión:

$$\begin{aligned} L_{1i} &= \widehat{\beta}_1 + t(gl, \alpha/2)\widehat{\sigma}(\widehat{\beta}_1) = \widehat{\beta}_1 + t(gl, \alpha/2)\widehat{\sigma}\sqrt{\tfrac{1}{Sxx}} \\ L_{1d} &= \widehat{\beta}_1 + t(gl, (1-\alpha/2))\widehat{\sigma}(\widehat{\beta}_1) = \widehat{\beta}_1 + t(gl, (1-\alpha/2))\widehat{\sigma}\sqrt{\tfrac{1}{Sxx}} \end{aligned} \tag{10.29}$$

   donde $t(gl, \alpha/2)$ es el punto correspondiente a la distribución t para $gl = n - 2$ y $\alpha/2$.

2. Intervalo de confianza para β_0. Con un nivel de significancia del $100(1-\alpha)\,\%$:

$$L_{0i} = \widehat{\beta}_0 + t(gl, \alpha/2)\widehat{\sigma}(\widehat{\beta}_0) = \widehat{\beta}_0 + t(gl, \alpha/2)\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{Sxx}}$$
$$L_{0d} = \widehat{\beta}_0 + t(gl, (1-\alpha/2))\widehat{\sigma}(\widehat{\beta}_0) = \widehat{\beta}_0 + t(gl, (1-\alpha/2))\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{Sxx}} \tag{10.30}$$

.

Ejemplo 10.8

Los años de servicio, así como la capacitación son elementos relevantes para alcanzar un salario dentro de las empresas. Se tomó una muestra de 14 ingenieros de proceso en diferentes empresas que ocupan un puesto similar. Se les preguntó por los años de servicio y su sueldo, y la información recopilada se muestra en la tabla del ejemplo. El sueldo está expresado por 1000 en pesos mexicanos. Las varibles X en años y Y sueldo.

X	19	18	9	17	18	16	11	5	11	8	15	13	21	13
Y	55.8	55.6	30.9	40.8	50.4	35.9	30.9	10.5	20	15.1	40.1	20.3	55.8	30.8

1.-Proporcione el modelo de regresión e interprete el parámetro β_1. 2.-Realice la prueba de hipótesis sobre β_1, y construya el intervalo de confianza para β_1.

Solución

Inciso 1. El resumen estadístico para obtener los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de los parámetros, aplicando las expresiones 10.12 y 10.11 es:

$$\overline{X} = 13.857,\ \overline{Y} = 35.207,\ S_{xx} = 281.770,\ S_{xy} = 855.012,\ S_{yy} = 3075.41.$$

Así, se tiene que:

$$\widehat{\beta}_1 = \frac{855.012}{281.770} = 3.035,\quad \widehat{\beta}_0 = 35.207 - (3.035)(13.857) = -6.845$$

El modelo se describe mediante la expresión:

$$\widehat{Y} = -6.845 + 3.035X$$

Inciso 2. Procedimiento metodológico para la prueba de hipótesis de la pendiente es como sigue:

1. Planteamiento de la hipótesis sobre la pendiente β_1, dado que el modelo crece conforme pasan los años, se propone la siguiente hipótesis:

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 > 0$$

2. El nivel de significancia propuesto para realizar esta prueba es $\alpha = 0.05$. Entonces el valor crítico es $t_{cd} = t(gl, (1-\alpha)) = t(10, 0.95) = 1.781$. Se sustituye este valor en $\widehat{\beta}_{1cd} = \beta_1 + t(gl, (1-\alpha/2))ES(\widehat{\beta}_1)$ y completando se tiene que, $\widehat{\beta}_{1cd} = \beta_1 + t(gl, (1-\alpha/2))ES(\widehat{\beta}_1) = 0 + 1.781(0.377) = 0.671$.

3. Dado que $\widehat{\beta}_m = 3.035$, se rechaza la hipótesis nula ya que $\widehat{\beta}_m > \widehat{\beta}_{1cd}$.

4. Por lo tanto existe una relación lineal, que indica que cada año de experiencia el salario aumenta tres mil pesos, 3035.00 pesos. El procedimiento para probar hipótesis empleando la $t - Student$ se describe en la figura 10.29.

Cálculos para obtener el error estándar de $\widehat{\beta}_1, ES(\widehat{\beta}_1)$

$$ES\left(\hat{\beta}_1\right) = \sqrt{CM_{error}}\sqrt{\frac{1}{S_{xx}}}$$
$$SC_{error} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \doteq 3071.410 - 2594.476 = 480.934$$
$$CM_{error} = \frac{SC_{error}}{n-2} = 40.073$$
$$\sqrt{\frac{1}{S_{xx}}} = \sqrt{\frac{1}{281.77}} = 5.957x10^{-2}$$
$$ES(\widehat{\beta}_1) = \sqrt{40.073}(5.957x10^{-2}) = 0.377$$

Hipótesis
$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 > 0$
Se rechaza H_0 si $t_m > t_{cd}$
$t_m = 8.055 > t_{cd} = t(12, 0.95) = 1.781$
$t_m = \frac{3.035-0}{0.377} = 8.055$
o se rechaza $H_0, valor - p < \alpha = 0.05$
$valor - p = P(t > 8.055) = 0.000$

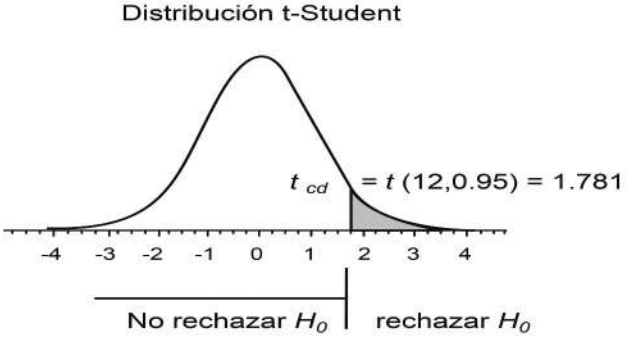


Figura 10.29 Procedimiento para probar la hipótesis β.

Intervalo de confianza para el modelo

El análisis de regresión se emplea también para pronosticar y predecir los valores de la variable independiente. A partir del modelo de regresión, para un valor dado de la variable $X = x^*$ se obtiene un valor Y^*. Entonces, se procede a construir un *estimador puntual* o de *intervalo de confianza* sobre la media μ de la respuesta Y. Se plantean las siguientes preguntas:

1. ¿Cuál es el procedimiento para obtener un intervalo de confianza del modelo?

2. Si el modelo resulta significativo, ¿se puede usar este para predecir? Si ¿cómo? Mediante un *intervalo de predicción* de Y dada X.

Intervalo de confianza de la respuesta media Y para un valor dado de X

Cuando se propone un valor específico de la variable independiente X, es de interés estimar el valor de la respuesta para ese valor de X; suponga que el valor propuesto es $X = x^*$, con ello la respuesta esperada es $Y = \beta_0 + \beta_1 x^*$, y se estima por $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x^*$. Entonces resulta que para cada valor dado de X, $\widehat{Y}^*$ es una variable aleatoria que tiene una media condicionada a un valor de X y se denota por $\mu_{Y|X^*} = E(\widehat{Y}^*) = E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$. Para tener una mejor idea observe la figura 10.30, donde $\mu = Y = \beta_0 + \beta_1 X$.

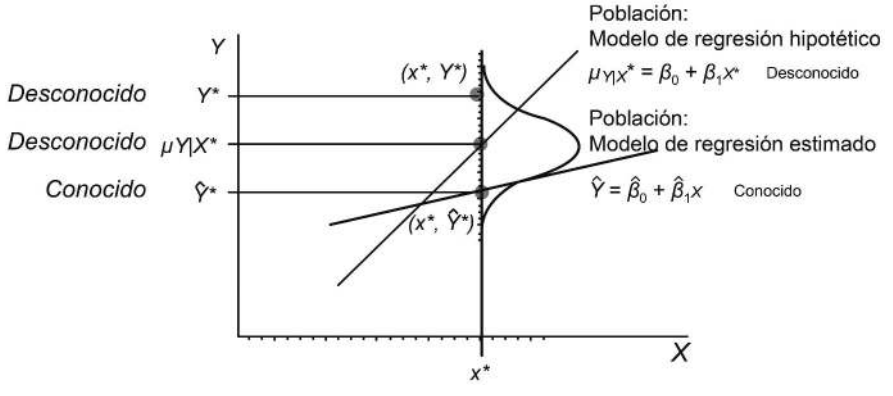


Figura 10.30 Panorama estadístico de la relación población y muestra en el proceso de estimación.

Escenario estadístico

Con estas condiciones estamos en la situación similar a la planteada en los capítulos 7 y 8 donde se desarrolló el procedimiento para realizar la estimación de la media μ. Es decir, se construye una expresión, tal como la que sigue:

$$(Estimador + DP_i ES(estimador), Estimador + DP_d ES(estimador))$$

donde $ES(estimador)$ es el error estándar del estimador. En general, los DP_i y DP_d corresponden a las probabilidades de una distribución a la izquierda y derecha respectivamente, en particular para la $t - Student = DPlpha/2))$. Considerando esta nueva sugerencia, se procede de manera similar a la estudiada con anterioridad, para estimar de manera puntual o por intervalo un parámetro.

Comentario: los intervalos de confianza y predicción tienen la misma forma que los estudiados antes, ellos sólo difieren de la desviación estándar. Se puede decir que obtener el $ES(estimador)$ consiste en el trabajo teórico que hay que desarrollar, el cual se realiza en cursos de métodos estadísticos más avanzados. Aquí se escriben las fórmulas que se consiguen de esa tarea técnica.

Existen dos posibles estrategias prácticas para obtener el *intervalo de confianza* para la media condicionada de Y.

1. Se mantiene fijo el valor de x^*, así se obtendrán muchos valores de $\widehat{Y}^*$. Entonces, con una confianza de $(1-\alpha)\,\%$ de confianza la media $\mu_{Y|X^+}$ estará contenida en el intervalo especificado.
2. Si se toman muchas muestras de la pareja de valores (X, Y) el intervalo de confianza con base en cada muestra, el $(1-\alpha)\,\%$ de ellos contendría a $\mu_{Y|X^+}$ en particular para el valor dado de $X = x^*$.

Distribución muestral de $\widehat{Y}^*$

La distribución muestral $\widehat{Y}^*$ es similar a las distribuciones muestrales presentadas con anterioridad, y en referencia a los dos casos citados en el punto previo.

Considerando que la variable ε en el modelo $Y = \beta_0 + \beta_1 X + \varepsilon$, se distribuye como una normal, entonces $\widehat{Y}^*$ se distribuye como una normal con:

- Media

$$\mu_{Y|X^+} = E(\widehat{Y}^*) = E(\widehat{\beta}_0 + \widehat{\beta}_1 x^*) = \beta_0 + \beta_1 x^* \text{ y}$$

- Varianza

$$\sigma^2(\widehat{Y}^*) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}} \right)$$

La varianza estimada es:

$$\widehat{\sigma}^2(\widehat{Y}^*) = \widehat{\sigma}^2 \left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}} \right) \tag{10.31}$$

donde $\widehat{\sigma}^2 = CM_{error}$. El error estándar del estimador es:

$$ES(\widehat{Y}^*) = \widehat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}} \right)} \tag{10.32}$$

- Ideas sobre el intervalo de predicción. El punto fino, de nueva cuenta, es determinar la varianza sobre la predicción para poder obtener el intervalo de confianza. Observe de la figura 10.30, la distancia de Y^* a $\widehat{Y}^*$, ésta tiene dos componentes, a saber: la distancia de Y^* al modelo de regresión de la población y luego la distancia comprendida entre este modelo poblacional y la línea de $\widehat{Y}^*$. Así la varianza de la primer distancia es $\widehat{\sigma}^2$ y la de la segunda es $\widehat{\sigma}^2_{\widehat{Y}^*}$. De tal manera la suma es:

$$\widehat{\sigma}^2 + \widehat{\sigma}^2(\widehat{Y}^*) = \widehat{\sigma}^2 + \widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}\right) = \widehat{\sigma}^2\left(1 + \left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}\right)\right)$$

  La desviación estándar es:

$$\widehat{\sigma}\sqrt{\left(1 + \left(\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}\right)\right)}$$

Los límites de confianza para $\beta_0 + \beta_1 X^*$

La meta es construir los límites del intervalo de confianza para media de Y dada X $\mu_{Y|X^+}$, esto es:

$$L_i \leq \mu_{Y|X^+} \leq L_d$$

Empleando la expresión 10.32, los límites son:

$$L_i = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + t(gl, \alpha/2)ES(\widehat{Y}^*)$$
$$L_i = \widehat{\beta}_0 + \widehat{\beta}_1 x^* + t(gl, (1-\alpha/2))ES(\widehat{Y}^*)$$

Finalmente el intervalo de confianza para la respuesta esperada $\beta_0 + \beta_1\ X^*$ con un nivel de $100(1-\alpha)\,\%$ confianza es:

$$\left(\widehat{\beta}_0 + \widehat{\beta}_1 x^* + t(gl, \alpha/2)\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}}, \widehat{\beta}_0 + \widehat{\beta}_1 x^* + t(gl, (1-\alpha/2))\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}}\right) \quad (10.33)$$

La aplicación de estas expresiones y la ejemplificación de la estimación se realizarán en el ejemplo 8.

Intervalo de predicción de la respuesta Y **para un valor dado de** X, **un nivel de confianza del** $(1\text{-}\alpha)\,\%$

Una vez que se ha construido el modelo de regresión: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$, si se da un valor de $X = X^*$, se obtendrá un valor de predicción $\widehat{Y}$. En este caso se puede construir un intervalo para esa predicción con un nivel de $100(1-\alpha)\,\%$ confianza. En este trabajo, sólo se indicará la fórmula:

$$\left(\widehat{\beta}_0+\widehat{\beta}_1x^*+t(gl,\alpha/2)\widehat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{X})^2}{S_{xx}}},\widehat{\beta}_0+\widehat{\beta}_1x^*+t(gl,(1-\alpha/2))\widehat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x^*-\overline{X})^2}{S_{xx}}}\right) \tag{10.34}$$

Bandas del $(1\text{-}\alpha)\,\%$ **de confianza para la predicción de** Y **dado** X **y la media de** Y **dado** X

Véase la figura 10.31.

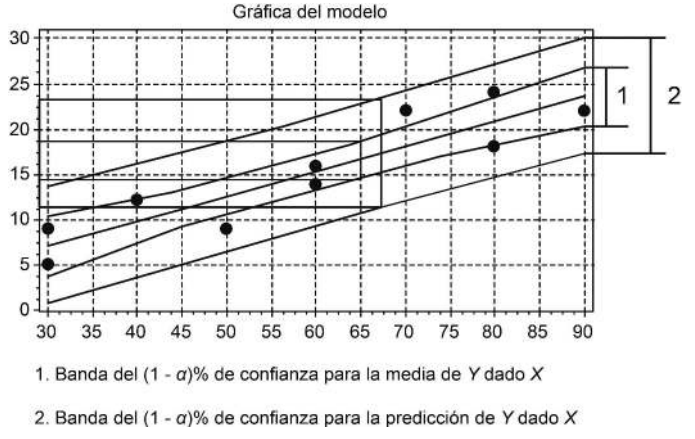


Figura 10.31 Bandas de confianza para la predicción de Y dado X y la media Y dado X.

Observaciones:

1. Repetir el proceso descrito en la figura 10.30 para todos los valores de X dentro del rango de valores del problema estudiado, generan las bandas descritas en la figura 10.31. Recuerde que el rango comprende al valor mínimo y máximo de X.
2. Los intervalos de predicción de $\widehat{Y}$ son más anchos que los intervalos de confianza de $\mu_{Y|X^+}$
3. Ambos intervalos crecen cuando x^* se aleja de la media $\overline{X}$. Analice ese efecto considerando la fórmula: $A=\sqrt{\frac{1}{n}+\frac{(x^*-\overline{X})^2}{S_{xx}}}$, por ejemplo, si $x^*=\overline{X}$, entonces $A=\sqrt{\frac{1}{n}}$

Hipótesis estadística para $\mu_{Y|X^+}$

Las hipótesis estadísticas para la predicción se muestran por:

$$H_0:\beta_0+\beta_1x^*=\mu_{Y|X^+0},\qquad H_1:\beta_0+\beta_1x^*\neq\mu_{Y|X^+0},$$

el estadístico de prueba que permite contrastar esta hipótesis es:

$$t_c = \frac{\widehat{\beta}_0 + \widehat{\beta}_1 x^* - \mu_{Y|X^+0}}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \overline{X})^2}{S_{xx}}}}, gl = n - 2,$$

10.4.2 Reporte estadístico del modelo de regresión en *CalEst*

En la investigación de la relación entre dos variables el *diagrama de dispersión* es una técnica gráfica para entender un análisis estadístico formal. Este análisis se realiza en *CalEst* en módulo de regresión, en el cual aparecen tres opciones: la regresión simple, la regresión múltiple y la regresión avanzada. En esta sección se abordará la primera; como paso inicial se anotan los datos en una hoja de trabajo, mediante la creación o apertura de un archivo en *CalEst* o en otra hoja electrónica.

Ejemplo 10.9

Una actividad importante en la esfera de la economía es la venta por llamadas telefónicas. Existe una organización que administra este tipo de servicios, y de la selección aleatoria de un sistema se tiene el registro de 10 equipos de la venta realizada de un producto. En la tabla siguiente se describe el número de llamadas y la cantidad del monto de las ventas, las cuales se multiplica por 1000.

Equipos	**1**	**2**	**3**	**4**	**5**	**6**	7	**8**	**9**	**10**
X_i: llamadas	30	30	40	50	60	60	70	80	80	90
Y_i: ventas	9	5	12	9	14	16	22	18	24	22

Responda a las siguientes cuestiones:

1. Capture la información en una hoja de datos con el fin de resolver el ejemplo usando la calculadora, CalEst.
2. Presente la información en un diagrama de dispersión.
3. Obtenga el modelo a través de los resultados generados en el reporte de salida de la calculadora.
4. Plantee las hipótesis que va a probar y verifíquelas; use el reporte del inciso anterior.
5. Observe la tabla del análisis de la varianza (andeva) e interprétela.
6. Obtenga el intervalo de confianza para los parámetros del modelo.
7. Interprete los coeficientes de correlación y determinación.
8. Indique los valores del intervalo de confianza para la media del modelo; usando el calculador pruebe con varios valores de α e interprete los resultados.

9. Anote los límites del intervalo de confianza para un valor predicho; también use el calculador considerando diferentes valores para α

10. Escriba un reporte de los valores predichos, el valor de los residuales y haga un análisis con estos valores; comente sus resultados.

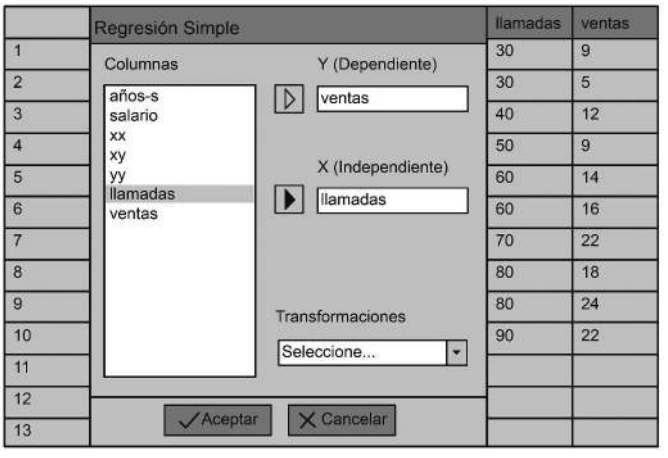


Figura 10.32 Hoja de captura y descripción del módulo de regresión, para realizar el análisis estadístico del modelo.

Solución mediante el uso de CalEst

Solución a las cuestiones 1 y 2. Capturar los datos, la hoja aparece en la figura 10.32; una vez realizada esa actividad, se llama la opción *regresión simple* que está en uno de los módulos que contiene el CalEst. En ésta se llenan los cuadros para la variable dependiente *ventas* en este caso, y la variable independiente *llamadas*, tal y como se observa en la figura 10.32.

A continuación se oprime el botón **Aceptar**. A partir de aquí se presenta el escenario que comprende el reporte estadístico; en principio se exhiben cuatro diagramas que forman parte del análisis gráfico del modelo de regresión, éstos son: 1) el diagrama de dispersión con el modelo; 2) observadas vs predichos; 3) residuales vs X; 4) residuales vs predichos; cada uno se irán explicando en el marco de la solución del problema. Para efectos prácticos, se puede aislar cada gráfica para su análisis o para imprimir un reporte; en la figura 10.33 se despliegan las cuatro, y el círculo permite guardar la gráfica o copiarla; el cuadro da la opción para estimar de manera dinámica los intervalos de confianza para la media del modelo y el predicho. La gráfica superior de la izquierda corresponde al diagrama de dispersión y el modelo, las líneas corresponden a los intervalos del 95 % de confianza para la media del modelo y del valor predicho que se explicará más adelante.

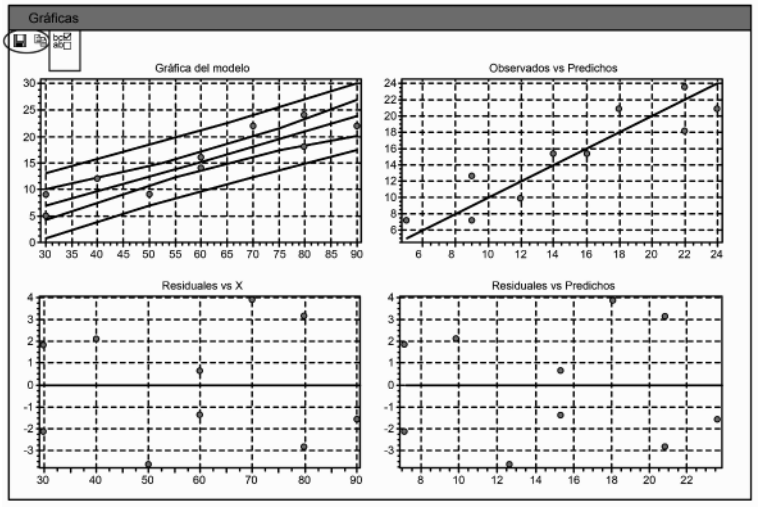


Figura 10.33 Descripción de los cuatro diagramas.

Solución a la cuestión 3. Atrás de estas gráficas se halla una hoja que completa el análisis estadístico, esta contiene los valores estimados de los parámetros, el error estándar, el cálculo de los estadísticos de prueba y el $valor - p$, con estos dos últimos elementos se realiza la prueba de hipótesis. Abajo de esta reseña, siguen los cálculos correspondientes a la tabla del andeva, junto con un resumen que incluye al coeficiente de determinación. Por último, se tienen los intervalos con un nivel de confianza del 95 % para los parámetros. Los resultados se despliegan en la figura 10.34 y para llegar ahí se han aplicado las fórmulas 10.18 y 10.17 para estimar los parámetros y 10.10 para obtener el modelo, por lo que el modelo es:

$$Y_i = -1.07 + 0.274X_i$$

Interpretación: en el modelo lineal en que se está estudiando el coeficiente de la variable independiente X es en particular el más importante. Así, el valor de $\widehat{\beta}_1 = 0.274$ indica que una llamada en X, la venta que se obtiene en Y es 274 pesos.

Solución a la cuestión 4. Revise el procedimiento que se propuso para la prueba de hipótesis ya que aquí se han elaborados las operaciones con ayuda del calculador y se presenta el análisis de manera resumida. La hipótesis planteada es: $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$. Para verificarla se obtiene el estadístico de prueba t_m. Observe en la figura 10.34 que éste es:

$$t_m = \frac{0.274}{0.044} = 6.213$$

Este valor de t_m se compara con el valor de $t_{cd} = t(gl, \alpha) = t(8.0.05) = 1.860$, consulte este valor en las tablas o el calculador de la distribución $t - Student$, en CalEst. Se puede observar que $t_\alpha > t_{cd}$.

Por lo que se concluye que se rechaza la hipótesis nula. Tome en consideración las siguientes cuentas $\widehat{\sigma}^2 = CM_{error} = 7.957$, $S_{xx} = 40.9$ y $ES(\widehat{\beta}_1) = \sqrt{CM_{error}}\sqrt{1/S_{xx}} = 0.044$. De la figura 10.34 se tiene la alternativa del $valor-p$, puesto que se tiene que $valor-p = 0.0003 < \alpha = 0.05$, y se confirma el resultado.

Solución a la cuestión 5. Se realiza la prueba para $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 > 0$, ahora empleando la tabla del análisis de la varianza, que da lugar al estadístico F, expuesto en la tabla 10.12, y descrito por la expresión 10.26, de nueva cuenta la figura 10.34 indica que:

$$RV = \frac{CM_{regresión}}{CM_{error}} = \frac{307.247}{7.957} = 38.615$$

Puesto que $RV > F_c = F(gl_n, gl_d, \alpha) = F(1, 8, 0.05) = 5.317$, se concluye rechazar H_0, lo que indica que hay un efecto lineal. Como se ha visto tanto la prueba con el estadístico $t - Student$ y como con el F, dan resultados equivalentes.

Solución a la cuestión 6. Recuerde que, los intervalos para los parámetros $\widehat{\beta}_1$ y $\widehat{\beta}_0$ se obtienen mediante las expresiones 10.29, y 10.30. Entonces de la figura 10.34 observe que $L_{1i} = 0.1724$, $L_{1d} = 0.3757$ para β_1 y $L_{0i} = -7.4108$, $L_{0d} = 5.2689$ para β_0, así:

$$0.1724 \leq \beta_1 \leq 0.3757$$
$$-7.4108 \leq \beta_0 \leq 5.2689$$

Solución a la cuestión 7. El coeficiente de correlación $r = 0.9102$, indica que existe una estrecha relación lineal entre los datos. El coeficiente de determinación $R^2 = 0.8284$, se interpreta diciendo que 82 % de los datos son explicados por el modelo. El coeficiente R^2_{aju} se explicará en el siguiente apartado.

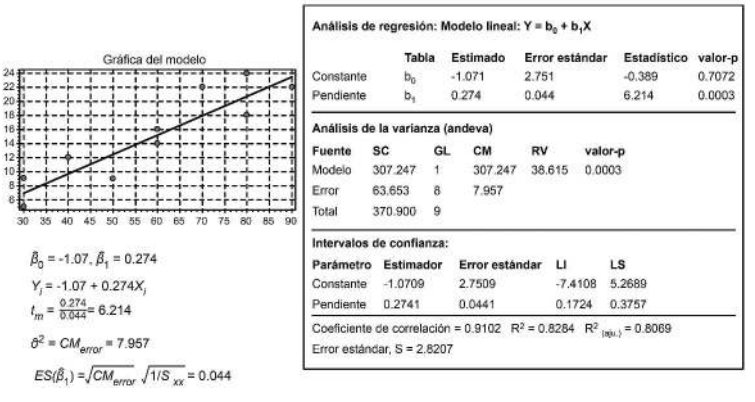


Figura 10.34 Resultados que genera el CalEst, con ellos se puede realizar el análisis estadístico sobre el modelo.

Solución a las cuestiones 8 y 9. En la figura 10.33 se describe el análisis gráfico del modelo proporcionado por el *CalEst*. En la gráfica superior izquierda aparece el modelo con las bandas de confianza del 95 % tanto para la media del modelo como para el valor predicho. Note que el cuadro que se indica

con las letras ab, brinda la posibilidad de cambiar el nivel de confianza para ambos casos.

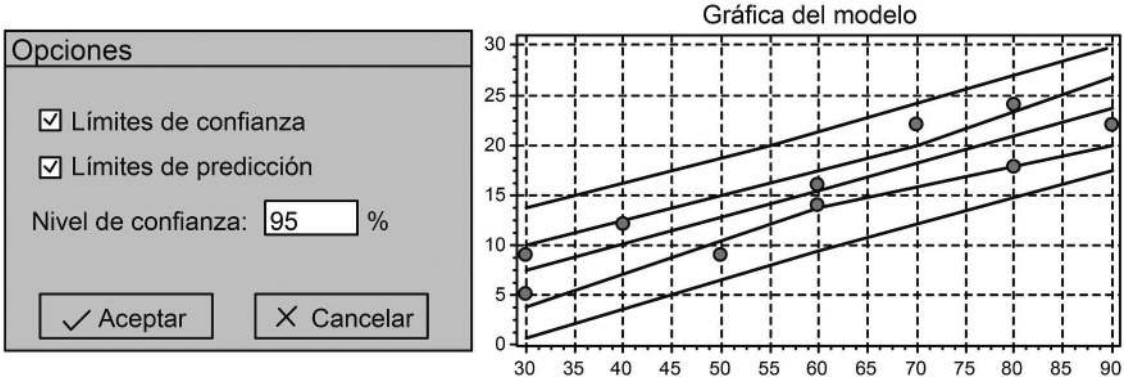


Figura 10.35 Salida del CalEst para mostrar la estimación por intervalo para la predicción y la media.

La línea de regresión que se ajustó a los datos descritos en el ejemplo 1 es $\widehat{Y} = -1.07 + 0.274X$, la venta correspondiente cuando el número de llamadas es $X^* = 65$ se estima por $\widehat{\beta}_0 + \widehat{\beta}_1 X^* = -1.07 + (0.274) * 65 = 16.745$. La desviación estándar se estima mediante la siguiente expresión:

$$2.8207\sqrt{\frac{1}{10} + \frac{(65 - 5.9)^2}{40.9}} = 0.893$$

El intervalo de confianza del 95 % para la media de la venta Y con al número de llamadas $X^* = 65$ es:

$$16.745 + t(8, 0.025)(0.893), 16.745 + t(8, 0.975)(0.893) = (14.600, 18.889).$$

Interpretación: con un 95 % de confianza las ventas Y está entre $(14.600, 18.889)$ considerando que $X^* = 65$, para un valor no observado de X.

Ahora se estimará el intervalo para un valor predicho, considere nuevamente $X^* = 65$ número de llamadas, entonces el valor predicho es: $\widehat{\beta}_0 + \widehat{\beta}_1 X^* = -1.07 + (0.274) * 65 = 16.745$. La figura 10.35 complementa el análisis estadístico de este ejemplo. Para reproducirlo oprima el botón tercero en la segunda franja. La predicción se obtiene usando el botón $\widehat{y}$. Un intervalo de confianza del 95 % de confianza es:

$$\left(16.74 - 2.306(2.8207)\sqrt{1 + \tfrac{1}{10} + \tfrac{(65-5.9)^2}{40.9}}, 16.74 - 2.306(2.8207)\sqrt{1 + \tfrac{1}{10} + \tfrac{(65-5.9)^2}{40.9}}\right)$$
$$= (16.74 - 6.85, 16.74 - 6.85) = (9.89, 23.59).$$

Comentario: Si usa la opción que se muestra en la figura 10.36, puede practicar dando diferentes valores de X para observar cómo cambian los valores de ambos intervalos cuando X se aleja de la media.

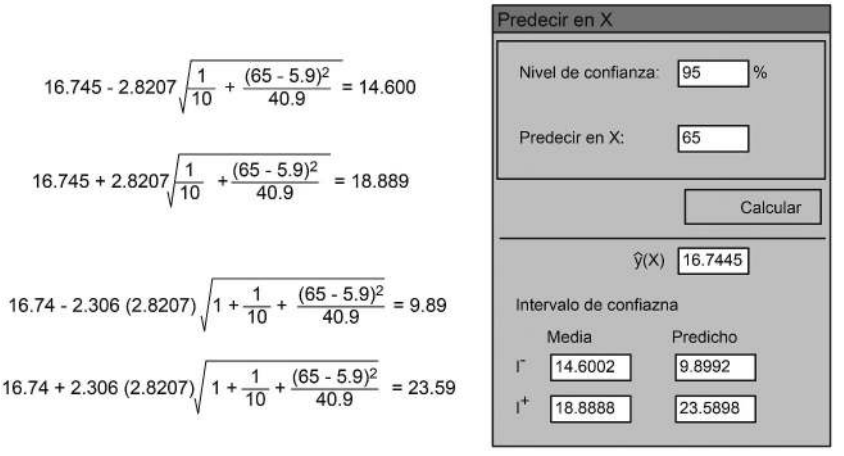


Figura 10.36 Nivel de confianza para la media y predicción de Y dado X.

Solución a la cuestión 10. Vea primero el apartado de residuales que está a continuación. En la figura 10.32, la gráfica de la derecha superior muestra la relación entre los valores reales y los predichos e indica qué tanto explica el modelo. Las siguientes dos gráficas completan el análisis del modelo mediante la evaluación de los residuales, en ellas se puede interpretar lo siguiente: si existe homogeneidad en la varianza, así como dispersión de los residuales y puntos aberrantes. Abajo de estas gráficas hay una quinta gráfica sobre los residuales que interpreta la dependencia con respecto al tiempo.

Residuales

¿Qué procedimiento estadístico se sigue para evaluar los supuestos del modelo? ¿Cuál es la importancia de ε en el modelo? $\varepsilon = Y - (\beta_0 + \beta_1 X)$.

Las desviaciones individuales entre las observaciones Y_i y los valores ajustados $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ se llaman *residuales*, y se denotan con e_i. Los residuales se expresan por:

$$e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i).$$

Los residuales permiten verificar los supuestos estadísticos en los que se basa el método de mínimos cuadrados y son de utilidad para evaluar otras características del modelo, estos aspectos se verán posteriormente. Una propiedad de los residuales es $\sum_{i=1}^{n} e_i = 0$. La suma de cuadrados de los residuales es:

$$SC_{error} = \sum_{i=1}^{n} e_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \tag{10.35}$$

Esta expresión es útil para estimar la varianza, así la varianza σ^2 se estima por:

$$\widehat{\sigma}^2 = \frac{SC_{error}}{n-2}$$

La gráfica en la figura 10.33 del extremo inferior izquierdo permite evaluar si el modelo es lineal, la varianza es homogénea. Si esa gráfica fuera como las de la figura 10.37, se tienen ejemplos de un modelo no lineal, y las varianzas son no homogéneas.

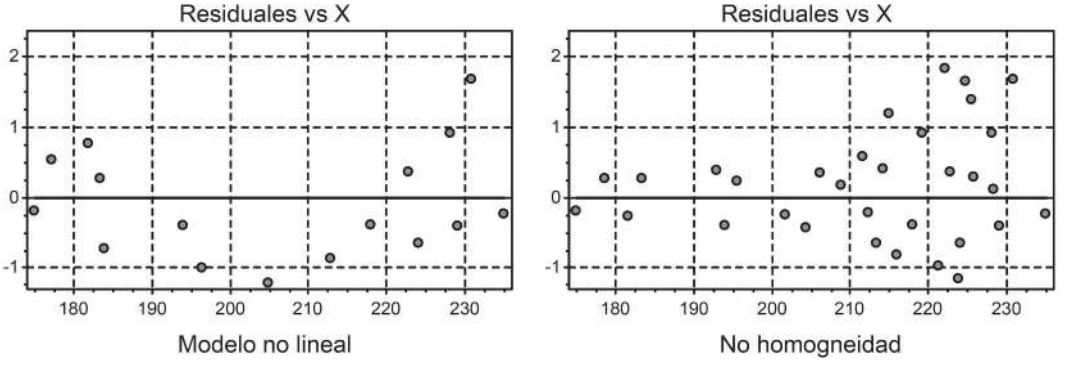


Figura 10.37 Dos casos en que la gráfica de residuales dan lugar a detectar situaciones ajenas a la linealidad y supuestos.

10.5 Resumen

Resumen modelación estadística

Diagrama de dispersión	Gráfica que se utiliza para representar datos cuantitativos bivariados. En los ejes se describen las variables y los datos son dibujados por puntos.
Tabla de contingencia	Tabla cuyos renglones representan los posibles valores de una variable, y sus columnas describen los posibles valores para una segunda variable. Las celdas en la tabla son el número de veces que cada par de valores ocurre.

Coeficiente de correlación muestral r $-1 \leq r \leq 1$	Cálculo numérico que mide qué tan estrecha es la relación lineal entre dos variables. Si $r = 1$ o próximo a 1, indica que existe una fuerte relación positiva entre las variables. Si $r = -1$ o próximo a -1, indica que existe una fuerte relación negativa entre las variables. Si $r = 0$ o próximo a 0 indica que no existe ninguna relación entre las variables.
Desviación	La diferencia entre un dato observado y el valor predicho por la línea de regresión de mínimos cuadrados.
La técnica de mínimos cuadrados	Esta técnica encuentra la ecuación de la recta que minimiza la suma de cuadrados entre el valor observado y la línea.
$\widehat{Y}$	$\widehat{Y}$ (Y gorro) es el valor predicho de Y para un valor seleccionado de X.

10.6 Complemento didáctico

Aplicaciones de la estadística

Un elemento didáctico que resulta de interés es conocer aplicaciones de la estadística en estudios reales. En este complemento se presentan varias actividades relacionadas con diferentes tipos de información, cuyo análisis permite identificar los conceptos estadísticos ahí empleados.

10.7 Ejercicios

Ejercicios 1

10.1 Pida a 20 de sus compañeros que se pesen en una báscula y midan su estatura con una cinta métrica.

1. Registre los datos en una tabla.
2. Describa esos datos en un diagrama de dispersión.
3. Dibuje los diagramas de caja para la estatura y el peso, arriba del eje horizontal y del lado derecho

del eje vertical, respectivamente.

4. Trace las medianas para los datos de cada variable, ¿Qué observa?

5. Realice la prueba de las medianas.

10.2 Se ha sugerido que hay una relación entre el número de horas sin dormir y la habilidad para completar una tarea. Para explorar esa idea se toma una muestra de 12 personas. La prueba se organizó de tal manera que cada 3 personas no durmieran durante 15 horas; otras 3, 18 horas; 3 más 21 horas y, finalmente, 3 personas no durmieran por 24 horas. A continuación se les preguntó cuántas tareas pudieron realizar.

Sujeto	1	2	3	4	5	6	7	8	9	10	11	12
Horas	15	15	15	18	18	18	21	21	21	24	24	24
Tareas	13	9	15	8	12	10	5	8	7	3	3	4

1. Construya el diagrama de dispersión. Coloque "Horas sin dormir" en el eje horizontal y "Tareas realizadas" en el eje vertical. ¿Qué observa?

2. ¿Cuál es la forma de la relación?

3. ¿Existe una relación fuerte entre las variables?

4. Use la opción de datos bivariados para el inciso a.

10.3 En la compañía encargada de suministrar energía eléctrica piensan que las variables de consumo de energía (kilowatt hora) y de tamaño (metros cuadrados) de una casa están relacionadas. Con el fin de conocer esta posibilidad se tiene el registro de 12 casas de una zona residencial para el primer bimestre del 2005.

Área	282	250	235	200	180	175
Consumo	1975	1952	1894	1675	1750	1590

Área	160	140	135	120	100	90
Consumo	1320	1210	1089	1122	980	896

1. Dibuje el diagrama de dispersión con estos datos.

2. ¿Podría afirmar que existe una relación entre estas variables?

10.4 Las mujeres comenzaron a participar en las pruebas de 100 metros planos desde los Juegos Olímpicos de 1928. Aunque se sabe que el desarrollo físico de mujeres y hombres es diferente, puede evaluarse la relación entre los tiempos efectuados por ambos géneros. Si se establece una relación entre las variables, ésta puede interpretarse tanto por el paso del tiempo como por la evolución competitiva del ser humano.

En el siguiente cuadro se muestran los datos para mujeres y hombres en la competencia de 100 metros planos a partir de los Juegos Olímpicos celebrados de 1928 hasta los de 2004.

Año	Mujeres	Hombres	Año	Mujeres	Hombres
1928	12.2	10.8	1972	11.07	10.14
1932	11.9	10.38	1976	11.08	10.06
1936	11.5	10.3	1980	11.06	10.25
1948	11.9	10.3	1984	10.97	9.99
1952	11.65	10.79	1988	10.54	9.92
1956	11.82	10.62	1992	10.82	9.96
1960	11.18	10.32	1996	10.94	9.84
1064	11.49	10.06	2000	10.75	9.87
1968	11.08	9.95	2004	10.93	9.85

10.5

1. Trace un diagrama de dispersión (aquí use un diagrama de puntos porque el año será una referencia y, cuando se habla de periodos, se clasifica mejor como una variable cualitativa), donde el eje horizontal corresponda al año y el eje vertical a los tiempos de las mujeres (señale los puntos con un color verde), luego agregue el tiempo registrado por los hombres con color rojo. Escriba sus observaciones de la relación de los tiempos con respecto a los años. ¿Se observa la misma tendencia en los tiempos de mujeres y hombres?

2. ¿Qué tendencia siguen los datos? Al pasar 5 Juegos Olímpicos, ¿por cuánto redujeron los tiempos mujeres y hombres?

3. Dibuje un diagrama de dispersión considerando como variables los tiempos. Anote el tiempo de las mujeres en el eje horizontal y el tiempo de los hombres en el eje vertical. Para cada eje use un rango de valores entre 9.5 y 12.5. Trace una línea a 45 grados que pase por el vértice del cuadrado que está a la derecha y abajo. ¿De qué lado de esta línea quedan los puntos? ¿En qué año se da la marca femenina que supere al mayor tiempo realizado por un hombre?

4. Use la opción datos bivariados para realizar el diagrama de dispersión del inciso c.

Ejercicios 2

10.6 En una revista o un periódico busque un artículo donde haya una relación entre variables. Comente con sus compañeros lo que encontró y preséntelo en clase.

10.7 En un centro educativo están preocupados por mejorar el rendimiento académico de sus estudiantes. La psicóloga de la escuela cree que si un alumno es hábil en lectura su desempeño en matemáticas será mejor. La información permitirá que la dirección corrija los planes de estudios, por lo que es importante establecer si existe una relación entre las calificaciones en lectura y matemáticas. De modo que se con-

sideraron a 20 estudiantes a quienes se les aplicó una evaluación de 100 preguntas de cada tema. Los resultados registrados fueron:

Estudiante	1	2	3	4	5	6	7	8	9	10
Lectura	47	71	64	35	43	60	38	67	56	59
Matemáticas	42	81	68	43	50	75	47	69	57	59
Estudiante	11	12	13	14	15	16	17	18	19	20
Lectura	67	57	69	38	54	76	53	40	47	23
Matemáticas	57	54	75	38	59	63	57	40	52	22

1. Trace un diagrama de dispersión y un diagrama de caja para cada variable.
2. Calcule el coeficiente de correlación r.
3. Los alumnos que obtuvieron 50 puntos o menos en lectura, ¿estuvieron igual o mejor en matemáticas? Los alumnos que obtuvieron 60 puntos o más en lectura, ¿estuvieron mejor en matemáticas?
4. Use su opción Datos bivariados para resolver los incisos a y b.
5. Realice la prueba de la mediana y compárela con el resultado que obtuvo en el inciso b (opcional).

10.8 En muchas escuelas de nivel bachillerato hay talleres para capacitar a los estudiantes en áreas técnicas. Uno de esos talleres es el de mecanografía. Al finalizar el curso se aplica un examen, el cual consiste en evaluar la velocidad (número de palabras por minuto) y la precisión (total de errores) de escritura en máquina. La prueba se aplica a una muestra de 20 estudiantes. Las variables de velocidad y de precisión de escritura se describen a continuación:

Practicante	1	2	3	4	5	6	7	8	9	10
Velocidad	68	72	91	47	52	75	63	55	65	35
Precisión	8	2	14	9	13	12	3	0	14	9
Género	1	1	2	1	2	2	1	1	2	2
Practicante	11	12	13	14	15	16	17	18	19	20
Velocidad	84	45	58	61	69	22	46	55	66	71
Precisión	0	14	14	12	2	2	5	5	13	2
Género	1	2	2	2	1	1	1	1	2	1

1. Si fuera el encargado de contratar a un practicante, ¿cuáles serían los mejores candidatos según los datos obtenidos? Use el diagrama de dispersión para apoyar su respuesta.
2. Calcule e interprete el coeficiente de correlación r.
3. Identifique los puntos en el diagrama de dispersión que corresponden a los números uno y dos, pues ésta es una tercer variable, que corresponde al género, de tal modo que el número uno identifica a una mujer y el dos a un hombre. ¿Qué puede concluir a partir de estos datos?

4. Use la opción Datos bivariados para obtener el coeficiente de correlación y el correspondiente diagrama de dispersión.

10.9 Una clínica pública monitorea con regularidad la presión de la sangre de 12 mujeres hipertensas. Las lecturas de la presión sistólica y diastólica son:

Mujer	1	2	3	4	5	6	7	8	9	10	11	12
Sistólica	148	120	164	140	130	175	142	138	154	180	170	132
Diastólica	82	80	90	80	85	94	90	85	92	110	98	80

1. Elabore un diagrama de dispersión. ¿Qué observa?
2. Calcule e interprete el coeficiente de correlación.
3. Use la opción Datos bivariados para obtener el coeficiente de correlación r y el diagrama de dispersión.

10.10 En un bachillerato están preocupados por la salud de sus profesores, de modo que seleccionaron a 20 de ellos de manera aleatoria para verificar sus niveles de colesterol.

Profesor	1	2	3	4	5	6	7	8	9	10
LDL	110	145	113	128	122	110	145	113	128	122
HDL	30	16	30	33	33	30	36	42	33	39
Profesor	11	12	13	14	15	16	17	18	19	20
LDL	105	112	133	135	133	125	133	125	139	141
HDL	36	35	46	36	35	32	38	32	47	42

1. Dibuje un diagrama de dispersión, ¿nota alguna tendencia de estos datos?
2. Calcule el coeficiente de correlación r.
3. Grafique las medianas para cada variable, ¿qué observa?

10.11 Tipos de colesterol:

HDL=Lípidos de alta densidad: Este tipo de colesterol es el que podríamos decir que es el "bueno"de la familia. Sirve para facilitar el flujo sanguíneo ya que lubrica las paredes de los vasos.

LDL=Lípidos de baja densidad: Podríamos llamarle "malo", puesto que al perder la densidad, queda como si fuera agua sucia con muchas partículas de deshecho en suspensión, las cuales pueden irse adhiriendo a las paredes arteriales.

El colesterol total correspondería prácticamente a una suma de los dos anteriores, y a alguna otra pequeña fracción que no se tiene en cuenta más que en contadas ocasiones.

10.12 Tanto para los partidos políticos como para el instituto encargado de organizar las elecciones, es de suma importancia estudiar la relación entre el porcentaje de votantes registrados y el porcentaje de personas que votaron durante las elecciones intermedias. Por ello, se tomó una muestra aleatoria de 10 municipios y los datos resultantes fueron:

1. Dibuje un diagrama de dispersión. ¿Observa alguna tendencia en estos datos?
2. Calcule el coeficiente de correlación r.
3. A partir de los datos obtenidos, ¿podría decir si las personas que pueden votar muestran interés por las elecciones? ¿Por qué?

10.13 Los precios al consumidor de 20 medicinas en dos farmacias que denominaremos A y B, se muestran a continuación:

Medicina	Farmacia A	Farmacia B	Medicina	Farmacia A	Farmacia B
1	55.65	87.00	11	83.30	130.00
2	134.40	233.00	12	297.50	467.00
3	14.60	24.56	13	204.40	357.68
4	294.70	429.00	14	297.50	467.00
5	75.95	115.00	15	204.40	357.68
6	168.38	288.00	16	71.80	112.75
7	76.90	109.90	17	63.40	96.00
8	30.98	49.00	18	56.93	89.00
9	107.95	176.40	19	179.53	282.53
10	104.63	182.70	20	147.88	243.40

10.14

1. Trace un diagrama de dispersión de la relación de precios entre estas dos farmacias. ¿Existe alguna correlación? ¿Ésta es positiva o negativa? ¿Cómo la interpreta?
2. Trace una recta a 45 grados que pase por la unión de los ejes horizontal (precios de la farmacia A) y eje vertical (precios de la farmacia B) y señala qué observa. Obtenga una conclusión.
3. Trace dos líneas perpendiculares al eje horizontal que pasen por los precios 100 y 200. También trace dos líneas perpendiculares al eje vertical que pasen por los precios 100 y 200. En promedio, ¿cuánto paga de más en la farmacia B cuando el precio de las medicinas en la farmacia A?

   a) Está entre 100 y 200 pesos.

   b) Es más de 200 pesos.

Ejercicios 3

10.15 Grafique en un papel cuadriculado

1. La línea recta: $\widehat{Y} = 5 + 3X$; localice los puntos $X = 1$ y $X = 4$. ¿Cuál es el intercepto? ¿Cuál es la pendiente?

2. La línea recta: $\widehat{Y} = 12 - 2X$; localice los puntos $X = 0$ y $X = 4$. ¿Cuál es el intercepto? ¿Cuál es la pendiente?

10.16 El dueño de una refaccionaría ha determinado que la ganancia mensual (Y) que se obtiene de la venta de acumuladores para una marca específica de autos está dada por la línea ajustada: $\widehat{Y} = -154 + 12X$, multiplicada por mil pesos, donde X denota el número de baterías vendidas en un mes.

1. Si se venden 41 acumuladores en un mes, ¿cuál es la ganancia?

2. ¿Al menos cuántos acumuladores necesita vender para obtener ganancia mensual?

10.17 Los siguientes datos son los valores para el número de semanas que un estudiante está en un curso de lectura rápida y la rapidez que adquiere en el número de palabras que lee por minuto.

Semana	X	2	3	4	6	8
Rapidez	Y	49	86	109	164	193

1. En un papel cuadrícula dibuje un diagrama de dispersión.

2. Trace a ojo la línea que considere como la que mejor representa la relación entre X y Y.

3. Usando esa línea, ¿cuál es la rapidez que el estudiante gana en la semana 5?

4. Si se selecciona una persona de manera aleatoria, ¿quién tiene 3 semana en el curso?, ¿cuál es la rapidez ganada?

5. ¿Cuál es la rapidez ganada al aumentar en una unidad la variable X: una semana?

6. Ahora dibuje la línea: $\widehat{Y} = 10.1 + 23.9X$; usando esta recta repite las preguntas c-e.

7. Discuta cuál de las dos líneas es mejor.

10.18 Identifique la variable X, y la variable de respuesta Y, en las siguientes situaciones:

1. El objetivo de un estudio es relacionar el nivel de monóxido de carbono en la sangre para una muestra de fumadores, a los que se les preguntó por la media de cigarros que fuman al día.

2. Una analista de mercado desea relacionar los gastos que se realizaron haciendo la promoción de un producto y la cantidad posterior de la venta del producto.

3. Un agrónomo investiga la razón de crecimiento de un hongo en relación al nivel de humedad del medioambiente.

10.19 Un sociólogo en una universidad está interesado en conocer el salario que perciben sus egresados años después de haberse recibido.

Años	X	5	6	7	8	9	11
Salario	Y	25470	26280	26770	27530	28270	30050

1. ¿Cuál es la variable de entrada y cuál la de respuesta?
2. En un papel cuadriculado describa el diagrama de dispersión para los datos.
3. ¿Existe una relación lineal entre el número de años de egresado y el salario? Si es así, describa la relación.
4. Encuentre la línea de regresión de mínimos cuadrados.
5. Explique qué le indica este modelo al sociólogo sobre los años de egresado y el salario.

10.20 Una agencia de viajes está interesada en conocer cómo las tarifas aéreas, en pesos, se relacionan a las longitudes de los vuelos, en kilómetros. La agencia plantea la siguiente hipótesis: que a mayor distancia de vuelo, es mayor la tarifa aérea. Los datos recabados por la agencia son:

Longitud	X	3560	2100	1875	3490	1480	3075
Tarifa	Y	4558	2885	2670	4473	2190	3954

1. ¿Cuál es la variable de entrada? ¿Cuál es la variable de respuesta?
2. Construir un diagrama de dispersión.
3. A partir del diagrama, ¿se puede concluir que la hipótesis de agencia es correcta?, ¿por qué?
4. Escriba la ecuación de la línea de regresión de mínimos cuadrados.

10.21 Se estudia el tiempo de vida (en horas) de seis componentes electrónicos sometidos a diferentes temperaturas (^0C). Los datos son:

Temperatura	X	50	100	150	200	250	300
Tiempo de vida	Y	875	884	762	424	365	128

1. Encuentre la línea de regresión de mínimos cuadrados.
2. Utilizando la línea de regresión ajustada, predecir el tiempo de vida si la temperatura es $210(^0C)$.

10.22 En un proceso químico se realiza un estudio experimental para conocer la relación entre el número de impurezas Y y la temperatura $X(^0C)$ en la fase de reacción del proceso. Se reporta el siguiente resumen estadístico:

$$n = 9, \sum_{i=1}^{9} X_i = 1260, \quad \sum_{i=1}^{9} Y_i = 312, \quad S_x = 6000, \quad S_y = 2350, \quad S_{xy} = -3710$$

1. Encuentre la línea de regresión de mínimos cuadrados.

2. Utilizando la línea de regresión ajustada, predecir el tiempo de vida si la temperatura es $160(^0C)$.

10.23 En la materia de comprensión de lectura se registran las calificaciones en una muestra de 16 estudiantes para observar si existe una relación entre los resultados a la mitad del semestre (X) y el finalizar el semestre (Y).

X	81	75	71	61	96	56	85	18	70	77	71	91
Y	80	82	83	57	99	30	68	56	40	87	65	86
X	88	79	77	68								
Y	82	57	75	47								

1. En una hoja cuadriculada elabore el diagrama de dispersión trace a ojo la línea recta que considere mejor se ajuste a estos datos. Usando ésta prediga el valor de $\widehat{Y}$ si $X = 60$.

2. Determine la ecuación de la línea de regresión de mínimos cuadrados y use ésta para predecir el valor de la calificación final si $X = 75$.

Ejercicios 4

10.24 El tipo de cambio del peso mexicano frente al dólar durante los últimos once años ha variado sobremanera. En este caso, se establece una relación entre la variable "Años" y la variable "Tipo de cambio". A continuación se presenta esa información del año 1994 al año 2004 tomando como fuente el Banco de México del año 1994 al año 2004.

Año	94	95	96	97	98	99	00	01	02	03	04
Tipo de	3.39	7.00	7.59	7.95	9.24	9.56	9.50	9.32	9.75	10.75	11.39
Cambio											

1. Construya el diagrama de puntos, con el año en el eje horizontal y el tipo de cambio en el eje vertical. ¿Qué observa?

2. ¿Cuál es la tendencia de la relación entre los años y el tipo de cambio?

3. ¿Existe una relación fuerte entre las variables?

10.25 Al finalizar el año escolar, un profesor acostumbra pedir a sus estudiantes que evalúen su sistema de enseñaza. La escala de evaluación es A: Excelente, B: Muy bueno, C: Bueno, D: Regular y E: Malo. Además, les pide que anexen el promedio de la calificación final que obtuvieron en su materia durante ese año. Abajo se presenta la información que registró de una muestra de 16 alumnos, donde las variables son: evaluación del sistema de medición (cualitativa) y calificación (cuantitativa).

Evaluación	B	C	C	A	B	D	C	C
Calificación	8.2	7.4	8.6	9.0	8.6	6.2	8.2	9.0

Evaluación	C	E	A	C	D	B	B	E
Calificación	7.0	5.4	8.6	8.2	7.8	7.4	9.0	6.6

1. Grafique estos datos asignando la evaluación de los estudiantes en el eje horizontal y la calificación en el eje vertical.

2. Comente sus observaciones del diagrama de puntos con tus compañeros.

10.26 La administración quiere establecer tiempos de garantía para ello realiza un prueba sobre las denominadas baterías. Se ha probado que el tiempo de vida, medido en días, de un acumulador se puede predecir (por ejemplo) midiendo la carga que se le proporciona a la batería (en miliamperes entre hora, mAh). Se realiza una prueba forzando las condiciones normales del funcionamiento de seis acumuladores para predecir el tiempo de vida ante esta situación. Los datos para seis acumuladores son:

Corriente	X	17.9	23.6	30.9	56.1	61	77
Tiempo de vida	Y	245	220	215	211	161	135

1. Encontrar la línea de regresión de mínimos cuadrados para este conjunto de datos. ¿Qué tanto cambia el tiempo de vida al cambiar la corriente en una unidad? Predecir el tiempo de vida si la corriente es de 20 mAh.

2. Reproduzca este ejercicio en el CalEst, complete la evaluación estadística e interprete.

10.27 Se desea estudiar cómo afecta el fumar durante el embarazo al peso de un recién nacido. En una muestra aleatoria de 16 mujeres fumadoras que dieron a luz, se les pidió que proporcionaran una estimación del número de cigarros en promedio que fumaban al día y el peso (en gramos) de sus bebés al nacer. Los datos registrados fueron:

Cigarros	22	16	4	19	42	8	12	30
Peso al nacer	2900	3260	3670	3120	2760	3800	3440	2940

Cigarros	14	16	5	20	32	2	15	48
Peso al nacer	3800	3670	3850	2990	2720	3580	3210	2490

Mostrar una representación gráfica de estos datos. Obtener la línea de regresión de mínimos cuadrados. Predecir el peso de un recién nacido si la madre fuma en promedio 28 cigarros.

Ejercicios 5

10.28 En el ejercicio 2 de las actividades de aprendizaje 4.4. se tiene el registro del profesor para 16 estudiantes. En las variables Evaluación y Calificación, en la tabla de abajo se presenta la información

que el profesor ha registrado para 82 estudiantes.

A	B	C	D	E
8.5, 7.7, 7.8,	7.2, 8.4, 7.9,	7.0, 6.6, 6.8,	4.6, 6.4, 5.7,	3.3, 3.6, 3.9,
7.9, 9.9, 8.9,	7.8, 7.1, 9.1,	7.9, 7.8, 5.8,	6.1, 4.4, 5.2,	4.3, 4.1, 4.1,
8.9, 8.1, 9.3,	7.9, 7.7, 7.8,	7.3, 7.2, 7.1,	5.9, 6.2, 5.7,	7.6, 4.2, 4.2,
8.9, 8.4, 8.1,	7.9, 6.8, 8.5,	6.6, 6.2, 7.9,	5.9	4.6, 4.8, 3.8
7.9, 7.8, 8.1,	7.6, 6.5, 7.8,	7.1, 8.3, 5.8,		
8.4, 8.1, 8.8,	7.2, 7.6, 6.9	5.7, 6.2, 6.3,		
9.5, 9.6		7.9, 6.6, 8.1,		
		6.7		

Evaluación: A: Excelente, B: muy Bueno, C: Bueno, D: Regular, E: Malo

1. Trace el diagrama de puntos. ¿Qué observa?

2. En una nueva gráfica, anexe los diagramas de caja para cada letra de la evaluación. ¿Qué puede concluir?

10.29 En los reportes que el Instituto Nacional de Estadística, Geografía e Informática (INEGI) presenta anualmente, se muestra la distribución de establecimientos para espectáculos. Éstos incluyen establecimientos de tipo teatral, deportivo, taurino y recreativo. En la siguiente tabla se indica el año y el número de establecimientos registrados. Haga un diagrama de puntos para esos datos y comente sus observaciones.

Año	1991	1993	1995	1997	1999	2001	2002	2003
Número	218	318	392	443	405	428	429	561

10.30 Los tiempos (en minutos) de la prueba de maratón para hombres registrada en la historia de los Juegos Olímpicos se muestra a continuación. Las variables de la muestra son "Año en que se realizó la prueba" y "Tiempo registrado"

Año	1900	1904	1908	1912	1920	1924	1928	1932
Tiempo	179.5	208.5	175.2	156.6	152.4	161.2	152.6	151.4
Año	1936	1948	1952	1956	1960	1964	1968	1972
Tiempo	149.2	154.5	143.0	145.0	135.2	132.1	140.3	132.2
Año	1976	1980	1984	1988	1992	1996	2000	
Tiempo	129.6	131.0	129.2	130.3	133.2	132.4	130.1	

1. Trace un diagrama de puntos.

2. Indique y comente la tendencia de la relación entre estas variables.

3. ¿En qué año se obtuvo la mejor marca?

10.31 Una tienda tiene 8 grupos de precios para películas en DVD y Blueray. Los siguientes datos muestran el número de ventas y el precio de cada grupo.

Número de ventas	520	450	480	510	490	450	560	510
Precio	155	160	165	170	150	175	145	140

1. Realice una estimación para encontrar un estimado de la varianza de los términos de error de la línea de regresión de la población.
2. Realice una estimación para encontrar la estimación por mínimos cuadrados de la pendiente de la línea de regresión de la población.
3. Encuentre un intervalo de confianza del 90 % de la pendiente de la línea de regresión de la población.

10.32 Una compañía quiere saber qué tan efectivo es promocionar sus productos en internet. El experimento lo realiza comparando el porcentaje de cambio en el gasto de publicidad en contraste con el porcentaje de aumento de las ventas de sus 10 productos, X : Aumento en el gasto de publicidad en internet (%), Y : Aumento en las ventas (%). Los valores medidos fueron:

X	1	5	13	10	9	12	2	4	6	1
Y	2.5	7.2	9.9	8.8	10.4	10.9	4.5	2.6	7.6	3.1

1. Usando mínimos cuadrados, estime la regresión lineal del incremento en ventas en relación al aumento del gasto en publicidad.
2. Encuentre el intervalo de confianza del 90 % para la pendiente de la línea de regresión de la población.

Ejercicios 6

10.33 Se realizó una encuesta para evaluar la satisfacción de los usuarios de diferentes servicios. Los entrevistados deberían contestar si el tipo de servicio ofrecido fue excelente, bueno, regular o malo. La siguiente tabla recoge la información que se obtuvo de la encuesta

Evaluación del servicio	Larga distancia	Llamada local	Potencia	Tv Cable	Teléfono celular
Excelente	264	444	131	215	29
Bueno	934	981	398	378	524
Regular	242	156	102	66	198
Malo	123	98	24	49	14

10.34

1. Estime los totales para cada renglón y cada columna e interprete los valores.
2. Estime las proporciones de toda la tabla y de algunas interpretaciones de sus resultados.

3. Estime las proporciones por renglón (Evaluación del servicio). Represente sus resultados en un diagrama de barras y explique sus conclusiones.

10.35 Existe interés en conocer si hay alguna relación en el rendimiento de los estudiantes en el examen de selección para ingresar a la preparatoria con el promedio en la materia de matemáticas X_1 y la de español X_2 que obtuvieron en los tres años de secundaria, el rendimiento se evaluó por el número de respuestas correctas en el examen. Los resultados de nueve estudiantes seleccionados al azar son los mostrados en la siguiente tabla, aciertos en el examen de selección X_1 y el promedio en secundaria X_2.

$Y = aciertos$	89	78	66	62	52	49	47	44	38
$X_1 = mat.promedio$	78	89	70	74	72	68	70	65	62
$X_2 = esp.promedio$	89	81	80	72	70	67	68	69	68

10.36 En un proceso se quiere conocer la relación que existe entre el tiempo de mezclado y la velocidad del equipo con la densidad. Un ingeniero realiza varias pruebas, sus resultados se muestran en la siguiente tabla:

prueba	tm	vel	den	prueba	tm	vel	den
1	5	100	3.1	11	8	200	3.2
2	5	100	3.3	12	8	200	3.5
3	5	200	2.6	13	9	100	2.8
4	5	200	2.4	14	9	100	2.6
5	7	100	2.5	15	9	200	3.1
6	7	100	2.6	16	9	200	3.0
7	7	200	3.0	17	10	100	3.2
8	7	200	3.3	18	10	100	3.4
9	8	100	2.4	19	10	200	2.5
10	8	100	2.3	20	10	200	2.4

Proponga un modelo de regresión múltiple y realice un análisis estadístico completo para evaluar el modelo.

10.37 Un investigador estudia el efecto de la razón de carga (X_1) y temperatura (X_2) en la vida de un nuevo tipo de celda de poder. Un experimento se realiza para tres niveles de X_1 (6, 1.0 y 1.4 ampers) y de X_2 (10, 20, 30 ^{o}C). Los factores que se refieren a la descarga de la celda Y se midió en términos del número de ciclos de carga-descarga en que la celda se mantiene antes de fallar. Los datos se muestran en la siguiente tabla:

razón de carga	tem			número de ciclos
X_1	X_2	x_1	x_2	Y
0.6	10	-1	-1	150
1.0	10	0	-1	86
1.4	10	1	-1	49
0.6	20	-1	0	288
1.0	20	0	0	157
1.0	20	0	0	131
1.0	20	0	0	184
1.4	20	1	0	109
0.6	30	-1	1	279
1.0	30	0	1	235
1.4	30	1	1	224
$\overline{X}_1 = 1$	$\overline{X}_2 = 20$			

Se propone el modelo: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$. Obtenga:

1. La tabla de coeficientes.
2. La tabla del ANDEVA, la tabla de suma de cuadrados condicionales.
3. Obtenga el valor de la varianza y de los coeficientes de determinación.

10.38 Se somete un material a cierta temperatura, se desea conocer las impurezas generadas en un proceso químico. A continuación se muestran los datos.

X Temperatura (^0C)	Y Impurezas (%)
90	18.4, 17.6, 18
100	11.7, 10.3
110	7.7, 8.3
120	6.5, 6.7
130	6.6, 7.2, 6.7

1. Escriba el modelo lineal.
2. Haga la prueba de hipótesis sobre los parámetros del modelo.
3. Planteé y pruebe la hipótesis sobre el modelo, ¿qué puede concluir?
4. Interprete el coeficiente de determinación.
5. Interprete el valor de la estimación del parámetro de la temperatura.
6. Estime el error estándar del parámetro de la variable temperatura.

7. Observe la gráfica de residuales, ¿qué puede concluir?

8. Agregando al modelo el término cuadrático: Use el módulo de regresión avanzada, modelo polinomial de orden 2, ¿qué puede concluir?

Parámetro	Coeficiente	error std	estadístico t	p
coeficiente	176.834	11.850	14.923	0.000
X	-2.802	0.219	-12.811	0.000
X^2	0.012	0.0009	11.588	0.000

Índice analítico