Carlos Véliz Capuñay

Estadística para la administración y los negocios

PEARSON

CENTRUM
CENTRO DE NEGOCIOS
PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Estadística para la administración y los negocios

Autor

Carlos Véliz Capuñay
Doctor en Ciencias y Magíster en Matemáticas
Pontificia Universidad Católica del Perú
Doctor en Ingeniería Industrial
Universidad Nacional Mayor de San Marcos, Perú

Revisión Técnica

Juan Carlos Del Valle Sotelo
Magíster en Ciencias
Tecnológico de Monterrey
Campus Estado de México

Prentice Hall

Perú • Argentina • Bolivia • Brasil • Chile • Colombia • Costa Rica
España • Guatemala • México • Puerto Rico • Uruguay • Venezuela

Datos de catalogación bibliográfica

VÉLIZ CAPUÑAY, CARLOS

Estadística para la administración y los negocios

Primera Edición

Pearson Educación, México, 2011

ISBN: 978-607-32-0142-1

Área: Administración y finanzas

Formato: 18.5 x 23.5 cm Páginas: 472

GERENTE EDITORIAL:
Clara Andrade

COORDINADORA EDITORIAL:
Carla Soto
carla.soto@pearsoned.cl

EDITOR:
Daniel Soria

CORRECCIÓN DE TEXTOS:
Daniel Soria / Alejandro Sabag

DISEÑO Y DIAGRAMACIÓN:
Rosario Capuñay R.

DISEÑO DE PORTADA:
Pablo Araya

Primera Edición, 2011

ISBN: 978-607-32-0142-1
Impreso en México / *Printed in México*

Prentice Hall
es una marca de

PEARSON

A Mirtha,
José Carlos
y Jessi

Índice

Acerca del autor

Carlos Véliz Capuñay es doctor en Ciencias, Pontificia Universidad Católica del Perú; doctor en Ingeniería Industrial, Universidad Nacional Mayor de San Marcos; magíster en Matemáticas, Pontificia Universidad Católica del Perú; y bachiller en Matemáticas, Pontificia Universidad Católica del Perú. Ha realizado además estudios de Estadística Matemática en la Universidad de Grenoble, Francia.

Ha escrito libros de análisis matemático y estadística. Es consultor en análisis de datos y modelos estadísticos en general de diversas entidades particulares y del Estado.

Ha realizado asesorías para diversas empresas públicas y privadas como Booz Allen & Hamilton S. A., Banco Mundial, Superintendencia de Banca y Seguros del Perú, Oficina Nacional Previsional del Perú, entre otras. Es especialista en análisis de datos, desarrollo de modelos estadísticos y desarrollo de modelos actuariales.

Actualmente, se desempeña como profesor principal en la Pontificia Universidad Católica del Perú en la Facultad de Ciencias e Ingeniería y en el Posgrado de Estadística, en la rama de Estadística. Es coordinador del diploma de especialización en Estadística Aplicada de la Pontificia Universidad Católica del Perú. Es investigador y profesor en el Área Académica de Operaciones de la escuela de negocios de la PUCP, CENTRUM Católica.

Introducción

Una vasta experiencia en la enseñanza universitaria y en la aplicación de la estadística en diferentes campos de la ciencia, tecnología y empresarial ha sido el impulso para escribir esta obra.

Con este libro se trata de introducir al lector en los conceptos básicos de la estadística, utilizando una serie de ejemplos para que los estudiantes de diferentes campos, como el de la administración y negocios, comprendan el papel relevante que tiene esta ciencia en el planteamiento de estrategias que permiten el mejoramiento de los procesos.

En la actualidad, las empresas disponen de gran cantidad de datos producto del desarrollo de diferentes instrumentos de medición y de la influencia de la tecnología de la información; sin embargo, estos datos, para ser transformados en información, deben ser analizados, y en este sentido ayudan los procedimientos estadísticos que en este texto se desarrollan.

El texto comprende tres partes básicamente. En la primera parte se desarrollan procedimientos de la estadística descriptiva, poniéndose énfasis en los diferentes métodos que permiten el análisis exploratorio de los datos. La primera parte comprende los capítulos 1, 2, 3 y 4. La segunda parte comprende los capítulos 5 y 6, y se refiere a los principios básicos de la teoría de la probabilidad y los modelos probabilísticos más importantes para la construcción de

los modelos estadísticos. Los temas que corresponden a la tercera parte son: la estimación de parámetros, las pruebas de hipótesis y el estudio de algunos modelos, como la regresión y los diseños experimentales. Estos temas se encuentran en los capítulos 7, 8 y 9.

En el texto también se introducen algunos temas complementarios muy importantes en el análisis de los datos; estos temas son: simulación, la teoría de decisión y una introducción al muestreo estratificado.

Agradezco el apoyo que CENTRUM CATÓLICA me ha brindado para la realización de esta tarea.

Carlos Véliz Capuñay
Profesor principal PUCP

CAPÍTULO 1

Descripción de los datos

En la actualidad, con la invención de una variedad de instrumentos de medida y el uso de diferentes medios de comunicación, es posible acceder a una gran cantidad de datos. En particular, las empresas tienen la oportunidad de transformar estos datos en información para la toma de decisiones; sin embargo, existe aún falta de conocimiento de las técnicas estadísticas que son necesarias para estos fines. En este texto se presenta una serie de métodos estadísticos que ayudan en este sentido.

En este capítulo se muestran los tipos de datos con los que trata la estadística, y que se presentan en diferentes campos, en particular en el terreno de la administración y los negocios. Aquí se ofrecen técnicas descriptivas y gráficas interpretables que ayudan en la exploración de los datos para convertirlos en información.

CONTENIDO

1.1 La estadística: ciencia de la variabilidad

A menudo escuchamos o leemos frases tales como:

"La producción en el sector agropecuario creció 5.8% en 2002".

"En este banco los préstamos se otorgan en menos de 24 horas, en promedio".

"Las lámparas que vendemos tienen una duración promedio de 1,500 horas".

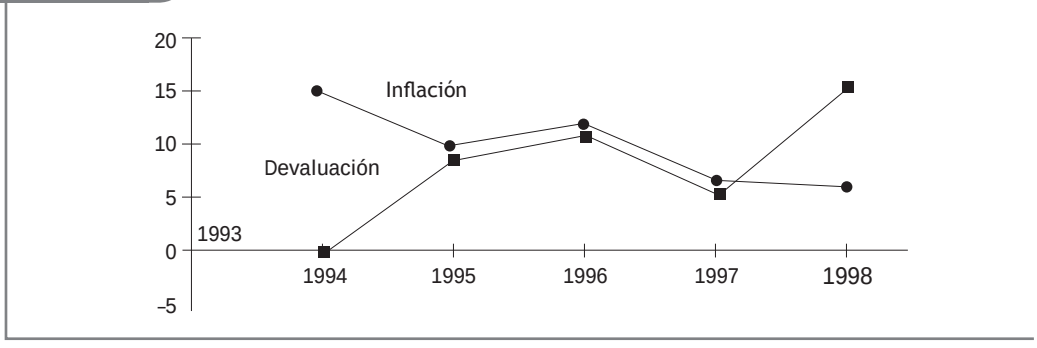
"Los reclamos son atendidos en 24 horas, en promedio".

"En esta ciudad el promedio de la edad de las mujeres es 25 años".

En diarios de la capital han aparecido ilustraciones como la siguiente:

Según las fuentes que proporcionaron el Banco Central de Reserva y el Instituto Nacional de Estadística e Informática, los índices de devaluación e inflación durante los años 1994,..., 1998, fueron como aparece en la siguiente figura.

FIGURA 1.1 *Inflación – Devaluación*



Estas afirmaciones y gráficos se basan en resultados de carácter numérico llamados *datos*, que tienen a la *variabilidad* como característica esencial, y que generalmente surgen de situaciones reales de incertidumbre o al observar alguna característica de los objetos en los cuales se tiene interés. La estadística proporciona las herramientas y modelos para el análisis de este tipo de datos, *permitiendo obtener y entender la información contenida en ellos*. Los datos con los que trata la estadística se producen en casi todas las ramas de la ciencia y la tecnología, de ahí su importancia en todos los campos del conocimiento. El manejo científico de la información estadística es una característica de nuestro tiempo. Se espera que en un futuro cercano la estadística sea necesaria como la lectura y escritura de nuestro lenguaje habitual.

> *Los datos con los que se trabaja en estadística provienen de fenómenos y experiencias cuyos resultados pueden ser diversos; es decir, tienen variabilidad. El estadístico estudia procedimientos o métodos para recolectar, presentar y analizar datos.*

Los artículos producidos por una fábrica no siempre tienen las mismas características, tienen variabilidad; por esta razón no necesariamente cumplen el mismo nivel de exigencia. Por ello es necesario controlar la variabilidad dando lugar a lo que se llama "control de calidad del producto".

No todas las personas viven la misma cantidad de años, existe variabilidad en los tiempos de vida. Estudiando esta variabilidad, los especialistas en seguros pueden establecer, por ejemplo, los precios de las pólizas de seguros de vida.

El tiempo necesario para satisfacer la demanda de un determinado servicio no siempre es constante, varía. Se estudia la variabilidad de estos tiempos para tratar de disminuirlos y lograr mayor satisfacción de la clientela.

En resumen, la variabilidad aparece en muchos procesos. La finalidad del análisis estadístico es explicarla.

EJEMPLO. *Tipos de empresas. Variabilidad*

En la Tabla 1.1 se observan datos que corresponden a 15 empresas nacionales. No todas las empresas son industriales, no todas tienen 50 empleados y no todas han ganado 2 millones de dólares.

TABLA 1.1 *Datos de 15 empresas nacionales*

Empresas	*Tipos de empresas*	*Número de empleados*	*Ganancias anuales en miles de dólares*
1	Servicios	50	2,000
2	Minera	1,000	80,000
3	Industrial	65	3,000
4	Minera	400	50,000
5	Servicios	45	4,000
6	Industrial	300	3,500
7	Servicios	35	2,500
8	Agrícola	80	1,000
9	Servicios	45	2,000
10	Servicios	35	2,500
11	Minera	600	60,000
12	Servicios	35	2,000
13	Agrícola	100	1,800
14	Industrial	280	3,500
15	Industrial	500	1,800

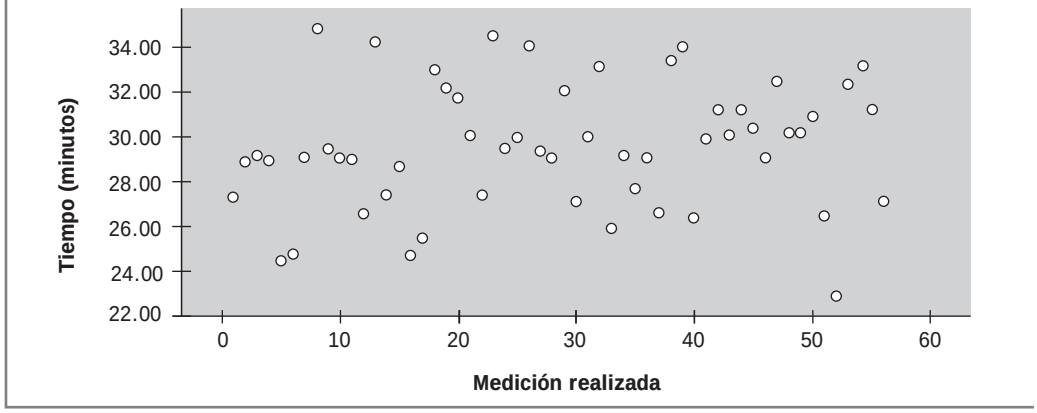
EJEMPLO. *Variabilidad de los tiempos de uso de una cabina de Internet*

El siguiente es un conjunto de datos que corresponden a los tiempos, en minutos, que utilizaron 56 usuarios de una cabina de Internet. Los tiempos fueron registrados de manera secuencial.

27.27	29.44	25.43	29.96	25.85	29.90	30.18
29.89	29.05	33.04	34.09	29.16	31.20	30.92
29.17	28.98	32.21	29.36	27.66	30.08	26.42
28.92	26.52	31.75	29.04	29.06	31.22	22.82
24.43	34.28	30.04	32.09	26.58	30.38	32.38
24.70	27.38	27.36	27.06	33.44	29.05	33.23
29.05	28.66	34.55	29.99	34.06	32.48	31.23
34.90	24.64	29.20	33.17	26.33	30.20	27.09

Se observa que 23 de las 56 personas (41.07%) usaron Internet más de 30 minutos.

FIGURA 1.2 *Variabilidad*



En la Figura 1.2 se representan los tiempos de utilización de Internet. En el eje horizontal se indica el número de la medición realizada y en el eje vertical la medición respectiva. Se observa que los tiempos de utilización varían alrededor de 30 minutos, aproximadamente.

La variabilidad que se observa alrededor de 30 minutos también puede representarse usando un modelo como el siguiente:

$$Dato = 30 + residuo$$

El valor 30 puede considerarse como la parte *regular* del modelo mientras que el *residuo* es la parte *aleatoria.*

1.2 Estadística descriptiva y estadística inferencial

En este primer capítulo se desarrollan técnicas esenciales de la *estadística descriptiva*, métodos que tienen que ver con la codificación, tabulación, representación gráfica y la síntesis numérica de datos que presentan variabilidad. Las técnicas que se estudian sirven para organizar y presentar los datos antes de pretender cualquier interpretación. Se hace uso de este tipo de estadística cuando se dice que 30 de los 100 miembros que componen la empresa SGS son ingenieros de sistemas. Esta afirmación está referida a los 100 miembros de esa empresa, pero de ninguna manera sugiere que el 30% de los miembros de *todas* las empresas del país son ingenieros de sistemas.

Sin embargo, la estadística descriptiva ayuda muchas veces a la identificación de modelos teóricos que sirven para extender las propiedades que se observan en los datos recolectados a conjuntos más amplios, de los cuales estos forman parte. La extensión de las propiedades a toda la población cae en el dominio de la *estadística inferencial*. Se usa la estadística inferencial si, por ejemplo, después de tomar una muestra aleatoria de 120 empresas peruanas y encontrar que el 20% de ellas son de servicios, se puede inferir de manera muy aproximada que entre el 18% y el 22% de las empresas en el Perú son de servicios. En la extensión de las propiedades muestrales a toda la población juega un papel muy importante la *teoría de la probabilidad*.

1.3 Datos, población, unidad de análisis, muestra y variable

Los datos que son estudiados en la estadística pueden ser de tipo *cualitativo* o de tipo *cuantitativo*.

Los datos que indican, por ejemplo, el sector (servicios, industrial, etc.) en donde opera una empresa son de tipo *cualitativo*. *Los datos cualitativos tienen que ver con atributos y cualidades*.

En una empresa, los datos que indican las edades de los clientes son de tipo *cuantitativo*.

Los datos *cuantitativos* se refieren a observaciones o mediciones numéricas.

Población, *unidad de análisis* y *variable* son conceptos en los que se basan las técnicas estadísticas para el análisis de datos. Si se estudia la opinión que tienen los ciudadanos de un país respecto de una ley recién promulgada, la población estará formada por todos los ciudadanos del país. Cada uno de los ciudadanos es una unidad de análisis, y la variable está relacionada con la opinión acerca de la ley. Si se desea estudiar la variación de la edad de los habitantes de la región sur, habrá que referirse a la

población formada por los habitantes de la región sur, a las unidades de análisis (cada uno de los pobladores de la región) y a la variable edad. Si se estudia el grado de conservación de los monumentos históricos de una ciudad, la población estará formada por los monumentos históricos de la ciudad, la unidad de análisis es cada uno de los monumentos históricos y la variable es el grado de conservación. Si para esta misma población y para estas mismas unidades de análisis se estudia la antigüedad que estos tienen, la variable es el tiempo que tiene cada uno de los monumentos.

Se llama *población* o *universo* a cualquier conjunto de elementos de los que se desea obtener *información*.

Una población no necesariamente está formada por personas, y en ella pueden definirse una o más variables.

A cada elemento de la población se le denomina *unidad estadística* o *unidad experimental*.

Dependiendo del número de sus elementos, una población puede ser *finita* o *infinita*.

Cuando la población es pequeña (se considera que una población es pequeña si el número de sus elementos es menor o igual que 100,000) es razonable realizar un censo; esto significa, observar cómo es toda la población respecto de la propiedad que se desea estudiar. Sin embargo, en muchos casos las poblaciones no son pequeñas, y realizar un censo es casi imposible, ya sea por falta de tiempo o por escasez de recursos, sobre todo económicos. En tales casos, se desarrollan procedimientos que permiten deducir el conocimiento de toda la población, a partir de un subconjunto de *n elementos*. Cada subconjunto de la población se llama *muestra*.

Las muestras interesantes de la población, desde el punto de vista estadístico, son aquellas cuyos elementos tuvieron igual posibilidad de ser elegidos. Estos subconjuntos se llaman *muestras aleatorias.*

Muchos estudios estadísticos relativos a una población comienzan con la *elección* de una muestra mediante *procesos de muestreo*; luego se describen y se resumen los datos recogidos (*estadística descriptiva*). Posteriormente, los resultados obtenidos de la muestra se generalizan a toda la población (*estadística inferencial*).

Las medidas características de una muestra se llaman *estadísticos*, las medidas características de la población se llaman *parámetros*. El sueldo promedio de un grupo de trabajadores que constituyen una muestra de 200 personas tomadas de todas las empresas del país es un estadístico, mientras que el sueldo promedio de todos los trabajadores de todas las empresas en el país es un parámetro. Otros estadísticos y otros parámetros serán estudiados más adelante.

Variables

Muchos de los conjuntos de datos cuya variabilidad se desea estudiar resultan al asignar a cada uno de los elementos de una población un número real. Los números asignados muchas veces aparecen como resultado de lecturas directas de instrumentos o de codificaciones. Se establece así una correspondencia entre las unidades estadísticas y los números reales, resultando lo que se llama *variable*.

> *Una variable es una correspondencia que a cada elemento de una población le asigna un número real.*

La variable puede interpretarse como una característica o propiedad que difiere o varía de una observación a otra. El uso de una variable permite al investigador "operar" con los conceptos teóricos, después de transformar en números las características de los elementos de una población.

> *El establecimiento de una variable es una de las etapas más importantes de toda tarea estadística. Con este proceso se relaciona el mundo real con la matemática.*

Los números que indica la balanza cuando se mide el peso de un grupo de personas son los valores de una variable que puede denotarse con *Y* y llamarse *peso.*

Si se desea estudiar cómo varía la ocupación de un grupo de personas, podemos asignar, por ejemplo: el número **1** si es obrero, el número **2** si es técnico, el número **3** si es profesional y el número **4** si no tiene ninguna de las ocupaciones. Se tiene así la variable *O* = ocupación, definida en el grupo de personas y cuyos valores son: 1, 2, 3 y 4.

Los datos recolectados y que posteriormente se analizan para obtener información se presentan en tablas como la que aparece a continuación.

En la Tabla 1.2 se presenta una lista de 10 clientes del banco BB que han obtenido préstamos de parte de este. Ahí se indican los valores recogidos de las siguientes variables

Estado civil: S (soltero), C (casado), O (otro)
Sueldo: sueldo mensual
Edad: edad del cliente
Tarjetas: posesión de tarjetas de crédito. S (sí), N (no)
Préstamo: valor del préstamo, en dólares, otorgado por el banco

TABLA 1.2 *Base de datos. Información del banco BB*

Nombres	*Estado civil*	*Sueldo*	*Edad*	*Tarjetas*	*Préstamo*
A. Arce	S	5,400	32	N	50,000
B. Baca	S	6,500	43	S	40,000
C. Casti	C	7,200	45	S	30,000
D. Dejo	S	4,800	34	N	20,000
E. Este	O	3,900	29	S	15,000
F. Farr	C	5,200	45	N	40,000
G. Guer	S	6,000	29	S	30,000
H. Hus	C	5,200	32	S	20,000
I. Irri	C	5,500	37	N	10,000
J. Juan	C	6,500	42	S	15,000

La información relativa a una población se obtiene también mediante cuestionarios. Cada pregunta puede relacionarse con una variable, como en el caso siguiente:

¿Está usted de acuerdo con la ley antitabaco?

Las respuestas pueden ser: sí, no, no sabe/no contesta. El entrevistado debe dar solo una respuesta, la cual puede ser codificada, respectivamente, con los valores 1, 2 y 3, correspondiendo estos valores a una variable que puede llamarse "ley antitabaco".

Observaciones

1. Usualmente, al conjunto de los valores que una variable toma en los distintos elementos de una población se le llama también *población.* Así, podremos decir: "la población formada por los salarios de las 1,000 personas que trabajan en una empresa".
2. Además de las poblaciones formadas por elementos que en su totalidad "existen en la realidad", se considera también poblaciones cuyos elementos no siempre pueden disponerse en una lista. Tal es el caso de la población formada por todos los tiempos necesarios para que los automóviles hechos en la fábrica MG pasen de 0 km/h a 100 km/h.

1.4 Escalas de medida. Tipos de escalas de medida

El establecimiento de una variable, es decir, la asignación de números a los elementos de una población, se hace siguiendo determinadas reglas. Así, a cada persona se le asigna su coeficiente de inteligencia, calculando previamente un número que

corresponde a respuestas ante determinadas situaciones típicas. En general, cuando se asigna un número se dice que se ha establecido una *escala de medida*, que se ha realizado una *medición*.

> *La medición es una versión formalizada de una observación, es la descripción del mundo.*

Generalmente, se consideran cuatro tipos de escalas de medida: *nominales*, *ordinales*, *de intervalo* y *de razón* o *cociente,* y es importante conocerlos, pues permiten una mejor aplicación de los métodos estadísticos.

Escala nominal

> *Se tiene una escala **nominal** cuando el número asignado funciona solo como etiqueta.*

Las operaciones que se realicen con los números asignados, así como el orden que entre ellos se pueda establecer, no tienen significado alguno con relación al atributo que se mide. Con una escala de este tipo se consigue clasificar las unidades estadísticas estudiadas, agrupándolas en clases equivalentes llamadas *modalidades* o *categorías.*

EJEMPLO. *Especialidad de los estudiantes*

> Se considera la población formada por todos los estudiantes de una universidad y se define en esta a la variable "especialidad". Se asigna a los alumnos de Ciencias el número 1, a los de Letras el número 2 y a los de Arte el número 3. Se tiene así una escala nominal con las modalidades: "ciencias", "letras" y "arte". Los elementos que forman la población han quedado clasificados, agrupándolos en tres clases: la clase de ciencias, la de letras y la de arte.
>
> Es cierto que 2 es mayor que 1, pero no se puede hacer la misma comparación entre los alumnos de Letras y de Ciencias.

Escala ordinal

> *Se tiene una escala **ordinal** cuando el orden de los números asignados a las unidades estadísticas refleja diferentes grados de la propiedad o el atributo que es objeto de medición.*

Si se ordena por categorías a los vendedores de una tienda, asignándole el número 1 al empleado que logró las ventas más altas, el 2 al empleado que logró ventas intermedias y el 3 a los que lograron las ventas más bajas, se tendrá una escala ordinal; es decir, una escala en la que los vendedores son categorizados según los tres rangos de venta. Sin embargo, aun cuando $3 - 2 = 2 - 1$, la diferencia entre las

ventas logradas por los vendedores de categoría baja y las ventas realizadas por los de categoría intermedia no es en general igual a la diferencia entre las ventas logradas por los vendedores de categoría intermedia y las ventas logradas por los vendedores de categoría alta.

EJEMPLO. *Escala ordinal. Las escalas tipo Likert y diferencial semántico*

En las ciencias sociales es muy común tratar de medir el sentir que tienen los individuos acerca de productos o servicios. Existen diversas escalas para este propósito; entre estas se tienen las escalas tipo *Likert* y la escala de *diferencial semántico*.

Las escalas tipo *Likert* se usan frecuentemente para medir el sentir de los individuos utilizando frases que resaltan los aspectos positivos y negativos del producto o servicio. El entrevistado deberá indicar su grado de acuerdo o desacuerdo con las frases que se indican.

La escala de *diferencial semántico* se usa generalmente para medir el sentir de los individuos frente a diferentes atributos de un producto o servicio, estableciendo comparaciones al utilizar una escala bipolar.

EJEMPLO. *Calidad del transporte*

Para conocer la percepción de los usuarios de transporte frente al servicio que se brinda, se escribieron las siguientes frases:

Los servicios de transporte son de mala calidad (resalta el aspecto negativo).

Los servicios de transporte ofrecen seguridad (resalta el aspecto positivo).

El entrevistado deberá indicar su grado de acuerdo o desacuerdo con cada una de las frases, señalando el número 1 si está en completo desacuerdo, el 2 si está en desacuerdo,..., y el 5 si está en completo acuerdo.

TABLA 1.3 *Calidad de transporte. Escala de Likert*

Frases	1	2	3	4	5
Los servicios de transporte son de mala calidad					
Los servicios de transporte ofrecen seguridad					

De acuerdo a lo estudiado, la escala de Likert es una escala ordinal.

EJEMPLO. *Calidad de servicios bancarios*

Para estudiar la actitud de los individuos frente a los servicios que se prestan a los clientes en una cierta oficina bancaria, se consideraron los siguientes atributos: amabilidad y ambiente.

Definiendo el polo negativo y el polo positivo de cada atributo:

Nada amables-Muy amables.

Ambiente desagradable-Ambiente agradable.

Los individuos deberán evaluar los atributos dando una puntuación del 1 (para el polo negativo) al 7 (para el polo positivo).

TABLA 1.4 *Diferencial semántico*

Frases	1	2	3	4	5	6	7	*Frases*
Nada amables								Muy amables
Ambiente desagradable								Ambiente agradable

Al igual que la escala de Likert, la escala de diferencial semántico es una escala ordinal.

Escala de intervalo

*La escala **de intervalo** es una escala para la cual el* 0 *no indica ausencia de la propiedad que se estudia, y en donde, además de tener en cuenta el orden de los números asignados a las unidades estadísticas, sus diferencias tienen sentido en la realidad. Las distancias que separan a los valores asignados pueden cuantificarse gracias al establecimiento de alguna unidad de medición estándar (años, horas, kilogramos, centímetros, etcétera).*

Las escalas de Fahrenheit (F) y de Celsius (C) usadas para medir la temperatura son escalas de intervalo. En estas escalas, los números mayores corresponden a mayores temperaturas. Es decir, el orden de los números asignados tiene significado, pero también tienen sentido expresiones como "la temperatura de hoy es 5 grados C más alta que la de ayer", suponiendo, por ejemplo, que la temperatura de ayer fue de 45 grados C y la de hoy es de 50 grados C. La diferencia entre las temperaturas tiene sentido.

Escala de razón

> *La escala de **razón** o **cociente** es una escala para la cual el cero indica ausencia de la propiedad que se estudia, y en donde, además de tener significación las diferencias de los números asignados, sus cocientes o razones reflejan los cocientes entre las cantidades de las propiedades que se miden.*

Para las escalas de razón, además del cociente entre los números asignados, tiene sentido hablar también del orden y las diferencias que se puedan establecer entre ellos. Toda escala de razón tiene las características de las escalas de intervalo y ordinales.

La escala que corresponde a la variable "sueldo" es de razón. Tiene también las características de la escala ordinal. Una persona que gana $ 1,200 tiene mayor "poder adquisitivo" que otra que gana $ 400. Esta escala también tiene las características de la escala de intervalo; la "distancia" que separa a aquellas personas con sueldos de $ 700 y $ 750 es la misma que la existente entre aquellos que ganan $ 320 y $ 370. Por último, si una persona gana $ 2,000 y otra gana $ 1,000, se puede decir que *la primera gana dos veces lo que gana la segunda*. No se puede decir lo mismo con respecto a las escalas de temperaturas Fahrenheit y Celsius. 80 grados F corresponden a 25 grados C, mientras que 40 grados F corresponde a 5 grados C. La razón de 80 a 40 es 2; sin embargo, la razón de 25 a 5 no es 2, es 5. La razón en este caso no tiene sentido.

Los físicos han creado una escala de temperaturas llamada "absoluta" o de "Kelvin". En esta escala el cero es absoluto; esta escala de temperaturas sí es de razón.

1.5 Clasificación de las variables

De acuerdo al tipo de escala que se utiliza para asignar valores a los elementos de una población, una variable puede ser *cualitativa* o *cuantitativa.*

> *Una variable es **cualitativa** o **categórica** si la asignación de valores solo tiene sentido cuando se usa una escala nominal u ordinal.*

El "lugar de procedencia" de un conjunto de personas es una variable cualitativa.

> *Una variable es **cuantitativa** o **numérica** si la asignación de sus valores tiene sentido para la escala de intervalo o para la escala de razón.*

Son variables cuantitativas:

- La variable que a cada persona le hace corresponder su salario.
- La variable que a cada país le asigna su producto bruto interno (PBI).

Las variables cuantitativas se clasifican a su vez en *discretas* y *continuas*.

> *Una variable cuantitativa es **discreta** si el conjunto de los valores que ella puede tomar es finito o infinito numerable* (*un conjunto es infinito numerable si es infinito y sus elementos se pueden contar*). *Es decir, se pueden poner en correspondencia biunívoca con el conjunto de los números naturales 1, 2, 3...*

Las siguientes variables son cuantitativas discretas:

- El número de quejas diarias que se reportan en una empresa por mal servicio en las entregas de los productos a los clientes.
- El número de artículos que compra una persona en un centro comercial.

> *Una variable cuantitativa es **continua** si sus valores pueden ser cualquiera de los valores de un intervalo.*

La variable que indica el tiempo de vida de los aparatos electrónicos es una variable continua. Los valores que puede tomar son no negativos, y por no saber cuál será el mayor, se considerará que estos valores pueden variar en el intervalo $[0, +\infty[$.

Muchos datos que teóricamente corresponden a una variable continua, en la realidad y por las limitaciones de los instrumentos, son tratados como si correspondieran a una variable discreta. El peso de una persona generalmente se expresa en kilogramos y a lo más en décimas de kilogramo por limitaciones de las balanzas que se utilizan para hacer la medición; sin embargo, es una variable continua.

Por otro lado, una variable cuantitativa discreta o continua, en la práctica, también puede ser considerada como una variable cualitativa. Por ejemplo, si la edad de una persona se clasifica en menos de 20 años, entre 20 y menos de 60 años, y más de 60 años, se tendrá que esta variable puede tratarse como cualitativa con las modalidades correspondientes a: “joven”, “adulto” y “anciano”, respectivamente.

> *De acuerdo al papel que cumplen en una investigación, las variables pueden clasificarse en **independientes**, **dependientes** y **de control**.*

Una variable *independiente* es una variable cuyos valores pueden ser manipulados por el investigador. La variable que se mide para observar el efecto de los valores de una variable independiente se llama *variable dependiente*.

Así, para ver cómo influye una droga A en el tratamiento de una enfermedad, el investigador aplica diferentes dosis al enfermo para observar si este mejora o no. La variable cuyos valores son las distintas dosis de droga suministrada es una variable independiente, mientras que la variable que indica el grado de mejoría del enfermo es una variable dependiente. En este caso, los valores de la variable independiente pueden ser manejados por el experimentador.

Las variables de *control* ayudan a comprender la relación entre una variable independiente y una dependiente.

Algunas veces se puede observar que los valores de una variable aumentan (disminuyen) cuando los valores de otra también aumentan (disminuyen); sin embargo, no es suficiente para indicar que la característica que mide una es la causa de la característica que mide la otra variable. En una encuesta se determinó que la altura de una persona y sus ingresos estaban relacionados: la gente más alta ganaba más; sin embargo, esto no fue suficiente como para indicar que la relación era causal. No se necesitaba ser alto para tener mejor remuneración, pues un estudio posterior, en donde se introdujo la variable sexo como variable de control, indicó que los hombres eran más altos que las mujeres y que los ingresos de estas, a pesar de la igualdad de condiciones de trabajo, eran más bajos que los de los hombres. De este modo se determinó que las variables "estatura" e "ingreso" estaban relacionadas por su vínculo con la variable sexo. Existía tan solo una *relación espuria*, de tipo práctico (*relación estadística*), pero no una relación de *causa-efecto* (*relación causal*).

1.6 Análisis exploratorio de datos

Para comenzar a entender "lo que dicen los datos" recolectados es necesario usar técnicas exploratorias que describan de una manera rápida su comportamiento. En la exploración se utilizan tablas, gráficos, resúmenes numéricos, etcétera. La exploración ayuda al investigador en la identificación de algún *modelo probabilístico*. Esto permite en varias ocasiones extender los resultados de una muestra a toda la población (*estadística inferencial*). Muchas veces las suposiciones planteadas se confirman o se rechazan a partir del estudio de gráficos simples o de procedimientos que no necesitan una matemática formal, pero sí una dosis de sentido común y cuidado.

En la sección anterior se introdujo una manera de clasificar a los datos; así se establecieron los datos cuantitativos y cualitativos. La importancia de establecer el tipo de dato de que se trata es de singular importancia, pues de ello dependerá el procedimiento que se usará para el análisis. Los resúmenes numéricos que se puedan establecer cuando se tienen datos cualitativos se refieren al conteo de las observaciones que existen en cada categoría. Las operaciones aritméticas que se pudieran hacer con los números que resultan de las codificaciones que se realizan con los datos cualitativos no tienen significación. Si se trata de datos cuantitativos, las operaciones aritméticas, como la adición, multiplicación y división, que con ellos se puedan realizar, sí tienen sentido.

Existe una serie de paquetes computacionales estadísticos, muchos de los cuales, aparte de realizar diferentes tipos de análisis, tienen excelentes presentaciones de gráficas y cuadros. Se citan algunos de estos: SPSS, SAS, MINITAB, S-PLUS, R, STATGRAPHICS, BMDP, STATPAC, SYSTAT, STATISTICA, etcétera. Otros paquetes,

como el Excel, sin ser de tipo estadístico, permiten también construir tablas, gráficos y realizar análisis estadísticos. Sin embargo, es necesario prevenir al usuario acerca de que muchos inconvenientes pueden surgir cuando no se toman en cuenta las premisas de los procedimientos que se aplican, cuando no se entienden los procedimientos y los modelos o cuando se tienen limitaciones para entender los resultados que se obtienen al aplicar las herramientas computacionales.

Cuando se usan gráficos y tablas para representar los datos, generalmente existe pérdida de información, por lo que se deben tomar las precauciones del caso; sin embargo, esta desventaja resulta compensada frente a la facilidad de interpretación.

Tablas de distribución de frecuencias. Representación gráfica

Las tablas de distribución de frecuencias describen cómo se distribuyen los valores de un conjunto de datos cuando se organizan en clases o categorías. En estas tablas se muestra el número de elementos de cada clase y la proporción que existe en cada una de ellas. Una tabla de frecuencias describe entonces la variabilidad de los datos. Además de las tablas, se usan diferentes tipos de gráficos, que informan de una manera rápida y concisa acerca de la variabilidad de los valores de la variable.

Caso en donde la variable es cualitativa

EJEMPLO. *Empresas que cotizan en la bolsa de valores local*

En la Tabla 1.5 se registran 143 empresas que cotizan en la bolsa de valores local, de acuerdo al sector al que pertenecen.

TABLA 1.5 *Tipos de empresas*

Tipo de empresa	*Frecuencia: número de empresas en el sector*	*Frecuencia relativa: proporción de empresas en cada sector*	*Frecuencia relativa en %: porcentaje de empresas en el sector*
(1) Empresas industriales	53	0.371	37.1
(2) Empresas agrarias	68	0.475	47.5
(3) Empresas mineras	15	0.105	10.5
(4) Empresas financieras	7	0.049	4.9
Total	143	1	100.00

La *frecuencia* (*absoluta*) de cada categoría o modalidad indica las veces que se ha observado dicha categoría en el conjunto de datos considerado. En este ejemplo, las frecuencias absolutas indican el número de empresas que pertenecen a cada sector. De esta manera se observa que 68 de las empresas pertenecen al sector agrario.

La *frecuencia relativa* de cada categoría o modalidad indica la proporción de las veces que se ha observado dicha categoría en el conjunto de datos considerado. Es igual *al cociente entre la frecuencia absoluta de la categoría y el número total de observaciones*. La frecuencia relativa puede expresarse como *porcentaje* y usarse para comparar dos o más distribuciones de datos.

Las frecuencias se representan gráficamente usando barras rectangulares o mediante sectores circulares.

En el gráfico de los sectores circulares de la Figura 1.3, cada sector corresponde a una modalidad y su correspondiente ángulo central es $\theta_i = 360^o n_i/T$, en donde n_i es la frecuencia absoluta de la modalidad y T es el total de datos. La modalidad "empresas industriales", por ejemplo, está representada por un sector circular cuyo ángulo central es igual a

$$(360^o)(53)/143 = 133.42^o$$

En el gráfico de barras, cada barra rectangular con el mismo ancho corresponde a una modalidad; su altura puede ser medida en unidades de frecuencia absoluta o de frecuencia relativa. En la Figura 1.4, la modalidad o categoría "empresas industriales" está representada por una barra vertical de altura igual a 53 unidades.

FIGURA 1.3 *Gráfico de sectores circulares*

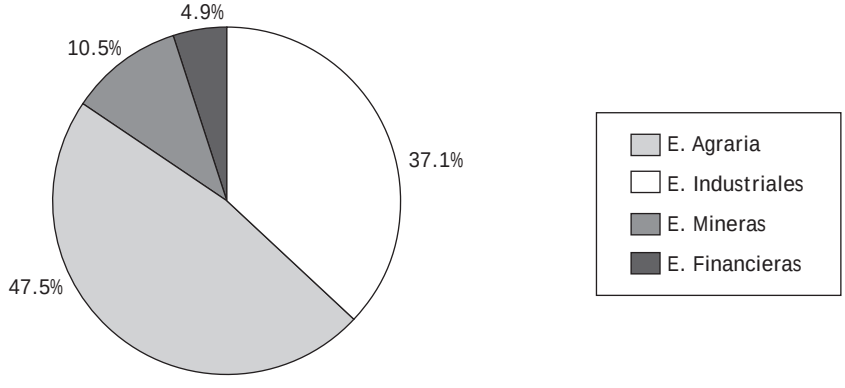
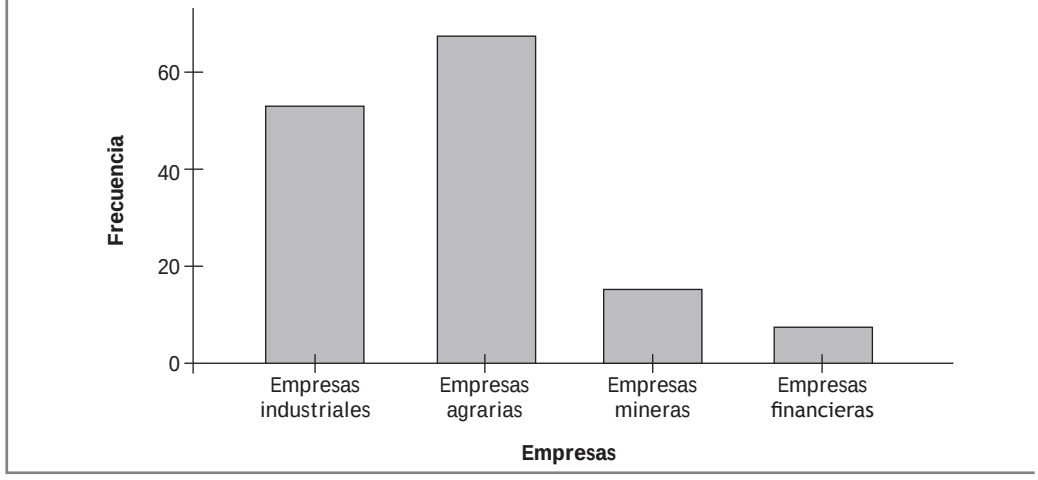


FIGURA 1.4 *Gráfico de barras*



EJEMPLO. *Diagrama de Pareto y el control de calidad*

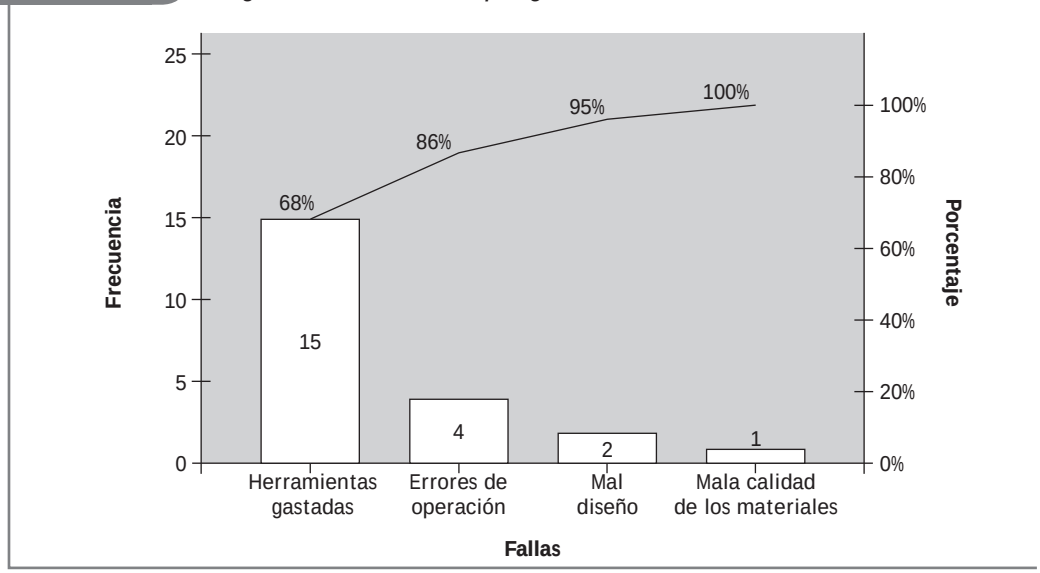
Con el fin de mejorar la calidad de un producto o servicio, se listan las causas que contribuyen a que el producto no cumpla con las especificaciones establecidas. Las frecuencias de estas causas se representan mediante barras. Si las barras se grafican de mayor a menor frecuencia, se obtiene un diagrama llamado de Pareto. Por ejemplo, en la fabricación de un cierto producto se han detectado las siguientes causas de fallas: herramientas gastadas, errores de operación, mal diseño y mala calidad de los materiales. Las frecuencias de estas fallas se muestran a continuación, en la Tabla 1.6.

TABLA 1.6 *Causa de fallas*

Causa de falla	*Frecuencia*	*Frecuencia acumulada*
Herramientas gastadas	15	15 (68.18%)
Errores de operación	4	19 (86.36%)
Mal diseño	2	21 (95.45%)
Mala calidad de materiales	1	22 (100.00%)

La gráfica de las frecuencias se muestra a continuación, en la Figura 1.5, en un diagrama de Pareto.

FIGURA 1.5 *Diagrama de Pareto. El polígono indica frecuencia acumulada*



Este gráfico se llama así en honor a Vilfredo Pareto, economista italiano (1824-1923), quien introdujo un análisis para el estudio de la distribución de la riqueza. Los principios de este análisis se aplicaron posteriormente para estudiar las causas que dan lugar a diversos problemas, estableciéndose que de un grupo de factores que contribuyen a un efecto común, frecuentemente unos cuantos son los responsables de la mayor parte del efecto ("el 80% de los problemas se deben al 20% de las causas").

En este ejemplo, si se pone atención en los primeros tipos de falla (50% de causas de fallas), se podrá reducir por lo menos el 86% de los problemas.

Caso en donde la variable es cuantitativa discreta

EJEMPLO. *Seguridad en la empresa*

El siguiente es un reporte de 105 semanas acerca del número de accidentes, por semana, ocurridos en la fábrica de alimentos Nutre.

0 0 1 0 0 1 3 5 1 1 2 2 2 1 5 3 4 1 1 2 2 2 2 3 3 3 2 2 3 1 1 2 2 4 2 2 2 2 2 4 2 2 2
1 1 4 3 3 3 3 4 3 3 3 2 2 1 3 4 1 1 2 5 2 2 1 3 1 1 3 3 0 0 1 2 3 1 1 2 2 1 1 1 2 2 2
2 2 2 2 2 0 0 2 2 2 2 1 1 2 3 3 5 0 0

De la forma como están presentados los datos no es fácil extraer información. A lo más se puede decir que el número de accidentes en las semanas registradas varía de 0 a 5. Es necesario organizar los datos para así obtener mejor información. Para ello se utilizan las tablas de distribución de frecuencias.

Una tabla de distribución de frecuencias, como la Tabla 1.7, es fácil de construir. En la primera columna se escriben los diferentes valores de la variable que se han observado. En este caso, 0, 1, 2, 3, 4 y 5 accidentes por semana.

La *frecuencia* (absoluta), es decir el número de veces que se repite cada valor de la variable en el conjunto de datos observado, aparece en la segunda columna. Así por ejemplo: el 0 se repite 10 veces (10 semanas sin accidentes), el 1 se repite 25 veces (25 semanas con 1 accidente), etcétera.

En la tercera columna, la *frecuencia relativa* expresa la proporción en que se da cada valor, en relación con el total de los datos. Por ejemplo, en el 38.09% se dan 2 accidentes cada semana.

En la cuarta columna, la *frecuencia acumulada* (absoluta) hasta un determinado valor se interpreta como el número de datos acumulados hasta ese valor. En la quinta columna, la *frecuencia acumulada relativa* hasta un determinado valor indica la proporción de datos acumulados hasta dicho valor. Así por ejemplo, en 75 de las semanas observadas ocurrieron 2 accidentes a lo más y en un tercio de las semanas observadas (0.3333) ocurrió 1 o ningún accidente por semana.

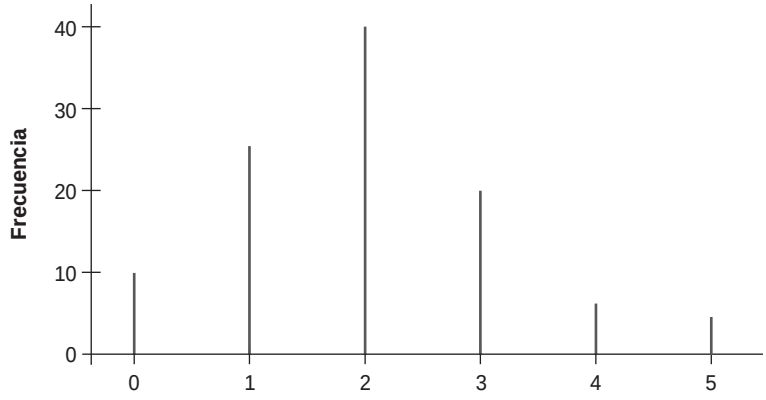
TABLA 1.7 *Distribución de las* 105 *semanas de acuerdo al número de accidentes*

Número de accidentes	*Frecuencia del número de semanas*	*Frecuencia relativa*	*Frecuencia acumulada*	*Frecuencia acumulada relativa*
0	10	0.0952	10	0.0952
1	25	0.2381	35	0.3333
2	40	0.3809	75	0.7142
3	20	0.1905	95	0.9047
4	6	0.0571	101	0.9618
5	4	0.0382	105	1.0000
Total	$n = 105$	1		

La Tabla 1.7 muestra que en la mayor parte de semanas ocurrieron de 1 a 3 accidentes.

En el gráfico de la Figura 1.6 se representa el gráfico de bastones para las frecuencias. Este se usa para representar la columna de frecuencias; indica la forma de la distribución de la variable. En esta gráfica cada "bastón" o segmento de recta tiene una altura proporcional a la frecuencia absoluta del valor de la variable. De igual manera se procede para hacer gráficos de bastones para los otros tipos de frecuencias.

FIGURA 1.6 *Gráfico de bastones para la distribución del número de accidentes por semana*



Caso en donde la variable es cuantitativa y continua

Para organizar conjuntos de valores que corresponden a variables continuas se usan los intervalos de clase. Se procede como en el siguiente ejemplo.

EJEMPLO. *Pagos por siniestros*

El conocimiento de la distribución de los pagos realizados por una compañía de seguros por distintos siniestros ocurridos en un cierto periodo es muy importante por diferentes razones; una de ellas puede ser el establecimiento de un "margen de solvencia" (reservas que la compañía debe tener para hacer frente a los siniestros). Los siguientes son los valores, en unidades monetarias (u.m.), pagados por la compañía de seguros Segur durante el mes de marzo.

10.00	20.30	177.00	94.50	65.70	76.40
60.00	65.60	220.30	40.20	87.20	100.10
236.60	72.30	230.50	92.20	54.90	87.20
65.30	60.20	71.10	99.50	89.40	87.30
36.60	30.20	45.20	88.50	78.90	92.10
129.00	200.30	186.20	213.00	180.90	200.50
134.00	101.20	156.20	123.00	145.20	160.30
121.00	200.10	261.20	67.00	78.10	78.30
178.00	219.30	203.40	78.60	65.20	80.20
78.80	54.30	82.20	35.60	48.20	70.30

La variable en estudio es la variable continua *X = pago por siniestro en u.m.*

Encontramos que esta variable registra valores comprendidos entre 10 u.m. y 261.20 u.m.

Si se construye una tabla de frecuencias para estos datos, tal como se hizo para el caso discreto, se tendría una tabla con 60 filas (una por cada valor), y no brindaría ninguna información respecto de la manera como se distribuyen los valores de la variable. Por ello, y para un mejor estudio, los estadísticos han ideado un procedimiento mediante el cual se agrupan los datos en "*clases*" que se obtienen al dividir el intervalo $I = [10.00, 261.20]$, cuyos extremos son el dato mayor y el dato menor, en subintervalos de igual longitud. Estos subintervalos se llaman *intervalos de clase.*

A la diferencia entre el dato mayor y el dato menor se le llama *rango* de la variable. En este ejemplo, el rango es 261.20 – 10.00 = 251.20.

Si se divide, por ejemplo, el intervalo I en $k = 10$ intervalos de clase de igual longitud, cada uno de estos intervalos medirá $h = \frac{261.20 - 10.00}{10} = 25.12$ unidades, valor que se aproxima a 26 unidades por exceso con la finalidad de cubrir todo el intervalo *I.*

Se conviene en que los intervalos de clase sean abiertos por la izquierda y cerrados por la derecha, a excepción del primero, que será cerrado por ambos extremos. En el ejemplo, los intervalos de clase son:

$$I_1 = [10.00, 36.00] \quad I_2 =]36.00, 62.00] \quad \ldots\ldots \quad I_{10} =]244.00, 270.00]$$

La Tabla 1.8 es un arreglo de la información por columnas. La primera columna está formada por los intervalos de clase antes obtenidos.

En la segunda columna aparecen las llamadas *marcas de clase,* que son los puntos medios de cada intervalo de clase. Cada marca de clase es un representante de los datos que están en el respectivo intervalo de clase. En la tercera columna de la tabla se indica la frecuencia o el número de datos que se observaron en cada intervalo de clase. Así, en el intervalo de clase $I_1 = [10.00, 36.00]$ hay 4 datos. Es decir, en marzo se realizaron 4 pagos mayores o iguales que 10.00 y menores o iguales que 36.00.

En el intervalo]36.00, 62.00] la frecuencia es 8 y la marca de clase es 49.00; entonces, se puede considerar que 49.00 se repite 8 veces.

En la cuarta columna la *frecuencia relativa* indica la proporción de datos en cada intervalo.

En la quinta columna, el valor de la *frecuencia acumulada* hasta un determinado intervalo de clase se interpreta como el número de datos acumulados hasta ese intervalo.

En la sexta columna, el valor de la *frecuencia acumulada relativa* hasta un determinado intervalo de clase indica la proporción de datos acumulados hasta el intervalo respectivo.

TABLA 1.8 *Tabla de frecuencias que indica la distribución de los* 60 *pagos realizados en marzo*

Montos pagados en u.m. durante el mes de marzo	*Marcas de clase* x_i	*Frecuencia de pagos* n_i	*Frecuencia relativa* f_i	*Frecuencia acumulada* N_i	*Frecuencia acumulada relativa* F_i
[10.00, 36.00]	23.00	4	0.067	4	0.067
]36.00, 62.00]	49.00	8	0.134	12	0.201
]62.00, 88.00]	75.00	19	0.317	31	0.518
]88.00, 114.00]	101.00	8	0.134	39	0.652
]114.00, 140.00]	127.00	4	0.067	43	0.719
]140.00, 166.00]	153.00	3	0.050	46	0.769
]166.00, 192.00]	179.00	4	0.067	50	0.836
]192.00, 218.00]	205.00	5	0.080	55	0.916
]218.00, 244.00]	231.00	4	0.067	59	0.983
]244.00, 270.00]	257.00	1	0.017	60	1.000

La pregunta que a menudo se plantea tiene que ver con el número de intervalos de clase a usar. Si se eligen pocos intervalos, la longitud de cada intervalo de clase resulta grande, y las marcas de clase no siempre son buenas representaciones de los valores que están en el intervalo, perdiéndose considerable información. Si se eligen muchos intervalos, la longitud de cada uno de ellos resulta pequeña, y, en este caso, la organización de los datos puede requerir de mayor trabajo, con el riesgo de realizar interpretaciones erradas de los resultados. Existen diferentes reglas para determinar el número de intervalos de clase. Una de ellas es la regla de Sturges, que recomienda tomar el mayor entero próximo al valor $1+3.3 \log T$ como número máximo de intervalos, si el número de datos T es una potencia de 2.

La Tabla 1.9 muestra otra regla práctica para determinar el número de intervalos.

TABLA 1.9 *Regla práctica para determinar el número de intervalos a usar en una tabla de distribución de frecuencias*

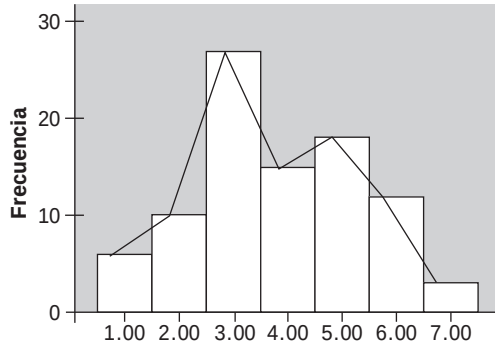
Número de observaciones	*Número de intervalos de clase*
Menor o igual que 25	De 5 o 6
Entre 25 y 50	De 7 a 14
Mayor o igual que 50	De 15 a 20

Esta metodología puede aplicarse también para variables cuantitativas discretas que tienen muchos valores diferentes.

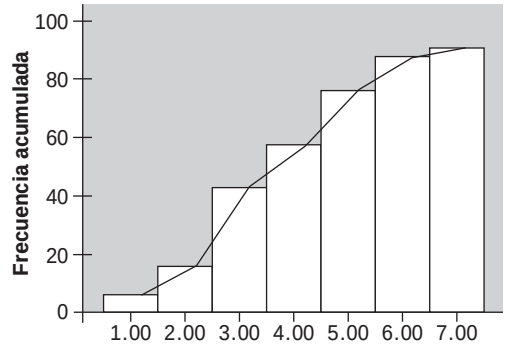
Las frecuencias que se indican en una tabla de distribución por intervalos de clase se representan gráficamente, colocando los intervalos de clase en el eje *X* y construyendo sobre cada uno de estos barras rectangulares yuxtapuestas de tal manera que la altura de cada una de ellas sea proporcional a la frecuencia respectiva que se desea representar. Los gráficos que resultan son llamados *histogramas.* Los histogramas de frecuencia absoluta y de frecuencia acumulada se representan en la Figura 1.7a y Figura 1.7b, respectivamente.

La distribución de los datos puede observarse también trazando los denominados polígonos de frecuencias. Estos gráficos se obtienen uniendo, mediante segmentos, los puntos medios de los lados superiores de los rectángulos del histograma. En el caso de las frecuencias acumuladas, la poligonal que se obtiene se llama *ojiva.* Los polígonos de frecuencia suavizados indican la forma de la distribución de los datos; estos sugieren el "modelo teórico" que puede servir para el análisis de los datos.

FIGURA 1.7 *Histograma de frecuencias e histograma de frecuencias acumuladas*



(a) Histograma y polígono de frecuencias



(b) Histograma de frecuencias acumuladas y polígono de frecuencias acumuladas. Ojiva

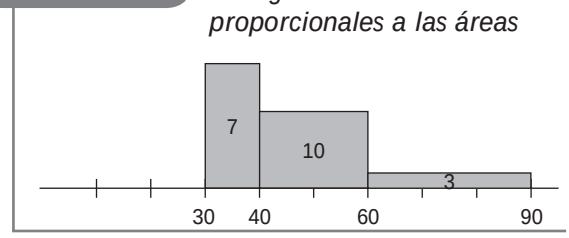
Observación

1. Si no se tienen los datos originales y solo se cuenta con la información tabulada, para facilitar la interpretación, se supone que los datos se distribuyen de manera uniforme en cada intervalo.
2. En los histogramas construidos se considera que los intervalos de clase son de igual longitud y que las frecuencias son proporcionales a las alturas de los rectángulos. Si los intervalos de clase son de longitudes diferentes, las frecuencias deberán ser proporcionales a las áreas de los rectángulos y no a las alturas de los rectángulos.

TABLA 1.10 *Tabla de frecuencia*

Intervalos de clase	*Frecuencia*
[30, 40]	7
]40, 60]	10
]60, 90]	3

FIGURA 1.8 *Histograma. Las frecuencias proporcionales a las áreas*



Los histogramas pueden tomar diferente forma. Algunas veces, los histogramas son simétricos con colas a la derecha y a la izquierda (Figura 1.9a). Se dice que estos histogramas tienen la forma de una *distribución norma*l.

La forma de los histogramas también puede ser sesgada con cola a la derecha (Figura (1.9b). En este caso los datos están concentrados en el extremo izquierdo, y van disminuyendo gradualmente en número hacia el extremo derecho del eje horizontal. Los histogramas asimétricos, como el de la Figura 1.9d, son sesgados con cola a la izquierda. En este caso los polígonos resultantes representan una distribución que tiene sus datos concentrados a la derecha, y van disminuyendo gradualmente cuando el eje horizontal se recorre de derecha a izquierda. Por ejemplo, la distribución de los datos que indican el número de días que se requiere para vender una casa, generalmente, tiene una distribución que se representa como el histograma que aparece en la Figura 1.9d. Esto, porque una casa no se vende tan rápidamente. Los salarios, a menudo, tiene una distribución como la que indica la Figura 1.9b. (La mayoría gana poco y la minoría gana mucho).

Algunas recomendaciones, como las siguientes, se pueden tener en cuenta para interpretar un histograma.

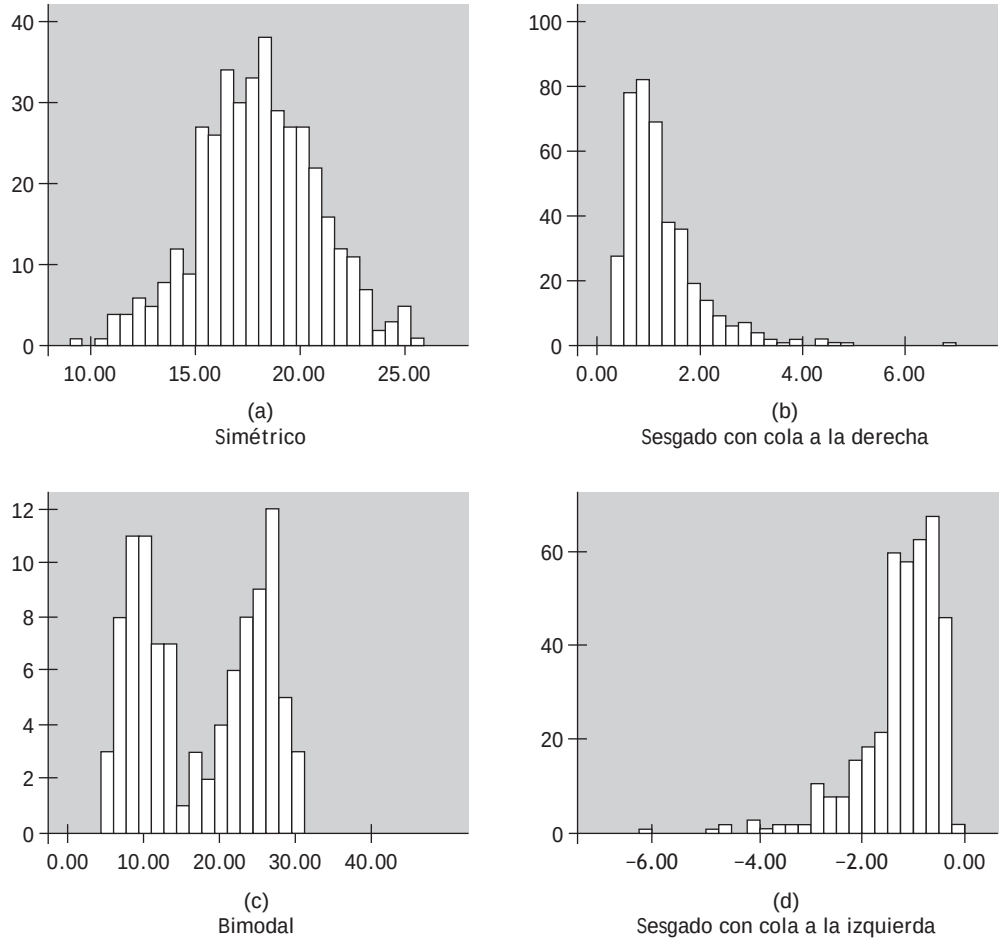
1. Observar las barras de mayor frecuencia, es decir, la "tendencia central" de los datos.
2. Estudiar el punto en donde se "centran" los datos.
3. Estudiar la variabilidad.
4. Analizar la forma del histograma. Observar si el histograma es simétrico, sesgado, multimodal o presenta depresiones.
5. Observar la existencia de "datos raros", es decir, medidas muy extremas. Los datos raros reflejan a menudo situaciones especiales que es preciso investigar (datos incorrectos, mediciones realizadas sobre elementos que no pertenecen a la población o proceso en estudio).

También es necesario considerar las limitaciones que tiene esta herramienta fundamental. Algunas son las siguientes:

6. Con un histograma es difícil detectar tendencias que suceden en el tiempo. Para ello es preferible usar las gráficas de series de tiempo o las cartas de control (gráficos que se explican más adelante).

7. Los histogramas no son adecuados para comparar la variabilidad de dos grupos o más grupos de datos. Son preferibles los "diagramas de cajas", que se verán más adelante.

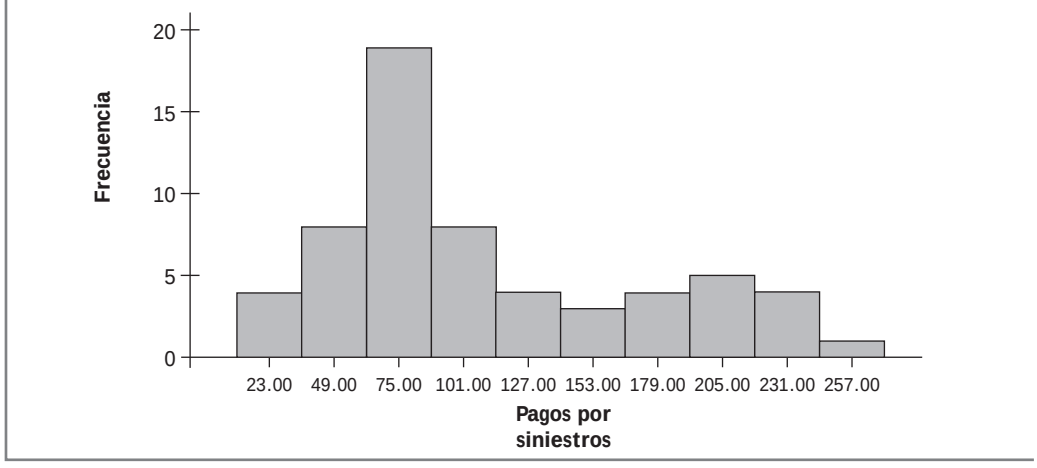
FIGURA 1.9 *Diferentes tipos de histogramas: simétricos, sesgados, bimodales*



EJEMPLO. *Histograma de los pagos por siniestros*

Utilizando los datos correspondientes a los pagos realizados por la compañía de seguros Segur por diferentes siniestros, se obtiene el histograma de la Figura 1.10.

FIGURA 1.10 *Pagos por siniestros*



En el histograma se observa que los pagos menores por siniestro suceden con mayor frecuencia.

Diagrama de tallos y hojas

Los histogramas y los gráficos de bastones muestran la manera como varían los datos. Sin embargo, estos no permiten la identificación de los valores particulares. El diagrama de tallos y hojas muestra los datos y además la manera como estos están distribuidos.

> *El diagrama de tallos y hojas se obtiene básicamente ordenando los datos en intervalos de clase, como en un histograma, y separando las cifras que forman cada dato en dos partes: una con las primeras cifras para formar el* ***tallo,*** *y la otra con las cifras restantes para formar las* ***hojas.***
>
> *Cada tallo conjuntamente con sus hojas forma una* ***rama.***

Así por ejemplo, si se tiene el dato 68, la cifra 6, de las decenas, conformará el tallo, mientras que la cifra 8, de las unidades, formará *la* hoja.

Tallo	*Hoja*
6	8

Con los siguientes datos, correspondientes a medidas tomadas en un laboratorio:

21 22 27 29 32 33 36 37 38 39 42 44 44 46 46 46 48 49 49 51 52 54 55 56 64 65 66 70 71 71 75 102

se construye el diagrama de tallos y hojas correspondiente. Se consideran intervalos de clase de longitud 10, tomando las cifras de las decenas como tallos y las cifras de las unidades como hojas. Por ejemplo, los valores 21, 22, 27 y 29 (comprendidos en el intervalo [20, 30]) están en la primera rama. La segunda rama comprende los valores que están en el intervalo]30, 40]. La frecuencia de esta rama es 6.

2	1279
3	236789
4	244666899
5	12456
6	456
7	0115
10	2

Cuando existan muchas hojas en un tallo se pueden formar otros tallos. A partir del tallo 2 por ejemplo, se pueden crear los tallos 2* y 2•, considerando como hojas del tallo 2*, los datos de 0 a 4 y como hojas del tallo 2• los datos de 5 a 9.

EJEMPLO. *Diagrama de tallos y hojas*

El diagrama de tallos y hojas, que se muestra a la derecha, corresponde a los siguientes datos:

4.12, 4.12, 4.26, 4.26, 4.26, 4.27, 4.28, 4.30, 4.31, 4.31, 4.32, 4.32, 4.33, 4.35, 4.42, 4.4.3, 4.51, 4.62

2	41	22
5	42	66678
6	43	0112235
2	44	23
1	45	1
1	46	2

En este caso, 41| 2 representa, por ejemplo, al dato 4.12. Sin embargo, podría corresponder también a datos que han sido “redondeados”, como 4.1234, 4.1242, etcétera.

1.7 Variabilidad multidimensional

Se han presentado algunos métodos para estudiar la variabilidad de los datos de una sola variable; sin embargo, cuando se realizan experimentos o encuestas se mide generalmente no una sino varias características de los elementos que se observan. En este caso los datos corresponden a varias variables y por ello se llaman *datos multivariados*. Dependiendo del tipo de cada una de las variables, estos datos se presentan en tablas y se representan mediante gráficos como los que se indican a continuación.

Los diagramas de dispersión para dos variables

Los diagramas de dispersión para dos variables son representaciones gráficas en el plano cartesiano de pares de datos (x, y), correspondientes a dos variables cuantitativas. Se obtienen colocando el valor x del par en el eje de las abscisas y el valor y en el eje de las ordenadas.

En la Figura 1.11 aparece un diagrama de dispersión de los pares

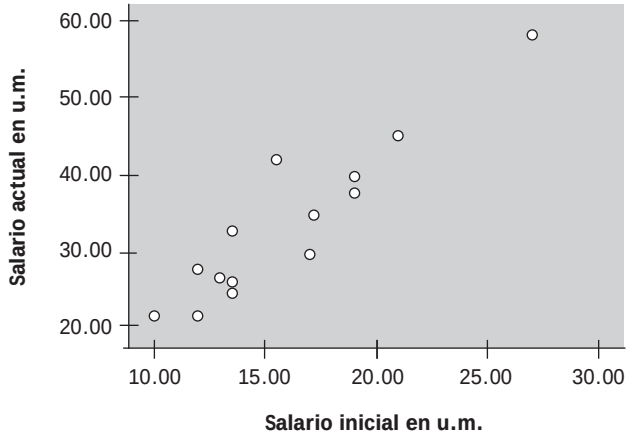
(10, 22), (12, 22), (12, 28), (13, 27), (13.5, 25), (13.5, 26), (13.5, 33), (15.5, 42), (17, 30), (17.2, 35), (19, 38), (19, 40), (21, 45), (27, 58)

cuyas coordenadas corresponden, respectivamente, a las variables:

X: *salario inicial e Y*: *salario actual,* ambas en unidades monetarias (u.m.) y definidas para un grupo de 14 personas.

En el diagrama de dispersión se observa que los valores de *X* crecen conjuntamente con los valores de *Y*. Ello no implica necesariamente que la causa de un mayor salario actual sea un mayor salario inicial.

FIGURA 1.11 *Diagrama de dispersión*



Distribución conjunta de frecuencias. Tablas cruzadas o tablas de contingencia

Para dos variables categóricas *X* con valores $x_1,\ldots, x_m$ e *Y* con valores $y_1,\ldots, y_n$, su distribución conjunta de frecuencias indica el número de veces, n_{ij}, que las variables toman de manera simultánea los valores x_i e y_j, respectivamente. Las frecuencias n_{ij} se escriben en una tabla llamada *tabla cruzada o tabla de contingencia*.

Para iniciar una estrategia de ventas, una tienda al menudeo lleva a cabo una encuesta a 500 clientes en donde se recoge la información relativa a las siguientes variables categóricas:

Género (X) , con las categorías: 1: Mujer y 2: Varón

Horario de compra (Y), con las categorías: 1: Mañana, 2: Tarde, 3: Noche.

La distribución conjunta de estas variables se registra en la Tabla 1.11, de contingencia

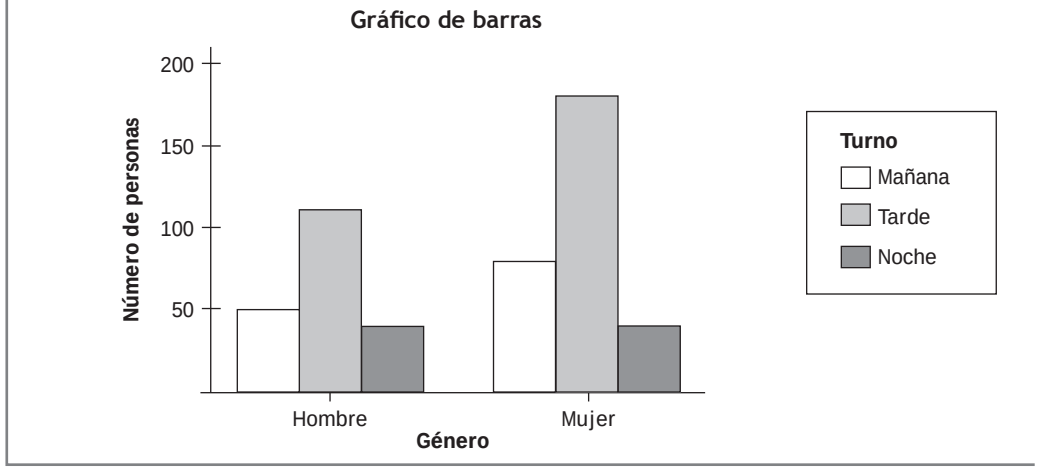
TABLA 1.11 *Tabla de contingencia. Distribución conjunta: género vs. horario de compra*

		Y			***Total***
		1: Mañana	2: Tarde	3: Noche	
X	1: Hombre	50	110	40	200
	2: Mujer	80	180	40	300
	Total	130	290	80	500

La frecuencia conjunta del par (1, 2) es 110. Indica que el número de clientes, de los 500 encuestados, que son hombres y que compran en la tarde es 110.

Los totales indican las frecuencias de cada variable. Las frecuencias de X son: 200 para los hombres y 300 para las mujeres. Las frecuencias de Y son: 130 para el turno de la mañana, 290 para la tarde y 80 para la noche. En general, estas frecuencias se llaman *frecuencias marginales*. La representación gráfica aparece en la Figura 1.12. Es una tabla de barras agrupadas.

FIGURA 1.12 *Barras agrupadas*



EJEMPLO. *Control de calidad: defectos en placas de circuitos*

La tabla de contingencia 1.12 corresponde a la distribución conjunta de las variables discretas: *X* = Tipo de defectos principales e *Y* = Tipo de defectos secundarios de 73 placas de circuitos grabados. En esta tabla (1.12) se muestran las frecuencias, expresadas en porcentajes con relación a los totales en las filas. Resultados análogos se obtienen procediendo con las columnas.

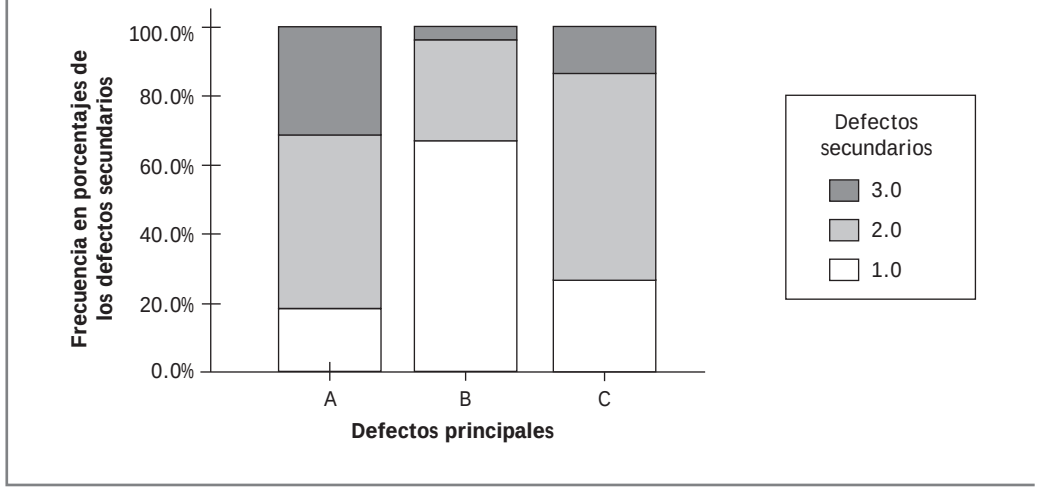
TABLA 1.12 *Tabla de contingencia. Defectos principales * Defectos secundarios*

			Defectos secundarios			
			D	E	F	**Total**
Defectos principales	A	Frecuencia	3	8	5	16
		% de defectos principales	18.8%	50.0%	31.3%	100.0%
	B	Frecuencia	18	8	1	27
		% de defectos principales	66.7%	29.6%	3.7%	100.0%
	C	Frecuencia	8	18	4	30
		% de defectos principales	26.7%	60.0%	13.3%	100.0%
Total		Frecuencia	29	34	10	73
		% de defectos principales	39.7%	46.6%	13.7%	100.0%

El análisis de los valores de las variables puede hacerse a partir de la tabla de porcentajes o utilizando el gráfico de la Figura 1.13. Por ejemplo, usando uno u otro medio, podemos concluir que el defecto secundario D ocurre con más frecuencia conjuntamente con el defecto principal B. Que el defecto secundario E ocurre con mayor frecuencia conjuntamente con el defecto principal C y que el defecto secundario F ocurre con menor frecuencia con el defecto principal A. Nótese que la distribución de frecuencias de los defectos secundarios cambian cuando cambian los defectos principales. Por ello diremos que los defectos secundarios varían con los defectos principales.

Los porcentajes de las categorías D, E y F (defectos secundarios) condicionados a cada categoría A, B o C (defectos principales) se llaman *perfiles fila*. Por ejemplo, los porcentajes 26.7, 60.0 y 13.3 de los defectos D, E y F, respectivamente, conforman el perfil fila condicionado al defecto principal C.

FIGURA 1.13 *Perfiles fila condicionados a los defectos principales*



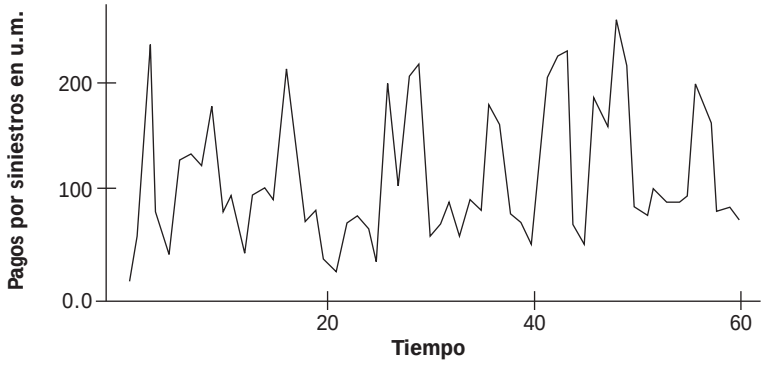
Una recomendación básica a tener en cuenta cuando se analizan datos de una tabla de contingencia es saber si estos son resúmenes de otras tablas de contingencia. Es necesario analizar los datos de las tablas que se usaron para obtener los agregados, pues algunas veces las conclusiones que se obtienen de las tablas agregadas contradicen a las que se obtienen desagregando los resultados.

Gráficos de series de tiempo

En ejemplos anteriores, se consideró que los datos habían sido recolectados al mismo tiempo. Para referirse a este tipo de datos, los analistas los llaman *datos de tabulación transversal*. Cuando los datos de una misma variable se recogen a lo largo del tiempo, se tiene una *serie de tiempo*.

El gráfico que a continuación aparece (Figura 1.14), al unir los puntos que resultan cuando en el eje horizontal de un plano cartesiano se representa el tiempo, en meses, y en el eje vertical se representan los pagos, en unidades monetarias (u.m.), por siniestros realizados por la compañía de seguros Segur, durante 60 meses consecutivos, es una serie de tiempo.

FIGURA 1.14 *Serie de tiempo de los pagos por siniestro*



Gráfica de radar o telaraña

Las gráficas de radar o telaraña se utilizan para ilustrar las diferencias entre un estado actual y un estado ideal. El gráfico consta de una serie de polígonos o círculos concéntricos en donde el número de radios graficados corresponde al número de los ítems sobre los cuales se realiza la comparación.

En el siguiente caso se ilustran las calificaciones realizadas por un cliente a un banco en siete rubros principales:

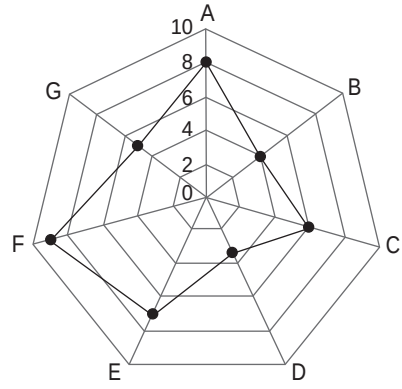
A: Seguridad
B: Parqueo
C: Facilidad de acceso
D: Atención al cliente
E: Prontitud
F: Confianza
G: Diversidad de productos

La escala en cada ítem es del 1 al 10. Lo ideal corresponde a la calificación 10.

Cliente i

Ítem	*Escala*
A	8
B	4
C	6
D	3
E	7
F	9
G	5

FIGURA 1.15 *Gráfica de radar o telaraña*



APLICACIÓN: Problemas en el banco B

Una serie de problemas han sido detectados en la entidad bancaria B; sin embargo, dos de ellos se han señalado con mayor frecuencia, los cuales, por su efecto en la economía y en la imagen de la empresa, tienen la primera prioridad para su solución. Estos problemas son:

1. La pérdida de clientes (*churn* en el idioma inglés) y
2. La morosidad de los clientes que han recibido un préstamo.

Los expertos en el análisis de datos han recomendado que para analizar el primer problema era necesario conocer la "migración" en estos últimos meses y elaborar cuestionarios que permitan investigar sobre la satisfacción de los clientes con los servicios que brinda el banco, así como los factores que pueden incidir en la decisión de los clientes de irse a otra entidad de este tipo. Las variables que pueden servir para el estudio de la migración son: la que indica los bancos en los cuales los clientes realizaban con mayor frecuencia sus operaciones bancarias hace tres meses y la que indica los bancos en donde realizan sus operaciones con mayor frecuencia en la actualidad.

Para el segundo problema, y al revisar la literatura, se ha detectado que dos características podrían ayudar a discriminar a los buenos deudores (que cumplen puntualmente con el pago de las cuotas) de los malos deudores: la edad y el sueldo. Dado que solo se trata de dos variables, esta suposición podría ser validada con la ayuda de un gráfico.

La base de datos que se utilizará en el análisis es una muestra de 410 clientes, construida a partir de la base de datos del banco. Esta muestra ha sido complementada con una encuesta aplicada a los 410 clientes y contiene las siguientes variables:

1. Educación: profesional (1) , no profesional (2)
2. Morosidad: no moroso (0), moroso (1)
3. Edad del cliente
4. Sueldo del cliente
5. Banco en donde realizaba sus operaciones con mayor frecuencia hace tres meses: banco A (1), banco B (2) y banco C (3).
6. Banco en donde realiza con mayor frecuencia sus operaciones en la actualidad: banco A (1), banco B (2) y banco C (3).

Las distribuciones de las variables "banco anterior" y "banco actual" aparecen a continuación:

TABLA 1.13 *Banco anterior*

		Frecuencia	*Porcentaje*
Banco	A	264	64.4
	B	111	27.1
	C	35	8.5
	Total	410	100.0

FIGURA 1.16 *Banco anterior*

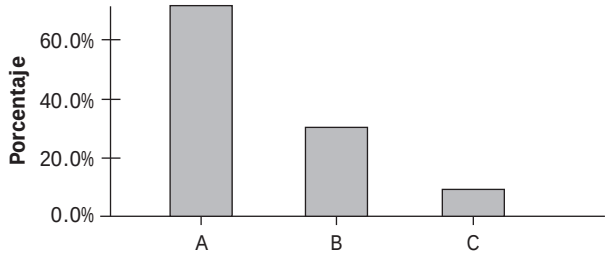
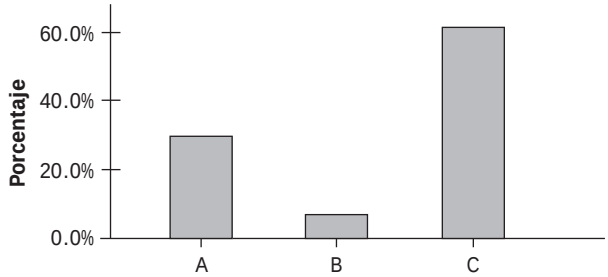


TABLA 1.14 *Banco actual*

		Frecuencia	Porcentaje
Banco	A	123	30.0
	B	31	7.6
	C	256	62.4
Total		410	100.0

FIGURA 1.17 *Banco actual*



La variable "banco anterior" indica que el 64.4% realizaban sus operaciones en el banco A (1), el 27.1% en el banco B (2) y el 8.5% en el banco C (3).

La variable "banco actual" indica que en la actualidad el 30% realizan sus operaciones en el banco A, el 7.6% en el banco B y el 62.4% en el banco C. Como se puede observar, el banco B ha sufrido una fuerte disminución en el porcentaje de clientes que eran atendidos en este banco hace tres meses.

Para averiguar cómo ha sido la transferencia de clientes en estos tres últimos meses se ha construido la siguiente tabla de contingencia (Tabla 1.15) con las variables "banco anterior" y "banco actual". Se observa que solo el 6.3% ha permanecido en B y la mayoría de sus antiguos clientes (el 62.2%) ha migrado al banco C (3).

TABLA 1.15 *Tabla de contingencia. Banco anterior * Banco actual*

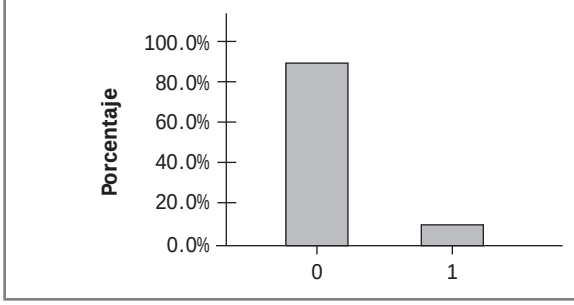
			Banco actual			
			A	B	C	**Total**
Banco anterior	A	Frecuencia	78	23	163	264
		% de banco anterior	29.5%	8.7%	61.7%	100.0%
	B	Frecuencia	35	7	69	111
		% de banco anterior	31.5%	6.3%	62.2%	100.0%
	C	Frecuencia	10	1	24	35
		% de banco anterior	28.6%	2.9%	68.6%	100.0%
Total		Frecuencia	123	31	256	410
		% de banco anterior	30.0%	7.6%	62.4%	100.0%

Con respecto a los morosos y de acuerdo a la distribución de la variable "morosidad" se puede indicar que 10.2% de los clientes que han recibido un préstamo son morosos (1).

TABLA 1.16 *Morosidad*

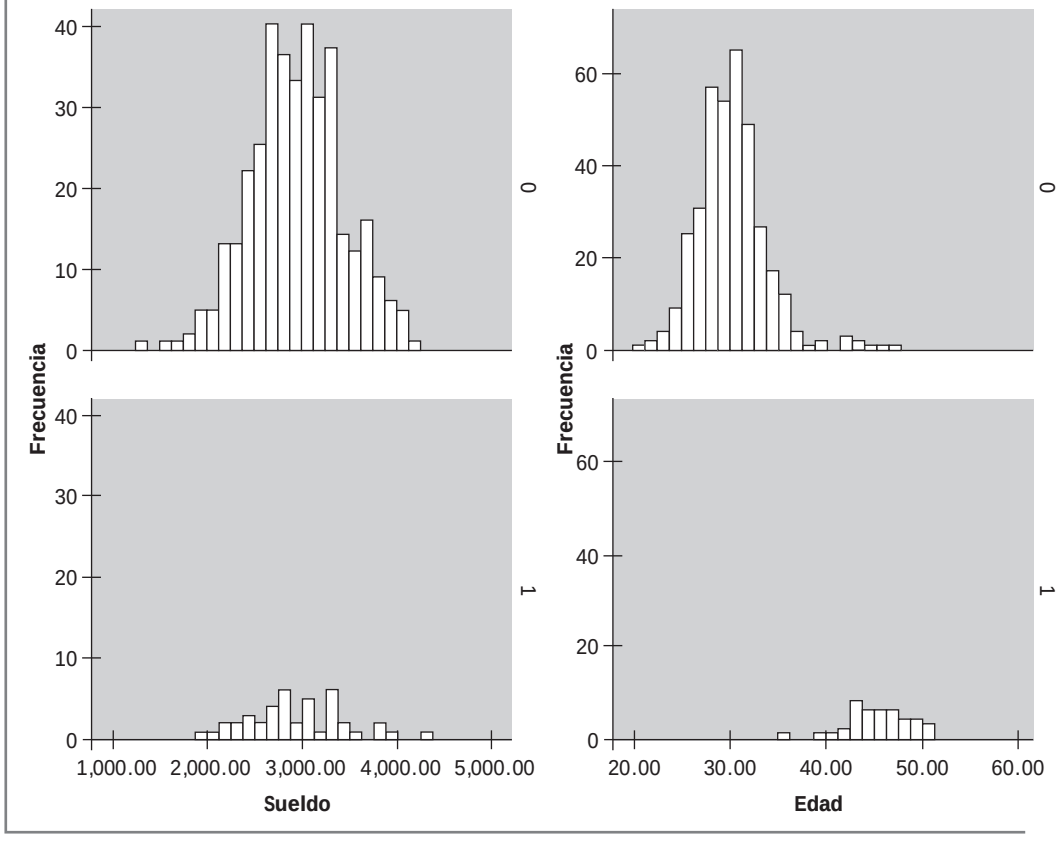
	Frecuencia	**Porcentaje**
Morosidad 0	368	89.8
1	42	10.2
Total	410	100.0

FIGURA 1.18 *Morosidad*



Graficando el histograma de las variables "sueldo" y "edad" para cada una de las categorías de la variable "morosidad", se observa que al parecer la única variable que discrimina a los morosos (1) de los no morosos (0) es la variable edad.

FIGURA 1.19 *Distribuciones del sueldo y edad en los morosos (1) y no morosos (0)*

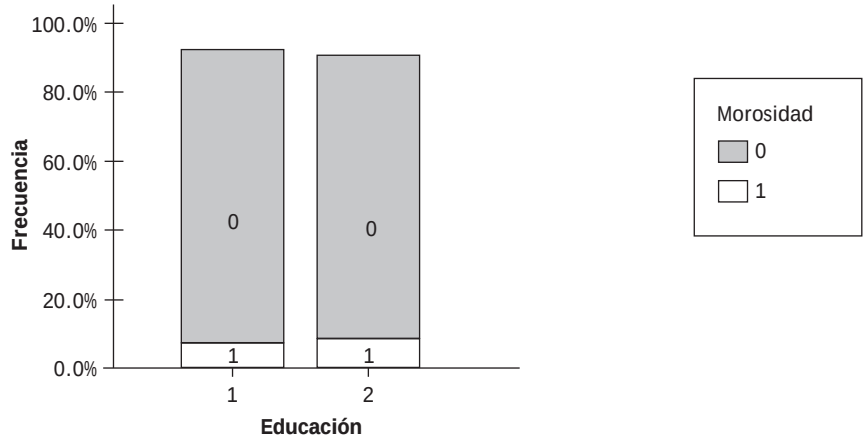


Observando la tabla de contingencia versus morosidad (Tabla 1.17), se tiene que la distribución de los morosos y no morosos, prácticamente, es igual en ambas categorías de la variable educación. La variable educación no es un factor discriminante de los morosos y no morosos.

TABLA 1.17 *Educación * Morosidad*

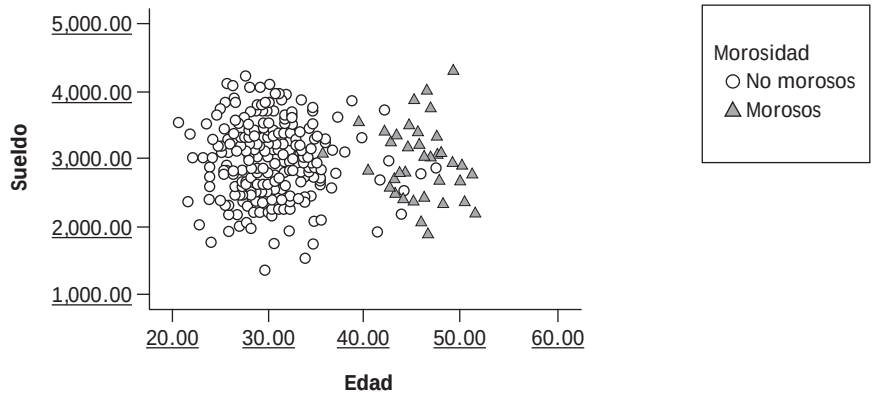
			Morosidad		
			0	1	***Total***
Educación	1	Frecuencia	183	17	200
		% de educación	91.5%	8.5%	100.0%
	2	Frecuencia	190	20	210
		% de educación	90.5%	9.5%	100.0%
Total		Frecuencia	373	37	410
		% de educación	91.0%	9.0%	100.0%

FIGURA 1.20 *Distribuciones de los morosos y no morosos en cada una de las categorías de la variable "educación"*



La suposición anterior toma fuerza al graficar el diagrama de dispersión (Figura 1.21) entre las variables “edad” y “sueldo” e indicando en este diagrama las categorías de la variable “préstamo”: morosos (1) y no morosos (0).

FIGURA 1.21 *Al parecer, la variable edad es la más discriminante*



LA ESTADÍSTICA EN LA EMPRESA

El empirismo y la intuición han sido por mucho tiempo las metodologías que imperaban en la toma de decisiones de la empresa y los negocios. En la actualidad, estos métodos son insuficientes y limitados para resolver problemas competitivos relacionados con la calidad y producción y para analizar la cantidad de datos con los que hoy en día se dispone. Muchas empresas se han visto obligadas a sustituir estos métodos tradicionales con metodologías basadas en el tratamiento científico de la información, que realmente las ayuden en la toma de decisiones.

Las empresas modernas conscientes de este panorama han ido incorporando en su gestión cuadros de profesionales, entre los que se cuentan a los estadísticos, lo que ha permitido el desarrollo de muchos procesos modernos que han elevado la producción y en muchos casos con costos más bajos.

Aún falta que este progreso se experimente no solo en el aspecto productivo sino también en lo que se refiere a la gestión administrativa de la empresa. En este sentido, el papel del profesional de la estadística es fundamental para la empresa, sobre todo si se sabe que todos los procesos que se desarrollan y que generalmente se desean mejorar tienen a la variabilidad como característica común.

LA ESTADÍSTICA EN LA EMPRESA

El periódico *La Verdad*

La última encuesta realizada por una agencia de publicidad indicó que el 70% de los ciudadanos opina que *La Verdad* es el diario que marca la pauta de la credibilidad en el país.

Cada día, más de 550,000 lectores buscan en *La Verdad* información sobre el acontecer nacional y mundial, sobre la economía, finanzas y deportes; últimamente ha incorporado una sección muy bien presentada sobre ciencia y tecnología.

Los atributos que han permitido el posicionamiento de *La Verdad* en la mente de los lectores son: la diversidad y gran calidad de la información que han hecho posible además la diferenciación de los otros diarios del país.

El diario *La Verdad* fue creado por el patriarca de una de las familias más importantes del país hace 100 años, y desde su inicio, el registro de la vida del país ha sido el motivo para que hoy día sea objeto de consulta en todas las bibliotecas del mundo. Además, por el interés que esta empresa periodística tiene por la renovación, por el acercamiento a la juventud y a la familia y por la conservación del medio ambiente, ha recibido una serie de premios y menciones otorgados por varias renombradas instituciones nacionales y extranjeras.

Hoy en día, *La Verdad* tiene un lugar en Internet que es visitado diariamente por tres millones de internautas ávidos de las noticias y de la lectura especializada dirigida a la juventud, a la política, etcétera.

La preocupación de *La Verdad* por el análisis estadístico de la información se observa semanalmente en la serie de trabajos estadísticos relacionados con diferentes campos como: el medio ambiente, la política, la economía, la educación, la industria, etc. Aparte de ello, esta empresa hace uso de la estadística para conocer el perfil de los diferentes conglomerados de sus lectores y así poder llegar a ellos con mayor eficacia.

EJERCICIOS

1. Se indica que el porcentaje de personas mayores de 20 años que no tienen empleo en una ciudad de 5 millones de habitantes es 6%. Los resultados fueron obtenidos a partir de un cuestionario aplicado a 2,000 personas de la ciudad, elegidas de entre las personas mayores de 20 años.
   a) ¿Sobre qué población se ha realizado la encuesta? ¿Cuáles son las unidades estadísticas?
   b) ¿Cuál ha sido la muestra utilizada?
2. ¿Una muestra grande es necesariamente una buena muestra?
3. En los siguientes casos, indicar el tipo de escala a usar si se trata de:
   a) Asignar a los distritos de la capital su código postal.
   b) Asignar a cada profesor de una universidad su categoría (profesor auxiliar, profesor asociado o profesor principal).
   c) Asignar a cada uno de los objetos producidos en una fábrica su peso.
4. Se eligieron al azar 10 alumnos de una escuela de administración y se midió la estatura de cada uno de ellos.
   a) ¿Los datos corresponden a una variable cualitativa o cuantitativa?
   b) ¿Los datos corresponden a una variable discreta o continua?
5. Un profesor propone a sus alumnos del curso de aritmética una prueba con 10 ejercicios de cálculo. Si al resolverla alguno de los ejercicios presenta algún error, el profesor califica toda la prueba con 0, de otro modo la califica con 20. ¿Qué tipo de escala empleó?
6. A los elementos *w, y, z* de una población se les asignaron los números 8, 16 y 30, respectivamente, para medir cierto atributo. Si en lugar del número asignado a *w* se le asignará el número 20 y el 40 a *y*, ¿qué número corresponderá asignarle a *z* si la escala usada es:
   a) nominal? b) ordinal? c) de intervalo?
7. En un censo de alumnos de una universidad se ha considerado como variables importantes para obtener cierto tipo de información a las siguientes: facultad de pertenencia, año de ingreso, nacionalidad, tiempo de residencia en el país, número de semestres que lleva en la universidad, grado en la escala de pensiones, grado de instrucción del padre, número de hermanos e ingreso mensual familiar promedio. Hacer una clasificación de las variables que se usaron en el censo, indicando la escala apropiada para determinar los valores de estas variables.
8. Con la finalidad de medir el nivel de satisfacción de sus clientes, un banco usó una encuesta de opinión en donde se consultó acerca del servicio en general del banco. Las respuestas a las preguntas podían ser: pésimo, regular, bueno y excelente. ¿Qué tipo de variable representa a las respuestas, cualitativa o cuantitativa? ¿Qué escala se usó?
9. A un fumador se le solicita que diga el orden de preferencia de cuatro tipos de cigarrillos de marcas A, B, C y D. Indicar el tipo de escala utilizada.
10. En una encuesta relacionada con automóviles aparecen las siguientes cuestiones:
    Los automóviles RWC:
    Consumen mucha gasolina - - - - - - - - - Consumen poca gasolina

No son ruidosos - - - - - - - - - - - - - - - - Son muy ruidosos

Los que responden deben indicar su tendencia hacia las categorías opuestas de la dimensión que se estudia. ¿Qué tipo de escala se usó?

11. La asociación de salas de cine realizó una encuesta entre 500 personas que asistieron a la última función de la semana. Algunas de las preguntas fueron:
    a) ¿Cuántas veces asistió al cine las últimas 12 semanas?
    b) ¿Cuál es su edad?
    c) ¿Asiste al cine solo o acompañado?
    d) ¿Considera usted que el precio de entrada es caro o barato?

    Indicar el tipo de variables a que dan lugar las preguntas anteriores.

12. Usando una encuesta, la Oficina Nacional de Turismo colecciona datos de los visitantes que llegan al país por vía aérea.

    Entre las cuestiones que se plantearon en el cuestionario están las siguientes:
    a) Este es mi: primer, segundo, tercer, cuarto, quinto, etcétera viaje.
    b) El motivo de este viaje es turismo, negocios, salud, otros.
    c) El número de días que permaneceré en el país es...
    d) Planeo estar en: hotel, casa de familiares, campamento, otro.

    ¿Cuál es la población que está siendo estudiada?

    Indicar el tipo de escala relacionada con cada cuestión.

13. Una encuesta para estudiar la preferencia de los electores por el candidato NN a la alcaldía de la ciudad AA se llevó a cabo entre 600 electores de la ciudad. El resultado fue que el 45% prefieren al candidato NN.
    a) ¿Cuál es la población que corresponde a esta encuesta?
    b) ¿Cuál fue la muestra para esta encuesta?
    c) ¿Por qué fue necesario realizar una encuesta?
    d) Indicar la variable que corresponde a los datos recogidos.

14. Determinar el tipo de cada una de las variables relacionadas con:
    a) El número de automóviles por hogar en una ciudad.
    b) El número de llamadas telefónicas realizadas por mes por una persona.
    c) El tiempo que demora un aviso comercial en la TV.
    d) Cantidad de dinero que las universidades gastan mensualmente en libros.
    e) Carrera universitaria elegida por un estudiante.
    f) Tiempo utilizado en atender un cliente en la ventanilla de un banco.
    g) Número de empleados que tiene una empresa minera.
    h) Número de clientes que llegan entre las 10 a.m. y las 12 p.m. a un banco local.

15. Dar un ejemplo de una variable numérica que sea continua pero que, en la práctica, se trate como variable discreta.

16. En una encuesta se pidió lo siguiente:

    Marcar con una X el cuadro que corresponde a su nivel de ingreso.

    ☐ Menos de 1,000 ☐ Entre 1,000 y 2,000

    ☐ Entre 2,000 y 3,000 ☐ 3,001 a más

    Indicar el tipo de variable correspondiente. ¿Qué tipo de escala o de medición se usó?

17. ¿A qué tipo de variable da lugar la siguiente pregunta que aparece en una encuesta?

    En conjunto, ¿diría usted que se siente muy feliz, feliz, poco feliz o nada feliz?

18. ¿Cuántas y cuáles variables definiría usted para recibir las respuestas a la siguiente pregunta que aparece en una encuesta?

    Pregunta: A su criterio, ¿cuáles de los siguientes problemas aquejan al país?

    a) corrupción b) narcotráfico c) inseguridad d) falta de servicios

    ¿De qué tipo son las variables definidas?

19. ¿Cuántas y cuáles variables definiría usted para recibir las respuestas a la siguiente pregunta que aparece en una encuesta?

    Cuando acude a un centro comercial a realizar sus compras de vestido y complementos, ¿cuáles de los siguientes factores considera más importantes? Ordenar, colocando en primer lugar aquel que considere más importante y en último lugar aquel que considere menos importante.

    Número de tiendas en el centro comercial

    Prestigio de las tiendas

    Prestigio del propio centro comercial

    Distancia de su casa al centro comercial

    Superficie comercial

    Parqueo

    Ofertas

20. Escriba dos frases positivas y dos frases negativas que permitan evaluar la actitud de los individuos frente a los servicios de seguridad ciudadana que presta el gobierno nacional.

21. Utilizando la escala de diferencial semántico y usando dos atributos, evaluar la actitud de los individuos frente a la corrupción en el país.

22. En la tabla siguiente aparecen datos correspondientes a 10 universidades.

Universidad	*EST99*	*EST00*	*GRA99*	*GRA00*	*NUPRO*
Universidad 1	8,000	9,500	500	480	400
Universidad 2	2,000	1,800	100	120	80
Universidad 3	4,000	4,200	200	190	300
Universidad 4	3,500	3,200	350	300	180
Universidad 5	1,800	1,500	170	180	80
Universidad 6	6,000	6,200	500	600	380
Universidad 7	5,000	5,500	200	180	300
Universidad 8	12,000	10,000	480	500	500
Universidad 9	2,200	1,500	80	75	150
Universidad 10	4,700	4,000	200	200	250

En esta tabla se indican las siguientes variables:

EST99 = Número de estudiantes en el año 1999.

EST00 = Número de estudiantes en el año 2000.

GRA99 = Número de graduados en el año 1999.

GRA00 = Número de graduados en el año 2000.

NUPRO = Número de profesores en el año 2000.

a) A partir de las variables indicadas, generar las siguientes variables:

VRET: Variación relativa, en %, entre el número de estudiantes en 2000 con respecto a 1999 (VEST/EST99)*100.

PROALU: Número de profesores por alumno en el año 2000.

PGRA99: Porcentaje de graduados en 1999 respecto del total.

PGRA00: Porcentaje de graduados en 2000 respecto del total.

TIPO: Tipo de universidad.

La variable TIPO es una variable categórica que clasifica a las universidades de acuerdo al número de profesores. Las categorías para esta variable son como sigue:

pequeña: hasta 150 profesores, mediana: de 151 a 250 profesores y grande: más de 250 profesores.

b) Indicar:

Las universidades que en el año 2000 han tenido el menor y el mayor número de estudiantes.

Las universidades que han tenido el mayor crecimiento de estudiantes en 2000 respecto del número de estudiantes en 1999.

Las universidades que tienen el mayor número de profesores por alumno.

El lugar que ocupa la universidad 5 de acuerdo al número de profesores.

c) Hacer un gráfico adecuado para estudiar la variabilidad de la variable TIPO.

23. Indicar la estructura de la base de datos que se genera al consultar a un grupo de personas sobre el orden que asignarían a los siguientes factores al comprar un refresco: sabor, envase, precio, volumen del líquido y marca. Indicar el tipo de variables que se han considerado.

24. Mediante un gráfico de barras, representar la información que aparece en la siguiente tabla, correspondiente a la canasta de gastos mensuales de un estudiante de una universidad de la capital.

Ítem	*Cantidad en dólares*
Comida	100
Transporte	50
Útiles	40
Entretenimiento	45
Lavandería y otros	65
	Total = 300

25. Un estudiante de derecho quiere mostrar que sus compañeros de especialidad son los que deciden en las votaciones, pues son mayoría. Para ello recogerá la información y la presentará gráficamente.
    a) ¿Cuál es la variable que se debe medir? ¿De qué tipo es?
    b) ¿Qué tipo de diagrama es el más adecuado para representar la información?

26. En cuatro regiones, las reservas de petróleo, estimadas en miles de barriles, fueron como sigue:

Región	*Reservas*
Región 1	38.7
Región 2	22.6
Región 3	8.8
Región 4	60.0

    Representar, usando un diagrama de sectores circulares, la información anterior.

27. El sida es una de las principales causas de mortalidad en el mundo y afecta seriamente al continente latinoamericano. Según el informe de Onusida, el número de personas con el virus del sida en 1997 fue como sigue.

País	*Con sida*	*País*	*Con sida*	*País*	*Con sida*
Argentina	120,000	El Salvador	18,000	México	180,000
Brasil	580,000	Guatemala	27,000	Perú	72,000
Colombia	72,000	Haití	190,000	R. Dominicana	83,000
Ecuador	18,000	Honduras	43,000	Venezuela	82,000

    Representar la información mediante una gráfica adecuada.

28. Treinta y cinco usuarios de la compañía de teléfonos pagaron facturas con las siguientes valorizaciones en dólares:

    35.04 20.55 18.58 20.53 20.75 16.00 18.00 21.56 21.51 21.30 20.59 18.80 20.35 17.78 20.16 21.35 20.20 18.98 16.52 17.08 19.55 21.07 20.61 20.82 16.72 20.63 15.14 19.60 21.13 24.25 20.73 20.02 20.94 17.43 17.19.

    a) Resumir la información construyendo una distribución de frecuencias y un histograma. Usar nueve intervalos de clase.
    b) Resumir la información construyendo una distribución de frecuencias y un histograma. Usar tres intervalos de clase.

    Comparar los histogramas e indicar la conveniencia o inconveniencia de usar muchos o pocos intervalos de clase.

29. Una compañía tiene dos fábricas de vestidos. En cada fábrica 20 empleados completan la misma tarea, registrándose los siguientes tiempos, en minutos:

Fábrica A	*Fábrica B*
17 24 16 28 11 25 5 7 10 30 9	7 26 11 9 24 17 14 32 5 28 19
4 31 21 13 35 26 21 19 28	25 28 31 12 31 27 25 36 19

a) Construir las distribuciones de frecuencias y los histogramas para cada fábrica, usando cinco intervalos de clase.

b) Construir la distribución de frecuencias para el conjunto de todos los datos y el histograma correspondiente, usando cinco intervalos de clase.

c) Comparar los histogramas hallados en a) y b) y comentar los resultados.

30. Para determinar las preferencias por los servicios de una compañía SS, dedicada al reparto de productos, se realizó en cierta ciudad una encuesta a 50 negocios que usan los servicios de SS. Los siguientes resultados indican las veces que cada negocio usó los servicios de la compañía en un lapso de dos meses consecutivos.

10 2 25 3 14 5 17 1 21 6 16 12 11 10 10 26 15 14 29 23 12 7 28 9 3 0 7 4 15 17 14 16 2 8 8 12 11 10 10 27 14 17 7 33 3 22 15 10 12 11

a) ¿Qué porcentaje de negocios hicieron uso de los servicios de la compañía menos de 12 veces?

b) ¿Qué porcentaje de negocios hicieron uso de los servicios de la compañía entre 12 y 18 veces?

c) ¿Qué porcentaje de negocios hizo uso de los servicios de la compañía más de 18 veces?

31. Los siguientes datos proporcionan los ingresos anuales, en miles de dólares, de 50 personas:

7.9 10.3 45.7 95 43.0 56.0 38.0 6.7 48.0 30.5 25.0 40.0 30.0 25.5 50.0 17.1 25.5 43.5 31.6 59.0 41.5 13.5 12.0 9.2 42.0 41.9 35.0 11.7 55.3 27.0 58.4 57.0 29.6 38.5 26.0 16.5 18.0 24.9 20.0 28.0 28.5 36.4 39.5 5.0 9.0 5.0 6.9 7.0 12.0 8.3

a) Construir un histograma de frecuencias relativas con cinco intervalos de clase para los ingresos anuales.

b) Interpretar el histograma.

c) Usar las marcas de clase para estimar la proporción de ingresos que están entre 12,500 y 52,500 dólares.

d) Estimar el porcentaje de sueldos que son inferiores a 50,000 dólares.

e) Estimar el porcentaje de sueldos que son superiores a 40,000 dólares.

32. El número de periódicos que un canillita vendió durante los últimos 24 días fue como sigue:
21 16 30 42 5 33 26 28 45 17 28 39 32 8 34 27 33 27 26 24 28 16 21 19

a) ¿Cuál es el porcentaje de días en los que el canillita vendió más de 20 periódicos?

b) Usar el método de los intervalos de clase para obtener una tabla de distribución de frecuencias con cuatro intervalos de clase y responder la pregunta anterior. Comentar los resultados con respecto a los métodos usados.

33. Un agente de seguros ha registrado en la siguiente tabla las frecuencias de las ventas mensuales de pólizas de seguros, en dólares, ofrecidas durante el mes anterior.

Venta mensual	*Frecuencia*	*Venta mensual*	*Frecuencia*
[10,000, 12,000]	2	[18,000, 20,000[	6
]12,000, 14,000]	4	]20,000, 22,000]	8
]14,000, 16,000]	7	]22,000, 24,000]	2
]16,000, 18,000]	5	]24,000, 26,000]	1

a) Construir el histograma y el polígono de frecuencias relativas. Interpretar los resultados.
b) Graficar la ojiva correspondiente.
c) Indicar las características de la distribución indicada.

34. Para cada uno de los 11 gerentes de igual número de empresas se ha consignado su sueldo mensual, en dólares (*Y*), y el tiempo de servicios (*X*) en la empresa.

Gerente	1	2	3	4	5	6	7	8	9	10	11
Tiempo de servicios, en años. X	4.2	3.8	3.2	2.21	3.0	4.5	2.7	4.3	2.9	4.2	2.6
Sueldo, en dólares. Y	4,700	3,200	3,100	2,500	3,300	3,800	3,200	3,500	2,200	3,600	2,300

Graficar el diagrama de dispersión, tomando en el eje *X* el tiempo de servicio y en el eje *Y* los sueldos. Comentar los resultados.

35. En la siguiente tabla se indican los motivos de quejas sobre el servicio en un banco local.

Motivo de quejas	***Frecuencia***
Demora en la atención en ventanilla	23
Dificultad en el parqueo	27
Horario de atención inadecuado	43
Falta de información general	28
Local no apropiado	10

Utilizar un diagrama de Pareto para indicar en cuáles de los problemas hay que focalizarse para disminuir las quejas sobre el servicio.

36. Los costos originados por las fallas reportadas durante el proceso de elaboración de un producto fueron como se indica en la siguiente tabla. Representar esta información mediante un gráfico de Pareto.

Fallas	***Costos en miles de dólares***	***Costo acumulado***
Desperdicios	500	500
Retrabajo	350	850
Materia prima dañada	80	930
Reinspecciones	20	950

37. Registrar las ventas realizadas por algún centro comercial durante 30 días consecutivos. Usar una gráfica de serie de tiempo. Indicar si las ventas crecen o decrecen durante el tiempo de registro.

38. En la siguiente tabla de contingencia se muestra la distribución de 500 administradores de acuerdo a su ingreso mensual, en miles de dólares, y a los años de servicio.

Tiempo de servicio	*Ingreso mensual en miles de dólares*			
	Menos de 2	*Entre 2 y 4*	*De 4 a más*	*Total*
Menos de 1 año	450	1,200	450	2,100
De 1 a menos de 3 años	200	2,500	380	3,080
De 3 a menos de 5 años	100	1,000	450	1,550
Más de 5 años	80	800	550	1,430
Total	830	5,500	1,830	10,160

a) Calcular los porcentajes por filas y columnas.

b) Construir un gráfico de barras al 100% para comparar la distribución de sueldos en cada tiempo de servicio.

c) Indicar si se observa alguna relación entre los ingresos y el tiempo de servicio.

39. En el reporte anual de la empresa farmacéutica Formal se indicó que las ventas netas y la inversión en publicidad, ambas en miles de dólares, durante los años comprendidos entre 1999 y 2008, fueron como sigue:

Año	*Ventas netas*	*Publicidad*
1999	9,700	1,189
2000	10,570	1,230
2001	11,150	1,290
2002	11,415	1,250
2003	11,080	1,360
2004	12,760	1,540
2005	16,860	1,945
2006	16,900	2,240
2007	18,280	2,312
2008	20,220	2,400

a) Representar gráficamente los datos que aparecen en la tabla.

b) Comentar el comportamiento de manera conjunta de ambas variables a lo largo del tiempo. ¿Sugieren los datos que un aumento en los gastos en publicidad están acompañados con un aumento en las ventas?

40. Se llevó a cabo una encuesta para averiguar si la frecuencia con que acudían los clientes a un centro de ventas estaba relacionada con su satisfacción respecto de los servicios que ahí se les brindaba. Para la frecuencia se consideraron dos categorías: "asiduo" y "no asiduo", y para la satisfacción se usaron las categorías: "satisfecho" y "no satisfecho". Los resultados fueron como sigue:

	Asiduo	*No asiduo*
Satisfecho	400	200
No satisfecho	180	120

a) ¿Se podría indicar que existe alguna relación de tipo estadístico entre la frecuencia de visita y el grado de satisfacción? ¿Se pueden extender las conclusiones obtenidas a la población de donde procede la muestra?

b) Con el fin de obtener mejor información se consideraron los clientes que concurrían los días lunes o los días jueves, pero no ambos a la vez, y así se obtuvo lo siguiente:

Lunes

	Asiduo	***No asiduo***
Satisfecho	250	80
No satisfecho	120	40

Jueves

	Asiduo	***No asiduo***
Satisfecho	150	120
No satisfecho	60	80

¿A la luz de la nueva división de los datos, se puede ampliar la información? ¿Se pueden extender las conclusiones obtenidas a la población de donde vino la muestra?

41. Aplicar la siguiente encuesta a un grupo de personas que acuden a los restaurantes 1 y 2. Construir la base de datos. Representar gráficamente los resultados. Analizar los resultados.

a) Indique la frecuencia con que acude a los dos restaurantes que se indican.

	Restaurante 1	Restaurante 2
5 o más veces por semana	(1)	(1)
3 a 4 veces por semana	(2)	(2)
1 a 2 veces por semana	(3)	(3)
0 veces a la semana	(4)	(4)

b) Indique la suma que gasta

	Restaurante 1	Restaurante 2
Más de 100 u.m.	(1)	(1)
De 50 a 100 u.m.	(2)	(2)
De 20 a menos de 50 u.m.	(3)	(3)
Menos de 20 u.m.	(4)	(4)

c) Respecto de la comida que le sirvieron la última vez, a usted:

	Restaurante 1	Restaurante 1
Le agradó mucho	(1)	(1)
Le agradó	(2)	(2)
No le agradó mucho	(3)	(3)
Le desagradó completamente	(4)	(4)

d) Indique el restaurante que coincide con cada descripción.

	Restaurante 1	Restaurante 2	Sin opinión
Es fácil llegar	(1)	(2)	(3)
Dispone de playa de estacionamiento	(1)	(2)	(3)
Buena calidad de la comida	(1)	(2)	(3)
Precios cómodos	(1)	(2)	(3)
Carta variada	(1)	(2)	(3)
Personal amable	(1)	(2)	(3)
Limpieza	(1)	(2)	(3)
Rapidez en la atención	(1)	(2)	(3)

e) Indique la importancia de cada concepto en la elección del restaurante.

	No es importante				Muy importante
Es fácil llegar	(1)	(2)	(3)	(4)	(5)
Dispone de playa de estacionamiento	(1)	(2)	(3)	(4)	(5)
Buena calidad de la comida	(1)	(2)	(3)	(4)	(5)
Precios cómodos	(1)	(2)	(3)	(4)	(5)
Carta variada	(1)	(2)	(3)	(4)	(5)
Personal amable	(1)	(2)	(3)	(4)	(5)
Limpieza	(1)	(2)	(3)	(4)	(5)
Rapidez en la atención	(1)	(2)	(3)	(4)	(5)

RESPUESTAS A LOS EJERCICIOS

1. a) Población: conjunto formado por las personas mayores de 20 años b) Muestra: conjunto formado por 2,000 personas mayores de 20 años. **2.** No necesariamente. **3.** a) Nominal b) ordinal c) de intervalo y de razón. **4.** a) Cuantitativa b) continua. **5.** Ordinal. **6.** a) Cualquier número diferente de 20 y 40 b) cualquier número mayor que 40 c) 75. **7.** Nominal, de razón, nominal, de razón, de razón, ordinal, ordinal, de razón, de razón, respectivamente. **8.** Cualitativa, ordinal. **9.** Ordinal. **11.** a) Cuantitativa, discreta b) cuantitativa, continua c) cualitativa. **12.** a) Ordinal b) nominal c) de razón d) nominal. **13.** a) Electores para la alcaldía de la ciudad AA b) conjunto de 600 electores de la ciudad d) variable que se puede denotar con X y que toma el valor 1 si el elector votará por NN y 0 si el lector no votará por NN. **17.** Ordinal. **18.** Observar que la pregunta puede ser contestada con una o más respuestas. **22.** b) La universidad 9 ha tenido el menor número de estudiantes en el año 2000. La universidad 6 ha tenido el mayor número de accidentes en el año 2000 c) Usar gráficos de bastones o de barras. **25.** a) Medir la variable que indica el número de estudiantes por facultad. Esta variable es cuantitativa. b) Bastones o barras. **27.** Usar un diagrama de barras cuyas alturas son el número de personas con el virus. **30.** a) 50% b) 20% c) 18% 11. a) 75%. **34.** Observar que los puntos están más o menos alineados.

CAPÍTULO 2

Resumen numérico de los datos

John Tukey

John Tukey nació en New Bedford, Massachussets, EE. UU., en 1915. Murió en el año 2000. En la Universidad Brown obtuvo sus grados de bachiller y maestría en la especialidad de Química en 1936 y 1937, respectivamente. Posteriormente, en la Universidad de Princeton, Tukey obtuvo los grados de máster y doctor en Matemáticas en los años 1938 y 1939, respectivamente.

Después de trabajar 10 años en la Universidad de Princeton como profesor de Matemáticas, en 1965 Tukey comandó el Departamento de Estadística, recién creado, pero a su vez fue miembro del *staff* técnico de los laboratorios Bell, con quienes trabajó hasta el año 1985.

La contribución de Tukey a la estadística es muy importante. Fue uno de los líderes en el campo del análisis de datos. Escribió una serie de libros y trabajos técnicos y científicos en matemáticas, estadística y teoría de la información (introdujo la palabra *bit*).

Durante su vida profesional John Tukey recibió una serie de honores, como la Medalla Nacional de Ciencias de EE. UU., la medalla de honor de IEEE y la medalla James Madison de la Universidad de Princeton; asimismo, fue el primero en recibir el premio Samuel Wilks de la American Statistical Association.

CONTENIDO

2.1 Introducción

Usando tablas de distribución para los datos es posible realizar cierto tipo de análisis de su variabilidad; sin embargo, también es necesario considerar resúmenes numéricos que ayuden a tener una idea más precisa de la manera como los datos están distribuidos.

La mayor parte de los valores de una variable muestran una tendencia a agruparse alrededor de un *valor central*. Este valor, que de alguna manera tipifica al conjunto, se llama medida de *tendencia central.*

Entre las medidas de tendencia central están: *la moda, la media aritmética, la mediana* y *la media geométrica.* Para el buen uso de estas medidas deberá analizarse su significancia y el tipo de variable a las que se aplican.

En muchos de los procesos que se desarrollan en la industria y en el mundo de los negocios, las medidas de tendencia central son las primeras características que se revisan para investigar si estos cumplen con especificaciones previamente indicadas.

Además de las medidas de tendencia central se tienen las medidas de dispersión, que indican cuán diferentes son entre sí los datos u observaciones. Algunas de estas medidas son: *la desviación estándar, la varianza, el coeficiente de variación, el rango* y *el rango intercuartil.*

Para conocer la "forma" como están distribuidos los valores de una variable se usan las medidas de *simetría* y de *apuntamiento* o *curtosis*.

2.2 Medidas de tendencia central

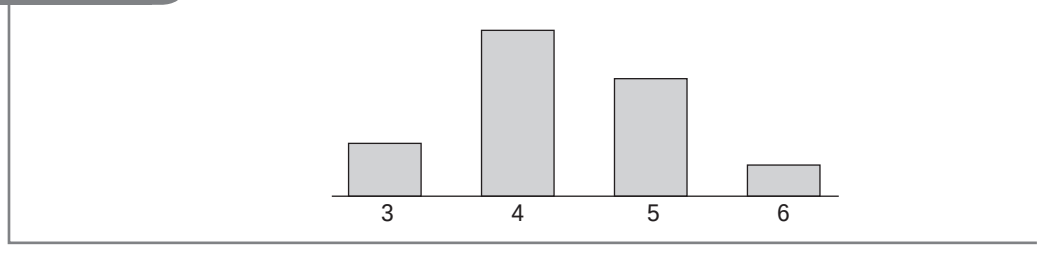
La moda

Para un grupo de datos, la ***moda*** M_0 *es el dato que más se repite.*

La moda es una medida que tiene sentido para cualquier tipo de escala.

La moda del conjunto de datos 3, 3, 4, 4, 5, 5, 5, 4, 4 y 6 es 4.

FIGURA 2.1 *La moda es 4*



Un conjunto de datos puede tener una moda (unimodal), dos modas (bimodal), etcétera, o puede no tener ninguna moda, si la frecuencia es la misma para todos los datos.

La media aritmética

Si $x_1,\ldots, x_n$ es un grupo de datos que corresponden a una muestra de una población, la ***media aritmética*** *o simplemente la* ***media*** *de estos valores es el número*

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

La idea es que este número represente a cada uno de los valores del conjunto. Esto sucederá en la medida en que la distribución sea simétrica y tenga poca dispersión.

Si el conjunto de datos corresponde a toda la población la media aritmética se denota con μ.

A μ también se le llama *media poblacional* mientras que a $\bar{x}$ (que corresponde a la muestra) se le llama *media muestral*.

La media tiene sentido para datos medidos con la escala de intervalo y de razón.

De no indicar lo contrario, los conjuntos de datos que usaremos a menudo corresponderán a muestras de una población.

EJEMPLO. *Tiempo promedio en realizar una tarea*

Los siguientes datos corresponden al tiempo, en minutos, que utilizan 30 empleados en realizar una tarea.

4.1 2.2 6.7 2.9 5.0 3.2 3.7 3.4 4.0 7.5 3.1 8.0 2.4 7.6 6.2 8.7 4.5 4.7 6.1 3.5 2.7 4.5 3.9 5.1 3.0 4.6 4.6 3.6 4.0 3.7

La media de estos datos es $\bar{x} = \frac{4.1 + \ldots + 3.7}{30} = 4.5733$.

Así, el tiempo medio que utiliza cada empleado en realizar la tarea es de 4.6 minutos, aproximadamente.

La pregunta que surge es ¿qué tan bien representa 4.5733 a cada uno de estos datos? La respuesta se tendrá cuando se estudien las medidas de dispersión alrededor de la media.

Media para datos agrupados. Media ponderada

Si los valores $x_1,\ldots, x_k$ se repiten $n_1,\ldots, n_k$ veces, respectivamente, entonces la media de los valores puede calcularse con:

$$\bar{x} = \frac{x_1 n_1 + \ldots + x_k n_k}{n_1 + \ldots + n_k} = \sum_{i=1}^{k} f_i x_i$$

donde $f_i = n_i/(n_1 + \ldots + n_k)$ es la frecuencia relativa del valor x_i.

En este caso, a $\bar{x}$ se le llama *fórmula de la media para datos agrupados*, y corresponde a una *suma ponderada de los datos*. La ponderación de cada dato es su frecuencia relativa, f_i. La ponderación indica la importancia del dato en el conjunto.

EJEMPLO. *Tiempos de servicio*

Los tiempos de servicio, en años, en la empresa ABC de seguros, que emplea a 150 personas, están distribuidos de la manera como se indica en la siguiente tabla.

TABLA 2.1 *Tiempos de servicio*

Tiempo de servicio en años	*Frecuencia*
4	50
5	70
6	30

La media de los tiempos de servicio es $\bar{x} = \dfrac{4(50) + 5(70) + 6(30)}{150} = 4.86$ años.

En promedio, las 150 personas tienen 4.86 años de servicio en la empresa ABC.

Cuando la distribución de valores se expresa en una tabla de intervalos de clase, *la media ponderada se aproxima usando las marcas de clase x'_i, en lugar de los valores x_i.*

$$\bar{x} \approx \frac{\sum_{i=1}^{k} (x'_i) n_i}{n_1 + \ldots + n_k}$$

En general, si para indicar la importancia de los valores $x_1, x_2,\ldots, x_k$ se asocian las ponderaciones $w_1, w_2,\ldots, w_k$, respectivamente, la *media ponderada* se define como el número:

$$\frac{w_1 x_1 + \ldots + w_k x_k}{w_1 + \ldots + w_k}$$

EJEMPLO. *Aumento de precios*

El aumento del precio de un bien puede ser insignificante para unas personas pero muy importante para otras. Si el aumento del precio del consumo de electricidad es de 20%, el de la vivienda es de 10% y el de la alimentación es de 3%, entonces se tendrá que la media aritmética de los aumentos es 11%; sin embargo, a una persona que dedica el 1% de su sueldo a consumo de electricidad, el 9% a vivienda y el 90% a alimentación, el promedio de aumento que le interesa es la media ponderada de los porcentajes de aumento, con las ponderaciones respectivas: 1/100, 9/100 y 90/100.

Propiedades de la media

Algunas propiedades importantes de la media son las siguientes.

a) *Si el conjunto de m datos,* $x_1,\ldots, x_m$ *tiene media* $\bar{x}$ *y el conjunto de n datos,* $y_1,\ldots, y_n$ *tiene media* $\bar{y}$*, entonces el conjunto de los m + n datos:* $x_1,\ldots, x_m, y_1,\ldots, y_n$*, tiene media* $\dfrac{m\bar{x} + n\bar{y}}{m + n}$.

Si 20 empleados ganan 2,000 dólares mensuales en promedio y 30 obreros ganan 500 dólares mensuales en promedio, entonces los 50 trabajadores ganan 1,100 dólares en promedio.

b) $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$. Es decir, la suma de las desviaciones de los datos respecto de la media es 0.

La mediana

Al calcular la media de un conjunto de datos que forman una muestra, los valores extremos (valores muy grandes o muy pequeños en relación con los demás) pueden influenciar fuertemente en el resultado, desvirtuando de este modo la utilidad de la media aritmética como valor que caracteriza a los datos.

TABLA 2.2 *Salarios*

Sueldos (x_i)	*Frecuencia* (n_i)
100	4
120	5
2,000	1
3,000	1

En la Tabla 2.2 se indica la distribución del sueldo, en dólares, de 11 personas.

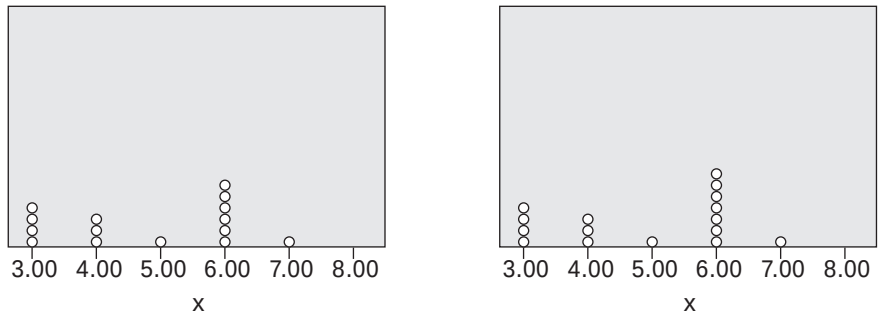
Según esta tabla, la media de los sueldos en la muestra es 545.45 dólares; sin embargo, 9 de las 11 personas tiene una remuneración menor o igual a 120. Ello indica que la media no representa bien a los 11 sueldos. Los valores extremos 2,000 y 3,000 han influido fuertemente en el resultado (si se construye el polígono de frecuencias de la tabla anterior, podrá notarse que este tiene un sesgo con cola a la derecha).

Si una distribución de datos tiene un sesgo (con cola a la derecha o con cola a la izquierda) es mejor utilizar una medida de tendencia central que no sea sensible a los valores extremos. Una de estas medidas es la *mediana*.

> *Para un grupo de n datos ordenados que conforman una muestra, la* ***mediana*** M_e *es el valor que ocupa la posición central si n es impar, y es el promedio de los dos datos centrales si n es par.*

La mediana del conjunto de datos: 3, 3, 3, 3, 4, 4, 4, 5, 6, 6, 6, 6, 6, 6 y 7 es 5, y la mediana del conjunto de datos: 3, 3, 3, 3, 4, 4, 4, 5, 6, 6, 6, 6, 6, 6, 6 y 7 es 5.5.

FIGURA 2.2 *En* (*a*) *la mediana es* 5 *y en* (*b*) *la mediana es* 5.5

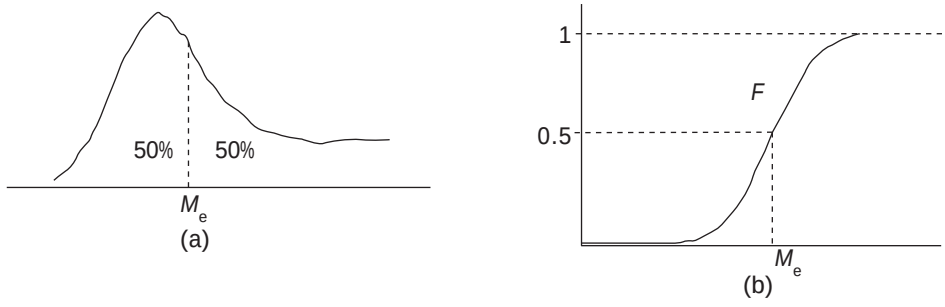


La mediana M_e es una medida de centralización resistente. Al variar uno de los datos que no sean los de la posición central, la mediana no varía; por ello se usa como resumen numérico de grupos de datos cuya forma del polígono de frecuencias no es simétrica.

La mediana M_e es un valor que está en el centro de los datos ordenados. Esto generalmente no ocurre con la media ni con la moda. Sin embargo, cuando el polígono de frecuencias es simétrico, los tres valores coinciden. Cuando el polígono es asimétrico, la media se ubica generalmente hacia el lado de la cola más larga.

Los siguientes gráficos (Figura 2.3) proporcionan una idea de la posición de la mediana. A partir de la ojiva *F* de un grupo de datos se puede obtener una aproximación de la mediana. Hasta la mediana, la frecuencia acumulada relativa es 0.5, hasta la mediana se acumula el 50% de los datos.

FIGURA 2.3 *Representación de la mediana*



La media geométrica. Inflación promedio mensual

La media geométrica

A menudo se trabaja con cantidades que cambian con el tiempo, y en muchos de estos casos es útil encontrar una tasa promedio de cambio, como por ejemplo la tasa de crecimiento promedio de los precios de la canasta familiar, en cinco meses consecutivos.

Para el cálculo de la tasa promedio de cambio se usa la *media geométrica.*

> *La* ***media geométrica*** *de los* ***valores positivos*** $x_1, x_2, \ldots, x_n$ *se define como:*
>
> $$G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

EJEMPLO. *Tasa promedio de rendimiento*

Si se invierte en un negocio la cantidad de $ 100, y al finalizar el primer mes esta inversión se reduce al 50%, y luego, al finalizar el segundo mes, se recupera a su valor original de $ 100, se tendrá que al término de los dos meses la tasa de rendimiento es 0.

Formalmente:

Al finalizar el primer mes se tiene (1 – 0.5)100.

Al finalizar el segundo mes se tiene (1 + 1)(1 – 0.5)100 = 100.

El cambio, al finalizar los dos meses, ha sido 0.

Los valores –0.5 y 1 son las *tasas de rendimiento* durante el primero y segundo mes, respectivamente.

Se desea ahora encontrar la tasa promedio mensual de rendimiento de la inversión. Si se usara la media aritmética para este fin, la tasa promedio sería $(-0.5 + 1)/2 = 0.25$.

Si la tasa promedio es 0.25, al finalizar el segundo mes el capital invertido al inicio se transformaría en $(1 + 0.25)(1 + 0.25)100 = 156.25$. Valor que no corresponde a la realidad.

Para encontrar la tasa promedio de rendimiento adecuada, R, para los dos meses, hacer lo siguiente.

Si R es la tasa de rendimiento promedio mensual, esta debe representar a cada una de las tasas mensuales de crecimiento, y así se tendrá que:

al finalizar el primer mes se tendrá $(1 + R)100$ *y*
al finalizar el segundo mes se tendrá $(1 + R)(1 + R)100$.

Esta última expresión debe ser igual a $(1 + 1)(1 - 0.5)100$, luego,

$$(1 + R)^2 100 = (1 - 0.5)(1 + 1)100$$

Se tendrá entonces que la tasa promedio es $R = \sqrt[2]{(1 - 0.5)(1 + 1)} - 1 = 0$.

En general, si se invierte una cantidad A durante n periodos y las tasas de rendimiento son, respectivamente, $R_1, \ldots, R_n$, entonces:

$$R = [(1 + R_1)\ (1 + R_2)\ \ldots\ (1 + R_n)]^{1/n} - 1$$

expresa la tasa de rendimiento promedio por periodo.

EJEMPLO. *Ejemplo de aplicación para el cálculo de la inflación mensual y acumulada*

La ***inflación mensual o tasa de inflación mensual*** R_i, en el mes i, es la ***variación*** del *"precio PC_i, de la canasta familiar"*, en el mes i, con respecto del mes anterior $i - 1$. Expresado en porcentaje se tiene:

$$R_i = \frac{PC_i - PC_{i-1}}{PC_{i-1}}\ 100\%$$

Si la inflación mensual fue del 10% en el mes i, entonces, respecto del mes anterior, el precio de la canasta familiar aumentó el 10%.

Nótese que si $R_i = \left(\frac{PC_i}{PC_{i-1}} - 1\right) \bullet 100\%$, y que $\frac{PC_i}{PC_{i-1}} = \frac{R_i}{100} + 1$.

Si en el mes de noviembre de 1990 el precio de la canasta fue $ 865 y en el mes de diciembre del mismo año fue $ 1,070, entonces, en diciembre, la variación porcentual mensual de la canasta fue igual a:

$$R_{Dic} = \left(\frac{1070 - 865}{865}\right) 100\% = 23.69\%$$

La *inflación acumulada* entre el periodo *a* y el periodo *i* ($a \leq i$), expresada en porcentajes, es el valor:

$$IAC_{i/a} = ((PC_i - PC_{a-1})/PC_{a-1})100\% = (PC_i/PC_{a-1})100\% - 100\%$$

Considerando que $PC_i/PC_{a-1} = \frac{PC_i}{PC_{i-1}}\frac{PC_{i-1}}{PC_{i-2}} \ldots \frac{PC_a}{PC_{a-1}}$ y que $\frac{PC_k}{PC_{k-1}} = \frac{R_k}{100} + 1$, se tiene que la inflación acumulada entre el periodo *a* y el periodo i ($a \leq i$), en términos de las inflaciones mensuales, es:

$$IAC_{i/a} = \left[\left(\frac{R_i}{100} + 1\right)\left(\frac{R_{i-1}}{100} + 1\right) \ldots \left(\frac{R_a}{100} + 1\right)\right] 100\% - 100\%$$

EJEMPLO. *Inflación promedio e inflación acumulada*

Según el Instituto Nacional de Estadística e Informática, la inflación mensual en porcentajes, entre mayo de 1990 y septiembre de 1990, fue como lo indica la Tabla 2.3.

TABLA 2.3 *Inflación entre mayo de 1990 y septiembre de 1990*

Mayo	*Junio*	*Julio*	*Agosto*	*Septiembre*
28.0	23.1	24.6	25.1	26.0

La inflación promedio mensual para los meses comprendidos entre mayo de 1990 y septiembre de 1990 fue:

$$R = ((1 + 0.280)(1 + 0.231)(1 + 0.246)(1 + 0.251)(1 + 0.260))^{1/5} - 1 = 0.2534$$

En porcentaje, la inflación promedio fue 25.34%.

La inflación acumulada entre los meses de mayo y junio de 1990 fue igual a:

$$\left(\frac{23.1}{100} + 1\right)\left(\frac{28.0}{100} + 1\right)100 - 100 \cong 0.5757 \ (57.57\%)$$

2.3 Medidas de dispersión

Las medidas de dispersión indican la variación de los datos alrededor de una medida de tendencia central. Permiten verificar si determinadas medidas de tendencia central son significativas o no; es decir, si son confiables o no. Por ejemplo, cuando la dispersión de un grupo de datos numéricos es muy grande la media aritmética no tiene mucha significación, no representa bien los datos.

Las medidas de dispersión se usan algunas veces como medidas de riesgo. Las ganancias de una empresa que son extremadamente grandes pero también extremadamente pequeñas (incluso negativas) indican un riesgo muy alto para los accionistas y acreedores de una empresa.

El rango

> *El **rango** o recorrido de un conjunto de datos es la diferencia entre el dato mayor y el dato menor* (*longitud del intervalo en donde varían los datos*).

El rango del conjunto de datos:

23, 24, 24.5, 24.6, 24.7, 24.9, 25, 26.9, 27, 28, 100

es 100 − 23 = 77

El rango no indica la manera como están distribuidos los datos. Los conjuntos de datos *A* y *B*, que a continuación se indican, tienen rangos iguales; sin embargo, las distribuciones de sus datos son diferentes. En el primer conjunto los datos están concentrados en el lado derecho; en el segundo conjunto, están concentrados en el lado izquierdo.

A: 1 2 3 4 5 6 7 7 7 8 8 9 9 9 10 10 10 10 10

B: 1 1 1 1 1 1 1 2 2 2 2 2 2 3 4 5 6 7 8 9 10

La varianza y la desviación estándar

Una medida de dispersión muy importante es la ***varianza***. La varianza indica cómo están dispersos los datos respecto de su media. Esta medida explica gran parte de la información contenida en los datos. *Si* $x_1,\ldots, x_n$ *es un conjunto de n datos correspondientes a una muestra de una población, cuya media es* $\bar{x}$, entonces *su **varianza** se denota con* s^2 *y se define como el número no negativo.*

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Si el conjunto de datos corresponde a toda una población de tamaño *N*, la *varianza* se denota con σ^2 y se define como:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

A σ^2 se le llama también ***varianza poblacional*** mientras que a s^2 se le llama ***varianza muestral***.

En ambos casos, el valor de la varianza no corresponde a la misma escala de los datos, pues considera los cuadrados de las desviaciones y no las desviaciones mismas. Este es un problema que se resuelve si se considera como medida de dispersión a la raíz cuadrada de la varianza. Así resulta la desviación estándar, número que tendrá la misma escala que los datos.

> *Para la población y para la muestra, la raíz cuadrada de la varianza se llama* ***desviación estándar*** *poblacional y muestral, respectivamente.*

La desviación estándar se denota con *s* para el caso muestral y con σ para el caso poblacional.

Si el conjunto de datos formado por 1, 3, 5 y 7 corresponde a una población, entonces la media es $\mu = 4$ y la varianza σ^2 es $\frac{(1-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2}{4} = 5$.
La desviación estándar es $\sqrt{5} = 2.2360$.

El intervalo $[\mu - \sigma;\ \mu + \sigma] = [1.764;\ 6.2360]$ contiene a los valores 3 y 5; es decir, al 50% de los datos.

Generalmente la mayor parte de los datos está a una distancia igual a dos desviaciones estándar de la media. Pocos están a una distancia mayor a dos desviaciones estándar respecto de la media.

EJEMPLO. *Distribución del número de empleados*

La siguiente tabla indica cómo se distribuye el número *X* de empleados para una muestra de 34 empresas de calzado.

TABLA 2.4 *Número de empleados*

Número de empleados, x_i	3	5	7
Número de empresas, n_i	10	15	9

En promedio hay $\bar{x} = [10(3) + 15(5) + 9(7)]/[10 + 15 + 9] = 4.9412$ empleados por empresa.

La varianza del número de empleados es igual a:

$$s^2 = \frac{(3 - 4.9412)^2(10) + (5 - 4.9412)^2(15) + (7 - 4.9412)^2(9)}{10 + 15 + 9 - 1} = 2.299$$

EJEMPLO. *Tiempo dedicado a ver televisión*

Para una muestra de 28 niños, se han medido los tiempos, en horas por día, que estos dedican a ver televisión. La distribución respectiva es como se indica en la siguiente tabla. Hallar, en forma aproximada, la media y la desviación estándar de los tiempos que los 28 niños dedican a ver televisión cada día.

TABLA 2.5 *TV en casa*

Tiempo en horas	*Marca de clase* x_i	*Número de niños* n_i
[2, 4]	3	5
]4, 6]	5	7
]6, 8]	7	10
]8, 10]	9	4
]10, 12]	11	2

Solución

Usando las marcas de clase, se tiene que la media del tiempo que los 28 niños dedican a ver televisión se puede aproximar de la siguiente manera:

$$\bar{x} \approx \frac{\sum_{i=1}^{n} x_i n_i}{n} = \frac{(3)5 + (5)7 + (7)10 + (9)4 + (11)2}{5 + 7 + 10 + 4 + 2} = 6.36 \text{ horas, aproximadamente } (x_i \text{ es la marca de clase}).$$

De igual manera, la varianza del tiempo que los 28 niños ven televisión se aproxima con

$$s^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n - 1} \approx \frac{(3 - 6.36)^2 5 + (5 - 6.36)^2 7 + (7 - 6.36)^2 10 + (9 - 6.36)^2 4 + (11 - 6.36)^2 2}{27} =$$

5.349 y la desviación estándar es $\sqrt{5.349} = 2.3128$ horas, aproximadamente.

Los niños que conforman la muestra ven, en promedio, 6.36 horas de televisión por día con una desviación estándar de 2.3128 horas.

Propiedades de la varianza. Propiedad de Chebyshev

a) Para el caso de una población de tamaño N, la varianza se puede expresar como:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

La igualdad se obtiene al desarrollar la expresión $(x_i - \mu)^2$ en la relación

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}.$$

b) Si todos los datos son iguales a una constante c, la media es igual a c y, por tanto, su varianza es igual a 0 (no hay dispersión).

c) Si a cada uno de los datos $x_1,\ldots, x_n$ se le suma una constante b, entonces la media de los datos transformados $x_1 + b,\ldots, x_n + b$ es igual a la media de los datos originales más la constante b, mientras que su varianza es igual a la varianza de los datos originales.

d) Si cada uno de los datos $x_1,\ldots, x_n$ es multiplicado por una constante k, entonces la varianza de los datos transformados: $kx_1,\ldots, kx_n$ es igual a la varianza de los datos originales multiplicada por el cuadrado de la constante k.

e) La propiedad de Chebyshev indica lo siguiente:

> *Para un conjunto de datos con media μ y desviación estándar σ,* ***no importa cuál sea la forma de su distribución****, la proporción de ellos que caen en el intervalo* $[\mu - k\sigma, \mu + k\sigma]$ *es mayor o igual a* $1 - (1/k^2)$*. Esta propiedad se cumple para k mayor que 1.*

Así, en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ hay por lo menos el $(1 - (1/2^2))100\% = 75\%$ de los datos y en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$ hay por lo menos el $(1 - (1/3^2))100\% = 89\%$ de los datos.

EJEMPLO. *Consumo de gasolina*

Si en una refinería la producción de gasolina, durante N días, tiene una media de 150,000 galones con una desviación estándar de 1,000 galones, hallar la proporción de días cuya producción de gasolina se espera que esté entre 148,000 y 152,000 galones.

Solución

El intervalo [148,000, 152,000] es igual al intervalo [150,000 − 2(1,000), 150,000 + 2(1,000)].

Usando la propiedad de Chebyshev con $k = 2$, se tiene que la proporción de días cuya producción se espera que esté en el intervalo indicado es al menos igual a $1 - (1/2^2) = 0.75$.

Cuando la distribución tiene la forma de una campana, se aplica la regla empírica conocida como 68-95-99.

El coeficiente de variación

Es indudable que un grupo de datos que tiene media 50 y desviación estándar 10 tiene mayor dispersión que un grupo de datos que tiene media 5,000 y desviación estándar 10. Esto indica que la desviación estándar no es la única base para comparar la dispersión de dos grupos de datos. Es necesaria una medida relativa que tome en cuenta la dispersión con respecto a la magnitud de la media. Esta medida es el ***coeficiente de variación***, que se define como:

$$CV = \frac{s}{|\bar{x}|} \times 100\%$$

cuando se trata de valores de una muestra.

Con el coeficiente de variación se expresa, en porcentaje, la desviación estándar en "medias aritméticas".

De dos grupos de datos, el más homogéneo es el que tiene menor coeficiente de variación.

El coeficiente de variación se define también para el caso poblacional como

$$CV = \frac{\sigma}{|\mu|} \times 100\% \ .$$

En la práctica se considera que un grupo de datos tiene baja dispersión si su coeficiente de variación es menor que el 10%.

EJEMPLO. *¿Qué tipo de acciones comprar?*

Juan desea comprar acciones de la empresa A o de la empresa B, que cotizan en la bolsa de valores de la ciudad. Los informes indican que en la última semana estas acciones han rendido en promedio la misma cantidad. También se conoce que las desviaciones estándar de estos valores son: 4 dólares y 8 dólares para A y B, respectivamente. Considerando que en las situaciones financieras la fluctuación en torno a su rendimiento promedio es el *riesgo* de las acciones, es recomendable que Juan compre acciones de la empresa A, pues estas acciones tienen menor desviación estándar (una medida del riesgo es la desviación estándar).

Si los rendimientos promedios de las acciones no son iguales, para comparar los riesgos usar el coeficiente de variación.

Datos "normales". La regla 68-95-99

Se usa la expresión "datos normales" para referirse a los conjuntos de datos cuyos histogramas alcanzan su máximo en la media y que de manera simétrica decrecen a ambos lados de este punto. Los histogramas de esta distribución tienen la forma de una campana.

Para un conjunto de datos normales, cuya media y desviación estándar son μ y σ se cumplen las siguientes características empíricas:

1. Aproximadamente, el 68% de los datos está en el intervalo $[\mu - \sigma, \mu + \sigma]$.
2. Aproximadamente, el 95% de los datos está en el intervalo $[\mu - 2\sigma, \mu + 2\sigma]$.
3. Aproximadamente, el 99.7% de los datos está en el intervalo $[\mu - 3\sigma, \mu + 3\sigma]$.

La propiedad 3 permite afirmar lo siguiente:

4. Si *A* y *B* son el mínimo y el máximo, respectivamente, de un conjunto de "datos normales", la desviación estándar puede ser aproximada con $(B - A)/4$. Esta propiedad puede aplicarse aun cuando los datos no son normales si se tiene en cuenta la propiedad de Chebyshev.

Se acostumbra decir que los datos que están más allá de dos desviaciones estándar de la media son *infrecuentes* o *poco comunes.*

Datos tipificados o estandarizados

Generalmente los datos recolectados provienen de mediciones que tienen diferentes unidades de medida, dificultando las comparaciones. Para resolver este problema *los datos se estandarizan* como sigue.

Dado un conjunto de datos $x_1,\ldots, x_n$, la *estandarización* de estos consiste en expresar en desviaciones estándar el alejamiento de cada uno de ellos, respecto de la media $\bar{x}$:

$$\frac{x_1 - \bar{x}}{s}, \frac{x_2 - \bar{x}}{s}, \ldots, \frac{x_n - \bar{x}}{s}$$

Cada valor estandarizado es una *medida de posición* respecto de la media.

Si el valor estandarizado del valor de una variable es 2.2, entonces el dato está a 2.2 desviaciones estándar de la media.

La media y la varianza de los datos estandarizados son iguales a 0 y 1, respectivamente.

EJEMPLO. *Comparando las notas*

Se supone que en el aula A la media de las notas de Matemáticas Financieras es 13 y su desviación estándar es 2, mientras que en el aula B las notas del mismo

curso tienen media de 16 y desviación estándar de 1. Si un alumno del aula A tiene nota 12 y otro alumno del aula B tiene nota 14, entonces sus notas estandarizadas son: –0.5 y –2.0, respectivamente.

En el aula A el alumno con nota 12 tiene mejor posición que el alumno del aula B, que tiene la nota 14, pues –0.5 es mayor que –2.0.

Característica de concentración: índice de Gini

Para analizar el grado de desigualdad en el reparto de los salarios entre los distintos grupos de una población se usa el *índice de concentración de Gini*.

Para establecer el índice de concentración de Gini de un grupo de datos, se considera la Tabla 2.6, en donde se observa la distribución de los salarios de 58,058 empleados de una región.

TABLA 2.6 *Distribución de los salarios de* 58,058 *empleados de una región*

Inter. de clase	n_i	F_i	S_i	q_i
[100, 120[	2,413	0.0416	253,365	0.0293
[120, 130[	4,342	0.1164	525,382	0.0903
[130, 140[	8,642	0.2652	1,192,596	0.2284
[140, 150[	13,300	**0.4942**	1,888,600	**0.4473**
[150, 160[	14,500	**0.7440**	2,276,500	**0.7112**
[160, 170[	10,200	0.9196	1,652,400	0.9027
[170, 180[	4,093	0.9901	732,647	0.9876
[180, 190[	443	0.9978	81,955	0.9971
[190, 200]	125	1.0000	24,875	1.0000
Total	58,058		8,628,320	

En esta tabla, n_i es la frecuencia de asalariados en el *i-ésimo* intervalo de clase y F_i es la frecuencia acumulada relativa. La columna indicada con S_i expresa la suma total de los n_i sueldos que se encuentran en el respectivo intervalo. La columna indicada con q_i expresa la proporción de la masa total monetaria que han ganado los $n_1 + n_2 + \ldots + n_i$ primeros asalariados. Esto es:

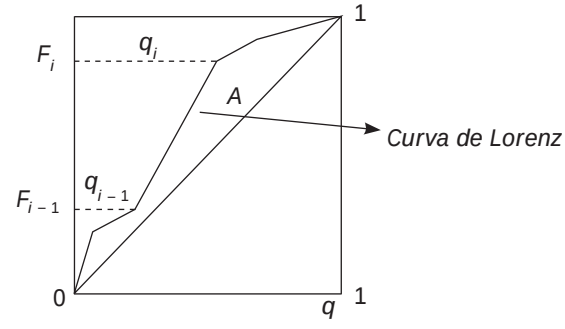
$$q_i = \frac{\sum_{j=1}^{i} S_j}{S}$$

en donde S es el total de masa monetaria ganada por todos los empleados.

En la Tabla 2.6 se puede leer, por ejemplo, que el 49.42% de los empleados reciben el 44.73% de la masa total de los salarios, que el 74.40% de los trabajadores perciben el 71.12% del total de los salarios, etcétera.

El máximo grado de concentración sucede cuando un solo asalariado percibe el total de la masa monetaria. El mínimo grado de concentración se obtiene cuando todos los asalariados se reparten por igual el total de la masa monetaria.

FIGURA 2.4 *Curva de Lorenz*



En un plano cartesiano, graficando y uniendo los puntos (q_i , F_i) se obtiene la *curva de concentración de Lorenz* (Figura 2.4).

El **índice de concentración de Gini**, que se denota con *G**, se define como *el doble del área de la región A, comprendida entre la curva de concentración de Lorenz y la bisectriz principal.*

Observaciones

1. El índice de Gini es un valor comprendido entre 0 y 1.
2. El mínimo valor del índice de Gini es igual a 0, y se obtiene cuando todos los grupos se reparten por igual la masa monetaria.
3. Si un grupo, digamos el primero, gana toda la masa monetaria, el índice de Gini es igual a 1.

Si la curva de concentración tiende a los lados del rectángulo, existe una *fuerte concentración*; un alto porcentaje de los asalariados se reparte una pequeña parte de los salarios mientras que un pequeño porcentaje de ellos se reparte casi la masa total de los salarios. Si la curva de concentración tiende a la bisectriz del cuadrado se dice que existe *una concentración débil*; el índice de Gini tiende a 0 y los salarios son casi idénticos para toda la masa de trabajadores.

El índice de Gini de un grupo de datos se encuentra calculando el área *A* de la región comprendida entre la bisectriz y la curva de concentración. Se tiene que:

$$\text{Área } A = 0.5 = \left[\sum 0.5(q_i + q_{i-1})(F_i - F_{i-1})\right]$$

Luego:

$$\text{Índice de Gini} = 2(\text{Área } A) = 1 - \left[\sum (q_i + q_{i-1})(F_i - F_{i-1})\right]$$

Para los datos registrados en la tabla de distribución de los 58,058 empleados el índice de Gini es igual a 0.4399. La concentración de los salarios es baja.

Percentiles de un grupo de datos. Rango intercuartil

Para definir los percentiles o los cuantiles de un conjunto de *n* datos, se comienza por ordenar los datos de tal manera que $x_{(1)}$ sea el menor, $x_{(2)}$ sea el que le sigue y así sucesivamente, hasta tener el mayor valor.

> *Para cada valor k entre* 1 *y* 100*, el* ***percentil k*** *es el dato* P_k *que está en la posición* (*n.k*/100) + 0.5*, si esta expresión es un número entero.*
>
> *Si* (*n.k*/100) + 0.5 *no es un número entero, el* ***percentil k*** *es el promedio de los dos datos cuyas posiciones son las más cercanos a dicho valor.*

El percentil tiene significación si la escala en que están expresados los datos es ordinal, de intervalo o de razón.

EJEMPLO. *Percentiles*

Ordenando el conjunto: 15 18 23 15 16 25 17 19 21 35 25, que corresponde al gasto semanal, en unidades monetarias (u.m.), que 11 jóvenes realizan en transporte, se tiene:

15 15 16 17 18 19 21 23 25 25 35

El percentil 50 es la observación cuya posición es el número entero (11 × 50/100) + 0.5 = 6.0. Observando los datos se tiene que el percentil 50 es 19. Es decir, la mitad de los jóvenes del conjunto tienen un gasto en transporte no superior a 19 u.m.

El percentil 25 es la observación cuya posición es (11)(25)/100 + 0.5 = 3.25. Como este valor no es entero, el percentil 25 es la media aritmética de los datos cuyas posiciones son 3 y 4. Entonces el percentil 25 es el promedio de 16 y 17. Esto es, 16.5. Es decir, el 25% de los 11 jóvenes gasta en transporte menos de 16.5 u.m.

El percentil 75 es la observación cuya posición es (11 × 75/100) + 0.5 = 8.75. Como este valor no es entero, el percentil 75 es la media aritmética de los datos que están en las posiciones 8 y 9, respectivamente.

El percentil 75 es el promedio de los datos 23 y 25. Esto es, 24. El 75% de los 11 jóvenes no supera las 24 u.m. de gasto en transporte.

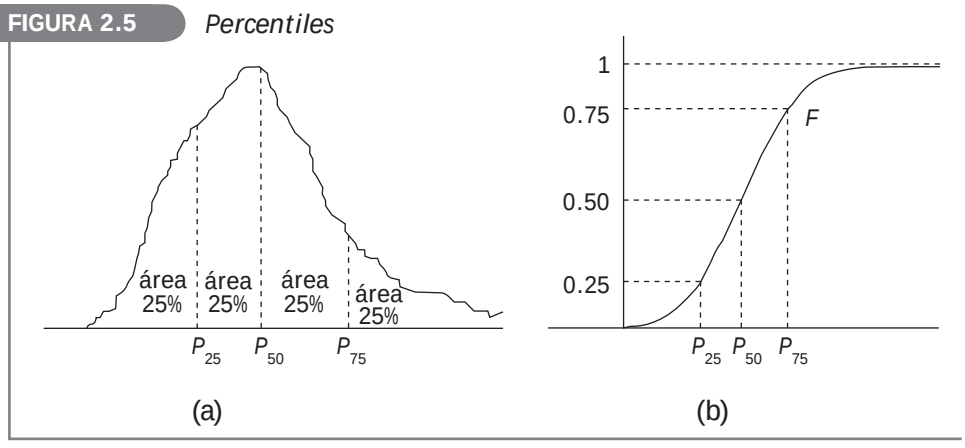
A los percentiles 25, 50 y 75 se les llama *primer cuartil*, *segundo cuartil* y *tercer cuartil*, y se les denota con Q_1, Q_2 y Q_3, respectivamente.

Nótese que el percentil 50 es igual a la mediana.

Al percentil *k* también se le llama *cuantil de orden k*/100. Así, el percentil 20 es el cuantil de orden 0.2.

Notemos también que, aproximadamente, el *k*% de los datos es menor o igual al percentil *k*.

FIGURA 2.5 *Percentiles*



Utilizando los percentiles es posible formar intervalos cuyas longitudes suelen usarse como índices de dispersión de los datos. El intervalo que más se usa es el que tiene como extremos el primer cuartil y tercer cuartil ($[Q_1, Q_3]$). La longitud de este intervalo se llama ***rango intercuartil***.

Las ventajas del rango intercuartil, como medida de dispersión, son: la rapidez en su cálculo y su estabilidad ante fluctuaciones de los datos extremos. Un dato extremo puede cambiar sensiblemente pero no afecta al rango intercuartil.

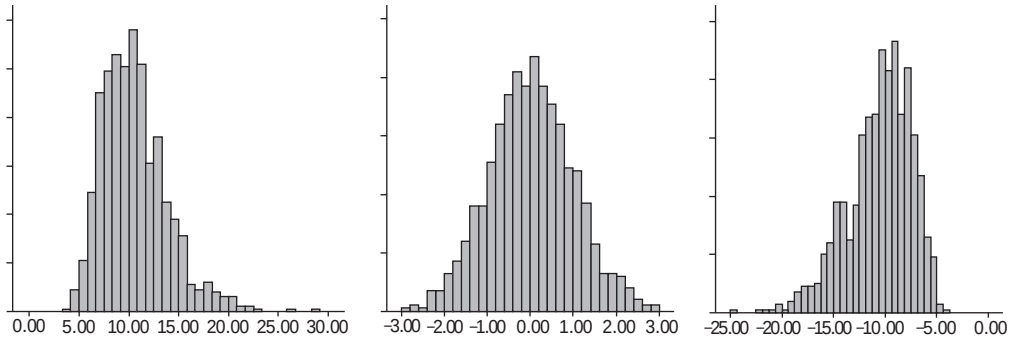
2.4 Medida de simetría

Una medida de la simetría de la distribución de un grupo de datos es *el* ***coeficiente de simetría de Fisher****.* Este coeficiente se define como:

$$s_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^2}, \quad \text{donde } s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Este coeficiente de simetría es igual a 0 cuando la distribución es simétrica. Si el coeficiente de simetría es diferente de 0, la distribución es asimétrica.

FIGURA 2.6 *Simetría y asimetría*



Asimétrica, s_k = 1.013 Simétrica, s_k = 0.050 Asimétrica, s_k = -0.744

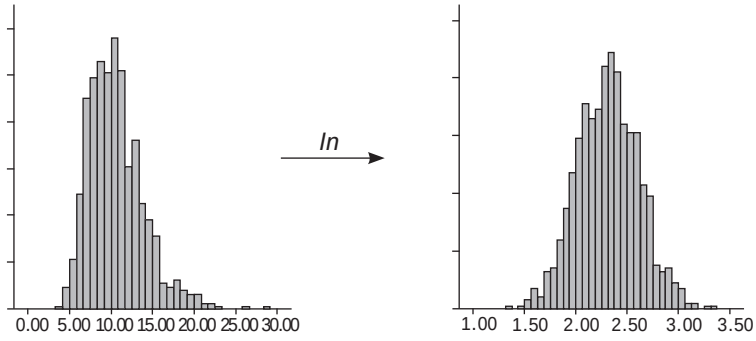
Observación

Mediante **transformaciones** adecuadas aplicadas a los datos se puede lograr, en muchos casos, que la distribución resultante sea aproximadamente como la de los datos normales. Generalmente se eleva cada dato a una determinada potencia *p* de la secuencia:

...−3, −2, −1, −1/2, −1/3, −1/4, 1/4, 1/3, 1/2, 1, 2, 3,...

o se aplica a cada uno de los datos el logaritmo natural.

FIGURA 2.7 *Transformación logaritmo natural*



Distribución de X Distribución de $Y = \ln(X)$

Transformaciones de este tipo son muy importantes, pues muchos procedimientos estadísticos que se aplican a los datos son válidos cuando se distribuyen "normalmente".

La elección de la potencia adecuada puede obtenerse, en primera instancia, por ensayo, observando el histograma de los datos transformados. En la Figura 2.7 se muestra la distribución de los datos de una variable X y de los datos transformados $Y = ln(X)$, cuya distribución se aproxima a la de los datos normales.

2.5 Medida de curtosis

El ***coeficiente de curtosis de Fisher*** de un grupo de datos, $x_1,\ldots, x_n$ es una medida del apuntamiento o agudeza de su polígono de frecuencias. Se define como:

$$k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4} - 3$$

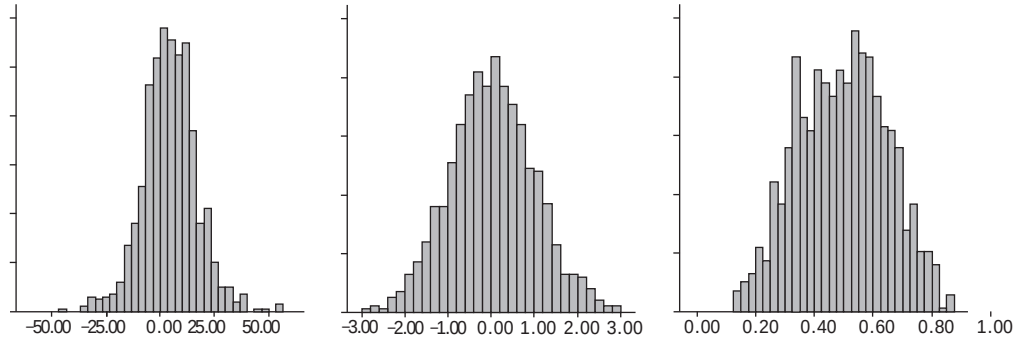
Si la curtosis es igual a 0, la distribución es *mesocúrtica*. Las distribuciones normales son mesocúrticas.

Si la curtosis es mayor que 0, la distribución es *leptocúrtica*. Las leptocúrticas son más apuntadas que las mesocúrticas.

Si la curtosis es menor que 0, la curva es *platicúrtica*. Las platicúrticas son más achatadas que las mesocúrticas.

Las distribuciones platicúrticas presentan menor concentración alrededor de la media, mientras que las leptocúrticas presentan mayor concentración alrededor de la media.

FIGURA 2.8 *Curtosis*



Leptocúrtica, $k = 1.362$ Mesocúrtica, $k = 0.010$ Platicúrtica, $k = -0.620$

EJEMPLO. *Tiempo de atención*

La distribución que aparece en la Tabla 2.7 corresponde a los tiempos, en minutos, que utiliza el empleado de una ventanilla de un banco para atender a 100 personas.

TABLA 2.7 *Tiempos de atención*

Tiempo de atención	*Número de personas*
[2.70, 2.80]	1
]2.80, 2.90]	2
]2.90, 3.00]	4
]3.00, 3.10]	7
]3.10, 3.20]	6
]3.20, 3.30]	9
]3.30, 3.40]	12
]3.40, 3.50]	10
]3.50, 3.60]	15
]3.60, 3.70]	11
]3.70, 3.80]	7
]3.80, 3.90]	6
]3.90, 4.00]	6
]4.00, 4.10]	4

El histograma y un cuadro resumen de las principales medidas aparecen a continuación.

El coeficiente de variación ($s/\overline{x}$) es igual a 9%, aproximadamente, e indica baja dispersión.

El coeficiente de simetría de Fisher, cercano al cero, indica una distribución simétrica. La distribución es platicúrtica (curtosis = −0.591).

FIGURA 2.9 *Histograma*

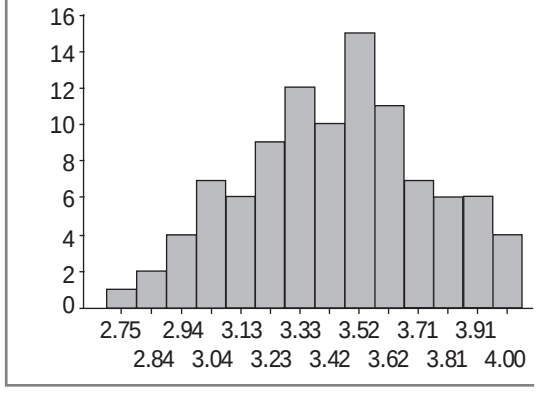


TABLA 2.8 *Descriptivos*

N (número de datos)	100
Mínimo	2.75
Máximo	4.05
Media	3.4710
Desviación estándar	0.3085
Varianza	0.095
Simetría	–0.096
Curtosis	–0.591
Coeficiente de variación	0.09

2.6 El gráfico de caja (*box plot*). Datos discordantes

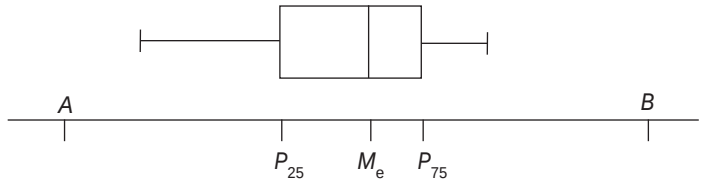
En los *gráficos de caja* se representan la mediana, el primer y tercer cuartil, lo que permite tener una idea de la distribución de los datos.

Un ***gráfico de caja*** es un rectángulo como el de la Figura 2.10. El lado izquierdo indica el percentil P_{25} (primer cuartil) y el lado derecho indica el percentil P_{75} (tercer cuartil). En el rectángulo se indica la ubicación de la mediana M_e mediante un segmento.

En un gráfico de caja se observa principalmente:

- la centralización
- la dispersión
- la simetría de la distribución

FIGURA 2.10 *Gráfico de caja*



La longitud, *d,* del largo de la caja corresponde al rango intercuartil.

A partir de cada uno de los puntos que representan a los percentiles 25 y 75 se determinan, respectivamente, el punto extremo izquierdo $A = P_{25} - 1.5(d)$ y el punto extremo derecho $B = P_{75} + 1.5(d)$.

Si un dato es menor que A o mayor que B, se considera que es un dato *discordante o atípico* (*outlier*, en el idioma inglés). Un dato discordante es un dato "que parece no ir con el resto". Todo dato de este tipo se representa en el gráfico de cajas con *. Por ejemplo, en el conjunto 6, 66, 70, 68, 67 y 69, que corresponde a un grupo de edades, obviamente, el dato 6 es discordante.

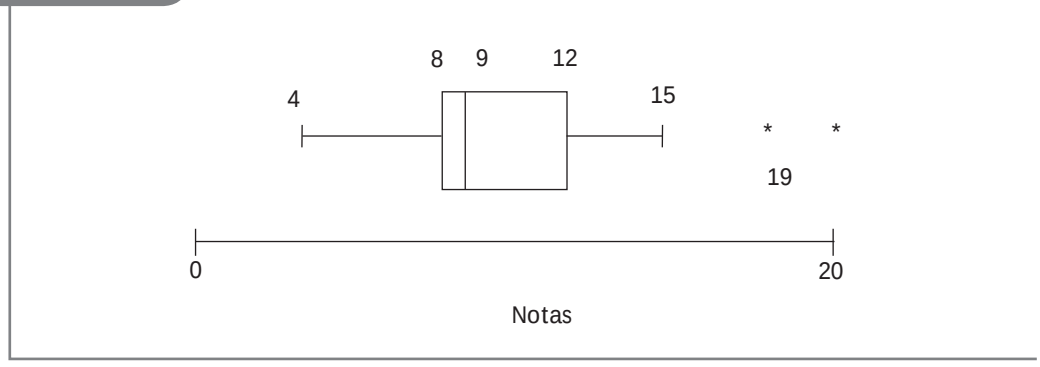
Para mejor conocimiento de la distribución de los datos, en cada lado de la caja se grafica un segmento o "bigote". El bigote del extremo izquierdo va del lado izquierdo de la caja al menor dato que está entre el punto *A* y el percentil 25. El bigote del extremo derecho va del lado derecho de la caja al mayor dato que está entre el percentil 75 y el punto *B*.

La influencia de un dato discordante en la variabilidad de los datos puede ser muy importante. En el anterior ejemplo, el rango del conjunto es 64; sin embargo, si no se considera el dato discordante el rango es 4.

EJEMPLO. *Las notas de los alumnos*

La Figura 2.11 representa un gráfico de caja correspondiente a 58 notas de los alumnos de un curso de Historia. Observando el gráfico de caja se puede indicar que el polígono de frecuencias es asimétrico con cola a la derecha. Existe mayor variabilidad en el conjunto de datos que están por encima de la mediana.

FIGURA 2.11 *Las notas de los alumnos*



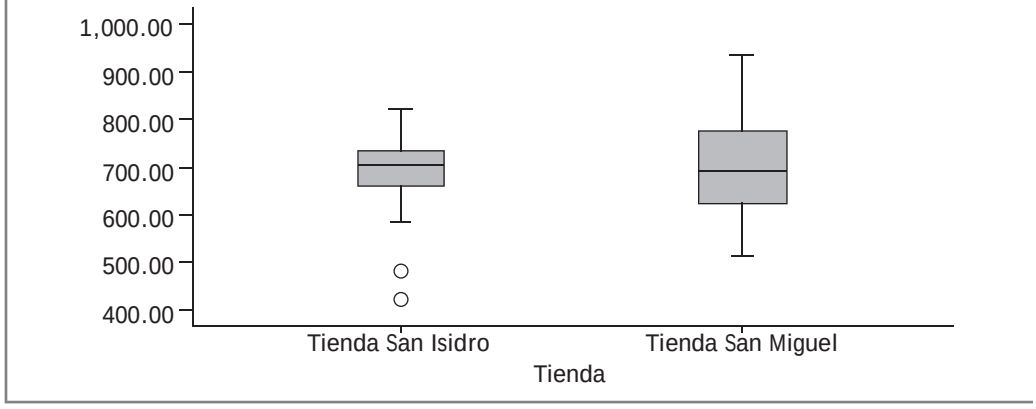
Se observa que:

- la mediana de los datos es 9.
- el percentil 25 es 8.
- el percentil 75 es 12.
- el rango intercuartil es 4.
- las notas 19 y 20 son datos discordantes.

EJEMPLO. *Gráficos de cajas y ventas diarias*

Gráficos (Figura 2.12) de cajas de las ventas diarias, en cientos de dólares, en dos tiendas de una cadena de comercialización de productos.

FIGURA 2.12 *Gráfico de caja, uno por cada tienda*



Aparición de datos discordantes

En muchos de los análisis a realizar será necesario determinar si cierto valor es realmente un dato discordante.

Un dato discordante generalmente aparece:

a) Por observación incorrecta, por anotación incorrecta o por introducción incorrecta del dato en el computador.
b) Cuando la observación es de una población diferente a la población de la cual viene el resto de las observaciones.
c) Cuando la medida es correcta pero el evento es raro. Por ejemplo, lo que pagó el seguro por la tragedia del 11 de septiembre.

Establecida la existencia de un dato discordante, sigue el problema de qué hacer con este valor. Se recomiendan algunas acciones: a) removerlo, b) transformarlo, c) dejarlo tal como está o d) realizar el análisis primero con el dato discordante y luego sin el dato discordante y comparar los resultados.

Remover el dato discordante significa no considerar toda la información, esconder el problema que podría existir. Un dato discordante puede ser, por ejemplo, un salario muy alto. Borrar este dato podría llevar a ignorar, tal vez, la existencia de un presunto evasor de impuestos.

La transformación de los datos puede consistir, por ejemplo, en calcular la raíz cuadrada de cada dato. Esta acción "jala" a los valores y hace que el dato discordante sea más consistente con el resto de los datos. El problema de este remedio es que los datos transformados pueden carecer de significado.

Dejar el dato discordante tal como está implica reconocer que en el conjunto de valores pueden aparecer estos datos por azar. Suponiendo que el diseño para la colección de datos está bien hecho, las medidas calculadas reflejarán lo que realmente sucede en la realidad.

Realizar el análisis con y sin los datos discordantes para luego comparar los resultados permite observar el efecto de estos valores y tener una mayor información del problema para tomar decisiones.

APLICACIÓN: El caso de la venta de ropa

Topirop es una empresa que tiene un centro de venta de ropa para caballeros situado en un centro comercial bien concurrido. Topirop desea que sus antiguos clientes que han dejado de comprar en los últimos meses continúen comprando con ellos; por ello ha enviado cupones promocionales de descuento a los clientes cuya última compra fue hace seis meses. A la vez que averigua si su promoción ha dado efecto, Topirop quisiera tener información de sus clientes para de este modo idear un mejor servicio. La empresa ha recabado información de 400 clientes correspondiente a las siguientes variables:

Edad del cliente
Género del cliente: hombre (0), mujer (1)
Medio de pago: en efectivo, con tarjeta de crédito
Valor de la compra
Tipo de cliente: sin cupón (0), con cupón (1)
Forma de pago: con dinero en efectivo (0), con tarjeta de crédito (1)
Intención de compra (comprará en Topirop en el presente mes): no (0), sí (1)

Usted puede sugerir el análisis de los datos que Topirop puede realizar después de recabar los datos de la encuesta.

LA ESTADÍSTICA EN LA EMPRESA

Hoteles Marion

Hoteles Marion es una empresa líder en hospitalidad; tuvo su origen en un negocio familiar que se dedicaba a la venta de postres en un pequeño lugar de la ciudad La Luz, en el año 1940. Hoy en día la empresa sigue siendo familiar, posee 60 hoteles de 5 estrellas y maneja propiedades de vacaciones en 40 países del mundo.

Hoteles Marion tiene aproximadamente 100,000 empleados, y en varios de los países en donde se desarrolla ha sido declarada como la empresa en la que mejor se trabaja.

El constante entrenamiento de los empleados para su mejor desempeño y la convicción del espíritu de servicio son las principales características que, desde su creación, han forjado la identidad del personal de la empresa.

Como en toda la industria hotelera, el conocimiento de sus clientes (de dónde provienen, cuánto gastan, cuánto tiempo permanecerán alojados, etc.), así como de su mercado, es la preocupación constante de Marion.

Marion crea constantemente campañas directas a través del correo electrónico, promociones de planes estacionales, propaganda personalizada, la clasificación de sus clientes en segmentos con perfil bien definido, la determinación con precisión del número de habitaciones para reservas, etc.

La base de datos de sus clientes, que la empresa ha formado a lo largo de estos años, es la fuente principal que Marion utiliza para elaborar sus estrategias de mejor servicio. En esta labor la estadística es la herramienta fundamental.

EJERCICIOS

1. Calcular la media, la mediana y la moda de los datos 500, 400, 600, 400, 200, 160, 380, 400, 180 y 420, que corresponden al tiempo, en minutos, que 10 usuarios de teléfonos celulares emplearon en el mes pasado. Indicar el percentil 70 y el percentil 90. Comentar los resultados. ¿Cuál de los promedios calculados describe mejor el centro de la distribución de los datos?

2. La siguiente tabla presenta la información de 53 pequeñas empresas de construcción de acuerdo al número de empleados que tienen.

Número de empresas	*Número de empleados*
11	12
16	13
17	14
9	15

   Usar la información de la tabla para calcular la media, la mediana y la desviación estándar del número de empleados.

3. El siguiente conjunto de datos corresponde a una muestra de los valores de una acción determinada en la bolsa de valores de una región, en unidades monetarias:

   90 63 20 18 12 60 24 28 14 11 85 29 25 8 10 86 16 25 6 11 80 16 20 16 6

   a) Calcular la media $\bar{x}$, la mediana y la desviación estándar s de los valores.

   b) ¿Qué porcentaje de datos está en el intervalo $[\bar{x} - 2s, \bar{x} + 2s]$?

   c) Observando los valores de la media y la mediana, ¿se podría decir que la información corresponde a un conjunto de datos normales?

4. En un grupo de 30 datos cuyos valores están entre 4 y 8, 15 de ellos tienen valores menores o iguales a 5. Si 3 valores iguales a 30 son agregados a los anteriores, calcular, aproximadamente, la moda, la media y la mediana de los 33 valores.

5. Las estaturas de un grupo de personas tienen media 1.65 m y desviación estándar 7 cm. ¿Considera usted que una persona cuya estatura es 1.80 m es inusualmente alta?

6. Los datos siguientes registran los tiempos, en minutos, que demora una oficina en dar trámite a 50 documentos que ha recibido.

   400 392 358 304 108 156 438 60 360 168 448 224 576 384 194 216 120 208 232 72 264 168 128 256 72 136 168 308 340 64 480 114 80 246 224 184 104 112 184 152 152 536 224 464 72 152 168 288 264 208

a) Construir el histograma de frecuencias relativas. Graficar el polígono de frecuencias relativas. Indicar las características de la distribución. ¿Se podría decir que los datos presentados son normales?

b) Calcular la media, varianza, desviación estándar y el coeficiente de variación de los datos.

c) Calcular la mediana, utilizando directamente los datos y usando la ojiva de frecuencia acumulada relativa.

d) Calcular los percentiles P_{10}, P_{25}, P_{75} y P_{90}.

e) Graficar el *box plot* de los datos.

f) Hallar el porcentaje de tiempos que están entre 200 y 400 minutos.

7. En cierto lugar se construyó una planta de tratamiento de aguas con una capacidad de procesamiento de 4,500,000 galones por día. La demanda de agua de esta planta, durante un mes en miles de galones por día, fue como sigue:

   4887 5535 4450 4852 5330 4448 4817 5152 4377 4816 5152 4363 4784 5142 5058 4763 4289 4188 4737 5041 5035 4724 4057 3992 4670 4998 3918 4666 4993 4525

   a) ¿Se puede considerar que la distribución de estos datos es simétrica?

   b) ¿Se puede considerar que estos datos son normales?

   c) Usando la frecuencia acumulada, indicar con qué frecuencia la demanda excedió la capacidad de procesamiento.

8. La siguiente tabla registra la distribución del número de accidentes por mes que ocurren en una empresa productora de alimentos.

Número de accidentes	***Número de meses observados***
0	3
1	6
2	4
3	5
4	2
5	1

   a) Calcular la media y la varianza de los datos.

   b) ¿Considera usted que han ocurrido datos de manera inusual?

9. Una regla de decisión para fijar el límite de velocidad en una carretera supone que el límite conveniente puede sobrepasarse en 15% de las veces. Siguiendo esta regla, ¿cuál límite de velocidad se recomienda para una carretera de la cual se ha recabado la siguiente información?

Velocidad en km/hora	*Vehículos*
[30, 40]	1
]40, 50]	3
]50, 60]	18
]60, 70]	20
]70, 80]	30
]80, 90]	20
]90, 100]	10
]100, 110]	5
]110, 120]	3
Más de 120	1

10. El número de computadoras que se malogran por día en una empresa tiene como media 6 y de mediana 10.

    a) Indicar el número que debería reportarse para indicar la situación descrita.

    b) ¿A qué se deberá la discrepancia observada entre la media y la mediana?

11. El 70% del personal de una compañía son varones y el resto, mujeres. El promedio de los sueldos del grupo de varones es $ 700 y el promedio del grupo de mujeres es $ 600. Hallar el sueldo promedio del personal de la compañía.

12. La media y la varianza de los tiempos $x_1,\ldots, x_n$, utilizados en realizar n tareas similares, son: 14 y 2.89, respectivamente. El costo por realizar cada tarea es $y_i = 20 + 0.5x_i + 0.1x_i^2$. Hallar la media de los costos, aproximadamente.

13. En una empresa pública, el promedio de los sueldos de los obreros es 40 unidades monetarias (u.m.), y el de los empleados 50 u.m. Si la empresa decide aumentar 20 u.m. a cada empleado y obrero, hallar el promedio general de los sueldos actuales (considerando el aumento), suponiendo que el número de obreros es 10% del número de empleados.

14. Para una población de 800 datos, la media aritmética y la desviación estándar halladas son 9.496 y 0.345, respectivamente. Una revisión de los resultados mostró que en lugar del valor 9.56 se introdujo 1.56. Recalcular la media aritmética y la desviación estándar.

15. Un banco ha clasificado a sus clientes según el monto de sus depósitos. En la clase A ubicó a los clientes de mayor monto, en la clase C ubicó a los de menor monto y el resto fueron ubicados en la clase B. Los de la clase A constituyen el 30%, los de la B el 40% y los de la C el

30%. El banco cobra una comisión fija de $ 3 por mantenimiento más una comisión variable de 4%, 3% y 2% del monto del depósito, según estén en C, B y A, respectivamente. Si el monto promedio de todos los depósitos en el banco es $ 1,250 para los de clase A, de los de la clase B es $ 1,100 y de los de la clase C es $ 500, hallar la comisión promedio que el banco recibirá por cada uno de sus clientes.

16. Treinta ladrillos de cierto tipo se sometieron a presión para determinar su calidad. La resistencia, en kg/cm^2, de cada uno de los ladrillos fue como sigue:

    107 104 110 125 132 137 125 122 114 120 115 117 129 131 128 129 126 127 126 129 131 106 131 132 116 117 112 104 122 139

    a) Graficar un histograma de frecuencias y comentar la variabilidad observada.

    b) Hallar la media, la varianza y el índice de simetría de la distribución.

    c) El secado de los ladrillos se realiza exponiéndolos al sol. Los correspondientes días de secado para los 30 ladrillos de la muestra anterior fue como sigue:

       7 8 7 12 12 15 15 12 11 12 13 15 15 13 15 13 15 15 8 11 11 15 15 10 14 15 14 13 15 13

    Los expertos en la materia afirman que, a mayor número de días de secado, la resistencia es mejor. ¿Validan los datos la afirmación de los expertos?

17. Se observaron los siguientes tiempos (en minutos), utilizados por 30 operarios en hacer una tarea similar:

    7.0 9.0 11.4 7.2 10.2 13.5 17.0 14.0 14.5 8.0 9.1 9.4 13.1 8.5 10.4 15.5 12.0 11.0 11.2 9.6 9.2 9.5 15.6 8.4 10.8 13.0 12.5 12.4 10.5 7.8

    Construir una distribución de frecuencias con cinco intervalos de igual longitud, y a partir de dicha distribución estudiar, usando medidas estadísticas adecuadas, las siguientes afirmaciones:

    a) Calculando una medida central representativa de la distribución hallada, se deduce que el tiempo promedio de ejecución de la tarea por operario llega a superar los 11 minutos.

    b) La mitad de los operarios no demoró más de 10.8 minutos.

    c) Un operario cualquiera nunca demoró un tiempo mayor que el tiempo medio en más de dos desviaciones estándar.

18. Durante cuatro años consecutivos un banco aumentó su capital en 5%, 8%, 11% y 15%, respectivamente. Indicar la tasa de aumento promedio anual del capital.

19. En cierto lugar, en 1992 la población fue 18,000,000, en 1993 fue 20,000,000 y en 1994 fue 25,500,000. Determinar la tasa de crecimiento promedio de la población en dichos tres años. Establecer por qué la media aritmética no es adecuada para indicar el cambio promedio de crecimiento.

20. En un trayecto de 300 kilómetros, el rendimiento de tres automóviles por galón de gasolina fue como sigue:

Automóvil	*Kilómetros por galón*
A	60
B	40
C	50

a) Hallar el rendimiento promedio de kilómetros por galón de gasolina.

b) Si a cada uno de los tres automóviles se les coloca cinco galones de gasolina y realizan el recorrido hasta que la gasolina se agota, ¿cuál es el rendimiento promedio de kilómetros por galón de gasolina?

21. En una experiencia sobre consumo de gasolina se examinaron 20 automóviles en un recorrido de 100 kilómetros, tanto en ciudad como en carretera. Los datos obtenidos corresponden al rendimiento en kilómetros por galón y aparecen a continuación:

Ciudad: 43, 55, 65, 45.5, 46.8, 55, 62, 54, 50, 52, 48, 56, 47, 49, 50, 47, 48, 45.3, 49.5, 48.3

Carretera: 48, 60, 68, 59, 50, 58.5 65,4 55, 52, 54.5, 55, 58, 49, 62, 65, 57, 49, 55, 60, 62

Comparar las distribuciones de ambos conjuntos de datos e indicar las conclusiones a las cuales se puede llegar respecto al rendimiento de los vehículos probados en la ciudad y la carretera. Indicar si es posible extender los resultados para todos los vehículos.

22. Para obtener tiras de tela cuya especificación indica que la longitud en milímetros debe estar en el intervalo [495, 505] se usa una cortadora, la cual es manejada por un operador. Con la finalidad de analizar si este proceso satisface estas especificaciones, se obtuvieron las mediciones de 60 cortes que a continuación se indican:

495.44 499.81 498.62 498.19 490.72 491.16 498.41 508.16 499.06 498.41 498.29 494.19 497.57

507.13 495.63 497.76 491.07 492.37 505.06 503.67 502.91 500.07 495.59 507.58 498.66 499.92

506.81 498.92 498.40 503.47 495.09 499.98 505.28 493.07 498.59 496.10 498.42 494.30 505.73

506.77 493.88 499.83 501.99 500.13 502.03 500.63 498.41 504.13 500.33 500.29 501.53 494.03

488.03 503.97 505.37 502.04 495.14 497.74 511.73 496.63

a) Hallar la media y la mediana, y usando estas, indicar si la tendencia central del proceso es adecuada.

b) Calcular la desviación estándar del proceso.

c) Obtener el histograma de las medidas e interpretar el gráfico resultante.

d) Observando los resultados anteriores, indicar si el proceso de corte es adecuado.

23. Para conocer el nivel de satisfacción de los clientes de un banco se aplicó a 50 de ellos una encuesta que consistía de 10 preguntas. Cada una de las preguntas evaluaba un aspecto del servicio que el banco proporcionaba a los clientes, y se sumaron los puntajes para obtener un puntaje total. Los resultados fueron como sigue:

    42 24 71 76 66 45 53 95 49 52 75 74 43 86 67 50 58 77 75 28 85 53 66 58 63
    41 76 55 49 33 74 56 50 50 32 53 33 37 68 73 54 68 88 64 75 43 47 47 59 44

    a) Calcular las medidas de tendencia central y de dispersión. Comentar los resultados.

    b) Hacer el histograma correspondiente. Interpretar.

24. Dos famosos estadísticos, Freedman y Diaconis, mostraron que la longitud D de los intervalos de clase que hacen mínima la mayor distancia entre el histograma de frecuencias relativas y la *función de densidad* (función que suaviza al histograma) está determinada por:

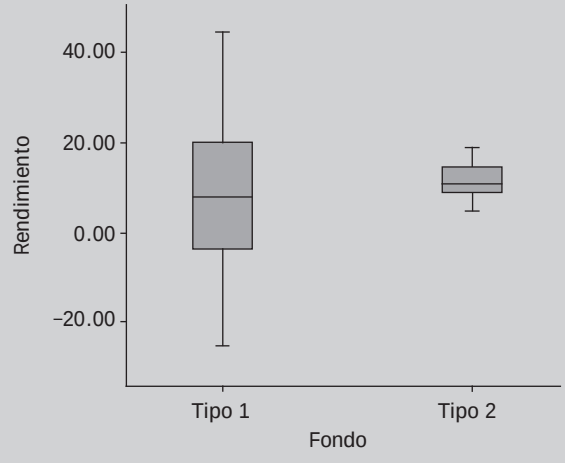
$$D = (P_{75} - P_{25})\left(\frac{\log n}{n}\right)^{1/3}$$

    De este modo el número de intervalos de clase a tomar es $(x_{max} - x_{min})/D$).

    Usando la regla de Freedman y Diaconis graficar el histograma para los siguientes datos:

    158 160 179 180 192 193 194 195 195 200 202 205 210 212 215 229 230 237
    240 244 245 247 259 270 301 301 315 321 337 415 434 444 485 496 529 537
    624 707 705 710

25. En la siguiente figura aparece la gráfica de cajas para los fondos de inversión de una administradora de fondos de pensiones durante el año 2006. La caja de la izquierda corresponde a los fondos de tipo 1 y la caja de la derecha corresponde a los fondos de tipo 2.



a) Describir cada una de las distribuciones de los fondos de cada tipo.

b) Comparar el riesgo de los dos tipos de fondos de inversiones.

26. El índice de Gini puede usarse también para analizar la concentración de habitantes de una determinada región. A partir de la siguiente información recogida en una cierta región, ¿podría decirse que existe alta concentración de habitantes en la región?

Habitantes	*Número de distritos*	*Población total*
[500, 1,000]	320	192,000
]1,000, 1,500]	130	143,000
]1,500, 2,000]	110	187,000
]2,000, 2,500]	80	192,000
]2,500, 3,000]	30	79,500
]3,000, 3,500]	10	34,500

RESPUESTAS A LOS EJERCICIOS

1. a) media = 364, mediana = 400, moda = 400. **2.** media = 13.45, mediana = 13, moda = 14, desviación est. = 1.01 **3.** media = 31.16, mediana = 20, desviación estándar = 27.82 **5.** Sí. **6.** b) media = 234.68, varianza = 17387.47, coeficiente de variación = 56.20% c) mediana = 208.00 d) percentil 10 = 72.80, percentil 25 = 134.00. **9.** Usar la frecuencia acumulada para determinar la respuesta. **10.** a) La mediana. **11.** 670. **12.** 46.889 aprox. **14.** media = 9.506, desviación estándar = 0.200 aprox. **15.** 100 aprox. **18.** 9.68% **20.** a) Para recorrer los 300 km A usa 300/60 galones, B usa 300/40 galones y C usa 300/50 galones. En promedio, para recorrer los 300 km se necesitan (1/3)(300/60 + 300/40 + 300/50). En promedio se necesitan 3/(1/60 + 1/40 + 1/50) galones para recorrer los 300 km. Esta última expresión se llama *media armónica de* 60, 40 y 50. b) Usar la media aritmética.

CAPÍTULO 3

Relación entre variables. Medidas de correlación y asociación

Adrien Legendre

Adrien Legendre nació en París, Francia, en 1752, hijo de una acaudalada familia. En 1770, a la edad de 18 años recibió el grado de matemáticas y física.

Entre 1775 y 1780 se desempeñó como profesor en la Escuela Militar de París. En 1783 fue incorporado a la Academia de Ciencias de París y posteriormente, en 1783, fue becario de la Royal Society.

Como consecuencia de la Revolución Francesa perdió su fortuna y tuvo que buscar trabajo, desempeñándose como profesor en el Instituto de Marat. En este periodo, escribió uno de los mejores textos de Geometría de la época.

En 1805, Legendre publicó de una manera clara y elegante el método de mínimos cuadrados, siendo este concepto uno de sus principales contribuciones a la estadística. Este método fue posteriormente relacionado por Francis Y. Edgeworth con el concepto de correlación formulado por F. Galton.

A. Legendre murió en París en 1833.

CONTENIDO

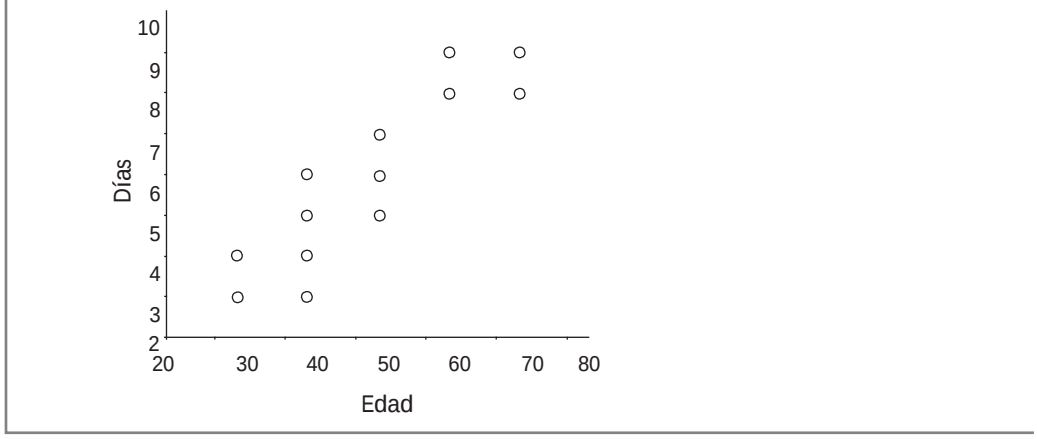
3.1 Introducción

En diversos problemas vinculados con el análisis de la información es importante estudiar las relaciones que puedan existir entre dos o más variables. Preguntas como: ¿las personas con mayor poder adquisitivo tienen mayor grado de educación? o ¿las empresas familiares pagan mejor a sus empleados? están referidas a las relaciones que podrían existir entre las variables "poder adquisitivo" y "educación" o "tipo de empresa" y "sueldo". Las relaciones entre variables pueden explorarse algunas veces usando gráficos adecuados, pero también existen medidas que indican no solo la existencia de la relación sino también la fuerza de esta. A estas medidas se les llama *medidas de correlación* cuando las variables en estudio son numéricas u ordinales y *medidas de asociación* si las variables son nominales.

3.2 El índice de correlación de Pearson

Al colocar en el eje *X* las edades de un grupo de pacientes de un hospital y en el eje *Y*, el número de días que cada uno de ellos ha necesitado para recuperarse después de una determinada operación, se obtiene el siguiente *diagrama de dispersión*.

FIGURA 3.1 *Diagrama de dispersión*



Nótese que a mayor edad del paciente acompaña mayor número de días necesarios para recuperarse. Podemos decir que las variables edad y estadía *covarían de manera positiva.*

Si ocurriera que a mayor edad es menor la estadía se dice que ambas variables *covarían de manera negativa*.

Para confirmar si dos variables cuantitativas covarían o no, se usa un índice que se llama *covarianza*.

Si se tienen los pares de valores $(x_1, y_1),\ldots, (x_n, y_n)$ de las variables *X e Y,* la *covarianza* entre estos valores se define como:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

La covarianza es el promedio de todos los productos de las desviaciones de las variables respecto de sus medias, e indica la relación lineal que entre ellas puede existir. Sin embargo, esta medida no indica la *fuerza de la relación* que pueda existir entre las variables. Un valor positivo muy grande o muy pequeño puede deberse simplemente a las unidades de medición y no a que exista mayor o menor grado de la relación. Se necesita, por tanto, una medida que, siendo acotada, no tenga en cuenta las unidades de medición. El artificio, ya utilizado, para obtener la medida adecuada consiste en expresar la covarianza en unidades de desviación estándar. Así se obtiene el *índice de correlación lineal de Pearson* como medida para medir la fuerza de la relación entre dos variables numéricas.

El ***índice de correlación lineal de Pearson*** o simplemente *índice de correlación* se define como:

$$r_{xy} = \frac{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}\sqrt{\sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}}} = \frac{Cov(X, Y)}{s_x s_y}$$

El índice de correlación se puede escribir como $r_{xy} = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right)}{n-1}$.

La última relación indica que la correlación es la covarianza de las variables estandarizadas.

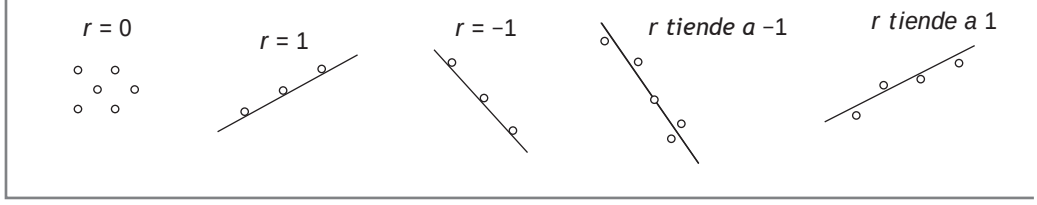
El índice de correlación mide la fuerza de la relación *lineal* entre dos variables. Un índice de correlación alto indica que una línea recta "ajusta bien" a la nube de puntos. Un índice de correlación cercano a 0 indica ausencia de relación lineal.

En general se cumplen las siguientes propiedades, que se pueden demostrar:

a) El índice de correlación *r* está entre –1 y 1.
b) Si el índice de correlación es igual a 0, no existe relación lineal; sin embargo, puede existir una relación no lineal (cuadrática, cúbica o más complicada).

c) Si *r* tiende a −1 o a 1, los puntos tienden a estar más alineados. Cuando *r* es igual a 1 o a −1, los puntos están perfectamente alineados.

FIGURA 3.2 *Distintos valores de r*



EJEMPLO. *Ingresos y egresos*

A continuación, en la Tabla 3.1 se presentan: los ingresos (*X*) y los egresos (*Y*) de cinco familias.

TABLA 3.1 *Ingresos vs. egresos de cinco familias*

Ingresos	150	180	150	200	250
Egresos	120	170	140	170	200

En la siguiente tabla (3.2) aparecen los valores de *X* e *Y,* así como los cálculos necesarios para hallar la covarianza entre *X* e *Y*.

TABLA 3.2 *Tabla de cálculos*

	x	y	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
1	150	120	1,440	1,296	1,600
2	180	170	−60	36	100
3	150	140	720	1,296	400
4	200	170	140	196	100
5	250	200	2,560	4,096	1,600
Total	930	800	4,800	6,920	3,800

La covarianza es $Cov(X, Y) = (1/4)\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = (1/4)(4800) = 1200$, e indica que los valores de *X* e *Y* covarían de manera positiva, pero no indica qué tan fuerte es la relación lineal entre *X* e *Y*. Para analizar la fuerza de la relación se calcula el índice de correlación.

El índice de correlación es:

$$r_{xy} = \frac{(1/4)\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5}\frac{(x_i - \bar{x})^2}{4}}\sqrt{\sum_{i=1}^{5}\frac{(y_i - \bar{y})^2}{4}}} = \frac{1200}{\sqrt{\frac{6920}{4}}\sqrt{\frac{3800}{4}}} = 0.9360$$

e indica "alta correlación positiva" entre los ingresos y los egresos. Se tiene que a valores altos de los ingresos acompañan valores altos de los egresos; sin embargo, no se infiere que quien gana mucho necesariamente siente el deseo de gastar mucho.

3.3 La recta de regresión de mínimos cuadrados

Ahora la idea es expresar mediante una relación matemática la relación lineal que podría existir entre los valores de *X* e *Y*. El *modelo de regresión lineal*, que se desarrolla más adelante, será la ayuda más importante para este propósito. Por ahora hallaremos la recta que "mejor ajusta" a la nube de puntos y que formará parte del modelo. Esta recta se llama recta de ***mínimos cuadrados*** o ***de regresión*** de *Y* en *X*, y se determina a partir de los pares (x_1, y_1) (x_2, y_2),..., (x_n, y_n) de *X* e *Y* usando el *método de mínimos cuadrados*.

La ecuación de la recta de mejor ajuste es de la forma $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ y los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ se determinan de tal manera que la suma:

$$SCE = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

sea mínima. De ahí el nombre de recta de mínimos cuadrados.

Haciendo los cálculos necesarios se encuentra que los coeficientes $\hat{\beta}_0$ y $\hat{\beta}_1$ son como sigue:

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X, Y)}{V(X)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

La ecuación de la recta encontrada se puede escribir como $y = (\overline{y} - \hat{\beta}_1\overline{x}) + \frac{Cov(X, Y)}{V(X)}\, x$ o simplemente como $y = \overline{y} + r\,\frac{s_Y}{s_X}\,(x - \overline{x})$, en donde r es el índice de correlación entre X e Y y s_X y s_Y son las desviaciones estándar de X e Y, respectivamente.

Notas

1. La última relación indica que la pendiente de la recta de regresión de mínimos cuadrados es proporcional al índice de correlación y tienen el mismo signo. Esta recta pasa por el punto $(\overline{x}, \overline{y})$.
2. Las diferencias $(y - \hat{y})$, entre los valores observados, y, y los correspondientes valores ajustados, $\hat{y}$, se llaman *residuales*.

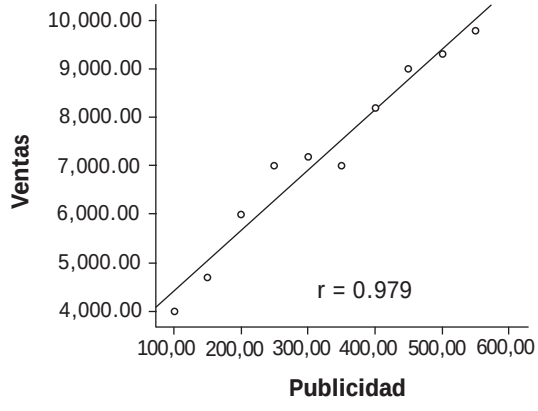
Si se escribe $y = \hat{y} + (y - \hat{y})$ se tiene que cada observación se puede expresar como el ajuste más el residual. El ajuste representa el patrón general que gobierna a los datos, mientras que el residual representa las desviaciones de este patrón.

EJEMPLO. *Publicidad vs. ventas*

Los gastos en publicidad, así como las ventas, en u.m., de 10 empresas industriales fueron registrados y representados en un diagrama de dispersión (Figura 3.3) como el que aparece a continuación.

FIGURA 3.3 *Recta de regresión de Y en X*

Publicidad	*Ventas*
100.00	4,000.00
150.00	4,700.00
200.00	6,000.00
250.00	7,000.00
300.00	7,200.00
350.00	7,000.00
400.00	8,200.00
450.00	9,000.00
500.00	9,300.00
550.00	9,800.00



La media y la desviación estándar de la variable publicidad, denotada con X, en u.m, son 325.00 y 151.00, respectivamente.

La media y la desviación estándar de la variable ventas, denotada con Y, en u.m, son 7,220.00 y 1,926.00, respectivamente.

El índice de correlación entre la publicidad y las ventas es 0.979. Un crecimiento en la publicidad acompaña a un crecimiento en las ventas, y la recta de la regresión de *Y* en *X* bien puede ajustar a los datos.

La ecuación de la recta de regresión de *Y* en *X* es $y = 3170.303 + 12.461x$.

Con la ecuación de la recta de mejor ajuste, se puede estimar las ventas que se esperan obtener cuando se invierte, por ejemplo, 380.00 u.m. en publicidad. El valor de las ventas esperadas es igual a:

$$Y = 3170.303 + 12.461\ (380.00) = 7{,}905.483 \text{ u.m.}$$

3.4 Índices de correlación para variables en escala ordinal

Estas medidas se calculan a partir del orden que se pueda establecer entre los valores de las variables. Entre los índices de correlación de este tipo están: los índices *de Spearman, de gamma de Goodman y Kruskal, de Sommers, de Tau-b de Kendall* y *de Tau-c de Kendall*. El más conocido es el índice de correlación de Spearman, que se desarrolla a continuación.

El índice de correlación de Spearman

El índice de ***correlación de rangos de Spearman*** se usa cuando las variables son ordinales y tienen muchas categorías.

Para definir el índice de correlación de rangos de Spearman entre un grupo de valores de las variables ordinales *X* e *Y* se procede de la siguiente manera:

1. Se consideran los órdenes de los valores: $x_1,\ldots, x_n$ de *X*. De igual manera para los valores correspondientes $y_1,\ldots, y_n$ de *Y*. Si existen dos o más valores iguales, los órdenes de cada uno de estos son iguales al promedio de los órdenes que les correspondería en el caso de que fueran diferentes (por ejemplo, a los valores 38, 37, 39, 40, 40 y 46 les corresponde, respectivamente, los órdenes, 2, 1, 3 4.5, 4.5, 6). A los órdenes se les llama *rangos*.
2. Se denota con $u_1,\ldots, u_n$ a los rangos que les corresponde a los valores $x_1,\ldots, x_n$ y con $v_1,\ldots, v_n$ a los rangos correspondientes a $y_1,\ldots, y_n$.
3. El índice de correlación de rangos de Spearman se define como el número:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n)(n^2 - 1)}$$

en donde d_i es la diferencia, $u_i - v_i$, entre los rangos respectivos de *X* e *Y*.

Se demuestra que cuando no existen valores repetidos de *X* o de *Y,* el índice de correlación de Pearson entre los rangos es igual al índice de correlación de rangos de Spearman. *Si el número de repeticiones es pequeño ambos índices son aproximadamente iguales*.

La interpretación del índice de correlación de rangos de Spearman es la misma que la del índice de correlación de Pearson.

EJEMPLO. *Calificaciones de dos jurados*

Siete postulantes a un empleo son calificados por dos jurados. Para determinar si existe relación entre las actitudes de los jurados, se determina a continuación el índice de correlación de Spearman de las calificaciones. Las calificaciones, así como los rangos R_i correspondientes, aparecen en la Tabla 3.3.

TABLA 3.3 *Calificación de los jurados*

Postulante	*Jurado* 1	*Jurado* 2	R_1	R_2	d_i	d_i^2
A	44	58	1	1	0	0
B	39	42	2	2	0	0
C	36	18	3	7	-4	16
D	35	22	4	6	-2	4
E	33	31	5	5	0	0
F	29	38	6	3.5	2.5	6.25
G	22	38	7	3.5	3.5	12.25
Total						38.50

El índice de correlación de Spearman es $\rho = 1 - \dfrac{6\sum_{i=1}^{n} d_i^2}{(n)(n^2 - 1)} = 1 - \dfrac{6(38.5)}{7(49 - 1)} = 0.31.$

El resultado indica que existe baja correlación entre las calificaciones de ambos jurados.

3.5 Medidas de asociación para variables con escala nominal

Cuando las variables en estudio son nominales, la relación que pueda existir entre ellas se mide con las *medidas de asociación.*

Si no es posible distinguir con precisión entre una variable independiente de clasificación y una variable de respuesta, la medida se dice que es *simétrica***.** Si una de las variables actúa como variable de respuesta, se dice que la medida es *asimétrica.*

Diferencia entre porcentajes

Un estadístico sencillo que permite medir la relación entre dos variables categóricas es *la diferencia entre porcentajes, calculada en sentido contrario al que hayan sido hallados los porcentajes*. Este método no es muy riguroso, sin embargo, es muy utilizado.

EJEMPLO. *¿La elección de una carrera depende del género?*

Para estudiar la relación entre la variable "género" y la variable "especialidad" entre los estudiantes de una universidad, se obtuvo la siguiente tabla (3.4) de contingencia.

TABLA 3.4 *Carrera vs. género*

	Hombres	*Mujeres*	*Total*
Ciencias	22	18	40
Letras	3	32	35
Total	25	50	75

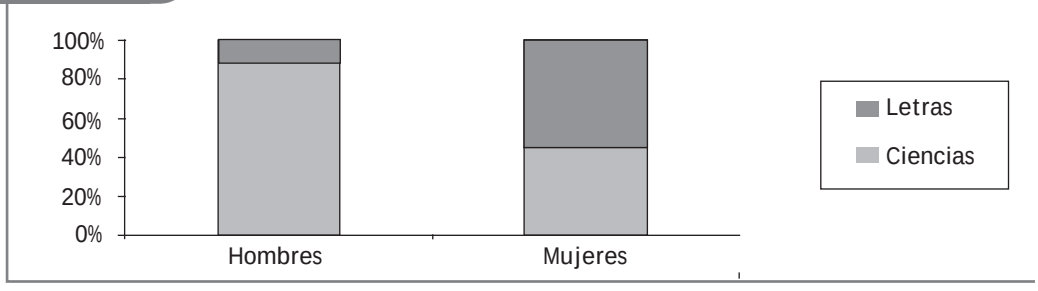
Hallando los porcentajes en el sentido de las columnas obtenemos:

TABLA 3.5 *Porcentaje en columnas*

	Hombres	*Mujeres*
Ciencias	88.00 %	45.00 %
Letras	12.00 %	55.00 %
Total	100.00 %	100.00 %

Se observa que la distribución de los hombres, de acuerdo a su especialidad, es diferente a la distribución de las mujeres, de acuerdo a su especialidad. Los resultados indican que, en el conjunto de datos analizado, los varones eligieron de manera diferente su especialidad que las mujeres. Al parecer el género influye en la elección de la especialidad. Un gráfico que ilustra este resultado es el gráfico de barras al 100%.

FIGURA 3.4 *Gráfico de barras al 100%. Al parecer, la especialidad depende del género*



El coeficiente phi para tablas de contingencia 2 x 2

Este coeficiente se aplica para tablas de contingencia 2×2; sus valores están comprendidos entre 0 y 1. Este índice puede ser mayor que la unidad si las tablas no son 2×2 ; por ello no es aplicable en estos casos. Cuando el coeficiente es 0, se considera que los valores de las variables no están relacionados. Si su valor es 1, la relación es perfecta.

El coeficiente phi, que se denota con ϕ, se define como:

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

Los valores que aparecen en la relación anterior se indican en la Tabla 3.6.

TABLA 3.6 *Tabla 2 x 2*

		Factor Y	
Factor X	*A*	*D*	Total
I	*a*	*b*	*a + b*
II	*c*	*d*	*c + d*
Total	*a + c*	*b + d*	*n*

EJEMPLO. *Recordación y deseo de compra*

Se desea estudiar la relación que existe entre el deseo de compra y la recordación de una propaganda realizada, relativa a un producto. El estudio se llevó a cabo con 150 personas. La distribución fue como sigue:

TABLA 3.7 *Recordación*

Recordación deseo	***Sí recuerda***	***No recuerda***	***Total***
Compra	60	20	80
No compra	20	50	70
Total	80	70	150

El valor del coeficiente phi es $\phi = \dfrac{60 \bullet 50 - 20 \bullet 20}{\sqrt{80 \bullet 70 \bullet 80 \bullet 70}} = 0.4642.$

Este valor indica baja asociación entre la recordación y el deseo de compra, entre los elementos de la muestra. ¿Se puede generalizar este resultado a la población de donde proviene la muestra?

El coeficiente V de Cramer para tablas de contingencia m x n

El coeficiente *V de Cramer* se define como $V = \sqrt{x^2/N(k-1)}$, donde k es el mínimo entre el número de filas y de columnas de la tabla de contingencia de N entradas.

TABLA 3.8 *Tabla de contingencia*

	1	*2*	*...*	*n*	*Total*
1	x_{11}	x_{12}	...	x_{1n}	$x_{1.}$
...	...				...
m	x_{m1}	x_{m2}	...	x_{mn}	$x_{m.}$
	$x_{.1}$	$x_{.2}$	...	$x_{n.}$	N

En esta relación, el valor x^2 se define como:

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(x_{ij} - (x_{i.}x_{.j}/N)\right)^2}{(x_{i.}x_{j.}/N)}$$

Este número se conoce con el nombre de *valor ji-cuadrado*, y como se verá más adelante, es una medida que indica la fuerza de la relación entre las variables *X* e *Y*.

Si las tablas son de orden 2×2, el coeficiente *V* de Cramer es igual al coeficiente phi. Se interpreta de manera similar al coeficiente ϕ.

El coeficiente de contingencia

Una medida de asociación más apropiada que el coeficiente de Cramer, con rango teórico entre 0 y 1, es el *coeficiente de contingencia C.* Este coeficiente se define como:

$$C = \sqrt{\frac{x^2}{x^2 + N}}$$

y se interpreta como el coeficiente ϕ.

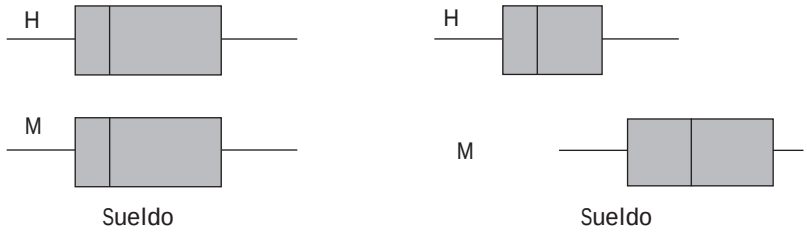
3.6 Medidas de asociación de variables con diferentes escalas de medida

El análisis de la asociación entre dos variables se realizó considerando que las variables tienen la misma escala; sin embargo, existen situaciones en donde estas pueden tener diferentes escalas de medida. Sucede, por ejemplo, si se analiza la asociación

entre la variable “género” (con escala nominal) y la variable “sueldo” (con escala de intervalo). Una solución a este caso puede ser el de considerar la categorización de la variable "sueldo" para luego aplicar las medidas estudiadas; sin embargo, una mejor solución consiste en observar si el hecho de estar en una u otra categoría de la variable “género” no afecta a la distribución de los valores de la variable “sueldo”.

Generalmente se utiliza como referente de comparación de las distribuciones a una de las medidas de tendencia central, por ejemplo, la media acompañada con otras medidas adicionales, como los índices de dispersión. Utilizando la media aritmética de la variable “sueldo” y observando las dispersiones, se pueden realizar comparaciones en las categorías de la variable género.

FIGURA 3.5 *Comparación de medias*



Los dos gráficos de cajas de la izquierda indican que los sueldos tienen igual distribución en ambas categorías de la variable género; se dice que estas variables *son independientes*. En los dos gráficos de cajas de la derecha se observa que la distribución de los sueldos, incluyendo las medias y dispersiones, es diferente para ambas categorías de la variable “género”; se dice que las variables *no son independientes*.

Si bien es cierto que el procedimiento brinda buena información de la distribución de una variable en cada una de las categorías, es necesario medir el llamado *tamaño del efecto* como una medida de la fuerza de la asociación. Un índice que mide la fuerza de la asociación entre una variable de intervalo y una variable nominal es el *índice de asociación de Cohen*.

El índice de Cohen

Para una variable *Y* con escala de intervalo o de razón y una variable *X* categórica con dos categorías *A* y *B*, el índice *de Cohen* es el número:

$$d = \frac{\bar{y}_A - \bar{y}_B}{sc}$$

donde $sc = \sqrt{\dfrac{\sum_{k=1}^{n_1}(y_k - \bar{y}_A)^2 + \sum_{k=1}^{n_2}(y_k - \bar{y}_B)^2}{n_1 + n_2 - 2}}$ es un estimador de la desviación estándar de los valores de Y, n_1 es el número de elementos en la categoría A y n_2 es el número de elementos en la categoría B.

$\bar{y}_A$ es la media aritmética de los valores de Y en la categoría A, etcétera.

Un valor de d igual a 0 indicaría que no hay efecto diferencial sobre la variable Y de la variable cualitativa.

El signo positivo o negativo de d indica la dirección en que se produce el cambio de nivel.

El valor de d es un índice que no tiene una cota, por ello es difícil indicar el tamaño de la asociación.

EJEMPLO. *Compras y género*

Los datos que a continuación se indican corresponden al valor de las compras, en dólares, que un grupo de hombres y un grupo de mujeres realizaron en una tienda de departamentos.

TABLA 3.9 *Comparación de medias*

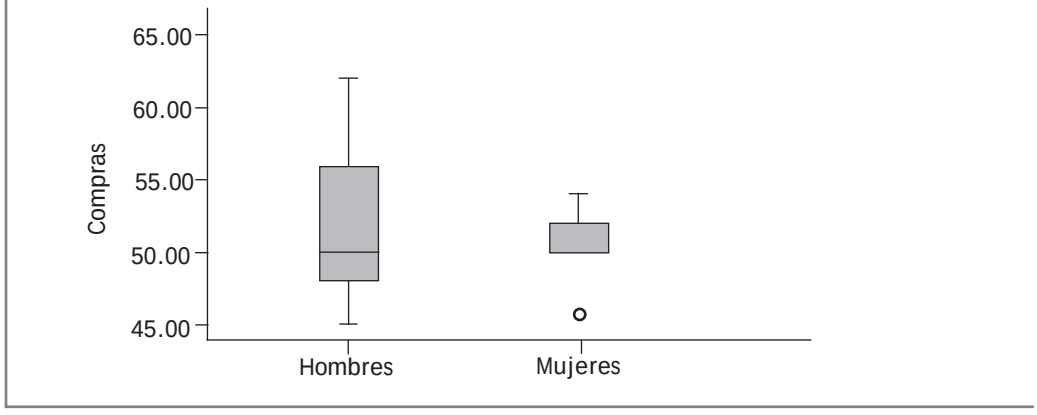
Mujeres		***Hombres***	
Sujeto	Compras	Sujeto	Compras
1	56	1	46
2	48	2	50
3	62	3	50
4	50	4	52
5	45	5	54

Se observa que la media de las compras para las mujeres es 52.2 dólares y la media de las compras para los hombres es 50.4 dólares.

El valor de sc es 5.2440.

El índice de Cohen es $d = \dfrac{52.2 - 50.4}{5.2440} = 0.3432$.

FIGURA 3.6 *Comparación de medias usando cajas*



Este valor indica que la distancia de separación entre las medias de las compras del grupo de hombres y mujeres es 0.3432 veces la desviación estándar de las compras, aproximadamente. Al parecer el valor de *d* es un valor pequeño, lo que indica independencia de la variable "compras" y la variable "género".

Coeficiente de correlación biserial puntual

El coeficiente de *correlación biserial puntual* se establece cuando se desea analizar la relación entre dos variables, una de las cuales, *X,* es categórica con dos categorías *A* y *B*, y la otra, *Y,* es cuantitativa. Este índice se define como:

$$r_{bp} = \frac{(\bar{y}_A - \bar{y}_B)pq}{s_Y}$$

en donde:

$\bar{y}_A$ = media aritmética de Y en la categoría A

$\bar{y}_B$ = media aritmética de Y en la categoría B

p = proporción de casos en la categoría A

q = proporción de casos en la categoría B

s_Y = desviación estándar de los valores de Y

Los valores del coeficiente de correlación biserial puntual varían entre –1 y 1 y su interpretación es análoga a la de los casos anteriores.

Medidas de asociación entre una variable en escala de intervalo o de razón y una variable cualitativa. El coeficiente eta

Se supone que una muestra formada con los sueldos, en u.m., de los empleados de tres empresas A, B y C es la siguiente:

1,050, 750, 900, 900, 750, 1,000, 1.300, 700, 850, 1,000, 1,000, 1,200, 1,200, 1,200, 1,350, 900, 1,050, 900, 1,350,

Una estrategia para predecir el sueldo de un empleado de cualquiera de las tres empresas es usar la media de todos los sueldos del grupo. Sin embargo, el error de predicción puede ser considerable si la variabilidad de los sueldos es grande. El error puede disminuirse si se conoce a cuál de las empresas pertenece el empleado. Si esto sucede, se dice que la variable "empresa" se puede usar para "explicar" a la variable sueldo.

En general, para analizar si la variabilidad de una variable cuantitativa (como el "sueldo") es explicada por las diferentes categorías de una variable cualitativa (como la "empresa"), se usa el *coeficiente eta*, que se define como:

$$\eta = \sqrt{\frac{SST - SSC}{SST}}$$

en donde *SST* es la "suma de cuadrados total"; esto es, la suma de los cuadrados de todas las desviaciones de los valores de la variable cuantitativa respecto de su media. Esta suma mide la variación total.

La expresión *SSC* es la suma de los cuadrados de las desviaciones de los valores, respecto de su media, en cada uno de los grupos que las categorías de la variable cualitativa determinan.

El numerador de la fracción que está dentro del radicando representa la *reducción debida a la variable cualitativa,* que describe a los grupos.

El numerador es 0 si la variable cualitativa no reduce en absoluto a la variación observada de los valores de la variable cuantitativa. En tal caso el valor eta es 0. Si la inclusión de la variable cualitativa en el análisis reduce al máximo la variación observada de la dependiente, el numerador es igual al denominador; el valor eta es igual a 1 y las variables son totalmente dependientes.

El cuadrado de eta expresa la proporción de la variabilidad de la variable *dependiente Y* que es explicada por la variable *independiente X*.

En el caso de los sueldos, la media de todos los sueldos es 1,026.47 y la suma de cuadrados total es:

$$SST = (1050 - 1026.4705)^2 + \ldots + (1350 - 1026.4705)^2 = 715588.235$$

Si:

en la empresa A los sueldos son: 1,050, 750, 900, 900, 750;
en la empresa B los sueldos son: 1,000, 1,300, 700, 850, 1,000, 1,000, 1,200 y
en la empresa C los sueldos son: 1,200, 1,200, 1,350, 900, 1,050, 900, 1,350,

se tendrá que la media de los sueldos en la empresa *A* es 862.50, en *B* es 1,008.3 y en *C* es 1,135.71.

La "suma de cuadrados" en *A* es:

$$(1050 - 862.50)^2 + \ldots + (750 - 862.50)^2 = 61875$$

La "suma de cuadrados" en *B* es:

$$(1000 - 1008.33)^2 + \ldots + (1200 - 1008.33)^2 = 242083.33$$

La "suma de cuadrados" en *C* es:

$$(1200 - 1135.71)^2 + \ldots + (1350 - 1135.71)^2 = 218571.4$$

La suma de la suma de cuadrados para las tres empresas es *SSC* = 522,529.767.

La reducción de la varianza, "debido a las empresas", es:

$$SST - SSC = 715{,}588.2350 - 52{,}259.7670 = 193{,}058.468$$

El coeficiente eta, tomando al sueldo como variable "dependiente" y a la empresa como variable "explicativa", es igual a 0.519, lo que indica que el conocimiento de la empresa a la cual pertenece el empleado reduce la variabilidad de la respuesta, en este caso el sueldo.

3.7 Medida de acuerdo: el coeficiente de Kappa

El coeficiente *kappa* es una medida que se aplica a variables categóricas. Se refiere al grado de acuerdo o desacuerdo en la clasificación subjetiva de eventos realizada por dos entidades que se llaman *jueces*. Para el cálculo de este coeficiente es necesario *tener una tabla de contingencia con las mismas categorías para filas y columnas*. Cada categoría indica la calificación otorgada por los jueces para el mismo sujeto. Por ejemplo, las calificaciones que realizan dos gerentes para elegir un nuevo empleado.

El *coeficiente kappa* se define como $\kappa = \frac{D_0 - D_e}{n - D_e}$, en donde *n* es el número de elementos en la muestra, $D_0 = \sum_i a_{ii}$ corresponde al *número de acuerdos en la muestra* y $D_e = \frac{1}{n}\sum_i n_{i.} n_{.i}$ corresponde al *número de acuerdos esperados si las calificaciones de los jueces fueran independientes.*

El coeficiente *kappa* es una medida cuyos valores están entre –1 y 1. Valores próximos a 1 indican total acuerdo entre los jueces. Valores próximos a –1 indican total desacuerdo

más allá de la independencia de las observaciones, de manera que la respuesta menos probable de un juez es, precisamente, la que ha registrado el otro. Si el valor del coeficiente es 0, las calificaciones de los jueces son totalmente independientes.

La Tabla 3.10 indica una escala de valoración propuesta por Landis y Koch, y sirve para interpretar diferentes valores de κ.

TABLA 3.10 *Escala de valoración de Landis y Koch*

κ	***Grado de acuerdo***
0.00 a 0.20	Insignificante
0.21 a 0.40	Regular
0.41 a 0.60	Moderado
0.61 a 0.80	Sustancial
0.81 a 0.90	Excelente

EJEMPLO. *¿Los médicos concuerdan en sus diagnósticos?*

Dos médicos, A y B, estudiando un determinado tipo de análisis, dictaminaron los siguientes resultados para 72 pacientes.

TABLA 3.11 *Acuerdos*

		Médico B			
		Bueno	Medio	Malo	Total
Médico A	Bueno	6	4	2	12
	Medio	12	8	4	24
	Malo	18	12	6	36
	Total	36	24	12	72

Determinar el grado de asociación de las calificaciones de los dos médicos.

Para esta tabla se tiene:

$D_0 = 6 + 8 + 6 = 20$, $D_e = (1/72)(12 \times 36 + 24 \times 24 + 36 \times 12) = 20.01$

El valor del coeficiente *kappa* es:

$$\kappa = \frac{D_0 - D_e}{n - D_e} = \frac{20.00 - 20.01}{72 - 20} = 0, \text{ aproximadamente,}$$

lo que indica un acuerdo no significativo. Las calificaciones de los jueces son "independientes".

LA ESTADÍSTICA EN LA EMPRESA

La firma Farmacon

Farmacom es una empresa transnacional farmacéutica dedicada al negocio de la protección de la salud de sus consumidores; su fortaleza se refleja en la ayuda que presta a muchas personas para que tengan una vida más larga y de mejor calidad.

Farmacom es la marca reconocida, confiable y respetada en el mundo desde hace 100 años. Fue creada en Alemania, y casi desde sus inicios instaló sucursales en nuestro país, participando en su desarrollo industrial.

El catálogo de medicinas que comercializa Farmacom supera los 4,000 productos, los que contribuyen para que muchos millones de personas en el mundo tengan una vida mejor. En la lista de medicinas que Farmacom comercializa figuran productos que han dado renombre a sus descubridores en el ámbito científico. El logro más importante de esta pujante empresa es la lista de vacunas que sus científicos han desarrollado y que con gran eficacia han comercializado.

Farmacom es conocida por su apoyo a diferentes instituciones que ayudan a la niñez necesitada.

Para Farmacom, la utilización de la estadística como herramienta de apoyo en el descubrimiento, desarrollo e introducción de nuevos productos es de primera prioridad y fundamental. Esta tarea comprende la identificación de diferentes fuentes de variabilidad en los procesos, la estimación de parámetros en diferentes modelos que se usan y el diseño de las experiencias que se realizan para determinar la influencia de diferentes factores en una serie de mediciones que se llevan a cabo.

Farmacom usa también la estadística como auxiliar del análisis de los procesos relacionados con el mercado y la identificación de las necesidades de sus clientes. La segmentación de sus clientes, el posicionamiento de sus productos y la predicción de valores de diferentes variables en el futuro son algunas de las innumerables tareas que la estadística ayuda a realizar.

EJERCICIOS

1. Con los datos de la siguiente tabla, relacionados con los ingresos (X) y los consumos (Y) en dólares de cinco personas.

X_i	200	300	400	600	900
Y_i	180	270	320	480	700

   a) Construir la gráfica de dispersión de los puntos (x, y).

   b) Calcular el índice de correlación de Pearson.

2. Usar el coeficiente phi para medir la relación que pueda existir entre el factor género y la edad entre los usuarios de cabinas de Internet. La edad está expresada en dos categorías: los que tienen edad comprendida entre 10 y 20 años y los que son mayores de 20 pero no mayores de 30 años. El estudio se realizó con 200 personas y la distribución de los datos fue como sigue.

Edad	*Varón*	*Mujer*	*Total*
$10 \leq Ed \leq 20$	75	30	80
$20 < Ed \leq 30$	45	50	70
Total	120	80	200

   Comparar el resultado anterior con el coeficiente V de Cramer. Comentar.

3. Los pares de valores de las variables X e Y, así como las frecuencias para cada uno de estos, es como sigue: (3, 1) con frecuencia 5, (5, 3) con frecuencia 4 y (7, 6) con frecuencia 8. Hallar el índice de correlación correspondiente.

4. Las correlaciones llamadas *ecológicas* se realizan entre proporciones o medias, y son muy utilizadas en política. Deben utilizarse con sumo cuidado.

   En un estudio en un determinado país, se calculó la correlación entre la renta y el nivel educativo de los hombres de 25 a 64 años. El resultado fue $r = 0.4$. Posteriormente, y tomando en cuenta las siete regiones en que estaba dividido el país, se calculó la correlación entre las *siete medias* de la renta y las *siete medias* del nivel educativo correspondientes. El resultado fue $r = 0.93$, muy distinto del 0.4 inicial. ¿Cuál resultado explica mejor la realidad? ¿Por qué?

5. Un estudio determinó que para 11 países la correlación entre el número promedio de cigarrillos (por persona) y el porcentaje de muertes por cáncer de pulmón era igual a 0.9. Esta cifra se tomó como demostración de la importante relación entre tabaco y cáncer. ¿Qué comentarios se pueden realizar al respecto?

6. Con la finalidad de estudiar la relación que pudiera existir entre las calificaciones en el curso de Responsabilidad social que han recibido los alumnos de una universidad de los profesores A, B y C, se recogió la siguiente información, relativa a las calificaciones:

   Profesor A: 15, 08, 12, 12
   Profesor B: 10, 13, 07, 15, 10, 14, 12
   Profesor C: 12, 12, 13, 09, 13, 09, 14

   Indicar si existe una relación entre las calificaciones y los profesores que califican.

7. Dos psicólogos realizaron una clasificación de 200 personas utilizando el test de Rorschach. Indicar el grado de acuerdo entre los psicólogos.

Psicólogo 2	*Psicólogo 1*		
	Psicótico	*Neurótico*	*Orgánico*
Psicótico	60	4	3
Neurótico	5	45	7
Orgánico	2	9	65

8. Existe el interés de encontrar variables que expliquen la gravedad de los accidentes de tránsito. Una de ellas podría ser la clase social. Por ello se clasifica la gravedad de los accidentes en los niveles de clase social: baja (1), media (2) y alta (3). Los datos que se obtienen al recoger los expedientes de 144 accidentados, teniendo en cuenta a la variable "clase social" y la gravedad de los accidentes, son:

Clase social	*Gravedad*			*Total*
	1	**2**	**3**	
A	9	6	4	19
B	21	32	51	103
C	7	6	8	21
Total	37	44	63	144

Usando una medida adecuada, indicar si la clase social se puede considerar como una variable que explica la gravedad del accidente.

9. En un centro comercial se investiga si la posición en donde se ubican los productos de cierta marca influye en las ventas diarias. Para ello, en determinados días se ubicaron los productos en el primer anaquel y se anotó el número de productos vendidos. De igual modo se hizo para el segundo y tercer anaquel. Los datos recogidos aparecen en la siguiente tabla de contingencia.

Número de productos vendidos	*Anaquel*		
	1	**2**	**3**
Mayor que 100	5	3	4
Entre 50 y 100	10	11	12
Menor que 50	40	35	32
Total	55	49	48

Interpretar los resultados.

10. Dos ingenieros prueban, cada uno, 110 motores asignando las calificaciones de "bueno", "medio" y "malo". Como resultado de las calificaciones se obtuvieron los siguientes datos.

		Ingeniero B			
Ingeniero A		Bueno	Medio	Malo	Total
	Bueno	18	11	6	35
	Medio	11	30	9	50
	Malo	2	3	20	25
	Total	31	44	35	110

Decir si existe acuerdo o desacuerdo entre los ingenieros.

11. Los resultados de diversos estudios muestran que existe una correlación negativa entre las horas que la gente ve televisión y las calificaciones que obtienen en pruebas de lectura. ¿Es el hábito de ver la televisión la causa de que disminuya la capacidad de lectura de la gente?

RESPUESTAS A LOS EJERCICIOS

1. a) El índice de correlación es 0.9991. **2.** phi = 0.245, V de Cramer = 0.245 y coeficiente de contingencia = 0.238. **6.** eta = 0.0334. **7.** kappa = 0.774. **8.** Spearman = 0.090. **10.** kappa = 0.418.

CAPÍTULO 4

Introducción a las series de tiempo

EL NIÑO

Los avances de los estudios de la ciencia sobre el fenómeno de El Niño –producto de los esfuerzos internacionales que se iniciaron con el año 1972 y por el impacto que éste tuvo en el mercado de proteínas baratas y luego, en 1983, por sus consecuencias tan pronunciadas en diferentes partes del mundo– permiten hoy en día detectar y predecir las condiciones oceanográficas y meteorológicas que caracterizan este fenómeno en forma casi inmediata.

El Niño, un fenómeno que los peruanos hasta hace poco pensábamos era peculiar exclusivamente de nuestras costas, hoy se reconoce como un fenómeno global.

FUENTE: Instituto Geofísico del Perú.

CONTENIDO

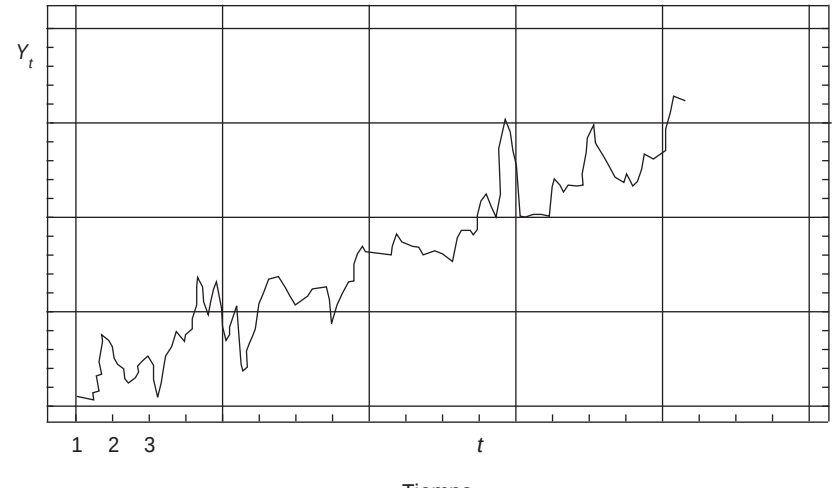
4.1 Introducción

Siempre se ha deseado pronosticar eventos futuros con la finalidad de planificar, prever o prevenir. Ello ha motivado la introducción de técnicas de predicción que se basan en información pasada posible de ser cuantificada y que obedece a "patrones" que no cambian con el tiempo. Algunas de estas técnicas se desarrollan en este capítulo, y permiten en muchos casos la construcción de modelos simples que explican la estructura que tienen las observaciones pasadas de una variable para luego prever sus valores futuros.

Una *serie de tiempo* es un conjunto de observaciones, Y_1, Y_2,..., Y_n, de una variable, obtenidas en los tiempos 1, 2,..., n, respectivamente.

La Figura 4.1 representa una serie de tiempo. En el eje de las X se indican los tiempos y en el eje de las Y, las observaciones correspondientes de la variable.

FIGURA 4.1 *Representación gráfica de una serie de tiempo*



Aplicaciones de las series de tiempo se pueden encontrar en el campo de la economía, la geografía, la física, la mercadotecnia, la sociología, la demografía, etcétera.

EJEMPLO. *Ejemplos de series de tiempo*

Los registros diarios de las precipitaciones pluviales en una región forman una serie de tiempo. Los registros de las ventas que realiza una empresa durante los días del mes de junio constituyen una serie de tiempo. El consumo de energía eléctrica a lo largo de un mes en una ciudad es una serie de tiempo. El número de nacimientos en una población, registrados diariamente a lo largo de un año, es una serie de tiempo.

4.2 Objetivos del estudio de una serie

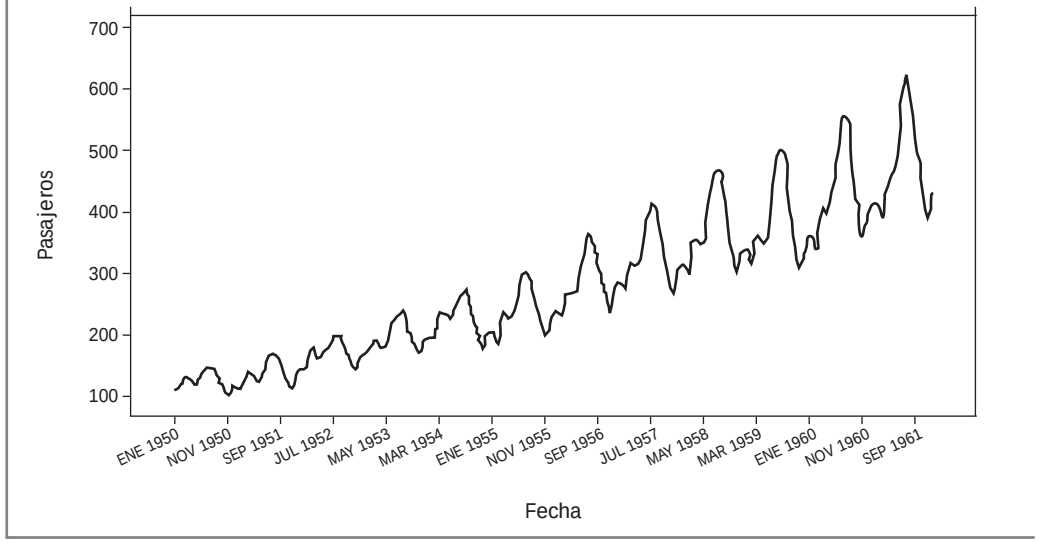
El análisis de una serie de tiempo se realiza, entre otras razones, para descubrir el proceso que dio lugar a los valores observados para así predecir los valores futuros. Los análisis también apuntan muchas veces a evaluar el efecto de algún fenómeno en el desarrollo de la serie, como por ejemplo para constatar la eficacia de una medida económica en la disminución de la inflación.

El primer paso en el proceso que podríamos llamar de elaboración de pronósticos es la recolección de datos. Esta etapa deberá realizarse con sumo cuidado; los datos que se obtengan deberán medir lo que realmente se desea y deben provenir de fuentes fidedignas, así mismo deben ser precisos y recolectados en el momento oportuno.

La utilización de un gráfico que represente a los valores de la serie ayuda a descubrir patrones que existen detrás de la serie. El siguiente gráfico (Figura 4.2) representa las ventas mensuales de pasajes por vía aérea en un país determinado, entre los años 1950 y 1961. Se observa que:

- Los valores de la serie crecen a lo largo del tiempo alrededor de una recta. La serie crece a largo plazo.
- Existen ciertas variaciones estacionales (patrones que se repiten en determinados meses del año).
- La amplitud de las oscilaciones alrededor de la recta va en aumento a medida que avanza el tiempo (la varianza de la serie no es constante a lo largo del tiempo).

FIGURA 4.2 *Serie de pasajeros*

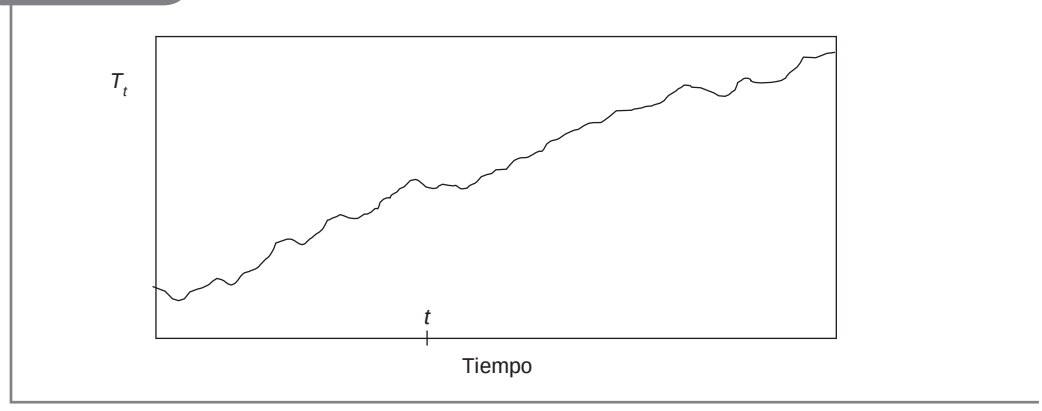


4.3 Modelos básicos para análisis de una serie de tiempo

Entre los modelos que a menudo se usan en el estudio de las series de tiempo están los de *tendencia*, los *estacionales* y los *estacionarios*. Se considera que muchos otros modelos son combinaciones de los anteriores.

Los modelos de ***tendencia***, T_t, aparecen cuando, a largo plazo, los valores de la serie crecen o decrecen. Estos modelos están representados por curvas suaves, como las curvas polinomiales. Generalmente la población de un país se incrementa; en este caso, un modelo de tendencia puede usarse de manera satisfactoria.

FIGURA 4.3 *Modelo de tendencia*



Los modelos se consideran que son *estacionales*, S_t, si los valores de la variable están influenciados por factores que tienden a repetirse cada cierto tiempo, en periodos menores a un año. Por ejemplo, las ventas de helados, que se incrementan en los meses de verano y decrecen en los otros, pueden modelarse usando modelos estacionales. Estos modelos reflejan las condiciones del clima, los días festivos, etcétera (ver Figura 4.4).

Los modelos ***estacionarios***, I_t, se utilizan para movimientos irregulares que fluctúan alrededor de una recta horizontal, con dispersión constante y provocados por factores esporádicos e imprevisibles, como desastres naturales, huelgas, epidemias, etcétera (ver Figura 4.5).

FIGURA 4.4 *Modelo estacional*

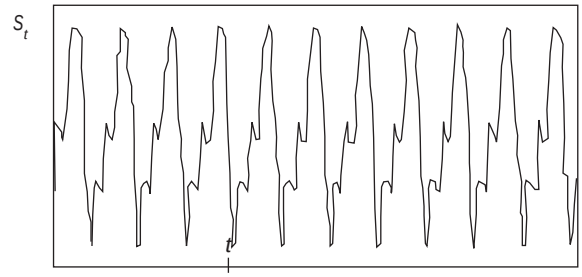
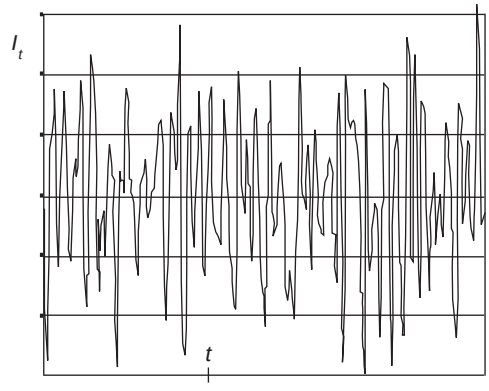


FIGURA 4.5 *Modelo estacionario*



Existen otros modelos que describen las características de series que se repiten después de tiempos largos (mayores de un año). Estos modelos se llaman *modelos cíclicos.* En la práctica, cuando no se involucran grandes periodos de tiempo, es difícil aplicar estos modelos, por ello se engloban dentro de los modelos de tendencia, considerándose de esta manera que la mayoría de las series pueden estudiarse usando modelos que son combinaciones de los modelos de tendencia, de estacionalidad y de estacionariedad.

Así, algunas series pueden expresar como $Y_t = T_t + S_t + I_t$ (modelo aditivo), donde cada sumando es una "componente" de la serie; otras pueden expresarse como:

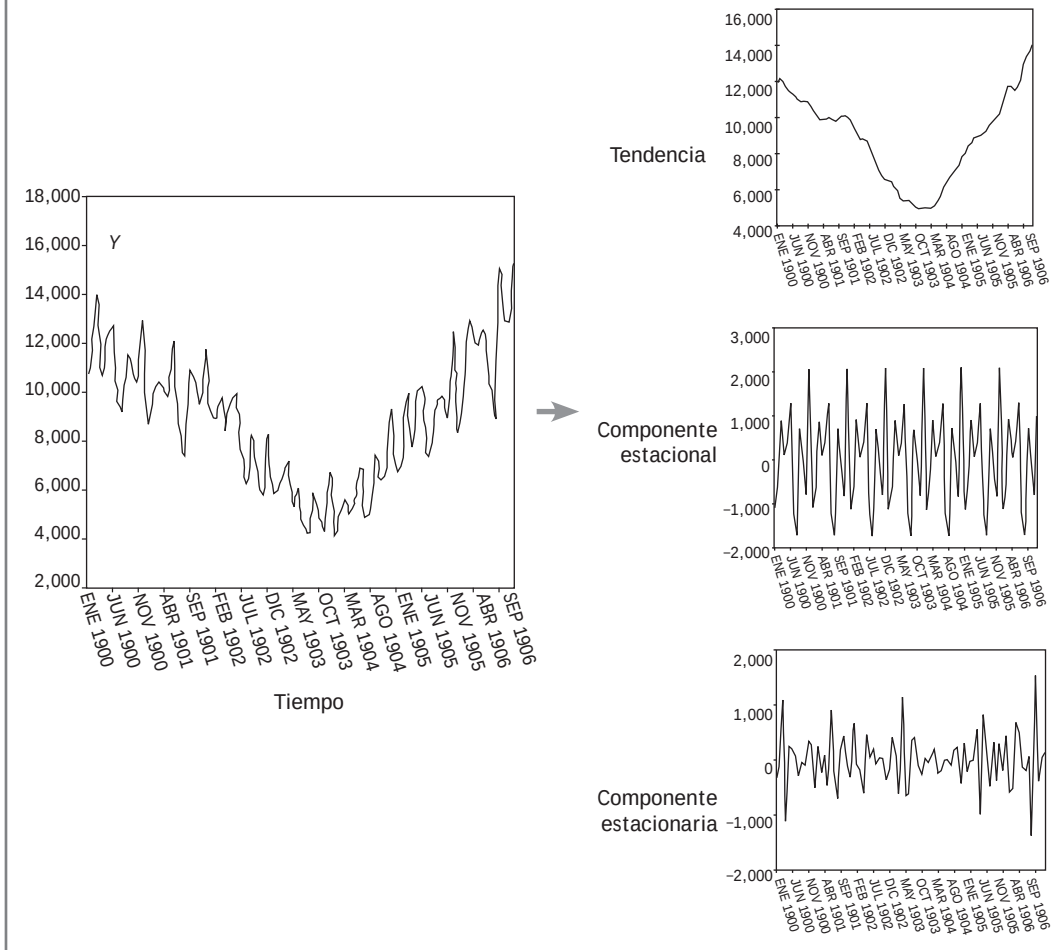
$$Y_t = T_t * S_t * I_t \text{ (modelo multiplicativo)}$$

En este tipo de modelos, la tendencia se mide en las mismas unidades que la serie Y_t; no así la componente estacional, la cual corresponde a números índices que varían alrededor de 0, en el caso del modelo aditivo, o alrededor de 1, en el caso del modelo multiplicativo.

En el caso del modelo aditivo, si el índice estacional es positivo, el valor de la serie está por encima del valor de la tendencia, y por debajo de la tendencia si este es negativo.

En el caso del modelo multiplicativo, si el índice estacional es mayor que 1, el valor de la serie está por encima del valor de la tendencia y por debajo de la tendencia si este es menor que 1.

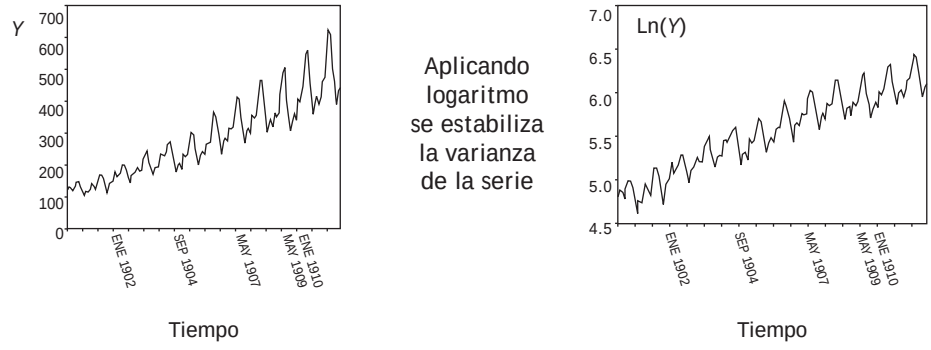
FIGURA 4.6 *Descomposición de una serie*



El modelo aditivo se usa cuando la varianza de la serie es constante a lo largo del tiempo. Si la varianza de la serie no es constante a lo largo del tiempo, se usa el modelo multiplicativo.

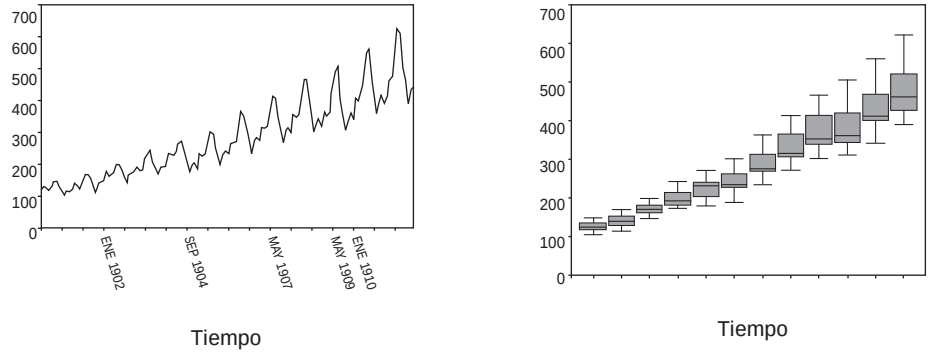
Algunas veces será necesaria la aplicación de algunas transformaciones como el logaritmo, para estabilizar la varianza de la serie. Otras transformaciones se usan, por ejemplo, para eliminar la tendencia de la serie.

FIGURA 4.7 *Transformación logaritmo*



El uso de gráficos auxiliares, como los gráficos de cajas, ayuda en el análisis de la serie. Para la siguiente serie, por ejemplo, los gráficos de cajas, obtenidos al usar grupos de 12 valores consecutivos, indican que la varianza de la serie no es constante.

FIGURA 4.8 *Estudio de la varianza de la serie*



4.4 Análisis de la tendencia: métodos de suavización

Diversos métodos básicos se usan para el estudio de la tendencia. Entre ellos están:

a) Los promedios móviles

b) Los métodos de suavización exponencial y

c) Los métodos de regresión

La elección del método a usar dependerá de la pericia del analista y del conocimiento de la naturaleza de los datos.

Métodos de promedios móviles

Este método, por su fácil aplicación, se usa a menudo para realizar pronósticos rápidos y de bajo costo. Permite, a partir de los datos originales, la obtención de una nueva serie más suave que refleja la tendencia sin el efecto de otras componentes como la estacionalidad o la parte irregular.

El método de promedios móviles se utiliza cuando se considera que son los últimos datos los que más influyen en el pronóstico de la serie.

Dada la serie $Y_1, Y_2,..., Y_n$ y el número natural *d,* la serie formada por:

$$\frac{Y_1 + ... + Y_d}{d}, \frac{Y_2 + ... + Y_{d+1}}{d}, \frac{Y_3 + ... + Y_{d+2}}{d}, ..., \text{etcétera,}$$

se llama serie de ***promedios móviles*** *de amplitud d.*

Nótese que en la construcción de esta nueva serie se va incorporando una nueva observación y abandonando la observación más antigua.

Para la serie de valores 120, 200, 240, 260, 300, 450, la serie de promedios móviles de amplitud $d = 3$ es:

$$\frac{120 + 200 + 240}{3}, \frac{200 + 240 + 260}{3}, \frac{240 + 260 + 300}{3}, \frac{260 + 300 + 450}{3}$$

En este caso, y en general cuando la amplitud es impar, la serie se llama *de promedios móviles centrados*.

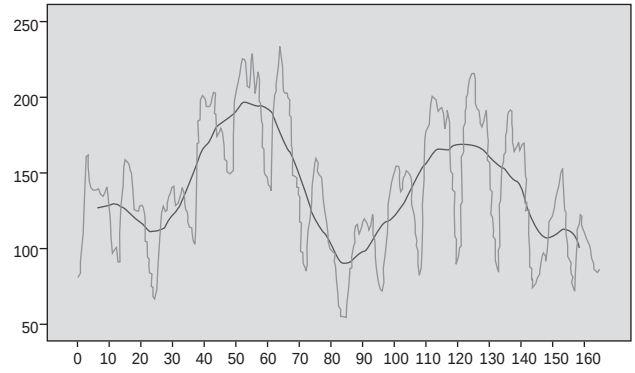
Si *d* es par la serie de promedios móviles no es centrada. Para obtener una serie centrada se hallan los promedios móviles de amplitud 2 de los promedios móviles no centrados.

EJEMPLO. *Ventas de electricidad en 166 días*

En la Figura 4.9 se muestran dos series; la serie más suave ha sido obtenida con los promedios móviles de amplitud 12 de la otra serie, que representa las ventas de fluido eléctrico realizadas por una empresa eléctrica durante 166 días. Algunos de los datos de la serie original, así como algunos valores de la serie de promedios móviles, se presentan a continuación.

FIGURA 4.9 *Promedios móviles*

Ventas	*Promedios móviles*
80.50	…
84.60	…
126.60	…
162.00	…
140.90	…
137.90	…
139.80	126.50
136.60	127.60
134.30	128.05
140.80	128.15
127.10	128.63
96.40	129.63
101.50	129.42
90.10	128.32
…	…
84.50	111.88
71.90	109.23
107.80	105.17
123.00	100.16
109.90	…
105.80	…
99.90	…
86.30	…
84.60	…
86.20	…



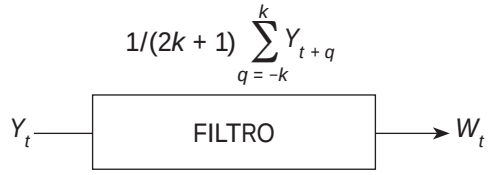
Si se desea, el primer promedio móvil centrado se halla calculando el promedio de los dos primeros promedios móviles no centrados de orden 12:

$$(1/2)\left[\frac{80.50 + 84.60 + \ldots + 96.40}{12} + \frac{84.60 + 84.60 + \ldots + 101.50}{12}\right] = 126.50$$

y así sucesivamente.

Se observa que los promedios móviles, actuando como un filtro, no dejan pasar la parte irregular ni la estacional, "suavizando" de este modo a la serie original.

FIGURA 4.10 *Filtro lineal*



En general, nótese que si la amplitud es 1, la serie de los promedios móviles sigue a la serie original, y que a medida que esta es mayor, la suavización es más severa.

La serie de promedios móviles es apropiada para predecir el valor futuro de la serie original cuando esta no tiene estacionalidad. Un valor futuro de la serie original se predice con el último valor de la serie suavizada. Si la serie tiene parte estacional, se usa el método de descomposición para la predicción, el cual se desarrolla más adelante.

Medida de la precisión de las predicciones

Para comparar la precisión de dos o más modelos aplicados a una serie se utilizan medidas como la ***media cuadrática del error***. Este valor es igual al promedio de los cuadrados de los residuales entre los valores observados de la serie original y los valores correspondientes de la serie suavizada.

Media cuadrática del error = $(1/k)\sum_{i=1}^{k}(Y_i - W_i)^2$; Y_i es el valor de la serie original, W_i es el valor de la serie suavizada.

Como criterio se tiene lo siguiente: de dos modelos aplicados a una serie, se utiliza el que tiene menor media cuadrática del error.

Método de suavización exponencial simple

Este método se aplica a series que tienen una tendencia que se orienta a la horizontalidad y que no poseen componente estacional.

Para una serie de valores Y_1, Y_2,..., Y_n, la serie de ***suavización*** exponencial simple se define como:

$$S_{t+1} = \alpha Y_t + (1 - \alpha)S_t\,,\ t = 1, 2,..., n$$

en donde $0 < \alpha < 1$.

El valor S_{t+1} puede usarse para predecir el valor de la serie original en el tiempo $t + 1$.

La relación puede escribirse como:

$$S_{t+1} = S_t + \alpha(Y_t - S_t)$$

Se aprecia que el nuevo pronóstico es igual al pronóstico anterior ajustado en α veces el error del pronóstico anterior.

Si en la última ecuación reemplazamos el valor S_t por su expresión $\alpha Y_{t-1} + (1 - \alpha)S_{t-1}$, y así sucesivamente, se tendrá

$$S_{t+1} = \alpha Y_t + \alpha(1 - \alpha)\ Y_{t-1} + \alpha(1 - \alpha)^2 Y_{t-2} + \alpha(1 - \alpha)^3 Y_{t-3} + ...$$

Se observa que el valor de predicción de la serie en el tiempo $t + 1$ puede considerarse como el promedio ponderado de las observaciones Y_t, Y_{t-1}, Y_{t-2},..., con ponderaciones α, $\alpha(\alpha - 1)$, $\alpha(\alpha - 1)^2$, ..., respectivamente.

Cuando α se acerca a 1, las ponderaciones correspondientes a observaciones más alejadas de la observación actual son cada vez menores que 1, y la serie suavizada prácticamente sigue a la serie original; en cambio, un valor α cercano a 0 equivale a tomar un número considerable de observaciones para la predicción.

La dificultad en el uso de los modelos exponenciales es la elección del valor de α. En general, un criterio que se puede usar para una buena elección de este parámetro es el de considerar el valor α de tal modo que la media cuadrática de los errores sea la menor posible. Sin embargo, puede considerarse para su elección que cuanto más pequeño sea, mayor será la suavización de la serie resultante.

En la práctica, el valor aconsejable para α es un valor comprendido entre 0.1 y 0.4.

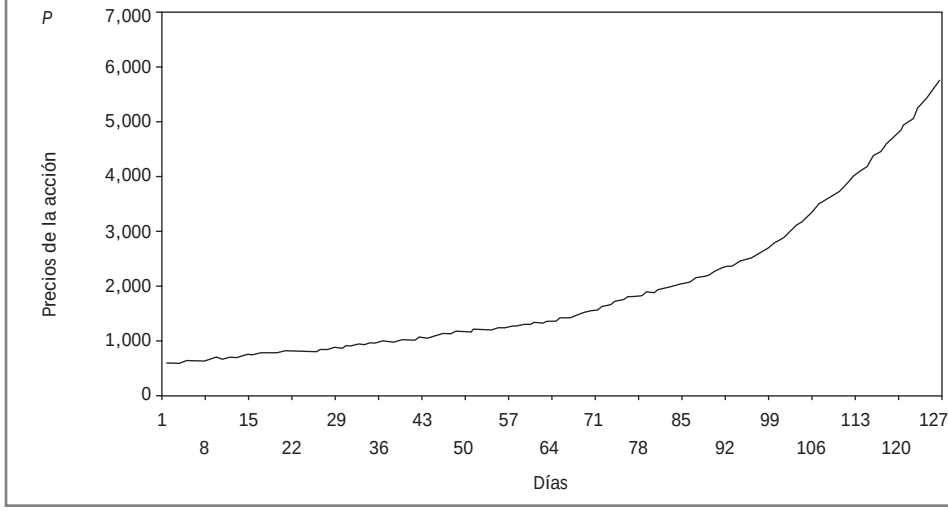
EJEMPLO. *Datos de Ledolter* 1983

La siguiente serie corresponde al valor de una variable de producción *P* durante 128 días consecutivos.

1 - 8	601.00	604.00	620.00	626.00	641.00	642.00	645.00	655.00
9 - 16	682.00	678.00	692.00	707.00	736.00	753.00	763.00	775.00
17 - 24	775.00	783.00	794.00	813.00	823.00	826.00	829.00	831.00
25 - 32	830.00	838.00	854.00	872.00	882.00	903.00	919.00	937.00
33 - 40	927.00	962.00	975.00	995.00	1,001.00	1,013.00	1,021.00	1,028.00
41 - 48	1,027.00	1,048.00	1,070.00	1,095.00	1,113.00	1,143.00	1,154.00	1,173.00
49 - 56	1,178.00	1,183.00	1,205.00	1,208.00	1,209.00	1,223.00	1,238.00	1,245.00
57 - 64	1,258.00	1,278.00	1,294.00	1,314.00	1,323.00	1,336.00	1,355.00	1,377.00
65 - 72	1,416.00	1,430.00	1,455.00	1,480.00	1,514.00	1,545.00	1,589.00	1,634.00
73 - 80	1,669.00	1,715.00	1,760.00	1,812.00	1,809.00	1,826.00	1,871.00	1,892.00
81 - 88	1,946.00	1,983.00	2,013.00	2,045.00	2,048.00	2,097.00	2,140.00	2,171.00
89 - 96	2,208.00	2,272.00	2,311.00	2,349.00	2,362.00	2,442.00	2,479.00	2,528.00
97 - 104	2,571.00	2,634.00	2,684.00	2,790.00	2,890.00	2,964.00	3,085.00	3,159.00
105 - 112	3,237.00	3,358.00	3,489.00	3,588.00	3,624.00	3,719.00	3,821.00	3,934.00
113 - 120	4,028.00	4,129.00	4,205.00	4,349.00	4,463.00	4,598.00	4,725.00	4,827.00
121 - 128	4,939.00	5,067.00	5,235.00	5,408.00	5,492.00	5,653.00	5,828.00	5,965.00

La gráfica de los valores de la serie se aprecia en la Figura 4.11. La serie presenta una tendencia no lineal.

FIGURA 4.11 *Serie de precios de la producción P*



Analizaremos el cambio porcentual diario de los valores de la serie:

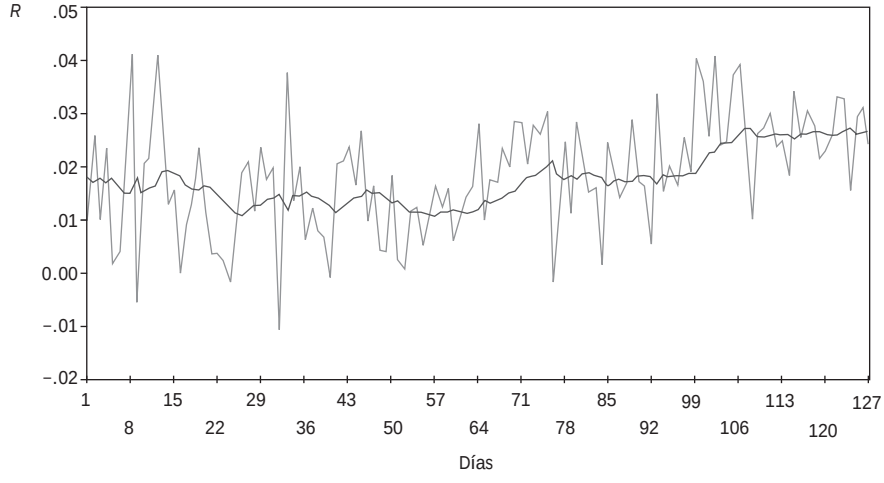
$$R_t = \frac{P_{t+1} - P_t}{P_t} * 100$$

La serie de los cambios porcentuales cambia muy poco en la tendencia. Ello permite suavizar a esta serie con un alisamiento exponencial simple. El valor del parámetro estimado en este modelo es $\alpha = 0.100$.

A continuación se muestran los 10 primeros valores de esta serie, los pronósticos y los residuales al cuadrado. El error medio cuadrático resultante es EMC = 0.000087.

FIGURA 4.12 *Serie de los cambios porcentuales diarios y serie suavizada*

R_t	*Pronóstico*	*Residual*	*Residuales al cuadrado*
.0050	.0144	−.0095	.0001
.0258	.0135	.0123	.0002
.0096	.0147	−.0051	.0000
.0234	.0142	.0092	.0001
.0016	.0151	−.0136	.0002
.0047	.0137	−.0091	.0001
.0153	.0128	.0024	.0000
.0396	.0131	.0265	.0007
−.0059	.0158	−.0217	.0005
.0202	.0136	.0067	.0000



Métodos de regresión

La tendencia de una serie que no contiene parte estacional puede expresarse como una función polinómica cuya variable independiente es el tiempo. Si la correlación lineal entre los valores de la serie Y_t y los valores que representan el tiempo es alta, se puede usar la ecuación de regresión lineal para representar a la tendencia:

$$Y_t = \beta_0 + \beta_1 t$$

4.5 Métodos de descomposición de una serie

Uno de los métodos clásicos para el estudio de las series es el método de descomposición. Este procedimiento permite, como su nombre lo indica, la descomposición de la serie en sus componentes: la tendencia, la estacionalidad y la parte irregular estacionaria.

Para aplicar el método de descomposición de la serie es necesario tener en cuenta la dispersión (la varianza) de los valores de la serie a lo largo del tiempo.

Si la serie tiene varianza constante a lo largo de su tendencia podrá usarse el modelo ***aditivo***; en cambio, si la serie no tiene varianza constante el modelo adecuado será el modelo ***multiplicativo.*** Sin embargo, a las series se les puede aplicar transformaciones, como el logaritmo, que estabilizan la varianza. De esta manera bastará con estudiar el modelo aditivo.

Se considera que la serie $Y_1,\ldots, Y_n$ puede ajustarse con el *modelo aditivo* si sus valores se pueden expresar como:

$$Y_t = T_t + S_t + I_t$$

en donde:

T_t es un modelo de tendencia que cambia suavemente. Describe el comportamiento de la serie a largo plazo.

S_t es un modelo estacional de periodo d y que cumple con $S_1 + \ldots + S_d = 0$.

I_t corresponde a un modelo estacionario. Esta componente contiene la parte aleatoria de la serie.

Se supone además que las componentes que constituyen la serie son independientes.

TABLA 4.1 *Descomposición de la serie*

Cuatrimestres	*T*	Y_t	W_t	$r_t = Y_t - W_t$	S_t	D_t
1 cuatrimestre	1	9.3	---	---	1.91	7.4
2 cuatrimestre	2	8.4	9.3	−0.9	−0.30	8.7
3 cuatrimestre	3	10.3	11.3	−1.0	−1.61	11.9
1 cuatrimestre	4	15.3	13.5	1.8	1.91	13.4
2 cuatrimestre	5	14.9	15.1	−0.2	−0.30	15.2
3 cuatrimestre	6	15.0	16.5	−1.5	−1.61	16.6
1 cuatrimestre	7	19.5	18.5	1.0	1.91	17.6
2 cuatrimestre	8	21.1	20.3	0.8	−0.30	21.4
3 cuatrimestre	9	20.4	22.9	−2.5	−1.61	22.0
1 cuatrimestre	10	27.1	24.3	2.8	1.91	25.2
2 cuatrimestre	11	25.4	26.5	−1.1	−0.30	25.7
3 cuatrimestre	12	27.0	...	...	−1.61	28.6

Procedimiento para descomponer una serie

Para ilustrar el método de descomposición con el modelo aditivo, usaremos la serie Y_t, con 12 valores, que aparece en la Tabla 4.1 (tercera columna), y cuyo periodo es $d = 3$.

1. Aplicando el método de los promedios móviles, con amplitud igual al periodo de la componente estacional, se elimina el efecto de las componentes estacional e irregular.

Si el periodo es impar, digamos $d = 2q + 1$, se recomienda usar los promedios móviles centrados.

Si el periodo de la componente estacional es par, $d = 2q$, se usan los promedios móviles de amplitud 2 de los promedios móviles de amplitud $2q$, con la finalidad de centrar los valores de la serie de los promedios móviles.

Para la serie Y_t, la serie de los promedios móviles de amplitud 3, W_t, aparece en la cuarta columna.

2. Separar de la serie original, por sustracción, los resultados obtenidos en el paso 1 para aislar la estacionalidad y la componente irregular. Así se obtienen la serie de valores $r_i = Y_i - W_i$ que aparece en la quinta columna.

3. Estimar la componente estacional. Esta estimación consiste en calcular los índices estacionales

$$S_1, S_2, S_3$$

cuya suma deberá ser igual a 0.

El primer índice estacional se obtiene promediando los valores de la quinta columna correspondientes al primer cuatrimestre:

$$\frac{1.8 + 1.0 + 2.8}{3} = 1.87$$

El segundo índice estacional se obtiene promediando los valores de la quinta columna correspondientes al segundo cuatrimestre:

$$\frac{-0.9 - 0.2 + 0.8 - 1.1}{4} = -0.35$$

El tercer índice estacional se obtiene promediando los valores de la sexta columna correspondiente al tercer cuatrimestre:

$$\frac{-1 - 1.5 - 2.5}{3} = -1.66$$

Para que los índices estacionales sumen 0, a cada valor obtenido en la anterior operación se le resta su promedio, [1.87 + (−0.37) + (−1.66)]/3. Como resultado se obtienen los siguientes índices estacionales corregidos:

$$S_1 = 1.91,\ S_2 = 0.30 \text{ y } S_3 = -1.61.$$

Estos índices estacionales se repiten en cada año. Aparecen en la sexta columna.

4. A cada valor de la serie original se le resta el índice estacional correspondiente, obteniéndose la serie *desestacionalizada* $D_t = Y_t - S_t$. Esta serie no contiene la componente de estacionalidad. Aparece en la séptima columna.

5. En ausencia de la componente de estacionalidad la serie se suaviza (si es posible, se usa la regresión).

En este caso, la tendencia se calcula aplicando la suavización por regresión a los puntos (t, D_t) con $t = 1, 2, 3, \ldots, 12$, obteniéndose la ecuación de la recta:

$$f(t) = 5.49 + 1.89t$$

La serie original puede aproximarse con esta función más la serie de los índices estacionales:

$$Y_t \approx f(t) + S(t)$$

Estimado así el modelo, los valores de predicción pueden escribirse como $\hat{Y}_t = 5.49 + 1/89t + S_t$, en donde S_t es el índice estacional correspondiente al tiempo *t.*

El valor de predicción en el tiempo 13, en donde el índice estacional tiene el valor 1.91, es:

$$\hat{Y}_{13} = f(13) + S_1 = 5.49 + 1.89(13) + 1.91 = 31.97$$

APLICACIÓN: El caso de la editora Liber

La empresa editora Liber desarrolló un centro muy importante para la publicación y venta de obras literarias y textos escolares. Por cambio de giro, los dueños de la empresa decidieron poner en venta el centro ofertándolo públicamente. Liber llevó a cabo una evaluación previa del negocio, la cual comprendió varios aspectos, entre los que se consideró: el estudio del área de mercado a partir de pronósticos sobre la población escolar y la población de profesionales, y el análisis de la serie de las ventas de textos y obras literarias realizadas los últimos años, con el objetivo de realizar pronósticos para los próximos tres meses.

La serie Y_t de las ventas mensuales que Liber realizó se presenta en la Tabla 4.2 y se representa en la Figura 4.13. Se observó que la serie tenía aproximadamente una varianza constante, presentando estacionalidades con un periodo aproximadamente igual a $d = 2q = 12$.

TABLA 4.2 *Ventas*

Año	*E*	*F*	*M*	*A*	*M*	*J*	*J*	*A*	*S*	*O*	*N*	*D*
1999	700	650	635	675	750	800	725	650	675	750	800	975
2000	750	725	675	700	825	850	825	700	700	800	825	1,000
2001	775	775	750	735	810	870	805	745	750	825	875	1,050
2002	815	775	780	760	850	920	855	810	795	865	960	1,090
2003	850	810	765	750	870	950	875	850	835	895	1,090	1,210
2004	925	840	825	800	890	1,000	920	860	855	930	1,090	1,220
2005	945	895	845	845	915	1,015	960	875	895	995	1,120	1,260

Utilizando el método de descomposición la serie de las ventas se descompuso en tres series:

- La tendencia T_t.
- La serie de los índices estacionales, S_t.
- La serie de los errores, que corresponde a los residuales entre la serie original de las ventas y la suma de la tendencia y la serie de los índices estacionales.

Estas series aparecen en la Tabla 4.3.

La serie de la tendencia puede ser ajustada con la recta $f(t) = 716.329 + 3.219t$.

Los valores futuros de la serie pueden aproximarse usando el ajuste lineal de la tendencia y la estacionalidad: $Y_t \approx 716.329 + 3.219t + S_t$.

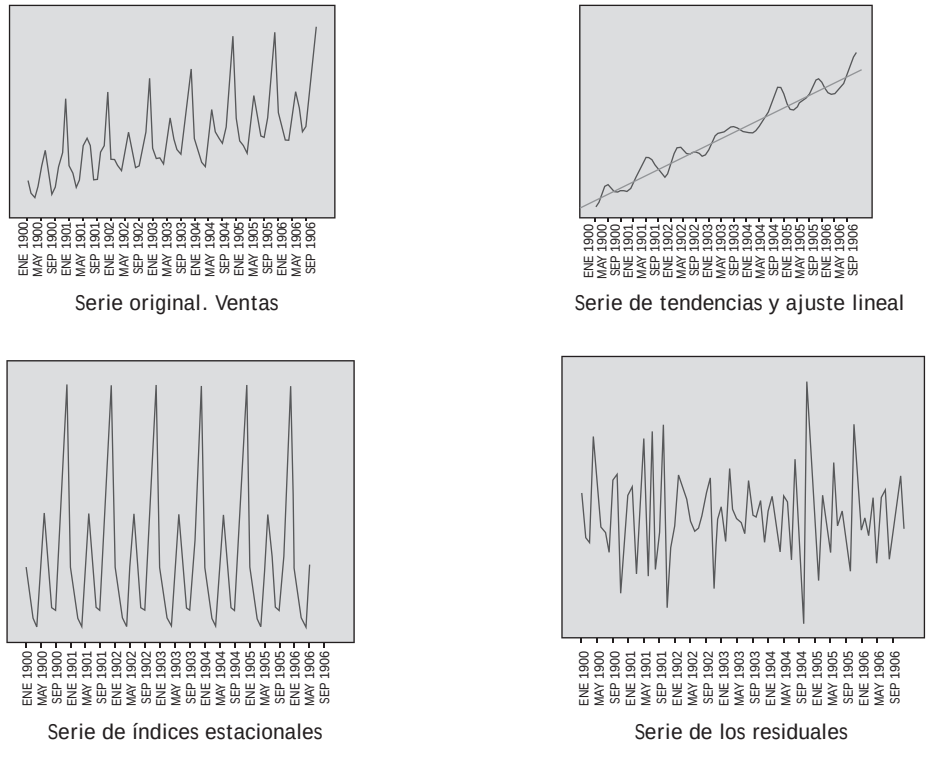
Por ejemplo, el valor de las ventas pronosticadas para enero de 1985 es:

$$Y_{85} \approx 716.329 + 3.219(85) + S_{85} = 716.329 + 3.219(85) + (-9.554) = 980.390$$

TABLA 4.3 *Promedios móviles, índices estacionales y serie desestacionalizada*

	***Tendencia* (1)**	***Índices estacionales* (2)**	***Errores* (3)**
1	701.08796	−9.55440	8.46644
2	711.04745	−52.74884	−8.29861
3	730.96644	−85.83912	−10.12731
4	745.62693	−97.40162	26.77469
5	748.19637	−6.32523	8.12886
6	740.75810	63.64005	−4.39815
7	732.98032	−1.42940	−6.55093
8	732.99961	−69.83218	−13.16744
9	735.66165	−73.82523	13.16358
10	735.28356	−.63079	15.34722
11	733.20795	92.87616	−26.08410
12	740.39159	241.07060	−6.46219
13	752.22029	−9.55440	7.33410
...	...	...	...
73	960.55363	−9.55440	−5.99923
74	948.37384	−52.74884	−.62500
75	938.74421	−85.83912	−7.90509
76	936.73804	−97.40162	5.66358
77	938.19637	−6.32523	−16.87114
78	945.20255	63.64005	6.15741
79	951.86921	−1.42940	9.56019
80	960.22184	−69.83218	−15.38966
81	975.10610	−73.82523	−6.28086
82	993.61690	−.63079	2.01389
83	1013.89468	92.87616	13.22917
84	1024.03356	241.07060	−5.10417

FIGURA 4.13 *Serie original y sus componentes*



Serie original. Ventas

Serie de tendencias y ajuste lineal

Serie de índices estacionales

Serie de los residuales

APLICACIÓN: El caso del hotel Melodía

El caso que se aborda ahora corresponde a la administración de servicios; proviene del sector hotelería, pero bien puede aplicarse a la industria de los viajes.

Uno de los problemas administrativos que a menudo enfrentan los hoteles es el de las reservas no cubiertas. Muchos clientes hacen reservas y después no aparecen o cancelan las reservas pocos antes de su llegada. Este problema, conjuntamente con otros relacionados con la capacidad del hotel, así como la sobreventa, deben ser enfrentados diariamente por la administración; por ello el pronóstico del número de habitaciones que estarán ocupadas será muy importante, pues no quisieran tener habitaciones vacías, pero tampoco desean decirles "no" a los clientes por haber vendido más habitaciones de las que tiene el hotel.

El hotel Melodía, que tiene 500 habitaciones, no es ajeno a este problema. Todos los fines de semana el administrador de servicios urge de la predicción del número de habitaciones que serán ocupadas la próxima semana. Para ello toma en cuenta el número de habitaciones reservadas y las habitaciones ocupadas en los diferentes días de la semana (1: lunes,…,7: domingo). En la tabla siguiente se ha registrado además el índice de acierto entre la demanda y la reserva.

Prediciendo el índice de acierto (*índice de acierto* = *demanda/reserva*) y multiplicándolo por la reserva se tendrá la predicción de la demanda. El lector puede realizar la tarea del administrador usando uno de los métodos desarrollados para las series de tiempo.

Día de la semana	Demanda	Reservas	Índice de acierto	Día de la semana	Demanda	Reservas	Índice
1	377	388	0.97	4	386	393	0.98
2	404	402	1.01	5	451	425	1.06
3	501	476	1.05	6	445	470	0.94
4	480	523	0.91	7	482	509	0.94
5	475	520	0.91	1	365	382	0.95
6	489	518	0.94	2	432	450	0.96
7	453	457	0.99	3	487	511	0.95
1	440	364	1.20	4	470	514	0.91
2	331	376	0.88	5	485	505	0.96
3	453	426	1.06	6	478	512	0.93
4	410	405	1.01	7	469	535	0.87
5	498	446	1.11	1	468	433	1.08
6	472	488	0.96	2	410	457	0.89
7	450	475	0.94	3	460	486	0.94
1	445	422	1.05	4	470	481	0.97
2	460	452	1.01	5	458	428	1.07
3	475	483	0.98	6	486	482	1.00
4	487	517	0.94	7	410	465	0.88
5	450	510	0.88	2	475	456	1.04
6	465	517	0.89	3	400	419	0.95
7	485	446	1.08	4	402	401	1.00
1	365	380	0.96	5	390	395	0.98
2	328	362	0.90	6	420	435	0.96
3	337	367	0.91	7	481	509	0.94
4	365	382	0.95	1	...	365	0.91
5	349	374	0.93	2	...	310	1.10
6	333	365	0.912	3	...	470	...
7	318	357	0.89	4	...	350	...
1	334	366	0.91	5	...	487	...
2	387	393	0.98	6	...	441	...
3	320	358	0.89	...	...	...	...

EJERCICIOS

1. En 15 meses consecutivos las ventas que tuvo una empresa, en miles de dólares, fueron como sigue:

   70.3 70.0 80.2 60.9 80.7 80.20 90.0 100.4 90.4 80.2 100.8 100.7 90.4 80.6 110.2

   a) Graficar los valores de la serie.

   b) Suavizar la serie usando el método de promedios móviles de amplitud 3. Hallar el pronóstico para el mes 16.

   c) Suavizar la serie, usando el método de promedios móviles de amplitud 4. Centrar la serie suavizada.

   d) De las series halladas en b) y c) y usando la media cuadrática del error, indicar la serie suavizada que mejor ajusta a los datos.

2. Los siguientes datos corresponden al número de teléfonos móviles vendidos durante 20 meses consecutivos por la empresa telefónica Starmov.

   318,250 318,130 318,200 317,900 318,000 317,800 317,840 317,800 317,880 317,870
   317,700 318,000 317,900 318,000 318,200 318,100 318000 319,500 325,500 328,000

   a) Graficar los valores de la serie.

   b) Suavizar la serie utilizando el modelo exponencial simple y usando $\alpha = 0.1$, $\alpha = 0.3$, $\alpha = 0.5$, $\alpha = 0.7$.

   c) Evaluar la precisión de los pronósticos que se pueden realizar con cada una de las series en b). ¿Con cuál de las series halladas en b) se obtiene un mejor pronóstico?

3. Los siguientes datos corresponden al número de trabajadores en la industria metalmecánica en cuatro años consecutivos.

	Año 1	*Año* 2	*Año* 3	*Año* 4
Enero	37,700	35,690	374,200	38,600
Febrero	37,600	36,440	37,500	38,843
Marzo	37,780	36,200	37,460	39,200
Abril	37,300	36,136	37,630	39,600
Mayo	37,090	36,600	37,560	39,700
Junio	36,870	36,420	37,780	39,500
Julio	37,070	36,410	37,500	39,800
Agosto	36,240	36,700	37,600	40,300
Septiembre	36,800	37,300	38,100	40,200
Octubre	36,400	37,000	38,500	40,340
Noviembre	36,230	36,600	38,500	40,280
Diciembre	36,047	36,400	38,450	41,200

   Usar el modelo exponencial simple con $\alpha = 0.3$ para suavizar la serie. Utilizar el modelo estimado para predecir el número de trabajadores de la industria metalmecánica en enero del año 5. Evaluar el ajuste realizado.

4. El número de visitantes a un centro arqueológico, durante seis años consecutivos, fue como sigue.

	Año 1	*Año* 2	*Año* 3	*Año* 4	*Año* 5	*Año* 6
Periodo I	680,940	753,990	824,700	919,030	985,910	1,044,770
Periodo II	752,530	818,160	902,200	988,830	1,052,230	1,122,180
Periodo III	752,580	796,720	874,120	942,130	990,590	1,038,590
Periodo IV	818,160	954,720	1,044,090	1,011,960	1,177,500	1,234,720

a) Graficar la serie de valores.
b) Aplicar el método de descomposición para analizar la serie del número de visitantes.
c) Usar los resultados de la descomposición para predecir el número de visitantes en el periodo I del año 7.

5. A continuación se indican las exportaciones mensuales de petróleo, en miles de barriles, desde el país VV, durante tres años consecutivos.

	Año 1	*Año* 2	*Año* 3
Enero	2,700	2,800	2,950
Febrero	2,800	3,050	3,100
Marzo	3,050	3,150	3,300
Abril	3,300	3,450	3,800
Mayo	3,400	3,650	3,900
Junio	4,150	4,300	4,900
Julio	4,600	5,000	5,200
Agosto	3,900	4,350	4,300
Septiembre	3,400	3,600	3,900
Octubre	3,400	3,700	3,950
Noviembre	3,300	3,550	3,800
Diciembre	2,950	2,200	3,500

a) Graficar la serie de valores.
b) Analizar la serie por el método de descomposición.
c) Usar la tendencia y la estacionalidad para predecir las exportaciones en el mes de enero del año 4.

6. Los registros de pasajeros, en miles, que llegaron al país fueron como sigue.

	***Bimestre* 1**	***Bimestre* 2**	***Bimestre* 3**	***Bimestre* 4**	***Bimestre* 5**	***Bimestre* 6**
Año 1	3,845	3,860	4,600	4,400	4,700	4,700
Año 2	4,400	4,390	4,100	4,700	5,100	5,010
Año 3	4,870	4,800	5,350	5,150	5,400	5,410
Año 4	5,065	5,170	5,800	5,580	6,810	5,930
Año 5	5,675	5,840	6,500	6,330	6,680	6,520
Año 6	6,010	5,850	6,410	6,180	6,650	6,470
Año 7	6,220	6,100	6,810	6,540	6,920	6,750

a) Graficar la serie de valores.
b) Aplicar el método de descomposición estacional para analizar la serie.
c) Evaluar la precisión de los pronósticos que se hallaron a partir del análisis realizado en b).

7. La venta de tractores durante siete años consecutivos en una región fue como sigue.

	***Año* 1**	***Año* 2**	***Año* 3**	***Año* 4**	***Año* 5**	***Año* 6**	***Año* 7**
Periodo I	620	560	460	660	650	860	950
Periodo II	470	440	310	440	510	620	710
Periodo III	360	310	249	340	440	505	560
Periodo IV	510	410	505	560	705	740	850

a) Suavizar la serie usando el método de promedios móviles de amplitud 4.
b) Suavizar la serie usando el modelo de exponencial simple. Usar $\alpha = 0.4$.
c) Comparar la precisión de los pronósticos realizados con las series hallados en a) y b).

8. El consumo de energía en kilowatts-hora de una empresa de calzado de lunes a jueves y en seis lapsos de cuatro horas durante el día fue como sigue.

Lapsos	**1**	**2**	**3**	**4**	**5**	**6**
Lunes				125,000	113,500	41,500
Martes	19,300	33,200	99,500	124,000	112,100	48,500
Miércoles	31,200	36,500	120,200	160,000	127,800	74,200
Jueves	27,800	33,100	154,500			

a) Graficar la serie.
b) Usar el método de descomposición para analizar la serie.
c) Graficar la tendencia, la parte estacional y la parte irregular de la serie.
d) Usar la tendencia y la serie de índices estacionales para predecir el consumo de energía en el lapso 4 del día jueves. Evaluar la precisión de los pronósticos.

RESPUESTAS A LOS EJERCICIOS

1. b) Para pronosticar el valor de la serie para el mes 16 usar el último promedio móvil. **2.** Elegir la serie de suavización exponencial para la cual la media cuadrática de los errores es la menor. **3.** Para enero del año 5 el pronóstico es $S_{49} = S_{48} + 0.3(41200 - S_{48})$. La serie S es la suavización exponencial. El ajuste realizado se evalúa con la media cuadrática del error. **4.** Observar que la serie tiene varianza constante. Usar el modelo aditivo. Utilizar periodos de tamaño 4. Se obtienen así 4 índices estacionales.

CAPÍTULO 5

Probabilidad

Andrei Kolmogorov

Andrei Kolmogorov nació en Rusia, el 25 de abril de 1903. Muy joven ingresó a la Universidad Estatal de Moscú y se graduó en 1925.

Kolmogorov revolucionó las probabilidades introduciendo los axiomas que definen la probabilidad de un evento. Estos axiomas permitieron la amplia aplicación de las probabilidades a diferentes campos como la física, biología e ingeniería. Sin embargo, sus aportes también se produjeron en el campo de las ecuaciones diferenciales, en donde diferentes propiedades llevan su nombre.

Su aporte a las matemáticas no solo se circunscribió al campo universitario, pues también se interesó en las matemáticas escolares, promoviendo diferentes programas de entrenamiento para los profesores de esta área.

Andrei Kolmogorov murió el 20 de octubre de 1987, cuando aún era profesor de la Universidad Estatal de Moscú.

CONTENIDO

5.1 Introducción

Muchos eventos que tienen que ver con el azar son situaciones que diversas empresas deben tener en cuenta si desean ser competitivas y eficientes. Los bancos se enfrentan a menudo con este tema al no tener la seguridad de que los clientes a los cuales se les ha hecho un préstamo lo devuelvan, o cuando las tarjetas de crédito que otorgan pueden ser usadas de manera fraudulenta. Parecido es el panorama que se presenta para las empresas de seguros, las que muchas veces tienen que lidiar con clientes que usan los seguros indebidamente. Las empresas industriales no son ajenas a estos hechos; la calidad de los productos que estas elaboran es el resultado de una serie de experiencias aleatorias.

Los modelos que se usan para analizar situaciones como las anteriores y para aquellas cuyos resultados aparecen como consecuencia de "experiencias al azar y de la incertidumbre en general", son dados por la teoría de la probabilidad. Estos modelos se llaman *modelos probabilísticos*; permiten el estudio de procesos cuya característica es la variabilidad y se caracterizan principalmente porque describen los patrones que rigen los resultados, pero no los resultados en sí. Con estos modelos no es posible tener una fórmula que indique si una empresa tendrá éxito o no al realizar una determinada operación, pero sí se puede argumentar, por ejemplo, que la empresa tendrá mayor posibilidad o no de tener éxito en la operación. El concepto más importante para obtener un modelo probabilístico es el de probabilidad.

5.2 La probabilidad

Para obtener este concepto se desarrollan previamente las siguientes ideas.

Experimento aleatorio

> *Un **experimento aleatorio** es toda experiencia que se puede repetir indefinidamente, obteniéndose resultados diferentes e imprevisibles aun cuando se repita de la misma manera.*

Existen experiencias cuyos resultados son imprevisibles y que no pueden repetirse cuantas veces se desee; tal es el caso del experimento consistente en observar si cierto día lloverá o no en un lugar determinado a las 8 a. m. Aun cuando esta experiencia no se puede repetir cuantas veces se desee, el mismo día y en la misma hora, se puede imaginar que se trata de la primera de una serie ilimitada de experiencias semejantes.

Los siguientes experimentos son aleatorios:

- El lanzamiento de una moneda para luego anotar el resultado.
- La elección de una familia, para luego anotar el número de hijos que tiene.
- La medición del peso de los artículos fabricados por un mismo proceso.
- La llegada de un autobús a un paradero para luego anotar el tiempo de llegada.
- La realización de un negocio para luego anotar la ganancia obtenida.

Espacio muestral

Para modelar el patrón de resultados de una experiencia aleatoria es necesario conocerlos previamente. *El conjunto de todos los resultados de una experiencia aleatoria se llama espacio muestral.*

El espacio muestral se denota con Ω.

EJEMPLO. *Variabilidad de los resultados de una experiencia*

Una experiencia se repite hasta que sea exitosa (e). El espacio Ω de este experimento aleatorio está formado por los siguientes resultados, en donde *f* indica fracaso:

e, la experiencia es exitosa en la primera realización.

fe, la experiencia es exitosa en la segunda realización.

ffe, la experiencia es exitosa en la tercera realización, etcétera.

EJEMPLO. *Midiendo las ventas en una empresa*

Si se elige una empresa para luego anotar sus ventas *X,* en dólares, los resultados que conforman el espacio muestral están en el intervalo [0, *a*]. En este intervalo, el valor *a* puede ser un número como 0, 100,000, 3,700,000... No sabiendo a ciencia cierta cuál será el valor *a*, conviene tomar como espacio muestral al intervalo [0, +∞[.

Evento

*Un **evento** de un experimento aleatorio E es cualquier subconjunto del espacio muestral.*

En particular, el espacio muestral Ω y el conjunto vacío Φ son eventos, pues son subconjuntos de Ω.

Un evento puede conocerse enumerando sus elementos o usando alguna proposición que los describa.

EJEMPLO. *Experiencia con dos resultados posibles*

Al lanzar una moneda para luego ver el resultado, se obtiene el espacio muestral:

Ω = {*c*, *s*}, donde *c* corresponde a cara y *s* corresponde a sello.

Los eventos que se pueden obtener son: {c}, {s}, Ω y Φ.

Si al realizar un experimento aleatorio el resultado pertenece a un evento *A* se dice que el resultado *favorece* al evento *A* o que el evento *A se realiza*[1]. Si al lanzar un dado se obtiene un número par, entonces el evento *A* = {2, 4, 6} se realiza. El espacio muestral Ω se llama también suceso *siempre cierto*, siempre se realiza, mientras que Φ se llama suceso *imposible*, nunca se realiza.

Si con cada elemento del espacio muestral se forman subconjuntos unitarios, se tendrán los eventos llamados *eventos elementales*.

Operaciones con eventos

Entre los eventos de un espacio muestral, se puede definir operaciones tales como la *intersección de eventos, reunión de eventos* y *complemento de un evento*.

Para los eventos *A* y *B*, la ***intersección*** de *A* y *B*, que se denota con $A \cap B$, es el evento formado por los elementos comunes a los eventos *A* y *B* (ver Figura 5.1a).

El evento $A \cap B$ se realiza si *A* y *B* se realizan a la vez.

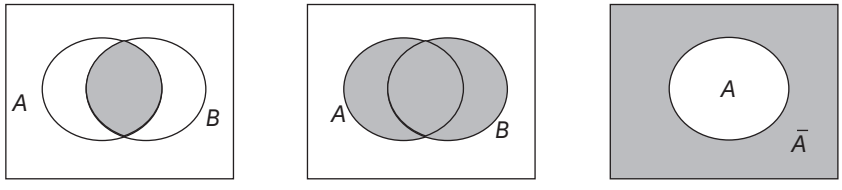
Si $A \cap B = \Phi$, se dice que *A* y *B* son *mutuamente excluyentes* o *disjuntos*.

La *reunión* o *unión* de los eventos *A* y *B*, que se denota con $A \cup B$, es el evento formado por los elementos del espacio muestral que están en *A*, en *B* o en ambos eventos a la vez (ver Figura 5.1b).

El evento $A \cup B$ se realiza si al menos uno de los eventos *A* o *B* se realiza.

El *complemento* de *A*, que se denota con $\bar{A}$, es el evento formado por los elementos de Ω que no están en *A* (ver Figura 5.1c). El evento $\bar{A}$ se realiza cuando el evento *A* no se realiza.

FIGURA 5.1 *Operaciones*



a) $A \cap B$ b) $A \cup B$ c) $\bar{A}$

[1] También se acostumbra decir que el evento sucede u ocurre.

EJEMPLO. *Prohibido fumar*

Si A representa al evento que indica a todos las personas que fuman y B representa a todas las personas que padecen de enfermedades bronquiales, entonces $A \cap B$ es el evento que representa a todas las personas que fuman y que padecen de enfermedades bronquiales, $A \cup B$ es el evento que representa a todas las personas que fuman o que padecen de enfermedades bronquiales y $\bar{A}$ es el evento que representa a todas las personas que no fuman.

La probabilidad de un evento

Para medir la posibilidad de la ocurrencia de un evento se usa la probabilidad. Históricamente, para llevar a cabo este objetivo se usó el *concepto clásico de probabilidad.* Este concepto, ideado por Pierre S. Laplace (1749-1827), tuvo su origen en los juegos de azar y se aplica a experiencias aleatorias para las *cuales el espacio muestral es finito y con resultados igualmente posibles*. Así se definió la probabilidad de un evento A como el cociente:

$$\frac{\textit{Número de resultados que favorecen al evento } A}{\textit{Número de resultados posibles}}$$

Al lanzar un dado equilibrado, la probabilidad de que aparezca un número par es igual al cociente:

$$\frac{\textit{Número de resultados que favorecen al evento ``par''}}{\textit{Número de resultados posibles}} = \frac{3}{6} = 0.5$$

Por mucho tiempo se manejó este concepto clásico de probabilidad; sin embargo, no incluía situaciones como, por ejemplo, la probabilidad de que un determinado aparato dure un tiempo comprendido entre 100 y 200 horas. El matemático ruso Andrei Kolmogorov (1903-1987) amplió este concepto mediante una lista de axiomas consistentes con los conceptos mencionados anteriormente, y que de manera racional resuelven el problema de medir la incertidumbre, aun cuando el número de resultados posibles de la experiencia sea infinito. Estos axiomas se comprenden si se toma en cuenta que el espacio muestral es el *evento seguro*, el que siempre se realiza, de ahí que si se desea determinar una medida o ponderación para este evento, esta debe ser máxima. En oposición, el evento vacío debe tener la menor ponderación, pues este es el que *nunca se realiza*. Si decidimos que la ponderación de Ω debe ser 1 y que la de Ω debe ser 0, se tendrá, en forma natural, que para cualquier evento A, la ponderación debe ser un número entre 0 y 1, y más seguridad existirá que ocurra este evento cuando la ponderación esté más cercana a 1.

La ***probabilidad*** *de que ocurra un evento A se denota con P*(A) *y se define como el número que cumple con los siguientes axiomas de Andrei Kolmogorov.*

1. Es un número no negativo: $P(A) \geq 0$.
2. La probabilidad del espacio muestral es 1: $P(\Omega) = 1$. Esto indica que la probabilidad de un evento que ocurrirá con certeza es igual a 1.
3. Si A_1 y A_2 son eventos que no tienen elementos en común (mutuamente excluyentes), entonces, la probabilidad de la unión de estos es la suma de las probabilidades de cada uno de ellos.

   Simbólicamente: $P(A_1 \cup B_2) = P(A_1) + P(B_2)$, si A y B no tienen elementos en común.

Las siguientes propiedades se deducen de los axiomas.

4. *La probabilidad del evento vacío es igual a* 0. Esto indica que la probabilidad de un suceso imposible es 0.
5. *La probabilidad de cualquier suceso es un número entre* 0 y 1.
6. *La probabilidad del complemento de un evento A es igual a* $1 - P(A)$. Esto indica que la probabilidad de que un evento A no ocurra es igual a 1 menos la probabilidad de que el evento ocurra.
7. *En general, la probabilidad de la unión de dos eventos A y B es:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Asignación de probabilidades

Al lanzar una moneda para luego anotar los resultados se tiene el espacio muestral {C, S}, donde C denota "cara" y S denota "sello". Los eventos que se pueden obtener son:

$$\{C\}, \{S\}, \{C, S\} \text{ y el conjunto vacío, } \Phi$$

Las probabilidades que podemos asignar a cada uno de estos eventos pueden ser, por ejemplo:

$$P(\{C\}) = 0.4,\ P(\{S\}) = 0.6,\ P(\{C, S\}) = 1 \text{ y } P(\Phi) = 0$$

Los valores asignados cumplen con los axiomas de la probabilidad; sin embargo, un mejor conocimiento acerca de la moneda que se usa en la experiencia podría llevar a una mejor asignación de los valores. Así por ejemplo, si se indica que la moneda es equilibrada podríamos asignar los valores de la siguiente manera:

$$P(\{C\}) = 0.5,\ P(\{S\}) = 0.5$$

Los axiomas y propiedades relativas a la probabilidad no indican cómo deben asignarse las probabilidades a los diversos resultados de una experiencia aleatoria; solamente establecen las limitaciones de las formas en que esto puede hacerse. Algunas veces la adecuada asignación antelada es posible, como en el caso del lanzamiento de una moneda equilibrada, en donde a cada resultado se le asignó 0.5 como valor de probabilidad. Otras veces, sin embargo, tal asignación previa es imposible, por lo que será

necesario lanzar la moneda varias veces, digamos 200, y asignar a cada resultado su respectiva frecuencia relativa. Estas probabilidades pueden ser mejoradas observando un mayor número de lanzamientos. En general se puede asignar probabilidades provisionales para luego "afinarlas" a partir de una mayor información de la experiencia que se estudia.

Aparte de la *asignación clásica de Laplace,* que es apropiada cuando todos los resultados experimentales tienen la misma posibilidad de aparecer, existen otras maneras de asignar probabilidad.

1. Considerando el conocimiento previo que se tiene del evento

Conociendo la composición de una población según su género, se puede indicar la probabilidad de que una persona elegida al azar sea mujer. Si el 55% de los habitantes de una ciudad son mujeres, entonces la probabilidad de que al elegir, al azar, a una persona de la ciudad, se puede aproximar con 0.55.

Conociendo que un dado es equilibrado, se puede asignar 1/6 como probabilidad a cada uno de los resultados que se obtienen al lanzarlo.

Si en una ciudad de 950,000 habitantes, según los datos del censo, la distribución de la renta anual, en dólares, es como sigue:

TABLA 5.1 *Frecuencia*

Renta anual	*Frecuencia*
[0 20,000]	800,000
]20,000 40,000]	100,000
]40,000 50,000]	50,000

entonces, al evento que indica que una persona tiene una renta de entre 20,000 y 40,000 dólares se le puede asignar la probabilidad 100,000/950,000 = 0.1052.

2. Observando el valor hacia el cual se acerca la frecuencia relativa del evento en una serie prolongada de experimentos repetidos y bajo las mismas condiciones.

Esta manera de asignar probabilidades está basada en el principio estudiado por Jacques Bernoulli (1654–1705) y publicado posteriormente con el nombre de *la ley débil de los grandes números* (también se conoce como *teorema de Bernuolli*).

Si después de realizar n veces un experimento aleatorio se observa que el evento A sucedió n_A veces, entonces la probabilidad de A se aproxima con n_A/n. A esta probabilidad se le llama *probabilidad frecuencial.*

Esta asignación es la más indicada cuando se trata de estimar la proporción de veces que se presentará cada resultado de una experiencia.

EJEMPLO. *Asignación de probabilidad*

Asignar una probabilidad al evento descrito por "una persona compra un determinado producto".

Solución

Al seleccionar una persona al azar hay dos posibilidades: que la persona compre el producto o que la persona no compre el producto.

La probabilidad de que compre el producto se puede asignar tomando una muestra al azar de *n* personas y usando la proporción de personas en la muestra que comprarán el producto. Esta operación, que corresponde a la asignación por frecuencias relativas, es cada vez mejor, a medida que la muestra es más grande.

EJEMPLO. *Asignación de probabilidad*

Para analizar la tenencia de aparatos de televisión y teléfono en una población se realizó una encuesta en un grupo de 800 personas tomadas al azar. La distribución conjunta de los resultados fue como aparece en la Tabla 5.2.

TABLA 5.2 *Distribución conjunta*

TV-Teléfono	*Sí tiene*	*No tiene*	*Total*
Sí tiene	280	120	400
No tiene	150	250	400
Total	430	370	800

Si la muestra es "representativa" de la población, y de acuerdo a los resultados, se puede asignar probabilidades a los siguientes eventos en toda la población:

A: Tener TV

B: Tener teléfono

$\overline{B}$ *:* No tener teléfono

$A \cap B$*:* Tener TV y teléfono

$A \cup B$*:* Tener TV o teléfono

de la siguiente manera:

$P(A) = 400/800 = 0.5000 \quad P(B) = 430/800 = 0.5375$
$P(\overline{B}) = 1 - P(B) = 370/800 = 0.4625 \quad P(A \cap B) = 280/800 = 0.3500$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{400}{800} + \frac{430}{800} - \frac{280}{800} = \frac{550}{800} = 0.6875$$

3. De manera subjetiva

Esta asignación se basa en las creencias de la persona que realiza la experiencia. Esta asignación generalmente toma en cuenta cualquier evidencia que se tenga a mano, así como el sentir acerca de la situación que se estudia. Se aplica cuando se tienen pocos datos y los resultados no son igualmente probables. La probabilidad de tener éxito en la venta de un libro puede ser 0.6 para el autor y 0.3 para el editor. La asignación del autor puede basarse en su optimismo y en su autoestima, mientras que la del editor, en su sentido empresarial.

5.3 Probabilidad condicional y eventos independientes

Probabilidad condicional

Para el cálculo de la probabilidad de un evento *A* no se asumen condiciones especiales aparte de las que definen el experimento. Sin embargo, se puede calcular la probabilidad de un evento a la luz de la información que brinda la realización de otro evento *B* ya llevado a cabo. Así, la probabilidad de que una persona compre un artículo puede ser calculada a partir de la información adicional de que la persona escuchó por radio un aviso relacionado con el artículo. De este modo, el espacio muestral queda reducido al conjunto de las personas que escucharon el aviso. La probabilidad de que una persona compre el artículo, sabiendo que escuchó el aviso por radio, es una *probabilidad condicional.*

La probabilidad del evento A, sabiendo que el evento B ha sucedido, se llama ***probabilidad condicional de A dado B****, se denota con P(A|B) y se define como:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad si\ P(B) \neq 0$$

Cuando $P(B) = 0$ *se define* $P(A|B) = 0$.

EJEMPLO. *Probabilidad conjunta y probabilidad marginal*

Entre las preguntas que se realizaron a 500 personas de un área metropolitana para estudiar el efecto de la propaganda en el deseo de compra de un nuevo producto estaban las siguientes: ¿desearía usted comprar el producto? ¿Cuál fue el medio de propaganda por el cual usted se enteró de la existencia del producto?

La distribución de frecuencias de las respuestas fue registrada en la Tabla 5.3.

TABLA 5.3 *Distribución conjunta*

Deseo de compra	Medio de propaganda			
	TV	Radio	Periódico	Total
Sí	150	80	50	280
No	40	100	80	220
Total	190	180	130	500

Utilizando la información que aparece en la Tabla 5.3 y escribiendo:

A: persona vio la propaganda en la TV,
B: persona manifiesta que comparará el producto,

se puede asignar las siguientes probabilidades en toda la población:

$P(A) = 190/500 = 0.38$

$P(A \cap B) = 150/500 = 0.30$

$P(B|A) = \dfrac{150/500}{190/500} = 0.7894$

siempre que la muestra sea representativa.

La probabilidad condicional $P(B|A)$ equivale a calcular la probabilidad del evento "tener deseo de compra", restringido al espacio muestral descrito por "vio la propaganda por TV".

En este contexto, a la probabilidad de ocurrencia de dos o más eventos a la vez, como el referido por "ver la propaganda por TV y manifiesta que comprará el producto", se le llama *probabilidad conjunta*, y a la probabilidad de un evento simple, como el referido por "ver la propaganda por TV", se le llama *probabilidad marginal.*

Eventos independientes

Si la probabilidad del evento *A* no depende de la realización del evento *B* se dice que *A es independiente de B,* y en tal caso se puede escribir que la probabilidad condicional de *A* dado *B* es igual a la probabilidad de *A.*

Cuando los eventos *A* y *B* son independientes, la probabilidad de que ocurran los eventos *A* y *B* de manera simultánea se calcula simplemente con la siguiente *regla de multiplicación:*

$$P(A \cap B) = P(A)P(B)$$

EJEMPLO. *Eventos independientes*

Los resultados que aparecen en la Tabla 5.3 sugieren que el deseo de compra no es independiente del tipo de propaganda, pues $P(\textit{Deseo de compra} \mid \text{TV}) \neq P(\textit{Deseo de compra})$, y así para los otros medios.

EJEMPLO. *Sueldo y estatura*

Es razonable considerar, por ejemplo, que el sueldo de una persona es independiente de su estatura. Por ello, si el 10% de las personas gana más de 1,000 dólares y el 20% mide más de 1.65 m, y si se consideran los eventos:

A: La persona gana más de 1,000 dólares y
B: La persona mide más de 1.65 m,

entonces, la probabilidad del evento $A \cap B$ se puede considerar que es igual a:

$$P(A)P(B) = (0.10)(0.20) = 0.02$$

Nota

En general, la probabilidad $P(A \cap B)$ de que ocurran los eventos A y B simultáneamente se calcula con:

$$P(A|B)P(B) \text{ si } P(B) \neq 0 \quad \text{o con} \quad P(B|A)P(BA) \text{ si } P(A) \neq 0$$

EJEMPLO. *Sistemas en serie*

Un sistema de dos componentes, C_1 y C_2, se dice que *está en serie* si las componentes están conectadas de la siguiente manera:

FIGURA 5.2 *Componentes en serie*

Si se tratara de una conexión eléctrica y cada componente es una resistencia, el paso de corriente eléctrica desde el punto A hasta el punto B sucede siempre que ambas componentes funcionen. Cuando las componentes funcionan de manera independiente, la probabilidad R de que pase corriente de A a B es igual a la probabilidad de que funcione la componente C_1 por la probabilidad de que funcione la componente C_2. Esto es:

$$R = P(C_1)P(C_2)$$

EJEMPLO. *Proceso de fabricación en serie*

La empresa que administra Armando Paredes fabrica para la venta un producto alimenticio utilizando dos máquinas A y B, las que funcionan en serie. En una primera etapa de la fabricación se utiliza la máquina A, para luego pasar en una segunda etapa a usar de manera independiente la máquina B. Finalizada esta segunda etapa el producto se considera terminado.

Armando Paredes considera que, por razones obvias, es importante hacer un mantenimiento a todo el sistema, pero duda en llevarlo a cabo después de las 1,000 horas de funcionamiento.

Si la probabilidad de que una máquina de tipo A funcione después de 1,000 horas es 0.9 y la probabilidad de que una máquina de tipo B funcione más de 1,000 horas es 0.95, y como las máquinas están conectadas en serie, entonces, la probabilidad de que todo el sistema funcione correctamente después de las 1,000 horas de uso es $P = (0.90)(0.95) = 0.855$.

Ahora que Armando Paredes conoce la probabilidad de buen funcionamiento después de 1,000 horas de uso, podrá tomar la decisión de llevar a cabo el mantenimiento del sistema o no.

APLICACIÓN: Evaluación de riesgos. Préstamos bancarios

Los resultados de este ejemplo pueden situarse en otro contexto, por ejemplo, en el campo de la medicina; sin embargo, el desarrollo se limita esta vez a evaluar la validez de un modelo para predecir si un cierto individuo que ha recibido un préstamo bancario será moroso o no.

Se consideran N personas, de las cuales, en la realidad, s son morosas y e no lo son. Se supone que al aplicar un modelo de clasificación para ver quién es moroso y quién no lo es, y al confrontarlo con la realidad, resulta lo siguiente:

TABLA 5.4 *La realidad vs. la predicción*

Predicción	***Lo real***		
	Moroso	***No moroso***	***Total***
Moroso	*a*	*b*	*m*
No moroso	*c*	*d*	*r*
Total	*s*	*e*	*N*

En la Tabla 5.4, a indica el número de personas morosas para las cuales el modelo pronosticó que eran morosas, etc. Nótese que si el modelo de predicción funciona perfectamente, es decir, si con él se lograran predicciones siempre correctas, se tendrá que $c = b = 0$.

Para evaluar la validez del modelo se definen los siguientes *índices de validez*:

La *tasa de falsos positivos PFP,* que es igual a la probabilidad de que el modelo indique que la persona es morosa cuando realmente no lo es. Se calcula con:

$$PFP = b/e$$

La *tasa de falsos negativos PFN,* que es igual a la probabilidad de que el modelo indique que la persona no es morosa cuando realmente lo es**.** Este índice se calcula con:

$$PFN = c/s.$$

La *especificidad* (*Esp*), que es igual a la probabilidad de que el modelo indique que la persona es morosa cuando realmente lo es. Se calcula con *a/s.*

La *sensibilidad* (*Sens*), que es la probabilidad de que el modelo indique que la persona no es morosa cuando realmente no lo es. Se calcula con *d/e*.

¿Cómo deben ser estas probabilidades para tener un buen modelo?

APLICACIÓN: El caso de la editora La Luz. Evaluación de la eficacia de los avisos publicitarios

La empresa editora La Luz, para mejorar sus ventas, se ha visto precisada a emitir una serie de avisos publicitarios en los medios periodísticos. Después de cierto tiempo de realizada la publicidad, la empresa desea evaluar la eficacia de la propaganda.

El problema que se presenta es cómo medir la eficacia de la publicidad.

Los expertos indican que una manera de medir la eficacia de la propaganda es comparar el cociente de la probabilidad de compra habiendo leído los avisos y la probabilidad de no compra habiendo leído los avisos con el cociente de la probabilidad de compra y la probabilidad de no compra. Si el primer cociente es mayor que el segundo, se puede concluir que la propaganda ha sido efectiva. Es decir, si se indica con:

L: Leyeron el aviso.

NL: No leyeron el aviso.

C: Compraron.

NC: No compraron.

se debe cumplir la siguiente relación: $\frac{P(C|L)}{P(NC|L)} > \frac{P(C)}{P(NC)}$

El departamento de marketing de la editora llevó a cabo una encuesta entre 900 personas para indagar sobre los eventos *L, NL, C y NC.* Los datos recolectados aparecen en la Tabla 5.5:

TABLA 5.5 *Resultados de la encuesta*

	L	*NL*
C	400	300
NC	50	150

Esto permitió decir que los avisos que emitió La Luz fueron efectivos.

5.4 El teorema de la probabilidad total y el teorema de Bayes

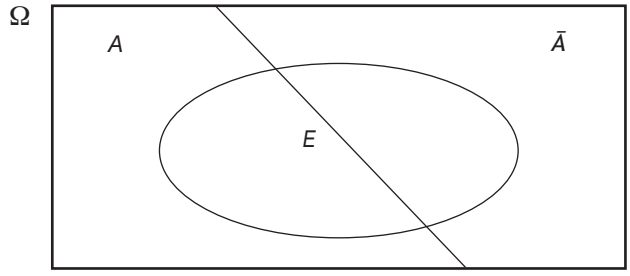
Teorema de la probabilidad total

El teorema de *la probabilidad total* muestra cómo calcular la probabilidad de un evento en función de las probabilidades de otros eventos que ocurren en condiciones mutuamente excluyentes y que influyen en la realización o no del evento.

La probabilidad de tener éxito (E) en un negocio puede conocerse si se tiene información acerca de la apertura (A) o no ($\bar{A}$) de un negocio similar. El cálculo de la probabilidad es como sigue:

$$P(E) = P(E \cap A) + P(E \cap \bar{A}) = P(E|A)P(A) + P(E|\bar{A})P(\bar{A})$$

FIGURA 5.3 *El espacio muestral es la unión de A y su complemento*

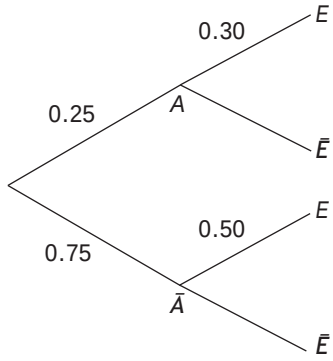


Si la probabilidad de tener éxito en el negocio cuando se abre otro similar es $P(E|A) = 0.30$ y la probabilidad de tener éxito en el negocio cuando no se abre otro similar es $P(E|\bar{A}) = 0.50$, y si se sabe que la probabilidad de que se abra un negocio similar es $P(A) = 0.25$, entonces la probabilidad de tener éxito en el negocio es:

$$P(E) = (0.30)(0.25) + (0.50)(0.75) = 0.45$$

La información suele presentarse usando un *árbol de probabilidades* como el de la Figura 5.4. Cada rama representa un resultado parcial de la experiencia y cada *nudo terminal*, un punto del espacio muestral. Sobre algunas ramas del árbol se ha escrito la probabilidad condicional del resultado que representa dada la experiencia anterior. La probabilidad de cada nudo terminal se obtiene multiplicando los números escritos sobre las ramas que conducen a él.

FIGURA 5.4 *Los nodos terminales que indican éxito aparecen con E*



$$P(E) = P(E|A)P(A) + P(E|\bar{A})P(\bar{A}) =$$
$$(0.30)(0.25) + (0.50)(0.75)$$

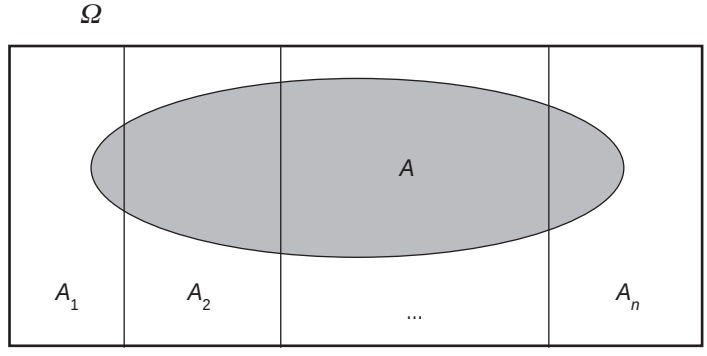
La probabilidad de tener éxito en el negocio se halla multiplicando las probabilidades que están sobre las ramas que van del inicio hasta el terminal *E* (éxito) y luego sumando los productos.

El resultado anterior es un caso particular del siguiente teorema, llamado *teorema de la probabilidad total.*

Si $A_1,\ldots, A_n$ son eventos disjuntos cuya reunión es igual al espacio muestral Ω, entonces se cumple que para cualquier evento A:

$$P(A) = P(A \cap A_1) + P(A \cap A_2) + \ldots + P(A \cap A_n) =$$
$$P(A|A_1)P(A_1) + P(A|A_2)\,P(A_2)\ldots + P(A|A_n)P(A_n)$$

FIGURA 5.5 *A y la partición del espacio muestral*



APLICACIÓN: El caso de los empleados que "chatean" en horas de trabajo

El uso de los computadores y de Internet se incrementa cada día en las empresas. La utilización de estos medios en el análisis de la información para el mejor conocimiento de los clientes, la predicción de su perfomance, la comunicación con sus clientes y en general el tratamiento inteligente de sus negocios le otorgan a la empresa gran ventaja competitiva. Sin embargo, también se presentan diversos problemas por el mal uso de estos adelantos, como la utilización del chat en horas de oficina, causando en muchos casos pérdidas a la empresa.

La asociación de empresas de servicios se ha propuesto tratar de aportar en la solución del problema del "chateo" y considera, para iniciar el estudio, realizar una encuesta para conocer el porcentaje de empleados que utilizan el "chat" en horas de oficina en la empresa en donde trabajan. La pregunta a plantear en la encuesta es "¿chatea usted en el lugar de su trabajo?". La empresa contratada para la realización de la encuesta considera que, por la naturaleza de la pregunta, es posible que las personas que responden se sientan incómodas ante esta pregunta y no den la respuesta correcta, sino que mientan. Por ello ha decidido utilizar un procedimiento de consulta que no incomode a la persona a la cual se le pregunta.

Este procedimiento contiene un mecanismo de aleatorización, como el lanzamiento de una moneda equilibrada, y consiste en plantear dos preguntas: la pregunta delicada (¿chatea en su trabajo?) y otra que no cause ningún problema en contestar y cuya probabilidad de la respuesta se conozca (por ejemplo, "¿el último número de su DNI es par?"). Se pide al entrevistado que lance la moneda y que conteste la pregunta delicada si sale cara y que conteste la pregunta inocua si sale sello. El entrevistador no sabrá cuál es la pregunta que se contestó, y de este modo el entrevistado no se sentirá incómodo al responder. El procedimiento se llama *respuesta aleatorizada*.

Supongamos que al usar este mecanismo, de 500 entrevistados, 200 respondieron "sí". Entonces, usando las probabilidades totales, se tendrá que

$$P(\textit{responder sí}) = P(\textit{responder sí}|D)P(D) + P(\textit{responder sí}|I)P(I)$$

en donde D es el evento "respondió la pregunta difícil" e I es el evento "respondió la pregunta inocua".

Despejando la probabilidad de responder la pregunta difícil, $P(\textit{responder sí}|D)$, se tiene:

$$P(\textit{responder sí}|D) = \frac{P(\textit{responder sí}) - P(\textit{responder sí}|I)P(I)}{P(D)}$$

Reemplazando los valores conocidos se tiene que:

$$P(\textit{responder sí}|D) = \frac{(200/500) - (0.5)(0.5)}{0.5} = 0.30$$

Usando este resultado, la encuestadora pudo indicar que un estimador del porcentaje de los que chatean en horas de trabajo es 30%.

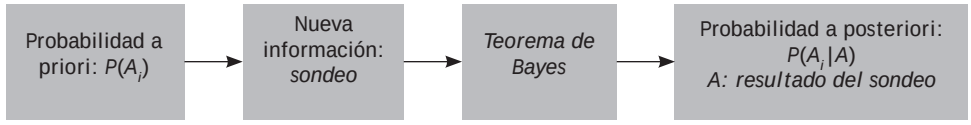
El teorema de Bayes

Con el teorema de Bayes[2] se obtiene una metodología que permite mejorar el conocimiento previo que el investigador tiene acerca de un evento A_i, con la incorporación de información adicional A de carácter aleatoria. Aplicando este procedimiento, el investigador hará decisiones más correctas y tendrá mejores resultados.

En este contexto, la información previa acerca de A_i se sintetiza con su probabilidad, ($P(A_i)$), la que se llama *probabilidad a priori de* A_i, y de lo que se trata es de conocer la probabilidad condicional de A_i conocida la información adicional A. La información rectificada de A_i obtenida a la luz de A, que finalmente se obtiene, se llama *probabilidad a posteriori de* A_i.

Un industrial ha creado un producto y piensa, a priori, que con este acaparará el 15% del mercado; sin embargo, una encuesta tomada a un conjunto de clientes potenciales elegidos al azar le indica que de 10 clientes solo 1 tiene la intención de comprar el producto. A la luz de esta nueva información, el industrial revisará la probabilidad de adquisición del producto que le ayudará a tomar nuevas decisiones.

FIGURA 5.6 *Proceso de Bayes*



Enunciado del teorema de Bayes

La probabilidad a posteriori de A_i, *a la luz de la información A, se calcula con:*

$$P(A_i|A) = \frac{P(A|A_i)\,P(A_i)}{P(A)} = \frac{P(A|A_i)\,P(A_i)}{\sum_{k=1}^{n} P(A|A_k)P(A_k)} \text{ , para } i = 1,\ldots, n$$

Se suponen las mismas condiciones indicadas en el teorema de la probabilidad total.

Según esta igualdad, la probabilidad de A_i se revisa a la luz del conocimiento del evento A utilizando como insumo la probabilidad previa de A_i que a priori se tenía.

[2] El teorema de Bayes se le atribuye al reverendo Thomas Bayes (1702-1761), de origen inglés.

EJEMPLO. *Artículos defectuosos*

Basado en su experiencia, el encargado del almacén de una planta de ensamblaje de computadoras ha indicado que en un lote de chips, el 30% pueden venir de la planta A, el 50% pueden venir de la planta B y el resto, de la planta C. Por otro lado se conoce que:

El 1% de los chips que provienen de la planta A son defectuosos.

El 2% de los chips que provienen de la planta B son defectuosos.

El 3% de los chips que provienen de la planta C son defectuosos.

Al seleccionar un chip del lote para ser utilizado en un ensamblaje, este resultó defectuoso. A la luz de este hecho, se desea revisar la información dada por el encargado del almacén.

Para facilitar el cálculo se usarán las siguientes notaciones:

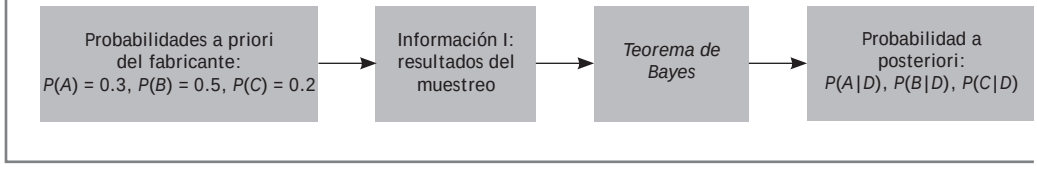
D = el chip utilizado es defectuoso.

A = el chip proviene de A.

B = el chip proviene de B.

C = el chip proviene de C.

FIGURA 5.7 *Proceso de Bayes*



Utilizando el teorema de Bayes y a la luz del hecho de que un chip seleccionado resultó defectuoso, se tiene que las probabilidades revisadas son:

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{(0.01)(0.30)}{0.019} = 0.1579$$

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{(0.02)(0.50)}{0.019} = 0.5263$$

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} = \frac{(0.03)(0.20)}{0.019} = 0.3158$$

(La probabilidad de que un chip seleccionado del lote sea defectuoso (*D*) se calculó con:

$$P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) =$$
$$(0.01)(0.30) + (0.02)(0.50) + (0.03)(0.20) = 0.019)$$

Al parecer, y de acuerdo a las probabilidades obtenidas a posteriori, las probabilidades, que indicó el encargado del almacén, no son las adecuadas.

EJEMPLO. *La prueba del alcoholímetro*

Un laboratorio somete a los choferes que sufren accidentes de tránsito a la prueba del alcoholímetro. ¿Cuál es la probabilidad de que un chofer de esta población esté ebrio, dado que el resultado del alcoholímetro fue positivo?, si se ha determinado que:

– Cuando un chofer está ebrio, la prueba proporciona resultado positivo en el 95% de los casos.

– Cuando un chofer no está ebrio, la prueba proporciona resultado negativo en el 94% de los casos.

– El 2% de los conductores que sufren accidentes manejan ebrios.

Solución

El espacio muestral está formado por todos los choferes que sufren accidentes.

Sean los eventos *E*, descritos por "el chofer está ebrio", y *T*, descritos por "el test es positivo".

Por el teorema de Bayes se tiene que a partir de la información adicional *T*, la probabilidad corregida de *E* es:

$$P(E|T) = \frac{P(T|E)P(E)}{P(T|E)P(E) + P(T|\bar{E})P(\bar{E})} = \frac{(0.95)(0.02)}{(0.95)(0.02) + (1 - 0.94)(0.98)} = 0.2442$$

La prueba indica que solo el 24.42% de los choferes cuya prueba es positiva están ebrios, o de manera equivalente, el 75.6% de los resultados positivos de la prueba corresponden a choferes que no están ebrios.

APLICACIÓN: El clasificador de Bayes

Una de las principales preocupaciones de las entidades financieras es la recuperación de los préstamos que hacen a sus clientes. Ello ha motivado que se utilicen modelos de clasificación que permiten la mejor elección de las personas que deban beneficiarse de los préstamos de la entidad. Uno de los modelos de clasificación muy utilizado, por su fácil aplicación, es el llamado clasificador de Bayes. Este modelo se basa en el teorema de Bayes y utiliza predictores (variables que sirven para predecir valores de una variable dependiente, que en este caos será denotada con *Y*) que son categóricos. El modelo supone que los predictores son independientes en cada una de las clases de *Y*, y funciona bien cuando se tiene una base de datos muy grande; sin embargo, con la finalidad de ilustrar la manera como se aplica, usaremos una base de datos ficticia y pequeña, de 10 clientes, del banco Biscao, también ficticio. En esta base, las variables "sexo" y "sueldo" son las variables predictoras, mientras que la variable *Y* es la variable dependiente que indica si el cliente es buen pagador (1) o si es mal pagador (0). La variable "sueldo" ha sido categorizada en las categorías: alto (A), bajo (B) e intermedio (C), para así aplicar el método.

Cliente	*Sexo*	*Edad*	*Y*
1	M	C	1
2	M	C	1
3	M	B	0
4	H	A	0
5	H	A	0
6	M	B	1
7	H	C	0
8	H	C	0
9	H	C	0
10	H	C	0

El objetivo del clasificador de Bayes es la estimación de la probabilidad que tiene cada cliente de pertenecer a cada una de las clases de la variable dependiente, a la luz de los valores de los predictores. Luego se clasifica a cada cliente en la clase de *Y* que tiene la mayor probabilidad estimada.

A continuación, el cálculo de la estimación de la probabilidad que tiene el cliente 1 de pertenecer a cada una de las clases de *Y*, conociendo los valores de los predictores:

$P(Y = 1|Sexo = M \text{ y } Sueldo = C)$
$P(Sexo = M \text{ y } Sueldo = C|Y = 1)P(Y = 1)/P(Sexo = M \text{ y } Sueldo = C) = (2/3)(3/10)/(2/10) = 1$
$P(Y = 0|Sexo = M \text{ y } Sueldo = C) = P(Sexo = M \text{ y } Sueldo = C|Y = 0)P(Y = 0)|P(Sexo = M \text{ y } Sueldo = C) = 0$
Como la mayor probabilidad estimada corresponde a la clase 1 de *Y*, el cliente es ubicado en esta categoría.

Análogamente para los otros clientes.

Habiendo sido "entrenado" el modelo, este podrá ser usado para predecir a qué clase pertenecen los nuevos clientes que solicitan un préstamo a la entidad. Ello ayudará a tomar la decisión de darle o no el préstamo solicitado.

Cabe indicar que estos modelos son evaluados antes de ser aplicados. Una manera de hacerlo es comparar las predicciones con los valores de *Y* que se tienen en la base datos antes de la predicción.

APLICACIÓN: El caso de la prueba de Elisa

Al igual que para evaluar modelos de riesgos, en medicina es muy importante evaluar la validez de las pruebas "diagnósticas". Estas pruebas permiten determinar si un individuo tiene o no una cierta enfermedad. Una de estas pruebas es la prueba de Elisa, que detecta si una persona tiene o no el virus del VIH.

El Ministerio de la Salud está evaluando la validez de la prueba de Elisa que recientemente se aplicó a 500 personas, de las cuales 20 tienen realmente el virus VIH. Los resultados obtenidos aparecen en la Tabla 5.6.

TABLA 5.6 *Realidad vs. predicción*

Estado real	*Resultado de la prueba*		
	Negativo	*Positivo*	*Total*
No tiene el virus	477	3	480
Tiene el virus	2	18	20
Total	479	21	500

Según estos resultados, la prueba indicó, erróneamente, que de las 480 personas que no tienen el virus, 3 tienen el virus, etc.

Si $R+$ representa al evento descrito por "el resultado de la prueba es positivo" y V representa al evento descrito por "la persona tiene el virus", entonces la especificidad (*Espc*) de la prueba es la probabilidad $P(R+|V)$ de que la prueba resulte positiva cuando la persona tiene el virus. Esta probabilidad es 18/20 = 0.9000.

La sensibilidad (*Sens*) de la prueba es la probabilidad $P(\overline{R}+|V)$ de que la prueba resulte negativa cuando la persona no tiene el virus. Esta probabilidad es 477/480 = 0.99375.

La tasa de falsos positivos PFP y la tasa de falsos negativos PFN se calculan de la siguiente manera:

$$\text{PFP} = 1 - \textit{Sens} = 0.00625$$
$$\text{PFN} = 1 - \textit{Espc} = 0.1000$$

Otra manera de evaluar la prueba es con las probabilidades que a continuación se calculan y que se llaman valores predictivos.

- La probabilidad de que la persona tenga el virus (V) dado que la prueba fue positiva ($R+$), que se calcula usando también el teorema de Bayes, con:

$$P(V|R+) = \frac{P(R+|V).P(V)}{P(R+|V).P(V) + P(R+|\overline{V}).P(\overline{V})} = \frac{\textit{Espc}.P(V)}{\textit{Espc}.P(V) + (1 - \textit{Sens})(1 - P(V))}$$

- La probabilidad de que la persona no tenga el virus, dado que la prueba fue negativa ($\overline{R}$), que se calcula usando también el teorema de Bayes, con:

$$P(\overline{V}|\overline{R}+) = \frac{\textit{Sens}.(1 - P(V))}{\textit{Sens}.(1 - P(V)) + (1 - \textit{Espc}.(1 - P(V))}$$

Nótese que los valores predictivos se pueden calcular si se conoce la probabilidad $P(V)$ de que una persona tenga el virus. A esta probabilidad se le llama prevalencia de tener el virus.

Si, por ejemplo, los reportes indican que el 2% de la población tiene el virus (la prevalencia es 2%), el Ministerio de la Salud podrá calcular los valores predictivos de la prueba de Elisa.

EJERCICIOS

1. Indicar el espacio muestral de los siguientes experimentos aleatorios.
   a) E_1: Elección de tres artículos producidos por una empresa, de uno en uno y sin restitución, para luego anotar si cada uno de ellos es defectuoso, (D), o no (O).
   b) E_2: Elección de un automóvil producido por una fábrica para luego anotar el espacio recorrido después de consumir un galón de gasolina.

2. Un experimento consiste en seleccionar al azar cuatro empresas que operan en el país y observar si son empresas industriales.
   a) Indicar el espacio muestral.
   b) Enumerar los elementos de los eventos que se describen a continuación:
      A: por lo menos tres de las cuatro empresas seleccionadas son industriales.
      B: a lo más dos de las cuatro de las empresas seleccionadas son industriales.

3. El señor Pérez debe pasar por tres entrevistas consecutivas para ingresar a trabajar en una empresa. Las personas encargadas de las entrevistas son: Hugo, Paco y Luis, en ese orden. Sean los eventos descritos por las proposiciones que se indican:
   A: el veredicto de Hugo es favorable.
   B: el veredicto de Paco es favorable.
   C: el veredicto de Luis es favorable.
   Usando *A*, *B* y *C*, escribir los eventos descritos por:
   a) Ninguno de los veredictos es favorable.
   b) Todos los veredictos son favorables.
   c) Por lo menos dos veredictos son favorables.
   d) El veredicto de Hugo es favorable.

4. Se tienen dos computadoras de tipo A y dos de tipo B. Si se eligen al azar y de una sola vez tres computadoras:
   a) ¿Cuál es el número de elementos que tiene el espacio muestral?
   b) ¿Cuál es el número de elementos que tiene el evento cuyos elementos están formados por dos computadoras de tipo A y una de tipo B?

5. Indicar el procedimiento a seguir para asignar probabilidades a los eventos representados por las siguientes expresiones:
   a) Mañana bajará el dólar.
   b) Existe en el país una persona que recorre 100 m en menos de 8 segundos.
   c) El próximo mes las ventas de la bebida gaseosa A serán mayores que las ventas de la bebida gaseosa B.
   d) Un cliente de la tienda Seas tiene cuenta corriente en el banco Bilbá.
   e) El próximo lunes, las acciones de la minera Patocha tendrán mayor precio.
   f) De cinco personas que entren en el centro comercial, tres comprarán algún producto.

6. Una computadora puede quedar fuera de servicio por acción de un virus, por falla en el hardware o por ambas causas a la vez. Para dos computadoras se tiene lo siguiente:

   Para la computadora 1: $P(virus) = 0.07$, $P(falla\ en\ el\ hardware) = 0.10$ y $P(ambas\ causas) = 0.06$.

   Para la computadora 2: $P(virus) = 0.09$, $P(falla\ en\ el\ hardware) = 0.12$ y $P(ambas\ causas) = 0.06$

   ¿Cuál computadora tiene la mayor probabilidad de quedar fuera de servicio?

7. La probabilidad de que Juan vaya a una determinada cita es 0.4, de que Pedro vaya a la misma cita, 0.6, y de que ambos vayan a la cita, 0.2. ¿Cuál es la probabilidad de que Juan o Pedro vayan a la cita?

8. Una encuesta para conocer los hábitos de consumo de comidas rápidas en relación con la edad fue realizada en un grupo de 1,000 personas. Los resultados indican que de las 300 personas que tienen entre 20 y 30 años, 250 gustan de las comidas rápidas, mientras que de las personas mayores de 30 años, 300 prefieren las comidas rápidas. Asignando probabilidades adecuadas, estimar:

   a) La probabilidad de que una persona, elegida al azar, prefiera las comidas rápidas.

   b) La probabilidad de que una persona sea menor de 30 años y que no prefiera las comidas rápidas.

9. Un alumno de la universidad UU debe llevar en el segundo ciclo de estudios los cursos de Filosofía, Matemáticas y Lengua. Si la probabilidad de aprobar el curso de Filosofía es 0.65; el de Lengua, 0.55; el de Matemáticas, 0.5; el de Filosofía y Matemáticas, 0.3; el de Filosofía y Lengua, 0.35; el de Matemáticas y Lengua, 0.3; y los tres a la vez, 0.2; calcular:

   a) La probabilidad de aprobar por lo menos dos cursos.

   b) La probabilidad de aprobar por lo menos un curso.

   c) La probabilidad de no aprobar curso alguno.

10. Una caja contiene 100 vacunas. La probabilidad de que al menos una no sea efectiva es 0.05, y de que haya al menos dos no efectivas es 0.01. ¿Cuál es la probabilidad de que la caja contenga

    a) todas las vacunas efectivas?

    b) exactamente una vacuna no efectiva?

    c) a lo más una vacuna no efectiva?

11. Si la probabilidad de un evento A es p, entonces se define "el chance de que ocurra A" como la razón de p a $1 - p$. A menudo el chance se expresa como cociente de dos factores que no tienen un factor común, y si es más probable que no ocurra un evento, se acostumbra dar el chance de que no ocurra en lugar de que sí ocurra. ¿Cuál es el chance a favor o en contra de la ocurrencia de un evento si su probabilidad es a) 3/8, b) 0.07 y c) 0.4?

12. Dadas las probabilidades $P(A) = 0.5$, $P(B) = 0.7$ y $P(A \cap B) = 0.15$:

    a) Decir si los eventos A y B son independientes

    b) Hallar $P(A|B)$

    c) Hallar $P(B|A)$.

    d) ¿Es $P(A|B) = P(B|A)$?

13. Si A y B son dos eventos mutuamente excluyentes con $P(A) = 0.5$ y $P(B) = 0.7$:
    a) Hallar $P(A|B)$.
    b) Decir si A y B son independientes.

14. Se ha determinado que las probabilidades de que un televidente vea los programas A o B son: 0.5 y 0.4, respectivamente. Si se asume que cada persona ve los programas independientemente uno del otro, ¿cuál es la probabilidad de que un televidente vea por lo menos uno de los programas?

15. Para que un postulante sea admitido a una escuela de negocios debe pasar con éxito al menos dos exámenes consecutivos de los tres a que es sometido en forma alternada y ante dos personas A y B. Se supone que los exámenes son independientes. Por experiencia se sabe que el 40% de los postulantes aprueban el examen con A, mientras que solo el 35% aprueban el examen con B. Si a cada postulante se le permite escoger a la persona con quien iniciar los exámenes, ¿qué recomendar al postulante, iniciar con A o con B?

16. Se está estudiando la relación del uso de los teléfonos móviles y la edad. Se ha determinado que el 30% de las personas tienen edades entre los 20 y 30 años, de los que usan teléfono móvil el 20% tienen entre 20 y 30 años, y se sabe además que el 70% de las personas usan el teléfono móvil. Hallar la probabilidad de que una persona que tiene entre 20 y 30 años use el teléfono móvil. Si se desea hacer propaganda sobre el uso de los celulares, ¿será necesario tomar en cuenta que una persona tiene entre 20 y 30 años?

17. Cada vez que un vendedor ofrece uno de sus artículos para la venta, la probabilidad de que lo compren es 0.2. Hallar la probabilidad de que el vendedor realice:
    a) Ninguna venta en cuatro ofrecimientos.
    b) Cuatro ventas en cuatro ofrecimientos.
    c) Tres ventas en cuatro ofrecimientos.

18. Para analizar la relación que podría existir entre la tenencia de tarjetas de crédito y la variable género, se aplicó una encuesta a 200 personas. Se encontraron los resultados que se muestran en la siguiente tabla.

	Con tarjeta de crédito	***Sin tarjeta de crédito***
Varones	45	85
Mujeres	30	40

    Usando la información que se muestra, elaborar una tabla de probabilidades conjuntas y luego:
    a) Escribir las probabilidades marginales correspondientes a las variables "género" y "tenencia de tarjeta de crédito".
    b) Si una persona es varón, ¿cuál es la probabilidad de que tenga tarjeta de crédito?
    c) ¿Se puede decir que la tenencia de tarjetas de crédito es independiente del hecho de ser varón o mujer?

19. La empresa de calzado Rex está planeando abrir una nueva tienda en un lugar determinado. El analista de la empresa ha establecido que la probabilidad de tener éxito en este nuevo local es 0.7 si no se abre un negocio similar en las cercanías del lugar; pero si esto ocurre, la probabilidad de éxito será solo de 0.20. ¿Cuál es la probabilidad de que la empresa Rex tenga éxito en el negocio si la probabilidad de apertura de un negocio similar es 0.40?

20. La compañía de teléfonos asegura que la fiabilidad del servicio es tal que cuando se marca correctamente el número deseado se tienen 19 chances sobre 20 de obtener tono al otro lado de la línea.

    a) ¿Cuál es la probabilidad de obtener tono al otro lado de la línea en tres intentos a lo más si se supone que cada vez se marca el número correcto?

    b) Si se estima que en el primer intento existe 1 chance sobre 10 de marcar un número equivocado (error de manipulación), para el segundo intento existe 1 chance sobre 100 de equivocación y para el tercer intento existe 1 chance sobre 1,000 de equivocación, ¿cuál es la probabilidad de obtener tono al otro lado de la línea en tres ensayos a lo más?

21. En la siguiente tabla se presenta la distribución de 125 hogares de acuerdo con los ingresos de sus jefes de familia y con el hecho de ser propietarios de teléfonos y de aparatos de televisión. A partir de la información, elaborar las probabilidades conjuntas y contestar las siguientes preguntas.

	Hogares con ingresos de $ 1,000 o menos		***Hogares con ingresos de más de $ 1,000***	
	Con teléfono	Sin teléfono	Con teléfono	Sin teléfono
Con TV	27	20	18	10
Sin TV	18	10	12	10

    a) ¿Cuál es la probabilidad de elegir un hogar con TV?

    b) Si una familia con ingresos de más de $ 1,000 tiene teléfono, ¿cuál es la probabilidad de que tenga TV?

    c) ¿Cuál es la probabilidad de elegir a una familia que tenga TV, dado el hecho de que tiene teléfono?

    d) ¿Son independientes los eventos "tener TV" y "tener teléfono"?

    e) ¿Son independientes los eventos "ingresos de menos de $ 1,000" y "ser propietario de TV"?

22. Una encuesta realizada por una tienda de artefactos para el hogar ha revelado la información que aparece en la siguiente tabla, y que se refiere a las proporciones de los clientes que compran o no algún producto y que visitan la tienda de manera frecuente o infrecuente.

Frecuencia de visita	*Compra*	*No compra*
Frecuente	0.40	0.12
Infrecuente	0.38	0.10

a) Hallar la probabilidad de que un cliente visite frecuentemente la tienda.
b) Hallar la probabilidad de que un cliente visite frecuentemente la tienda y compre un producto.
c) Hallar la probabilidad de que un cliente que visita frecuentemente la tienda compre un producto.

23. Para probar la efectividad de un modelo para detectar a los malos deudores de todos aquellos que han recibido un préstamo, un banco, usando una base de datos que agrupa a 500,000 clientes, ha obtenido los siguientes resultados:

En la base de datos se registran 130 personas que son malos pagadores.

Con el modelo se predicen correctamente 80 malos pagadores.

Con el modelo se predicen correctamente 499,850 buenos pagadores.

Hallar:

a) La "tasa falsa positiva" del modelo.
b) la "tasa falsa negativa" del modelo.
c) La especificidad del modelo.
d) La sensibilidad del modelo.

24. Una persona interviene en 100 ocasiones independientes en un negocio cuya probabilidad de fracaso es 0.3. Hallar la probabilidad de que un fracaso ocurra en una o más ocasiones.

25. La probabilidad de que una determinada acción suba de precio cada día es 0.1. Hallar la probabilidad de que la acción suba en tres días consecutivos. ¿Cuál es la probabilidad de que suba recién el tercer día, después de comprada?

26. En una urna hay dos bolas rojas y una negra. Hugo, Paco y Luis (en ese orden) deben sacar, uno después del otro, una bola sin restituirla posteriormente. ¿Cuál de las tres personas tiene mayor probabilidad de sacar la bola negra?

27. Un estadístico ha ideado un modelo para detectar los cobros fraudulentos en los seguros de salud. Según los estudios realizados, el modelo detecta el 80% de los fraudes. También indica incorrectamente defraudadores en el 5%. La información que se tiene de diferentes fuentes es que el 10% de los reclamos de seguros son fraudulentos. Hallar la probabilidad de que un reclamo que es detectado como fraudulento por el modelo sea en realidad fraudulento.

28. En el caso de aplicación de la editora La Luz, suponer que el departamento de marketing de la empresa solo informa que:

$P(L|C) = 0.60$. Es decir, el 60% de los que compraron leyeron el aviso.

$P(L|NC) = 0.30$. Es decir, el 30% de los que no compraron leyeron el aviso.

A partir de esta información, ¿podrá la editora concluir que los avisos fueron efectivos?

29. Según una encuesta realizada:
    - El 90% de los hombres (H) que tienen cáncer pulmonar (C) son fumadores (F).
    - El 70% de mujeres (M) que tiene cáncer pulmonar son fumadoras.
    - La frecuencia de cáncer pulmonar es 4×10^{-4} para los hombres y de 10^{-4} para las mujeres.
    - La proporción relativa de fumadores es 5 veces más elevada en los hombres que en las mujeres.

    ¿Se puede concluir que una mujer fumadora tiene más propensión a contraer cáncer pulmonar que un hombre fumador?

30. El 40% de las resistencias que se utilizan en una fábrica son de la marca A y el resto son de la marca B. El 1% de las resistencias de la marca A y el 2% de la marca B son defectuosas. Si de un lote de tales resistencias adquiridas por la fábrica se elige una al azar, ¿cuál es la probabilidad de que esta resulte defectuosa?

31. Un empresario tiene la posibilidad de intervenir en una de tres licitaciones diferentes: A, B y C, con probabilidades 0.5, 0.3 y 0.2, respectivamente. El empresario sabe que si interviene en la licitación A tiene el 80% de posibilidad de obtener dictamen favorable, si interviene en la licitación B, el dictamen puede ser favorable con probabilidad 0.4, y si interviene en la licitación C, el dictamen le puede favorecer con probabilidad 0.6. Si el empresario ha recibido un dictamen favorable, ¿cuál es la probabilidad de que haya intervenido en la licitación B?

32. Una empresa está estudiando la posibilidad de construir una granja en cierto sector agropecuario. La compañía considera de gran importancia la construcción de un reservorio en las cercanías del lugar. Si el gobierno aprueba este reservorio, la probabilidad de que la compañía construya la granja es 0.9, de otra manera la probabilidad es de solo 0.2. El presidente de la compañía estima que hay una probabilidad de 0.6 de que el reservorio sea aprobado.
    a) Hallar la probabilidad de que la compañía construya la granja.
    b) Si la granja fue construida, hallar la probabilidad de que el reservorio haya sido aprobado.

33. Al contestar una pregunta de opción múltiple y con cinco distractores (posibles respuestas), de los cuales solo uno es la respuesta correcta, puede ocurrir que la persona que responda correctamente conozca realmente la respuesta o responda al azar. La probabilidad de que conozca la respuesta es 0.6, y de que recurra al azar, 0.4. Si un estudiante marcó correctamente la respuesta, ¿cuál es la probabilidad de que conozca la respuesta?

34. Un lote de chips contiene 2% de defectuosos. Cada chip es probado antes de ser enviado para su venta; sin embargo, el inspector de calidad no es totalmente confiable: la probabilidad de que el inspector diga que el chip es bueno dado que realmente está bueno es 0.95, y la probabilidad de que el inspector diga que el chip es defectuoso dado que está defectuoso es 0.94.

    Si el inspector indica que un chip está bueno, ¿cuál es la probabilidad de que este sea defectuoso?

35. Una compañía de seguros clasifica a las personas en una de tres categorías respecto al riesgo: bueno, promedio o malo. Sus registros indican que la probabilidad de que una persona se vea involucrada en un accidente en el lapso de un año es de 0.05, 0.15, y 0.30, respectivamente. Si 20% de las personas están clasificadas como de riesgo bueno, 50% como promedio y 30% como malo, ¿qué proporción de las personas tienen accidentes durante un año? Si el poseedor de una póliza no tiene un accidente durante un año, ¿cuál es la probabilidad de que haya sido clasificado como de riesgo promedio?

36. Después de analizar los factores que podrían incidir en el mal funcionamiento de las computadoras que ensambla, una empresa decide estudiar con mayor detalle los lotes de chips para video que la fábrica Visualchip le envía. La empresa Visualchip le indica a la empresa ensambladora que los lotes que le envía tienen 1% de chips defectuosos con probabilidad 0.97, o 2% de defectuosos con probabilidad 0.03. Para tratar de comprobar esta información y ante la duda, la ensambladora toma una muestra al azar de 4 chips y encuentra que todos estos chips son defectuosos. A la luz de la prueba realizada, la ensambladora quisiera realizar un reclamo a la empresa que envía los chips. ¿Recomendaría usted a la ensambladora que lleve a cabo el reclamo?

37. Una empresa recibe billetes de tres bancos: A, B y C. De A recibe el 60% de todos los billetes, de B, el 30%, y el resto de C. Se ha determinado que la proporción de billetes falsos que provienen de A es 0.1%, de B, 0.2% y de C, 0.1%. En cierta ocasión, al recibir una cantidad de billetes, resultó que uno era falso. ¿De qué banco se puede sospechar que proviene el billete falso?

38. Para conminar a sus deudores una compañía utiliza: el teléfono, visita personal y correo. De los datos registrados se sabe que al 20% se le sugiere por vía telefónica que paguen, 20% son visitados personalmente y al resto se les envía una carta. Las probabilidades de recibir respuesta positiva al aplicar estos métodos son: 0.6, 0.8 y 0.4, respectivamente. Si acaban de informar que un cliente acaba de hacer efectivo el pago de una deuda, ¿cuál es la probabilidad de que se le haya visitado personalmente?

39. Una prueba para diagnosticar una cierta enfermedad E dio los siguientes resultados para $N = 130$ personas.

Resultado de la prueba	***Estado***		
	No enfermo	***Enfermo***	***Total***
Negativo	0	5	5
Positivo	55	70	125
Total	55	75	$N = 130$

A la luz de estos resultados:

a) Evaluar esta prueba de diagnóstico.

b) ¿Sirve esta prueba para descartar un diagnóstico? ¿Sirve esta prueba para aceptar un diagnóstico?

c) Si se considera que la prevalencia de la prueba es 0.5, ¿cuáles son los valores predictivos de la prueba?

40. Se va a perforar un pozo de petróleo en cierto lugar. El terreno ahí es roca con probabilidad 0.53, arcilla con probabilidad 0.21 o arena con probabilidad 0.26. Un examen geológico da un resultado positivo con una exactitud del 35% en el caso de que sea roca, con una exactitud del 48% si es arcilla y con 75% si es arena. ¿Cuál es la probabilidad de que sea roca si la prueba resulta positiva?

41. El encargado del almacén de una planta de ensamblaje de computadoras ha determinado que, según su experiencia, el lote de 1,000 chips que se ha recibido puede tener 5% de defectuosos con probabilidad 0.6, 6% de defectuosos con probabilidad 0.3 y 7% con probabilidad 0.1. Para revisar las probabilidades indicadas por el encargado, el lote se somete a una inspección por muestreo, tomando 5 chips al azar. De los 5 chips observados resultó que ninguno es defectuoso. Indicar las probabilidades a posteriori.

42. Un médico tiene duda entre tres enfermedades E_1, E_2 y E_3, posibles en un paciente. Observando el estado general del paciente, al médico le parece que la probabilidad de que suceda la enfermedad E_1 es el triple de la probabilidad de que suceda cada una de las otras dos enfermedades. Sin embargo, ordena un examen de sangre, el que se sabe resulta positivo en el 10% de los casos cuando E_1 es la causa de la dolencia, en el 90% de los casos cuando la causa de la dolencia es E_2 y en el 60% de los casos cuando la causa de la dolencia es E_3. Si el resultado del análisis fue positivo, ¿cuál es la probabilidad final de cada enfermedad? A la luz de los resultados, ¿se puede afirmar que la probabilidad de que suceda la enfermedad E_1 es el triple de la probabilidad de que suceda cada una de las otras dos enfermedades?

43. Una empresa desea lanzar un nuevo producto. Puede suponerse que la demanda del nuevo producto puede ser alta (θ_1) o baja (θ_2), y que las ganancias que la empresa puede esperar son: cinco y un millón de pesos, respectivamente. La información inicial que tiene la empresa permite suponer que $p(\theta_1) = 0.4$, $Y(\theta_2) = 0.6$. Un equipo de marketing ofrece a la empresa disminuir la incertidumbre respecto de la demanda realizando una encuesta. La encuesta puede aconsejar el lanzamiento ($x = 1$) o no ($x = 0$). Suponiendo que:

$$p(x = 1|\theta_1) = 0.9 \ , \ p(x = 1|\theta_2) = 0.2$$

calcular el precio máximo que debe pagarse por la encuesta.

RESPUESTAS A LOS EJERCICIOS

1. a) {(O, O, O), (O, O, D), (O, D, O), (D, O, O), (O, D, D), (D, O, D), (D, D, O), (D, D, D)} b) $[0, +\infty[$

3. a) $\bar{A} \cap \bar{B} \cap \bar{C}$ c) $(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C) \cup (A \cap B \cap C)$. **4.** a) El espacio muestral tiene cuatro elementos. **6.** La probabilidad de que la computadora 1 quede fuera de servicio es 0.11, para la segunda es 0.15. **7.** 0.8 **8.** a) 0.55. **9.** a) 0.55 b) 0.95. **10.** a) 0.95 b) 0.04 c) 0.99. **11.** a) 3 a 5. **12.** a) no d) no. **13.** a) 0 b) no. **14.** 0.7. **15.** Con B. **16.** 0.4666. **17.** a) 0.4096 b) 0.016, c) 0.0256. **18.** a) P(varón) = 0.65, P(mujer) = 0.35 b) 0.3465. **19.** 0.5 **20.** a) 0.9998 b) 0.9996. **21.** a) 0.60 y c) 0.60. **22.** a) 0.52 b) 0.40 c) 0.7692. **23.** a) 0.00004. **24.** 1 aprox. **25.** 0.001, 0.081. **26.** Los tres tienen igual probabilidad. **27.** 0.64.

28. Como $\frac{P(C|L)}{P(NC|L)} = \frac{P(L|C)P(C)}{P(L|NC)P(NC)} = \frac{0.6P(C)}{0.3P(NC)} > \frac{P(C)}{P(NC)}$ la editora sí podrá decir que la propaganda fue efectiva.

29. $\frac{P(C|M \cap F)}{P(C|H \cap F)} = \frac{P(C \cap M \cap F)}{P(M \cap F)} \frac{P(H \cap F)}{0.9P(H \cap C)} = \frac{0.7 \times 10^{-4}}{P(F|M)} \frac{5P(F|M)}{0.9 \times 4x10^{-4}} = \frac{3.5}{3.6}$ Como 3.5/3.6 es menor que 1, la respuesta es negativa; sin embargo, este valor es muy cercano a 1, lo que indica que una mujer fumadora tiene prácticamente la misma probabilidad de contraer cáncer pulmonar que un hombre fumador. **30.** 0.016. **31.** 0.1875. **32.** a) 0.62 b) 0.8709. **33.** 0.8823. **34.** 0.0013. **36.** A partir de la información proporcionada, las probabilidades revisadas son 0.6690 y 0.3310. **37.** De A o de B. **39.** a) especificidad = 0, sensibilidad = 0.9333. **40.** 0.3854. **42.** Usando las condiciones indicadas se tiene que la probabilidad de tener la enfermedad E1 es 0.6. Aplicar el teorema de Bayes para calcular la probabilidad de que suceda cada enfermedad dado que el análisis salió positivo. **43.** Calcular las probabilidades revisadas: $p(\theta_1|x = 1)$ y $p(\theta_2|x = 1)$. El precio máximo a pagar es 5 $p(\theta_1|x = 1)$ + 1 ($p(\theta_2|x = 1)$).

CAPÍTULO 6

Variables aleatorias y distribución de probabilidad

Karl F. Gauss

Karl F. Gauss nació en Brunswick, Alemania, en 1777. Desde muy pequeño fue atraído por el cálculo. Es famosa la anécdota que narra que a los ocho años realizó de una manera instantánea el cálculo de la suma de todos los números del 1 al 100.

Los trabajos de Gauss estuvieron relacionados con varias disciplinas. Su tesis doctoral trató sobre la teoría de los números complejos y el teorema fundamental del álgebra. Su teoría de números es considerada uno de los trabajos más brillantes desarrollados en matemáticas. En estadística, dos de los principales aportes fueron el método de mínimos cuadrados y fundamentalmente la teoría de la distribución normal, conocida, en su honor, como la "distribución de Gauss".

Conjuntamente con Wilhelm Weber, Gauss inventó, cinco años antes que Samuel Morse, un telégrafo.

A los treinta años, Gauss asumió la dirección del observatorio de la Universidad de Brunswick, en donde permaneció hasta su deceso.

Karlf F. Gauss murió en 1855 en Gottingen.

CONTENIDO

6.1 Introducción

Generalmente, los resultados de los experimentos aleatorios presentan patrones que pueden ser estudiados con mayor facilidad si se asocian a números reales, y para ello se analiza la manera como estos ocurren. En esta sección se desarrollan distribuciones teóricas o empíricas que pueden servir para este propósito.

Habiendo asignado, por ejemplo, el número X de artículos defectuosos que aparecen en cada lote de 5 artículos seleccionados al azar y habiendo observado que en 100 de tales lotes la frecuencia relativa de lotes con 0, 1, 2, 3, 4 o 5 artículos defectuosos es como se indica en la Tabla 6.1, podremos establecer un modelo empírico que proporcione la probabilidad de que ocurra cada uno de estos valores de la variable.

TABLA 6.1 *Tabla de frecuencias*

Número de artículos defectuosos	*Lotes*	*Frecuencia relativa*
0	80	0.80
1	10	0.10
2	6	0.06
3	2	0.02
4	1	0.01
5	1	0.01
Total	100 lotes	1.00

Estas probabilidades son:

$P[X = 0] = 0.8$, $P[X = 1] = 0.10$, $P[X = 2] = 0.06$, $P[X = 3] = 0.02$, $P[X = 4] = 0.01$ y $P[X = 5] = 0.01$

Estas probabilidades, que expresan la manera como se reparte la *masa unitaria en los distintos valores de X*, es lo que se llama *distribución de probabilidad de X*, y por la manera como se han obtenido se considera que es un *modelo empírico*.

Cuando se conoce la probabilidad p de que un artículo sea defectuoso, la repartición de la masa unitaria también se puede hacer usando el modelo teórico llamado *modelo binomial.* Según este modelo, las probabilidades $P[X = 0]$, $P[X = 1]$, $P[X = 2]$, $P[X = 3]$, $P[X = 4]$ y $P[X = 5]$ se calculan con:

$$P[X = k] = C_k^5 p^k (1 - p)^{5-k} \text{ para } k = 1, 2, 3, 4, 5, \text{ y en donde } C_k^5 = \frac{5!}{(5 - k)!k!}.$$

Estos modelos se introducen formalmente a partir de los conceptos que a continuación se desarrollan.

6.2 Variables aleatorias

*Una **variable aleatoria** es una función que asigna un número real a cada resultado de un experimento aleatorio.*

Las variables aleatorias se denotan con letras mayúsculas: X, Y, etc. Los valores de las variables se denotan con letras minúsculas: x, y, etc.

EJEMPLO. *Variables aleatorias*

En la Tabla 6.2 se indican ejemplos de variables aleatorias que se pueden definir en los espacios muestrales de los respectivos experimentos aleatorios.

TABLA 6.2 *Variables aleatorias*

Experimento aleatorio	*Variable aleatoria*
Elegir al azar un lote de 10 artículos	*X = Número de artículos defectuosos en el lote*
Fabricación de un artículo	*X = Tiempo de realización del artículo*
Elegir una persona al azar	*X = Edad de la persona*
Elección de una familia al azar	*X = Número de personas que conforman la familia*

EJEMPLO. *Variable aleatoria. Tamaño de las empresas*

Consideremos una experiencia aleatoria que consiste en elegir 50 empresas al azar del grupo de todas las empresas del país. Considerando que una empresa es pequeña si tiene menos de 10 trabajadores, a cada empresa seleccionada asignémosle un número de la siguiente manera:

$$X = \begin{cases} 1 \text{ si la empresa es pequeña} \\ 0 \text{ si la empresa no es pequeña} \end{cases}$$

Esta relación define una variable aleatoria.

Variables aleatorias discretas y continuas

Las variables pueden ser discretas y continuas.

Una variable aleatoria es discreta si el conjunto de sus valores es finito o infinito, pero que se puede contar.

Las variables aleatorias discretas generalmente se usan cuando se desea describir el *número de veces* que ha ocurrido un suceso.

EJEMPLO. *Variables aleatorias discretas*

Si cada vendedor de una empresa ofrece su producto cinco veces al día, entonces la variable *X* que indica el número de veces que el vendedor tiene éxito (realiza una venta) es una variable aleatoria discreta.

Si para cada persona de una comunidad se considera el número *X* de periódicos que lee diariamente, entonces *X* es una variable aleatoria discreta.

> *Una variable aleatoria X es continua si puede tomar cualquiera de los valores de un intervalo.*

Si cada vez que ocurre una venta en una determinada empresa se considera el *valor X* de la venta realizada, entonces *X* es una variable aleatoria continua. Los valores que puede tomar *X* pueden ser cualquiera de los valores del intervalo $]0, +\infty[$.

EJEMPLO. *Variables aleatorias continuas*

El tiempo *X* en que se fabrica un determinado artículo es una variable aleatoria continua.

El voltaje medido de una batería de un vehículo es una variable aleatoria continua.

Nota

Al decir que una variable aleatoria continua puede tomar cualquiera de los valores de un intervalo, no estamos indicando que se tiene que encontrar cada uno de los valores del intervalo en los datos que se tienen. Así por ejemplo, es posible que en un conjunto de personas no se observe una que mida 1.637 m. Sin embargo, debemos considerar la posibilidad de que exista una persona con tal estatura en toda la población.

Corresponde ahora establecer los principios para modelar los patrones que podrían encontrarse en los valores de las variables. Estos patrones están relacionados con la variabilidad de los resultados de la variable. Para facilitar el estudio comenzamos con las variables aleatorias discretas.

Distribución de probabilidad de una variable aleatoria discreta

Se trata de establecer ponderaciones, usando la probabilidad, que indiquen el patrón de variabilidad de los valores de una variable aleatoria *X*. Cuando tales ponderaciones se tengan, se habrá establecido un modelo matemático o ley de probabilidad más comúnmente llamado *distribución de probabilidad* para *X.*

Si la variable es discreta, es decir, si toma los valores $x_1, x_2,..., x_n,...$, la distribución de probabilidad es definida por un conjunto de valores $P[X = x_i]$ que cumplen las siguientes propiedades:

a) Cada uno de estos valores es no negativo.

b) La suma de estos valores es igual a 1.

Cada expresión $P[X = x_i]$ representa un valor de probabilidad que se lee "probabilidad de que X tome el valor x_i".

La distribución de probabilidad de X constituye un modelo que describe una realidad de manera simple pero aproximada.

Los valores $P[X = x_i]$ pueden elegirse:

a) de manera empírica, a partir de una serie de repeticiones de la experiencia que da lugar al espacio muestral, en donde está definida la variable;

b) a partir de algún modelo teórico previamente elegido o

c) de manera subjetiva, a partir de la creencia del decisor.

EJEMPLO. *Modelando la distribución a partir de las frecuencias relativas*

Un sociólogo, al observar 100 familias seleccionadas al azar, ha determinado que en un distrito:

Ochenta familias no tienen hijos,
once familias tienen un hijo,
cuatro familias tienen dos hijos,
tres familias tienen tres hijos,
dos familias tienen cuatro hijos.

TABLA 6.3 *Distribución de probabilidad*

Número de hijos	*Probabilidad*
0	80/100
1	11/100
2	4/100
3	3/100
4	2/100

Basado en esta experiencia, el sociólogo podría indicar que para familias del mismo distrito se puede establecer que la *variable aleatoria X*, correspondiente al número de hijos, tiene la siguiente *distribución* o *ley de probabilidad de la variable:*

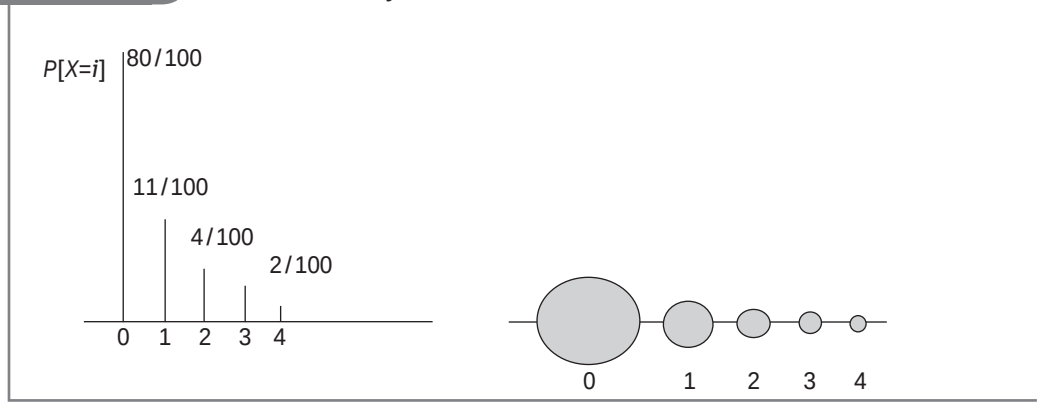
$$P[X = 0] = 80/100,..., P[X = 4] = 2/100$$

La distribución o ley de probabilidad puede representarse gráficamente mediante "bastones", como se indica en la Figura 6.1, o usando la idea de que la masa total unitaria se distribuye en "pequeñas masas puntuales o ponderaciones" sobre los valores de la variable.

Usando el modelo se puede obtener la probabilidad de determinados eventos como, por ejemplo, que una familia tenga 1 o 2 hijos:

$$P[\textit{uno o dos hijos}] = P[X = 1] + P[X = 2] = \frac{11}{100} + \frac{4}{100} = \frac{15}{100} = 0.15$$

FIGURA 6.1 *Gráfico de la ley de X*



EJEMPLO. *Modelando la distribución de los valores a partir de un modelo teórico*

El número de clientes que llegan a un banco entre las 9 a.m. y las 10 a.m. es una variable aleatoria X y sus valores pueden ser 0, 1, 2, 3,...

La distribución de la variable, descrita por la probabilidad de que la variable tome cualquiera de los valores k, $P[X = k]$ para $k = 0, 1, 2, 3,...$, puede determinarse utilizando un modelo teórico llamado *de Poisson con parámetro* λ. El valor de λ se interpreta como el promedio del número de clientes que llegan al banco entre las 9 a.m. y las 10 a.m.

El modelo de Poisson se aplica en muchas situaciones, y es muy importante en los procesos de producción de una empresa. En una empresa de producción en línea, el número de artículos defectuosos que se producen por día se distribuyen como la distribución de Poisson.

Esperanza y varianza de una variable aleatoria discreta

La distribución de probabilidad de una variable aleatoria es una función que proporciona una descripción completa del comportamiento de los valores de la variable;

sin embargo, en muchas ocasiones, se precisa de una descripción más concisa de estos. La *esperanza* y la *varianza* de una variable son resúmenes numéricos que pueden usarse con tal objetivo.

> *La* ***esperanza*** *o* ***valor esperado*** *de una variable aleatoria discreta, X, con valores* $x_1,\ldots, x_n$ (*n puede ser infinito*) *y cuya función de distribución de probabilidad es* $P[X = x_i]$, *se denota con* $E(X)$ *y se define como:*
>
> $$E(X) = \sum_i x_i P[X = x_i]$$

La esperanza de *X*, llamada también *media de la población*, es la suma ponderada de los valores de la variable y resume en un solo número parte de la información contenida en la función de distribución de probabilidades.

La esperanza de una variable se puede extender a funciones de la variable. Si $g(X)$ es una función de la variable *X con distribución* $P[X = x_j]$, se define la esperanza de $g(X)$ como el número:

$$E\{g(X)\} = \sum_j g(x_j)P[X = x_j]$$

que se interpreta como la suma ponderada de los valores $g(x_j)$.

De este modo, puede encontrarse, por ejemplo, la esperanza del cuadrado de los valores de una variable o del costo de realizar una tarea en función del tiempo de realización de esta.

El valor esperado de la variable describe la tendencia central de una variable aleatoria; sin embargo, no informa de la dispersión de sus valores. Una medida que ayuda en este sentido y que indica la manera como están dispersos los valores de la variable respecto de la media es la *varianza*. Este número es el promedio ponderado del cuadrado de las desviaciones respecto de la media.

La *varianza de una variable aleatoria discreta X,* que se denota con $V(X)$, se define como:

$$V(X) = \sum_j (x_j - E(X))^2 P[X = x_j]$$

A la varianza de *X* se le denota con σ_X^2.

Al número $\sqrt{V(X)}$ se le llama *desviación estándar* de *X* y se le denota con σ_X.

Si no hay lugar a confusión se usa simplemente σ y σ^2 para la desviación estándar y varianza, respectivamente. A menudo se utiliza la desviación estándar de la variable, pues tiene la misma dimensión que la variable. Si *X* proporciona valores en metros, la desviación estándar también estará en metros.

El análisis conjunto de la esperanza y de la varianza incrementan el conocimiento del patrón de variabilidad de los valores de la variable.

Repitiendo lo indicado anteriormente, la mayoría de los valores de la variable se encuentran en el intervalo [*media* − 2.*desviaciones estándar, media* + 2.*desviaciones estándar*]. Los valores que no caen en este intervalo se consideran como *infrecuentes o poco comunes*.

EJEMPLO. *Negocios*

Cada vez que se invierte en un negocio se gana $ 2,000, con probabilidad 0.2, se gana $ 1,500 con probabilidad 0.7 y se pierde $ 3,000 con probabilidad 0.1.

Si se denota con X a la variable ganancia, se tiene que su distribución es como indica la Tabla 6.4.

TABLA 6.4 *Ganancia en el negocio*

X	2,000	1,500	−3,000
$P[X = x]$	0.2	0.7	0.1

Entonces, el valor esperado de la utilidad X es igual a:

$$E(X) = 2{,}000 \times 0.2 + 1{,}500 \times 0.7 + (-3{,}000) \times 0.1 = 1{,}150$$

Cada vez, el inversor puede ganar $ 2,000 o $ 1,500, pero también puede perder $ 3,000. Si realiza muchas veces el negocio, ganará, en promedio, $ 1,150 por vez, aproximadamente.

El cálculo de la varianza de la ganancia X aparece en la Tabla 6.5.

TABLA 6.5 *Varianza y desviación estándar*

Ganancia	*Probabilidad*	*(Ganancia-1,150)²*Probabilidad*
2,000	0.2	144,500
1,500	0.7	85,750
−3,000	0.1	1,722,250
	Varianza	1,952,500
	Desv. estándar	1397.3188

EJEMPLO. *Venta de vehículos*

La empresa Golf, dedicada a la venta de vehículos, ha determinado que la demanda X mensual del número de vehículos cuyo valor de venta es 30,000 dólares tiene la siguiente distribución.

TABLA 6.6 *Demanda de vehículos*

X	0	1	2	3	4	5
p	0.2	0.3	0.4	0.06	0.03	0.01

Si por cada vehículo vendido la empresa gana el 10% pero incurre en gastos fijos cada mes de 2,000 dólares en total, entonces la utilidad mensual es:
$U = 0.1(30{,}000)X - 2{,}000$.

El valor esperado mensual de la utilidad es:

$$E(U) = 0.1(30{,}000)E(X) - 2{,}000 = 0.1(30{,}000)(1.45) - 2{,}000 = 2{,}350$$

Propiedades de la esperanza y la varianza de una variable aleatoria discreta

La esperanza de una variable aleatoria X, discreta, tiene las siguientes propiedades:

1. $E(a) = a$, *a constante*
2. $E(aX) = aE(X)$, *a constante*. En general: $E(ag(X)) = aE(g(X))$
3. $E(aX + b) = aE(X) + b$, a y b *constantes*

Las siguientes propiedades son básicas en el desarrollo de los diferentes cálculos que se realizan sobre la varianza.

Si X es una variable aleatoria discreta, se cumple:

1. $V(X) \geq 0$
2. *Si* $X = a$, $V(X) = 0$
3. $V(aX + b) = a^2V(X)$
4. $V(X) = E(X^2) - [E(X)]^2$

APLICACIÓN: El caso del negocio de las flores

Rosa Flores tiene una pequeña empresa que se dedica a la venta de flores para regalo por Internet. Rosa recibe los pedidos por este medio y luego los envía al almacén que posee en los suburbios de la ciudad. El almacén es surtido por floricultores de la zona. Las flores que al terminar el día no son vendidas se rematan a los fabricantes de aceites para la industria de los perfumes, pero a un precio menor del que se venden las flores para regalo. Últimamente, como el negocio ha crecido, Rosa desearía ampliar los pedidos a los floricultores, pero teniendo en cuenta que las flores son artículos perecederos ha decidido hacer un estudio que le permita hacer su pedido de manera adecuada.

Cada día Rosa recibe en su almacén 300 ramos de flores para la venta de parte de los floricultores. Por cada ramo de flores que Rosa vende, gana $ 40, y pierde $ 10 por cada ramo no vendido al finalizar el día. Después de revisar sus registros de datos, Rosa ha podido determinar que la demanda diaria X de los ramos de flores tiene la ley de probabilidad descrita en la Tabla 6.7.

TABLA 6.7 *Distribución de las ventas*

x	100	200	300
f(x)	0.1	0.6	0.3

>>>

(*vende 100 con probabilidad* 0.1, *etc*.). Se observa que la demanda tiene un valor esperado de $E(X) = 220$.

La ganancia de Rosa es una variable aleatoria G que es función de la variable aleatoria X y es igual a:

$G = 40(X) - 10(300 - X)$

Según esto, la esperanza de la ganancia G es igual a:

$E(G) = 40E(X) - 10(300 - E(X)) = 8{,}000$

Algunos días, la ganancia de Rosa será $ 2,000 (cuando venda 300 ramos), pero en promedio ("a la larga") ganará $ 8,000 por día.

Como la demanda puede ser de 100 o 200 o 300 ramos, Rosa deberá pedir a los floricultores una de estas cantidades, pero de tal manera que la utilidad esperada por día sea la máxima.

Si Rosa pide $k = 100$ ramos, la ganancia es $G = 40(100) = 4{,}000$, cualquiera que sea la demanda, y su valor esperado es también 4,000.

Si Rosa pide $k = 200$ ramos, la ganancia es:

$$G = \begin{cases} 40(100) - 10(100) \textit{ si la demanda es } X = 100 \\ 40(200) \textit{ si la demanda es } X = 200 \text{ o } 300 \end{cases}$$

y el valor esperado de la ganancia es:

$E(G) = (3{,}000)(P(X = 100)) + (8{,}000)(P(X = 200 \text{ o } X = 300)) = 7{,}500$

Si Rosa pide $k = 300$ ramos, la ganancia es:

$$G = \begin{cases} 40(100) - 10(200) \textit{ si la demanda es } X = 100 \\ 40(200) - 10(100) \textit{ si la demanda es } X = 200 \\ 40(300) \qquad\qquad \textit{si la demanda es } X = 300 \end{cases}$$

y el valor esperado de la ganancia es:

$E(G) = (2{,}000)(P(X = 100)) + (7{,}000)(P(X = 200) + (12{,}000)P(X = 300) = 8{,}000$

Se deduce que Rosa debe pedir 300 ramos para la venta si desea maximizar su ganancia diaria esperada.

6.3 Algunos modelos probabilísticos para variables aleatorias discretas

Se estudian ahora ciertos modelos teóricos de distribución de probabilidad para variables aleatorias discretas que suceden a menudo en los campos de la industria y de los negocios.

Distribución binomial

Este modelo teórico se utiliza para describir el patrón que tiene el número X de elementos con cierta propiedad que se encuentran en un cierto conjunto de tamaño n.

El número de personas que tienen tarjeta de crédito en una lista de 20 clientes de un banco, el número de ventas realizadas en tres visitas diarias realizadas por un vendedor de seguros, pueden modelarse con la distribución binomial.

> *Una variable aleatoria discreta X, con valores* 0, 1, 2,..., *n, sigue* ***la distribución binomial con parámetros n*** *y* ***p*** *si su ley de probabilidad se define como:*
>
> $$P[X = k] = C_k^n p^k q^{n-k}$$

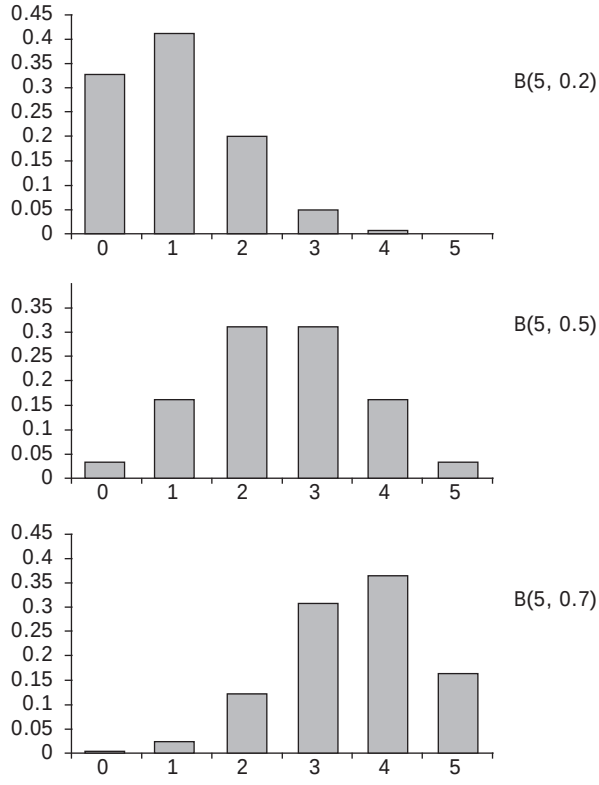
con $C_k^n = \dfrac{n!}{(n-k)!k!}$, $0 < p < 1$, $q = 1 - p$ *y para* $k = 0, 1, 2, \ldots, n$.

Si X sigue la distribución *binomial con parámetros n y p,* se escribe $X \sim B(n, p)$.

TABLA 6.8 *Ley binomial para diferentes parámetros*

Valores k	***B*(5, 0.2) *P*(*X* = *k*)**	***B*(5, 0.5) *P*(*X* = *k*)**	***B*(5, 0.7) *P*(*X* = *k*)**
0	0.32768	0.03125	0.00243
1	0.4096	0.15625	0.02835
2	0.2048	0.3125	0.1323
3	0.0512	0.3125	0.3087
4	0.0064	0.15625	0.36015
5	0.00032	0.03125	0.16807

FIGURA 6.2 *Ley binomial para diferentes parámetros*



En general, si $p = 0.5$, la gráfica de la distribución binomial es simétrica; si $p > 0.5$, la gráfica es asimétrica con una cola a la izquierda; y si $p < 0.5$, la gráfica es asimétrica con una cola a la derecha.

¿Cuándo usar la distribución binomial?

Para aplicar este modelo probabilístico deben cumplirse los siguientes requisitos:

1. El proceso de donde vienen los resultados tiene un número *n* fijo de repeticiones o ensayos.
2. Las repeticiones deben realizarse de manera independiente.
3. Cada resultado es una *experiencia de Bernoulli*; es decir, solo tiene dos resultados posibles, por ejemplo, *éxito* y *fracaso.*
4. La probabilidad *p* de éxito permanece constante en cada ensayo.

Satisfechas estas condiciones, el modelo que puede atribuirse al número de éxitos, *X*, en las *n* repeticiones es el modelo binomial con parámetros *n* y *p.*

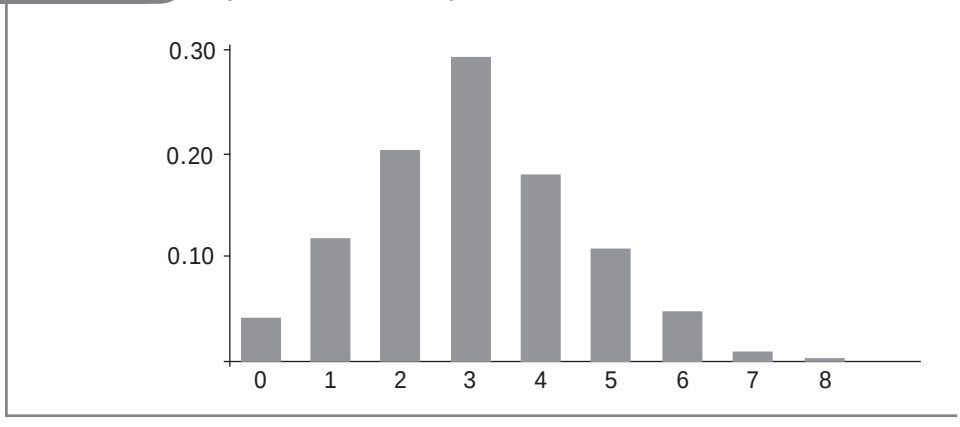
EJEMPLO. *Las tarjetas de crédito*

Si el 30% de los clientes de un banco tienen tarjeta de crédito, entonces podemos decir que $p = 0.3$ es la probabilidad de que un cliente del banco tenga tarjeta de crédito. La probabilidad de observar k personas que tienen tarjeta de crédito de una lista de 10 clientes del banco, seleccionados al azar, se puede calcular usando la distribución binomial, de la siguiente manera:

$$P[X = k] = C_k^{10} 0.3^k 0.7^{10-k} \quad \text{para } k = 0, 1, 2, \ldots, 10$$

donde X es el número de personas que tienen tarjeta de crédito de una lista de 10.

FIGURA 6.3 *Ley binomial* $n = 10$, $p = 0.3$



Se escribe $X \sim B(10, 0.3)$.

Las distintas probabilidades, calculadas con la fórmula anterior para $k = 1, 2, \ldots, 10$ se representan en la Figura 6.3.

La probabilidad de que 5 clientes tengan tarjeta de crédito (de los 10) es igual a:

$$P[X = 5] = C_5^{10} (0.3)^5 (0.7)^{10-5} = 0.1029$$

La probabilidad de que el número de clientes que tienen tarjeta de crédito sea menor que 2 es igual a:

$$P[X < 2] = P[X = 0] + P[X = 1] = C_0^{10} 0.3^0 0.7^{10} + C_1^{10} 0.3^1 0.7^9 = 0.1493$$

Nota

En el ejemplo anterior, para aplicar el modelo binomial se deberá suponer que la elección de cada uno de los 10 clientes se hizo de manera independiente. Esta suposición no hubiera sido posible si los 10 clientes se eligen de un grupo pequeño, por ejemplo

de 20 clientes y sin restitución. Si cada extracción se realiza *sin* restitución, la probabilidad *p* en cada una de ellas no es constante, por lo tanto no es aplicable el modelo binomial; sin embargo, cuando *N* es grande y *n* es pequeño respecto de *N*, se puede considerar que la probabilidad es constante para cada extracción, y por lo tanto la distribución de *X* se puede aproximar con la binomial. En la práctica, la aproximación es buena si $n \leq 0.1N$ (estrictamente *X* sigue la distribución *hipergeométrica*, que será analizada posteriormente). Estos resultados se obtienen igualmente cuando se toman todos los elementos a la vez.

EJEMPLO. *Modelando el número de chips defectuosos*

Para controlar la calidad de determinado tipo de chips que se fabrican en la empresa FAB, y en donde se espera que el 10% sean defectuosos, se tomaron 1,000 muestras de 20 chips cada una. Los resultados que indican las frecuencias del número de chips defectuosos obtenidos aparecen a continuación.

En la Tabla 6.9 se lee que en el 12.9% de las muestras no aparecen defectuosos, que en el 27.4% de las muestras aparece 1 defectuoso, etc.

TABLA 6.9 *Frecuencia observada*

Defec.	*Frec.*	*Frec. relati.*
0.00	129	0.129
1.00	274	0.274
2.00	305	0.305
3.00	184	0.184
4.00	68	0.068
5.00	28	0.028
6.00	9	0.009
7.00	3	0.003
Total	1,000	100.0

Frecuencia calculada, usando la distribución binomial

Defectuosos: (k)	*Probabilidad: P[X = k]*
0.00	0.12157665
1.00	0.27017034
2.00	0.28517981
3.00	0.19011987
4.00	0.08977883
5.00	0.03192136
6.00	0.00886704
7.00	0.00197045
8.00	0.00035578
9.00	5.2708E-05
10.00	0.0000000

Teóricamente, los resultados de esta experiencia se pueden modelar con la binomial de parámetros $n = 20$ y $p = 0.10$.

Notar que las frecuencias observadas son aproximadamente iguales a las frecuencias calculadas con el modelo teórico binomial.

Esperanza y varianza de una variable aleatoria con distribución binomial

Si una variable X tiene ley binomial con parámetros n y p entonces se cumple que su esperanza y su varianza son:

$$E(X) = np \text{ y } V(X) = npq, \text{ en donde } q = 1 - p$$

EJEMPLO. *Ventas de seguros de vida*

El número X de seguros de vida que se venden en 10 visitas diferentes se puede modelar con la distribución binomial si se conoce que cada vez la probabilidad de realizar una venta es 0.4.

Usando esta distribución, se tiene que la probabilidad de que al realizar 10 visitas se vendan 4 seguros de vida es 0.2508. Si diariamente se visitan 10 de tales clientes, se espera que en el 25.08% de los días se vendan 4 seguros de vida.

También se tiene que el valor esperado de la variable X es 4. Si por cada visita la probabilidad de que se venda un seguro de vida es 0.4, entonces por cada 10 visitas se venderán en promedio 4 seguros de vida.

EJEMPLO. *Modelando la ocupación en hoteles*

Cada día que un cliente se aloja en el hotel Estrella debe pagar un costo fijo de $ 100. Además, si ocupa una habitación de tipo A deberá pagar $ 50 más, y si ocupa una habitación de tipo B deberá pagar $ 20 más. De los registros del hotel se deduce que la probabilidad de que un cliente ocupe una habitación de tipo A es 0.4 y de tipo B es 0.6. Si un día se presentaron 10 clientes:

a) ¿Cuál es la probabilidad de que todos se hayan alojado en habitaciones de tipo B?

b) Indicar el total T de dinero que se recaudó por los 10 clientes. Hallar el valor esperado de T. Interpretar el resultado.

c) Hallar la probabilidad de que un día cualquiera se haya recaudado $ 1,320.

Solución

a) Sea X el número de clientes, de los 10, que se alojan en habitaciones de tipo B. Se tiene que X sigue una ley binomial de parámetros $n = 10$ y $p = 0.6$. La probabilidad de que los 10 se hayan alojado en habitaciones de tipo B es:

$$P[X = 10] = C_{10}^{10}\, 0.6^{10} 0.4^{0} = 0.0060$$

b) El total de dinero T que se recauda por los 10 clientes es igual a:

$$T = 1{,}000 + 20X + 50(10 - X),\ X = 1,\ldots,\ 10$$

o

$$T = 1{,}500 - 30X,\ X = 1,\ldots,\ 10$$

La esperanza de T es $E(T) = 1{,}500 - 30E(X) = 1{,}500 - 30(10)(0.6) = 1{,}320$.

c) Para recaudar T = \$ 1,320 es necesario y suficiente que $X = 6$. Luego,

$$P[\text{“Recaudar \$ 1,320”}] = P[X = 6] = C_6^{10}0.6^6 0.4^4 = 0.2508$$

EJEMPLO. *Las calificaciones de los alumnos*

Una prueba de aptitud consta de 10 preguntas con 5 alternativas cada una, de las cuales solo una es la correcta. La calificación se realiza de la siguiente manera: por cada pregunta correctamente respondida, el que responde recibe 2 puntos y por cada pregunta mal contestada recibe k puntos. Se desea determinar el valor de k de tal manera que la nota esperada de un alumno que responde al azar las 10 preguntas sea 0.

Solución

Sea X el número de preguntas acertadas de las 10 respondidas al azar. La variable X tiene distribución binomial con parámetros 10 y 0.2. La calificación es:
$C = 2X + k(10 - X) = (2 - k)X + 10k$. El valor esperado de C es $E(C) = (2 - k)(10)(0.2) + 10k$.

El valor de k para el cual el valor esperado es 0 es −0.5. Deberá descontarse 0.5 puntos por cada pregunta mal respondida.

EJEMPLO. *Recordación de la marca*

El encargado de la publicidad de la empresa Tucmesa indica que el 70% de las personas sí recuerdan el nombre de la marca de la empresa; sin embargo, al consultarse a 8 agentes de ventas solo 2 decían recordar la marca de la empresa. ¿Es compatible este resultado con el indicado por el encargado de la publicidad?

Solución

Si el 70% de las personas recuerda el nombre de la marca, la probabilidad de que de 8 agentes 3 o menos de ellos recuerden la marca es:
$C_8^2 0.7^2 0.3^6 + C_1^8 0.7^1 0.3^4 + C_0^8 0.7^0 0.3^8 = 0.011$.

Si lo que indica el encargado de la publicidad es verdadero, encontrar un resultado como el que se ha obtenido en la consulta es casi imposible; sin embargo, se ha obtenido. Podemos decir que el último resultado es incompatible con la afirmación del encargado de la publicidad; al parecer este no dice la verdad.

JAMES BERNOULLI

James Bernoulli nació el 27 de diciembre de 1654 en Bassel, Suiza. Perteneció a una familia de renombrados matemáticos. Estudio teología, sin embargo, fue atraído por las matemáticas, en especial por el cálculo. En 1690 usó por primera vez la palabra "integral" para describir la inversa del diferencial.

Algunos de sus trabajos sobre cálculo combinatorio fueron publicados después de su muerte en el libro *Ars Conjectandi*, obra considerada como la precursora de la teoría de probabilidad.

Bernoulli se desempeñó en sus comienzos como profesor en la Universidad de Bassel en la rama de Filosofía y Mecánica, pero terminó como profesor de matemáticas hasta su muerte, en 1705.

Distribución de Poisson

Con esta distribución se puede modelar, por ejemplo, el número de llamadas telefónicas que se reciben en una central telefónica entre las 10 a.m. y 12 m. El número de personas que llegan a un banco entre las 9 a.m. y las 10 a.m. puede ser modelado también con una distribución de Poisson.

Esta distribución fue ideada por Simeon Denis Poisson (1781–1840), y la expresión que la describe es la siguiente:

$$P[X = k] = \frac{e^{-\lambda}(\lambda)^k}{k!}\ ,\ k = 0,\ 1,\ldots,$$

en donde X es la variable que indica el número de eventos que ocurren en un intervalo de longitud unitaria y λ es una constante mayor que 0.

Esta distribución se llama distribución de Poisson de tasa λ.

Cuando una variable X sigue una distribución de Poisson con tasa λ se escribe $X \sim P(\lambda)$.

¿Cuándo se usa este modelo?

El modelo de Poisson se usa para modelar, por ejemplo:

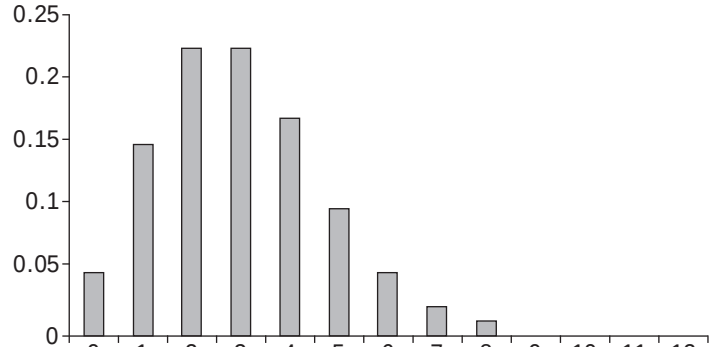
- El número de accidentes de tránsito que ocurren los lunes entre las 8 y las 9 de la mañana en la intersección de dos avenidas determinadas.
- El número total de inundaciones durante un año en una región determinada.
- El número total de artículos defectuosos en una línea de producción en un intervalo determinado.
- El número de vehículos que llegan a una estación de gasolina entre las 8 a.m. y 9 a.m.
- El número de personas que acceden a Internet entre las 11 a.m. y 12 m.

A continuación se presenta la distribución de Poisson con tasa $\lambda = 3$.

TABLA 6.10 *Distribución de Poisson con $\lambda = 3$*

Valores x	*P(X = k)*
0	0.04978707
1	0.14936121
2	0.22404181
3	0.22404181
4	0.16803136
5	0.10081881
6	0.05040941
7	0.02160403
8	0.00810151
9	0.0027005
10	0.00081015
11	0.00022095
12 o más	5.5238E-05

FIGURA 6.4 *Distribución de Poisson con parámetro $\lambda = 3$*



EJEMPLO. *Llegadas a un banco*

La tasa de llegada de los clientes a un banco es igual a 20 por hora. Si se considera que el número de llegadas al banco se puede modelar con la distribución de Poisson:

a) Hallar la probabilidad de que entre las 10 a.m. y las 11 a.m. lleguen al banco 10 clientes.

b) Hallar la probabilidad de que el primer cliente llegue 10 minutos después de abierto el banco.

Solución

a) Si X es la variable aleatoria que indica el número de clientes que llegan entre las 10 a. m. y las 11 a. m., se tendrá que la probabilidad de que en ese intervalo lleguen 10 clientes es igual a:

$$P[X = 10] = \frac{e^{-20}20^{10}}{10!}$$

b) Si la tasa de llegada es 20 clientes por hora, se puede considerar que la tasa de llegada es 20/6 clientes por cada 10 minutos. Luego, si X es la variable que indica el número de clientes que llegan en intervalos de 10 minutos, se tendrá que la probabilidad de que el primer cliente llegue después de 10 minutos de abierto el banco equivale a la probabilidad de que en el intervalo [0, 10] es:

$$P[X = 0] = \frac{e^{-20/6}(20/6)^0}{0!} = 0.0356$$

Esperanza y varianza de una variable aleatoria con ley de Poisson

Para una variable aleatoria que tiene distribución de Poisson con tasa en el intervalo I de longitud unitaria, se cumple:

$$E(X) = \lambda \quad y \quad V(X) = \lambda$$

La tasa λ puede aproximarse observando el número de eventos que ocurren en varios intervalos de longitud unitaria y luego tomando el número promedio de eventos ocurridos por intervalo.

Observamos que en un proceso de producción en línea que se modela con la distribución de Poisson, si se reduce el promedio de productos defectuosos, en un intervalo de longitud unitaria disminuirá la variabilidad del número de artículos defectuosos.

EJEMPLO. *El negocio de taxi*

A una estación de taxis llegan clientes a solicitar el servicio a razón de 15 por hora. Si en la estación existen 5 taxis entonces el número X de clientes que solicitan servicio por taxi sigue una distribución de Poisson con tasa $\lambda = 15/5 = 3$ clientes/hora por taxista.

Si el servicio de taxi cuesta 10 dólares por cliente, entonces el ingreso que obtiene cada taxi es $C = 10X$. El valor esperado del ingreso por taxi es $E(C) = 10E(X) = 30$ dólares.

Distribución de Poisson como aproximación de la distribución binomial

La distribución binomial indica la probabilidad de que un número particular de observaciones en n repeticiones de un experimento exhiban alguna propiedad. Cuando el

número de repeticiones es muy grande y la probabilidad *p* de exhibir tal propiedad es un valor pequeño, la distribución binomial puede aproximarse con una distribución *de Poisson de parámetro* $\lambda = np$. Esta propiedad se expresa formalmente de la siguiente manera:

> *La distribución binomial con parámetros n y p puede aproximarse mediante la distribución de Poisson cuando n es suficientemente grande y p es pequeño.*

Si una variable *X* tiene distribución binomial con parámetros *n y p*, se calcula con:

$$P[X = k] \approx \frac{e^{-\lambda}(\lambda)^k}{k!} \text{ con parámetro } \lambda = np \text{ cuando } n \geq 30 \text{ y } np \leq 5$$

EJEMPLO. *Aproximación de la binomial mediante la Poisson*

En una caja hay 100 resistencias, las cuales proceden de una fábrica que produce el 1% de defectuosas. Hallar la probabilidad de que ninguna resistencia sea defectuosa.

Si *X* denota al número de resistencias de la caja que son defectuosas, se tendrá que esta variable puede modelarse con una distribución binomial de parámetros $n = 100$ y $p = 0.01$. De este modo puede escribirse:

$$P[X = 0] = C_0^{100}\, 0.01^0 0.99^{100} = 0.99^{100} = 0.366$$

El valor de los parámetros satisfacen las condiciones que permiten aproximar $P[X = 0]$ usando la distribución de Poisson con $\lambda = (100)(0.01) = 1$. Así se tiene:

$P[X = 0] \cong e^{-1} = 0.3679$, valor aproximadamente igual al anterior.

Distribución hipergeométrica

Se dice que una variable aleatoria X discreta tiene ***distribución hipergeométrica con parámetros N, r y n*** *si su ley de probabilidad es:*

$$P[X = k] = \frac{C_k^r C_{n-k}^{N-r}}{C_n^N}$$

y los valores k de la variable cumplen con $máx[0, n + r - N] \leq k \leq min[n, r]$.

Cuando *X* sigue la distribución hipergeométrica con parámetros *N, r* y *n,* se escribe $X \sim H(N, r, n)$.

¿Cuándo se usa este modelo?

Considérese una población finita con *N* elementos, *r* de los cuales tienen la característica *A* y los $N - r$ restantes tienen la característica *B*.

Se eligen al azar y de una sola vez n elementos de la población y se considera la variable aleatoria X, que indica el número de elementos con la característica A entre los n elegidos.

La variable X tiene, como se puede comprobar, la distribución hipergeométrica de parámetros N, r y n.

El resultado de cada extracción depende de lo que haya sucedido en la anterior, y a diferencia del caso de las extracciones con restitución, estos resultados no son independientes.

Esperanza y varianza de una variable aleatoria con ley hipergeométrica

Para una variable aleatoria X cuya distribución es hipergeométrica, H(N, r, n), se cumple que:

$$E(X) = n(r/N) \text{ y } V(X) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Si N es un valor suficientemente grande y n es un valor pequeño respecto de N, $\left(\frac{N-n}{N-1}\right)$ se aproxima a 1, y así la esperanza y la varianza tienen la misma forma que en el caso de la distribución binomial.

APLICACIÓN: El caso de Taxiseguro

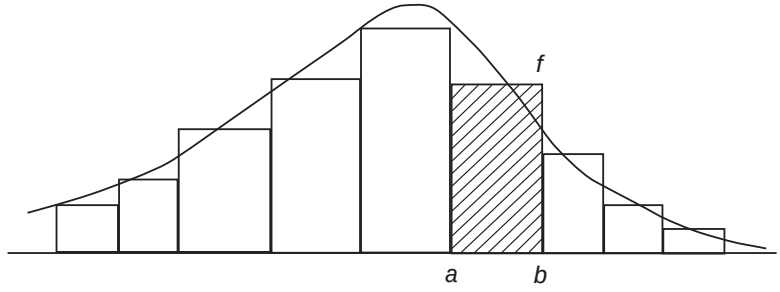
La empresa Taxiseguro tenía el monopolio del servicio de taxis en una pequeña ciudad. Los taxis eran llamados por teléfono a una central de servicio de donde acudían para realizar el respectivo servicio. Últimamente han incursionado otras empresas que ofrecen hacer el servicio de manera gratuita si el taxi no llega al llamado antes de los 15 minutos. Taxiseguro ha comenzado a tener una baja en el número de servicios que realiza y se ha determinado que ello se debe a la oferta del servicio gratuito que las otras compañías realizan. Como Taxiseguro no desea quedarse atrás, quisiera ofrecer también una oferta similar a los clientes, pero para ello deberá revisar la información contenida en la base de datos relacionada con los tiempos utilizados para llegar al domicilio del solicitante. Según la información obtenida, y considerando las diferentes dificultades actuales del tránsito, la empresa ha determinado que la probabilidad de que se llegue desde la central de servicio hasta el domicilio del solicitante antes de los 15 minutos es 0.95. Si cada taxi de la empresa realiza 20 servicios por día a diferentes puntos de la ciudad, es de esperar que uno de los servicios a realizar deberá hacerse de manera gratuita. Este resultado se ha determinado, pues el número de veces que Taxiseguro llegará en un tiempo mayor que 15 minutos por día es una variable aleatoria que puede modelarse con una distribución binomial de parámetros $n = 20$ y $p = 0.05$. Taxiseguro ha decidido, a base de este resultado y de su presupuesto, que llevará a cabo la oferta.

Función de densidad de una variable aleatoria continua

En esta sección se desarrollan los principios básicos para estudiar la variabilidad de los valores de una variable aleatoria continua *X*. Usando las probabilidades, se puede modelar, por ejemplo, la variabilidad de los valores que corresponden a las estaturas de las personas, los tiempos necesarios para realizar una tarea específica, la tasa de retorno diaria del valor de una acción, etc.

Consideremos un histograma para valores de *X* cuyas longitudes de los intervalos de clase son iguales y en donde la altura de cada rectángulo es igual a la frecuencia relativa dividida entre la longitud del intervalo. La suma de las áreas de los rectángulos construidos así es igual a 1. Si el histograma se construye de tal manera que los intervalos de clase sean cada vez más pequeños, el polígono de frecuencias correspondiente *tiende* a una función continua, que cambia suavemente. A esta función, que se utiliza para "modelar el comportamiento" de los valores de *X,* se le llama *función de densidad de probabilidad* de *X* o simplemente *función de densidad de X*.

FIGURA 6.5 *Área de la región sombreada* ≈ $P[a \leq X \leq b]$



Usando la función de densidad es posible calcular la probabilidad de que un valor de la variable esté en un determinado intervalo. La probabilidad de que la variable *X* tome valores en el intervalo [*a, b*] se denota con $P[a \leq X \leq b]$, y es igual al área bajo la gráfica de *f*, entre las rectas $x = a$ y $x = b$ y por encima del eje horizontal. De este modo se analiza la variabilidad de los valores de la variable.

Una serie de modelos probabilísticos teóricos para variables continuas se presentan más adelante, y pueden ayudar en la descripción del patrón de la variabilidad de los datos que resultan en muchas situaciones reales.

Considerando la manera como se ha obtenido la función de densidad, será fácil entender las siguientes propiedades.

La *función de densidad* de una variable aleatoria continua X es una función $f(x)$ que cumple las siguientes propiedades:

a) Los valores de f son mayores o iguales que 0. ($f(x) \geq 0$).

b) El área bajo la gráfica de la función y por encima del eje X es 1.

c) La probabilidad de que la variable X tome valores entre a y b se denota con $P[a \leq X \leq b]$, y es igual al área comprendida entre la gráfica de f, el eje X y las rectas paralelas al eje Y que pasan por los valores a y b.

Observaciones

1. Los valores de la función de densidad no son probabilidades. Pero esta función sí permite el cálculo de probabilidades.
2. Según c), la probabilidad de que una variable aleatoria continua tome un valor a es cero.
3. Cuando X toma todos sus valores en un intervalo $[a, b]$, simplemente se establece que $f(x) = 0$ para los valores que no están en $[a, b]$.

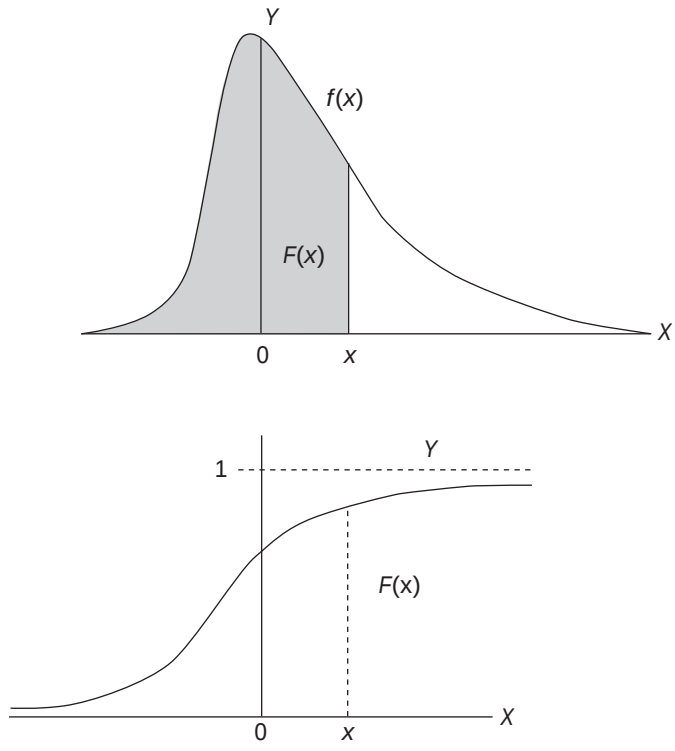
Función de acumulación o de distribución de una variable aleatoria continua

La variabilidad de los valores de una variable aleatoria X continua o discreta puede conocerse también a partir de las probabilidades acumuladas $P[X \leq x]$. Estas probabilidades determinan la llamada función de acumulación o de distribución.

> *Se define la* ***función de distribución o de acumulación*** *de una variable aleatoria continua X o discreta como:*
>
> $$F(x) = P[X \leq x], \text{ para todo número real } x$$

Esta función corresponde al área bajo la gráfica de la función de densidad, desde $-\infty$ hasta x.

FIGURA 6.6 *Función de densidad y de acumulación*



f(*x*): *Función de densidad de X* *F*(*x*): *Función de acumulación de X*

Esperanza y varianza de una variable aleatoria continua

Para una variable X continua, la *esperanza o valor esperado de X* se denota con $E(X)$ y se considera que es el valor alrededor del cual están "concentrados" sus valores.

Como en el caso discreto, con la esperanza de X, se resume en muchos casos la información contenida en la función de distribución de probabilidad. Otras veces será necesario agregar otros parámetros como la varianza de la variable.

También, como en el caso discreto, la esperanza de la variable es la medida de tendencia central más representativa y más usada.

Si la función de densidad de la variable es simétrica respecto de una cierta recta vertical $x = c$, la esperanza de la variable es igual a ese valor c.

La ley débil de los grandes números determina que la media aritmética de un conjunto de valores de X estará más cerca de la esperanza a medida que el número de datos es más grande. Este principio permitirá estimar la esperanza de una variable con la media aritmética de un conjunto de valores de X.

La *varianza de X,* que se denota con $V(X)$, es el valor que indica la dispersión de los valores de la variable alrededor del valor esperado de X.

Al número $\sqrt{V(X)}$ se le llama *desviación estándar* de X y se le denota con σ_X. La varianza se denota con σ_X^2. Si no hay lugar a confusión se usan simplemente σ y σ^2 para la desviación estándar y varianza, respectivamente.

Como en el caso de las variables aleatorias discretas, se considera que los valores de la variable que están fuera del intervalo $[\mu - 2\sigma, \mu + 2\sigma]$ serán considerados "infrecuentes o poco comunes".

6.4 Algunos modelos probabilísticos para variables aleatorias continuas

Para modelar el patrón de variabilidad que tienen los valores de una variable aleatoria continua existe una serie de distribuciones teóricas muy conocidas que son utilizadas en muchas situaciones reales. Las distribuciones que a continuación se presentan son algunos de los modelos teóricos más importantes en el estudio de la variabilidad; entre ellas destaca la distribución normal.

La distribución uniforme

Una variable aleatoria continua X tiene ***distribución uniforme*** *en el intervalo* $[a, b]$*, con* $a < b$*, si sus valores se dispersan uniformemente en el intervalo.*

La función de densidad de una variable con estas características es:

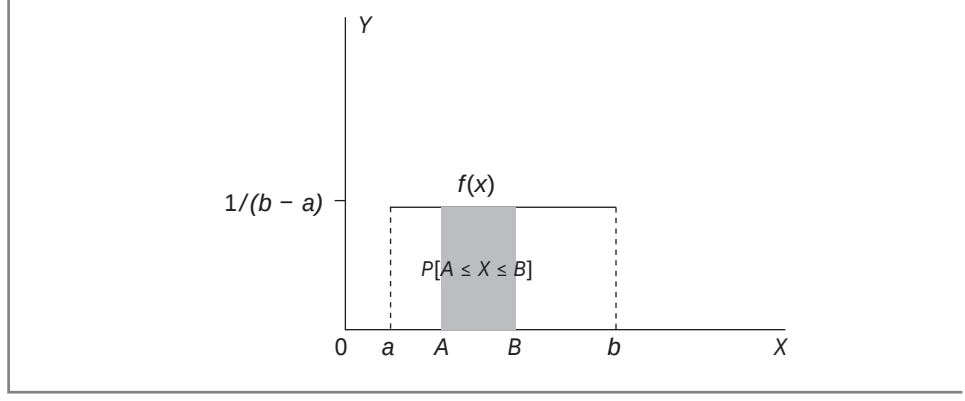
$$f(x) = \begin{cases} 1/(b-a) & \text{si } x \in [a, b] \\ 0 & \text{en el resto} \end{cases}$$

Si X tiene distribución uniforme en $[a, b]$, se escribe $X \sim U[a, b]$.

En general, las probabilidades de que X tome valores en dos intervalos de la misma longitud y contenidos en $[a, b]$ son iguales. En este sentido podemos decir que la masa de valores de X se distribuye de manera

uniforme en [*a, b*]. Esta propiedad sugiere utilizar este modelo cuando los valores de una variable pueden variar con igual posibilidad en un intervalo.

FIGURA 6.7 *Función de densidad de X*



Muchos casos reales pueden modelarse con la distribución uniforme. Presentamos a continuación los siguientes ejemplos.

EJEMPLO. *Números random*

Muchas de las calculadoras tienen un programa que proporciona "números random". Estos números tienen distribución aproximadamente uniforme en el intervalo [0, 1].

EJEMPLO. *Salida de trenes*

Entre las 7 a. m. y las 8 a. m., los trenes parten de una estación a los 20, 30 y 60 minutos después de las 7 a. m. Si la hora en que llega una persona a la estación sigue una distribución uniforme en el intervalo comprendido entre las 7 a. m. y las 8 a. m., ¿cuál es la probabilidad de que tenga que esperar a lo más 5 minutos la salida de un tren?

Solución

La persona espera a lo más 5 minutos en la estación si llega entre las 7:15 y 7:20, o entre las 7:25 y 7:30, o entre las 7:55 y 8:00 a. m.

Considerando que la distribución del tiempo *X* de llegada es uniforme, la probabilidad de que la persona llegue en cada uno de los intervalos de tiempo indicados es 5/60. La probabilidad de que tenga que esperar a lo más 5 minutos es $3\,\frac{5}{60} = 1/4$.

Esperanza y varianza de una variable con distribución uniforme

Para una variable aleatoria X que tiene distribución uniforme en el intervalo [a, b], se cumple:

$$E(X) = \frac{a+b}{2} \quad \text{y} \quad V(X) = \frac{(b-a)^2}{12}$$

La distribución normal

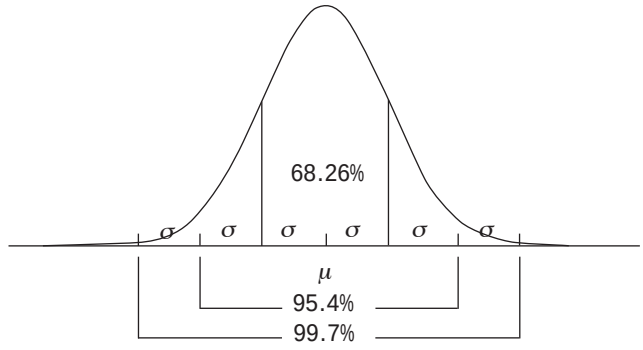
Esta distribución, conocida inicialmente como *la curva normal de errores,* fue introducida por el matemático francés Abraham De Moivre (1667–1754) y estudiada luego por el matemático francés Pierre Simón Laplace (1749–1827) y el matemático alemán Carl Friedrich Gauss (1777–1855). Gauss llegó a este modelo estudiando la distribución de ciertos errores de medición. En honor a estos matemáticos, la distribución también se conoce como de Moivre-Laplace-Gauss. La distribución normal es uno de los modelos probabilísticos más importante y el más utilizado en estadística. Diversos fenómenos aleatorios que se estudian en la realidad se pueden modelar con esta distribución. Las propiedades matemáticas que este modelo posee han contribuido al desarrollo de la estadística; sin embargo, estas son extendidas sin ningún reparo a cualquier situación, lo que ha contribuido también a un sinnúmero de errores en la utilización de este modelo.

Los resultados de las experiencias aleatorias cuyos histogramas presentan la forma de una campana (Figura 6.8) pueden ser descritos con la distribución normal.

Se dice que la variable X sigue una ***distribución normal*** *con parámetros* μ *y* σ *y se escribe* $X \sim N(\mu, \sigma^2)$ *si su función de densidad es:*

$$f(x) = \frac{1}{\sigma\sqrt{\alpha\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad \text{con } -\infty < \mu < +\infty,\ \sigma > 0$$

FIGURA 6.8 *Función de densidad de una variable con distribución normal*



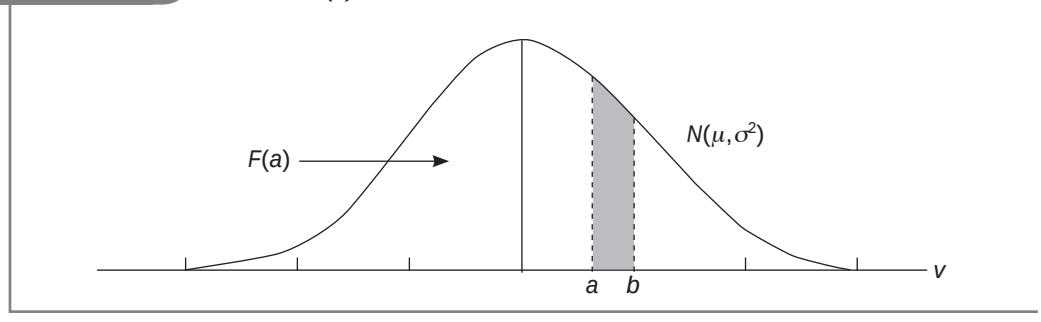
Las siguientes propiedades caracterizan a la distribución normal.

- La gráfica de *f* es simétrica respecto de la recta $x = \mu$ y se denomina *campana de Gauss.*
- El área bajo la curva $f(x)$ y por encima del eje X es 1.
- El área comprendida entre las rectas $x = a$ y $x = b$, que corresponde a la probabilidad $P[a \leq X \leq b]$, se calcula usando los métodos de integración; sin embargo, existen tablas, como la del apéndice A, que permiten su cálculo.
- La distancia del punto de inflexión (en donde cambia la concavidad de la curva) a la recta $x = \mu$ es igual a σ.
- La ordenada que corresponde a la abscisa μ está a una altura $1/\sigma \quad 2\pi$. Cuando σ es cada vez más pequeño la curva es más leptocúrtica, en cambio, cuando es cada vez más grande la curva es más platicúrtica.

A partir de la función de densidad se puede obtener la *función de acumulación* $F(x)$, que corresponde a la probabilidad de que la variable tome valores menores o iguales que *x.*

El valor de $F(a)$ corresponde al área bajo la curva de la función densidad, por encima del eje X y que está a la izquierda de la recta $x = a$.

FIGURA 6.9 *Valor de F(a)*



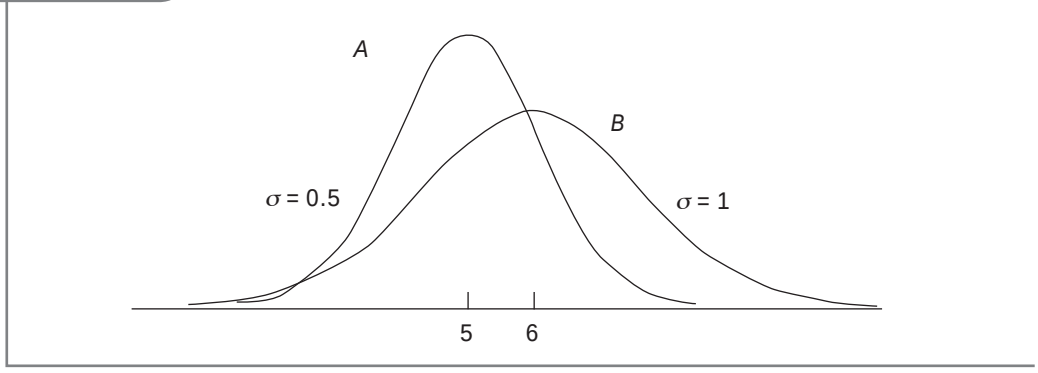
Esperanza y varianza de una variable aleatoria con distribución normal

Usando la función de densidad de una variable aleatoria X con distribución normal de parámetros μ y σ, se prueba que:

$$E(X) = \mu \quad \text{y} \quad V(X) = \sigma^2$$

La función de densidad de la distribución normal queda determinada conociendo la esperanza y la varianza de la variable. La Figura 6.10 representa las distribuciones del tiempo que dura el proceso de fabricación de un cierto artículo cuando se sigue el método A y el método B. La media y la desviación estándar para A son: 5 horas y 0.5 horas, respectivamente. Para B la media y la desviación estándar son: 6 horas y 1 hora, respectivamente.

FIGURA 6.10 *Distribución de dos procesos de fabricación*

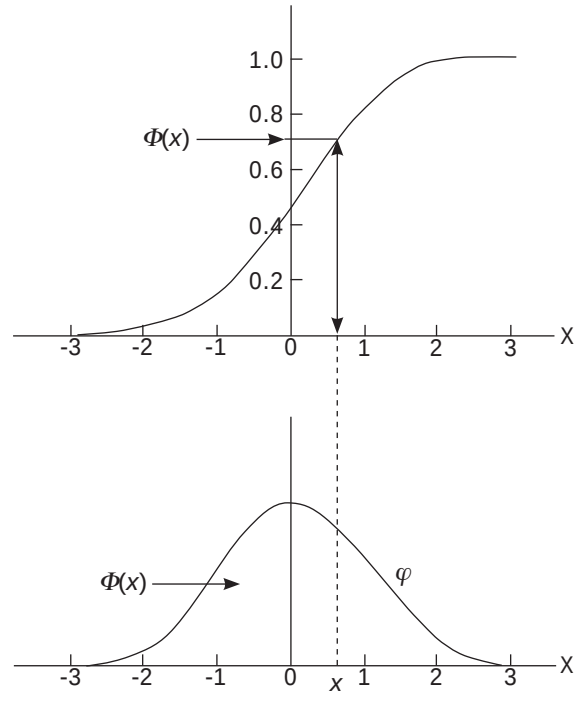


Otras de las características importantes de la distribución normal son las medidas de simetría y de curtosis. Se demuestra que para la distribución normal estándar, los índices de Fisher, que indican la simetría y el achatamiento, son iguales a 0. Precisamente estos índices sirven como referencia para estudiar la forma de otras distribuciones.

Cuando la esperanza es $\mu = 0$ y la desviación estándar $\sigma = 1$, la distribución $N(0, 1)$ se llama distribución *normal estándar o típica*.

Una variable con distribución *normal estándar* se denota con Z y sus funciones de densidad y acumulación, con φ y Φ, respectivamente.

FIGURA 6.11 *Función de acumulación Φ y función de densidad φ*



Usando la tabla del apéndice A se puede comprobar la regla 68-95-99:

- El área comprendida entre $\mu - \sigma$ y $\mu + \sigma$ es aproximadamente el 68.26% del área total.
- El área comprendida entre $\mu - 2\sigma$ y $\mu + 2\sigma$ es aproximadamente el 95.4% del área total.
- El área comprendida entre $\mu - 3\sigma$ y $\mu + 3\sigma$ es aproximadamente el 99.7% del área total.

Si se tiene un conjunto de datos que provienen de una distribución normal con desviación estándar desconocida, entonces se puede estimar la desviación estándar con $\hat{\sigma}$ = *Rango*/4.

EJEMPLO. *Modelando los errores de medición*

Un objeto fue pesado cien veces en una balanza. Los resultados fueron anotados con tres cifras significativas y fueron los siguientes:

19.5 20.0 20.5 20.9 20.8 21.2 21.3 21.5 22.1 22.1 22.3 22.4 22.5 22.5 22.6 22.8 22.8 23.0 23.1 23.1 23.2 23.2 23.3 23.3 23.3 23.6 23.7 23.8 23.8 23.9 23.9 23.9 24.0 24.0 24.0 24.0 24.0 24.1 24.1 24.1 24.3 24.5 24.5 24.7 24.7 24.7 24.8 24.8 24.8 24.9 24.9 25.0 25.0 25.1 25.1 25.1 25.2 25.4 25.5 25.5 25.7 25.7 25.7 25.8 25.8 25.8 25.8 25.9 25.9 26.1 26.2 26.2 26.3 26.3 26.4 26.5 26.6 26.6 26.6 26.7 26.8 26.8 26.9 27.0 27.1 27.1 27.1 27.4 27.5 27.5 27.6 27.7 27.7 28.0 28.6 28.6 28.7 29.0 29.4 29.7 30.8

La media de las mediciones es 25.0.

Los errores calculados, respecto de la media, son: −5.5 −5.0 −4.5,..., 3.6.

El histograma de los errores aparece a continuación. Una curva "normal" se ajustó a la poligonal suavizada. La curva normal ajustada correspondió a una distribución normal de media 0 y desviación estándar 2.19.

FIGURA 6.12 *Histograma para la normal de media* 0 *y d. s.* = 2.19

Las propiedades de la normal indican que:

El 68.26% de los errores se encuentra en el intervalo [−(1)(2.19), (1)(2.19)], aproximadamente.

El 95.40% de los errores se encuentra en el intervalo [−(2)(2.19), (2)(2.19)], aproximadamente.

El 99.70% de los errores se encuentra en el intervalo [−(3)(2.19), (3)(2.19)], aproximadamente.

PROPIEDAD (de tipificación o de estandarización)

La tabla del apéndice A corresponde a la distribución normal estándar; sin embargo, si se aplica la siguiente propiedad, puede usarse también para cualquier otra distribución normal.

Si X es una variable con distribución normal con media μ y desviación estándar σ, entonces la variable X estandarizada, $\frac{X-\mu}{\sigma}$, sigue la distribución N(0, 1).

La transformación $\frac{X-\mu}{\sigma}$ se llama de *estandarización* o de *tipificación,* y la variable que resulta suele denotarse con *Z.*

La estandarización de *X* equivale a expresar las desviaciones de *X* respecto de la media μ en unidades de su desviación estándar.

EJEMPLO. *Cálculos*

Para una variable aleatoria normal estándar *Z*, hallar:

a) $P[Z \leq 2.57]$

b) $P[0 \leq Z \leq 2]$

c) $P[0 \leq Z \leq 2.57]$

d) $P[-1 \leq Z \leq 1]$

e) $P[-2 \leq Z \leq 2]$

f) $P[-3 \leq Z \leq 3]$

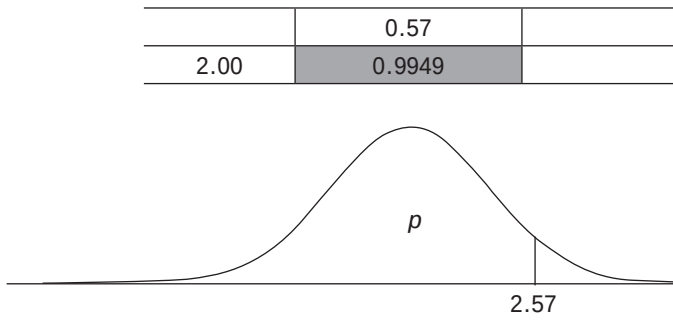
g) $P[-2 < Z < 1]$

h) Hallar el valor *z* de *Z* para el cual el valor de la función de acumulación es igual a 0.9279.

Solución

a) $P[Z \leq 2.57]$ equivale al valor de la función de acumulación en 2.57, $\Phi(2.57)$. Según la tabla del apéndice A, el valor de la función de acumulación es $\Phi(2.57) = 0.9949$.

FIGURA 6.13 $p = \Phi(2.57) = 0.9949$



Usando la tabla del apéndice A se tiene también:

b) $P[0 \le Z \le 2] = \Phi(2) - \Phi(0) = 0.9772 - 0.50 = 0.4772$

c) $P[0 \le Z \le 2.57] = \Phi(2.57) - \Phi(0) = 0.9949 - 0.50 = 0.4949$

d) Usando la simetría de la función de densidad de la normal, se tiene:

$P[-1 \le Z \le 1] = 2P[0 \le Z \le 1] = 2(\Phi(1) - \Phi(0)) = 2(0.8413 - 0.50) = 0.6826$

e) $P[-2 \le Z \le 2] = 2(\Phi(2) - \Phi(0)) = 0.9544$

f) $P[-3 \le Z \le 3] = 2(\Phi(3) - \Phi(0)) = 2(0.4987) = 0.9974$

g) $P[-2 < Z < 1] = \Phi(1) - \Phi(-2) = 0.8413 - [1 - \Phi(2)] = 0.8185$

h) Se trata de hallar el valor z para el cual $\Phi(z) = 0.9279$; hacer lo siguiente:

En el cuerpo de la tabla del anexo A, ubicar el valor del área más cercana a 0.9272, para luego identificar la puntuación z correspondiente. Esta puntuación es 1.46.

EJEMPLO. *Cálculos*

Para una variable aleatoria X cuya distribución es normal con media 12 y desviación estándar 2:

a) Hallar $P[X \le 15]$.

b) Hallar x_0, para el cual $P[X \le x_0] = 0.7734$.

Solución

a) Para calcular la probabilidad $P[X \le 15]$, se "estandariza" primero la relación $X \le 15$. Haciendo la operación de estandarización se tiene la relación equivalente

$$\frac{X - 12}{2} \le \frac{15 - 12}{2}$$

Ahora se puede escribir $P\left[\frac{X - 12}{2} \le \frac{15 - 12}{2}\right] = P[Z \le 1.5]$ (recordar que $\frac{X - 12}{2}$ es una variable normal estándar Z).

Entrando a la tabla de la normal estándar, se tiene que $P[Z \le 1.5] = 0.9332$.

Luego, $P[X \le 15] = P\left[\frac{X - 12}{2} \le \frac{15 - 12}{2}\right] = P[Z \le 1.5] = 0.9332$.

b) Para entrar a la tabla A, primero se estandariza la expresión $X \le x_0$:

$$\frac{X - 12}{2} \le \frac{x_0 - 12}{2}$$

El problema equivale a encontrar x_0 para el que se cumple:

$$p\left[\frac{X - 12}{2} \le \frac{x_0 - 12}{2}\right] = P\left[Z \le \frac{x_0 - 12}{2}\right] = 0.9332$$

Entrando en la tabla resulta que $\frac{x_0 - 12}{2} = 1.50$. Despejando se tiene:

$$x_0 = 2(1.50) + 12 = 15$$

EJEMPLO. *Tiempos de atención a los clientes*

Con la finalidad de obtener mejoras en el proceso de atención a los clientes y después de observar el histograma de un conjunto de datos que representan los tiempos, en minutos, que demoran los clientes en realizar trámites en una sección de las oficinas de un ministerio, se ha convenido ajustar una curva normal de media 12 y varianza 4. Usando este ajuste es posible, por ejemplo, conocer cuál es el porcentaje de clientes que se espera que demoren entre 10 y 14 minutos en realizar los trámites en las oficinas del ministerio.

En efecto, si X denota a la variable aleatoria que representa los tiempos de atención, se tendrá, según lo indicado, que $X \sim N(12, 4)$.

Aplicando la propiedad de estandarización y luego la tabla del apéndice A, se tiene:

$P[10 \le X \le 14] = P[(10 - \mu)/\sigma \le (X - \mu)/\sigma \le (14 - \mu)/\sigma] = P[-1 \le Z \le 1] =$
$\Phi(1) - \Phi(-1) = \Phi(1) - \{1 - \Phi(1)\} = 2\Phi(1) - 1 = 0.6826$

Se espera entonces que el 68.26% de los tiempos de servicios duren entre 10 y 14 minutos.

Usando la tipificación, se comprueba ahora que si X es una variable aleatoria con distribución normal de media μ y varianza σ^2, se cumple:

$$P[\mu - \sigma \le X \le \mu + \sigma] = P[-\sigma \le X - \mu \le \sigma] = P\left[-1 \le \frac{X - x}{\sigma} \le 1\right] = P[-1 \le Z \le 1] = 0.6826,$$

$P[\mu - 2\sigma \le X \le \mu + 2\sigma] = 0.9544$ y $P[\mu - 3\sigma \le X \le \mu + 3\sigma] = 0.9974$

EJEMPLO. *Seguros*

Una empresa de seguros ha determinado que los montos X pagados por siniestros en un mes determinado y para cierto tipo de seguros se pueden modelar con distribución normal de media de 500 dólares y desviación estándar de 50 dólares. De acuerdo a ello, la probabilidad de que el valor de un siniestro sea mayor que 550 dólares es igual a:

$$P[X > 550] = P\left[Z \ge \frac{550 - 500}{\sigma}\right] = 0.8413$$

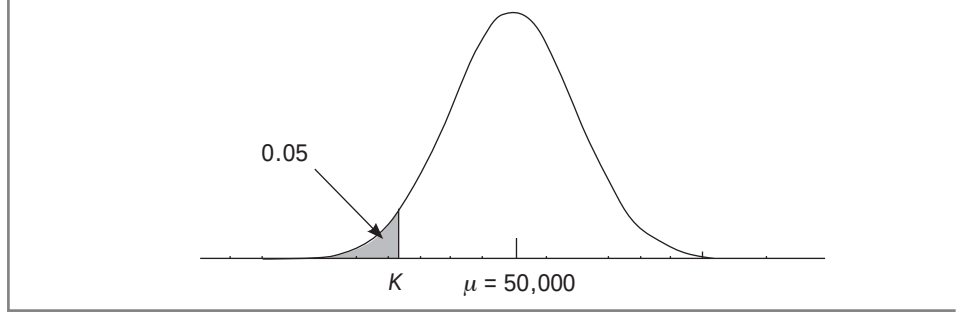
Se espera que el 84.13% de los pagos por siniestro sean mayores que 550 dólares.

EJEMPLO. *Garantía otorgada*

La compañía GMS acaba de fabricar un nuevo tipo de neumáticos para automóviles y desea promocionarlos para la venta, indicando una garantía que permita la devolución del costo del neumático si dura menos de *K* kilómetros, y de tal manera que esta se dé a lo más al 5% de los clientes que han adquirido el neumático. Para determinar el valor de *K,* los técnicos de la compañía han realizado una serie de pruebas y han determinado que la duración en kilómetros puede ajustarse con una distribución normal con un promedio de 50,000 y una desviación estándar de 2,000.

De acuerdo a lo indicado y si se escribe *X* = *duración, en kilómetros,* se tiene que el valor *K* debe cumplir con la siguiente exigencia: $P[X \leq K] = 0.05$.

FIGURA 6.14 *Distribución de la duración de los neumáticos*



Aplicando la estandarización en la relación, se tiene:

$$P\left[\frac{X - 50{,}000}{2{,}000} \leq \frac{K - 50{,}000}{2{,}000}\right] = 0.05$$

Considerando que $\frac{X - 50{,}000}{2{,}000}$ tiene distribución normal estándar y usando la tabla del apéndice A, se tiene que $\frac{K - 50{,}000}{2{,}000} = -1.645$. Despejando, resulta:

$K = 50{,}000 + (-1.645)\ (2{,}000) = 46{,}710$.

La garantía que deberá otorgar la empresa GMS será de 46,710 kilómetros.

EJEMPLO. *Porcentaje de artículos fuera de especificaciones*

En un proceso industrial que permite la fabricación de láminas de acero, se ha determinado que su grosor debería tener la siguiente especificación: 10.50 mm ± 0.10. Una manera de evaluar si se cumple esta especificación es usar los datos de una muestra y calcular el porcentaje de planchas cuyo grosor está entre 10.40 mm y 10.60 mm.

Si la forma del histograma permite indicar que una distribución normal puede ser ajustada a los datos que produce el proceso, será fácil calcular el porcentaje de las planchas producidas que satisfacen las especificaciones. Para ello será preciso estimar la media y la desviación estándar del proceso. La media del proceso se estima con la media de los datos, ($\overline{x}$), mientras la desviación estándar se estima con $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$.

Si Y es la variable que representa a las mediciones y esta tiene la distribución normal, la probabilidad de que una plancha cumpla con las especificaciones estará dada por:

$$p = P(10.40 \leq Y \leq 10.60) = P\left(\frac{10.40 - \overline{x}}{s} \leq Z \leq \frac{10.60 - \overline{x}}{s}\right)$$

donde Z es la variable con distribución normal estándar.

Si el proceso no tiene cambios en el futuro, se espera que el porcentaje aproximado de las planchas que cumplan las especificaciones sea igual a $p \times 100\%$.

Aproximación de la distribución binomial mediante la distribución normal

El llamado *teorema del límite central* permite la aproximación de la distribución binomial con parámetros n y p mediante la distribución normal cuando $np \geq 5$ y $nq \geq 5$.

Supondremos que se tiene la variable X con distribución binomial de parámetros n y p y se desea calcular la $P(X = k)$.

Se procede de la siguiente manera.

1. Como X es una variable aleatoria discreta, se hace la *corrección por continuidad*:
$$P(X = k) \approx P(k - 0.5 \leq X \leq k + 0.5)$$
2. Usando el valor esperado (np) y la desviación estándar ($\sqrt{npq}$) de X, se estandariza ($k - 0.5 \leq X \leq k + 0.5$) y así se tiene:

$$P(X = k) \approx P(k - 0.5 \leq X \leq k + 0.5) = P\left(\frac{k - 0.5 - np}{\sqrt{npq}} \leq \frac{X - np}{\sqrt{npq}} \leq \frac{k + 0.5 - np}{\sqrt{npq}}\right)$$

El *teorema del límite central*, que revisaremos más adelante, permite usar la distribución normal estándar a la distribución binomial estandarizada. Así se tiene la aproximación:

$$P(X = k) \approx P(k - 0.5 \leq X \leq k + 0.5) = P\left(\frac{k - 0.5 - np}{\sqrt{npq}} \leq \frac{X - np}{\sqrt{npq}} \leq \frac{k + 0.5 - np}{\sqrt{npq}}\right) =$$

$$P\left(\frac{k - 0.5 - np}{\sqrt{npq}} \le Z \le \frac{k + 0.5 - np}{\sqrt{npq}}\right)$$

Aplicando el mismo procedimiento se calculan las probabilidades: $P(a \le X)$, $P(a \le X \le b)$, $P(X \le b)$ no sin antes hacer las correspondientes correcciones por continuidad, respectivas:

$P(a \le X) \approx P(a - 0.5 \le X)$

$P(a \le X \le b) \approx P(a - 0.5 \le X \le b + 0.5)$

$P(X \le b) \approx P(X \le b + 0.5)$

EJEMPLO. *Aproximación de la binomial*

Si X es una variable aleatoria con distribución binomial con parámetros $n = 20$ y $p = 0.5$, entonces:

$$P(X = 7) \approx P\left(\frac{7 - 0.5 - 10}{\sqrt{5}} \le Z \le \frac{7 + 0.5 - 10}{\sqrt{5}}\right)$$

Recurriendo a la tabla de la normal estándar se tiene el resultado 0.0730. Calculando directamente con la binomial el resultado es 0.0739.

Propiedades de la distribución normal

A continuación, se enuncian algunas propiedades de la distribución normal, que son muy útiles en los cálculos a realizar.

1. *Si X tiene distribución normal con media μ y varianza σ^2, entonces la variable $Y = a + bX$ es una variable aleatoria que sigue la distribución normal con media $a + b\mu$ y varianza $b^2\sigma^2$.*

   En particular, si Z es una variable aleatoria normal estándar, entonces la variable $Y = \mu + \sigma^2 Z$ tiene distribución normal con media μ y varianza σ^2.

2. **"Propiedad reproductiva de la normal"**

   Si $X_i \sim N(\mu_i, \sigma_i^2)$, con $i = 1, 2, \ldots, n$, son n variables aleatorias independientes, entonces:

$$Y = \sum_{i=1}^{n} c_i X_i \sim N\left(\sum_{i=1}^{n} c_i \mu_i, \sum_{i=1}^{n} c_i^2 \sigma_i^2\right)$$

En particular se cumple que:

a) si X e Y son variables aleatorias independientes con distribuciones $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$, respectivamente, entonces la variable $U = X + Y$ tiene distribución $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

b) si $X_1, \ldots, X_n$ son variables aleatorias independientes, todas con la misma distribución $N(\mu, \sigma^2)$, entonces la variable aleatoria $\sum_{i=1}^{n} X_i$ tiene distribución normal $N(n\mu, n\sigma^2)$.

EJEMPLO. *Los gastos por cliente*

Se supone que la distribución de la cantidad de dinero, en dólares, que cada cliente, de manera independiente, gasta cada día en una tienda es $N(25, 0.4)$. ¿Cuál es la probabilidad de que 20 clientes de la tienda gasten más de 505 dólares en cada día?

Solución

Si denotamos, respectivamente, con X_1, X_2,..., X_{20} a lo que gasta cada cliente, se tendrá que la cantidad de dinero que los 20 clientes gastan es $\sum_{i=1}^{20} X_i$. Esta variable, de acuerdo a la propiedad b), tiene distribución normal con media 25(20) = 500 y varianza 20(0.4) = 8, y así, la probabilidad de que 20 clientes gasten más de 505 por día es:

$$P\left[\sum_{i=1}^{20} X_i > 505\right] = 1 - \Phi\left(\frac{505 - 500}{\sqrt{8}}\right) = 1 - \Phi(1.77) = 0.034$$

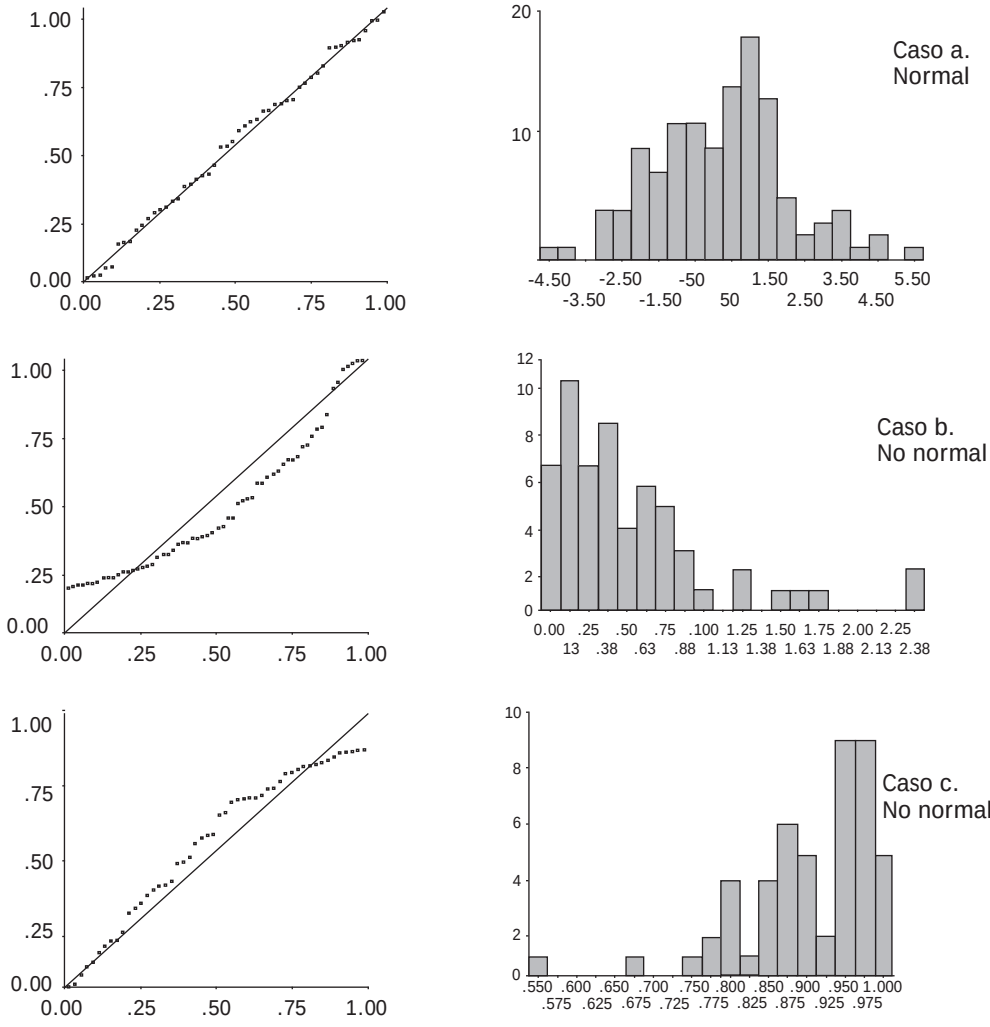
Ajuste de un modelo probabilístico normal a un conjunto de datos

En muchas ocasiones es necesario comprobar previamente si el conjunto de valores aleatorios $(x_1,..., x_n)$ con los cuales se trabaja provienen de una distribución normal. La forma básica (pero no estricta) de comprobación se obtiene observando el histograma o el diagrama de tallo y hojas. Si la suavización del polígono de frecuencias del histograma se acerca a la campana de Gauss, se puede considerar que el grupo de datos proviene de una distribución normal. Este procedimiento es útil cuando el número de observaciones es suficientemente grande, pero es difícil llegar a una conclusión semejante cuando el número de datos es pequeño. Si los datos no son muchos, se usan las *gráficas de probabilidad p – p*. Con estas gráficas se comparan las observaciones, previamente ordenadas, $x_{(1)}, x_{(2)},..., x_n)$ con los percentiles teóricos p_i de la normal. Si los datos provienen de una variable X con distribución normal, estos percentiles deben ser aproximadamente iguales a los percentiles $x_{(i)}$. Por ello, si los datos $x_1,..., x_n$ provienen de una distribución normal, los puntos cuyas coordenadas son $x_{(i)}$ y $p_{(i)}$ deben estar aproximadamente en la diagonal del plano cartesiano. Básicamente, la prueba es visual; se requiere de alguna experiencia y tener en cuenta algunas recomendaciones, como: poner énfasis en que los puntos estén alineados en los puntos centrales (entre los percentiles 33 y 67 aproximadamente).

Otra manera más precisa de comprobar si cierto conjunto de valores provienen de una distribución normal se realiza haciendo uso de pruebas inferenciales, como la prueba muy conocida de Kolmogorov Smirnov.

Los diagramas *p – p* y los histogramas que aparecen a continuación corresponden a un grupo de datos que provienen de una distribución normal (Figura 6.15a) y de distribuciones no normales (figuras 6.15b y 6.15c).

FIGURA 6.15 *Los gráficos de la izquierda corresponden a los diagramas p – p*



APLICACIÓN: El VaR (valor en riesgo)

El *VaR* o valor en riesgo es una de las aplicaciones de la distribución normal que se usa a menudo en el sector financiero. Para introducir este útil concepto, suponga que usted es el administrador de un activo cuyo valor actual es A_0, y cuya rentabilidad R_f al fin del próximo periodo (un periodo puede ser por ejemplo un mes) tiene un valor esperado $E(R_f)$ y su varianza es $V(R_f)$.

Si además se supone que la distribución de la rentabilidad final es normal, se tendrá que el valor final del activo A_f tambien tiene distribución normal con valor esperado y varianza

$E(A_f) = E(A_0(1 + R_f)) = A_0(1 + E(R_f))$ y $V(A_f) = V(A_0(1 + R_f)) = (A_0)^2V(R_f)$, *respectivamente.*

Si se denota con A_c al menor valor que puede tomar A_f con probabilidad $1 - \alpha$, el valor en riesgo *VaR* en un periodo es:

$$VaR = A_0 - A_c$$

Es decir, el valor en riesgo VaR es el mayor valor que se puede perder en un periodo, con probabilidad $1 - \alpha$.

¿Cuál es el valor en riesgo de un portafolio de 10 millones de dólares al final del mes si se supone que la rentabilidad mensual es normal con media 5% y varianza 10%?

La distribución exponencial

Un modelo teórico muy útil en el estudio del tiempo que transcurre entre dos eventos aleatorios es la *distribución exponencial.* Esta distribución se usa, por ejemplo, para analizar la variabilidad del tiempo transcurrido entre dos fallas consecutivas de un aparato electrónico, del tiempo transcurrido entre dos llegadas de clientes a un banco, etcétera.

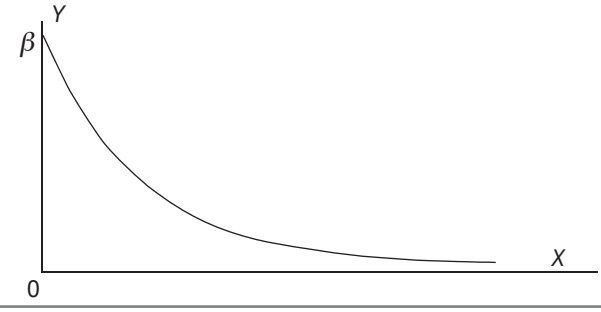
Una variable aleatoria continua X tiene distribución ***exponencial con parámetro*** β si su función de densidad está dada por:

$$f(x) = \beta e^{-x\beta} \quad x \geq 0, \; e \approx 2.7183$$

Se denota $X \sim Exp(\beta)$ cuando una variable tiene distribución exponencial con parámetro β.

La gráfica de la función de densidad aparece a continuación.

FIGURA 6.16 *Función de densidad exponencial con parámetro* β



Esperanza y varianza de una variable aleatoria con distribución exponencial

La esperanza y la varianza de una variable aleatoria X son, respectivamente:

$$E(X) = 1/\beta \quad y \quad V(X) = 1/\beta^2$$

Función de acumulación de una variable aleatoria con distribución exponencial

Se demuestra que la función de acumulación de una variable aleatoria que tiene distribución exponencial de parámetro β está definida por:

$$F(x) = P(X \leq x) = 1 - e^{-\beta x} \quad con \quad x \geq 0$$

Usando esta función se tiene que la probabilidad de que los valores de una variable X con distribución exponencial de parámetro β estén entre a y b es:

$$P[a \leq X \leq b] = F(b) - F(a) = e^{-\beta a} - e^{-\beta b}$$

EJEMPLO. *Modelando el tiempo de vida de una lámpara*

El tiempo de vida en horas, X, de un cierto tipo de lámparas tiene función de densidad:

$$f(x) = (1/100)e^{-x/100} \text{ con } x \geq 0$$

¿Cuál es la probabilidad de que una lámpara escogida al azar no tenga que ser sustituida durante las primeras 150 horas de uso?

Solución

Para que una lámpara no se sustituya durante las primeras 150 horas, su tiempo de vida X debe ser mayor que 50.

La distribución de X es exponencial con parámetro $\beta = 1/100$; por lo tanto, la probabilidad de que una lámpara no sea sustituida durante las primeras 150 horas es:

$$P[X > 150] = 1 - P[X \leq 150] = 1 - F(150) = 1 - [1 - e^{-(1/100)(150)}] = 0.2231$$

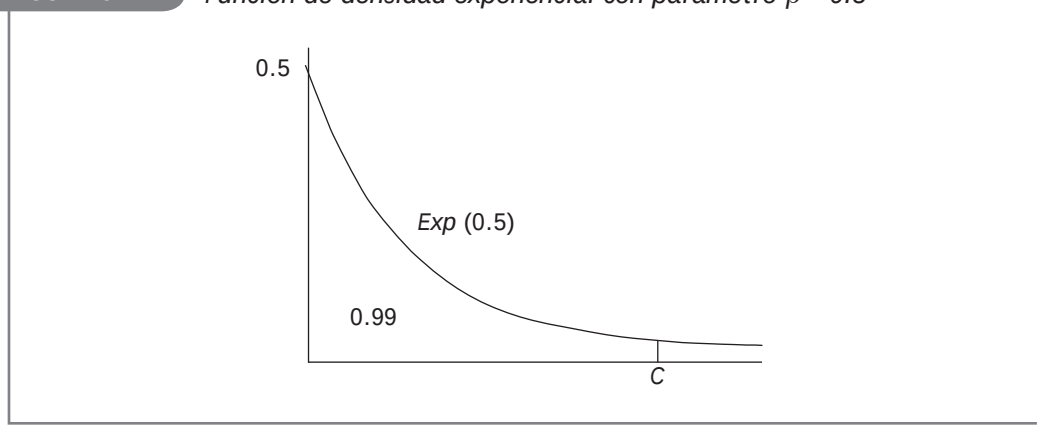
EJEMPLO. *Inventario de gasolina en una estación de servicio*

Una estación de suministro de gasolina recibe gasolina una vez por semana. Si su volumen semanal de ventas V, en miles de galones, se distribuye exponencialmente con parámetro $\beta = 0.5$, calcular la capacidad que debe tener el depósito de la estación si el distribuidor desea que, con una garantía del 99%, exista gasolina en el fin de semana.

Solución

Si C es la capacidad del tanque, debe cumplirse $P[V < C] = 0.99$.

FIGURA 6.17 *Función de densidad exponencial con parámetro $\beta = 0.5$*



Luego, $F(C) = 1 - e^{-0.5C} = 0.99$. Tomando logaritmo y despejando el valor de C, se tiene que $C = 9.2103$.

El tanque debe tener una capacidad de 9.2103 miles de galones para que, con probabilidad 0.99, exista gasolina en el fin de semana.

Propiedad. La distribución exponencial y la distribución de Poisson

Una propiedad muy útil por las diferentes aplicaciones indica que si el número de eventos que ocurren en un intervalo [*a, b*] *tiene distribución de Poisson con tasa* λ *en este intervalo, entonces el tiempo que transcurre entre dos eventos consecutivos tiene distribución exponencial con parámetro* λ (*con media 1/*λ*).*

EJEMPLO. *Servicio de recojo de mineral*

La tasa de llegada de camiones de carga para recojo de mineral en una mina es $\lambda = 10$ por hora. Si se asume que el número de llegadas se modela con la distribución de Poisson, entonces el tiempo T entre llegada y llegada de los camiones es una variable aleatoria que tiene distribución exponencial con parámetro $\lambda = 10$. Es decir, el tiempo promedio entre llegada y llegada de los camiones es:

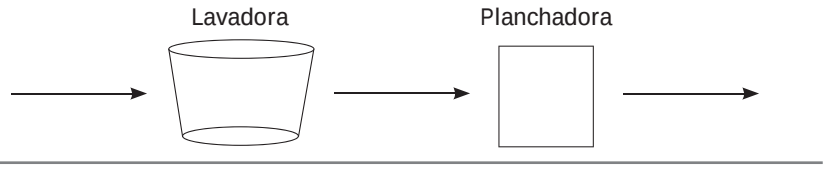
$$\frac{1 \text{ hora}}{10 \text{ camiones}} = 0.1 \text{ hora/camión}$$

La función de distribución del tiempo en horas entre llegada y llegada de los camiones es $F(x) = 1 - e^{-10x}$.

APLICACIÓN: El caso de la lavandería Limpia

El gerente de mantenimiento de la lavandería industrial Limpia quisiera contratar cada cierto tiempo un técnico para que revise la lavadora y la planchadora que conforman el sistema de lavado y planchado. El gerente piensa contratar al técnico cada diez meses; sin embargo, es posible que este tiempo sea inadecuado por ser muy temprano o muy tarde para la buena marcha del sistema formado por la lavadora y la planchadora. La empresa que vende las máquinas ha informado que la probabilidad de que lavadora dure a lo más t meses sin fallar está dada por $F(t) = 1 - e^{-\frac{1}{10}t}$, mientras que para la planchadora esta probabilidad está dada por $F(t) = 1 - e^{-\frac{1}{12}t}$.

FIGURA 6.18



Indicar un procedimiento (no es necesario que usted realice operaciones de cálculo) que permita recomendar al gerente de mantenimiento si es adecuado contratar cada seis meses al técnico para la revisión de las dos máquinas, considerando que estas funcionan de manera independiente.

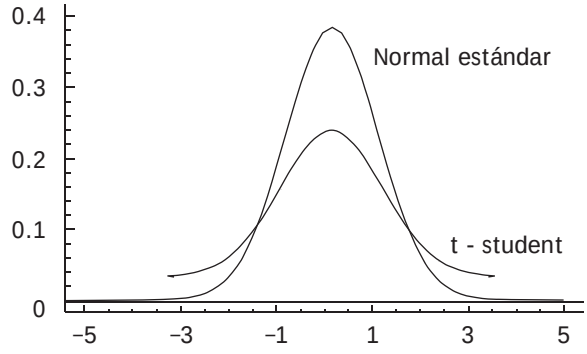
Otros modelos teóricos

Otros modelos teóricos que son usados en el desarrollo de la estadística inferencial son:

a) La distribución *t-student con n grados de libertad,* $t(n - 1)$.

Esta distribución se asemeja a la distribución normal estándar, salvo en las colas. Las colas en la *t-student* son "más pesadas" que las de la normal estándar. Son "más pesadas" en el sentido de que tienen mayor área y por lo tanto hay mayor probabilidad de encontrar valores extremos. *Se demuestra que si el número de grados de libertad crece indefinidamente la distribución t-student tiende a la distribución normal estándar.*

FIGURA 6.19 *Funciones de densidad normal estándar y t-student*



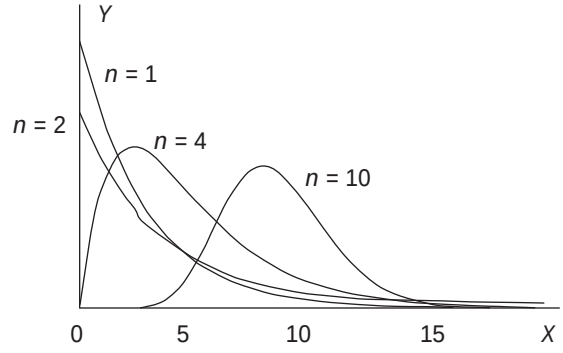
En la práctica se considera que cuando el número de grados de libertad *n* es mayor o igual a 30 la distribución *t-student* se aproxima muy bien a la distribución normal estándar.

En la tabla del apéndice B se encuentra tabulada la distribución *t* student con *n* grados de libertad.

b) La distribución *ji cuadrado con n grados de libertad*

Esta distribución es muy importante en el análisis de datos categóricos. La gráfica de las funciones de densidad para diferentes grados de libertad aparece a continuación (Figura 6.20).

FIGURA 6.20 *Gráficos de funciones de densidad ji-cuadrado con n grados de libertad*



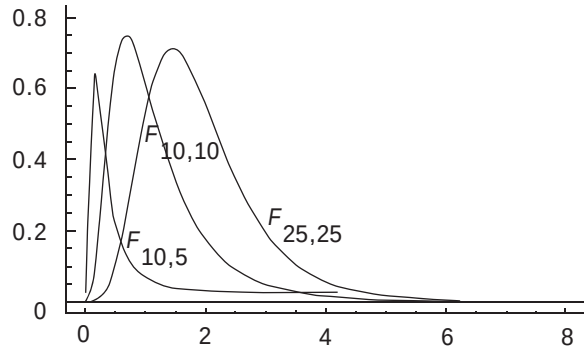
En la tabla del apéndice C se encuentra tabulada la distribución *ji cuadrado con n grados de libertad*.

c) La distribución *F de Snedecor con m, n grados de libertad*

Esta distribución se utiliza en el análisis de los diseños experimentales y del modelo de regresión lineal. En la siguiente gráfica (Figura 6.21) aparecen las funciones de densidad para distintos grados de libertad.

En la tabla del apéndice D se encuentra tabulada la distribución *F* para distintos grados de libertad.

FIGURA 6.21 *Funciones de densidad $F_{10,10}$, $F_{10,5}$ y $F_{25,25}$*



APLICACIÓN: El caso de Taxiseguro

La empresa Taxiseguro, de la cual nos ocupamos en la sección anterior, aplicó su oferta de servicio "gratis por tardanza" respondiendo a la competencia, pero su inventario contable ha comenzado a marcar números rojos (pérdidas) en estos últimos días, por lo que se ha decidido suspender la oferta y llevar a cabo, con ayuda de un profesional de la estadística, un estudio más a profundidad; y es así como se han anotado los tiempos utilizados para llegar de la central de servicios al domicilio del solicitante. El consultor ha indicado que la distribución de los tiempos utilizados tiene una distribución normal con media 12 minutos y una desviación estándar de 3 minutos. Con esta nueva información la empresa Taxiseguro ha determinado que la probabilidad de que el tiempo utilizado en llegar al domicilio del solicitante sea menor que 15 es:

$$P(X \leq 15) = P\left(Z \leq \frac{15 - 12.98}{2.28}\right) = P(Z \leq 0.8859) = 0.8121$$

De esta manera la probabilidad de que Taxiseguro tenga que hacer el servicio gratis es 0.1879.

Taxiseguro ha decidido hacer una revisión de sus servicios, pues, según la probabilidad calculada, se espera hacer 3.758 (20 × 0.1879) servicios gratis por día. Este resultado parece explicar el déficit que Taxiseguro ha comenzado a experimentar.

EJERCICIOS

1. El número X de vehículos vendidos por mes en un centro de ventas es una variable aleatoria con distribución de probabilidad:

x_i	0	2	3	4
p_i	0.1	0.4	c	0.2

   a) Hallar el valor de c.
   b) Hallar la probabilidad de que se vendan al menos dos vehículos.
   c) Hallar la probabilidad de que el número de vehículos vendidos sea menor o igual que 3.
   d) Representar gráficamente la ley de probabilidad de X.
   e) Hallar la ley de probabilidad de $Y = |X - 3|$.

2. Una prueba para medir la satisfacción con su trabajo fue aplicada a 50 empleados del sector textil, los que fueron seleccionados al azar. Los resultados fueron como se indica a continuación:

Puntuaciones de la prueba	*Número de empleados*
1	4
2	8
3	3
4	20
5	15

   a) Construir una distribución de probabilidad para la variable X, que indica las puntuaciones de la prueba, relacionadas con la satisfacción con el trabajo.
   b) Usando a) hallar la probabilidad de que X tenga un valor menor que 3; es decir, la probabilidad de que un empleado del sector textil tenga una puntuación menor que 3.
   c) Si se considera que un empleado estará satisfecho con su trabajo si su puntuación es 4 o 5, hallar la probabilidad de que un empleado del sector textil esté satisfecho con su trabajo.

3. La frecuencia de los pagos que realizó una empresa de seguros por accidentes automovilísticos en el último año fue como sigue:

Pagos en dólares	*Frecuencia*
0	245
1,000	30
2,000	15
4,000	5
6,000	3
8,000	2

   a) A partir de las frecuencias observadas, escribir un modelo de la distribución de probabilidad de la variable X que indica los pagos realizados.
   b) Usando la distribución definida en a), encontrar la "prima" que deberán pagar los clientes si se conviene que estos pagarán el valor esperado de los pagos realizados por la compañía.

4. Invirtiendo $ 10,000 en el fondo 1, el retorno mensual puede ser $ –100 en época de recesión, $ 70 si la economía es estable o $ 250 si la economía está en expansión. Si la misma inversión se realiza en el fondo 2, el retorno mensual puede ser $ –200 o $ 50 o $ 350 si suceden, respectivamente, las condiciones antes indicadas. Si la probabilidad de que suceda una recesión es 0.20, de que suceda una economía estable es 0.50 y de que la economía entre en expansión es 0.30, indicar en cuál de los fondos conviene invertir si como criterio se considera el valor esperado del retorno.

5. Una empresa tiene dos líneas de producción A y B para producir el mismo artículo. De los artículos que produce la línea A, el 1% es defectuoso, mientras que para la línea B el 2% es defectuoso. Si un cliente compra 2 artículos de la empresa, hallar la ley de probabilidad de la variable *X* que indica el número de artículos que resultaron defectuosos, considerando que el 40% de la producción total corresponde a la línea A y el resto a B.

6. La demanda mensual unitaria de un determinado producto en un centro comercial es una variable aleatoria que tiene la siguiente distribución de probabilidad:

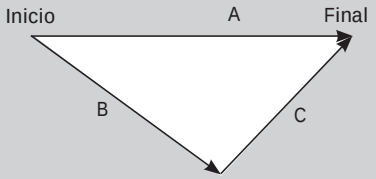
Demanda	*Probabilidad*
200	0.2
300	0.3
400	0.4
500	0.1

Cada unidad demandada produce una utilidad de $ 50; sin embargo, si la unidad al final del mes no es demandada genera una pérdida de $ 20.

a) Hallar la ganancia que se genera si el centro comercial ordena al iniciar el mes 300 unidades.

b) Indicar el número de unidades que debe ordenar el centro comercial al iniciar el mes, de tal manera que la utilidad esperada sea máxima.

7. Un proyecto de construcción de tarjetas de circuitos está compuesto de tres actividades: A, B, C. El orden en que las labores deben ser ejecutadas se presenta en el siguiente diagrama de redes (las líneas representan actividades). Esto es, la actividad B debe ser completada antes de que la C pueda comenzar.



La actividad A no depende, para su inicio, ni de B ni de C (se ejecutan simultáneamente), pero ambas, A y C, deben ser completadas antes de que el proyecto se considere terminado. El tiempo necesario para completar cada actividad es incierto. Sin embargo, se asignan probabilidades a los tiempos de terminación de las actividades, como se indica en la siguiente tabla.

Actividad	*Tiempo posible en semanas para terminar*	*Probabilidad*
A	4	0.50
	6	0.50
B	1	0.25
	3	0.75
C	2	0.80
	4	0.20

Hallar la ley de probabilidad del tiempo, *T*, necesario para terminar el proyecto.

8. José tiene opción de participar en dos tipos de negocios, I y II. Las utilidades para los negocios tipo I son: −4 , 5 y 9 con probabilidades 0.4, 0.2 y 0.4, respectivamente; y para los negocios de tipo II, son: −4, 5 y 9, con probabilidades respectivas de 0.3, 0.6 y 0.1.

   Si José debe participar una sola vez en uno de los negocios, ¿en cuál negocio le recomienda participar?

   Si José desea participar muchas veces en uno de los dos negocios, ¿en cuál negocio le recomienda participar?

9. En un lote hay dos artículos defectuosos y dos no defectuosos. Si los artículos se revisan al azar, uno después de otro, sin restitución y se considera la variable X que indica el número de artículos que se deben revisar a fin de sacar todos los defectuosos, hallar:

   a) La ley de probabilidad de X. b) El valor esperado de X.

10. Invirtiendo en el sector de construcción de viviendas una empresa ganaría $ 7,000,000 con probabilidad 0.70 y perdería $ 2,000,000 con probabilidad 0.30. Invirtiendo en el sector agrícola ganaría $ 6,000,000 si se construye un determinado reservorio, pero si no se construye, perdería $ 2,000,000. ¿Cuál debe ser la probabilidad de realización del reservorio para que a largo plazo sea indiferente el campo de inversión?

11. Una empresa de seguros suscribe una póliza con una compañía minera para pagar la cantidad A en el caso de que en un lapso de 1 año un determinado evento E ocurra. Si se estima que el evento ocurre con probabilidad p, en el lapso de un año, ¿qué porcentaje de A deberá cobrarse como prima al cliente de manera que el valor esperado de la ganancia de la empresa de seguros sea el 10% de A?

12. Una empresa comercializadora de televisores considera que estos tienen algún defecto con probabilidad 0.1. La compañía asegura que cada aparato vendido con algún defecto será reparado sin costo para el cliente. Si se han vendido dos televisores, ¿cuánto espera gastar la compañía en reparaciones si el costo por reparación se expresa como $C = Y^2 + Y + 5$, donde Y es el número de televisores con algún defecto?

13. Una fábrica vende focos en cajas de 50 unidades. La oficina de control de calidad hace una revisión de cada caja adoptando la siguiente política: selecciona 10 focos al azar; si el número de focos defectuosos que se encuentra es 0 o 1, la caja sale a la venta reemplazando previamente los defectuosos con un número igual de focos no defectuosos; de otro modo se revisa toda la caja saliendo al mercado con 0 defectos. ¿Cuál es la probabilidad de que una caja que antes de la revisión tenía 5 focos defectuosos salga al mercado con 0 focos defectuosos?

14. Una empresa petrolera ha sido designada para perforar pozos en cierta área. La probabilidad de que tenga éxito en una prueba es 0.3. Si X es la variable que indica el número de pozos perforados hasta encontrar el primer pozo productivo, hallar la probabilidad de que X tome el valor 5, es decir, la probabilidad de que el primer pozo productivo sea el quinto perforado.

15. El porcentaje de artículos que produce una máquina es 1%. Durante el proceso de producción, el control detiene la máquina cuando se produce el primer artículo defectuoso. ¿Cuál es la distribución de probabilidad de la variable X que indica el número de artículos producidos hasta la fabricación del primer artículo defectuoso?

16. La probabilidad de que un artículo que produce la empresa A sea defectuoso es 0.1. Si se venden 10 de tales artículos sin probar:
    a) Hallar la probabilidad de que los 10 artículos vendidos sean defectuosos.
    b) Si el comprador del artículo regresa las piezas defectuosas para su reparación, y el costo de reparación es $C = 4X + 2$, en donde X es el número de artículos defectuosos, ¿cuál es el costo esperado por reparación?
    c) Si el costo es $C = 4X + X^2 + 2$, ¿cuál es el costo esperado por reparación?

17. La probabilidad de que una aspirina que se hace en un laboratorio tenga algún defecto es 0.01. Las aspirinas se colocan en cajas de 12 pastillas cada una. Hallar la probabilidad de que:
    a) Una caja contenga 2 aspirinas defectuosas.
    b) Un paquete que contiene 6 cajas contenga al menos 2 cajas con 2 aspirinas defectuosas cada una.

18. Una empresa tiene el 45% de sus proyectos en el sector de minas y el resto en agricultura. La probabilidad de que un proyecto del sector de minas se realice es 0.6 y la probabilidad de que un proyecto de agricultura se realice es 0.7. ¿Cuál es la probabilidad que de 6 proyectos sectores, 4 se realicen?

19. Una empresa de turismo vende por adelantado boletos para realizar paseos por la ciudad en pequeños vehículos de 5 asientos; sin embargo, la empresa vende hasta 6 boletos, pues en el 20% de las veces los pasajeros no se presentan en el momento del embarque. Cada boleto cuesta $ 50, y si al presentarse alguno de los pasajeros encuentra el vehículo totalmente lleno la empresa le paga $ 100 (le devuelve el valor del pasaje más una penalidad). ¿Cuánto espera recaudar la empresa por viaje si se supone que cada vez vende 6 boletos?

20. Una empresa de ensamblaje de radios recibe grandes lotes de chips. El fabricante de los chips indica que la probabilidad de que un chip sea defectuoso es 0.01. Si al tomar una muestra de 10 chips la empresa ensambladora encontró 4 chips defectuosos, ¿diría usted que el fabricante dice la verdad? Justifique su respuesta.

21. Por controles anteriores se ha podido determinar que la proporción X de contribuyentes que defraudan al fisco tiene la siguiente distribución de probabilidad.

X	0.02	0.05
p	0.70	0.30

    En una muestra de 10 contribuyentes se encontró un defraudador. A base de este resultado, ¿cómo se modifican las probabilidades "a priori" que se presentan en la tabla?

22. Un pequeño hotel ha tenido amonestaciones de parte de la autoridad respectiva debido a una serie de reclamos de parte de sus clientes. El administrador del hotel era consciente de que los reclamos podían ser: no cumplimiento de reservaciones, mala atención en las oficinas, impuntualidad y otros.

    Con los datos que se disponían se elaboró un gráfico de Pareto y se detectó que el principal reclamo fue el incumplimiento de las reservaciones. Se decidió analizar este problema tomando en cuenta el número de habitaciones que tiene el hotel y la probabilidad de que un cliente haga efectiva la reservación.

    Teniendo en cuenta que el número de habitaciones que tiene el hotel es 20 y que la probabilidad de que un cliente haga efectiva la reserva es 0.9, indicar la probabilidad de que existan reclamos por incumplimiento de reserva si se decide aceptar 23 reservaciones.

23. Un banco recibe en promedio y por día cinco cheques que no tienen fondos. Si se considera que el número de cheques sin fondos que recibe el banco tiene distribución de Poisson, hallar la probabilidad de que en un día determinado reciba:
    a) 0 cheques sin fondos.
    b) Más de dos cheques sin fondos.

24. En una empresa que fabrica tarjetas de video se ha establecido que cada tarjeta que tenga más de 3 defectos no saldrá a la venta. Si en promedio cada tarjeta tiene 2 defectos, ¿qué porcentaje de tarjetas se espera que no salgan a la venta?

25. El número de buses que necesitan reparación cada día en una empresa de transporte sigue una distribución de Poisson con parámetro $\lambda = 6$. Cuando un bus entra en reparación otro bus entra en reemplazo del primero.
    a) ¿Cuál es la probabilidad de que en cualquier día existan uno o más buses en reparación?
    b) ¿Cuántos buses de repuesto es necesario tener si se requiere que la probabilidad de no tener uno de repuesto disponible para reemplazar sea a lo más 0.10?
26. Una compañía de seguros ha determinado que el número de accidentes que registran los automovilistas asegurados en su compañía tiene distribución de Poisson con media 0.5. Si un accidente automovilístico ocurre, el daño al automóvil representa el 20% de su valor de mercado con probabilidad 0.7, el 60% de su valor de mercado con probabilidad 0.25 y una pérdida total con probabilidad 0.05. Hallar el valor de la prima V que deberá cobrar la compañía para un automóvil que vale $ 7,000 para que la ganancia esperada sea 0, considerando que los daños los asume el asegurado a partir del segundo accidente si este ocurriera.
27. El 10% de los clientes que entran a una tienda de zapatos realiza una compra. Hallar, de manera aproximada, la probabilidad de que de un grupo de 40 clientes que ha entrado a la tienda 5 realicen una compra.
28. Se estima que después de una campaña publicitaria el promedio de la demanda de un artículo de tocador, que sigue la ley de Poisson con parámetro 3, será duplicada con probabilidad 0.8 y triplicada con probabilidad 0.2. Hallar la ley de probabilidad de la demanda después de la campaña.
29. Una empresa de seguros de vida ha determinado que dos de cada diez mil personas fallecen anualmente por accidentes de tránsito. Si la empresa tiene contratados 5,000 seguros de vida de esta modalidad y por cada una ha de pagar en un año al menos $ 100,000:
    a) ¿Cuál es la probabilidad de que la empresa tenga que pagar por lo menos $ 12,000,000 en un año?
    b) ¿Qué cantidad de dinero debe mantener en reserva la empresa para tener una probabilidad de 0.95 de poder pagar a todos los familiares de las personas aseguradas fallecidas por accidentes de tránsito?
30. La tasa de llegada de los clientes a un banco es de 2 clientes cada 8 minutos. Usando una distribución adecuada, calcular la probabilidad de tener que esperar más de 4 minutos hasta la llegada del primer cliente.
31. Una compañía petrolera ha sido designada para perforar pozos en cierta área. La probabilidad de que tenga éxito en una prueba es 0.3.
    a) Hallar la probabilidad de que el primer pozo productivo sea el quinto perforado.
    b) Hallar la probabilidad de no encontrar pozo productivo si su capital le permite perforar a lo más siete pozos.
32. Un vendedor de aparatos de TV ofrece una garantía de un año a los que los compren. Durante este tiempo el vendedor se compromete a reemplazar cada aparato, sin costo para el cliente, si este falla por primera vez. El fabricante estima que el tiempo de vida de los aparatos es una variable aleatoria con función de densidad

$$f(x) = \begin{cases} x/4 & \text{si } 0 \le x \le 2 \\ 1 - x/4 & \text{si } 2 \le x \le 4 \\ 0 & \textit{en otro caso.} \end{cases}$$

    en donde X es el tiempo, en años, de vida del aparato.

    ¿A qué porcentaje de los aparatos espera reemplazar el vendedor?

33. La demanda X de gasolina en miles de galones y por día en una refinería tiene función de densidad

$$f(x) = \begin{cases} 2cx, & \text{si } 0 < x \le 1 \\ c(3 - x), & \text{si } 1 \le x \le 3 \\ 0 \textit{ en otro caso.} \end{cases}$$

a) Hallar el valor de c.

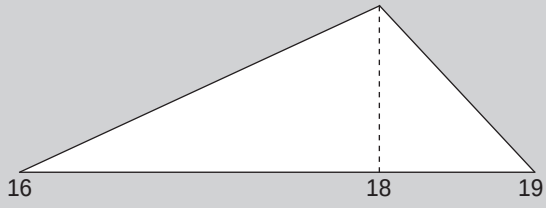
b) Graficar la función de densidad de X.

c) Hallar la probabilidad de que en un día cualquiera la demanda de gasolina sea 1,500 galones o más.

d) Calcular la probabilidad de que la demanda diaria esté entre 800 y 1,500 galones.

f) ¿Cuál debe ser el inventario mínimo que la refinería debe tener para que, con probabilidad 0.95, exista gasolina al final de cualquier día?

34. De acuerdo a su experiencia, el ejecutor de un proyecto piensa que la obra que le han encargado se terminará en el periodo comprendido entre 16 y 19 meses, y que es muy posible que se finalice alrededor de los 18 meses. Por ello considera razonable usar la distribución cuya gráfica aparece a continuación.



La distribución considerada se llama *distribución triangular.*

a) Hallar la altura del triángulo.

b) Hallar la probabilidad de que la obra sea terminada en el periodo comprendido entre 17.5 y 18.5 meses.

c) Hallar la probabilidad de que la obra sea terminada en un periodo menor que 17 meses.

35. Los errores de medición que se realizan con un aparato pueden ser modelados usando una variable X con función de densidad:

$$f(x) = \begin{cases} a + 2ax, & -0.5 \le x \le 0 \\ a - 2ax, & 0 \le x \le 0.5 \\ 0 \textit{ en otro caso} \end{cases}$$

Hallar el valor de a y luego la probabilidad de que el aparato realice un error mayor a 0.2.

36. En un proceso de fabricación de ciertos elementos es necesaria una etapa de secado. Se ha determinado que un modelo para estudiar la probabilidad del tiempo X, en minutos, de secado está determinado por la función de distribución:

$$f(x) = \begin{cases} 0, \textit{ si } x \le 0 \\ \\ 1 - e^{\frac{-x}{512}} \textit{ si } x > 0 \end{cases}$$

Encontrar el porcentaje de elementos que se espera que estén secos después de 10 minutos.

37. Las ventas diarias, en miles de dólares, en una tienda de artefactos electrónicos están distribuidas de la siguiente manera:

Ventas	*Días*
[70 75]	16
]75 80]	17
]80 85]	16
]85 90]	17
]90 95]	16

a) Indicar un modelo (distribución) teórico para estudiar la variabilidad que se podría usar para estudiar la variabilidad de las ventas.
b) Usando el modelo que se indicó en a), hallar la probabilidad de que las ventas sean menores que 82,000 dólares por día.
c) De 5 días tomados al azar, ¿cuál es la probabilidad de que en 3 de ellos las ventas sean menores que 85,500 dólares por día?

38. Se indica que un autobús llegará con "toda seguridad" al lugar de embarque entre las 8:00 y las 9:00 a.m. ¿Qué modelo se sugiere para describir la variable aleatoria que indica el tiempo de llegada? Usando el modelo sugerido, ¿cuál es la probabilidad de que el autobús llegue entre las 8:00 y las 8:10 a.m.?

39. Una llamada telefónica llegó a una central telefónica entre las 8:00 a.m. y las 8:05 a.m. Si la central estuvo ocupada durante dos minutos de ese intervalo, usar un modelo adecuado para indicar la probabilidad de que la llamada haya llegado cuando la central estuvo desocupada.

40. En la industria del petróleo, la temperatura de destilación T, en grados centígrados, es importante para determinar la calidad del producto final. Supongamos que T es una variable aleatoria distribuida uniformemente en el intervalo [150, 300] y que cuesta $ 1.5 producir un galón del producto. Si el petróleo destila a una temperatura menor que 200 grados centígrados, el producto se vende como gasolina a $ A por galón. Si se destila a una temperatura mayor que 200 grados centígrados, el producto es un aceite refinado que se vende a $ 1.2A por galón. Hallar el valor de A de tal manera que la ganancia esperada sea igual a $ 0.3 por galón.

41. Si $Z \sim N(0.1)$, hallar:

a) $P[Z \leq 1.7]$ b) $P[-2 \leq Z \leq -1.57]$ c) $P[Z > 25]$ d) $P[|Z| \leq 1.5]$ e) $P[Z = 2]$

42. Si Z es una variable aleatoria con distribución N(0, 1), hallar el valor z para el cual:

a) $P[Z \leq z] = 0.95$ b) $P[Z > z] = 0.01$ c) $P[Z \leq z] = 0.017$ d) $P[-z \leq Z \leq z] = 0.9298$

43. El precio promedio de las acciones que pertenecen a la empresa EE es 20 dólares, mientras que la desviación estándar es 2 dólares. Si el precio tiene distribución normal:

a) Hallar la probabilidad de que el precio de las acciones sea mayor a 22 dólares.
b) Hallar la probabilidad de que el precio de las acciones esté entre 19 y 22 dólares.

44. El diámetro X de las esferas de rodamiento fabricadas por una factoría debe estar entre 29.55 mm y 29.57 mm para que sean consideradas como buenas. Si se asume que los diámetros tienen distribución normal con media 29.56 mm y desviación estándar 0.01 mm:

a) Hallar el porcentaje esperado de esferas que son consideradas como buenas.
b) ¿Qué porcentaje de las esferas se espera que tengan diámetros mayores a dos desviaciones estándar de la media?
c) ¿Cuál debe ser el valor k para que el porcentaje esperado de los diámetros que estén en $[\mu - k\sigma, \mu + k\sigma]$ sea el 50%?

45. Un fabricante de televisores asegura que el tiempo medio de funcionamiento sin fallas de los aparatos es de 2 años con una desviación estándar de 0.25 años. Si el tiempo de vida de los aparatos sigue una distribución normal:
    a) ¿Cuál es la probabilidad de que el tiempo de buen funcionamiento de un televisor sea menor que 2.5 años?
    b) El fabricante garantiza que reemplazará gratis cualquier aparato de TV cuya duración sin fallas sea menor que k años. Aproximar k de tal modo que solo el 1% de los aparatos vendidos tenga que ser reemplazado, aproximadamente.

46. Una empresa minera ha determinado que el mineral que extrae de la mina que explota contiene un porcentaje X de cierto metal. De acuerdo a X la utilidad proyectada U será como sigue:

$$U = \begin{cases} 20 & si\ 40 < X \le 50 \\ 30 & si\ X \ge 50 \\ 0 & si\ X \le 40 \end{cases}$$

    Hallar la utilidad esperada de la empresa si se considera que la distribución de X tiene una distribución con tendencia central simétrica y desviación estándar igual a 4.

47. Una máquina automática para el llenado de paquetes de arroz puede regularse de modo que la cantidad media de arroz llenado sea la que se desee. Si la cantidad de arroz depositada se distribuye normalmente con desviación estándar igual a 10 gramos, ¿cuál debe ser la regulación media de modo que solo el 1% de los paquetes tengan un peso neto inferior a 990 g? Con la regulación media calculada, se escogen al azar y cada hora 4 paquetes que luego se pesan; si el promedio de estos no está entre 995 y 1,000 g, la máquina se detiene. Hallar la probabilidad de que la máquina se detenga.

48. El tiempo que se usa para reparar una máquina es una variable aleatoria cuya distribución es normal con esperanza 120 minutos y varianza 16. Si la reparación dura más de 125 minutos se incurre en una pérdida de $ 10,000; hallar la pérdida esperada total, incluyendo al costo de reparación si este es igual a $ 200.

49. X corresponde a las notas de un curso de Geografía y tiene distribución normal de media 12 y varianza 4, e Y corresponde a las notas de Historia y tiene distribución normal con media 13 y varianza 5.
    a) Hallar la distribución de $X + Y$ de $(X + Y)/2$ si se supone que X e Y son independientes.

    ¿Cuál es la probabilidad de que para un alumno que ha cursado las dos materias
    b) la suma de las notas sea mayor que 28?
    c) el promedio de sus notas esté entre 14 y 15?
    d) la nota de Historia sea mayor que la nota de Geografía?

50. Aproximar, usando la normal, las siguientes probabilidades:

$$P(12 \le X),\ P(11 \le X \le 15),\ P(X \le 13) \text{ y } P(X < 10)$$

    si X es una variable que tiene distribución binomial con parámetros $n = 100$ y $p = 0.3$.

51. En un centro de cómputo deben asignarse un número k de terminales de cómputo para 100 estudiantes. La probabilidad de que en un instante cualquiera un estudiante necesite un terminal es 0.5. Encontrar el número de terminales que deben asignarse si se desea que con probabilidad 0.95 un estudiante cualquiera encuentre al menos un terminal desocupado en el momento que concurra.

52. Una empresa dispone de 9 minutos para trasladar a su personal desde el punto A hasta el punto B. La empresa debe elegir entre el tipo de transporte R y el tipo de transporte S. Se ha determinado que el tiempo que se necesita para ir de A a B en el medio de transporte R es normal con esperanza 10 minutos y varianza 1, mientras que usando el medio de transporte S el tiempo que se usa es normal con media 10 y varianza 4. ¿Qué tipo de transporte se debe utilizar?

53. El tiempo que demora un empleado de una tienda en atender a un cliente tiene distribución normal con media 10 minutos y desviación estándar igual a 2 minutos. Si el empleado dispone de 44 minutos para atender a n clientes, hallar el número n de tal manera que esto sea posible con probabilidad 0.8413.

54. Un vendedor de computadoras ha determinado que el tiempo de vida de estas tiene distribución exponencial de media dos años.
    a) Hallar el porcentaje de computadoras vendidas que se espera que duren menos de dos años.
    b) Para promocionar la venta de estas computadoras el vendedor ofrece cambiarlas si estas duran menos de k años. Encontrar el valor de k de tal manera que el valor esperado de las computadoras devueltas sea a lo más el 5% de las vendidas.

55. La tasa de llegada de los camiones que abastecen un almacén es 2 cada hora. Si el número de camiones que llegan tiene distribución de Poisson y el tiempo que transcurre entre dos llegadas consecutivas se denota con T, hallar la probabilidad de que el tiempo que transcurre entre 2 llegadas consecutivas sea menor que 20 minutos.

RESPUESTAS A LOS EJERCICIOS

1. a) c = 0.3 d) $P[Y=0] = 0.3$, $P[Y=1] = 0.6$ y $P[Y=3] = 0.1$; b) 0.9 c) 0.8. **2.** a) $P(X=1) = 0.08$, $P(X=2) = 0.16$, $P(X=3) = 0.06$, $P(X=4) = 0.40$, $P(X=5) = 0.30$ b) 0.24 c) 0.7. **3.** b) 379.99. **4.** Calcular los valores esperados de los retornos para cada fondo y luego comparar. **6.** a) 13,600 b) calcular el valor esperado de la ganancia cuando se ordenan 200, 300, 400 y 500. Comparar y demandar la cantidad para la cual la ganancia esperada es mayor. **7.** La variable T puede tomar los valores 4, 5, 6 y 7, con probabilidades 0.1, 0.325, 0.425 y 0.15, respectivamente. **8.** Para una sola participación, calcular la probabilidad; para muchas participaciones, calcular la esperanza. **9.** a) Los valores que puede tomar X son: 2, 3 y 4 con probabilidades 4/24, 8/24 y 12/24, respectivamente b) 3.3333. **10.** 0.875. **11.** Deberá cobrarse $(0.1 + p)A*100\%$ **14.** 0.07203. **16.** b) La esperanza del costo es 6. c) De la fórmula de la varianza se tiene que la esperanza de x^2 es 1.9. La esperanza del costo es 7.9. **17.** a) 0.006 b) Calcular $P[Y \geq 2]$, en donde Y indica el número de cajas con dos aspirinas defectuosas cada una. La respuesta es 0.0005. **18.** Aplicar el teorema de la probabilidad total. La probabilidad pedida es 0.3286. **19.** $300P(X \leq 5) + (300 - 100)P(X = 6) = 273.79$, X = Número de pasajeros, de los 6 que han comprado boletos, que se presentan en el momento del embarque. **21.** Recalcular las probabilidades usando el teorema de Bayes. Las probabilidades revisadas son: 0.5525 y 0.4475. **22.** 0.5920. **23.** a) 0.0067. **24.** 0.3233. **25.** a) 0.9975 b) 9 buses para reemplazar. **26.** La prima debe ser 721,854 dólares. **27.** Usar la aproximación de la binomial con la distribución de Poisson con parámetro 400(0.01). **28.** Usar el teorema de la probabilidad total: $P(X = k) = \frac{e^{-6}6^k}{k!}\ 0.8 + \frac{e^{-9}9^k}{k!}\ 0.2$. **29.** a) Considerar que la distribución de la variable X, que indica el número de fallecidos de las 5,000 personas aseguradas, puede aproximarse mediante la distribución de Poisson de parámetro 1. **30.** 0.3678 **31.** Averiguar sobre la distribución geométrica y aplicarla. a) 0.07203 b) 0.657. **32.** 25% **33.** a) c = 1/3 c) 0.375. **34.** a) La altura es 2/3. **35.** a) a = 2 **36.** Observar que la distribución es la exponencial. El porcentaje esperado es 0.9806. **37.** Observando el histograma se puede usar la distribución uniforme. **38.** 0.1667. **40.** 1.59. **41.** a) 0.9554. **42.** a) 1.645. **43.** a) 0.1587. **44.** a) 68.26% b) 56% c) $k = 0.69$. **46.** La esperanza es $20P[40 < X < 50] + 30P[X \geq 50] = 18.944$. **47.** La media debe ser 1,013.66. **48.** 1,077.12 dólares. **49.** $(1/2)(X + Y)$ tiene distribución normal de media 12.5 de varianza 95/4. **50.** $P(12 \leq X) \approx P\left(\frac{12 - 0.5 - 100(0.3)}{\sqrt{100(0.3)(0.7)}} \leq Z\right)$ **51.** 58 aprox. **53.** $n = 4$. **54.** a) 0.6321 b) $k = 0.1026$ años.

CAPÍTULO 7

Estimación de parámetros

William Gosset

William Gosset nació en Canterbury, Inglaterra, en 1876. En Oxford estudió matemáticas y química, obteniendo el grado en Ciencias Naturales en 1899.

Después de graduarse trabajó en una cervecería de Arthur Guinness e hijos en Irlanda. En esta cervecería, Gosset comenzó a desarrollar una serie de métodos estadísticos aplicados al proceso de fabricación de la cerveza. La empresa envío a Gosset a trabajar a la University College en Londres para desempeñarse bajo la dirección de Karl Pearson.

Una de las principales contribuciones de Gosset es el desarrollo de la llamada distribución *t* de Student. Este estudio permitió el tratamiento inferencial de pequeñas muestras.

Gosset empleó el seudónimo de Student porque la empresa Guinness no le permitía publicar sus investigaciones.

Gosset murió en 1937, en Inglaterra, después de pasar toda su vida profesional con la cervecería Guinness.

CONTENIDO

7.1 Introducción

El objetivo de muchas tareas de investigación precisa la estimación de características numéricas de una o más poblaciones. Las estimaciones de parámetros, como el ingreso medio por hogar o como la proporción de personas con educación universitaria en un estado, se hacen efectivas usando muestras de la población. El muestreo proporciona las técnicas esenciales para el mejor diseño de las muestras y de este modo la mejor obtención de la información. En este capítulo se abordan inicialmente los conceptos relacionados con las propiedades de ciertas variables que se obtienen a partir de las muestras y que son útiles en la estimación de parámetros.

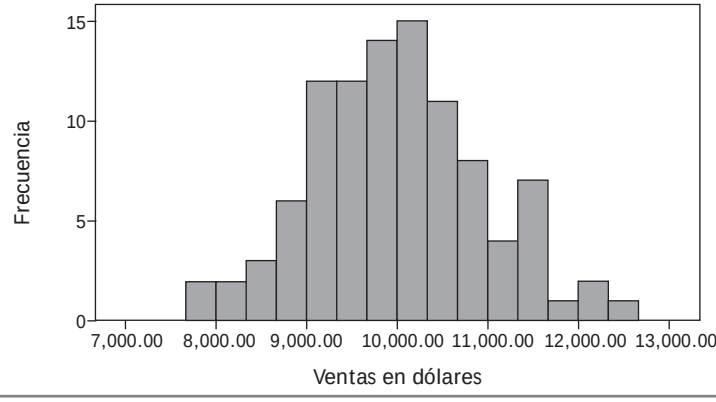
Para mejorar sus procesos, las empresas necesitan conocerlos, lo que implica el estudio de la manera como sus resultados varían. Pizzarun es un restaurante de comida italiana que, con el fin de "volverse más competitivo", decidió hacer entregas a domicilio de las pizzas solicitadas por teléfono. Pizzarun deseaba establecer un tiempo máximo de entrega. Si la entrega llegaba después de este tiempo máximo la empresa entregaba gratis el pedido. Con el fin de evitar, en lo posible, este tipo de entregas, Luiggi, gerente de Pizzarun, deberá conocer una serie de parámetros que caracterizan el proceso que comienza con el pedido, sigue con la preparación y termina con la entrega del pedido. Luiggi deberá recurrir a muestras para estimar estos parámetros.

Una presentación más formal del asunto de la estimación puede encuadrarse en la necesidad de conocer la distribución que sirve como modelo de los datos que se desea analizar. Estas distribuciones, que muchas veces se determinan a partir del análisis exploratorio de los datos y del conocimiento que el investigador tiene del problema que se estudia, están descritas por expresiones que dependen de parámetros que generalmente no se conocen y que deberán ser estimados.

Retailcom es una empresa que administra un centro comercial y desea estudiar la variabilidad de las ventas de lunes a viernes; para ello toma una muestra de las ventas de 100 días, sin considerar los sábados y domingos. Los valores de las ventas, en dólares, y el histograma aparecen en la Fig. 7.1.

FIGURA 7.1 *Distribución de datos*

12,492.05	11,649.23	10,383.81	9,396.35
8,753.61	9,330.08	8,817.98	9,064.77
9,723.35	10,201.65	10,545.90	10,999.58
10,640.40	11,772.62	8,800.99	9,508.99
8,535.52	8,162.81	11,603.00	10,995.87
10,105.34	9,313.93	10,750.35	9,161.41
10,190.27	10,327.02	10,372.32	10,461.95
10,056.47	8,938.73	10,083.69	8,608.46
11,513.31	11,225.15	10,718.24	9,638.02
9,772.81	10,057.77	9,954.67	9,695.50
10,282.66	9,578.39	9,999.39	9,983.92
9,739.87	9,620.55	11,532.81	11,315.31
9,452.61	12,248.23	8,681.17	10,537.22
9,506.94	11,411.55	9,360.95	8,583.25
12,321.04	9,975.27	9,870.18	10,806.65
10,475.94	8,042.20	9,166.78	10,798.44
10,296.29	7,804.94	9,083.95	11,223.09
9,473.67	7,860.84	9,603.46	10,966.93
10,292.85	9,784.57	9,086.35	11,537.94
10,809.97	11,380.81	9,037.68	10,505.21
10,072.35	9,595.09	10,109.96	10,002.23
10,634.16	9,245.47	9,130.21	10,002.29
9,720.15	10,587.15	9,125.74	10,169.51
9,891.32	10,366.61	11,210.05	9,202.96
9,833.46	8,752.65	9,551.58	9,682.60



Por la forma del histograma, se considera que la distribución normal es el modelo que, al parecer, puede usarse para describir la variabilidad de todas las ventas. La forma del modelo queda determinada; sin embargo, falta conocer la media μ y la desviación estándar σ, parámetros del modelo. Los valores de la muestra obtenida permitirán estimarlos.

A los parámetros del modelo se les llama ***parámetros de la población***.

EJEMPLO. *Media de los salarios*

Para conocer la media de los salarios de todos los trabajadores de una región se suman todos los salarios y el resultado se divide entre el número de trabajadores; esto se hace si se conocen todos los salarios. Si la información no está disponible, se puede obtener una muestra de salarios, y a partir de los valores que aparecen en la muestra, se procederá a estimar la media desconocida.

EJEMPLO. Control de calidad

Para conocer el porcentaje de artículos defectuosos de toda la producción de una empresa, se puede elegir 100 artículos fabricados, observar el número de defectuosos que hay entre los 100 y a partir de ello estimar el parámetro deseado. La población la forman todos los artículos que la fábrica produce, mientras que la muestra está constituida por los 100 artículos seleccionados.

Estos ejemplos ilustran la necesidad del uso de muestras para obtener estimadores de los parámetros. Claro está que se trata de obtener buenas estimaciones, y esto dependerá de la manera adecuada como se selecciona la muestra. La adecuación de la selección o diseño de la muestra está relacionada con el error que se produce al estimar el parámetro. Como el parámetro no se conoce, habrá que recurrir a la probabilidad para medir este error; de ahí que para la estimación estadística sea necesario integrar la aleatoriedad en la elección de los elementos de la muestra. Esto se consigue utilizando técnicas del *muestreo probabilístico*.

Existen diversos tipos de muestreo probabilístico, entre ellos se tiene *el muestreo aleatorio simple o básico*. El muestreo aleatorio simple o básico (m.a.s.) es la técnica más fácil para obtener muestras aleatorias y sirve como base para la aplicación de otros tipos de muestreo más elaborados, como: *el muestreo estratificado, el muestreo por conglomerados y los muestreos polietápicos*. Para el m.a.s., cada grupo de *n* elementos de la población tiene igual oportunidad de ser seleccionado.

Hay dos tipos de muestreo aleatorio simple: ***con reemplazo*** y ***sin reemplazo***. El m.a.s. con reemplazo consiste en reemplazar cada elemento seleccionado antes de realizar una nueva selección. Es de imaginar de qué se trata el m.a.s. sin reemplazo.

En este desarrollo se utiliza, para comenzar, el m.a.s. con reemplazo para construir las muestras. Con este tipo de muestreo resultan eventos que son más fáciles de tratar y de analizar; sin embargo, posteriormente se analizan los resultados obtenidos de muestras conseguidas con el m.a.s. sin reemplazo.

En la elección de muestras también se usa el muestreo no probabilístico, que se basa, en parte, en el juicio del responsable de la investigación, no se apoya en ninguna teoría probabilística y no permite el cálculo de los posibles errores que se cometen. Sin embargo, es preferido por los menores costos que produce y la facilidad del diseño. Entre los diversos tipos de muestreo no probabilísticos están: el muestreo *por conveniencia*, que se realiza acudiendo a poblaciones fácilmente accesibles, como por ejemplo entrevistas a la salida de un centro comercial, a la salida de un centro de votación (a "boca de urna"), etc.; el muestreo según el *criterio*, que se hace buscando elementos de la población más representativos (elección de una persona de una institución para que proporcione la información); el muestreo de *bola de nieve*, que se realiza cuando las poblaciones son pequeñas y no se dispone de la lista de los elementos de la población (a la persona que se le entrevista se le pide el nombre de una o más personas de la población que se estudia, para luego entrevistarlas y así aumentar la muestra); y el muestreo *por cuotas*, que se realiza dándole al entrevistador la libertad de elegir los elementos de la muestra pero bajo ciertos criterios, como las características y el número de personas a entrevistar.

7.2 Estimadores puntuales

Esencialmente, hay dos maneras de estimar los parámetros de una población: *puntualmente*, cuando se usa un único número como el valor más representativo del parámetro desconocido, y por *intervalos de confianza*, cuando se utiliza un intervalo o rango de valores que con cierta probabilidad contiene al valor del parámetro de la población que no se conoce.

Un parámetro poblacional θ es la característica numérica de la distribución de alguna variable aleatoria X, definida en la población. Visto así el asunto, para hallar un estimador de un parámetro se toma una muestra de tamaño n de la población y a cada elemento de la muestra se le aplica la variable, obteniéndose los valores $x_1, x_2,\ldots,x_n$.

Cualquier función de los valores $x_1, x_2,\ldots,x_n$ es una ***estimación puntual*** del parámetro.

Como la muestra es aleatoria, la estimación puntual de un parámetro es el valor de una variable aleatoria. Esta variable se llama ***estimador puntual*** del parámetro.

EJEMPLO. *Estimación puntual de la media*

Si se desea estimar la media μ de las edades de los pobladores de una región, se puede proceder de la siguiente manera:

1. Tomar una muestra aleatoria de *n* pobladores de la región, por ejemplo 5 pobladores.
2. Registrar los 5 valores de la variable *X* = edad. Las edades pueden ser: 40, 50, 12, 49 y 17.
3. Una estimación puntual de la media de la población puede ser la media $\bar{x} = \frac{40 + 50 + 12 + 49 + 17}{5}$. Otras expresiones, obtenidas a partir de los valores de la muestra, también pueden usarse; sin embargo, por razones que se explicarán se usará la media de estos valores.

EJEMPLO. *Estimación puntual de la proporción*

Si se desea estimar la proporción *p* de artículos defectuosos que produce una fábrica de artículos para el hogar, podremos tomar una muestra aleatoria de 200 artículos y registrar el número de aquellos defectuosos que en la muestra aparecen. Si en la muestra hay cuatro defectuosos, una estimación del parámetro es la proporción definida por la expresión $\hat{p} = \frac{4}{200}$.

(En este caso, la variable *X* es una variable que vale 1 si el artículo es defectuoso y 0 si no lo es. La proporción *p* que se trata de estimar es la media de *X*.)

La media y la proporción halladas son funciones de la muestra que bien pueden usarse como estimaciones de la media de la población y de la proporción poblacional, respectivamente; sin embargo, se deberá asegurar que estas son buenas estimaciones; buenas en el sentido de que están alrededor del parámetro a estimar y cerca de este. Esto obliga a revisar algunos criterios que servirán para evaluar la bondad de los estimadores.

Estimadores insesgados

*Un estimador puntual $\hat{\theta}$ del parámetro θ se llama **insesgado** si su valor esperado es igual al parámetro.*

Se muestra más adelante que la media y la proporción que se obtienen a partir de las muestras son estimaciones insesgadas de la media y la proporción de la población.

Cuando sea posible obtener más de un estimador insesgado para un solo parámetro,

la elección del mejor estimador dependerá de la varianza del estimador. En tal caso los siguientes conceptos servirán de ayuda.

Eficiencia de un estimador puntual

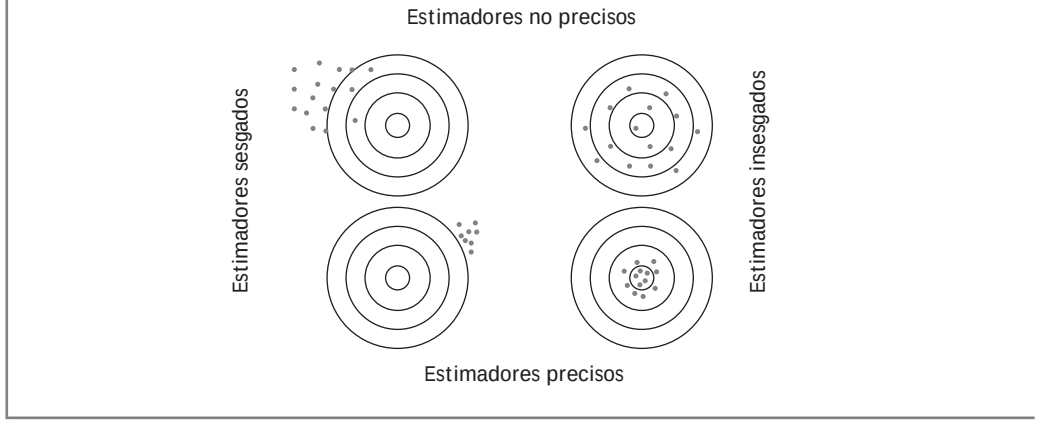
> *Dados dos estimadores puntuales insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$, de θ, se dice que $\hat{\theta}_1$* ***es más eficiente que*** *$\hat{\theta}_2$ si la varianza de $\hat{\theta}_1$ es menor que la varianza de $\hat{\theta}_2$.*

Error estándar de un estimador puntual

A la desviación estándar de un estimador puntual se le llama ***error estándar*** del estimador.

El error estándar es una medida de la *precisión* de un estimador; a menor error estándar, mayor es la precisión del estimador.

FIGURA 7.2 *Sesgo y precisión*



En la Figura 7.2, los centros de los círculos concéntricos representan al parámetro a estimar y los puntos que se observan, a los valores del estimador. Visto de esta manera, es intuitivamente claro que los buenos estimadores son los exactos y que tienen mayor precisión (menor varianza).

7.3 Distribuciones muestrales

La media muestral y la proporción muestral son resultados aleatorios que se obtienen a partir de los valores de una muestra de una variable aleatoria. Las distribuciones de estos valores se llaman ***distribuciones muestrales*** y sus características se estudian a continuación.

La media muestral

Usando una muestra aleatoria $x_1, x_2, \ldots, x_n$ de valores de la variable aleatoria X, se obtiene la ***media muestral****, que se define como* $\bar{x} = 1/n \sum_{i=1}^{n} x_i$.

La variable que representa a estos valores se denota con $\bar{X}$ y también se le llama media muestral.

EJEMPLO. *Las ventas de celulares*

Las ventas de celulares en tres días consecutivos en una tienda de aparatos de comunicación han sido 18, 22 y 14.

Si se toman al azar y con reemplazo las ventas de dos días se obtienen las muestras y las medias muestrales que a continuación aparecen.

Muestras	*Medias muestrales* $\bar{x}$
18, 18	18
18, 22	20
18, 14	16
22, 22	22
22, 14	18
22, 18	20
14, 14	14
14, 18	16
14, 22	18

Observamos que la media de todas las medias muestrales es igual a 18, valor que es igual a la media de los tres valores de la población. ¿Coincidencia? No. Es una propiedad de las medias muestrales.

Las propiedades de la distribución de las medias muestrales son las siguientes.

a) El valor esperado de la media muestral es igual a la media poblacional

Esta propiedad indica que si una población tiene media μ (la media de la variable *X* es μ), entonces la media de la media muestral $\bar{X}$ también es igual a μ.

De acuerdo a esta propiedad, la media muestral es un estimador insesgado de la media poblacional.

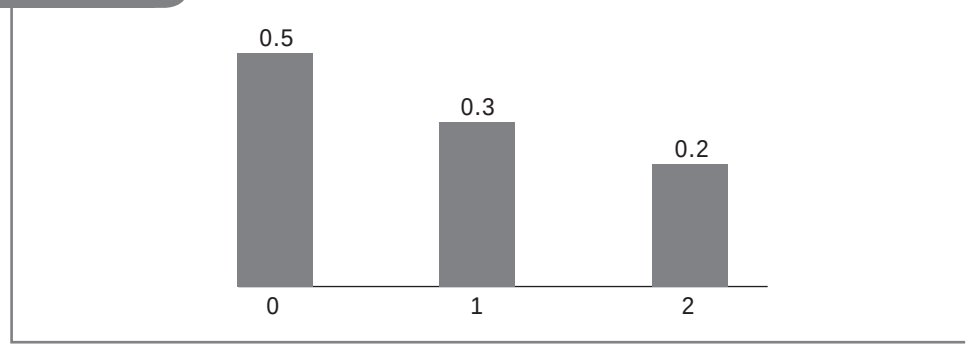
El siguiente ejemplo ilustrará esta propiedad.

Se ha observado que la demanda semanal *X* en una tienda de automóviles tiene la siguiente distribución.

TABLA 7.1 *Distribución de X*

Número de automóviles vendidos, X	0	1	2
$P[X = k]$	0.5	0.3	0.2

FIGURA 7.3 *Gráfico de la distribución de X*



El valor esperado de las ventas, por semana, es 0(0.5) + 1(0.3) + 2(0.2) = 0.7.

Si se forman muestras al azar con reemplazo de las ventas de dos días de la semana y se calculan sus medias muestrales respectivas se tendrán los resultados que se muestran en la Tabla 7.2.

TABLA 7.2 *Distribución de X*

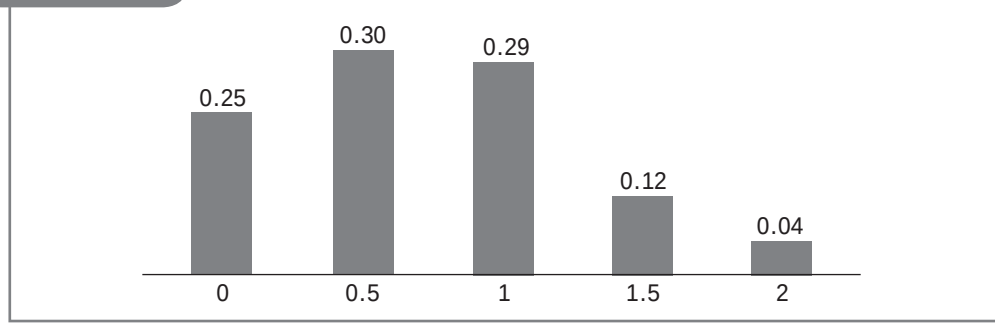
Muestra de las ventas de dos días cualesquiera	***Medias muestrales***
0 y 0	0
0 y 1	0.5
0 y 2	1
1 y 0	0.5
1 y 1	1
1 y 2	1.5
2 y 0	1
2 y 1	1.5
2 y 2	2

La distribución de probabilidad de las medias muestrales se resume en la Tabla 7.3 (notar que $P[\bar{X} = 0] = P[X = 0]\ P[X = 0] = (0.5)(0.5) = 0.25$, etcétera).

TABLA 7.3 *Distribución de la media muestral*

$\bar{X}$	0	0.5	1	1.5	2
$P[\bar{X} = k]$	0.25	0.30	0.29	0.12	0.04

FIGURA 7.4 *Distribución de la media muestral*



El valor esperado de la media muestral $\bar{X}$ es igual a (0)(25/100) + ... + (2)(4/100) = 0.7, valor que es igual a la media de X. Como indica la propiedad, la media de las medias muestrales es igual a la media de la población.

b) El error estándar de la media muestral

El error estándar de la media muestral de tamaño n, que se denota con $\sigma_{\bar{X}}$ y que es igual a la desviación estándar de la media muestral, dependerá de si la población es finita o infinita. Si la población es finita se supone que la muestra se construye con el muestreo aleatorio simple sin reemplazo. Si el muestreo que se usa es con reemplazo, el error estándar (EE) es igual al que se obtiene para poblaciones infinitas.

EE para población infinita

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

EE para población finita

$$\sigma_{\bar{X}} = \sqrt{\frac{N - n}{N - 1}}\frac{\sigma}{\sqrt{n}}$$

En donde:

σ es la desviación estándar de la población (de la variable X en estudio),

n es el tamaño de muestra y

N es el tamaño de la población.

El factor $\sqrt{\frac{N-n}{N-1}}$ se llama ***factor de corrección por población finita***. Este factor se aproxima con $\sqrt{1-\frac{n}{N}}$, y cuando el tamaño de la población es grande y el de la muestra es pequeña (en la práctica cuando $n \leq 0.1N$), este factor se aproxima a 1.

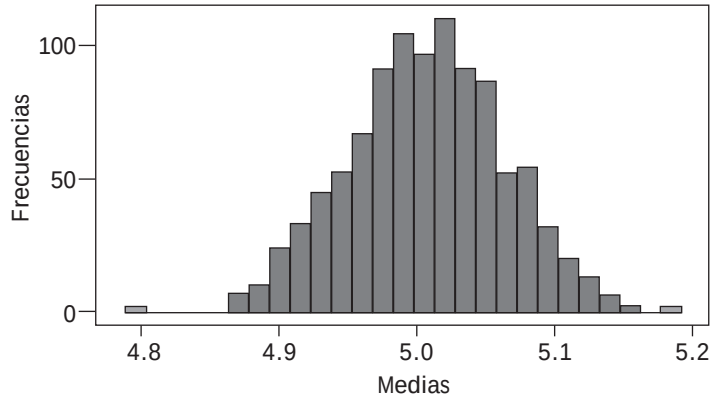
Cuando σ no se conoce, el error estándar se estima con $\frac{s}{\sqrt{n}}$, en donde $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

c) El teorema central del límite y la distribución de las medias muestrales

Este teorema, que es sumamente importante para la estadística, considera que los datos aleatorios que resultan de sumar diversos efectos aleatorios y sin relaciones entre sí tienen una distribución aproximadamente normal. Esto permite explicar la razón por la cual se considera a la distribución normal como modelo en las fluctuaciones de los valores de las acciones en la bolsa, en los promedios de los calificativos de los estudiantes o en las estaturas de las personas. Precisamente, las medias muestrales son datos aleatorios a los que se les puede aplicar esta propiedad. En este teorema también se basan muchos resultados de la estimación y de la prueba de conjeturas o hipótesis de parámetros de una población. El siguiente ejemplo ilustra esta propiedad.

De una población que consta de 1,000 mediciones se seleccionaron 50 muestras de tamaño 20 cada una. Para cada una de estas muestras se calculó la media muestral. El histograma de las 50 medias muestrales de tamaño 20 así obtenidas fue como aparece en la Figura 7.5.

FIGURA 7.5 *Distribución de la media muestral*



La forma "normal" del histograma obtenido no es casual; sucede en general cuando el tamaño de las muestras es grande y para cualesquiera que sea la forma de la distribución de los valores originales. Esta propiedad se debe precisamente al *teorema central del límite*.

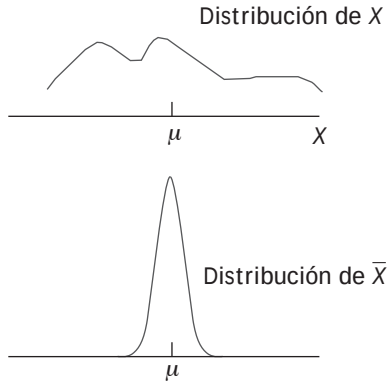
En general:

> *La media muestral $\bar{X}$ de una población de media μ y varianza σ^2 tiene distribución normal con media μ y varianza $\frac{\sigma^2}{n}$ cuando el tamaño de la muestra, n, es suficientemente grande.*

Como regla práctica, se considera que para obtener una aproximación razonable a la distribución normal el tamaño de la muestra debe ser mayor o igual que 30.

La siguiente gráfica (Figura 7.6) proporciona una idea de las relaciones de las características de *X* y de $\bar{X}$.

FIGURA 7.6 *Distribución de X y de la media muestral*



A partir de estos resultados, se puede indicar que:

1. La dispersión de los valores de la media muestral alrededor de la media poblacional es menor que la dispersión de los valores de la variable, pues el error estándar de la media muestral es menor que la desviación estándar de toda la población.
2. La precisión de la media muestral, medida con el error estándar, es mayor a medida que el tamaño de la muestra es más grande.

3. Para muestras grandes, de tamaño *n*, los valores estandarizados de las medias muestrales $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ se aproximan a la normal estándar, *N*(0, 1), no importando la población de donde provengan las muestras.

Cuando la población de donde provienen las muestras es normal, la distribución de las medias muestrales siempre es normal, no importando el tamaño de la muestra.

Las propiedades de la media muestral justifican el uso de la media muestral para estimar puntualmente la media poblacional.

EJEMPLO. *Tamaño de muestra para estimar el gasto promedio de los clientes en una tienda por departamentos*

Para estimar el promedio μ de las ventas de la tienda por departamentos De Rippley, el gerente de comercialización ha determinado previamente que el gasto de los clientes tiene una desviación estándar igual a 30 dólares. Para una primera estimación se tomó una muestra del valor de las compras de 36 clientes tomados al azar. Si se considera que el error de estimación es igual a $|\bar{x} - \mu|$, la probabilidad de que este sea menor que 5 dólares se calcula con:

$$P(|\bar{X} - \mu| < 5) = P(-5 < \bar{X} - \mu < 5) = P\left(\frac{-5}{30/\sqrt{36}} < \frac{\bar{X} - \mu}{30/\sqrt{36}} < \frac{5}{30/\sqrt{36}}\right) = P(-1 < Z < 1) = 0.6826$$

En este cálculo se considera que $30/\sqrt{36}$ es el error estándar de la media muestral y que $\frac{\bar{X} - \mu}{30/\sqrt{36}}$ es la media muestral estandarizada cuya distribución aproximada es la normal estándar, por el teorema del límite central.

Si el gerente deseara que el error sea menor que, por ejemplo, 2 dólares, con probabilidad 0.95, tal vez tenga que aumentar el tamaño de la muestra.

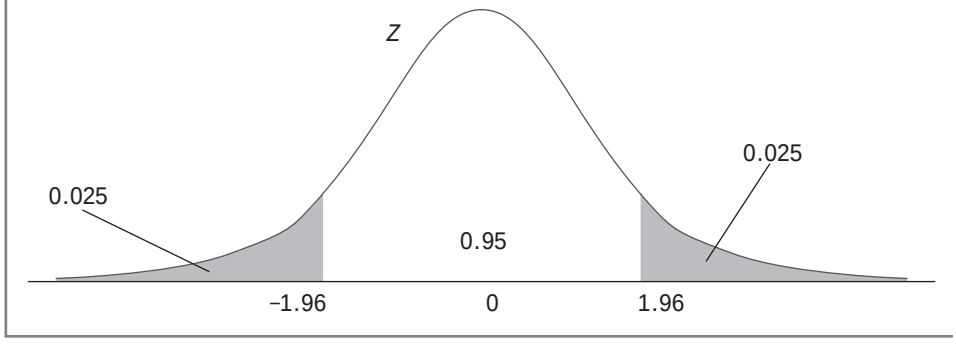
Veamos.

El gerente desea que $P(|\bar{X} - \mu| < 2) = 0.95$.

La expresión puede escribirse, de manera equivalente, como:

$$P\left(\frac{-2}{30/\sqrt{n}} < \frac{\bar{X} - \mu}{30/\sqrt{n}} < \frac{2}{30/\sqrt{n}}\right) = 0.95$$

FIGURA 7.7 *El cuantil de orden* 1 – 0.05/2 *de la distribución Z es* 1.96



Usando el cuantil 1.96 de orden 1 – 0.05/2 de la normal estándar, se tiene que $\frac{2}{30/\sqrt{n}} = 1.96$.

Despejando n de esta última igualdad se tiene que $n \approx 865$.

Luego, para obtener la estimación de la media, con la precisión requerida, es necesario tomar una muestra de 865 clientes.

EJEMPLO. *Rendimiento académico*

El jefe de la oficina de registro de una escuela de negocios indica que el rendimiento académico de los alumnos de la escuela tiene una distribución con promedio 17 y con una desviación estándar igual a 1.5. Como la escuela está en un proceso de acreditación, y el rendimiento es un factor importante en este proceso, un acreditador tomó al azar y con reemplazo el rendimiento de 50 alumnos, obteniendo una media muestral de 15. ¿Cuál será el informe del acreditador respecto del factor rendimiento?

Solución

Como el tamaño de la muestra es mayor que 30, la distribución de la media puede aproximarse con la distribución normal.

La media de la media muestral es igual a la media de la población. En este caso, la que indica el jefe de registro.

El error estándar de la media muestral es igual a $\frac{1.5}{\sqrt{50}} = 0.2121$.

Si lo que indica el jefe de registro, respecto de la media, es verdad, la probabilidad de obtener una media muestral menor o igual a 15 es:

$$P[\bar{X} \leq 15] = P\left[\frac{\bar{X} - 17}{1.5/\sqrt{50}} \leq \frac{15 - 17}{1.5/\sqrt{50}}\right] = P[Z \leq -9.4280] = 0$$

Bajo el supuesto del jefe de registro (la media de la población es 17 con una desviación estándar igual a 1.5), la probabilidad de encontrar una media muestral menor o igual a 15 es 0. Por ello el acreditador puede concluir que probablemente el informe del jefe de registro es falso.

La proporción muestral

Como en el caso de la estimación de la media de la población, para obtener la proporción muestral se sigue el siguiente proceso.

1. Se elige una muestra aleatoria de tamaño n de la población.
2. Con los valores de la muestra se calcula la *proporción* de los elementos de la muestra que tienen la propiedad. Esto es, se calcula el valor $\hat{p} = \frac{n_A}{n}$, en donde n_A es el número de elementos de la muestra que tienen la propiedad.

La proporción muestral $\hat{p}$ se usa como estimación puntual de la verdadera proporción poblacional p.

Los valores de $\hat{p}$ representados por la variable $\bar{P}$ corresponden a la variable aleatoria llamada ***proporción muestral de tamaño*** n. Algunas veces se usa indistintamente $\hat{p}$ y $\bar{P}$.

EJEMPLO. *Lectoría de periódicos*

Para estimar la proporción de personas que leen el periódico *La Tarde* en una ciudad se tomó una muestra de 200 personas y se comprobó que solo 40 de ellas leen el periódico.

La proporción muestral, 40/200 = 0.20, es una estimación de la verdadera proporción en toda la población de lectores del periódico en mención.

Las propiedades de la distribución de la proporción muestral son las siguientes.

a) El valor esperado de la proporción muestral es igual a la proporción en la población.

   Esta propiedad indica que la proporción muestral es un estimador insesgado de la proporción poblacional.

b) El error estándar de la proporción muestral

El error estándar de la proporción muestral, denotado con $\sigma\bar{P}$, dependerá de si la población es finita o infinita. Si la población es finita de tamaño *N*, se supone que la muestra se construye con el muestreo sin reemplazo. Si el muestreo que se usa es con reemplazo, el error estándar (*EE*) es igual al que se obtiene para poblaciones infinitas.

Se puede demostrar los siguientes resultados.

EE para población infinita ***EE para población finita***

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}} \qquad \sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

La diferencia entre las dos expresiones está en el factor de corrección por población finita. Si la muestra es pequeña, en relación con la población, ($n \leq 0.10N$), este factor es aproximadamente igual a 1 y las dos expresiones serán aproximadamente iguales.

Como *p* no se conoce, el error estándar se estima reemplazando a *p* con $\hat{p}$.

En la práctica, la proporción muestral se utiliza dentro del marco de poblaciones grandes, y en este desarrollo se supondrá, salvo indicación precisa, que la población es grande respecto de la muestra.

c) El teorema central del límite y la distribución de las medias muestrales

Como en el caso de la media muestral, y como consecuencia del *teorema central del límite*, se obtiene lo siguiente: si *el tamaño n de la muestra es suficientemente grande, la distribución de los valores de* $\bar{P}$ *estandarizados,* $\frac{\bar{p}-p}{\sigma_{\bar{p}}}$, *se aproxima a la distribución normal estándar*.

En la práctica, esta aproximación a la normal es válida cuando *n* es mayor que 30.

Las propiedades indicadas justifican el uso de la proporción muestral para estimar puntualmente la verdadera proporción en la población. La proporción muestral está alrededor de la proporción poblacional, y a medida que la muestra va creciendo la cercanía entre ellas es mayor.

En el ejemplo de lectoría de periódicos, se tiene lo siguiente:

a) El valor esperado de la proporción muestral es igual a la proporción de personas en toda la población que leen *La Tarde*.

b) El error estándar de la proporción muestral es aproximadamente igual a $\sqrt{\frac{(0.20)(1-0.20)}{200}} = 0.028$.

c) Aplicando las propiedades de la distribución de la proporción muestral, se tiene que $\frac{\bar{P}-p}{0.028}$ tiene una distribución aproximadamente igual a la normal estándar, por lo tanto:

$$P\left[\left|\frac{\bar{P}-p}{0.028}\right| < 1.96\right] = 0.95$$

De manera equivalente:

$$P\left[|\bar{P}-p| < (1.96)(0.028)\right] = 0.95$$

Esta relación indica que, con probabilidad 0.95, el error que se comete al estimar la verdadera proporción p con la proporción muestral $\hat{p}$ es menor que $1.96 \times 0.028 = 0.055$.

El siguiente ejemplo ilustra la manera como puede hallarse el tamaño de muestra que debe tomarse para estimar una proporción poblacional con la proporción muestral. En este ejemplo se puede notar que para ello es necesario conocer previamente el error de estimación y la probabilidad de que se cometa este error.

EJEMPLO. *Tamaño de muestra para la proporción*

En una ciudad se desea realizar una investigación para estimar la proporción p de individuos que conocen un refresco determinado. El investigador requiere que el error de estimación $|\hat{p}-p|$ sea menor que 0.05, con una probabilidad igual a 0.99. Se precisa el tamaño de la muestra.

Solución

Se requiere que $P(|\bar{P}-p| < 0.05) = 0.99$

Dividiendo entre el error estándar de la proporción muestral, se tiene:

$$P\left(\frac{|\bar{P}-p|}{\sqrt{p(1-p)/n}} < \frac{0.05}{\sqrt{p(1-p)/n}}\right) = 0.99$$

Como $\frac{\bar{P}-p}{\sqrt{p(1-p)/n}}$ tiene una distribución que se aproxima a la normal estándar, se cumple que $\frac{0.05}{\sqrt{p(1-p)/n}} = 2.58$ (el cuantil $1-0.05/2$ de la normal estándar es 2.58).

Despejando el valor de n, se tiene: $n = \dfrac{2.58^2\, p(1-p)}{0.05^2}$.

La dificultad que ahora se presenta es que no se conoce p (se desea estimar). Para salvar este problema se puede utilizar información pasada acerca de p. Si el investigador ha estimado un parámetro parecido en anterior ocasión, podrá usar este valor. Si no se tiene información anterior se podrá usar una muestra piloto, estimar p y luego usar esta información. Si no se desea usar una muestra piloto, se puede utilizar en la expresión el mayor valor que puede tomar $p(1-p)$. Este valor es 0.25.

Utilizando el mayor valor que puede tomar $p(1-p)$, se tiene que el tamaño de muestra a tomar es:

$$n \approx \frac{2.58^2(0.25)}{0.05^2} = 666$$

El investigador deberá tomar una muestra de 666 personas para estimar a p.

Otro de los parámetros que se precisa estimar a menudo es la varianza de la población. Este parámetro suele estimarse con la varianza muestral, que a continuación se indica.

La varianza muestral

A partir de los valores $x_1,\ldots, x_n$ de una muestra, se define la ***varianza muestral de tamaño*** n como el valor:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

La variable que representa a estos valores se llama *varianza muestral* y se denota con S^2.

Se demuestra que la esperanza de la varianza muestral es igual a la varianza de la población y que su error estándar disminuye a medida que la muestra crece.

Estas propiedades permiten usar a la varianza muestral como estimación puntual de la varianza de la población.

APLICACIÓN: El caso de Taxiseguro

Al asesor estadístico de Taxiseguro, con la finalidad de hallar la distribución del tiempo necesario en acudir al llamado de un servicio, tomó una muestra de 50 de estos tiempos. Los resultados en minutos fueron los siguientes:

12.40 10.44 13.49 15.55 15.40 16.47 8.63 12.53 15.19 10.83 11.62 9.62 9.31 11.04 11.45 8.76
11.86 13.24 14.32 12.27 12.35 12.26 15.69 12.83 12.63 11.97 16.94 14.73 17.75 14.53 16.32
12.34 14.08 14.80 16.84 12.83 11.95 14.35 12.24 14.52 10.11 11.31 11.40 13.12 12.94 13.06
12.35 17.39 9.52 11.53

La media muestral es 12.98 minutos aproximadamente, mientras que la desviación estándar muestral es de 2.28 minutos aproximadamente.

El asesor estadístico consideró, observando el histograma, que la distribución que se puede usar para modelar los tiempos transcurridos en acudir al servicio es la normal, con los parámetros iguales a las estimaciones antes mencionadas.

7.4 Estimación de parámetros por intervalos de confianza

Una de las desventajas de los estimadores puntuales es que no indican por sí mismos la precisión de la estimación realizada. Los estadísticos han diseñado procesos para construir intervalos que con el *nivel de confianza* $(1-\alpha)100\%$ contienen al parámetro. Este nivel o grado de confianza se refiere a la tasa de intervalos (construidos con el proceso) que contienen al parámetro. Si se seleccionan muchas muestras diferentes de tamaño n y con cada una de ellas se construye un intervalo al nivel de confianza $(1-\alpha)100\%$, se podrían tener $(1-\alpha)100\%$ de intervalos, aproximadamente, que contienen al parámetro. Cada uno de estos intervalos, cuyos extremos son valores que se obtienen a partir de una muestra, se llama ***intervalo al nivel de confianza* de $(1-\alpha)100\%$.**

El nivel de confianza de un intervalo se fija de antemano y generalmente se usa 95% y 99%.

Intervalo de confianza para la media de una población

El objetivo de esta sección es encontrar procedimientos para construir intervalos de confianza que permitan estimar la media μ de una población. Para ello se considera que:

1. Las muestras que se utilizan son aleatorias y provienen de poblaciones infinitas o de poblaciones finitas pero obtenidas con el m.a.s. con reemplazo.

2. La variable *X* que representa a la población tiene distribución normal de desviación estándar σ conocida o no.

   El supuesto de normalidad no es muy importante cuando la muestra es suficientemente grande (en la práctica, mayor que 30), pues en tal caso, la media muestral, por el teorema del límite central, tiene una distribución que tiende a la normal.

Para el muestreo aleatorio simple, el proceso de determinación de un intervalo de confianza comprende dos casos.

Caso 1. La desviación estándar de la población se conoce

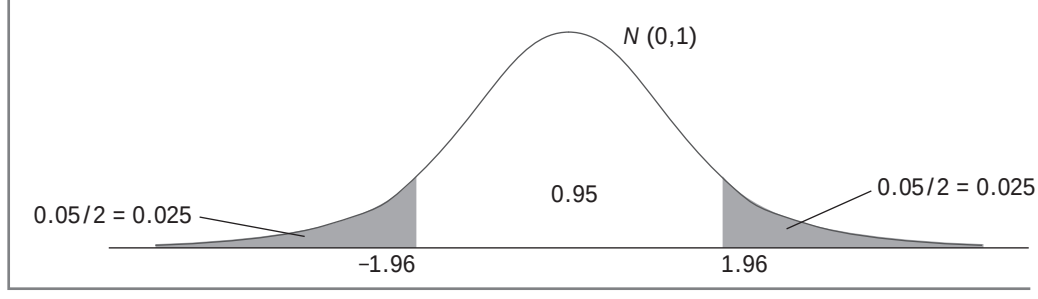
Teniendo la muestra $x_1,\ldots, x_n$ de la variable *X* y suponiendo que la desviación estándar σ de la población se conoce, el intervalo al nivel de confianza del 95% para la media poblacional μ es:

$$[\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}]$$

En esta expresión:

- $\bar{x}$ es la media muestral.
- 1.96 es el cuantil de orden 1 – 0.05/2 = 0.975 de la distribución normal estándar; esto es, el valor que deja por debajo de él, una área igual a 0.975, ($P(Z \leq 1.96) = 0.975$), y $\frac{\sigma}{\sqrt{n}}$ es el error estándar de la media muestral.

FIGURA 7.8 *El cuantil de orden* 1 – 0.05/2 = 0.975 *de la distribución Z es* 1.96



La semilongitud del intervalo, $1.96\sigma/\sqrt{n}$, se llama ***margen de error*** de la estimación.

El margen de error de la estimación puede interpretarse como el mayor error que se comete al estimar la media verdadera con la media muestral.

El intervalo puede escribirse como $\bar{x}$ ± *margen de error*. Cuanto más pequeño es el margen de error más cerca se encuentra la media muestral de la media poblacional que se desea estimar.

Un intervalo para la media μ, al nivel de confianza de 99% ($\alpha = 0.01$), es:

$$\left[\bar{x} - 2.58\ \frac{\sigma}{\sqrt{n}}, \bar{x} + 2.58\ \frac{\sigma}{\sqrt{n}}\right]$$

- El valor 2.58 es el cuantil de orden $1 - 0.01/2$ de la distribución normal estándar ($P(Z \leq 2.58) = 0.99$).
- El margen de error es $2.58\ \frac{\sigma}{\sqrt{n}}$.
- El intervalo se puede escribir como $\bar{x} \pm 2.58\ \frac{\sigma}{\sqrt{n}}$.

En general, el procedimiento para construir un intervalo de confianza para la media μ de una población normal, al nivel $(1 - \alpha)100\%$ y cuando la desviación estándar σ se conoce, es como sigue:

1. De la tabla de la normal estándar, calcular el cuantil de orden $1 - \alpha/2$. Este cuantil se denota con $z_{1-\alpha/2}$, ($P[Z \leq z_{1-\alpha/2}] = 1 - \alpha/2$).

   Para el nivel de confianza 95%, el cuantil es 1.96.

   Para el nivel de confianza 99%, el cuantil es 2.58.

2. Calcular el margen de error $z_{1-\alpha/2}\ \sigma/\sqrt{n}$.
3. El intervalo de confianza es $[\bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{n}]$. Intervalo que puede escribirse como $\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{n}$.

EJEMPLO. *Tiempo en el banco*

Se han registrado los tiempos que 100 clientes, tomados al azar, utilizan en sus distintas operaciones en un banco local. La media de la muestra fue de 10 minutos. Informaciones anteriores indican que la distribución de los tiempos utilizados en las distintas operaciones es normal con desviación estándar igual a 3 minutos. Estimar el promedio real μ del tiempo utilizado por los clientes con un intervalo al nivel de confianza de 99%.

Solución

El intervalo al nivel de confianza de 99% para la media poblacional es:

$$[10 - (2.58)(3/\sqrt{100}),\ 10 + (2.58)(3/\sqrt{100})] = [9.226, 10.774]$$

Los clientes utilizan para realizar sus operaciones en un banco local, en promedio, entre 9.226 y 10.774 minutos.

EJEMPLO. *Señal vs. ruido*

Desde un punto A hasta un punto B, un aparato transmite una "señal" que consiste en un número *U*. Por diferentes razones la señal llega con "ruido", de tal manera que se recibe el valor $Y = U + N$. El ruido *N* es una variable aleatoria con distribución normal de media 0 y varianza 1 (los valores recibidos *Y* tienen distribución normal de media *U* y varianza 1). Con la finalidad de reducir el ruido y de esta manera estimar con precisión a *U*, se envía el mismo valor 9 veces. Si los valores recibidos son: 4.0, 5.1, 5.8, 4.9, 7.0, 6, 5.5, 6.1, 5.3, se puede construir un intervalo al nivel de confianza 95% con la media de estos valores para la media *U* de *Y*. Este intervalo es $\left[5.52 - 1.96\ \frac{1}{\sqrt{9}},\ 5.52 + 1.96\ \frac{1}{\sqrt{9}}\right] = [4.87,6.17]$.

El margen de error de la estimación es $1.96\ \frac{1}{\sqrt{9}} = 0.6530$.

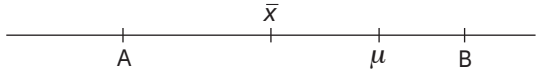
Se podría decir que el verdadero mensaje *U* estará entre 4.87 y 6.17 con un grado de confianza del 95%, y que el máximo error que se comete al estimar *U* con la media muestral, 5.52, es 0.6530.

Nótese que si mayor es el número de veces que se envía el mensaje, menor será el margen de error, y por tanto "el error de transmisión" de la señal será menor.

Tamaño de muestra para estimar la media de una población

El grado de confianza del intervalo $[\bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{n}]$ es de $(1 - \alpha)100\%$, lo que significa que si se estima el parámetro μ con $\bar{x}$, el margen de error es igual a $(z_{1-\alpha/2})(\sigma/\sqrt{n})$.

FIGURA 7.9 *Intervalo para* μ *al nivel de confianza* $(1 - \alpha)100\%$



$A = \bar{x} - (z_{1-\alpha/2})(\sigma/\sqrt{n})$, $B = \bar{x} + (z_{1-\alpha/2})(\sigma/\sqrt{n})$

Luego:

si se desea estimar la media de una población con la media muestral, con un margen de error igual a E, y con nivel de confianza de $(1 - \alpha)100\%$*, bastará tomar una muestra de tamaño n de tal modo que* $(z_{1-\alpha/2})(\sigma/\sqrt{n}) = E$.

Se deduce de esta igualdad que el tamaño de muestra a usar es:

$$n = \frac{(z_{1-\alpha/2}\,\sigma)^2}{E^2}$$

EJEMPLO. *Tamaño de muestra*

A partir de una muestra, se requiere estimar la media de las edades de las personas que trabajan en la industria metalmecánica, de tal modo que el margen de error sea 0.5, con nivel de confianza del 95%. ¿Cuántas personas deben incluirse en la muestra, si se supone normalidad y que la varianza de las edades de todas las personas que trabajan en la industria metalmecánica es 16?

Solución

Si se desea estimar μ con un margen de error de 0.5, con nivel de confianza 95%, bastará que el tamaño n de la muestra sea tal que $1.96(4)/\sqrt{n} = 0.5$.

Resolviendo la última ecuación, se tiene que el número de personas que se debe incluir en la muestra es $n = [(1.96)(4)/0.5]^2 = 246$.

Observación

Para calcular el tamaño de muestra es necesario conocer la desviación estándar de la población. Si esta no se conoce:

1. Suele usarse información proveniente de algún estudio anterior.
2. Suele usarse una muestra piloto. Se recomiendan al menos 31 valores para el tamaño de esta muestra. A partir de esta muestra se calcula el estimador s de la desviación estándar para ser usada en la fórmula correspondiente.
3. Si la población es normal o aproximadamente normal y el rango R de los valores de la muestra se conocen, suele usarse la aproximación $\sigma \approx R/4$.

Caso 2. La desviación estándar de la población no se conoce

En este caso se considera la construcción de intervalos de confianza para la media de una población cuando la desviación estándar no se conoce.

El supuesto que se requiere en este desarrollo, respecto de la muestra, es que esta provenga de una población normal. Este supuesto puede suprimirse dependiendo del tamaño de la muestra y de cuánto se aleja la distribución de la población de la distribución normal. En la práctica se considera que el tamaño de la muestra debe ser mayor que 30.

En general, el procedimiento para construir un intervalo para la media μ de una población normal, de un nivel de confianza de $(1 - \alpha)100\%$ y cuando la desviación estándar σ no se conoce, es como sigue:

1. Como no se conoce la desviación estándar σ, esta se estima con:

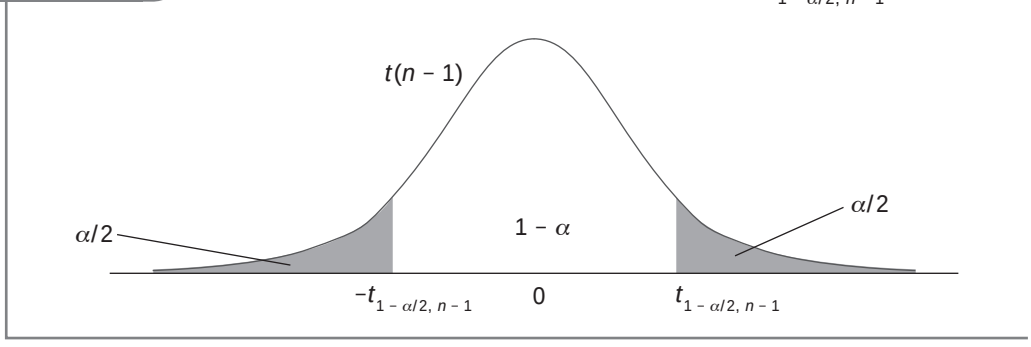
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

   Esta estimación origina una disminución del grado de confianza del intervalo, por lo que es necesario, para mantener un nivel de confianza deseado, ampliar el intervalo, utilizando la distribución *t de student* con $n - 1$ grados de libertad.

2. De la tabla de la *t* student se calcula el cuantil de orden $1 - \alpha/2$, $t_{1-\alpha/2,\, n-1}$.
3. Calcular el margen de error para la media muestral. En este caso es $t_{1-\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}}$.
4. El intervalo de confianza es $\left[\bar{x} - t_{1-\alpha/2,\, n-1}\ \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2,\, n-1}\ \frac{s}{\sqrt{n}}\right]$.

El intervalo puede escribirse como $\bar{x} \pm t_{1-\alpha/2,\, n-1}\frac{s}{\sqrt{n}}$.

FIGURA 7.10 *El cuantil* $1 - \alpha/2$, *de la distribución t* $(n - 1)$ *es* $t_{1-\alpha/2,\, n-1}$



Cuando la muestra es grande (mayor que 30), el cuantil $t_{1-\alpha/2,\, n-1}$ *se aproxima con el cuantil* $z_{1-\alpha/2}$ *de la normal y así:*

el intervalo al nivel de confianza del 95% se aproxima con:

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{n}}, \bar{x} + 1.96\frac{s}{\sqrt{n}}\right] \quad y$$

el intervalo al nivel de confianza de 99% se aproxima con:

$$\left[\bar{x} - 2.58\frac{s}{\sqrt{n}}, \bar{x} + 2.58\frac{s}{\sqrt{n}}\right]$$

EJEMPLO. *Promedio de precios cobrados*

Para una muestra de 25 corredores de bolsa se encontró que la media de los precios cobrados por una transacción de 5 acciones a $ 20 la acción fue $ 10. La desviación estándar poblacional fue $ 2.1. Hallar el intervalo de confianza de 95% para la media de todos los precios cobrados por una transacción de 5 acciones a $ 20 la acción. Se supone que los precios tienen distribución normal.

Solución

En este caso, $n = 25$, $\bar{x} = 10$ y $s = 2.1$.

El cuantil de orden $1 - 0.05/2$, de la distribución t con $n - 1 = 24$ grados de libertad, es $t_{1-0.05/2,\,24} = 2.797$.

El intervalo de confianza es:

$$[\bar{x} - t_{1-0.05/2,24}\, s/\sqrt{n},\ \bar{x} + t_{1-0.05/2,24}\, s/\sqrt{n}] = [10 - 2.797(2.1/\sqrt{25}),\ 10 + 2.797(2.1/\sqrt{25})] = [8.83,\ 11.17].$$

El promedio de los precios cobrados por una transacción de 5 acciones, a $ 20 la acción, está entre 8.83 y 11.17 dólares, con un nivel de confianza de 95%.

EJEMPLO. *Tiempo de procesamiento*

Para estimar el tiempo promedio, en segundos, que se necesita para procesar cierto tipo de programas computacionales, se consideraron 100 de estos programas. Se encontró que para los tiempos de procesamiento x_i:

$$\sum x_i = 1479.8 \qquad \sum (x_i - \bar{x})^2 = 1755.$$

El intervalo, al nivel de confianza del 95%, para la media del tiempo que dura el procesamiento de los programas se encuentra calculando la media y la desviación estándar muestral.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1479.8}{100} = 14.7980 \qquad s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1755}{99}} = 4.21$$

Por el teorema del límite central, la media muestral tiene una distribución que se acerca a la normal ($n \geq 30$). Por tanto, el intervalo al nivel de confianza de 95% para la media μ de todos los tiempos de procesamiento de los programas es:

$$[14.7980 - 1.96(4.21/10),\ 14.7980 + 1.96(4.21/10)] = [13.9728,\ 15.6179]$$

Intervalo de confianza para la media de una población pequeña

Los resultados anteriores se cumplen en cuanto se refieren a poblaciones grandes o cuando las poblaciones son pequeñas y se usa el m.a.s. con reemplazo; sin embargo, lo real es que sea necesario estimar parámetros en poblaciones finitas de tamaño *N* y utilizando muestras aleatorias obtenidas con el m.a.s. sin reemplazo.

Si la varianza σ^2 es conocida, el intervalo, al nivel de confianza de $(1 - \alpha)100\%$ para estimar la media μ de la población, es:

$$\left[\bar{x} - (z_{1-\alpha/2})\sigma\sqrt{\frac{1}{n}\left(\frac{N-n}{N-1}\right)}, \bar{x} + (z_{1-\alpha/2})\sigma\sqrt{\frac{1}{n}\left(\frac{N-n}{N-1}\right)}\right]$$

$\frac{N-n}{N-1}$ se llama factor de corrección por población finita y puede aproximarse con $\frac{N-n}{N-1} \approx 1 - \frac{n}{N}$, obteniéndose el intervalo de confianza:

$$\left[\bar{x} - (z_{1-\alpha/2})\sigma\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)}, \bar{x} + (z_{1-\alpha/2})\sigma\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)}\right]$$

Si la población es grande, en la práctica mayor que 100,000, y el tamaño de la muestra es menor que el 10% de la población, el factor de corrección por población finita puede obviarse.

Tamaño de muestra para estimar la media poblacional en el caso de una población pequeña

Si la población es pequeña de tamaño *N*, el tamaño de muestra para estimar la media poblacional, de tal manera que el margen de error sea igual a *E*, al nivel de confianza de $(1 - \alpha)100\%$, se calcula a partir de la relación $(z_{1-\alpha/2})(\sigma)\sqrt{\frac{1}{n}\left(1 - \frac{n}{N}\right)} = E$.

Despejando *n* se tiene que $n = \frac{N\sigma^2}{(N-1)\left(\frac{E}{z_{1-\alpha/2}}\right)^2 + \sigma^2} = \frac{\sigma^2}{\left(1 - \frac{1}{n}\right)\left(\frac{E}{z_{1-\alpha/2}}\right)^2 + \frac{\sigma^2}{N}}$, valor que puede escribirse como $n = \frac{n_{inf}}{1 + \frac{n_{inf}}{N}}$, en donde n_{inf} es el tamaño de muestra para poblaciones infinitas.

Muchas veces la desviación estándar de la población no se conoce, por lo que no será posible aplicar directamente esta regla. Para suplir esta deficiencia se estima la desviación estándar a partir de una muestra piloto o se usan resultados de experiencias anteriores.

Observaciones

Cuando el tamaño N de la población es grande, es suficiente tomar como tamaño de muestra al valor de n tal que $n = \frac{(z^2_{1-\alpha/2})(\sigma^2)}{E^2}$.

EJEMPLO. *Los salarios*

Para estimar el promedio de los salarios de 100 empleados de una compañía se tomó una muestra aleatoria de 50 de ellos. Para esta muestra se halló $\bar{x} = 84.1$ y $s = 11.0653$. Encontrar un intervalo al nivel de confianza del 95% para estimar la media de los salarios de todos los trabajadores de la compañía.

Solución

El intervalo al nivel de confianza del 95% para la media μ de la población es:

$$\left[84.1 - 1.96\ \frac{11.0653}{\sqrt{50}}\sqrt{1 - \frac{50}{100}},\ 84.1 + 1.96\ \frac{11.0653}{\sqrt{50}}\sqrt{1 - \frac{50}{100}}\right] = [81.9312,\ 86.2688]$$

Se tiene la confianza del 95% de que la media de la población esté entre los valores 81.9312 y 86.2688.

EJEMPLO. *Tamaño de muestra para estimar el tiempo que una empresa demora en contestar las quejas de sus clientes*

Se desea estimar el promedio de días que una empresa demora en responder a las quejas de 300 clientes. Aun cuando no se conoce la varianza poblacional del número de días que demoran las quejas, se sabe que varían aproximadamente en un intervalo de amplitud igual a 120 días. Hallar el tamaño de muestra necesario para estimar la media μ con un margen de error de 2 días y al nivel de confianza de 95%.

Solución

El tamaño de muestra para el caso de que la población fuera infinita o muy grande es:

$$n_{inf} = \frac{(1.96)^2(120/4)^2}{(2)^2} \approx 865$$

El tamaño de muestra para estimar la media del número de días que la empresa demora en responder a las quejas de sus clientes es $n = \dfrac{865}{1 + \dfrac{865}{223}} \approx 223$. Será necesario usar toda la población.

Intervalos de confianza para la diferencia de las medias de dos poblaciones normales e independientes

Cuando se desea obtener conclusiones acerca de la diferencia de las medias de dos poblaciones sin importar las medias en sí, se utilizan intervalos de confianza para la diferencia de medias. Así, para comparar los tiempos promedios que dos entidades financieras demoran para otorgar préstamos, se define una variable "tiempo" para cada entidad financiera, se selecciona una muestra aleatoria para cada variable y se calculan las medias muestrales respectivas. A partir de las medias muestrales se obtiene un intervalo de confianza que, con cierto grado de confianza, contiene a la diferencia de las medias de todos los tiempos para otorgar los préstamos.

Caso 1: Las varianzas de las variables se conocen

Se asume que se tienen dos variables *X* e *Y* normales (se puede prescindir de la suposición de normalidad cuando las muestras son grandes) e independientes, donde las medias respectivas μ_1 y μ_2 no se conocen pero sí las varianzas respectivas σ_1^2 y σ_2^2.

El intervalo para estimar la diferencia de las medias $\mu_1 - \mu_2$, al nivel de confianza $(1 - \alpha)100\%$*, es:*

$$\left[(\bar{x} - \bar{y}) - z_{1-\alpha/2}\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)},\ (\bar{x} - \bar{y}) + z_{1-\alpha/2}\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}\right]$$

en donde $\bar{x}$ e $\bar{y}$ son las medias muestrales de tamaños n_1 y n_2, de X e Y, respectivamente.

EJEMPLO. *Comparando dos procesos de fabricación*

Para comparar la eficacia de dos procesos A y B de fabricación respecto del tiempo necesario para construir un cierto tipo de artículos, se midieron los tiempos, en segundos, necesarios para construir 50 artículos, 25 con cada proceso. Los promedios muestrales fueron $\bar{x} = 290$ segundos e $\bar{y} = 280$ segundos, respectivamente. Por experiencias anteriores se sabe que las varianzas de los tiempos para A y B son 100 y 96, respectivamente. Hallar un intervalo al nivel de confianza del 99% para la diferencia de las medias de los dos procedimientos. Suponer que los tiempos necesarios en cada proceso son independientes y se distribuyen normalmente.

Solución

En este caso: $\bar{x} - \bar{y} = 10$, $\alpha = 0.01$ y $z_{1-\alpha/2} = 2.58$.

Con estos resultados, se tiene que el intervalo al nivel de confianza del 99% para la diferencia de los promedios de los tiempos que se utilizan para fabricar un artículo es:

$$\left[10 - 2.58\sqrt{(100/25) + (96/25)},\ 10 + 2.58\sqrt{(100/25) + (96/25)}\right] = [2.78,\ 17.22]$$

De acuerdo a este resultado se puede afirmar que, con un grado de confianza del 99%, el tiempo promedio de fabricación según el procedimiento A es mayor que el tiempo promedio según el proceso B.

Caso 2. Las varianzas no se conocen pero son iguales

Muestras pequeñas

Cuando se trata de muestras pequeñas, y las varianzas respectivas σ_1 *y* σ_2 *de las variables X e Y son desconocidas pero iguales a* σ*, un intervalo al nivel de confianza de* $(1 - \alpha)100\%$*, para la diferencia* $\mu_1 - \mu_2$*, es:*

$$\left[(\bar{x} - \bar{y}) - t_{1-\alpha/2}s_c\sqrt{(1/n_1 + 1/n_2)},\ (\bar{x} - \bar{y}) + t_{1-\alpha/2}s_c\sqrt{(1/n_1 + 1/n_2)}\ \right]$$

en donde:

$$s_C^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},\ \text{con}\ s_1^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x})^2}{n_1 - 2}\ \ y\ \ s_2^2 = \frac{\sum_{i=1}^{n_1}(y_i - \bar{y})^2}{n_2 - 1},\ \text{es}$$

un estimador de la varianza común y $t_{1-\alpha/2}$ *es el cuantil de orden* $1 - \alpha/2$ *de la distribución t con* $n_1 + n_2 - 2$ *grados de libertad.*

EJEMPLO. *Comparando ventas en dos secciones de una tienda*

Con el fin de comparar las ventas en cientos de dólares de dos secciones A y B en una tienda de ropa se registraron las ventas correspondientes a 9 operaciones de ventas en cada sección. Para la sección A se obtuvo: 32 37 35 28 41 44 35 31 34 y para la sección B: 35 31 29 25 34 40 27 32 31. Se supone que las varianzas de las ventas en las dos secciones son iguales.

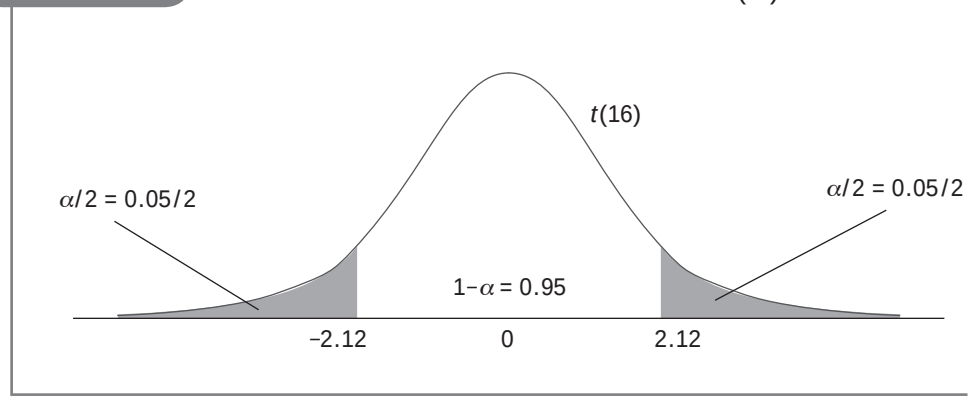
La varianza común se estima con $s_C^2 = \dfrac{8s_1^2 + 8s_2^2}{8 + 1 + 8 + 1 - 2} = \dfrac{195.56 + 160.22}{9 + 9 - 2} = 22.23.$

Para el nivel de confianza de 95% ($\alpha = 0.05$), el cuantil de orden 0.975 para la distribución *t* con 16 grados de libertad es 2.12.

El intervalo para estimar la diferencia de medias de las ventas en las dos secciones, $\mu_1 - \mu_2$, al nivel de confianza del 95%, es:

$$\left[(\bar{x} - \bar{y}) - t_{1-\alpha/2} s_c \sqrt{(1/n_1 + 1/n_2)},\ (\bar{x} - \bar{y}) + t_{1-\alpha/2} s_c \sqrt{(1/n_1 + 1/n_2)}\right] = [-1.05, 8.34]$$

FIGURA 7.11 *El cuantil de orden* 0.975 *de la distribución t*(16) *es* 2.12



El intervalo contiene valores positivos, negativos o cero. Por ello no se puede afirmar nada de la media de las ventas por cliente de la sección A, respecto de la media de las ventas por cliente de la sección B.

Muestras grandes

Cuando las muestras son mayores o iguales que 30, el intervalo para la diferencia de medias se determina, como en el caso anterior, usando la distribución normal.

Intervalos de confianza para la proporción *p*

Esta vez se describe el procedimiento para estimar, mediante un intervalo al nivel de confianza del $(1 - \alpha)100\%$, la proporción *p* de elementos de una población que tienen cierto atributo A.

El procedimiento es el siguiente.

1. Se requiere en primer lugar que la población de donde proviene la muestra sea infinita o muy grande, en la práctica mayor que 100,000, y que las muestras se obtengan con el m.a.s. con las condiciones adicionales $np \geq 5$ y $n(1 - p) \geq 5$. Estas condiciones permiten utilizar la distribución normal para aproximar la distribución de la proporción muestral, fundamental en este proceso.

2. De la muestra obtenida, calcular el estimador puntual p de p.
3. Calcular el cuantil $z_{1-\alpha/2}$ de orden $1-\alpha/2$ de la normal estándar.
4. Evaluar el margen de error de la estimación puntual p, determinado por $z_{1-\alpha/2}\sqrt{p(1-p)/n}$.
5. El intervalo aproximado, al nivel de confianza de $(1-\alpha)100\%$, para la proporción p es:

$$[p - z_{1-\alpha/2}\sqrt{p(1-p)/n},\ p + z_{1-\alpha/2}\sqrt{p(1-p)/n}]$$

Las relaciones $np \geq 5$ y $n(1-p) \geq 5$ no pueden ser verificadas en la realidad porque p no se conoce; sin embargo, se puede usar la estimación p para la verificación.

EJEMPLO. *A favor de GG*

En una encuesta de opinión realizada en una ciudad muy grande, 320 de un total de 400 personas entrevistadas se declararon a favor del candidato GG a la presidencia de la república.

a) Hallar un intervalo al nivel de confianza de 95% para estimar la proporción p de las personas que están a favor del candidato GG en toda la población.

b) Si la proporción de personas en toda la población que apoyan a GG se estima en 80%, ¿cuál es el margen de error de la estimación al nivel de confianza de 99%?

Solución

La proporción de las personas que votan por GG en la muestra es $p = 320/400 = 0.80$.

La población es grande, $np = 400(0.8) \geq 5$ y $400(1-0.8) \geq 5$; por lo tanto, los supuestos para construir un intervalo de confianza se cumplen.

a) El cuantil $z_{1-\alpha/2}$, de orden 0.975 de la normal para el que $P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 0.95$, es 1.96.

El margen de error es $1.96\sqrt{0.8(0.2)/400}$.

El intervalo aproximado para la proporción, al nivel de confianza del 95%, es:

$$[0.8 - 1.96\sqrt{0.8(0.2)/400},\ 0.8 + 1.96\sqrt{0.8(0.2)/400}] = [0.7608,\ 0.8392]$$

El grado de confianza de que la proporción de los que votan por GG en la población esté entre el 76.08% y 83.92% es 95%.

b) El intervalo al nivel de confianza del 99% para p, cuando la proporción muestral p es 0.80, es [0.7484, 0.8516]. El margen de error es 0.0516.

Tamaño de muestra para estimar una proporción en una población muy grande

El procedimiento que permite hallar el *tamaño de muestra* para estimar una proporción p en una población infinita o muy grande es análogo al que se usó para el caso de la media poblacional.

EJEMPLO. *Tamaño de muestra para estimar la proporción de internautas*

Se desea estimar, con un margen de error del 1% y con un nivel de confianza de 95%, a la proporción p de universitarios que acceden a Internet durante el día.

Solución

Si p denota a la proporción de universitarios que acceden a Internet durante el día, el intervalo al nivel de confianza del 95%, para la proporción de todos los universitarios con las características indicadas en la población, es:

$$[p - 1.96\sqrt{p(1-p)/n},\ p + 1.96\sqrt{p(1-p)/n}]$$

Como el margen de error debe ser 0.01, se debe cumplir $1.96\sqrt{p(1-p)/n} = 0.01$.

De la última igualdad se tiene $n = \dfrac{(1.96)^2\,\hat{p}\,(1-\hat{p}\,)}{(0.01)^2}$.

Como ya se indicó, para conocer el valor de n se necesita p; pero este valor no se conoce. Por ello, puede tomarse el mayor valor de $p(1-p)$ que es 0.25. Resulta así:

$$n = \frac{(1.96)^2\,(0.25)}{(0.01)^2} = 9{,}604$$

En general, para estimar la proporción p de los elementos que tienen un determinado atributo A en una población infinita o finita pero muy grande, dentro de un margen de error E y con nivel de confianza de $(1-\alpha)100\%$ *($\alpha > 0$ y pequeño), es suficiente tomar una muestra de tamaño n igual a:*

$$n = \frac{(z_{1-\alpha/2})^2\,(0.25)}{E^2}$$

Si se tiene alguna información anterior sobre p, esta se puede usar esta para determinar el tamaño de la muestra.

EJEMPLO. *Desempleados en el país*

Se desea estimar la proporción actual de desempleados en un país, con un margen de error del 1% y con un nivel de confianza de 95%. Hallar el tamaño de muestra a tomar si:

a) No se tiene ninguna información acerca de p.

b) En un censo realizado anteriormente la proporción de desempleados fue 0.2.

Solución

a) Si no se tiene ninguna información de p, el tamaño de muestra es:

$$n = \frac{(1.96)^2\,(0.25)}{(0.01)^2} \approx 9{,}604$$

b) Usando la información del censo, resulta $n = \dfrac{(1.96)^2\,(0.2)(0.8)}{(0.01)^2} \approx 6{,}147$.

Estimación del intervalo de confianza para la proporción en el caso de que la población sea pequeña

Si la población es pequeña, de tamaño N, y la muestra se obtiene con el m.a.s. sin reemplazo, el intervalo de confianza para estimar la proporción p se determina como antes, agregando en el margen de error el factor de corrección por población finita.

El intervalo aproximado, al nivel de confianza del $(1 - \alpha)$ 100%, es:

$$[p - z_{1-\alpha/2}\sqrt{p\,(1-p)/n}\,\sqrt{1-n/N},\ p + z_{1-\alpha/2}\sqrt{p\,(1-p)/n}\,\sqrt{1-n/N}]$$

EJEMPLO. *Intención de voto*

Para conocer la intención de voto en un grupo de 500 personas se tomó una muestra sin restitución de 250 personas, resultando que 42 "votarán por el candidato A". Hallar un intervalo al nivel de confianza 95% para estimar la proporción de personas de las 500 que votarán por A.

Solución

Se tiene que la proporción muestral de las personas que votarán por A es $p = 42/250 = 0.168$.

El intervalo al nivel de confianza del 95% para estimar la proporción de los 500 que votarán por A es:

$$[0.168 - (1.96)(0.016),\ 0.168 + (1.96)(0.016)] = [0.136,\ 0.199]$$

En general, el problema de determinar el tamaño de muestra para p en una población finita de tamaño N se resuelve considerando que p es la media de la variable cuyos valores son: 0 y 1. Así resulta que el tamaño de muestra n requerido para estimar a p, con un margen de error de E y para un nivel de confianza del $(1 - \alpha)100\%$, es:

$$n = \frac{p(1-p)}{\left(1 - \frac{1}{n}\right)\left(\frac{E}{z_{1-\alpha/2}}\right)^2 + \frac{p(1-p)}{N}}$$

Observación

Como en el caso de la media, la expresión equivale a $\frac{n_{inf}}{1 + n_{inf}/N}$, en donde n_{inf} es el tamaño de muestra que se obtiene cuando el muestreo es aleatorio simple con restitución o la población es muy grande.

Intervalos de confianza para la diferencia de dos proporciones p_1 y p_2 en dos poblaciones independientes

Se trata de estimar, con un intervalo de confianza, la diferencia $p_1 - p_2$ de las proporciones de elementos que tienen el atributo A en dos poblaciones grandes e independientes.

Una aproximación del intervalo al nivel de confianza del $(1 - \alpha)100\%$ *para la diferencia de proporciones* $p_1 - p_2$ *es:*

$$\left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

en donde $\hat{p}_1$ *y* $\hat{p}_2$ *son las proporciones muestrales que estiman, respectivamente, a* p_1 y p_2, *y* $z_{1-\alpha/2}$ *es el cuantil de orden* $1 - \alpha/2$ *de la normal estándar.*

EJEMPLO. *Fumar es dañino para la salud*

Se quiere saber si en una comunidad existe diferencia significativa entre la proporción p_1 de mujeres que fuman y la proporción p_2 de hombres que fuman. Para ello se realizó una encuesta anotándose que de 800 mujeres, 100 fuman, y de 600 hombres, 120 son fumadores. En tales condiciones, determinar un intervalo de estimación para la diferencia $p_1 - p_2$, al nivel de confianza de 99%.

Solución

Los valores respectivos de los estimadores para p_1 y p_2 son: $\hat{p}_1 = 100/800 = 0.125$ y $\hat{p}_2 = 120/600 = 0.20$.

El intervalo de confianza para la diferencia $p_1 - p_2$, al nivel del 99%, es:

$$\left[(0.125 - 0.20) - 2.58\sqrt{\frac{0.125(0.875)}{800} + \frac{0.20(0.80)}{600}},\ (0{,}125 - 0.20) + 2.58\sqrt{\frac{0.125(0.875)}{800} + \frac{0.20(0.80)}{600}}\right] =$$

[−0.126, −0.023].

El intervalo indica que, al nivel de confianza de 99%, se puede aceptar que p_1 es menor que p_2.

Intervalos de confianza para la diferencia de dos proporciones p_i y p_j en una sola población grande

El estudio anterior contemplaba la independencia de las muestras de las dos poblaciones relacionadas con las proporciones que se deseaban comparar; sin embargo, existen situaciones en donde se comparan proporciones de una sola población y a partir de una sola muestra. Tal caso se tiene cuando se requiere la comparación de las preferencias de los electores hacia dos candidatos a una representación pública en una comunidad.

Para comparar dos o más proporciones p_i y p_j en una población, a partir de los datos de una sola muestra de tamaño n, se utiliza el siguiente intervalo de confianza para la diferencia de dos proporciones.

$$\left[(\hat{p}_i - \hat{p}_j) - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}},\ (\hat{p}_i - \hat{p}_j) - z_{1-\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j}{n}}\right]$$

Así, para estimar la diferencia de las proporciones de las preferencias por dos candidatos A y B a la alcaldía de una ciudad, bastará con tomar una muestra aleatoria de los votantes en la ciudad, calcular en esta muestra las proporciones muestrales $\hat{p}_A$ y $\hat{p}_B$ que indican las preferencias de los ciudadanos por los candidatos A y B, respectivamente, en la muestra y luego calcular el intervalo indicado.

EJEMPLO. *Candidaturas*

En una encuesta realizada en 500 personas se determinó que 125 apoyaban al candidato A, 140 apoyaban a B y el resto apoyaban al candidato C.

El intervalo al nivel de confianza de 95% para la diferencia de proporciones de los que apoyan el candidato A y al candidato B es:

$$\left[(0.25 - 0.28) - 1.96\sqrt{\frac{0.25(0.75) + (0.28)(0.72) + (2)(0.25)(0.28)}{500}},\right.$$
$$\left.(0.25 - 0.28) + 1.96\sqrt{\frac{0.25(0.75) + (0.28)(0.72) + (2)(0.25)(0.28)}{500}}\right] =$$

[–0.0938, 0.0338]

De acuerdo al resultado obtenido, no se puede afirmar nada respecto de las proporciones; las proporciones pueden ser iguales o diferentes.

APLICACIÓN: El caso de la guerra de las colas

Las empresas KCola e ICola son dos de las compañías que surten al mercado local de refrescos desde hace muchos años. Últimamente han iniciado una serie de anuncios publicitarios por radio y televisión indicando que cada cual es la poseedora de mayor participación en el mercado de los refrescos. Los avisos son cada vez más numerosos, a tal punto que la prensa local ha bautizado a esta situación como la "guerra de las colas", en clara alusión al famoso film *La guerra de las galaxias*.

En los avisos publicitarios la KCola indica que a ella "la prefiere el 42% de los clientes, mientras que a su competidora ICola la prefiere el 40%". Con letras muy pequeñas, como suele suceder en estos avisos, se indica que esta afirmación se basa en los siguientes resultados obtenidos al tomar una muestra aleatoria simple de 600 clientes.

Refrescos	*Preferencia*
KCola	252
ICola	240
Otros	110
Total	600

Fiser, un avispado lector que ha fijado su atención en las pequeñas letritas de los avisos de las colas, dice que la diferencia que indica KCola es a nivel de muestra pero no necesariamente a nivel de población. Fiser toma su calculadora y obtiene el intervalo de confianza de la diferencia de las proporciones de los clientes que a nivel de población prefieren KCola e ICola al 95%.

El intervalo al 95% de confianza que Fiser calculó fue:

$$\left[(252/400) - 240/400) - 1.96\sqrt{\frac{(252/400)(148/400) + (240/400)(160/400) + (2)(252/400)(240/400)}{400}},\right.$$
$$\left.(252/400) - 240/400) - 1.96\sqrt{\frac{(252/400)(148/400) + (240/400)(160/400) + (2)(252/400)(240/400)}{400}}\right]$$

[–0.08365, 0.13364]

Los resultados no permiten a KCola decir que tiene la preferencia a nivel de población. La diferencia que se nota a nivel de muestra no es significativa.

LA ESTADÍSTICA EN LA EMPRESA

El empresa embotelladora Refrescola

La empresa embotelladora de refrescos Refrescola opera en el país desde el 1950, año en que firmó el convenio que le permitía embotellar y distribuir la bebida gaseosa que "refresca al mundo" hace 120 años en más de 200 países y cuya fórmula "secreta, única, energizante y refrescante" fue inventada en Atlanta, EE. UU.

La planta que Refrescola construyó para desarrollar el negocio y que aún funciona se encuentra situada en el centro, al sur de la capital, y desde este punto se reparte la bebida a todos los puntos del país.

El reconocimiento permanente que el mercado nacional le brinda a la marca de la bebida que Refrescola embotella y reparte se debe al marketing sofisticado e innovador y a las campañas y promociones publicitarias, que han permitido que la bebida sea la más reconocida y que participe en el 50% del mercado de las bebidas gaseosas. En la actualidad, Refrescola brinda trabajo en forma directa e indirecta a 25,000 personas y sus campañas a favor del medio ambiente han sido suficientes para que sea considerada una de las empresas líderes de la responsabilidad social.

La empresa Refrescola tiene entre sus trabajadores un grupo de profesionales para el estudio permanente del agua, del azúcar y de diversos ingredientes que se utilizan en la elaboración de la bebida. En este trabajo, así como en los relacionados con el marketing, Refrescola utiliza como herramienta fundamental a la estadística para llevar a cabo diseños experimentales, para el conocimiento de los clientes, para la predicción de ventas, etcétera.

EJERCICIOS

1. El costo promedio de un seguro anual para automóviles es $ 500 y la desviación estándar es igual a $ 20. Para una muestra de 200 de estos seguros hallar:
   a) El valor esperado de la media muestral.
   b) El error estándar de la media muestral.

2. En la puerta de un ascensor aparece el siguiente aviso: capacidad: "12 personas u 880 kg en promedio". Si en la población el peso de las personas tiene distribución normal con media 70 kg y con una desviación estándar de 10 kg, ¿cuál es la probabilidad de que el peso de 12 personas no supere la capacidad indicada en kilogramos?

3. Se desea estimar la media del tiempo que los taxistas manejan sus taxis durante el día. Para ello se tomó una muestra de 49 taxistas y se calculó la media muestral del tiempo de manejo al día. Hallar la probabilidad de que la media muestral calculada diste menos de 0.5 horas de la verdadera media si la desviación estándar de los tiempos de manejo durante el día es 2 horas (recordar que la media muestral, para un tamaño de muestra mayor que 30, tiene distribución aproximadamente normal).

4. Se desea estimar el volumen de refresco Maracuyá contenido en los envases que se venden en el mercado. Para ello se tomó al azar 300 envases y se determinó que la media muestral del volumen contenido era 0.98 litros con una desviación estándar igual a 0.10 litros. La media muestral se tomó como estimador de la media del volumen de refresco contenido en cada envase.
   a) Encontrar la probabilidad de que el error de estimación sea menor que 0.05 litros.
   b) ¿Cuál es el tamaño de muestra que se recomendaría para estimar la media del volumen contenido en los envases, de tal manera que el error de estimación sea menor que 0.03 litros con probabilidad 0.95?

5. Las ventas diarias que realiza una tienda es 80,000 unidades monetarias (u.m.) con una desviación estándar de 1,000 u.m. Hallar:
   a) La probabilidad de que la venta en un día cualquiera esté entre las 70,000 y 90,000 u.m.
   b) La probabilidad de que el promedio de ventas de 100 días sea menor que 79,000 u.m.
   c) La probabilidad de que el total de ventas en 50 días sea menor que 3,780,000 u.m.

6. El número de autos que vende diariamente un concesionario tiene distribución de Poisson con parámetro λ (el valor esperado de la distribución de Poisson es igual al parámetro λ).
   a) Indicar un estimador para λ si en 20 días la venta total de autos fue de 30.
   b) Indicar un estimador para λ si durante los últimos 30 días se han vendido: 0 autos durante 20 días, y uno en cada uno de los 10 días restantes.

7. Se desea estimar el tiempo, en minutos, que los alumnos de la universidad emplean cada día en llegar de su casa a la universidad. A fin de reducir la variabilidad, pues vienen de diferentes de distritos, se ha decidido considerar dos grupos: los que vienen del este y los que vienen del norte. Para llevar a acabo esta labor se ha seleccionado una muestra aleatoria en cada grupo, hallándose la siguiente información.

	Del este	***Del norte***
Población N_i	3,000	2,000
Muestra n_i	100	64
$\bar{X}_i$	60 minutos	80 minutos
S_i	6 minutos	12 minutos

   a) Hallar un estimador del total del tiempo usado en llegar de su casa a la universidad de los alumnos que vienen del este.

   b) Hallar un estimador del promedio de tiempo que usan los alumnos para llegar de su casa a la universidad.

8. Una distribuidora de revistas realiza envíos de cajas, cada una con 12 revistas de diversos títulos. El peso de cada revista tiene una media de 250 gr y una desviación estándar de 50 gr. Las cajas donde se depositan las revistas tienen un peso medio de 170 gr con una desviación estándar de 15 gr.

   a) Hallar la media y la desviación estándar de los pesos de las cajas que contienen 12 revistas cada una.

   b) Si las cajas se transportan en una avioneta y por razones de seguridad la carga debe pasar de 1,000 kg con probabilidad no mayor a 0.05, ¿cuál es el mayor número de cajas que se puede transportar en cada vuelo?

9. Se ha determinado que la proporción de estudiantes universitarios que siguen la carrera de Administración es 0.10. Si se toma una muestra de 500 estudiantes universitarios y se calcula la proporción de universitarios que estudian la carrera de administración, hallar:

   a) El valor esperado de la proporción muestral.

   b) El error estándar de la proporción muestral.

10. Se supone que la proporción de personas que consumen café instantáneo en un distrito es 0.40. Si se toma una muestra de 500 personas y se calcula la proporción muestral:

   a) ¿Cuál es la probabilidad de que la proporción muestral encontrada esté entre 0.39 y 0.41?

   b) ¿Cuál es la probabilidad de que la proporción muestral encontrada sea menor que 0.38?

11. Se ha estimado la proporción de personas que leen el periódico *ABC* utilizando una muestra de 100 lectores. Si la proporción muestral hallada ha sido 39%, encontrar, de manera aproximada, la probabilidad de que el error que se comete al estimar con este valor la verdadera proporción en toda la población sea menor que 0.05.

12. Después de realizar una prueba del sazonador Saz, los encargados de la comercialización desean estimar la proporción de amas de casa que comprarán el producto próximo a salir a la venta.

    a) ¿Cuál es el tamaño de muestra que se debe tomar si se requiere que el error sea menor que 0.05 con probabilidad 0.95 y se conoce que la proporción de los que compran un producto parecido es el 15% del mercado?

    b) ¿Cuál es el tamaño de muestra que se debe tomar si se requiere que el error de estimación sea a lo más 0.03 con probabilidad 0.99 y no se tiene ninguna información adicional?

13. Un inspector obtiene una muestra aleatoria de 10 cuentas por cobrar de las 500 cuentas de una empresa, registra el valor de cada una de ellas y verifica si corresponde a cierto tipo de ventas A. Los datos obtenidos fueron consignados como se indica en la siguiente tabla.

Cuentas	*Valor*	*Verificación*
1	142	Sí
2	335	Sí
3	290	No
4	219	Sí
5	212	Sí
6	168	Sí
7	305	No
8	188	Sí
9	221	No
10	310	No

    Estimar la cantidad total a cobrar en las 500 cuentas de la empresa. Hallar el error estándar de la media muestral. Estimar la proporción de cuentas que corresponden al tipo de ventas A. Hallar el error estándar para cada estimador.

14. Se desea estimar el promedio del tiempo en que los empleados en una empresa hacen una tarea, de tal manera que el error de estimación no sea mayor a 2 minutos con probabilidad 0.95. Si la desviación estándar de los tiempos necesarios para realizar la tarea es de 3 minutos, hallar el tamaño de muestra a tomar.

15. Se sabe que el 1% de las familias de una ciudad desea comprar un auto nuevo en el próximo año.

    a) Si se elige una muestra aleatoria de 100 familias de la población, ¿cuál es la probabilidad de que en la muestra elegida existan al menos ocho personas que desean comprar auto nuevo el próximo año?

b) ¿Cuál debe ser el tamaño de muestra que se debe tomar para que con probabilidad 0.95 la proporción muestral de las personas que desean comprar un auto nuevo diste de la proporción verdadera en 1% a lo más?

16. Una compañía de seguros se dispone asegurar a 36 personas contra un determinado riesgo. Si llamamos con X_i a la cantidad que la compañía tendrá que pagar al cliente i (X_i puede ser 0) y si el valor esperado que se paga al producirse una contingencia es $E(X_i) = 1000$ y la varianza es $V(X) = 2{,}500$ para todo i; hallar el valor C que debe cobrarse como prima a cada cliente si se desea ganar por lo menos $ 10,000 con probabilidad igual a 0.95.

17. La Oficina de Trabajo indica que, en promedio, los obreros estatales ganan 500 unidades monetarias (u.m.) con una desviación estándar de 20 u.m. ¿Cuál es la probabilidad de que la media de los 81 salarios de igual número de obreros estatales, seleccionados al azar, sea menor que 450 u.m.?

18. Se ha determinado que el tiempo necesario para atender a un paciente en el consultorio de un hospital tiene una distribución de probabilidad en forma de una montaña con una cola a la derecha, pues pocas veces ocurren largos tiempos de consulta. El tiempo medio de consulta por paciente es 0.3 horas con desviación estándar de 0.1 horas. Si cada médico planea atender 30 pacientes en una jornada de 12 horas, ¿qué probabilidad existe de que se cumplan los planes establecidos?

19. Una compañía hotelera observa que con probabilidad 0.12 una reserva de habitación no es cubierta. La compañía decide aceptar reservas por un 10% más de las 450 habitaciones que dispone. Calcular el porcentaje esperado de clientes con reserva que se quedarán sin habitación.

20. En cada operación que realiza una computadora, aproxima el resultado al entero más próximo cometiendo un error que es una variable aleatoria con distribución uniforme de media 0 y de rango 4. Si se ejecutan 50 operaciones, ¿cuál es la probabilidad de que la suma de los errores esté entre −4 y 4?

21. El peso medio de los pasajeros que abordan un avión es 60 kg con una desviación estándar de 6 kg. ¿Cuántos pasajeros deben considerarse para que, con probabilidad 0.95, el peso total de los pasajeros no supere los 2,766 kg?

22. Se considera que el 60% de los habitantes de cierto distrito están a favor de cierta ordenanza municipal. Si se seleccionan al azar 81 habitantes de tal distrito, ¿cuál es la probabilidad de que la fracción de las personas de la muestra a favor de la ordenanza difiera en valor absoluto de la verdadera proporción en no más del 1%?

23. Se desea estimar la proporción de electores que votarán a favor del candidato NN a la alcaldía de una ciudad. ¿Cuántos electores deberá contener una muestra para que, con probabilidad 0.95, la proporción muestral de los que votan por A se desvíe a lo más en 0.05 de la verdadera proporción?

24. El tiempo que demora una persona en hablar por teléfono es una variable aleatoria con media 3 minutos y desviación estándar 0.50 minutos. Hallar la probabilidad de que:

a) El tiempo total que demoran 49 personas sea mayor que 150 minutos.

b) El promedio del tiempo que demoran 64 personas sea mayor que 3.5 minutos.

c) Si por cada minuto se paga $ 0.50 y el costo fijo por llamada es igual a $ 0.80, calcular la probabilidad de que el costo al realizar 36 llamadas sea mayor que $ 45.

25. ¿Qué diferencia existe entre la estimación puntual y la estimación por intervalos?

26. ¿Cuáles son las ventajas de la estimación por intervalos comparada con la estimación puntual?

27. ¿De qué manera afecta a una estimación por intervalo

a) el aumento del tamaño de la muestra?

b) el aumento de la variabilidad en la población?

c) el aumento del nivel de confianza?

28. Para obtener un intervalo de confianza $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ para la media de una población se ha usado en tres ocasiones los siguientes niveles de confianza:

a) 95% b) 99% c) 90%.

En cada caso, ¿cuál es el valor de $z_{1-\alpha/2}$?

En cada caso, ¿cuál es el margen de error de la estimación si el error estándar es 2?

29. Se ha determinado que la cantidad de nicotina que tienen los cigarrillos de cierta marca se distribuye normalmente con desviación estándar igual a 1 miligramo. Para estimar la media de la cantidad de nicotina por cigarrillo de toda la producción, se tomó una muestra de tamaño 25 y se construyó el intervalo de confianza [9.6080, 10.3920]. ¿Con qué nivel de confianza se ha construido el intervalo?

30. Un banco desea estimar el monto medio de los préstamos para automóviles. La experiencia pasada revela que la desviación estándar es $ 250. Determinar una estimación, usando un intervalo de confianza de 95%, para la media poblacional si una muestra aleatoria de 100 préstamos tiene una media muestral de $ 2,000.

31. Con la finalidad de estimar el promedio de los sueldos de 1,000 trabajadores se tomó una muestra de tamaño 100, obteniéndose una media muestral igual a 400 dólares. Hallar el intervalo al nivel de confianza de 95% para la media de todos los sueldos si se supone que la distribución de estos es normal con desviación estándar igual a 30 dólares. Si la media de los sueldos se estima con la media muestral, ¿cuál es el margen de error?

32. Una compañía aérea desea conocer el promedio de asientos desocupados por vuelo de la ciudad A a la ciudad B. Para ello seleccionó al azar 225 registros en los archivos del año pasado, encontrando que la media y la desviación estándar del número de asientos desocupados eran iguales a 10 y 3, respectivamente. Estimar, usando un intervalo al nivel de confianza del 95%, el promedio del número de asientos desocupados por vuelo de la ciudad A a la ciudad B para el próximo año.

33. Una encuesta ha determinado que un aspecto importante relacionado con la satisfacción de los clientes de ventas por correo es la rapidez con que se entregan los pedidos. El gerente de ventas quiere conocer el promedio del tiempo que transcurre entre la llamada de un cliente haciendo un pedido y la contestación de la orden por parte de la compañía. A partir de las órdenes recientes se puede seleccionar una muestra. Se decide trabajar con un nivel de confianza del 95% y con un margen de error de 0.5 días. Como la variabilidad de los tiempos a nivel de toda la población no se conoce, se decide tomar una pequeña muestra para estimar la varianza. Así se encontró como estimación de la desviación estándar de la población al valor 2 días. Encontrar el tamaño de muestra a tomar para estimar la media de la población.

34. Para controlar la calidad de las varillas de aluminio que produce una fábrica para ensamblaje de bicicletas, el ingeniero encargado del control muestrea diariamente en la línea de producción un cierto número de varillas y construye 15 intervalos de confianza para la media de las varillas al nivel de confianza del 95%.
    a) Sea X el número (desconocido) de los intervalos que en efecto cubren la longitud media desconocida de las piezas producidas en el día, ¿cuál es la distribución para X?
    b) Calcular la probabilidad aproximada de que 12 de los 15 intervalos cubran la media verdadera.

35. Cinco determinaciones del pH de una solución dieron los siguientes resultados: 7.29, 7.95, 7.95, 7.50 y 7.94

    Hallar un intervalo de nivel de confianza del 99% de la media de todas las determinaciones del pH de la misma solución, si se supone que la variable de las determinaciones del pH es normal.

36. La ganancia que se obtiene al vender un dispositivo depende del estado de las dos piezas idénticas que lo conforman. Si ambas piezas están buenas la ganancia es 1 u. m. Si una de ellas está defectuosa, la ganancia es 0, y si ambas están defectuosas la ganancia es –1. Si la probabilidad de que una pieza cualquiera esté defectuosa es 0.01, indicar el intervalo que con probabilidad 0.95 contiene a la ganancia total al vender 100 dispositivos.

37. Hallar el tamaño de muestra que se debe tomar para estimar la media de una poblacional normal mediante la media muestral de modo que el error que se produzca no sea mayor que el 20% de la desviación estándar poblacional, con probabilidad 0.95.

38. Se planea una encuesta para conocer el tiempo promedio semanal que los niños ven televisión en una comunidad, con un error no mayor a 0.5 horas y con un nivel de confianza del 99%. Si el costo de administración de la encuesta es de $ 500 más $ 3 por entrevista, ¿cuál es el costo total que se debe presupuestar para la encuesta si se conoce que un estudio anterior indicó que la desviación estándar es de 2 horas?

39. Ochenta personas seleccionadas al azar del total de 225 egresados de la carrera de Administración reciben un sueldo promedio inicial de $ 900 mensuales con una desviación estándar de $ 100. Calcular un intervalo al 95% de nivel de confianza para el sueldo promedio inicial de los 225 graduados.

40. Hallar un intervalo de nivel de confianza del 95% para el ingreso promedio mensual de una comunidad que tiene 500 familias si para una muestra de 100 familias se obtuvo un ingreso mensual promedio de 30,000 pesos. Suponer que los ingresos tienen aproximadamente distribución normal con desviación estándar igual a 5,000 pesos.

41. Con el fin de conocer la proporción p de escuelas que tienen por lo menos una computadora, el Ministerio de Educación efectuó una encuesta que mostró que de 100 escuelas escogidas al azar, solo 20 tenían por lo menos una computadora. Considerando que el número de escuelas es muy grande, hallar:
    a) Un estimador puntual de la proporción p.
    b) Un intervalo al nivel de confianza del 95% para p.
    c) El tamaño de la muestra que se debería tomar para aproximar p de tal manera que con probabilidad 0.95 el error de estimación sea menor que 0.1.

42. Se ha determinado que de 100 personas tomadas al azar de una población de 200,000 habitantes, 20 han sido atacados por parasitosis. Mediante un intervalo de nivel de confianza del 95%, determinar la proporción de personas que han sido atacadas en la población de donde se tomó la muestra.

43. Una encuestadora estima que la proporción de ciudadanos que están a favor de cierta ley está en el intervalo descrito como $30 \pm 5\%$, al nivel de confianza del 99%. Si la estimación se hizo sobre la base de una muestra y tomando el mayor valor de $p(1 - p)$, ¿cuál ha sido el tamaño muestral que se ha usado?

44. De 100 momentos distintos, seleccionados al azar durante una semana de trabajo, se observa que un operador de una máquina realiza trabajo productivo en 80 de estas observaciones.
    a) Indicar un intervalo de confianza al 95% de confianza para la proporción de tiempo en que el operador realiza trabajo productivo.
    b) ¿Cuántas observaciones se necesitan para determinar la verdadera proporción de tiempo productivo durante la semana de trabajo con un margen no mayor de 5 puntos porcentuales y un nivel de confianza del 99%?

45. Una tienda de departamentos desea estimar el porcentaje de sus clientes que compran camisas. ¿De qué tamaño debe ser la muestra si quiere tener una certeza del 95% de que su estimador será correcto dentro de un margen de 2%?

46. Una encuesta realizada por encargo de un canal de TV indica que el candidato Pérez se ve más favorecido que el candidato García por un porcentaje de 43 a 37, con un margen de error igual a 4. El canal dice que como la diferencia de 6 puntos es mayor que el margen de error, sus lectores pueden estar seguros de que el candidato Pérez es el favorito del momento. ¿El razonamiento es correcto?

47. Una empresa que se encarga de medir el desempeño de una emisora radial quiere saber qué proporción de hogares están escuchando un cierto programa radial. Con este objeto planean realizar una encuesta a hogares escogidos aleatoriamente ¿De qué tamaño deberá ser la muestra si quieren tener un 90% de seguridad de que su estimado es correcto dentro de un margen de error del 2%?

48. Un lote de artículos tiene una proporción p de defectuosos. El valor de p no se conoce, pero se sabe que es menor o igual a 0.04. ¿Cuántos artículos deben ser tomados al azar para aproximar p con un error no mayor de 0.02 al nivel de confianza del 95%?

49. El promedio del número de latidos por minuto en 300 pobladores de la costa, tomados al azar, es de 80. En 240 pobladores de la sierra el promedio es 77 por minuto. Si la distribución del número de latidos es normal con desviación estándar de 3 minutos, para cualquiera de los pobladores, hallar un intervalo al nivel de confianza del 95% para el promedio de latidos de los pobladores de la costa y un intervalo de nivel confianza del 99% para la diferencia de las medias poblacionales en ambas regiones.

50. Para comparar las proporciones de artículos defectuosos producidos por dos líneas de producción, se seleccionan muestras aleatorias independientes de 200 artículos en cada línea. En la línea A se encontró el 10% de defectuosos y en la línea B, 14%. Hallar un intervalo, al nivel de confianza del 99%, para la diferencia de las proporciones de artículos defectuosos que producen ambas líneas.
    a) ¿Se podría indicar que la línea A produce una proporción de artículos defectuosos menor que la línea B?
    b) ¿Cuántos artículos de cada línea se debe seleccionar para que un intervalo al nivel de confianza del 95% para la diferencia real de proporciones tenga una amplitud de 0.02? Usar muestras de igual tamaño en cada línea.

51. Para comparar las actitudes de los jóvenes de dos naciones A y B respecto del peligro de una guerra nuclear, se tomó una muestra de 3,370 estudiantes en A y una muestra de 2,148 estudiantes en B. A cada estudiante se le preguntó si era posible una próxima guerra nuclear. 20% de los estudiantes en A y 29% de B contestaron afirmativamente.
    a) Calculando un intervalo al 99% de nivel de confianza, ¿se podría decir que las proporciones de los estudiantes de las dos naciones que piensan afirmativamente son diferentes?
    b) ¿Cómo podría reducirse el tamaño del intervalo de confianza a la mitad de su longitud?

52. Se desea conocer el efecto que producirá una campaña publicitaria para incrementar la proporción de compradores del refresco Limón. Muestras aleatorias e independientes de consumidores de refrescos se toman antes y después de la campaña publicitaria, y se les pregunta si han escuchado hablar de Limón. Si se desea estimar la diferencia de proporciones de los consumidores que reconocen al refresco en mención dentro de un margen de error del 5% con una probabilidad del 95%, ¿cuántas personas se deben encuestar antes y después de la campaña? (asumir que los tamaños de las muestras antes y después de la campaña son iguales).

53. Entre 600 de las 1,000 personas que trabajan en una mina se hizo una campaña para prevenir accidentes. Luego de un tiempo, 7 personas de las que recibieron instrucción sufrieron accidentes. De las personas no instruidas, 8 resultaron accidentadas. Use un intervalo al nivel de confianza del 95% para decir si las charlas dieron resultado.

RESPUESTAS A LOS EJERCICIOS

1. a) 500, b) 1.4142. **2.** 0.8758. **3.** 0.9199. **4.** a) 1 aprox., b) 43. **6.** a) 1.5, b) 1/3. **7.** b) 67.80. **8.** a) 3,170, b) 30,225. **9.** a) 0.10, b) 0.0135. **10.** 0.3519. **12.** a) 196 aprox., b) 1849 **13.** Estimación del total: 119,500. **14.** 9 aprox. **15.** 0 aprox. **16.** Al menos 1,291.49. **17.** 0 aprox. **18.** 1 aprox. **19.** 0 aprox. **20.** 0.3758. **21.** 45 aprox. **23.** El tamaño de muestra es $n = \frac{(1.96)^2\, p(1-p)}{(0.05)^2}$. Sin embargo, $p(1-p)$ no se conoce. Puede tomarse el mayor valor que esta expresión puede tener; esta es ¼, y así se tiene que n es aproximadamente igual a 385. **27.** a) Disminuye el margen de error, b) Aumenta el margen de error. **28.** a) 1.96. El margen de error es 3.92, b) 2.58. El margen de error es 5.16. **29.** 95% **30.** [1951, 2049]. **31.** 5.5810. **32.** [9.608, 10.392]. **33.** 62 aprox. **34.** a) Binomial con parámetros 15, 0.95. **35.** [7.3804, 8.0715]. **37.** 96. **38.** 821. **39.** [882.3692, 917.6307]. **41.** [0.1216, 0.2784]. **43.** 666. **45.** 2401. **48.** 369. **49.** Para la costa: [79.6605, 80.3394], para la diferencia: [2.4907, 3.5092]. **50.** El intervalo es [−0.1237, 0.0437].

CAPÍTULO 8

Pruebas de hipótesis

Jerzy Neyman

Jerzy Neyman nació en Rusia, en 1894. Comenzó sus estudios en la especialidad de Física en la Universidad de Kharkov, en 1912; sin embargo, su falta de habilidad en el laboratorio lo obligó a cambiarse a la carrera de Matemáticas.

Después de la Primera Guerra Mundial y luego de estar prisionero en Rusia, en 1921 fue a Polonia en virtud de un intercambio de prisioneros. En Polonia, en 1924, se doctoró en la Universidad de Warsav. En la Universidad de Kraków, Neyman trabajó con Karl Pearson entre 1924 y 1934. Posteriormente desempeñó cargos importantes en el University College de Londres hasta 1938. Neyman prosiguió su trabajo en la Universidad de Berkeley, en donde estableció el centro más importante en investigación teórica estadística.

Neyman ha sido considerado como el padre de la estadística moderna; son famosos sus trabajos sobre pruebas de hipótesis, intervalos de confianza y muestreo. Recibió importantes premios y honores.

Jerzy Neyman murió en 1981, en California, EE. UU.

CONTENIDO

8.1 Introducción

A menudo se leen o escuchan expresiones como las siguientes:

"Los cigarrillos de la marca DD tienen, en promedio, por lo menos 10 miligramos de nicotina".

"Esta pasta dental es recomendada por el 90% de los dentistas".

"El promedio del tiempo de realización de la tarea es mayor que 20 minutos".

"En promedio, las pilas de la marca A duran más que las pilas de la marca B".

Se trata de *conjeturas* referidas a parámetros (la media, la proporción, etc.) de una distribución asociada a una población. Provienen, a menudo, de muestras aleatorias, por ello es importante averiguar si se deben a la casualidad o reflejan la realidad.

Los procedimientos que permiten decidir el rechazo o no de tales conjeturas se llaman *pruebas o contrastes de hipótesis.*

8.2 Pruebas de hipótesis. Conceptos básicos

Una prueba de hipótesis se basa en el resultado obtenido de una muestra aleatoria, y su objetivo es probar si este resultado es significativamente diferente o no de lo que se afirma acerca de un parámetro de una población. Si se puede probar que la diferencia observada se debe al carácter aleatorio de la muestra, diremos que la diferencia no es significativa y que la conjetura no debe ser rechazada; si la diferencia observada no se debe a la aleatoriedad de la muestra, se dirá que la diferencia observada es significativa y la conjetura deberá ser rechazada.

La hipótesis nula y la hipótesis alternativa

La prueba o contraste de una hipótesis comienza por establecer lo que se llama la ***hipótesis nula***. Esta hipótesis, denotada con H_0, está referida generalmente a la conjetura que se hace sobre un parámetro de una población. Conjuntamente con la hipótesis nula se establece la ***hipótesis alternativa*** H_A, que expresa, como su nombre lo indica, alguna característica alternativa a la hipótesis nula.

Típicamente, la hipótesis nula es la que no tiene suficiente respaldo, y es posible que sea rechazada; sin embargo, esta debe mantenerse a menos que existan pruebas contundentes en contra de ella. En el caso de que suceda lo último, se aceptará la hipótesis alternativa. Si no existieran las pruebas de rechazo, la conclusión no será la aceptación de la hipótesis nula; simplemente se indicará que el procedimiento aplicado no es lo suficientemente fuerte como para rechazar la hipótesis nula.

Si se desea emplear las pruebas de hipótesis para sustentar alguna aseveración proveniente de una investigación, se recomienda escribir la aseveración de tal manera

que se convierta en la hipótesis alternativa. Redactada de esta manera, la decisión de rechazar la hipótesis nula para aceptar la hipótesis alternativa (la aseveración del investigador) podrá ser evaluada por el procedimiento de prueba que a continuación se desarrolla.

EJEMPLO. *¿El gerente de producción tiene razón?*

Al analizar un proceso, el gerente de producción de una empresa afirma que el promedio del tiempo de realización de una determinada tarea es al menos de siete minutos. Esta afirmación ha originado una serie de controversias con los trabajadores de la empresa, quienes aseguran que el tiempo promedio de realización de la tarea es menor que el indicado por el gerente.

La hipótesis nula H_0, del gerente, y la hipótesis alternativa H_A, de los trabajadores, están referidas a la media μ de la variable X, que indica el tiempo de realización de la tarea, y pueden escribirse de la siguiente manera:

Hipótesis nula: $H_0 : \mu \geq 7$

Hipótesis alternativa: $H_A : \mu < 7$

EJEMPLO. *Las malas prácticas*

Se determinó hace cierto tiempo que el 20 por ciento de vehículos en una ciudad no tiene los documentos en regla. Ante esta mala práctica de los conductores, la Dirección de Educación Vial llevó a cabo un plan de educación vial tendiente a solucionar este problema. Después de cierto tiempo, la dirección indicó que el plan había sido un éxito.

Las aseveraciones que en el enunciado se indican pueden traducirse en una hipótesis nula y una alternativa. La hipótesis nula corresponde a la proporción inicial de vehículos que no tenían los documentos en regla. La hipótesis alternativa es la aseveración de la Dirección de Educación. Esto es:

Hipótesis nula: $H_0 : p = 0.20$

Hipótesis alternativa: $H_A : p < 0.20$

Errores de tipo I y II. Nivel de significación

Como las pruebas de hipótesis se basan en la información obtenida en una muestra aleatoria y no en toda la población, existe el riesgo de tomar decisiones erróneas. Y es así que pueden cometerse los siguientes errores.

El ***error de tipo* I**, que ocurre al rechazar la hipótesis nula siendo esta verdadera.

El ***error de tipo* II**, que ocurre al no rechazar la hipótesis nula siendo esta falsa.

TABLA 8.1 *Errores en las pruebas de hipótesis*

	Estado real	
Decisión	H_o es verdadera	H_o es falsa
Rechazar H_o	**Error de tipo I**	
No rechazar H_o		**Error de tipo II**

Como no se conoce el estado real de la situación, es imposible medir *exactamente* los errores que se cometen; sin embargo, usando la probabilidad se podrá obtener cierta información acerca de ello.

A la probabilidad de cometer el error de tipo I se le denota con α y se le llama ***nivel de significación*** *de la prueba*. A la probabilidad de cometer el error de tipo II se le denota con β:

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ es verdadera}]$$

$$\beta = P[\textit{Aceptar } H_0 \mid H_0 \textit{ es falsa}]$$

El nivel de *significación* α proporciona una medida de la confiabilidad de la decisión de rechazar la hipótesis nula y aceptar la alternativa. Los valores que suelen usarse para α son: 0.05 y 0.01. Si el costo de cometer el error de tipo I es alto, se usará de preferencia el valor 0.01.

El *valor de* β es una medida de la confiabilidad de la decisión de aceptar la hipótesis nula; sin embargo, no es posible, en general, controlar este valor.

En muchas de las pruebas se especifica el nivel de significación α. En estos casos, si se decide rechazar la hipótesis nula será posible medir el error que se puede cometer, pero no se puede hacer lo mismo si la decisión es aceptar la hipótesis nula. Por ello, la hipótesis nula simplemente se rechaza o no se rechaza.

Estadísticos de prueba y región de rechazo

Siguiendo con el ejemplo *¿El gerente de producción tiene razón*?, que se refiere a los tiempos de realización de la tarea, tenemos lo siguiente.

Planteadas las hipótesis, el procedimiento para contrastarlas consiste en elegir una muestra al azar, por ejemplo, de 16 tareas similares, medir los tiempos de realización y luego calcular su media muestral $\overline{x}$ (parece razonable que se calcule este valor si se desea probar hipótesis relativas a la media de una población).

Si el valor de $\overline{x}$ encontrado es "significativamente" menor que 7 (el grado de discrepancia entre la hipótesis nula y la muestra es grande), se podrá tomar la *decisión de rechazar* la hipótesis nula. Por ejemplo, si para la muestra elegida la media muestral es menor que 2, podremos rechazar la hipótesis nula, pues al parecer es

casi imposible que siendo la media igual a 7 se pueda hallar una media muestral menor o igual que 2. Pero, ¿rechazaríamos la hipótesis si la media muestral encontrada es 6.8? ¿Se puede considerar ahora que el grado de discrepancia es grande? El problema es cómo medir el grado de discrepancia entre lo encontrado en la muestra y lo que se afirma en la hipótesis nula. Los estadísticos han establecido una medida de esta discrepancia. Esta medida se llama ***estadístico de prueba***. Con el estadístico de prueba se compara lo hallado en la muestra con lo que se indica en la hipótesis nula.

La forma del estadístico de prueba depende del parámetro sobre el cual se hace la conjetura, y lo recomendable es utilizar estadísticos cuya distribución sea conocida.

Para el ejemplo, el estadístico de prueba está determinado por la expresión:

$$z = \frac{\bar{x} - 7}{\sigma / \sqrt{n}}$$

Con este estadístico se compara la media muestral ($\bar{x}$) con la media (7) indicada en la hipótesis nula, en términos de la desviación estándar de la media muestral.

Si la probabilidad de encontrar valores menores o iguales al valor del estadístico de prueba es baja, digamos menores o iguales que 0.05 o 0.01, se considera que existe discrepancia suficiente entre lo observado en la muestra y lo indicado en la hipótesis nula, por lo tanto, la hipótesis nula deberá ser rechazada.

Observación

En el ejemplo, si en lugar de 7 se toma cualquier valor *A* mayor que 7, el valor $\frac{\bar{x} - A}{1/\sqrt{16}}$ será menor que $\frac{\bar{x} - 7}{1/\sqrt{16}}$. Por ello es suficiente analizar esta última expresión (con $A = 7$) para rechazar o no la hipótesis nula.

Esta observación permite plantear la hipótesis nula simplemente en términos de la relación igual:

$$H_0 : \mu = 7$$

Teniendo en cuenta esta observación, las hipótesis nulas se escribirán siempre como una igualdad.

El conjunto de valores del estadístico de prueba para los cuales se considera que existe discrepancia entre lo hallado en la muestra y lo que indica la hipótesis nula se llama ***región de rechazo***.

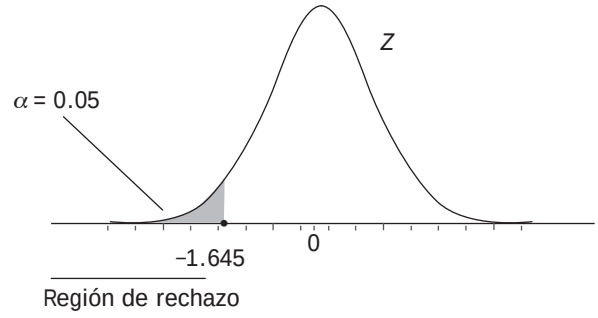
En el ejemplo, la región de rechazo está formada por valores menores que un cierto valor *C*, llamado *valor crítico de la prueba.* Este valor *C* se determina a partir de la distribución del estadístico de prueba cuando la hipótesis nula es verdadera

y del nivel de significación de la prueba. Por ejemplo, si el nivel de significación es $\alpha = 0.05$, el valor crítico de la prueba es el valor *C,* para el cual se cumple:

$$0.05 = P\left[\frac{\bar{x} - 7}{\sigma/\sqrt{n}} < C \mid H_0 \text{ es verdad}\right]$$

Si la población de donde proviene la muestra es normal, el estadístico de prueba tiene distribución normal estándar, y por ello el valor *C* es igual a –1.645, y así se tiene que la región de rechazo es el intervalo]–∞, –1.645[. El valor –1.645 es el cuantil de orden 0.05 de la normal estándar.

FIGURA 8.1 *Región de rechazo de la prueba*



Se concluye que la hipótesis nula se rechaza si el valor del estadístico de prueba es menor que –1.645; es decir, si $\frac{\bar{x} - 7}{\sigma/\sqrt{n}} < -1.645$. Esto equivale a decir que la hipótesis nula se rechaza si $\bar{x} < 7 - 1.645(\sigma/\sqrt{n})$.

En el ejemplo, para el nivel de significación $\alpha = 0.05$, si se encontrara que $\bar{x} = 6$, la hipótesis nula se rechazaría, pues el valor del estadístico de prueba $z = \frac{6 - 7}{1/\sqrt{16}} = -4$ es menor o igual a –1.645.

Obviamente, el deseo es que los errores que se cometan en una prueba de hipótesis sean mínimos; desafortunadamente, las relaciones entre las probabilidades de los dos tipos de errores son de tal naturaleza que cuando se reduce un tipo de error, se incrementa el otro. Lo que resta es poner más cuidado en el error que se considere más importante. Generalmente se conviene en fijar el nivel de significación α y construir una prueba para la cual se obtenga el menor valor de β. En el lenguaje estadístico, a las pruebas que siguen este criterio se les llaman *pruebas de hipótesis uniformemente más potentes*. Las pruebas que se presentan son de este tipo.

El valor *p* de la prueba

Como se indicó, el proceso de probar una hipótesis nula se reduce a medir el grado de discrepancia entre lo observado en la muestra y lo que se indica en la hipótesis nula. Este proceso equivale a calcular la probabilidad de tener un valor más extremo o igual que el valor observado en la muestra cuando la hipótesis nula es verdadera. A esta probabilidad se le llama ***grado de significación o valor*** *p.* Si este valor es un valor pequeño, digamos menor o igual que 0.05 o 0.01, se considera que existe un alto grado de discrepancia entre lo observado en la muestra y lo indicado en la hipótesis nula, y por lo tanto esta se rechaza.

En el ejemplo que se está tratando, usando la estandarización, se encuentra que la probabilidad de encontrar una media muestral menor o igual a 6, cuando la hipótesis nula es verdadera, es 0:

$$P[\bar{X} \leq 6 \mid H_0 \text{ es verdadera}] = P[\bar{X} \leq 6 \mid u = 7] = P\left(\frac{\bar{X} - 7}{1/\sqrt{16}} \leq \frac{6 - 7}{1/\sqrt{16}}\right) = P(Z \leq -4) = 0$$

Esto significa que el valor *p* de la prueba es 0. Por tanto, considerando que la hipótesis nula es verdadera, si se hallara una media muestral menor o igual que 6, lo más razonable será considerar que la hipótesis nula no es verdadera.

Puede considerarse que el *valor p* de la prueba es el riesgo que tiene el decisor cuando, al obtener un cierto valor de la media muestral, rechaza la hipótesis nula. Si el *valor p* es 0, el riesgo que tiene el investigador al rechazar la hipótesis nula es 0. Si el *valor p* fuera 0.30, el riesgo que tendría el investigador al rechazar la hipótesis nula cuando esta es verdadera sería 0.30; el riesgo sería alto; por ello, si esto ocurriera, el decisor se abstendría de rechazarla.

El cálculo del *valor p* depende de la distribución del estadístico de prueba, y es cada vez más pequeño a medida que la discrepancia entre la hipótesis nula y la media muestral observada sea mayor.

> *En la práctica, la hipótesis nula se rechaza cuando el valor de p es menor que el nivel de significación α dado.*

8.3 Pruebas de hipótesis relativas a medias y proporciones

Pruebas de hipótesis referentes a la media de una población

De manera resumida se presentan algunos procedimientos para contrastar hipótesis relativas a la media μ de una población normal.

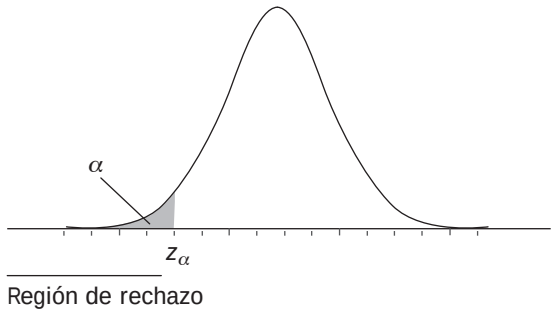
Caso 1. Población normal con desviación estándar σ conocida

En este caso, el proceso a seguir es el siguiente:

1. Plantear las hipótesis: nula $H_0 : \mu = \mu_0$ y alternativa H_A.
2. Indicar el nivel de significación α.
3. El estadístico de prueba en este caso es $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$. Teóricamente, este valor corresponde a una distribución normal estándar cuando la hipótesis nula es verdadera. Con este valor se compara la media muestral hallada con la media indicada en la hipótesis nula, en términos del error estándar de la media muestral.
4. Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu < \mu_0$, la regla de decisión es la siguiente.

FIGURA 8.2 *Región de rechazo de la prueba*



Rechazar la hipótesis nula si $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < z_\alpha$.

z_α es el cuantil de la normal estándar de orden α.

Si $\alpha = 0.05$, $z_\alpha = -1.645$ y la región de rechazo es $]-\infty, -1.645[$.

Si $\alpha = 0.01$, $z_\alpha = -2.33$ y la región de rechazo es $]-\infty, -2.33[$.

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu > \mu_0$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si $z_0 = \dfrac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}} > z_{1-\alpha}$.

  $z_{1-\alpha}$ es el cuantil de la normal estándar de orden $1 - \alpha$.

  Si $\alpha = 0.05$, $z_{1-\alpha} = 1.645$ y la región de rechazo es $]1.645, +\infty[$.

  Si $\alpha = 0.01$, $z_{1-\alpha} = 2.33$ y la región de rechazo es $]2.33, +\infty[$.

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu \neq \mu_0$, la regla de decisión es la siguiente:
  Rechazar la hipótesis nula si:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -z_{1-\alpha/2} \quad \text{o} \quad z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_{1-\alpha/2}$$

  $z_{1-\alpha/2}$ es el cuantil de la normal estándar de orden $1 - \alpha/2$.
  Si $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$ y la región de rechazo es $]-\infty, -1.96[\cup]1.96, +\infty[$.
  Si $\alpha = 0.01$, $z_{1-\alpha/2} = 2.58$ y la región de rechazo es $]-\infty, -2.58[\cup]2.58, +\infty[$.

Recogidos los valores de la muestra aleatoria y antes de aplicar los procedimientos se deberá comprobar si se cumplen los supuestos teóricos. En este caso, la población de donde proviene la muestra debe ser normal; sin embargo, la normalidad se puede relajar, pues lo que se requiere realmente es que la media muestral tenga distribución normal, y esto se puede lograr tomando muestras suficientemente grandes.

La Tabla 8.2 resume los casos que se presentan cuando la varianza se conoce.

TABLA 8.2 *Estadístico de prueba:* $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1)$

Hipótesis nula	*Hipótesis alternativa*	*Región de rechazo*
$H_0: \mu = \mu_0$ σ^2 conocida	$H_A: \mu < \mu_0$	$]-\infty, z_{\alpha}[$
	$H_A: \mu > \mu_0$	$]z_{1-\alpha}, +\infty[$
	$H_A: \mu \neq \mu_0$	$]-\infty, -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}, +\infty[$

Caso 2. Población normal con desviación estándar σ no conocida

1. Si la varianza no se conoce, el estadístico para una prueba de hipótesis con la hipótesis nula $H_0 : \mu = \mu_0$ es $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}}$, en donde s es el estimador de la varianza no conocida e igual a:

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

   Si la hipótesis nula $H_0 : \mu = \mu_0$ es verdadera, el estadístico de prueba corresponde a una variable con distribución *t-student* con $n - 1$ grados de libertad.

2. Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu < \mu_0$, la regla de decisión es la siguiente:

Rechazar la hipótesis nula si $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} < t_{\alpha,\,(n-1)}$

$t_{\alpha,(n-1)}$ es el cuantil de orden α de la distribución *t student* con $n - 1$ grados de libertad.

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu > \mu_0$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si $t = \dfrac{\bar{x} - \mu_0}{s / \sqrt{n}} > t_{1-\alpha,\,(n-1)}$

  $t_{1-\alpha,\,(n-1)}$ es el cuantil de orden $1 - \alpha$ de la distribución *t student* con $n - 1$ grados de libertad.

- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu \neq \mu_0$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} < -\,t_{1-\alpha/2,\,(n-1)} \quad \text{o} \quad t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} > t_{1-\alpha/2,\,(n-1)}$$

En la Tabla 8.3 se indican las zonas de rechazo cuando la varianza no se conoce.

TABLA 8.3 *Estadístico de prueba:* $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$

Hipótesis nula	*Hipótesis alternativa*	*Región de rechazo*
$H_0: \mu = \mu_0$ σ^2 no conocida	$H_A: \mu < \mu_0$	$]-\infty,\ t_{\alpha,\,(n-1)}[$
	$H_A: \mu > \mu_0$	$]t_{1-\alpha,\,(n-1)},\ +\infty[$
	$H_A: \mu \neq \mu_0$	$]-\infty,\ -t_{1-\alpha/2,\,(n-1)}[\cup]t_{1-\alpha/2,\,(n-1)},\ +\infty[$

EJEMPLO. *Ventas*

El gerente de ventas de un centro comercial, preocupado por la disminución de las ventas diarias de los productos que ofrecen, decidió llevar a cabo un plan de propaganda. Antes de la propaganda, el promedio de ventas diarias era de 10, en miles de dólares. Después de un cierto tiempo, el gerente ha optado por llevar a cabo la evaluación del plan, y para ello ha tomado una muestra aleatoria de las ventas de 9 días. Los valores de la media y de la desviación estándar muestrales han sido los siguientes:

$$\bar{x} = 10.48 \ \ y \ \ s = \sqrt{\frac{1}{8}\sum_{i=1}^{9}(x_i - \bar{x})^2} = 1.36$$

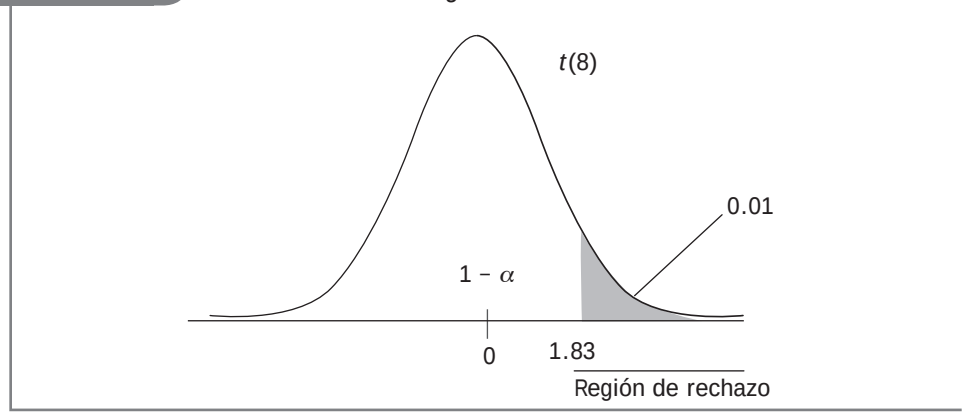
Suponiendo que no han influido otros factores diferentes a la propaganda realizada, y si esta ha sido efectiva, el gerente deberá mostrar que el valor de la media observada es significativamente mayor que 10 (valor de la media de las ventas diarias, en miles de dólares, antes de la propaganda) y que lo observado no se debe al azar. Es decir, deberá contrastar la hipótesis nula H_0: $\mu = 10$ frente a la hipótesis alternativa H_A : $\mu > 10$.

El gerente ha supuesto que la variable X, que representa las ventas después de la propaganda, es normal y ha decidido usar el nivel de significación $\alpha = 0.01$.

El valor del *estadístico de prueba* es $\dfrac{\bar{x} - \mu}{s / \sqrt{n}} = 1.058$.

Los valores del estadístico de prueba tienen distribución *t*-student con 9 – 1 = 8 grados de libertad, y según la Figura 8.3 y la tabla del apéndice B, la región de rechazo de la prueba, al nivel de significación $\alpha = 0.01$, es]1.83, +∞[.

FIGURA 8.3 *Distribución t con* 8 *grados de libertad*



Como el valor del estadístico no cae en la región de rechazo, podemos indicar que la diferencia que se observa entre el valor de la media muestral y la media de las ventas antes de la propaganda no es significativo. La información obtenida no permite indicar que la propaganda ha tenido efecto favorable.

EJEMPLO. *Proyecto de construcción de baterías para computadoras portátiles*

Un experto desea vender a una compañía de computación un proyecto para construir baterías para computadoras portátiles, que según él tendrán una duración promedio mayor que 5 horas. El responsable de la compañía duda de esta aseveración, por ello se propone, mediante una prueba de hipótesis, contrastar la hipótesis del experto versus la hipótesis H_A: $\mu < 5$. Se supone que el tiempo de duración X, en horas, es normal con una desviación estándar igual a 0.2 horas.

Para llevar a cabo el proceso de prueba, el experto propone que si su hipótesis es verdadera, el procedimiento utilizado la rechace con una probabilidad $\alpha = 0.05$. Por otra parte, el director de la compañía exige a su vez que si el tiempo de duración promedio fuera $\mu = 4.9$ la hipótesis nula sea aceptada por el procedimiento con una probabilidad $\beta = 0.01$ (la dirección se cuida de aceptar la propuesta del experto en la eventualidad de que esta no sea buena). Con estas condiciones, ¿cuál debe ser el tamaño de muestra a tomar y la región de rechazo correspondiente?

Solución

Supongamos que para una muestra de tamaño *n* (a determinar) el valor de la media muestral es $\bar{x}$.

Si $\bar{x} < C$ (*C* es una constante menor que 5, a determinar), se rechaza la hipótesis nula: $\mu = 5$.

La exigencia del experto es que la probabilidad de rechazar la hipótesis nula, siendo esta verdadera, debe ser $\alpha = 0.05$; luego:

$$P(\bar{X} < C, \text{ dado que } \mu = 5) = 0.05 \quad \text{(a)}$$

La exigencia del director es que la probabilidad de aceptar la hipótesis nula cuando $\mu = 4.9$ debe ser $\beta = 0.01$, de ahí que:

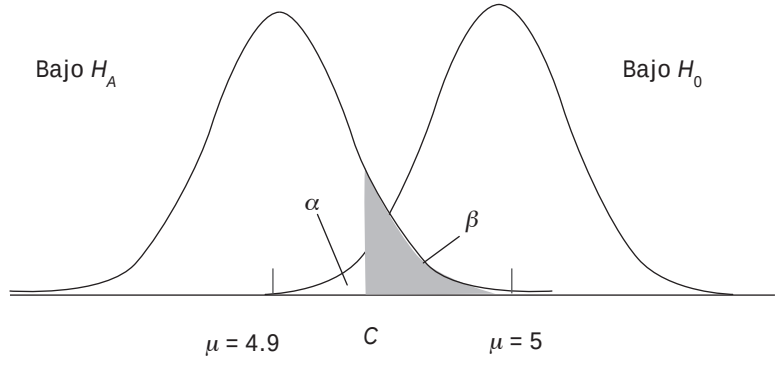
$$P[X \geq C, \text{ dado que } \mu = 4.9] = 0.01 \quad \text{(b)}$$

(Las relaciones (a) y (b) están representadas en la Figura 8.4.)

La relación (a) equivale a:

$$P\left[\frac{\bar{X} - 5}{0.2/\sqrt{n}} < \frac{C - 5}{0.2/\sqrt{n}}\right] = 0.05 \text{ o } P\left[Z < \frac{C - 5}{0.2/\sqrt{n}}\right] = 0.05$$

FIGURA 8.4 *Distribuciones normales para cada media*



La relación (b) equivale a:

$$P\left[\frac{\bar{X}-\mu}{0.2/\sqrt{n}} \geq \frac{C-4.9}{0.2/\sqrt{n}}\right] = 0.01 \quad \text{o} \quad P\left[Z \geq \frac{C-4.9}{0.2/\sqrt{n}}\right] = 0.01$$

Usando la tabla de la normal estándar $N(0, 1)$ y las relaciones anteriores se obtiene:

$$\frac{C-5}{0.2/\sqrt{n}} = -1.645 \quad \text{y} \quad \frac{C-4.9}{0.2/\sqrt{n}} = 2.33$$

Despejando C en ambas ecuaciones:

$$C = 5 + \frac{0.2}{\sqrt{n}}(-1.645), \quad C = 4.9 + = \frac{0.2}{\sqrt{n}}(2.33)$$

Resolviendo para n y C, se tiene:

$$n = \frac{(1.645 + 2.33)^2(0.2)^2}{(5-4.9)^2} \approx 63.04 \quad y \quad C = 4.96$$

El procedimiento a seguir es el siguiente: tomar una muestra del tamaño de 63 pilas construidas por el experto y calcular la media muestral del tiempo de duración. Si la media muestral es menor que 4.96, entonces se rechaza la hipótesis del experto.

Pruebas de hipótesis referentes a las medias de dos poblaciones normales

En muchos casos existe la necesidad de comparar dos procesos a partir de los resultados de dos muestras; así, se "comparan los tiempos" que duran dos procedimientos para fabricar un elemento electrónico usando dos máquinas diferentes A y B. El problema se traduce generalmente en analizar si las diferencias que se observan en las medias muestrales obtenidas para cada proceso indican una diferencia real entre las medias de las poblaciones o es que solo pueden ser atribuidas al azar. La prueba puede traducirse como el estudio de los efectos de un factor (máquina) en los tiempos de realización cuando se toman diferentes niveles (máquina A y máquina B). Los conceptos necesarios para el análisis de este caso y de otros en general se desarrollan a continuación.

Caso 1: Las desviaciones estándar de las variables se conocen

En resumen, para comparar dos procesos o tratamientos se puede seguir el siguiente procedimiento:

1. Asegurar que todos los factores (excepto aquel cuyo efecto se desea analizar) que pudieran influir en los resultados afecten de igual manera en las dos muestras a tomar. Si no es posible identificar todos los factores que podrían influir en los resultados, se deberá aplicar de manera aleatoria los tratamientos para evitar sesgos.

2. Recogidas las muestras x_{11}, $x_{12,\ldots,}$ x_{1n_1} y x_{21}, $x_{22,\ldots,}$ $x_{2n_2,}$ una en cada población y obtenidas al aplicar cada tratamiento, respectivamente, comprobar si se cumplen los supuestos teóricos. En este caso, las variables X_1 y X_2, que representan a las poblaciones de donde provienen las muestras, deben ser normales e independientes. Sin embargo, la normalidad se puede relajar, pues lo que se requiere realmente es que las medias muestrales tengan distribución normal, y esto se puede lograr tomando muestras grandes. La independencia de las poblaciones puede analizarse a partir de la procedencia de los datos.

TABLA 8.4 *Mediciones para cada tratamiento*

Tratamientos	*Muestras: mediciones*	*Medias muestrales*
Tratamiento 1	$x_{11}, x_{12},\ldots. x_{1n_1}$	$\bar{x}_1$
Tratamiento 2	$x_{21}, x_{22},\ldots. x_{2n_2}$	$\bar{x}_2$

3. El estadístico para comparar las medias μ_1 y μ_2 de X_1 y X_2, respectivamente, es:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

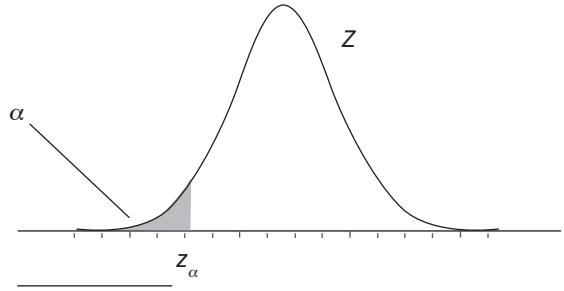
Si la hipótesis nula es verdadera y las varianzas: σ_1^2 de X_1 y σ_2^2 de X_2 se conocen, los valores del estadístico corresponden a una distribución normal estándar.

4. Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 < \mu_2$, la regla de decisión es la siguiente:
  Rechazar la hipótesis nula si:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_\alpha$$

FIGURA 8.5 *Región de rechazo*

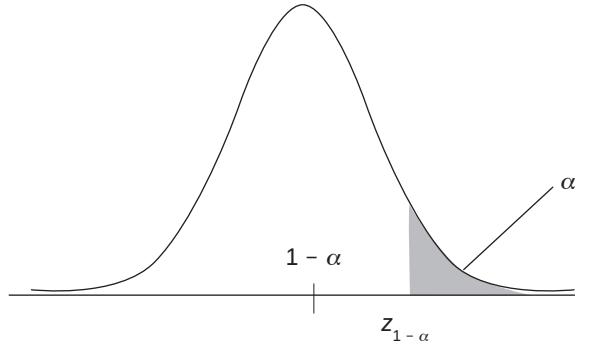


- Para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 > \mu_2$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha}$$

FIGURA 8.6 *Región de rechazo*

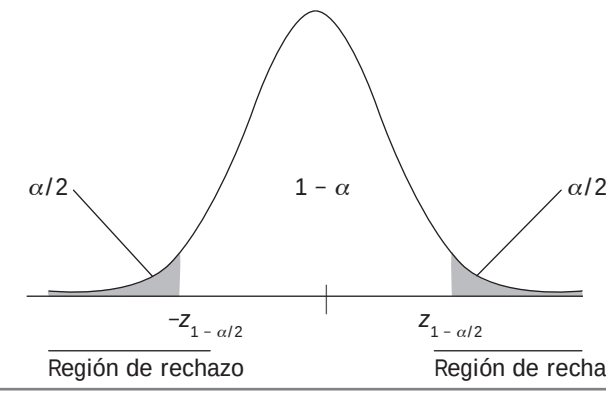


- Para contrastar la hipótesis nula $H_0 : \mu = \mu_0$ frente a la hipótesis alternativa $H_A : \mu \neq \mu_0$, la regla de decisión es:

  Rechazar la hipótesis nula si:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{1-\alpha/2} \quad \text{o} \quad z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha/2}$$

FIGURA 8.7 *Región de rechazo*



Las regiones de rechazo para las distintas hipótesis alternativas y para el nivel de significación α se resumen en la Tabla 8.5.

TABLA 8.5 *Estadístico de prueba:* $(\bar{x}_1 - \bar{x}_2)/\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$

Hipótesis nula	*Hipótesis alternativa*	*Región de rechazo*
H_0: $\mu_1 = \mu_2$	H_A: $\mu_1 < \mu_2$	$]-\infty, z_\alpha[$
	H_A: $\mu_1 > \mu_2$	$]z_{1-\alpha}, +\infty[$
	H_A: $\mu_1 \neq \mu_2$	$]-\infty, -z_{1-\alpha/2}[\cup]z_{1-\alpha/2} +\infty[$

EJEMPLO. *Empresas familiares y no familiares*

Con la finalidad de analizar si la propaganda por radio influye de manera diferente en las ventas de las empresas familiares y no familiares del mismo rubro se tomó al azar un grupo de nueve empresas familiares y un grupo de nueve empresas no familiares. Después de realizada la propaganda por un tiempo determinado, el promedio de las ventas del primer grupo resultó ser 40,000 dólares, mientras que para el segundo grupo el promedio fue 35,000 dólares.

Suponiendo que se tuvo cuidado de que no influyeran otros factores diferentes a la propaganda por radio, y que antes de que esta fuera realizada los promedios de las ventas para ambos tipos de empresa eran iguales. Contrastar la hipótesis nula de que las medias de las ventas de ambos grupos de empresas son iguales frente a la hipótesis alternativa de que las medias son diferentes, al nivel de significación 0.05. Además se supone también que las poblaciones son normales con varianzas iguales a 25.

Solución

En realidad se trata de evaluar si la propaganda por radio influye o no de manera diferente en las ventas de las empresas familiares y no familiares.

Se ha supuesto que las varianzas de las ventas son iguales a 25, y de acuerdo a lo pedido, deberá contrastarse la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 \neq \mu_2$, en donde μ_1 es la media de las ventas de las empresas familiares y μ_2 es la media de las ventas de las empresas no familiares.

El estadístico de prueba es $z = \dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} = \dfrac{40{,}000 - 35{,}000}{\sqrt{\dfrac{5{,}000^2}{9} + \dfrac{5{,}000^2}{9}}} = 2.1213.$

Para el nivel de significación $\alpha = 0.05$ el cuantil de orden $1 - 0.05/2$ es igual a $z_{1\,-\,0.05/2} = 1.96$.

Como el valor del estadístico de prueba es mayor que $z_{1 - 0.05/2}$, la decisión será rechazar la hipótesis nula. Las medias de las ventas son diferentes. La propaganda por radio influye de manera diferente en las ventas de las empresas familiares y no familiares.

Caso 2: Las desviaciones estándar no se conocen pero son iguales

Si las varianzas no se conocen pero son iguales, el estadístico de prueba es:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$$

en donde s^2 es el estimador de la varianza común definido con:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 < \mu_2$, la regla de decisión es:

  Rechazar la hipótesis nula si $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha, (n_1 + n_2 - 2)}$

- Para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 > \mu_2$, la regla de decisión es:

  Rechazar la hipótesis nula si: $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1 - \alpha, (n_1 + n_2 - 2)}$

- Para contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_A : \mu_1 \neq \mu_2$, la regla de decisión es:

  Rechazar la hipótesis nula si:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < -t_{1 - \alpha/2, (n_1 + n_2 - 2)} \quad \text{o} \quad t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{1 - \alpha/2, (n_1 + n_2 - 2)}$$

EJEMPLO. *Comparando dos procesos para elaborar un producto*

Una empresa está dedicada a la fabricación de telas de algodón. Lo primordial de las telas que se fabrican es su resistencia a la tracción. Aun cuando el procedimiento que en la actualidad utiliza la empresa produce resultados satisfactorios, existe la posibilidad de agregar un cierto porcentaje de nailon para sustituir parte del algodón en las telas. Este agregado daría como resultado un producto considerablemente más barato. Sin embargo, se sospecha que la resistencia disminuye. La empresa tomará la decisión de realizar el agregado siempre que la resistencia no disminuya.

Con la finalidad de ayudar en la decisión, y con la ayuda de los técnicos encargados de la fabricación, se tomaron 20 porciones de telas elaboradas en las mismas condiciones (se utilizó el mismo turno y los mismos operarios), pero 10 de ellas se elaboraron con un cierto porcentaje de nailon y 10 solo con algodón. Las resistencias obtenidas (en unidades de la escala del aparato de medida) son las que aparecen a continuación.

A_1	A_2
12.33	12.38
12.09	12.24
13.93	11.33
13.11	11.12
12.32	13.10
12.39	12.40
12.56	11.59
13.42	11.18
13.95	12.03
13.00	10.92

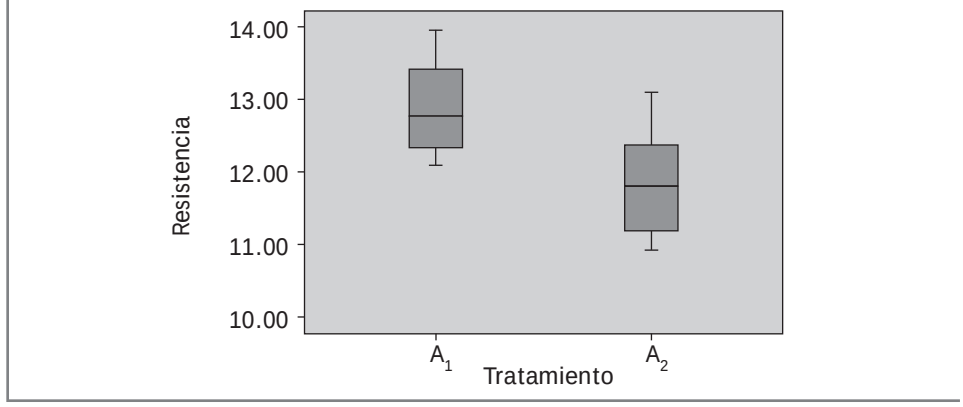
A_1: con mezcla

A_2: con mezcla

Las medias de estos grupos de datos son: $\overline{x}_1 = 12.9160$ y $\overline{x}_2 = 11.8290$, respectivamente.

Para iniciar el análisis se hicieron gráficos de cajas para cada muestra. Al parecer no hay mucha diferencia entre las dispersiones, pero sí en las medias; sin embargo, de la gráfica no se puede concluir en una primera instancia que la media para el primer tratamiento es mayor que la media para el segundo tratamiento. Se debe hacer la prueba de hipótesis.

FIGURA 8.8 *Comparación de los tratamientos*



Para realizar la prueba a nivel de la población, se debe contrastar la hipótesis nula $H_0 : \mu_1 = \mu_2$ versus la hipótesis alternativa $H_A : \mu_1 > \mu_2$, donde μ_1 y μ_2 son las medias de los procedimientos A_1 y A_2, respectivamente.

Para la prueba se usará el nivel de significación $\alpha = 0.05$, y se supondrá que las mediciones de las resistencias de las telas al aplicar ambos procedimientos tienen distribución normal.

Respecto a las varianzas se supondrá que son iguales.

El estadístico de prueba para comparar las medias de las poblaciones es:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

en donde $\bar{x}_1 = 12.9160$ y $\bar{x}_2 = 11.8290$.

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10 - 1)0.4784 + (10 - 1)(0.5012)}{10 + 10 - 2}} = 0.6998$$

Haciendo los cálculos se tiene que el valor del estadístico es 3.4730.

Los valores de este estadístico corresponden a una variable aleatoria con distribución *t* con 18 grados de libertad. La región de rechazo, al nivel de significación 0.05, es $]1.734, +\infty[$.

Como el valor del estadístico de prueba cae en la región de rechazo, la decisión es aceptar que con la mezcla se incrementa la resistencia de las telas. El riesgo de equivocarse al tomar esta decisión es a lo más el 5%.

Otra manera de analizar el asunto, compatible con el concepto del *valor p*, es el siguiente.

El valor del estadístico es 3.4730. De acuerdo a la tabla de la distribución t con 18 grados de libertad, la probabilidad de encontrar un valor mayor o igual a 3.4730, cuando la hipótesis nula es verdadera, es igual a 0.0013. Es decir, si las medias poblacionales fueran iguales sería poco probable hallar valores del estadístico de prueba como el encontrado. Es razonable suponer que las medias poblacionales son diferentes; es decir, que la diferencia que se observa en la medias muestrales es estadísticamente significativa.

Pruebas de hipótesis relativas a proporciones

Se trata de pruebas de hipótesis:

- relativas a la proporción *p* de elementos de una población que tienen cierto atributo.
- relativas a la comparación de dos proporciones de elementos que tienen cierto atributo en dos poblaciones *independientes.*

Hipótesis como "la proporción de analfabetos en el país es menor que el 6%" o "las proporciones de los artículos devueltos que son vendidos en dos tiendas diferentes son iguales" son revisadas en esta sección.

Prueba de hipótesis para una proporción

Para probar una hipótesis relativa a una proporción *p* de elementos de una población grande que tienen un cierto atributo, seguir el siguiente proceso:

1. Plantear las hipótesis: nula $H_0 : p = p_0$ y la hipótesis alternativa.
2. Indicar el nivel de significación α.
3. Tomar una muestra de tamaño *n y* calcular la proporción muestral $\hat{p}$.

   El estadístico de prueba para comparar lo hallado en la muestra con lo indicado en la hipótesis nula es:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

   (Este estadístico corresponde a la distribución normal estándar cuando la muestra es grande (en la práctica cuando $n > 30$).)

4. Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : p = p_0$ frente a la hipótesis alternativa $H_A : p < p_0$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} < z_\alpha$

  Si $\alpha = 0.05$, $z_\alpha = -1.645$ y la región de rechazo es $]-\infty, -1.645[$.
  Si $\alpha = 0.01$, $z_\alpha = -2.33$ y la región de rechazo es $]-\infty, -2.33[$.

- Para contrastar la hipótesis nula $H_0 : p = p_0$ frente a la hipótesis alternativa $H_A : p > p_0$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} > z_{1-\alpha}$

  Si $\alpha = 0.05$, $z_{1-\alpha} = 1.645$ y la región de rechazo es $]1.645, +\infty[$.
  Si $\alpha = 0.01$, $z_{1-\alpha} = 2.33$ y la región de rechazo es $]2.33, +\infty[$.

- Para contrastar la hipótesis nula $H_0 : p = p_0$ frente a la hipótesis alternativa $H_A : p \neq p_0$, la regla de decisión es:

  Rechazar la hipótesis nula si:

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} < -z_{1-\alpha/2} \quad \text{o} \quad z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}} > z_{1-\alpha/2}$$

  Si $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96$ y la región de rechazo es $]-\infty, -1.96[\cup]1.96, +\infty[$.
  Si $\alpha = 0.01$, $z_{1-\alpha/2} = 2.58$ y la región de rechazo es $]-\infty, -2.58[\cup]2.58, +\infty[$.

En la Tabla 8.6 se indican las regiones de rechazo.

TABLA 8.6 *Estadístico de prueba:* $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$

Hipótesis nula	*Hipótesis alternativa*	*Región de rechazo*
H_0: ´$p = p_0$	H_A: $p < p_0$	$]-\infty, z_\alpha[$
	H_A: $p > p_0$	$]z_{1-\alpha}, +\infty[$
	H_A: $p \neq p_0$	$]-\infty, -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}, +\infty[$

EJEMPLO. *Fumar es dañino para la salud*

Se sabe que en una ciudad el porcentaje de personas que fuman más de 20 cigarrillos diarios es $p = 0.16$. En 100 personas de la ciudad con afecciones pulmonares se observó que una proporción igual a 18% de ellos fumaba más de 20 cigarrillos por día. Se pregunta si en general los afectados fuman más cigarrillos en comparación con toda la población. Usar $\alpha = 0.01$.

Solución

Sea p la proporción de personas con afecciones pulmonares que fuman más de 20 cigarrillos.

Para responder a la pregunta deberá contrastarse la hipótesis nula:

H_0: "*Las 100 personas afectadas forman parte de la población general*", ($p = 0.16$), frente a la hipótesis alternativa H_A: $p > 0.16$.

Al nivel de significación 0.01, la regla de decisión consiste en rechazar la hipótesis nula si el valor del estadístico de prueba

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

cae en la región de rechazo $]2.33, +\infty[$.

El valor del estadístico de prueba es:

$$\frac{0.18 - 0.16}{\sqrt{0.16(0.84)/100}} = 0.5455$$

Como este valor no está en la región de rechazo, la hipótesis nula no se rechaza. No existe la suficiente información que permita indicar que el porcentaje de los que fuman en el grupo de las personas con afecciones pulmonares es mayor que el porcentaje de fumadores en toda la población.

EJEMPLO. *Circuitos. Control de calidad*

Un fabricante de circuitos para computadora asegura que el porcentaje de circuitos defectuosos que se producen durante el proceso es a lo más 8%.

Un cliente compra un lote muy grande de tales artículos, y para no revisar todo el lote conviene con el fabricante que aceptará el lote si al tomar una muestra al azar de 30 de tales circuitos encuentra que el número de defectuosos *X es menor o igual que* 1.

La situación puede plantearse como un procedimiento de prueba de hipótesis escribiendo como hipótesis nula lo que afirma el fabricante: $H_0 : p \leq 0.08$.

La hipótesis alternativa, atribuida al comprador y a su "carácter pesimista", puede escribirse como $H_A : p > 0.08$.

De acuerdo a lo pactado, existe el riesgo de que se rechace la hipótesis del fabricante siendo esta verdadera. Este riesgo está determinado por:

$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ es verdadera}] = P[X \geq 2 \mid p_0 = 0.08] = 1 - P[X \leq 1 \mid p_0 = 0.08]$

Asumiendo que la hipótesis nula es verdadera, y como la muestra es grande, se puede considerar que el número de circuitos defectuosos X en la muestra tiene distribución normal con media $np_0 = (30)(0.08)$ y varianza $np_0(1 - p_0) = (30)(0.08)(0.92)$. Usando la corrección por continuidad se tiene que:

$z = \dfrac{X - 0.5 - (30)(0.08)}{\sqrt{(30)(0.08)(0.92)}}$ tiene distribución normal, aproximadamente, y así:

$$\alpha = 1 - P[X \leq 1 \mid p_0 = 0.08] = 1 - P\left[Z \leq \frac{1 - 0.5 - (30)(0.08)}{\sqrt{(30)(0.08)(0.92)}}\right] = 1 - P[Z \leq -1.28] = 0.8997$$

Se concluye que siguiendo el procedimiento de control, y si el porcentaje de circuitos defectuosos es 8%, se tendrá que de cada 100 lotes que se reciban se rechazará aproximadamente el 90%. Al parecer el criterio que se sigue es muy estricto.

Prueba de hipótesis para dos proporciones en dos poblaciones independientes

Para probar una hipótesis relativa a la igualdad de dos proporciones p_1 y p_2 de elementos que satisfacen cierto atributo en dos poblaciones independientes, se procede como sigue:

1. Plantear las hipótesis: nula $H_0 : p_1 = p_2$ y la hipótesis alternativa.
2. Indicar el nivel de significación α.
3. Tomar una muestra de tamaño n_1 en la población 1 y una muestra de tamaño n_2 en la población 2 *y* calcular las proporciones muestrales $\hat{p}_1$ y $\hat{p}_2$ de los elementos que satisfacen la propiedad en estudio, respectivamente.

   El estadístico de prueba para comparar lo hallado en las muestras, de acuerdo a lo que se indica en la hipótesis nula, es:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1 - \bar{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

en donde $\bar{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ es un estimador de p si $p_1 = p_2 = p$.

Si la población es grande, este estadístico corresponde aproximadamente a la distribución normal estándar (en la práctica cuando $n \geq 30$).

(En realidad, el estadístico de prueba es $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, en donde p es el valor común en la relación $p_1 = p_2 = p$. Como p no se conoce, este parámetro se estima con $\overline{p}$.)

4. Al nivel de significación α:

- Para contrastar la hipótesis nula $H_0 : p_1 = p_2$ frente a la hipótesis alternativa $H_A : p_1 < p_2$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} < z_{\alpha}$

- Para contrastar la hipótesis nula $H_0 : p_1 = p_2$ frente a la hipótesis alternativa $H_A : p_1 > p_2$, la regla de decisión es la siguiente:

  Rechazar la hipótesis nula si $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} > z_{1-\alpha}$

- Para contrastar la hipótesis nula $H_0 : p_1 = p_2$ frente a la hipótesis alternativa $H_A : p_1 \neq p_2$, la regla de decisión es:

  Rechazar la hipótesis nula si:

  $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} < -z_{1-\alpha/2}$ o $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} > z_{1-\alpha/2}$

En la Tabla 8.7 se indican las regiones de rechazo.

TABLA 8.7 *Estadístico de prueba:* $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

Hipótesis nula	*Hipótesis alternativa*	*Región de rechazo*
$H_0: p_1 = p_2$	$H_A: p_1 < p_2$	$]-\infty, z_{\alpha}[$
	$H_A: p_1 > p_2$	$]z_{1-\alpha}, +\infty[$
	$H_A: p_1 \neq p_2$	$]-\infty, -z_{1-\alpha/2}[\cup]z_{1-\alpha/2}, +\infty[$

EJEMPLO. *Transporte por avión*

Se ha llevado a cabo un estudio para averiguar si existe alguna diferencia entre los porcentajes de personas mayores de 18 años de los países A y B que alguna vez han usado el avión como transporte. En una muestra aleatoria de 400 personas mayores de 18 años de A, 10 manifestaron que habían usado el avión alguna vez para transportarse, mientras que en una muestra de 600 personas mayores de 18 años de B, 20 habían usado el avión con el mismo fin. Se desea saber, al nivel de significación 0.05, si estos datos constituyen una evidencia para indicar que en A el porcentaje de las personas mayores de 18 años que han usado el avión para transportarse es menor que en B.

Solución

Si se indica con p_1 y p_2 las proporciones de personas mayores de 18 años que han usado alguna vez el avión para transportarse en A y B, respectivamente, se deberá contrastar la hipótesis nula $H_0 : p_1 = p_2$ versus la hipótesis alternativa $H_A : p_1 < p_2$.

La regla de decisión es rechazar la hipótesis nula a favor de la alternativa si el valor del estadístico de prueba

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

es menor que $z_\alpha = -1.645$. Es decir, si cae en la región $]-\infty, -1.645[$.

Usando los datos se tiene que $\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{(400)(10/400) + (600)(20/400)}{400 + 600} = 0.03$.

El valor del estadístico de prueba es:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(10/400) - (20/600)}{\sqrt{(0.03)(1-0.03)\left(\frac{1}{400} + \frac{1}{600}\right)}} = -0.7569$$

No es posible rechazar la hipótesis nula, pues el valor del estadístico de prueba no cae en la región de rechazo. Los resultados que se tienen de las muestras no evidencian que existe diferencia entre los porcentajes de personas mayores que 18 años que alguna vez han usado el avión para trasladarse.

APLICACIÓN: El caso Wonti

La empresa Wonti, dedicada a las ventas por menor, afronta una serie de reclamos de parte de los clientes que compran artículos alimenticios embolsados, los que provienen de la comercializadora Empak. Las bolsas que contienen los artículos, se supone, tienen un kilo de contenido; sin embargo, los clientes afirman que esto no es así. Wonti planea hacer un reclamo formal a Empak. En vista de que es posible que por azar algunos paquetes no cumplan con el peso especificado, antes de realizar el reclamo Wonti decidió emplear un procedimiento estadístico para tratar de analizar el peso del contenido que se deposita. El procedimiento consistió en elegir al azar 25 bolsas de alimentos envasados por Empak, y usando un peso patrón de un kilo se pesaron las 25 bolsas, registrándose un peso promedio de 1.050 kilos y una desviación estándar de 0.015 kilos. Al parecer, los resultados de la muestra indican que debe hacerse el reclamo; sin embargo, estos resultados podrían suceder por azar, y Wonti no quisiera equivocarse al realizar el reclamo. Por ello Wonti decidió contrastar la hipótesis $\mu = 1$ versus la hipótesis $\mu > 1$ con un nivel de significación 0.05.

Usted indicará si Wonti debe hacer la invocación.

APLICACIÓN: El caso de la ayuda social

La política de ayuda social a las comunidades con menos recursos económicos que realiza la sociedad Sosay contempla que dará apoyo a las comunidades que tengan un ingreso promedio menor que 300 dólares mensuales. Como no se cuenta con información registrada de las comunidades, se deberá recurrir al muestreo probabilístico, eligiéndose una muestra de tamaño *n*.

El lector debe determinar el tamaño de muestra a tomar considerando que Sosay desea que con probabilidad 0.05 se entregue la ayuda a comunidades que tengan ingresos cuyo promedio sea mayor o igual a 300 dólares mensuales, pero que con probabilidad 0.01 no se otorgue la ayuda a comunidades que tienen ingresos cuyo promedio sea igual a 280 dólares mensuales.

LA ESTADÍSTICA Y LA GESTIÓN DEL COMERCIO AL POR MENOR

Los estudiosos de las empresas dedicadas al comercio al por menor (al *retail*) indican que un alto porcentaje de los compradores en las tiendas al por menor deciden sus compras dentro del local de venta, y de igual manera respecto del tamaño de su compra. Esto obliga a la empresa a adoptar acciones que generalmente se sustentan en tres conceptos especiales: el conocimiento del comprador, las relaciones entre la empresa y el proveedor y la medición del desempeño.

La estadística permite situar al comprador, identificando, a través de los datos capturados, los patrones de compra y las ocasiones de consumo. De este modo se ajustan las ofertas y presentaciones en sala, se mejoran las ventas y así se construye o se mantiene la fidelidad de los compradores.

La estadística, como herramienta por excelencia de la decisión, permite, a través de sus modelos, el conocimiento de la demanda de los productos, de las series de productos que se compran, de la percepción de la calidad de los productos de parte de los clientes y la distribución de las preferencias de los compradores por las diferentes marcas.

Finalmente, la estadística también ayuda en el análisis de las mediciones del desempeño de las diferentes acciones que se llevan a cabo, permitiendo a la empresa la corrección de los errores.

8.4 La prueba de bondad de ajuste

Las pruebas de hipótesis estudiadas en secciones anteriores se refieren a parámetros de una distribución. En esta sección se estudia el procedimiento para probar conjeturas respecto a la manera como se distribuyen los elementos de una población en *k* categorías previamente indicadas.

Un caso ayudará a comprender cómo se aplica esta prueba.

La empresa FIDA es una de las tres empresas que hace varios años abastece de fideos al mercado nacional. Las otras dos empresas son FIDB y FIDC. Por mucho tiempo las tres empresas han copado el mercado en partes iguales; sin embargo, debido a las mejoras que ha estado introduciendo en la calidad de los fideos que fabrica, la empresa FIDC ha aumentado su participación a tal punto que la mayoría de empresas de investigación de mercados ha indicado que la participación actual es 20% para FIDA, 30% para FIDB y 50% para FIDC.

La empresa FIDA ha iniciado una investigación para comprobar si la nueva distribución de los clientes es como se indica; por ello precisa contrastar la hipótesis nula:

$$H_0 : p_A = 0.20,\ p_B = 0.30 \ \ y \ \ p_C = 0.50$$

versus la hipótesis alternativa $H_A : p_0$ *no es verdad* y donde:

p_A : es la participación en el mercado de los fideos de la empresa FIDA.

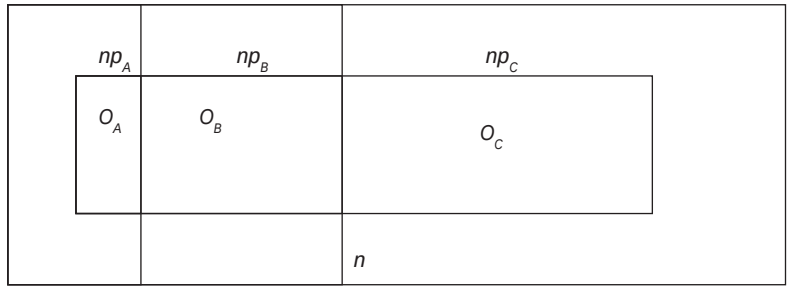
p_B : es la participación en el mercado de los fideos de la empresa FIDB.

p_C : es la participación en el mercado de los fideos de la empresa FIDC.

La hipótesis nula se refiere a la distribución de las tres categorías A, B y C (la suma de las probabilidades es igual a 1), y la manera como se lleva a cabo la prueba de hipótesis es análoga a la que antes se ha realizado:

1. Tomar una muestra de tamaño *n* en la población de todos los clientes de fideos.
2. "Comparar" las frecuencias observadas, O_A, O_B y O_C, de los que compran los fideos de FIDA, de FIDB y de FIDC, respectivamente, con las frecuencias np_A, np_B *y* np_C, que se esperan obtener en el caso de que la hipótesis nula sea verdadera.

FIGURA 8.9 *Muestra de tamaño n*



3. Para comparar las frecuencias observadas con las frecuencias esperadas, en el caso de que la hipótesis nula sea verdadera, se usa el estadístico:

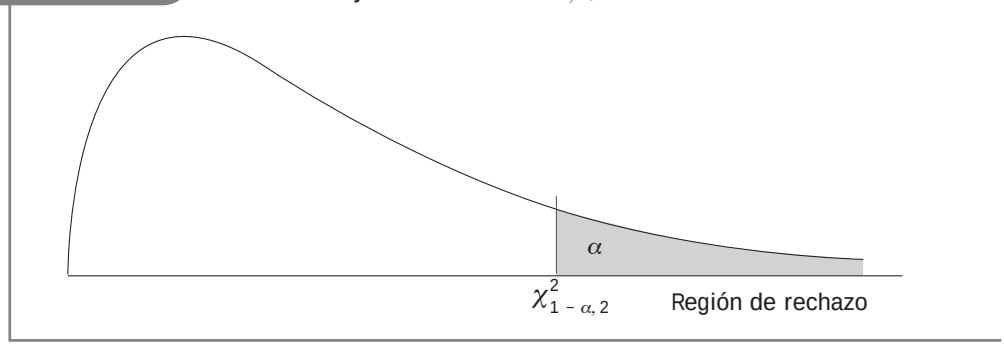
$$U^2 = \frac{(O_A - np_A)^2}{np_A} + \frac{(O_B - np_B)^2}{np_B} + \frac{(O_C - np_C)^2}{np_C}$$

   K. Pearson (1857-1936) demostró que este estadístico tiene aproximadamente la distribución ji-cuadrado, cuyo número de grados de libertad es igual al número de categorías menos 1. La aproximación es cada vez mejor si las frecuencias esperadas para cada clase son mayores o iguales que 5.

4. Tomando en cuenta la distribución del estadístico de prueba y al nivel de significación α, la decisión es la siguiente:

   Rechazar la hipótesis nula si el valor del estadístico U^2 es mayor que el cuantil de orden $1 - \alpha$ de la distribución ji-cuadrado con 2 grados de libertad.

FIGURA 8.10 *Distribución ji-cuadrado.* $\chi^2_{1-\alpha,2}$*, es el cuantil* $1\text{-}\alpha$



Esta prueba, que se refiere a la distribución misma y no a un parámetro de la distribución, se conoce como ***prueba ji-cuadrado*** o *prueba de bondad de ajuste.*

La empresa FIDC tomó una muestra de 80 clientes y encontró que 20 compran fideos de FIDA, 25 compran fideos de FIDB y el resto compran fideos de FIDC.

Las frecuencias esperadas de los clientes que compran los fideos de FIDA, FIDB y FIDC, si la hipótesis nula es verdadera, son: (80)(0.2) = 16, (80)(0.3) = 24, (80)(0.5) = 40, respectivamente.

Las frecuencias observadas y esperadas aparecen en la Tabla 8.8.

TABLA 8.8 *Tabla de contingencia*

Empresas	*Frecuencias observadas*	*Frecuencias esperadas si la hipótesis nula es verdadera*
FIDA	20	(80)(0.20) = 16
FIDB	25	(80)(0.30) = 24
FIDC	35	(80)(0.50) = 40
Suma	80	80

Usando el estadístico de prueba "se comparan" las frecuencias observadas con las frecuencias esperadas (obtenidas al usar la hipótesis nula).

El valor del estadístico es:

$$U^2 = \sum_{i=1}^{3} \frac{(O_i - np_i^0)^2}{np_i^0} = \frac{(20 - (80)(0.2))^2}{(80)(0.2)} + \frac{(25 - (80)(0.3))^2}{(80)(0.3)} + \frac{(35 - (80)(0.5))^2}{(80)(0.5)} = 1.67$$

El estadístico de prueba tiene distribución ji-cuadrado con 2 grados de libertad y al nivel de significación $\alpha = 0.05$ el cuantil de orden $1 - \alpha = 1 - 0.05 = 0.95$ es igual a 5.99. La región de rechazo es $]5.99, +\infty[$.

Como el valor calculado del estadístico no cae en la región, la hipótesis nula no se rechaza. La información encontrada en la muestra no permite a la empresa FIDA rechazar lo indicado por la empresa de mercados.

En general, para una población dividida en *k clases disjuntas*: $C_1,\ldots, C_k$, el contraste de la hipótesis nula:

H_0: *las probabilidades para las clases* $C_1,\ldots, C_k$, *son, respectivamente,*

$$p_1 = p_1^0,\ldots, p_k = p_k^0 \text{ ", con } p_1 + \ldots + p_k = 1$$

con la hipótesis alternativa:

H_A: *la hipótesis nula es falsa*

se realiza tomando una muestra de tamaño n y comparando las frecuencias, O_i, que resultan en cada categoría con las frecuencias esperadas, np_i^0, mediante el estadístico de prueba definido por:

$$U^2 = \sum_{i=1}^{k} \frac{(O_i - np_i^0)^2}{np_i^0}$$

Al nivel de significación α, la decisión es la siguiente:

Rechazar la hipótesis nula si el valor del estadístico U^2 es mayor que el cuantil de orden $1 - \alpha$ de la distribución ji-cuadrado cuyo número de grados de libertad es igual al número de categorías menos 1.

Observaciones

Si para calcular el valor del estadístico ji-cuadrado fuera necesario estimar *m* parámetros, su distribución es aproximadamente ji-cuadrado con $k - 1 - m$ grados de libertad.

Prueba de bondad de ajuste para una distribución normal

La prueba de ***bondad de ajuste*** se puede aplicar para confirmar si un grupo de datos proviene o no de una distribución normal. El procedimiento que a continuación se describe, mediante un ejemplo, consiste en partir a la población, de donde proviene el grupo de datos, en *k* categorías y comparar, mediante el estadístico ji-cuadrado, las frecuencias observadas en cada categoría con las frecuencias que se esperarían obtener si los datos provinieran de una distribución normal.

Las *k* categorías se obtienen al partir el conjunto de datos en cuestión en *k* subintervalos de clase y las proporciones p_i, a partir de las cuales se obtienen las frecuencias esperadas; estas se obtienen al suponer que los datos provienen de una distribución normal.

EJEMPLO. *Distribución de los precios de departamentos*

Usando los intervalos de clase, los precios en miles de dólares de 500 departamentos en una ciudad han sido distribuidos como sigue.

TABLA 8.9 *Distribución de precios*

Precios en miles de dólares	[43, 45]	]45, 47]	]47, 49]	]49, 51]	]51, 53]	]53, 55]	]55, 57]	]57, 59]
Número de departamentos	35	53	76	100	88	78	42	28

¿Se puede aceptar que los 500 precios recolectados provienen de una distribución normal? Usar el nivel de significación 0.01.

Solución

La hipótesis nula y alternativa pueden ser escritas, respectivamente, de la siguiente manera:

H_0 : los datos recogidos provienen de una población que tiene distribución normal.

H_A : la hipótesis nula es falsa.

La hipótesis nula se contrasta con la hipótesis alternativa usando el estadístico de prueba ji-cuadrado.

La información a usar se pueden organizar como en la Tabla 8.10.

TABLA 8.10 *Organización de los datos*

Intervalo: C_i	O_i	p_i	np_i	$(O_i - np_i)^2$	$\frac{(O_i - np_i)^2}{np_i}$
[43, 45]	35	0.0426	21.30	187.6900	8.8117
]45, 47]	53	0.0969	48.45	20.7025	0.4272
]47, 49]	76	0.1605	80.25	18.0625	0.2250
]49, 51]	100	0.2047	102.35	5.5225	0.0539
]51, 53]	88	0.1985	99.25	126.5625	1.2751
]53, 55]	78	0.1462	73.10	24.0100	0.3284
]55, 57]	42	0.0829	41.45	0.3025	0.0072
]57, 59]	28	0.0452	22.60	29.1600	1.2902
Total	n = 500				U^2 = 12.4187

La probabilidad p_i de que los datos estén en el intervalo C_i se calcula estimando previamente la media y la desviación estándar, $\overline{x}$ = 50.78, σ = 3.75, y suponiendo que la hipótesis nula es verdadera. Por ejemplo, para C_3, la proporción p_3 es igual a la probabilidad de que la variable X, que representa a los datos, esté en C_3 =]47, 49], suponiendo que esta tiene distribución normal:

$$p_3 = P[47 \le X \le 49] = P\left[\frac{47 - 50.78}{3.75} \le Z \le \frac{49 - 50.78}{3.75}\right] = 0.1605$$

Las frecuencias observadas se hallan calculando el producto del tamaño de la muestra, n, con las probabilidades p_i.

El valor del estadístico es:

$$U^2 = \sum_{i=1}^{8} \frac{(O_i - np_i)^2}{np_i} = 12.4187$$

que corresponde a una distribución ji-cuadrado con 8 – 1 – 2 = 5 grados de libertad (se han estimado dos parámetros).

Para el nivel de significación 0.01, la región de rechazo de la prueba es]15.1, +∞[.

Como el valor del estadístico de prueba no cae en la región de rechazo, se puede concluir que no existe suficiente evidencia que indique que los datos no provienen de una distribución normal.

8.5 Análisis de tablas de contingencia

Independencia de factores

Uno de los análisis interesantes que se realiza usando el estadístico ji-cuadrado es el relacionado con la prueba de ***independencia*** de dos variables categóricas, X e Y, asociadas a una población.

El procedimiento para llevar a cabo el análisis se indica a continuación, suponiendo previamente que las categorías respectivas de las variables X e Y son: {1,2,..., r} y {1,2,..., s}.

1. Plantear las hipótesis: H_0 : X e Y son independientes vs. H_A : X e Y no son independientes.

   La hipótesis nula significa que la probabilidad p_{ij} de que X tome el valor i y que la variable Y tome el valor j de manera conjunta es igual a $P[X = i]P[X = j]$, $\forall\ i, j$.

2. Tomar una muestra de n elementos y escribir las frecuencias que se observan en cada cruce de las categorías de las variables. La tabla de contingencia (8.11) aparece a continuación:

TABLA 8.11 *Tabla de contingencia*

$X \setminus Y$	1	2	...	s	*Suma*
1	O_{11}	O_{12}	...	O_{1s}	$O_{1.}$
2	O_{21}	O_{22}	...	O_{2s}	$O_{2.}$
...	...	...		...	
r	O_{r1}	O_{r2}	...		$O_{r.}$
Suma	$O._1$	$O._2$		$O._s$	n

3. Como en el caso de la prueba de bondad de ajuste, comparar las frecuencias anotadas en la tabla de contingencia con las frecuencias que se esperan obtener si la hipótesis nula es verdadera. La comparación se hace usando el estadístico ji-cuadrado.

$$U^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - np_i p_j)^2}{np_i p_j}$$

   En donde:

   O_{ij} es la frecuencia observada en la categoría *i* de *X* y la categoría *j* de *Y.*

   $np_i p_j = nP[X = i]P[Y = j]$ es la frecuencia esperada en el cruce de las categorías *i* y *j,* cuando la hipótesis nula es verdadera.

   El estadístico U^2 tiene una distribución ji-cuadrado cuyo número de grados de libertad es igual al producto (*número de filas* − 1)(*número de columnas* − 1), aproximadamente. Esta aproximación es mejor si $np_{ij} \geq 5$.

   (Observar que las *k* categorías referidas en la prueba de bondad de ajuste corresponden a las *r.s* celdas de la tabla de contingencia.)

4. Al nivel de significación α, la decisión es: la hipótesis nula se rechaza si el valor del estadístico de prueba es mayor o igual al cuantil de orden $1 - \alpha$ de la distribución ji-cuadrado cuyo número de grados de libertad es aproximadamente igual al producto (# *filas* − 1)(# *columnas* − 1).

   (Notar que, al parecer, el número de grados de libertad debería ser igual al número de celdas menos 1; sin embargo, por el número de parámetros que deben ser estimados el número de grados de libertad se reduce al producto (# *filas* − 1) (# *columnas* − 1).

EJEMPLO. *Evaluación de la política educativa*

Como parte de una investigación, se llevó a cabo una encuesta para evaluar si la aceptación de la política educativa del ministerio del ramo depende de los distintos niveles socioeconómicos. La encuesta se aplicó a 218 padres de familia, los que pertenecen a los estratos sociales A, B y C, de una pequeña comunidad. Planteando la pregunta: ¿está usted de acuerdo con la política educativa del ministerio? Y apuntando la clase social correspondiente, se obtuvieron las frecuencias que aparecen en la siguiente tabla (8.12) de contingencia.

TABLA 8.12 *Distribución conjunta*

	A	*B*	*C*	*Total*
Sí	22 (21.80)	24 (19.98)	20 (24.22)	66
No	50 (50.20)	42 (46.02)	60 (59.96)	152
Total	72	66	80	218

La idea es contrastar la hipótesis nula H_0: las respuestas no dependen de los estratos socioeconómicos frente a la hipótesis alternativa H_A: la hipótesis nula no es verdadera.

Las frecuencias esperadas se han escrito en la tabla entre paréntesis y la manera como se han aproximado se describe con el siguiente procedimiento.

Las probabilidades de los distintos niveles de cada variable o factor en toda la población no se conocen pero se pueden estimar con:

$$\hat{P}(A) = 72/218 \qquad \hat{P}(B) = 66/218 \qquad \hat{P}(C) = 80/218$$
$$\hat{P}(Si) = 66/218 \qquad \hat{P}(No) = 152/218$$

Si se admite la independencia de los factores (tipo de respuesta y estrato social), las probabilidades de los niveles cruzados se estiman como el producto de los estimadores de las probabilidades de cada uno de ellos:

$$\hat{P}(Sí \text{ y } A) = \frac{66}{218}\frac{72}{218} \qquad \hat{P}(No \text{ y } Sí) = \frac{152}{218}\frac{72}{218}$$
$$\hat{P}(Sí \text{ y } B) = \frac{66}{218}\frac{66}{218} \qquad \hat{P}(No \text{ y } B) = \frac{152}{218}\frac{66}{218}$$
$$\hat{P}(Sí \text{ y } C) = \frac{66}{218}\frac{80}{218} \qquad \hat{P}(No \text{ y } C) = \frac{152}{218}\frac{80}{218}$$

Multiplicando cada uno de estos valores con el total $n = 218$, se obtienen aproximadamente las frecuencias teóricas o esperadas E_{ij} (que aparecen en la Tabla 8.12 entre paréntesis). Así por ejemplo, la frecuencia esperada en el nivel cruzado *No* y *C* es:

$$218 \times P(No \text{ y } C) = 218(152/218)\ (80/218) = 59.96$$

Usando las frecuencias observadas y las frecuencias esperadas se obtiene el valor del estadístico de prueba U^2:

$$U^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(22 - 21.80)^2}{21.80} + \ldots + \frac{(60 - 59.96)^2}{59.96} = 1.897$$

Como se indicó, el valor encontrado expresa una medida de las desviaciones entre las frecuencias observadas y las frecuencias que se esperarían si la clase social no influyera en las respuestas. A simple vista el valor encontrado es “pequeño”, lo que indicaría que no existe influencia de las clases sociales en las respuestas. Pero, ¿qué tan pequeño es? Para responder, veamos la región de rechazo. Para esta prueba y al nivel de significación $\alpha = 0.01$, la región de rechazo que le corresponde, según la distribución ji-cuadrado con (3 – 1) (2 – 1) = 2 grados de libertad, es]9.210, +∞[.

Como el valor U^2 no cae en la región de rechazo, podemos indicar que lo observado en la muestra no permite decir que la clase social a la que pertenece un padre de familia influya en el tipo de respuesta.

Análisis de tablas con totales fijos en las filas o columnas

El modelo de las *tablas de contingencia r* × *s* corresponde al experimento de repartir al azar *n* bolas en *r* × *s* cajas. Algunas veces pueden resultar cajas vacías, lo cual, por razones prácticas, no es deseable. A partir de una muestra al azar de *n* artículos producidos por las tres fábricas F1, F2, F3, en donde no resultan artículos de F1, no se puede obtener información respecto de la relación que pueden tener las fábricas con la calidad de los artículos producidos. Sería deseable tomar de antemano una muestra de artículos de F1, otra de F2 y una tercera de F3; de este modo la prueba de independencia entre la clasificación de los artículos según la fábrica y según sean defectuosos o no se traduciría en probar que la proporción de defectuosos es la misma en las tres poblaciones formadas por las producciones de las tres fábricas. Esta prueba, que se llama *prueba de homogeneidad,* es una extensión de la prueba de hipótesis para comparar dos proporciones.

El siguiente es el procedimiento para llevar a cabo la prueba de homogeneidad para *s* poblaciones independientes.

1. Escribir la hipótesis nula $H_0 : p_1 = p_2 = \ldots = p_s$ (todas las proporciones de elementos con determinado atributo en cada una de las poblaciones son iguales), y la hipótesis alternativa H_A : H_0 no es verdad.
2. En cada población A_i, tomar una muestra de tamaño n_i y anotar las frecuencias observadas de los elementos con el atributo y sin el atributo, respectivamente, como se indica en la Tabla 8.13.

TABLA 8.13 *S poblaciones independientes*

	A_1	A_2	...	A_s	*Total*
Con el atributo	O_{11}	O_{12}	...	O_{1s}	$O_{1\cdot}$
Sin el atributo	O_{21}	O_{22}	...	O_{2s}	$O_{2\cdot}$
Muestras	n_1	n_2	...	n_s	n

3. El estadístico de prueba para comparar las frecuencias observadas con las frecuencias esperadas es el estadístico ji-cuadrado:

$$U^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde las frecuencias esperadas son iguales a $E_{ij} = \frac{O_{i\cdot}}{n} n_i$ $i = 1, 2$.

$O_{1\cdot}$ = Número de elementos con el atributo

$O_{2\cdot}$ = Número de elementos que no tienen el atributo

$n = n_1 + \ldots + n_s$

El estadístico de prueba tiene distribución ji-cuadrado con $s - 1$ grados de libertad.

4. Al nivel de significación α, la hipótesis nula se rechaza si el valor del estadístico de prueba es mayor que el cuantil $1 - \alpha$ de la distribución ji-cuadrado con $s - 1$ grados de libertad.

EJEMPLO. *Control de calidad de artículos que provienen de tres fábricas*

Con el fin de comparar la calidad de los artículos que se producen en tres fábricas: F1, F2 y F3, se tomaron al azar 200 artículos de F1, 300 de F2 y 400 de F3. El número de artículos defectuosos y no defectuosos que resultaron en cada una de las fábricas se muestra en la Tabla 8.14.

TABLA 8.14 *Tres fábricas*

	Fábrica F1	***Fábrica F2***	***Fábrica F3***	***Total***
Defectuosos	8	12	20	40
No defectuosos	192	288	380	860
Total	200	300	400	900

¿Se podría afirmar, al nivel de significación 0.05, que las proporciones de artículos defectuosos son iguales en las tres fábricas?

Solución

Se trata de contrastar la hipótesis nula:

$H_0 : p_1 = p_2 = p_3$, en donde p_1 = proporción de artículos defectuosos en F1, etcétera.

versus H_A: la hipótesis nula no es verdadera.

El estadístico de prueba es:

$$U^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

en donde:

O_{ij} es el número observado en la celda *i-j*.

E_{ij} es el número de elementos que se esperan en la celda *i-j* si la hipótesis nula es verdadera. Así por ejemplo, considerando que la hipótesis nula es verdadera, en la celda 1-1 se esperan:

$$E_{11} = \frac{O_{1\cdot}}{n}\, n_1$$

en donde:

$O_{1\cdot}$ es la suma de artículos defectuosos (8 + 12 + 20 = 40).

n es el total de las muestras tomadas en las tres fábricas (200 + 300 + 400).

$\frac{O_{1\cdot}}{n}$ = (0.10) es la estimación de la proporción común de artículos defectuosos si H_0 es verdadera.

n_1 es el número de datos que tiene la muestra tomada en la fábrica F1 (200 en la fábrica F1).

El estadístico de prueba, en este caso, tiene una distribución ji-cuadrado con (2 – 1) (3 – 1) = 2 grados de libertad si la hipótesis nula es verdadera.

Las frecuencias observadas y esperadas aparecen en la Tabla 8.15.

TABLA 8.15 *Frecuencias observadas y esperadas*

			Fábricas			*Total*
			F1	***F2***	***F3***	
Artículos	Defectuosos	Observados	8	12	20	40
		Esperados	8.9	13.3	17.8	
	No defectuosos	Observados	192	288	380	860
		Esperados	191.1	286.7	382.2	
	Total		200	300	400	900

El valor del estadístico de prueba es:

$$U^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(8 - 8.9)^2}{8.9} + \ldots + \frac{(380 - 382.2)^2}{382.2} = 0.523$$

Para α = 0.05, el cuantil de orden 1 – α = 0.95, de la distribución ji-cuadrado con 2 grados de libertad es igual a 5.99.

La decisión es la siguiente.

Al nivel de significación α = 0.05, la hipótesis nula no se rechaza, pues el valor del estadístico de prueba no es mayor que 5.99; no cae en la región de rechazo]5.99, +∞[. No existe suficiente evidencia para decir que las proporciones de artículos defectuosos en las tres fábricas son diferentes.

Por otro lado, el *valor p* de la prueba es igual a 0.770 (la probabilidad de encontrar un valor del estadístico de prueba mayor o igual a 0.523 es 0.770). Ello indica que, al igual que con el análisis anterior, no se puede rechazar la hipótesis nula. Si la hipótesis nula se rechazara, el error que se puede cometer al rechazar algo verdadero es 0.770.

APLICACIÓN: ¿Discriminación de género?

Juanita Pérez acaba de presentar un reclamo a la Dirección General de Trabajo. Juanita es una empleada altamente calificada de la empresa de servicios Servisur, y alega que en la última promoción realizada por esta empresa ella no ha sido ascendida, y piensa que esto se debe a que es mujer. Otras empleadas mujeres tampoco han sido ascendidas.

El hecho es que de 10 empleados candidatos al ascenso, siete fueron ascendidos y de esos siete, seis eran varones. El resto eran mujeres.

Para la promoción, realizada después de haberse aplicado una serie de pruebas y entrevistas a todos los empleados, se consideró como candidatos a los 10 empleados que mejor fueron calificados. Juanita era la cuarta mejor calificada, por lo que, según ella, debió ser ascendida.

La Dirección General de Trabajo, para tratar de resolver el reclamo de Juanita, recabó la información y realizó una prueba de independencia entre las variables X: género, con los niveles varón y mujer, e Y: ascenso con los niveles ascendió y no ascendió.

Los resultados mostraron que, en efecto, existe una relación de dependencia estadística entre las variables X e Y. En un inicio, la dirección le dio la razón a Juanita; sin embargo, los directivos de la empresa alegaron indicando que una relación estadística no implica una relación causal, por lo tanto ello no prueba que necesariamente existe discriminación de género. En vista de que el juzgamiento puede tener serias consecuencias, la Dirección General de Trabajo ha decidido seguir analizando los datos con la finalidad de investigar si existen o no otros factores confundidos en la situación o si incluyendo otros factores la asociación persiste como para indicar que los factores "ascenso" y "género" están significativamente relacionados. El alegato aún no termina y se planea ir a tribunales superiores.

EJERCICIOS

1. El tiempo promedio de duración de las llamadas que atiende una secretaria parece que ha aumentado con respecto al que se tenía hace dos años, el cual era igual a 3.8 minutos. Una muestra aleatoria de 100 llamadas revela un tiempo promedio igual a 4 minutos. Considerando que el tiempo de duración tiene distribución normal con desviación estándar igual a 0.5 minutos:
   a) Establecer la hipótesis nula y la hipótesis alternativa del problema con el fin de comprobar la sospecha.
   b) Al nivel de significación 0.05, contrastar las hipótesis establecidas en a).
2. Indicar cuándo se utiliza la distribución t en una prueba de hipótesis.
3. En una muestra aleatoria de 81 personas de una población normal con desviación estándar 4, se encontró que el promedio de las edades era 25 años.
   a) ¿Lo encontrado en la muestra permite afirmar que la media de las edades en la población es menor que 27? Para responder usar el nivel de significación 0.05.
   b) ¿De qué manera influye la desviación estándar en la prueba si esta aumenta o disminuye?
   c) ¿De qué manera influye el tamaño de la muestra en la prueba si esta aumenta o disminuye?
4. El promedio del tiempo de vida X de un determinado artículo que vende la empresa Sagaz es, según la misma compañía, mayor o igual que 2 años. Con la finalidad de comprobar esta afirmación se acuerda tomar una muestra de 64 de estos artículos y rechazar la afirmación de Sagaz versus la hipótesis $\mu < 2$ siempre que $\bar{x} < 1.9$. ¿Cuál es el riesgo que tiene la empresa de que rechacen su afirmación siendo esta verdadera? Se supone que el tiempo de vida para este caso tiene distribución normal de desviación estándar 0.3.
5. En el problema 3, hallar el valor p de la prueba.
6. Se afirma que, en el presente año, el 80% de los contribuyentes pagaron sus impuestos. Al parecer la información es exagerada, por lo que se tomó una muestra al azar de 200 contribuyentes, encontrándose que el número X de personas que pagaron sus impuestos fue 150. Para probar la hipótesis nula $p = 0.8$ versus la hipótesis alternativa $p < 0.8$, en donde p es la proporción de contribuyentes cumplidos, se usa la región de rechazo $\{X < 170\}$.
   a) Hallar la probabilidad de cometer el error de tipo I
   b) Encontrar la región de rechazo para $\alpha = 0.05$.
   c) Hallar la probabilidad de cometer el error de tipo II cuando $p = 0.6$
7. Después de escuchar las quejas de una serie de clientes acerca de los tiempos de duración de las pilas que vendía una compañía que afirmaba que duraban 18 horas en promedio, un comité de defensa del consumidor tomó una muestra aleatoria de 10 pilas y midió su duración. Los resultados fueron:

   18.4 19.0 17 17 18.6 17 19.2 17.0 18.5 16

   Suponiendo que la vida útil de las pilas sigue una distribución normal, contrastar la hipótesis del fabricante con la hipótesis de los clientes. Usar el nivel de significación 0.10. Halle el valor p de la prueba. Explicar.
8. Los directivos de un canal de televisión afirman que los avisos publicitarios que se propalan en el horario de la mañana duran en promedio 5 minutos. Establecer las hipótesis que plantearía un oyente para comprobar la afirmación de los directivos del canal. Para una muestra aleatoria de 36 avisos se encontró que el promedio de duración era 5.5 minutos con una

desviación estándar igual a 1.2 minutos. Considerando normalidad para el tiempo de los avisos, establecer la regla de decisión para probar la hipótesis del oyente, al nivel de significación 0.01.

9. Una oficina relacionada con la cobranza de impuestos comprobó que el 5% de las declaraciones juradas de impuestos eran inexactas. Después de un programa de educación aplicado a los contribuyentes se comprobó que de 1,124 declaraciones juradas, 45 eran inexactas. ¿Qué se puede concluir al nivel de significación $\alpha = 0.05$? ¿Existe suficiente evidencia para creer que el programa de educación ha rendido buenos resultados?

10. La oficina sectorial de agua potable está revisando la posibilidad de iniciar una campaña educativa en la ciudad para no hacer uso indiscriminado del agua potable. La campaña no será empezada si el promedio de agua potable consumida por familia es menor que 2,500 pies cúbicos por periodo. Ante la imposibilidad de conocer con exactitud si efectivamente el promedio es menor que 2,500, se toma una muestra aleatoria de n familias y se contrasta la hipótesis nula $H_0 : \mu = 2{,}500$ *vs.* $H_A : \mu < 2{,}500$. ¿Cuál debe ser el valor de n y cuál la decisión a tomar con la finalidad de que la probabilidad de cometer el error de tipo I sea igual a 0.05 y que la probabilidad de cometer el error de tipo II sea igual a 0.01 cuando el verdadero consumo de agua potable sea 2,300 pies cúbicos por periodo? Se supone que el consumo de agua potable en la ciudad tiene distribución normal con una desviación estándar igual a 50 pies cúbicos por periodo.

11. Un inspector de calidad decide aceptar una remesa de 20,000 lámparas si el porcentaje de las defectuosas en el lote es menor o igual que 10%. Si el inspector desea estar razonablemente seguro de que el lote satisface las especificaciones y pretende, observando una muestra de 1,000 lámparas escogidas aleatoriamente del lote, asumir un riesgo de 0.01 de aceptar una remesa que tiene 12% de lámparas defectuosas:
    a) ¿Cómo deben formularse las hipótesis?
    b) ¿Cuál es la regla de decisión para la prueba si el nivel de significación es 0.05?

12. Una fábrica de refrescos planea variar el sabor clásico de uno de sus refrescos por uno nuevo. Para realizar una prueba de sabor, 500 personas, tomadas al azar, prueban tres tipos de refrescos, dos de los cuales tenían el sabor clásico y uno tenía el nuevo sabor. A cada persona se le pide que indique el refresco con el nuevo sabor. Sea X el número de personas que responden correctamente; esto es, que seleccionan el refresco elaborado con el nuevo método. Si no existe diferencia entre los sabores, la probabilidad p de que el nuevo refresco sea seleccionado es 1/3; si existe diferencia se debe tener $p > 1/3$. ¿Cuál debe ser el menor valor A que debe tomar X para que con probabilidad $\alpha = 0.05$ se rechace la hipótesis nula H_0: $p = 1/3$ frente a la hipótesis H_A: $p > 1/3$?

13. Un agricultor viene sembrando maíz usando semilla de tipo A y obteniendo en promedio 100 kg por unidad de área con una desviación estándar de 6 kg. Se cree que una nueva semilla de tipo B dará mejor rendimiento. El agricultor desea usar la nueva semilla siempre que se aumente realmente la producción media. Para ello decide probar en 25 unidades de área que tiene las mismas condiciones que las que se usan con la semilla A. Suponiendo normalidad y usando un nivel de significación de 0.01:
    a) Indicar las hipótesis adecuadas.
    b) Indicar si el agricultor usará la semilla tipo B.

14. El límite legal de cierto desecho mineral que una factoría puede arrojar en las aguas de un río es de 6 mg en promedio por cada metro cúbico. Se toman 12 porciones de un metro cúbico cada una, obteniéndose $\sum_{i=1}^{12} x_i = 84$, en donde x_i es la cantidad de desechos por metro cúbico.

Si se supone que las cantidades x_i provienen de una distribución normal de media μ desconocida y desviación estándar $\sigma = 1.8$, probar la hipótesis $\mu = 6$ contra la hipótesis alternativa $\mu > 6$. Usar $\alpha = 0.05$.

15. Un químico ha desarrollado un material de plástico que, según él, tiene una resistencia media a la ruptura de 30 onzas por pulgada cuadrada. Para comprobar la bondad del método se tomaron 36 láminas del plástico en mención, hallándose que en cada una de estas la resistencia a la ruptura es, respectivamente:

    30.1 32.7 22.5 27.5 23.2 28.9 27.7 29.8 28.9 31.4 30.4 27.0 31.2 24.3 26.4 22.8 29.4 22.3 29.1 33.4 32.5 21.7 23.5 25.7 27.8 34.0 24.5 22.9 27.8 26.7 31.5 24.5 26.7 28.6 24.3 25.7

    Al nivel de significación $\alpha = 0.05$ y suponiendo normalidad:

    a) ¿Se admite la hipótesis del químico?

    b) Si se acepta la hipótesis del químico y la verdadera media es 28, ¿cuál es la probabilidad del error de tipo II que se comete?

16. La oficina de tránsito ha venido usando lámparas para semáforos de una marca cuya duración tiene una vida media de 1,000 horas con una desviación estándar de 90 horas. Existe la posibilidad de usar lámparas de otra marca que son más baratas que las anteriores a menos que su vida media sea inferior a 1,000 horas. Se prueban 100 lámparas de la nueva marca, obteniéndose un promedio de 989 horas de tiempo de vida. Si se supone que la desviación estándar de la nueva marca es igual a la de la marca anterior y que los tiempos de vida tienen distribución normal, decir si el departamento de tránsito debe comprar las lámparas de la otra marca. Usar $\alpha = 0.01$.

17. El gerente de ventas de una compañía dice que los vendedores tienen un promedio no mayor de $ 1,500 en las ventas diarias con una varianza igual a 900. Se seleccionan al azar 36 vendedores encontrándose una media muestral de $ 1,700 en las ventas.

    a) Usando el nivel de significación 0.05, probar la hipótesis nula $H_0 : \mu = 1500$ frente a la hipótesis alternativa $H_A : \mu > 1500$.

    b) ¿Se podría afirmar que las muestras de tamaño 36 pueden detectar una diferencia igual a $ 100 diarios en el promedio de ventas, por encima de lo que se indica en la hipótesis nula?

18. Para "comprobar" si un conductor que ha cometido una infracción de tránsito conduce embriagado, se realizan tres determinaciones de la concentración de alcohol en su sangre. Los resultados x_1, x_2 y x_3 se asume que provienen de una distribución $N(\mu, 0.06^2)$, donde μ es la verdadera concentración promedio. Si $\mu > 0.5$ la persona ha conducido embriagada. Se acuerda que si $\bar{x} > 0.5 + 1.645(0.06)/\sqrt{3}$, se acepta que el conductor ha conducido embriagado; de otro modo, no. Se trata de una prueba de la hipótesis nula H_0: $\mu = 0.5$ frente a la hipótesis alternativa $H_A : \mu > 0.5$ al nivel de significación $\alpha = 0.05$. Indicar cuál de las siguientes expresiones describe cabalmente la situación anterior

    a) A lo más el 5% de personas sometidas al test conducen embriagadas.

    b) A lo más el 5% de personas que no están embriagadas son consideradas como si lo estuvieran.

    c) A lo más el 5% de aquellos que son declarados embriagados por el test no están embriagados.

19. Los pobladores de un distrito han reclamado indicando que las balanzas de los vendedores de un mercado de abastos son fraudulentas. La municipalidad del distrito desea comprobar si las protestas son justificadas o no; para ello ha tomado una muestra de 16 puestos del mercado y pesado un kilo verdadero. La media muestral de los pesos obtenidos fue 1.02 kilos. Si se supone que los pesos tienen distribución normal con desviación estándar 0.03 kilos, ¿al nivel de significación del 5%, se justifican los reclamos?

20. Un fabricante de chips asegura que no más del 2% de los chips vendidos tienen algún defecto. Una compañía de artículos electrónicos compra una gran cantidad de estos chips. Para verificar lo que indica el fabricante, la compañía decidió probar una muestra de 49 de estos chips y encontró 3 chips defectuosos. ¿Debería rechazarse la indicación del fabricante? Usar el nivel de significación 0.05.

21. Se desea conocer si en promedio los pesos de las cargas de minerales que dos empresas mineras envían a un puerto diariamente son iguales. Para ello se tomaron dos muestras, una por cada empresa. Los resultados, en miles de toneladas, fueron como sigue:

    13.2 15.3 10.4 9.0 10.5 12.3 9.1 7.9 11.2 9.4, para la primera empresa
    12.3 13.5 14.1 10.1 12.1 7.2 9.3 8.6 9.5 10.2 9.5 9.7, para la segunda empresa

    Suponiendo normalidad de los pesos que se transportan, probar la hipótesis de igual de medias. Use el nivel de significación 0.01.

22. Probar la hipótesis $H_0 : p_1 = p_2$ versus la hipótesis $H_A : p_1 > p_2$, al nivel de significación 0.05, donde las proporciones p_1 y p_2 corresponden a dos poblaciones independientes. Los resultados para dos muestras independientes tomadas de las dos poblaciones fueron como sigue.

Muestra A	*Muestra B*
$n_1 = 300$	$n_2 = 450$
$p_1 = 0.45$	$p_2 = 0.40$

23. En un estudio sobre el uso de Internet entre hombres y mujeres se tomó una muestra de 250 mujeres y otra de 180 hombres y se encontró que 100 mujeres usaban Internet, mientras que en el caso de los hombres 120 hacían uso de este medio.
    a) Formular las hipótesis nula y alternativa si en el estudio la hipótesis de investigación afirmaba que era más probable que los hombres usaran Internet. Probar las hipótesis formuladas al nivel de significación 0.05.
    b) ¿Cuál es el valor *p* de la prueba y cuál es la conclusión?

24. En una muestra de 500 personas tomada en el presente año se observó que 200 de ellas consumían café instantáneo. Hace 4 años, en una muestra de 300 personas, se observó que 130 personas consumían el producto indicado.
    a) Establecer la hipótesis nula y la hipótesis alternativa para determinar si la proporción de las personas que consumen café instantáneo ha variado en este periodo de 4 años.
    b) Hallar el valor *p* de la prueba e indicar la conclusión de esta.

25. Para evaluar si un dado determinado es equilibrado, se arrojó este 120 veces, obteniéndose lo siguiente:

    el 1 apareció 12 veces, ($n_1 = 12$)
    el 2 apareció 34 veces, ($n_2 = 34$)
    el 3 apareció 25 veces, ($n_3 = 25$)
    el 4 apareció 15 veces, ($n_4 = 15$)
    el 5 apareció 18 veces, ($n_5 = 18$)
    el 6 apareció 16 veces, ($n_6 = 16$)

    Al nivel de significación $\alpha = 0.01$, ¿se podría indicar que el dado es equilibrado?

26. Un representante de ventas de una empresa de comercio afirma que las visitas que se producen en una de sus tiendas tienen igual frecuencia los cinco primeros días de la semana. Para comprobar esta aseveración, de los registros de la empresa se tomaron al azar 523 visitas realizadas hallándose los siguientes resultados.

Días de la semana	L	M	M	J	V
Número de visitas	109	92	120	87	115

Probar la información del representante. Usar α = 0.05.

27. Según una información oficial, el 20% de los habitantes de una región son analfabetos, el 63% tiene educación primaria, el 12% tiene educación secundaria y el 5%, educación superior. Para comprobar tal aseveración se toma una muestra de 100 individuos y se observa que 35 son analfabetos, 35 tienen educación primaria, 20 tienen educación secundaria y 10 tienen educación superior. ¿Confirman los datos la afirmación oficial? Use α = 0.05.

28. Cuatrocientos valores de las calificaciones, *X,* de un test dieron la siguiente distribución:

Calificaciones	75	80	85	90	95	100	105	110	115	120
Frecuencias	10	14	26	51	64	73	68	53	30	11

¿Se ajustan los datos a una ley normal? Use α = 0.05.

29. Un estudio ha revelado que en un día laborable cualquiera la distribución del total de clientes que se acercan a sacar dinero de los 5 cajeros que existen en un banco es 0.40, 0.15, 0.05, 0.10 y 0.30, respectivamente. Otro estudio realizado durante varios días no laborables reveló que en estos cajeros el número de transacciones de retiro fueron, respectivamente, 172, 58, 40, 30 y 100. ¿Muestran estos datos que la distribución del uso de los cajeros automáticos durante los días no laborables difiere a la de los días laborables. Use el nivel de significación 0.05.

30. En un estudio sociopolítico realizado en una ciudad se preguntó si se estaba o no de acuerdo con un proyecto relacionado con la construcción de vías de acceso a la ciudad. Tabulados los resultados se obtuvo la siguiente distribución tomando en cuenta las zonas en que está dividida la ciudad y las respuestas.

	De acuerdo	***Indeciso***	***En contra***
Región norte	79	69	89
Región sur	108	60	100
Región este	251	100	260
Región oeste	47	40	40

Probar la hipótesis de que la residencia geográfica de los que responden no está relacionada con las respuestas dadas. Use el nivel de significación 0.05.

31. Probar si la siguiente muestra de 50 datos proviene de una distribución normal de media 400 y desviación estándar 20, usando los intervalos de clase construidos con los percentiles 20, 40, 60 y 80 de la distribución.

399.9 421.2 372.3 394.4 384.4 393.7 377.2 422.9 427.1 375.6 399.3 408.0 400.5 408.1
402.5 393.6 416.5 401.3 401.2 406.2 376.1 352.1 415.9 421.5 408.2 380.6 391.0 446.9
386.5 390.3 421.2 419.4 377.7 435.2 409.6 386.7 397.8 422.8 420.7 357.5 434.2 391.1
409.2 398.4 371.9 417.4 390.5 381.9 360.2 415.3

32. De 40 alumnos elegidos al azar en la Universidad A, 8 siguen carreras técnicas; de 50 alumnos elegidos al azar en la Universidad B, 10 siguen carreras técnicas, y de 40 alumnos elegidos al azar en la Universidad C, 6 siguen carreras técnicas. ¿Se podría afirmar, al nivel de significación 0.05, que las proporciones de las personas que siguen carreras técnicas son las mismas en las tres universidades?

33. Para combatir la malaria se están experimentando con tres tipos de droga: A, B, y C. Se trataron 60 casos con A, 80 casos con B y 100 casos con C. La tasa de recaída para cada droga fue como sigue.

	A	*B*	*C*
Casos	60	80	100
Tasa de recaída	40%	30%	12%

A base de los datos, ¿se puede decir que existe evidencia de que las tasas de recaída de las tres drogas no son iguales? Use $\alpha = 0.01$.

34. Un inspector de calidad indica que las proporciones de tubos de plástico defectuosos que producen tres fábricas difieren entre sí. Para verificar si la indicación del inspector es correcta se toman muestras de 50, 60 y 80 tubos de cada fábrica y se encuentran 2, 10 y 17 tubos defectuosos, respectivamente. Probar la hipótesis del inspector al nivel de significación 0.05.

35. Con el fin de probar la efectividad de un aviso publicitario en la venta de un artículo, se aplicó una encuesta a 200 personas. Los resultados aparecen en la siguiente tabla.

	Compraron el artículo	*No compraron el artículo*	*Suma*
Leyeron el aviso	9	101	110
No leyeron el aviso	4	86	90

¿Presentan los datos evidencia suficiente como para indicar que el aviso no influyó en la compra del artículo? Use el nivel de significación de 0.05.

36. Una oficina de comunicaciones desea saber las preferencias de las mujeres por los avisos publicitarios en tres medios diferentes. Los resultados de una encuesta a una muestra de 130 mujeres fueron:

Mujeres		
A	Prefieren la televisión	25
B	Prefieren la radio	35
C	Prefieren el periódico	70

a) Se puede afirmar que las mujeres que están prefiriendo la televisión, la radio y el periódico son el 25%, 25% y 50%, respectivamente. Use el nivel de significación de 0.05.

b) Al nivel de confianza del 95%, ¿se puede afirmar que la proporción de las mujeres que prefieren la radio es mayor que la proporción de las mujeres que prefieren la televisión?

c) La misma encuesta se aplicó a una muestra de 100 hombres. Los resultados fueron como sigue.

Hombres		
A	Prefieren la televisión	70
B	Prefieren la radio	20
C	Prefieren el periódico	10

d) ¿Se puede afirmar que el porcentaje de mujeres que prefieren la radio es mayor que el porcentaje de hombres que prefieren la radio?

37. En un proceso de producción se fabrican artículos cuya calidad resultante puede ser de tres tipos: C1, C2 y C3. Cada artículo resultante puede necesitar aleatoriamente uno de los tres tipos de ajuste siguiente: A1, A2, A3. Los resultados obtenidos al tomar al azar 150 artículos fueron como sigue.

Ajuste	*Calidad*		
	C1	*C2*	*C3*
A1	26	10	14
A2	30	12	16
A3	18	8	16

a) Usar el nivel de significación 0.05 para decir si los tipos de calidad resultante se dan con igual frecuencia.

b) ¿Se puede afirmar que el tipo de calidad y el ajuste son independientes? Use el nivel de significación 0.05.

c) Usando un intervalo de confianza del 95% indicar si se puede aceptar que la frecuencia con la cual se produce el tipo de calidad C1 y se necesita un ajuste del tipo A3 es la misma que la correspondiente al tipo de calidad C2 y ajuste de tipo A1.

38. Con el fin de probar la efectividad de una vacuna contra cierta enfermedad, se realizó un experimento observando a 200 personas, 110 de ellas vacunadas y las otras 90 sin vacunar.

Los resultados obtenidos se observan en la siguiente tabla.

	Contrajeron la enfermedad	*No contrajeron la enfermedad*	*Suma*
Vacunados	9	101	110
Sin vacunas	4	86	90

¿Presentan los datos evidencia suficiente como para indicar que la proporción de personas vacunadas que contrajeron la enfermedad no es la misma que la proporción de personas que no se vacunaron y que contrajeron la enfermedad? Use la prueba ji-cuadrado con $\alpha = 0.05$.

RESPUESTAS A LOS EJERCICIOS

1. a) $H_0 : \mu = 3.80$ vs. $H_A : \mu > 3.80$ b) se rechaza la hipótesis nula. **3.** a) $H_0 : \mu = 27$ vs. $H_A : \mu < 27$. Valor del estadístico de prueba: −4.5. Se rechaza la hipótesis nula. La respuesta es afirmativa. **4.** El riesgo es igual a la probabilidad $P(\bar{X} < 1.9 \mid \mu = 2) = P\left(Z < \frac{1.9 - 2}{0.3/\sqrt{64}}\right)$. **5.** El valor p es igual a $P(\bar{X} < 25 \mid \mu = 27) \approx 0$ **6.** a) Calcular la probabilidad $P(X < 170 \mid p = 0.8) = P(\hat{p} < 170/200) = P\left(Z < \frac{170/200 - 0.8}{\sqrt{\frac{0.8 \times 0.2}{200}}}\right)$ b) Hallar C de tal modo que $0.05 = P(X < C \mid p = 0.8)$. **7.** La media muestral es 17.87 y la desviación estándar muestral es 0.9452. El estadístico de prueba es −0.4349. La región de rechazo se determina con la distribución t con 9 grados de libertad. Está en la cola izquierda. **8.** a) $H_0 : \mu = 5$ vs. $H_A : \mu > 5$ b) el valor del estadístico de prueba es $\frac{\bar{x} - 5}{1.2/\sqrt{36}} = 2.5$. La región de rechazo se determina con la distribución t con 35 grados de libertad. Como el tamaño de muestra es mayor que 30, se puede usar la distribución normal para hallar la región de rechazo. **9.** $H_0 : p = 0.05$ vs. $H_A : p < 0.05$. El valor del estadístico de prueba es −1.7040. **10.** Plantear las condiciones: $0.05 = P(\bar{X} < C \mid \mu = 2500)$ y $0.01 = P(\bar{X} > C \mid \mu = 2300)$, estandarizar y resolver las ecuaciones $\frac{C - 2500}{50/\sqrt{n}} = -1.645$ y $\frac{C - 2300}{50/\sqrt{n}} = 2.38$. El procedimiento será tomar una muestra del tamaño n encontrado y rechazar la hipótesis nula si la media muestral es menor que el valor de C hallado. **11.** a) $H_0 : p = 0.10$ vs. $H_A : p > 0.10$ b) Hallar el tamaño de muestra n y la regla de decisión considerando que $\alpha = 0.05$ y que β es 0.01 cuando $p = 0.12$. **12.** Usando el error tipo I, resolver la ecuación $\frac{A/500 - 1/3}{\sqrt{2/4500}} = 1.645$, donde A es el menor valor que debe tomar X para que con probabilidad 0.05 se rechace la hipótesis nula. **13.** a) $H_0 : \mu = 100$ vs. $\mu > 100$ b) El agricultor usará la semilla B si $\frac{\bar{x} - 100}{6/5}$ es mayor o igual a 2.33. **14.** $H_0 : \mu = 6$ vs. $H_A : \mu > 6$. El valor del estadístico de prueba es 1.9245. **16.** $H_0 : \mu = 1000$ vs. $\mu < 1000$. El valor del estadístico de prueba es −1.22. Decisión: si se rechaza la hipótesis nula, no usar las lámparas de la nueva marca. **17.** b) Si la probabilidad de cometer el error tipo II cuando $\mu = 1600$ es alta, la prueba no detecta una diferencia igual a $ 100 diarios en el promedio de ventas, por encima de lo que se indica en la hipótesis nula. **18.** La respuesta es b). **19.** Comenzar planteando las hipótesis $H_0 : \mu = 1$ vs. $\mu > 1$, en donde μ es la media de los pesos. Si se rechaza la hipótesis nula, las balanzas están adulteradas. **21.** Las medias muestrales respectivas son: 10.83 y 10.50. Calcular el estimador de la varianza común. El estadístico de prueba sigue una distribución t con 10 grados de libertad. **22.** $H_0 : p_1 = p_2$ vs. $H_A : p_1 > p_2$. El valor del estadístico de prueba es 1.33. La hipótesis nula no se puede rechazar. **25.** Plantear la hipótesis nula $H_A : p_1 = \ldots = p_6 = 1/6$ y la alternativa $H_A : H_0$ no es verdad. Comparar las frecuencias observadas con las frecuencias esperadas, usando el estadístico ji-cuadrado con 5 grados de libertad. Usar la prueba de bondad de ajuste de la hipótesis $p_i = 1/6$ para $i = 1,\ldots, 6$. **26.** Hipótesis nula: $p_L = 1/5$, $p_M = 1/5$, $p_{Mi} = 1/5$, $p_J = 1/5$ *y* $p_V = 1/5$, $H_A : H_0$ no es verdad. **29.** Hipótesis nula: $p_1 = 0.4$, $p_2 = 0.15$, $p_3 = 0.05$, $p_4 = 0.10$ *y* $p_5 = 0.30$. Aplicar la prueba de bondad de ajuste. **30.** Aplicar la prueba de independencia. Hipótesis nula: las respuestas son independientes de las regiones de procedencia. Hipótesis alternativa: la hipótesis nula es falsa. **32.** Probar la hipótesis de igualdad de proporciones usando la prueba de homogeneidad. **33.** Probar la hipótesis de igualdad de proporciones usando la prueba de homogeneidad. **34.** Aplicar la prueba de homogeneidad como el ejercicio 8 y 9. **35.** Prueba de independencia. Hipótesis nula: la compra es independiente de la lectura del aviso. Hipótesis alternativa: la hipótesis nula es falsa. **36.** a) probar la hipótesis: $p_1 = 0.25$, $p_2 = 0.25$ y $p_3 = 0.50$ usando la prueba de bondad de ajuste. b) usar un intervalo de confianza para la diferencia de proporciones *Pradio – Ptelevisión* en una sola muestra. d) usar un intervalo de confianza para la diferencia de proporciones *Pradio, hombres – Ptelevisión, mujeres* en dos muestras independientes. **37.** a) probar la hipótesis $p_1 = 1/3$, $p_2 = 1/3$ y $p_3 = 1/3$ (prueba de bondad de ajuste); b) prueba de independencia. Hipótesis nula: el tipo de calidad es independiente del tipo de ajuste. c) usar un intervalo de confianza para la diferencia de proporciones en una sola muestra.

CAPÍTULO 9

El modelo de regresión lineal

Francis Galton

Francis Galton nació en 1822 en el seno de una familia acomodada cuáquera. Sus estudios iniciales los hizo en medicina, los que fueron interrumpidos por sus deseos de viajar. Después de hacerlo, Galton inició estudios de matemáticas en Cambridge, que volvió a interrumpir en el tercer año para luego retomar los estudios médicos, pero estos tampoco fueron terminados. Sin tener un grado académico, Galton incursionó en diferentes campos, incluyendo biología, psicología, estadística y genética.

S. M. Stigler, en su obra *Francis Galton's Account of the Invention of Correlation* (1989), describió a Francis Galton como una "figura romántica en la historia de la estadística, tal vez el último caballero científico". Es a través de su obra *Hereditary Genius* (1869) que se perciben las primeras ideas sobre regresión. Posteriormente, en 1885, Galton presentó en la sección antropológica de la British Association for the Advancement of Science la primera descripción del fenómeno de regresión y su relación con la distribución normal. Este trabajo fue realizado en el contexto de un ejemplo empírico en donde se relacionaba el promedio de las estaturas de los padres con el promedio de las estaturas de los hijos. Algunos años más tarde Galton formuló el concepto de correlación.

Las ideas de Galton fueron tomadas y extendidas por otros investigadores como Francis Edgeworth, quien relacionó el método de mínimos cuadrados con los conceptos de correlación y regresión.

Murió en Inglaterra, en 1911.

CONTENIDO

9.1 Introducción

El propósito de este capítulo es la presentación de un modelo estadístico básico, muy popular y que integra una serie de herramientas ya estudiadas. Este es el *modelo de regresión lineal.*

El modelo de regresión lineal se usa para expresar la relación lineal que pueda existir entre los valores de una variable y los valores de un conjunto de una o más variables. Por ejemplo, un modelo de este tipo puede ser utilizado para explicar la variabilidad de las ventas de una empresa en términos de la inversión que se realiza en publicidad. De este modo se puede escribir:

Ventas en dólares = β_0 + β_1* (*inversión en publicidad en dólares*) + *Error aleatorio*

Para este modelo, se dice que las ventas constituyen una variable dependiente y que la inversión en publicidad es una variable independiente.

El modelo que trata de explicar la variable dependiente (ventas) mediante una relación lineal y usando solo una variable independiente (inversión en publicidad) se llama *modelo de regresión lineal simple*.

Algunos modelos de regresión incorporan más de una variable independiente, y su forma puede ser de lo más complicada posible. Los modelos de regresión que incorporan más de una variable independiente se llaman *modelos de regresión múltiple.*

Los modelos de regresión fueron introducidos por Laplace y Gauss. Posteriormente fueron usados por Galton en trabajos que trataban de explicar la relación de las estaturas de los padres con las de sus hijos, encontrando lo que él llamó *regresión a la media*, expresión usada para indicar "que los hijos de los padres altos, en promedio, no eran tan altos como los padres, y que los hijos de los padres bajos, en promedio, eran más altos que los padres". Había una regresión hacia el promedio.

Los modelos de regresión se aplican en casi todos los campos, como el de la ingeniería, de las ciencias físicas, de las ciencias económicas, de las ciencias de la vida, de las ciencias sociales, etc., y en muchos casos se utilizan para:

- Predecir rendimientos futuros de un proceso.
- Analizar la influencia de ciertos factores en los valores de una variable y de esta manera conocer, controlar y mejorar un proceso.

9.2 El modelo de regresión lineal simple

Esta sección se inicia considerando modelos de regresión que relacionan valores *x* de una variable independiente *X* con valores *y* de una variable dependiente *Y*, cuya forma es:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

La expresión $\beta_0 + \beta_1 x$ es la parte estructural lineal, mientras que ε resume la parte aleatoria que influye débilmente en la variable dependiente *Y*. A la ecuación $y = \beta_0 + \beta_1 x$, que expresa la estructura lineal, se le llama *ecuación de la recta de regresión lineal*. A los coeficientes β_0 y β_1 se les llama *coeficientes de regresión*.

Se considera de este modo que los puntos (*x, y*) no necesariamente están sobre la recta $y = \beta_0 + \beta_1 x$, sino que fluctúan aleatoriamente a su alrededor.

Se asume que ε es un valor aleatorio cuya distribución es normal de media 0 y varianza constante.

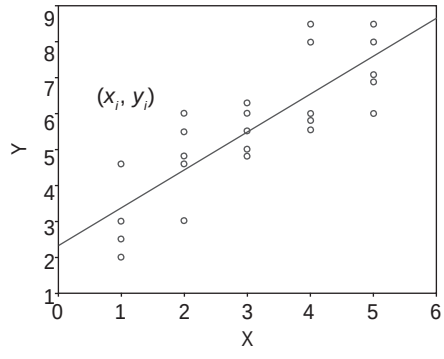
Al modelo establecido se le llama ***modelo de regresión lineal simple***.

Algunas veces se usarán las variables en lugar de los valores de las variables para describir el modelo. Así tendremos $Y = \beta_0 + \beta_1 X + \varepsilon$.

El coeficiente β_0 es el intercepto de la recta de regresión con el eje de las ordenadas, mientras que el coeficiente β_1 es la pendiente de esta recta; este representa *el incremento de la variable dependiente cuando la variable independiente se incrementa en una unidad*.

Si se grafican los valores *x* de *X* y los valores *y* correspondientes de *Y* se tendrá un *diagrama de dispersión,* como el que aparece a continuación (Figura 9.1). La gráfica de la recta proporciona una idea de la recta de regresión lineal. Los puntos observados están dispersos alrededor de esta recta.

FIGURA 9.1 *Diagrama de dispersión*



Aun cuando se use una única variable independiente, es posible que la relación entre los valores de la variable independiente y los valores de la variable dependiente tenga otros términos diferentes a los que determinan una relación lineal. Relaciones como las siguientes pueden tener lugar:

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$

$y = \beta_0 e^{\beta_1 x} \varepsilon$

$y = \beta_0 + \beta_1 x + \beta_2 \ln x + \varepsilon,$

etcétera.

En los procesos relacionados con los negocios es posible que exista más de una variable independiente, y en tales casos los modelos de regresión pueden ser como los siguientes:

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_3 x_1 x_2 + \varepsilon,$ etcétera.

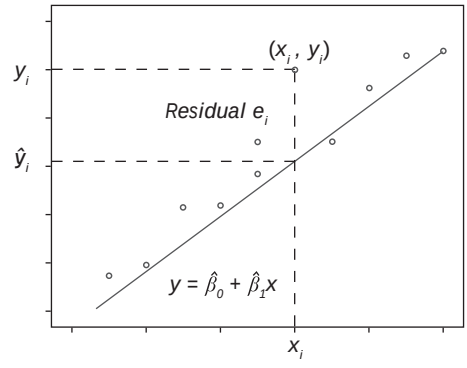
En cualquiera de los casos mencionados, para aplicar estos modelos habrá que seguir el camino de la mayoría de los procesos de investigación estadística. Este se puede resumir en los siguientes pasos:

1. Análisis *exploratorio* de los datos, resumen numérico de las variables, presentación del diagrama de dispersión, exploración de las relaciones entre las variables, etcétera.
2. *Formulación* del modelo. Este paso tiene que ver con el establecimiento de la forma de la relación que pueda existir entre las variables, pero sustentada con el conocimiento que se tenga de la realidad del proceso que se desea representar; es decir, con el "conocimiento del negocio".
3. *Estimación* del modelo. En buena cuenta, este paso tiene que ver con el "ajuste del modelo a los datos" que se tienen. Comprende la estimación de los parámetros, la evaluación del ajuste y el estudio de la precisión de los coeficientes.
4. *Estudio de la adecuación* del modelo, para analizar la bondad de ajuste del modelo a los datos.
5. *Evaluación* del modelo. Este paso tiene que ver con la comprobación de los supuestos del modelo.
6. *Informe y uso* del modelo.

Estimación del modelo de regresión lineal simple. El método de mínimos cuadrados

La estimación de los parámetros β_0 *y* β_1 se realiza, como se indicó en el capítulo 3, usando el *método de mínimos cuadrados* y los valores obtenidos de la muestra $(x_1, y_1),\ldots, (x_n, y_n)$.

FIGURA 9.2 *Recta de mínimos cuadrados*



(x_i, y_i) punto observado

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

El modelo estimado es $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Se considera que los residuales $y_i - \hat{y}_i$ son estimadores de los errores ε_i.

Aparte de los coeficientes de la ecuación de regresión es preciso estimar a la varianza σ^2 de ε. El método de mínimos cuadrados no proporciona directamente un estimador de este parámetro. Sin embargo, considerando que los residuos $e_i = (y_i - \hat{y}_i)$ se "parecen" a los errores aleatorios ε_i, y que la media de estos es 0, se usa como estimador de σ^2 a:

$$\hat{\sigma}^2 = \frac{SCE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Este estimador de σ^2 se llama *varianza residual* y mide la cantidad por la cual los valores verdaderos y_i difieren de los valores estimados $\hat{y}_i$.

EJEMPLO. *Salarios vs. gastos de alquiler*

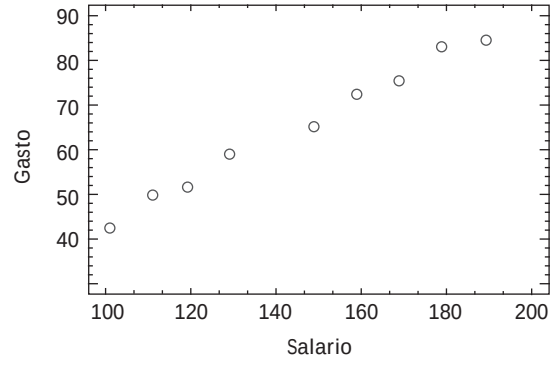
Para cada uno de los 10 obreros de una empresa se ha recabado: su salario diario en dólares (X) y el gasto diario, en dólares, del alquiler de vivienda (Y). Los datos obtenidos aparecen a continuación (Tabla 9.1).

TABLA 9.1 *Salarios vs. gastos de alquiler*

X	100	110	120	130	140	150	160	170	180	190
Y	45	51	54	61	66	70	74	78	85	89

El diagrama de dispersión correspondiente, que vemos a continuación (Figura 9.3), indica que existe una marcada relación lineal entre el salario y el gasto en alquiler, permitiendo el ajuste de un modelo de regresión lineal $y = \beta_0 + \beta_1 x + \varepsilon$.

FIGURA 9.3 *Diagrama de dispersión para el salario y el alquiler*



Usando:

$$\bar{x} = 145, \quad \bar{y} = 67.3, \quad \sum_{i=1}^{10} x_i y_i = 101570, \quad \sum_{i=1}^{10} x_i^2 = 218500$$

se obtienen los estimadores de β_1 y β_0, respectivamente:

$$\hat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{(1/n)\sum_{i=1}^{10} x_i y_i - \bar{x} \bullet \bar{y}}{(1/n)\sum_{i=1}^{10} x_i^2 - \bar{x}^2} = 0.4830, \quad \hat{\beta}_0 = \bar{y} - b\bar{x} = -2.7393$$

El modelo estimado es $y = -2.7393 + 0.4830x$.

Predicción de valores usando la recta de regresión

Uno de los usos del modelo de regresión es la predicción del valor de la variable dependiente cuando se conoce un valor particular de la variable independiente. La exigencia para que esta predicción sea adecuada es que la recta *ajuste bien a los datos* y que los valores para los cuales se desea predecir estén preferentemente en el rango de valores de los cuales se dispone.

El diagrama de dispersión es una ayuda primordial para evaluar el buen ajuste de la recta a los datos; sin embargo, una señal de buena adecuación del modelo a la data es el *índice de correlación lineal* entre la variable independiente y la variable dependiente. Si el índice de correlación es cercano a 1 o a –1 podremos indicar que la recta ajusta bien a los datos que se tienen.

Los valores observados de y, los valores estimados con la recta de regresión, $\hat{y}$, los residuales, $y - \hat{y}$, así como estos elevados al cuadrado, aparecen a continuación.

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
100.00	45.00	45.56364	−0.56364	0.32
110.00	51.00	50.39394	0.60606	0.37
120.00	54.00	55.22424	−1.22424	1.50
130.00	61.00	60.05455	0.94545	0.89
140.00	66.00	64.88485	1.11515	1.24
150.00	70.00	69.71515	0.28485	0.08
160.00	74.00	74.54545	−0.54545	0.30
170.00	78.00	79.37576	−1.37576	1.89
180.00	85.00	84.20606	0.79394	0.63
190.00	89.00	89.03636	−0.03636	0.00

La suma de los cuadrados de los residuales es igual a 7.22 y el estimador de la varianza de la parte aleatoria del modelo, σ^2, es igual a 7.22/8 = 0.9503.

Descomposición de la varianza: Anova

Si el modelo es adecuado para los datos, se puede usar el modelo estimado para predecir el valor de la variable dependiente *Y* cuando la variable independiente *X* toma un valor particular x_0. Cuando el modelo no ajusta bien a los datos la predicción del valor *Y* para cualquier valor de *X* es la media de los valores de *Y.* La pregunta es ¿por qué no usar *siempre* la media de todos los valores de *Y* que se tiene en la muestra para hacer cualquier predicción, sin importar el valor *x*? Para contestar esta pregunta consideremos la información contenida en la tabla de datos, referida a la cantidad *X,* en dólares, que una persona obtiene como salario diario y el gasto diario en vivienda, en dólares, *Y*, correspondiente.

TABLA 9.2 *Sueldo y gasto*

Salario: X	*Gasto en vivienda: Y*
100.00	45.00
110.00	51.00
120.00	54.00
130.00	61.00
140.00	66.00
150.00	70.00
160.00	74.00
170.00	78.00
180.00	85.00
190.00	89.00

Para cualquiera que sea el valor *x,* se usará la media, $\bar{y} = \frac{\sum y_i}{n} = 67.30$, para predecir el valor correspondiente de *Y.* De este modo, si una persona gana $x = 120$ o 160, el valor que se "esperaría" que gaste es 67.30.

Para evaluar el error total que se produce cuando se predicen los valores de *y* con la media $\bar{y}$, los estadísticos usan la *suma de cuadrados total*, $SCT = \sum (y_i - y)^2$.

Usando los datos, se tiene que:

$$SCT = \sum (y_i - \bar{y})^2 = (45 - 67.30)^2 + (51 - 67.30)^2 + \ldots + (89 - 67.30)^2 = 1932.100$$

Prácticamente, esta suma es la medida de la variación de los valores de *Y* alrededor de su media.

Por otro lado, si para cada valor de x_i se predice el valor de y_i usando la ecuación de regresión estimada, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, el error que se comete es la suma de cuadrados de los errores:

$$SCE = \sum (y_i - \hat{y}_i)^2$$

Usando los datos, se encuentra que $SCE = 7.2224$.

El error de predicción esta vez es menor, y la diferencia de este valor con respecto al anterior es igual a:

$$SCR = SCT - SCE = 1{,}932.100 - 7.2224 = 1{,}924.876$$

Se puede indicar que el uso de la variable *X* reduce el error de predicción; es decir, el conocimiento de *X* ayuda a explicar la variabilidad de los valores de *Y.*

Una medida de la ayuda que proporciona *X* en la predicción se determina con la proporción de *SCR*, respecto de la variación total, *SCT*. Esta proporción se denota con R^2. Para el ejemplo se tiene:

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCE}{SCT} = (0.998)^2 = 0.996$$

R^2 representa la proporción de la variación total que es explicada al usar los diferentes valores x_i. A R^2 *se* le llama ***índice de determinación***. En el ejemplo, el 99.6% de la variabilidad de los gastos en alquiler es explicada por los salarios. El 0.4% de la variación de *Y* alrededor de $\overline{Y}$ permanece sin explicación.

Se demuestra que *R,* la raíz cuadrada de R^2, es igual al *índice de correlación de Pearson* entre los salarios y los gastos en vivienda correspondientes.

A *SCR* se le llama *suma de cuadrados debida a la regresión* y se demuestra que $SCR = \sum(\hat{y}_i - \bar{y})^2$.

Los paquetes estadísticos reportan los resultados como los que aparecen en las tablas (9.3).

TABLA 9.3 *Salida del computador*

Resumen del modelo

Modelo	*R*	*R. cuadrado*	*R. cuadrado corregida*	*Error típ. de la estimación*
1	0.998[a]	0.996	0.996	0.95028

Anova

Modelo		*Suma de cuadrados*	*gl*	*Media cuadrática*	*F*	*Sig.*
1	Regresión	1,924.876	1	1,924.876	2,131.574	0.000
	Residual	7.224	8	0.903		
	Total	1,932.100	9			

Coeficientes

Modelo		*Coeficientes no estandarizados*		*Coeficientes estandarizados*	*t*	*Sig.*
		B	Error estándar	Beta		
1	(Constante)	−2.739	1.546		−1.771	0.114
	Salario	0.483	0.010	0.998	46.169	0.000

En el primer cuadro se indican el número del modelo (1), el índice de correlación *R* y el índice de correlación al cuadrado. También se indica el índice de correlación al cuadrado modificado, que para una variable coincide con el índice de correlación al cuadrado. Aparece así mismo el estimador de la desviación estándar del modelo (*error estándar del modelo*).

$$\hat{\sigma} = \sqrt{\frac{SCE}{n-2}} = 0.9503$$

En el segundo cuadro de resultados, titulado Anova, se indican: la suma de cuadrados según la regresión, la suma de cuadrados de los residuos, la suma de cuadrados total y los grados de libertad, df. Aparece también el valor *p* (*p-value*). Para el ejemplo, el valor *p* que aparece (0.000) indica que el modelo es adecuado para la predicción de *Y* y que el valor 0.996 de R^2, no ha sido obtenido por azar.

En el tercer cuadro, aparecen:

- En la columna **B**, el estimador de la constante, $\hat{\beta}_0 = -2.739$, y el estimador del coeficiente del salario, $\hat{\beta}_1 = 0.483$.
- En la columna **Error estándar,** el error estándar de $\hat{\beta}_0$, 1.546, y el error estándar de $\hat{\beta}_1$, 0.010.
- En la columna **coeficientes estándarizados,** los estimadores de los coeficientes del modelo que se obtiene al hacer la regresión entre los valores estandarizados de las variables dependiente e independiente.

En el último cuadro también se obtienen los extremos de los intervalos de confianza al 95% para cada coeficiente.

Adecuación del modelo

Uno de los usos del modelo de regresión es la predicción de los valores de *Y* a partir de los valores de *X* cuando el *modelo es adecuado*. El modelo es *adecuado* si la reducción del error de predicción es significativa al usar la regresión. Por ello, el modelo es adecuado *a nivel de muestra* si el índice de determinación R^2 tiende a 1.

El análisis de la adecuación del modelo a *nivel de población* se reduce al análisis de la significación estadística de la suma de cuadrados debido a la regresión (*SCR*) respecto de la suma de los cuadrados de los errores (*SCE*). Esta significación puede medirse con un estadístico de prueba cuyo valor *F* se reporta en la tabla Anova.

Para el modelo de regresión lineal simple el análisis de la adecuación del modelo a nivel de población se hace simplemente contrastando la hipótesis nula H_0: $\beta_1 = 0$ versus la hipótesis alternativa H_A: $\beta_1 \neq 0$.

Si no se rechaza la hipótesis nula podemos decir que no existe suficiente información para indicar que los valores de *X* explican a los valores de *Y.* Si se rechaza la hipótesis nula, el modelo es adecuado a nivel de población.

La prueba de este contraste se basa en el estadístico de prueba, $t = \frac{\hat{\beta}_1}{es(\hat{\beta}_1)}$, en donde $es(\hat{\beta}_1)$ es la desviación estándar o error estándar del estimador $\hat{\beta}_1$.

Cuando la hipótesis nula es verdadera, este estadístico tiene distribución *t student con n* − 2 grados de libertad .

Al nivel de significación α, la hipótesis nula se rechaza si $|t| > t_{1-\alpha/2,\, n-2}$, en donde $t_{1-\alpha/2,\, n-2}$ es el cuantil de orden $1-\alpha/2$ en la distribución *t* − *student* con *n* − 2 grados de libertad.

En el ejemplo, el valor del estadístico *t* es 30.650, valor muy significativo; la hipótesis nula se rechaza.

Validación del modelo: análisis de residuos

Las inferencias y luego las decisiones que se toman a partir de un modelo de regresión dependen del cumplimiento de los supuestos que se hicieron al definir el modelo; la comprobación de estos debe ser considerada como uno de los pasos importantes del proceso que se está desarrollando. Los supuestos se refieren al término ε y son los siguientes:

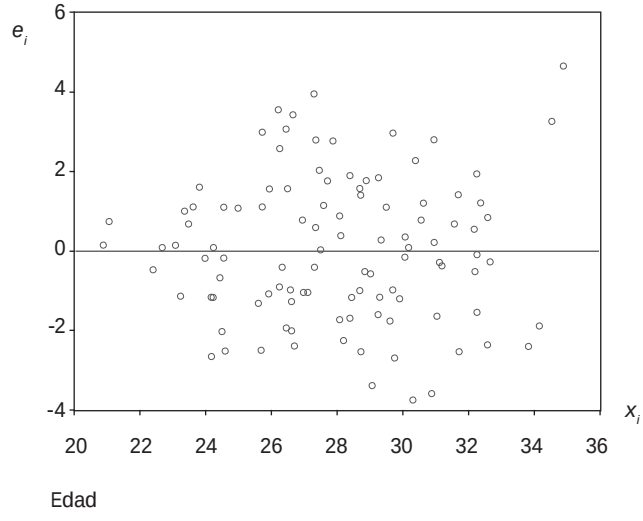
1. El término ε tiene media 0.
2. La varianza de ε es constante, cualquiera que sea el valor de la variable independiente.
3. El término ε tiene distribución normal.
4. Los valores de ε son no correlacionados.

La teoría estadística indica que las dos últimas suposiciones implican que los errores son independientes, siendo imposible extraer información de ellos, y que la mayor cantidad de la variabilidad de la variable respuesta o dependiente puede ser explicada por la parte estructural del modelo.

Las asunciones indicadas no se pueden comprobar directamente, pues los errores ε no se conocen; sin embargo, se pueden aproximar con los residuales $e_i = y_i - \hat{y}_i$. Los residuales pueden considerarse que son realizaciones de los errores aleatorios, y por lo tanto cualquier incumplimiento de las asunciones para los errores se observarán en los residuales.

Las gráficas de dispersión de los valores de x_i con los residuales e_i ayudan a verificar si se cumplen o no las asunciones 1, 2 y 3. Si estas se cumplen debemos esperar que en el diagrama de dispersión no existan tendencias y que la variabilidad de los residuales se mantenga aproximadamente constante (que exista *homocedasticidad*), como en la Figura 9.4.

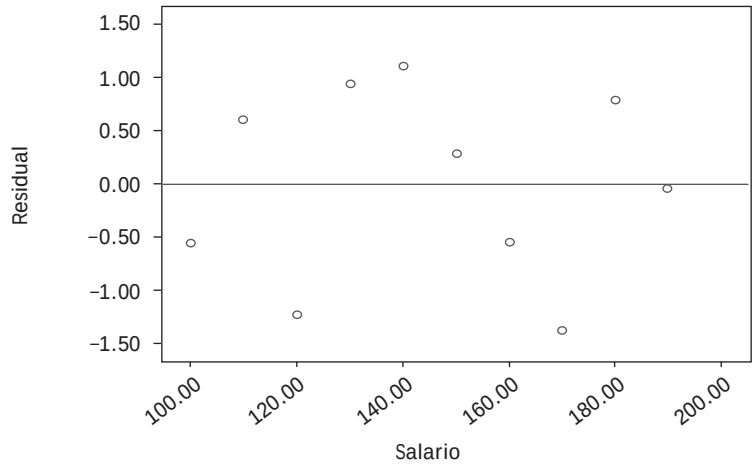
FIGURA 9.4 *Homocedasticidad*



En el caso de la regresión lineal simple se obtienen iguales resultados en el análisis si se estudia el diagrama de dispersión que resulta cuando en el eje horizontal se colocan los valores de Y o los valores de $\hat{Y}$ en lugar de los valores de la variable X.

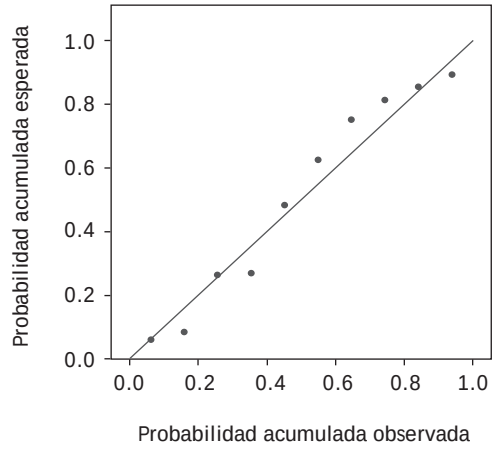
Para el ejemplo anterior, el diagrama de dispersión de los salarios vs. los residuales del modelo de regresión aparecen a continuación (Figura 9.5). Se observa que la dispersión de los residuales es constante para cualquier valor del salario y que no se observa patrón alguno que permita decir que de los residuales aún se puede extraer información.

FIGURA 9.5 *No existe patrón alguno*



El histograma de los residuos puede graficarse, pero cuando hay pocos puntos no se espera que brinden buena información. Una mejor alternativa es el gráfico *p – p* de los residuos, como el que aparece a continuación (Figura 9.6). Usando este gráfico se comparan los percentiles teóricos de la normal con los percentiles muestrales. Si los puntos están alineados alrededor de la recta diagonal del primer cuadrante, se considera que los residuales tienen la distribución normal.

FIGURA 9.6 *Gráfico p-p normal de los residuales*



Usos del modelo

Estimación de la respuesta media

A menudo interesa estimar la *respuesta media* de Y conocido el valor $x = x_0$ de X. La estimación de este valor, que se denota con $E(Y|x_0)$, puede hacerse de dos maneras: puntualmente y por intervalos de confianza.

- Un estimador puntual para $E(Y|x_0)$ es $(\hat{\beta}_0 + \hat{\beta}_1 x_0)$, cuando $x = x_0$.
- Si los supuestos del modelo se cumplen, y lo que se desea es hacer una inferencia que nos informe también de la cercanía del estimador de $E(Y|x_0)$ respecto de este parámetro, se usa el intervalo de confianza para $E(Y|x_0)$. Se demuestra que el intervalo de confianza para este valor, al nivel de confianza de $(1 - \alpha)100\%$, es:

$$\left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) - t_{1-\alpha/2,\, n-2}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}},\ (\hat{\beta}_0 + \hat{\beta}_1 x_0) + t_{1-\alpha/2,\, n-2}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right]$$

donde:

$t_{1-\alpha/2,\, n-2}$ es el cuantil de orden $1 - \alpha/2$ de la distribución t con $n - 2$ grados de libertad.

El valor $t_{1-\alpha/2,\, n-2}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$ es el margen de error del estimador.

En el ejemplo, el estimador puntual del gasto *promedio* de los empleados cuyo salario es $x_0 = 155$ es:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = -2.739 + (0.483)(155) = 72.1260$$

El intervalo con un nivel de confianza del 95% del gasto promedio de todos los empleados que ganan $x_0 = 155$ está dado por:

$$\left[72.1260 - (t_{1-0.05/2,\, 10-2})(0.9503)\sqrt{\frac{1}{10} + \frac{(155 - 145.00)^2}{8250.00}},\right.$$

$$\left.72.1260 + (t_{1-0.05/2,\, 10-2})(0.9503)\sqrt{\frac{1}{10} + \frac{(155 - 145.00)^2}{8250.00}}\right]$$

$t_{1-0.05/2,\, 10-2}$ es el percentil de la distribución t – *student* con 10 – 2 grados de libertad para el cual se cumple $P(t > t_{1-0.05/2,\, 10-2}) = 0.05/2$. Usando la tabla de la distribución t, se encuentra que este valor es 2.3060.

Reemplazando y haciendo los cálculos se tiene que [71.39654, 72.86406] es el intervalo de confianza, al nivel del 95%, para el gasto promedio de los empleados cuyo salario es 155 dólares.

Predicción de nuevas observaciones

El modelo de regresión también se usa para predecir una nueva observación de *Y* para un valor determinado de *X*. Esta predicción puede hacerse de dos formas: puntualmente y por intervalos de confianza.

- Si $x = x_0$, la mejor predicción puntual del valor correspondiente *y* de *Y* es $\hat{y} = (\hat{\beta}_0 + \hat{\beta}_1 x_0)$.
- El "intervalo de predicción" al nivel de confianza $(1 - \alpha)100\%$ para la nueva predicción de *Y* cuando $x = x_0$ es:

$$\left[(\hat{\beta}_0 + \hat{\beta}_1 x_0) - t_{1-\alpha/2,\, n-2}\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}},\ (\hat{\beta}_0 + \hat{\beta}_1 x_0) + t_{1-\alpha/2,\, n-2}\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}\right]$$

En el ejemplo, la predicción puntual del gasto para *un* individuo que gana $x_0 = 155$ dólares es igual al estimado del gasto promedio de todos los empleados que ganan esta cantidad.

El intervalo de confianza al 95% de la predicción del gasto de *un* individuo que gana $x_0 = 155$ dólares es:

$$\left[72.1260 - (2.3060)(0.9503)\sqrt{1 + \frac{1}{10} + \frac{(155 - 145.00)^2}{8250.00}},\right.$$

$$\left. 72.1260 + (2.3060)(0.9503)\sqrt{1 + \frac{1}{10} + \frac{(155 - 145.00)^2}{8250.00}}\right]$$

Es decir, el gasto promedio de *un* individuo que gana $x_0 = 155$ dólares está en el intervalo [69.81937, 74.44124] .

Observaciones

1. Notar que el margen de error de las estimaciones es menor a medida que el valor x_0, para el cual se desea predecir, está más cerca de la media de los valores de la variable independiente, *X*.
2. El intervalo de predicción de la media de los valores de la variable dependiente es más preciso que el intervalo de predicción de un valor individual de la variable dependiente.

APLICACIÓN: El caso de la empresa de transportes Tranper

La empresa de transportes Tranper transporta productos minerales de los centros mineros de la región a los centros de procesamiento, y viene operando desde hace mucho tiempo en el mismo lugar. Esta empresa desea establecer los precios que debe cobrar por el transporte de los productos como carga adicional en los camiones que están parcialmente cargados, por lo que ha decidido investigar acerca de los factores que podrían incidir en el costo al agregar carga adicional. La empresa cree que el único costo adicional en el que se incurre al agregar carga al camión es el costo adicional del combustible, pues los técnicos indican que el rendimiento del combustible bajaría al aumentar la carga. Para validar esta sospecha de la empresa y después de verificar que todos los camiones son idénticos y que realizan las entregas bajo las mismas condiciones, se ha determinado que una sola variable puede afectar el número de millas por galón de combustible, y que esta es el peso adicional de la carga.

La empresa cuenta con una gran base de datos en donde están registrados los pesos de la carga, los kilómetros recorridos y el número de galones de gasolina utilizados. De esta base de datos se ha tomado al azar una muestra de 18 viajes, obteniéndose la siguiente información, correspondiente a los pesos adicionales en cada viaje, así como los kilómetros recorridos por galón de combustible.

Los datos así como las salidas aparecen a continuación.

Y: km por galón	*X: peso en miles de kilos*
81	30
80	25
64	40
68	36
70	37
80	31
62	39
60	41
62	36
80	28
75	30
72	35
64	40
81	23
75	26
74	40
80	25
64	30

>>>

FIGURA 9.7 *Peso vs. kilómetros recorridos por galón*

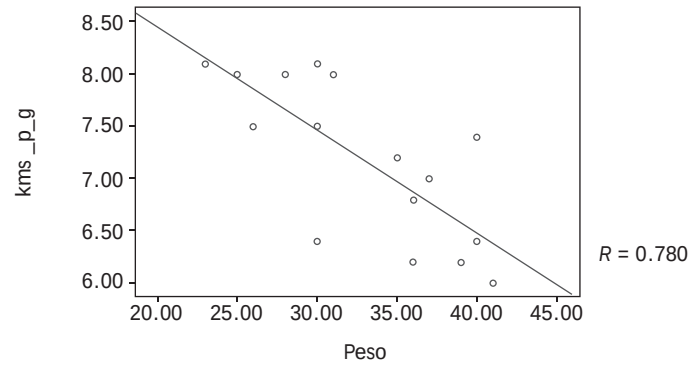


TABLA 9.4 *Peso vs. kilómetros recorridos por galón*

Resumen del modelo

Modelo	*R*	*R cuadrado*	*R cuadrado corregida*	*Error típ. de la estimación*
1	0.780[a]	0.608	0.584	0.49367

Los resultados, que el lector puede encontrar respecto de los coeficientes del modelo, permiten considerar que la ecuación de la recta de regresión obtenida, $y = 104.33 - 0.99x$, donde x se mide en kilos de peso por galón e y se mide en kilómetros por galón, es útil para el estudio, pues la pendiente −0.99 es significativamente diferente de 0, aun cuando el valor de r^2 es relativamente bajo.

La pendiente de la recta indica que por cada mil kilos de carga adicional el espacio recorrido se reduce en 0.99 kilómetros por galón de combustible.

Usando los resultados se calcula, como sigue, el costo de transportar 1,000 kilos adicionales a lo largo de 100 kilómetros, si, por ejemplo, se conoce que el galón de combustible cuesta 1.5 dólares.

Los valores de ***Y*** permiten indicar que en promedio se recorre 71.78 kilómetros por galón, luego el costo por transportar 1,000 kilos de carga a lo largo de 100 kilómetros es 1.5(100/(71.78)) = 2.089 dólares.

El costo que se tendría por transportar 1,000 kilos de carga adicional a la misma distancia es 100(1.5)/(71.78 − 0.99) = 2.12 dólares.

9.3 Regresión lineal múltiple

A menudo no es suficiente una variable independiente para explicar a la variable dependiente. Los gastos mensuales de una persona no necesariamente están explicados tan solo por el sueldo de la persona. Tal vez sean necesarias otras variables como la edad, los años de estudios, el número de hijos, etc. Un modelo que puede ayudar en este sentido es el que considera que los valores de la variable dependiente *Y* se pueden expresar en términos de los valores $x_1, x_2,..., x_k$ de las variables independientes respectivas $X_1, X_2,..., X_k$ mediante la relación:

$$y = \beta_1 + \beta_1 x_1 + ... + \beta_k x_k + \varepsilon$$

en donde ε es el *error aleatorio*, y que, como en el caso de la regresión lineal simple, se supone que tiene distribución normal de media 0 y varianza constante σ^2.

La estimación de los parámetros del modelo se realiza, como en el caso de la regresión lineal simple, con el método de los mínimos cuadrados, de tal manera que la suma de los cuadrados de los residuos:

$$SCE = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2$$

sea mínima.

El valor $MCE = \dfrac{SCE}{n - (k + 1)}$ se toma como estimador de σ^2.

El coeficiente β_i de cada variable independiente X_i mide el cambio que se obtiene en la variable dependiente cuando el valor de la variable independiente se cambia en una unidad, manteniendo constantes las otras variables independientes.

Una serie de paquetes estadísticos que permiten la estimación y el análisis de los modelos de regresión lineal han sido elaborados. Entre ellos están: el SAS, el SPSS, el MINITAB, el STATISTICA, el S-PLUS, etcétera.

EJEMPLO. *Gastos en el hogar*

Se desea modelar los gastos de los hogares *Y* en una ciudad en términos de los ingresos X_1 y el número de miembros en cada hogar, X_2, usando el modelo de regresión lineal múltiple $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, donde x_1, x_2 e *y* son valores de las variables X_1, X_2 *Y*, respectivamente.

Usando el método de mínimos cuadrados se obtienen los estimadores de los coeficientes:

$$\hat{\beta}_0 = 1{,}075.576, \quad \hat{\beta}_1 = 0.248, \quad \hat{\beta}_2 = 415{,}432, \quad \hat{\sigma}^2 = (502{,}59)^2$$

El estimador de la varianza de ε es $\hat{\sigma}^2 = (502.59)^2$.

El modelo estimado es $y = 1{,}075.576 + 0.248X_1 + 415{,}432X_2$.

Cada coeficiente $\hat{\beta}_j$, para $j = 1, 2$, se interpreta como un estimado del cambio del gasto Y cuando el valor de X_j se incrementa una unidad y la otra variable se mantiene fija. Así, cuando el ingreso se incrementa en una unidad el gasto se incrementa 0.248 unidades monetarias.

Gastos	*Ingresos*	*Miembros*
4,089.00	5,500.00	3
3,232.00	3,100.00	2
5,173.00	3,300.00	4
4,815.00	5,100.00	5
1,937.00	3,200.00	2
4,143.00	5,600.00	2
2,804.00	3,800.00	1
3,421.00	4,100.00	2
4,837.00	6,700.00	4
4,183.00	5,200.00	3
4,281.00	2,600.00	3
4,292.00	4,900.00	4
2,550.00	2,800.00	1
2,587.00	3,400.00	2
4,287.00	6,600.00	3
5,038.00	6,400.00	4
4,485.00	4,300.00	6
2,521.00	2,200.00	2
3,068.00	4,500.00	1
4,244.00	3,800.00	5
5,751.00	6,300.00	6
3,696.00	2,200.00	3
5,374.00	5,600.00	7
3,093.00	4,300.00	2

Como el valor de β_j puede verse afectado por las unidades de medida de X_j, se acostumbra medir la contribución real de la variable en la estimación del valor medio de *Y* con los "β_j estandarizados". Los estimadores de estos coeficientes se obtienen al efectuar la regresión entre los valores estandarizados de *Y* y los valores estandarizados de las variables independientes.

En el ejemplo anterior, el coeficiente estandarizado del "ingreso" es 0.340, mientras que el coeficiente estandarizado de la variable "miembros" es 0.680. Carentes de unidades, se puede decir que la variable "miembros" tiene mayor importancia en la explicación del ingreso que la variable "ingresos".

Descomposición de la varianza

Como en el caso de la regresión lineal simple, la suma de cuadrados total *SCT* se puede descomponer de la siguiente manera:

$SCT = SCR + SCE,$ en donde, como se recuerda:

$$SCT = \sum (y_i - \bar{y})^2$$

$$SCE = \sum (y_i - \hat{y}_i)^2$$

$$SCR = \sum (\hat{y}_i - \bar{y})^2$$

La proporción *SCR/SCT,* que se denota con R^2 y es igual a:

$$R^2 = \frac{SCR}{SCT} = \frac{SCT - SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

representa la proporción de la variación total de los *gastos en la muestra* que es explicada por los ingresos y el número de miembros. El valor R^2 se llama ***coeficiente de determinación***.

El coeficiente de determinación tiene valores entre 0 y 1, y su interpretación es como en la regresión lineal simple.

Para el ejemplo relativo al gasto, el valor de R^2 es igual a 0.778, indicando que el 77.8% de la variación de los valores *muestrales* de la variable *Y* es explicado en términos de las variables independientes.

(El coeficiente de determinación aumenta o permanece igual a medida que se añaden variables independientes al modelo. Por ello, y en forma alternativa al coeficiente de determinación R^2, se usa el coeficiente de *determinación corregido*, que se calcula con:

$$R_a^2 = 1 - \frac{n - 1}{n - (k + 1)} (1 - R^2)$$

Este coeficiente toma el valor 1 cuando el ajuste lineal es perfecto pero puede ser negativo cuando el ajuste es deficiente.)

Pruebas de hipótesis

¿El modelo es adecuado?

El valor de R^2 es una medida que permite analizar la adecuación del modelo a los *datos de la muestra*; una manera de medir si el modelo es adecuado para predecir los valores de *Y*, en términos de los valores de las variables independientes X_1,..., X_k y *a nivel de toda la población,* es mediante el contraste de las hipótesis:

$H_0 : \beta_1 = ... = \beta_k = 0$

H_A : *Al menos uno de los* β_j *es diferente de* 0

El rechazo de la hipótesis nula indicará que existe una relación significativa entre la variable dependiente y todas las variables independientes, y que al menos una de las variables independientes contribuye significativamente en la predicción. Es decir que el modelo es mejor para la predicción que el simple promedio de los valores de *Y*.

La prueba de hipótesis se basa en la "fórmula de descomposición de la varianza", que indica, como en el caso de la regresión lineal simple, que la variabilidad de los valores de *Y, SCT,* puede descomponerse en la suma de cuadrados según la regresión, *SCR,* más la suma de cuadrados de los errores, *SCE:*

$$SCT = SCR + SCE$$

El modelo es adecuado si la suma *SCR* es significativamente mayor que *SCE.*

El estadístico de prueba para esta situación es $F = \dfrac{SCR/k}{SCE/(n - (k + 1))}$.

Este estadístico corresponde a una variable aleatoria con distribución *F* con *k* grados de libertad para el numerador y $(n - (k + 1))$ grados de libertad para el denominador.

La hipótesis nula se rechaza si el valor del estadístico de prueba está en la zona de rechazo, correspondiente al nivel de significación de la prueba.

En la Tabla 9.5 se presentan los valores de *F* y su significación o valor *p*.

TABLA 9.5 *Anova*

Modelo		*Suma de cuadrados*	*gl*	*Media cuadrática*	*F*	*Sig.*
1	Regresión	18568613	2	9284306.6	36.754	.000
	Residual	5304718.7	21	252605.652		
	Total	28373332	23			

El valor del estadístico *F* es 36.754 y el correspondiente *valor – p* es 0.000 (la probabilidad de equivocarse al rechazar la hipótesis nula es 0.000). La hipótesis nula:

$$H_0 : \beta_1 = \beta_2 = 0$$

se rechaza (frente a la hipótesis alternativa: alguno de los coeficientes es diferente de 0), al nivel de significación 0.05. El modelo es adecuado a nivel de población.

Pruebas de hipótesis individuales

La prueba anterior no indica si determinado coeficiente es significativamente diferente de 0. Para analizar si determinada variable X_j contribuye o no significativamente en el valor promedio de *Y* se realiza una prueba de carácter "individual", contrastando la hipótesis nula $H_0 : \beta_j = 0$ versus la hipótesis alternativa $H_A : \beta_j \neq 0$. El valor del estadístico de prueba se indica en la columna *t* de la Tabla 9.6. Estas pruebas permiten al investigador eliminar o incluir otras variables que tal vez sean de utilidad para explicar la variable dependiente. El estadístico que se usa en este caso se basa en la distribución *t* student con $n - k - 1$ grados de libertad.

Para el ejemplo anterior, y usando los resultados de la tabla de los "coeficientes", se puede indicar que la hipótesis nula $H_0 : \beta_1 = 0$, que corresponde al "ingreso", se rechaza frente a la hipótesis alternativa $H_0 : \beta_1 \neq 0$, al nivel de significación 0.05. De igual manera se tiene para el coeficiente de la variable "miembros".

TABLA 9.6 *Estimación de los coeficientes*

		Coeficientes no estandarizados		***Coeficientes estandarizados***		
Modelo		B	Error típ.	Beta	***t***	***Sig.***
1	(Constante)	1075.576	370.797		2.901	0.009
	Ingreso	0.248	0.083	0.340	2.975	0.007
	Miembros	415.432	69.737	0.680	5.957	0.000

Usos del modelo de regresión múltiple para la predicción

A partir del modelo de regresión es posible estimar la respuesta media de los valores de la variable dependiente *Y*, que corresponden a valores fijos de las variables independientes.

La estimación puede hacerse puntualmente y por intervalo de confianza.

En el ejemplo, la estimación puntual del promedio del gasto, cuando los valores de las variables "ingreso" y "miembros" son 3,700.00 y 3, respectivamente, es:

$$\hat{y} = 1{,}075.576 + 0.248(3{,}700) + 415{,}432(3) = 3{,}238.2286$$

La predicción individual del gasto también puede hacerse puntualmente y mediante un intervalo de confianza.

Selección de variables en la regresión múltiple

La elección del mejor modelo de regresión no es un trabajo simple. El investigador comienza con una serie de variables explicativas potenciales, de las cuales debe elegir las que mejor ayudan a explicar a la variable dependiente y que finalmente deben ser incluidas en la ecuación final. En esta etapa, además de tener en cuenta el costo en el cual se puede incurrir al obtener y monitorear las variables, es necesario considerar el criterio del investigador y el conocimiento del problema que desea resolver. Se han desarrollado una serie de procedimientos para ayudar en la selección de las "mejores variables" (selección hacia adelante, selección hacia atrás, selección paso a paso, etc.); sin embargo, ninguno de ellos produce un modelo ideal.

Por otro lado, en la búsqueda del modelo ideal, es necesario tener en cuenta que el mejor modelo no es aquel que "modela" hasta las perturbaciones y para el cual el error de predicción es 0. Estos modelos que así se obtienen se llaman sobreajustados. Cuando estos se aplican a nuevos conjuntos de datos seleccionados de la misma población, su desempeño no es bueno, como sugieren los primeros resultados. A menudo aparecen modelos sobreajustados cuando se tienen pocos datos y muchas variables. Una regla práctica que puede ayudar a evitarlos es considerar al menos 10 observaciones por cada variable independiente.

Otra de las precauciones que se debe tener en cuenta en la búsqueda del mejor modelo se refiere a la significancia conjunta de los coeficientes. Una regresión que es estadísticamente significativa no necesariamente es útil. Existen situaciones en las cuales los coeficientes son significativamente diferentes de 0, de manera conjunta; sin embargo, explican solo una pequeña porción de la variabilidad. Una regla práctica para obtener una ecuación de regresión útil indica que para un nivel de significación 0.05, el valor del estadístico *F* calculado en la prueba Anova debe ser mayor a cuatro veces el valor crítico de la distribución *F.*

APLICACIÓN: El caso de la empresa Venagra

La empresa Venagra, relacionada con la venta de instrumentos para la agricultura, ha sufrido en estos últimos meses una baja en las ventas de sus productos. Después de varias reuniones entre los directivos de las diferentes secciones se ha determinado que un estudio de los diversos factores relacionados con los vendedores, y que podrían explicar la variabilidad de las ventas que se realizan, ayudaría en las decisiones a tomar para la solución de este problema. Para ello, el gerente de personal ha decidido usar las ventas del último mes como la variable dependiente (*Y*) y como variables independientes, que podrían explicar a la variable dependiente, a las siguientes:

X_1 = resultado de una prueba de aptitud para las ventas, en una escala del 0 al 100 (prueba)

X_2 = edad en años (edad)

X_3 = experiencia en años (experiencia)

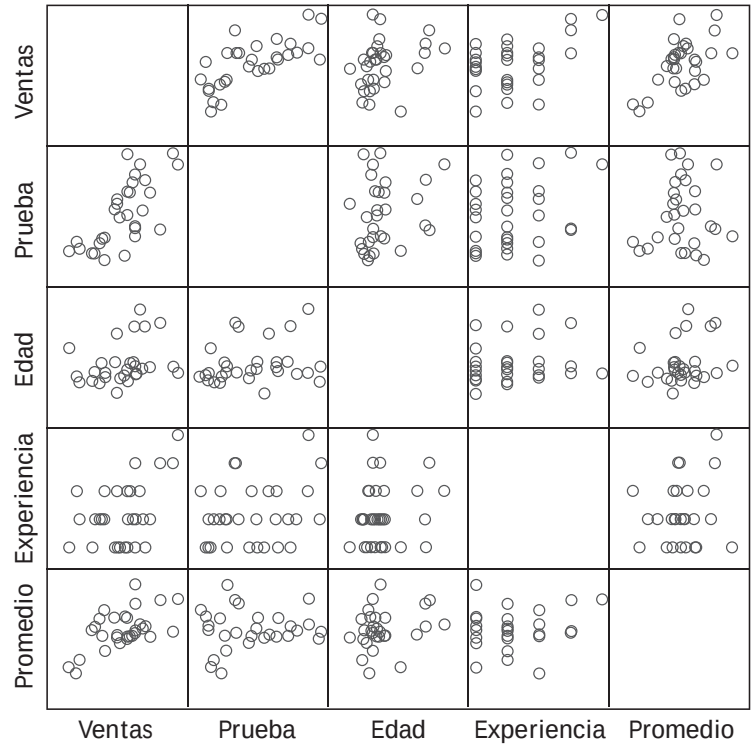
X_4 = promedio de las calificaciones de la escuela secundaria (promedio)

El gerente de personal tomó una muestra de 30 casos y los datos fueron los siguientes:

Ventas	*Prueba*	*Edad*	*Experiencia*	*Promedio*
45.00	11.00	26.32	0.00	14.09
48.00	20.00	26.07	1.00	14.70
62.00	26.00	27.61	0.00	16.52
72.00	32.00	31.45	3.00	15.69
60.00	65.00	27.90	2.00	13.75
62.00	81.00	26.87	1.00	14.07
56.00	43.00	26.47	0.00	13.40
59.00	66.00	27.17	0.00	14.74
65.00	49.00	27.34	1.00	14.18
68.00	65.00	27.49	1.00	13.73
55.00	55.00	25.17	0.00	13.69
46.00	11.00	27.05	1.00	14.27
54.00	50.00	27.96	0.00	14.75
77.00	100.00	27.56	3.00	14.01
66.00	76.00	31.16	0.00	14.30
36.00	13.00	29.20	0.00	12.11
59.00	45.00	27.28	2.00	13.76
39.00	21.00	26.71	2.00	11.79
61.00	74.00	27.98	1.00	13.79
50.00	5.00	26.60	2.00	15.16
58.00	9.00	26.76	0.00	14.80
59.00	99.00	26.26	1.00	13.62
50.00	25.00	27.01	1.00	12.97
55.00	59.00	30.49	2.00	13.84
49.00	24.00	27.89	1.00	13.78
64.00	90.00	32.65	2.00	14.40
62.00	35.00	31.11	1.00	15.50
40.00	15.00	26.16	1.00	12.50
62.00	33.00	27.01	3.00	13.94
79.00	90.00	26.98	4.00	15.75

>>>

FIGURA 9.8 *Diagramas de dispersión*



Los diagramas de dispersión que aparecen en la Figura 9.8 muestran la existencia de relaciones lineales entre la variable dependiente, "ventas", y las variables independientes, X_1, X_3, X_4. Se observa, por ejemplo, que, al parecer, los valores de la variable dependiente, "ventas", están alineados con los valores de la variable "prueba".

El valor de R^2 (0.811) indica que el modelo ajusta a los datos de la muestra. Este ajuste también sucede al nivel poblacional, como se puede observar al leer el valor del estadístico F = 26.848 en la tabla Anova. El valor de este estadístico es altamente significativo, permitiendo rechazar la hipótesis nula $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ frente a la hipótesis alternativa $H_A : H_0$ no es verdad.

Resumen del modelo

Modelo	*R*	*R cuadrado*	*R cuadrado corregida*	*Error típ. de la estimación*
1	0.901[a]	0.811	0.781	4.86997

>>>

TABLA 9.7 *Estimación de R y tabla Anova*

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2546.951	4	636,738	26.848	0.000
	Residual	592.915	25	23,717		
	Total	3139.867	29			

Los estimadores de los coeficientes permiten afirmar que la prueba de aptitud, la experiencia y el promedio de las calificaciones de la escuela secundaria explican significativamente los resultados de las ventas, por lo que la empresa deberá poner especial atención en estas variables en el momento de la toma de decisiones.

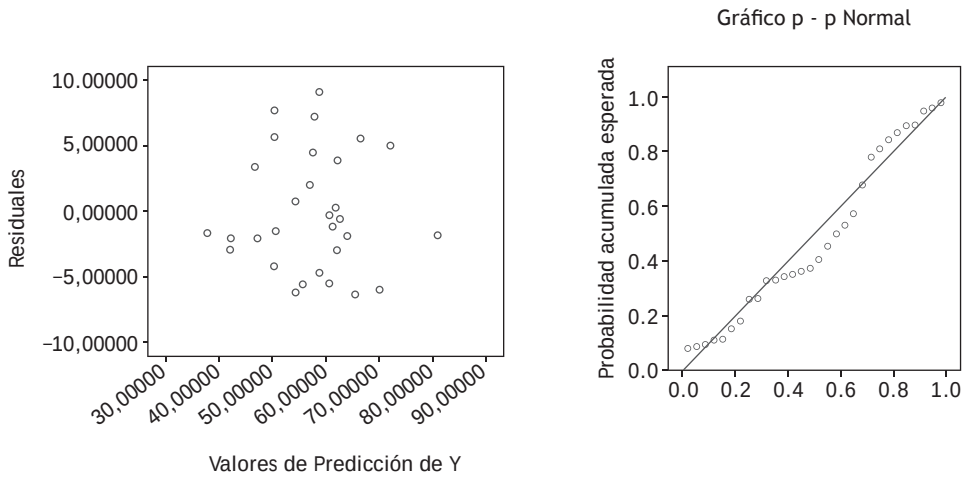
TABLA 9.8 *Estimación de los coeficientes*

		Coeficientes no estandarizados		Coeficientes estandarizados		
Modelo		B	Error típ.	Beta	t	Sig.
1	(Constante)	–28.164	16.970		–1.660	0.109
	Prueba	0.210	0.033	0.588	6.389	0.000
	Edad	0.066	0.527	0.011	0.124	0.902
	Experiencia	2.174	0.869	0.229	2.502	0.019
	Promedio	5.048	0.908	0.499	5.558	0.000

En la Fig. 9.9, el gráfico de los residuales versus los valores de predicción de las ventas, así como el gráfico P-P de estos indican que los supuestos de homocedasticidad y de normalidad, respectivamente, se cumplen. Ello valida las conclusiones obtenidas a nivel de población.

>>>

FIGURA 9.9 *Valores de predicción vs. residuales y gráfica p – p de los residuales*



APLICACIÓN: El caso de las tiendas Emco

La empresa EMCO opera 200 tiendas en diferentes lugares del país para la venta de artefactos electrónicos para el hogar. Los últimos informes indican que la curva de ventas mensuales *Y*, en miles de dólares, ha descendido a tal punto que ha tenido que cerrar algunas de las tiendas que regenta, en perjuicio de un buen número de empleados que han sido despedidos.

El gerente de comercialización, con el fin de enfrentar el problema que ha ocasionado el descenso de las ventas, ha determinado que es necesario un estudio estadístico que permita determinar las variables que pueden explicar la variabilidad de las ventas, como parte del análisis general, y es así como después de una serie de reuniones con los conocedores del negocio ha determinado que las siguientes variables podrían explicar las ventas:

X_1 = Área del terreno, en m^2, que ocupa la tienda

X_2 = Cantidad de habitantes que existen en 10 manzanas a la redonda de la tienda respectiva

X_3 = Número de líneas de buses que pasan por la calle en donde se encuentra la tienda

X_4 = Número de tiendas similares en 10 manzanas a la redonda que venden artefactos electrónicos para el hogar

>>>

La muestra contiene los siguientes datos:

Y	X_1	X_2	X_3	X_4
123.42	554.33	116,492	5	6
94.87	516.23	93,301	4	10
88.96	545.50	102,017	4	9
105.81	478.95	117,726	3	10
70.63	426.51	81,628	4	11
80.22	537.48	93,139	2	9
112.04	464.79	103,270	7	10
92.28	386.71	89,387	3	8
101.37	506.58	100,578	2	8
96.96	624.60	95,784	4	9
84.06	437.68	96,206	9	14
131.11	486.17	122,482	5	9
116.38	532.02	114,116	5	9
74.18	426.78	99,753	4	7
92.05	505.27	80,422	4	11
83.98	509.51	78,049	4	13
64.96	502.82	78,608	5	5
100.60	575.67	97,846	5	12
139.97	488.64	113,808	5	20
90.20	514.13	103,838	6	11
118.85	486.99	88,180	5	16
87.21	472.63	105,459	7	12
71.60	475.35	88,010	3	10
130.57	616.05	116,030	8	7
123.34	523.80	107,504	6	11
89.75	514.81	103,723	5	10
116.27	514.64	107,182	2	16

Al aplicar un modelo de regresión lineal de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

y realizando los cálculos, se obtuvo lo siguiente:

Los valores de los coeficientes de determinación R^2 y el coeficiente de determinación corregido son: 0.670 y 0.610, respectivamente. Las variables independientes explican de manera conjunta un porcentaje no muy alto de la variabilidad de las ventas. Es posible que se pueda usar otro modelo para elevar estos valores.

>>>

TABLA 9.9 *Resumen del modelo*

Modelo	*R*	*R cuadrado*	*R cuadrado corregida*	*Error típ. de la estimación*
1	0.818[a]	0.670	0.610	12.7080

Variables predictoras: (constante), X4, X2, X3, X1,
Variable dependiente: Y

El análisis de varianza (Anova) indica que el modelo de regresión lineal es adecuado. El valor del estadístico *F* es significativo (*F* = 11.157 y el valor de *p* es 0.000), lo que permite rechazar la hipótesis nula $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. la hipótesis alternativa H_A : H_0 *no es verdadera*.

TABLA 9.10 *Anova*

Modelo		*Suma de cuadrados*	*gl*	*Media cuadrática*	*F*	*Sig.*
1	Regresión	7207.237	4	1801.809	11,157	0.000
	Residual	3552.899	22	161.495		
	Total	10760.135	26			

El modelo estimado es $y = 72.662 + 0.079x_1 + 0.001x_2 + 0.168x_3 + 2.111x_4$.

Las variables X_2 y X_4, cantidad de habitantes y número de tiendas, respectivamente, son significativas (los valores *p* son: 0.000 y 0.013).

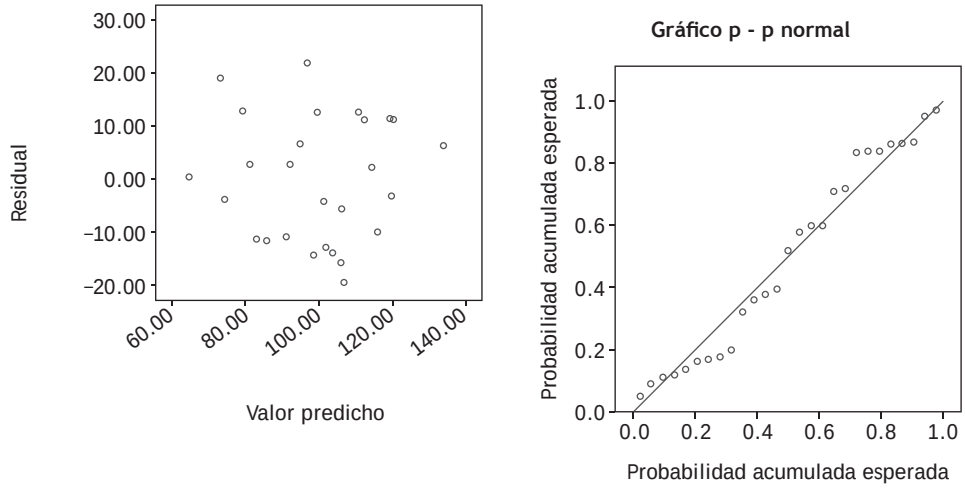
TABLA 9.11 *Estimación de los coeficientes*

		Coeficientes no estandarizados		*Coeficientes estandarizados*		
Modelo		B	Error típ.	Beta	*t*	*Sig.*
1	(Constante)	–72.662	29.847		–2.434	0.023
	X1	0.079	0.049	0.207	1.610	0.122
	X2	0.001	0.000	0.677	5.194	0.000
	X3	0.168	1.488	0.014	0.113	0.911
	X4	2.111	0.779	0.337	2.709	0.013

>>>

Al revisar las características de los residuos se puede indicar que al parecer estos tienen distribución normal y que gozan de la propiedad de homocedasticidad.

FIGURA 9.10 *Valores de predicción vs residuales y gráfica p – p plot de los residuales*



Conclusión: el gerente de comercialización deberá tener en cuenta la cantidad de población y el número de líneas de transporte que pasan alrededor de las tiendas en el momento de tomar una decisión.

Análisis de los residuales, detección de *outliers* y medidas influenciales

Los supuestos establecidos para el modelo de regresión lineal múltiple fueron los mismos que se indicaron para el modelo de regresión lineal simple; la comprobación de estas asunciones se realiza de igual modo que para este modelo.

El análisis de los residuales puede ayudar, algunas veces, a la detección de observaciones atípicas, que están fuera del patrón del resto de los datos. Si estas observaciones perturban la forma del modelo, se les llama *observaciones influenciales*. Existen índices, como la *medida de la distancia de Cook*, que ayudan en la detección de estos puntos.

Multicolinealidad

Si entre dos o más variables independientes de un modelo de regresión lineal existe una dependencia lineal se dice que existe *multicolinealidad* entre las variables.

En la práctica es muy común tener multicolinealidad. Esto supone información redundante para la predicción, lo que a simple vista parece beneficioso; sin embargo, la multicolinealidad lleva consigo la imprecisión en la estimación de los coeficientes del modelo, pues la multicolinealidad influye en la varianza de los estimadores. Algunas veces, cuando existe multicolinealidad aparecen coeficientes de las variables independientes con signos que no son coherentes con la teoría que respalda al modelo. Otras veces, la prueba de significación de todas las variables indica que el modelo es adecuado; sin embargo, las pruebas individuales indican que ninguna variable es significativa. Problemas como estos, originados por la multicolinealidad, obligan a usar procedimientos para detectarla. Uno de estos procedimientos se refiere al análisis de la correlación entre pares de variables independientes. Esta es una medida fácil de llevar a cabo; sin embargo, no es suficiente, pues una variable independiente puede estar correlacionada con un grupo de variables también independientes. Otro procedimiento, que tiene el mismo objetivo, consiste en observar la significación de los coeficientes de manera individual y de manera conjunta. *Generalmente cuando existe multicolinealidad, la contribución individual de cualquier variable no es significativa, mientras que la contribución de estas de manera conjunta sí lo es.*

Una medida adecuada para medir la multicolinealidad de una variable independiente con el resto de variables independiente es el *índice de inflación de la varianza (VIF)*. Este índice indica en cuánto la varianza de los coeficientes es *inflada* debido a la multicolinealidad.

Si el *VIF* es igual a 1, no existe multicolinealidad. Una regla práctica indica que cuando el VIF_i *es mayor que* 10 se puede decir que existe multicolinealidad severa.

Se demuestra que la varianza del estimador del coeficiente de la varianza de la variable X_i es directamente proporcional a VIF_i.

$$Error\ estándar\ de\ \hat{\beta}_i = SCE * VIF_i$$

De este modo, si el VIF_i es un valor grande, *el error estándar de* $\hat{\beta}_i$ es grande. Ello explica la falta de precisión de los coeficientes cuando el VIF_i es grande.

Algunos paquetes reportan como medida de multicolinealidad al índice de *tolerancia,* que es igual a la inversa del VIF_i: $TOL_i = 1/VIF_i$.

9.4 Modelos especiales de regresión

Modelos de regresión polinomiales

En muchas situaciones una respuesta *Y* depende de una variable independiente *X*, y de algunas de sus potencias. En tal caso los valores *y* pueden expresarse como:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \varepsilon$$

donde *k* es un entero positivo, $\beta_0, \beta_1, ..., \beta_k$ son parámetros a estimar y ε representa un valor aleatorio con las mismas asunciones del modelo general.

Casos particulares de estos modelos son los siguientes.

El *modelo de primer orden* en *x*: $y = \beta_0 + \beta_1 x + \varepsilon$, $k = 1$. Este modelo corresponde al modelo de regresión lineal simple.

El *modelo de segundo orden* en *x*: $Y = \beta_0 + \beta_2 x^2 + \varepsilon$, $k = 2$. Este modelo ajusta puntos que aproximadamente se desarrollan alrededor de una parábola.

La construcción del diagrama de dispersión puede ayudar en la elección del modelo que podría ajustar a los datos. Una regla práctica indica que el gráfico podría corresponder a un polinomio de orden *k* si la curva ajustada tiene $k - 1$ puntos, en donde existe un máximo o un mínimo local. Si el gráfico de la curva suavizada tiene dos puntos en donde hay un máximo o un mínimo local, entonces podría corresponder a un polinomio de tercer orden.

EJEMPLO. *Ajuste polinómico*

Usando los siguientes datos, ajustar un polinomio adecuado.

TABLA 9.12 *Datos*

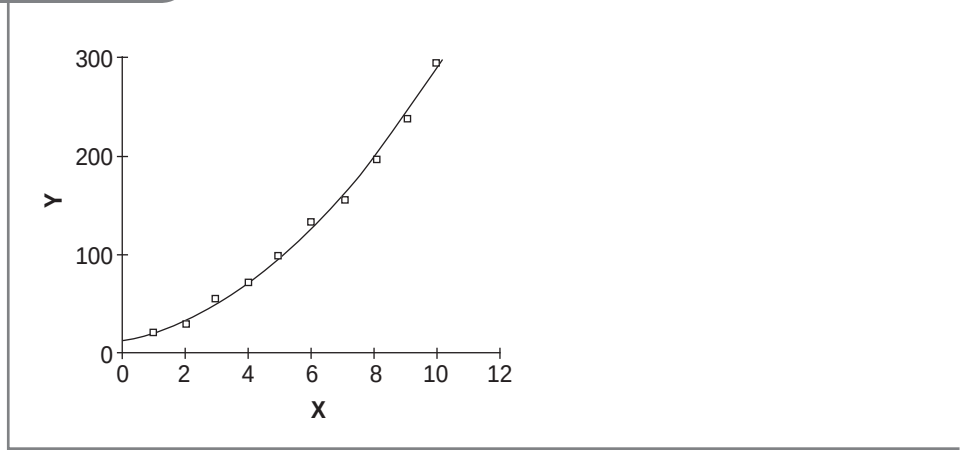
X	1	2	3	4	5	6	7	8	9	10
Y	20.60	30.80	55.00	71.40	97.30	131.80	156.3	197.3	238.70	291.70

El diagrama de dispersión de los puntos sugiere un polinomio de segundo orden como modelo:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Escribiendo $x_1 = x$ y $x_2 = x^2$, se tiene un modelo de regresión lineal múltiple.

FIGURA 9.11 *Ajuste no lineal*



A partir de los datos se tienen los siguientes resultados.

TABLA 9.13 *Índice de ajuste*

Modelo	***R***	***R cuadrado***	***R cuadrado corregida***	***Error típ. de la estimación***
1	0.999	0.999	0.998	3.69652

La curva ajusta casi perfectamente a los datos ($R^2 = 0.999$).

TABLA 9.14 *Anova*

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	415.201	3	138.400	90.856	.000[a]
	Residual	16.756	11	1.523		
	Total	431.958	14			

Los resultados que aparecen en la tabla Anova indican que el modelo es adecuado.

TABLA 9.15 *Estimación de los coeficientes*

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	12.643	4.348		2.908	0.023
	X	6.297	1.816	0.209	3.468	0.010
	X^2	2.125	0.161	0.795	13.209	0.000

Los estimadores de los coeficientes son:

$\beta_0 = 12.643$, $\beta_1 = 6.297$ y $\beta_2 = 2.125$

Los tres coeficientes son significativamente diferentes de 0.

El modelo estimado es $y = 12.643 + 6.297x + 2.125x^2$.

Modelos de regresión con variables independientes cualitativas

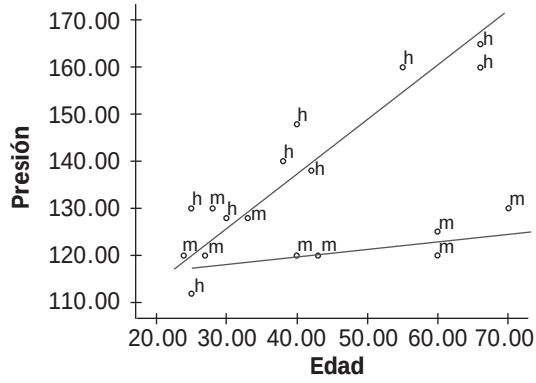
Cuando las variables independientes son cualitativas o categóricas, como sucede en muchas de las investigaciones sociales, y se desea incorporarlas en el análisis de regresión, es necesario introducir las variables llamadas *variables mudas o ficticias*, como en el siguiente caso.

EJEMPLO. *La presión sanguínea, la edad y el sexo*

Los datos siguientes corresponden a la presión sanguínea (Y), a la edad (X_1) y al sexo de las 18 personas que conforman la muestra. ¿Cómo influye la edad y el sexo de una persona en su presión sanguínea?

	Edad (X_1)	*Presión (Y)*	*Sexo*	X_2
1	25.00	112.00	H	0.00
2	25.00	130.00	H	0.00
3	42.00	138.00	H	0.00
4	55.00	160.00	H	0.00
5	30.00	128.00	H	0.00
6	40.00	148.00	H	0.00
7	66.00	165.00	H	0.00
8	60.00	160.00	H	0.00
9	38.00	140.00	H	0.00
10	28.00	130.00	M	1.00
11	24.00	120.00	M	1.00
12	60.00	120.00	M	1.00
13	33.00	128.00	M	1.00
14	40.00	120.00	M	1.00
15	70.00	130.00	M	1.00
16	43.00	120.00	M	1.00
17	60.00	125.00	M	1.00
18	27.00	120.00	M	1.00

FIGURA 9.12 *Presión, edad y sexo*



Codificando la variable "sexo" se tiene la variable $x_2 = \begin{cases} 1 \text{ si mujer } (m) \\ 0 \text{ si hombre } (h) \end{cases}$

Esta variable obtenida a partir de las dos categorías de la variable "sexo" se llama *variable ficticia o muda*. En el diagrama de dispersión de las variables “edad” y “presión” se observa que se puede ajustar dos rectas al conjunto de puntos del diagrama. Una para los hombres y otra para las mujeres.

Comenzando por ajustar un modelo de la forma $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, resulta que $R = 0.797$, $R^2 = 0.635$. Los coeficientes del modelo aparecen en la Tabla 9.16.

TABLA 9.16 *Estimación de los coeficientes*

	B	*s(B)*	*t*	*sig*
(Constante)	120.276	7.516	16.003	0.000
X_1	0.521	0.159	3.284	0.005
X_2	–18.898	4.770	–3.962	0.001

El modelo estimado es $y = 120.276 + 0.521x_1 + (-18.898)x_2$.

Observamos en el gráfico de dispersión que las rectas de ajuste para hombres y para mujeres no crecen de igual manera. En el caso de los hombres, la recta de ajuste crece más rápido que para el caso de las mujeres. Sin embargo, el modelo estimado no refleja esta situación, pues, según este modelo, para hombres y mujeres las rectas de ajuste tienen la misma pendiente, como puede observarse a continuación.

Para los hombres: $y = 120.276 + 0.521x_1$ (pendiente = 0.521).

Para las mujeres: $y = (120.276 - 18.898) + 0.521x_1$ (pendiente = 0.521).

Un modelo que mejor refleja la situación es el siguiente:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

El término x_1x_2 que ahora se agrega se llama término de *interacción*.

Para este modelo, $R = 0.954$ y $R^2 = 0.893$. Los estimadores de los coeficientes aparecen en la Tabla 9.17.

TABLA 9.17 *Estimación de los coeficientes*

	B	*s(B)*	*t*	*sig*
(Constante)	95.860	7.402	15.130	0.000
x_1	1.098	0.160	4.467	0.001
x_2	25.299	53.234	2.038	0.061
x_1x_2	–1.039	0.429	–2.401	0.031

Se observa ahora que el producto x_1x_2, considerado como variable independiente, es significativo, ayuda a explicar la variabilidad de la presión. El estimador de su coeficiente es $\hat{\beta}_3 = -1.031$. El modelo estimado es $y = 95.860 + 1.098x_1 + 25.299x_2 - 1.039x_1x_2$.

A partir de este modelo se obtiene que:

Para los hombres ($x_2 = 0$), la recta de ajuste tiene pendiente 1.098.

Para las mujeres ($x_2 = 1$), la recta de ajuste tiene pendiente 1.098 – 1.039 = 0.059.

Como resultado se tiene que el modelo con interacción es el que mejor ajusta a los datos.

Por lo general, es preferible comenzar con la estimación del modelo de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

estimarlo y luego probar la hipótesis nula $H_0 : \beta_3 = 0$. Si esta hipótesis no se rechaza, simplemente estimamos el modelo $E(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

EJEMPLO. *Ventas en tres regiones*

En general, cuando una de las variables independientes cualitativa tiene *k* categorías es necesario crear *k* – 1 variables ficticias.

Usando datos recolectados, una empresa de ventas de pizzas desea llevar a cabo un estudio para predecir las ventas (*Y*), en función de la variable *X,* que indica la región en donde se realiza la venta. La empresa tiene representaciones de venta en tres regiones: A, B, C.

A partir de la variable *X* se construyen dos variables ficticias:

$$x_1 = \begin{cases} 1 \text{ si } B \\ 0 \text{ si no} \end{cases} \qquad x_2 = \begin{cases} 1 \text{ si } C \\ 0 \text{ si no} \end{cases}$$

De este modo, si se tuvieran tres casos que pertenecen a las categorías B, A, C, respectivamente, las variables mudas tendrían los valores que se indican en la Tabla 9.18.

TABLA 9.18 *Valores de las variables mudas*

Casos	*Variable categórica*	x_1	x_2
1	B	1	0
2	A	0	0
3	C	0	1

Usando las variables mudas se tiene el modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

cuyos coeficientes será preciso estimar. Estimados los coeficientes y reemplazando los valores de las variables x_1, x_2 se tendrán los valores esperados de las ventas, para cada una de las regiones:

$y_A = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$

$y_B = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$

$y_C = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$

β_0 representa el valor esperado de las ventas en la región A, $\beta_0 + \beta_1$ representa el valor esperado de las ventas en la región B y $\beta_0 + \beta_2$ representa el valor esperado de las ventas en la región C. La región A resulta ser una categoría base, respecto de la cual se realizan las comparaciones de las ventas de las otras regiones.

LA ESTADÍSTICA EN LA EMPRESA

El banco Nacbank

El Banco Nacional del País (NACBANK) es una de las instituciones financieras de mayor antigüedad, que opera en el mercado nacional desde 1897. Se trata de uno de los principales bancos comerciales en el país, y trabaja con 5,000 millones de dólares de activos. El principal accionista de este banco es el grupo familiar Barry.

El Nacbank cuenta con una red de 120 oficinas en todo el país y 1,000 cajeros automáticos. Tiene 2,500,000 clientes activos, es el primer banco en saldos de tarjetas de crédito en el sistema financiero y ha entregado 800,000 tarjetas de crédito. En el rubro de crédito de consumo, el banco ha otorgado el 30% del crédito total brindado por el sistema bancario; sin embargo, posee el índice más bajo de morosidad, al haber desarrollado una serie de modelos de *scoring* que le permiten la elección de los mejores candidatos para recibir préstamos. Aparte de ello, la preocupación por conocer su mercado y a sus clientes ha sido una constante, lo que ha facilitado la elevación de las ventas de los diversos productos financieros que ha puesto a disposición del público. La obtención del modelo, así como el conocimiento de sus clientes y el entorno en donde se desempeña, ha sido posible gracias a la estadística.

EJERCICIOS

1. En la siguiente tabla se observa el número de determinados artículos demandados (X), así como su precio (Y), en dólares, en 10 ciudades diferentes.

X	38	62	73	85	100	122	50	70	98	123
Y	50	40	35	30	20	15	50	36	25	20

   a) Hacer el gráfico de dispersión.
   b) Ajustar un modelo de regresión lineal entre los artículos demandados y su precio.
   c) ¿Cuál es el precio que se espera cuando se demandan 70 artículos?

2. En el siguiente conjunto de datos se presentan las estaturas (X) de 10 ingenieros que se graduaron en la misma escuela y con puntuaciones parecidas. Asimismo se presentan los sueldos anuales (Y), en miles de dólares, que estos profesionales tienen en la actualidad.

X	64	65	66	69	70	72	72	74	74	75
Y	91	94	88	103	77	96	105	88	102	90

   a) Hacer el gráfico de dispersión.
   b) ¿Se puede decir que entre la estatura y el sueldo existe una relación lineal?
   c) ¿Se puede deducir de los resultados que la estatura es la causa del aumento o disminución de los sueldos?

3. Suponer que se trata de relacionar la satisfacción laboral, medida con una escala cuyos valores están entre 0 y 10, con el tiempo de servicio, medido en años. Los datos obtenidos fueron los siguientes:

Tiempo	8	4	12	9	16	14	10	15	22
Sat. laboral	5.6	6.3	6.8	6.7	7.0	7.7	7.0	8.0	7.8

   a) Estimar el modelo de regresión lineal entre el tiempo de servicios y la satisfacción laboral.
   b) Calcular el coeficiente de determinación y explicar su significado en el contexto de este problema.

4. Para analizar cómo los costos de los materiales (X) están relacionados con los costos de mano de obra (Y) en la construcción se recogieron datos correspondientes a ocho construcciones; estos se indican en la siguiente tabla.

Mano de obra: Y	***Costo de materiales: X***
293.20	270.10
283.10	265.12
285.10	260.15
280.10	270.10
293.50	280.20
278.10	260.20
310.15	290.20
305.50	285.12

a) Encontrar el índice de correlación lineal entre los costos de mano de obra y los costos de los materiales.

b) Estimar el modelo de regresión lineal que ajusta a los datos.

5. Una muestra aleatoria correspondiente a ocho meses de los precios del consumidor (IPC) y el índice industrial Dow-Jones (PDJ) fue recolectada y se consignó en la siguiente tabla:

PDJ: Y	*IPC: X*
560	13.00
600	14.50
610	12.70
580	14.50
700	12.50
550	14.90
510	15.00
700	12.80

a) Modelar la relación que pueda existir entre el IPC y el PDJ. Estimar el modelo.

b) ¿Indican los datos evidencia suficiente como para indicar que el IPC contribuye en la predicción del PDJ?

6. Un método para calibrar un instrumento de medición consiste en construir la recta de regresión entre las mediciones de este con un instrumento "patrón". Se determinaron 9 mediciones con el instrumento en mención y 10 mediciones correspondientes con el patrón:

Patrón: Y	*Instrumento: X*
19.00	19.02
15.70	15.00
18.10	18.50
25.30	25.00
20.20	21.00
24.80	25.00
18.10	17.90
21.90	22.02
27.10	28.00

a) Ajustar a los datos una recta de regresión de mínimos cuadrados. Comentar.

b) Hallar el índice de determinación e interpretar el resultado. Comentar.

7. Los datos de la tabla presentan las millas por galón que recorre un automóvil de prueba empleando gasolina de diversos niveles de octanos.

Rendimiento por galón: Y	*Número de octanos: X*
43.00	89
43.20	93
43.00	87
44.00	90
43.30	89
44.10	95
44.50	100
44.00	98

   a) Determinar el modelo de regresión lineal que ajuste a los datos. ¿Qué tan bueno es el ajuste?
   b) Hallar la suma de los cuadrados de los errores.
   c) ¿Sugieren los valores hallados en b) que el modelo ajusta bien a los datos?
   d) ¿Tiene alguna utilidad el modelo para predecir los valores de *Y*?
   e) Estimar puntualmente y por intervalo el rendimiento por galón de un automóvil cuando se usa un número de octanos igual a 92. Usar 95% de nivel de confianza.

8. En la siguiente tabla se tiene la información relativa a la cantidad de periódicos *Y*, en miles, que se venden en 15 ciudades de un país y el número de familias *X* que estas tienen.

Ciudad	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Número de periódicos	6.1	26.9	25.6	65.2	40.2	70.0	98.0	42.1	29.4	33.0	28.8	65.3	47.0	34.0	16.9
Número de familias	8.6	68.7	98.8	23.0	69.2	13.7	74.5	67.0	74.2	69.3	74.0	69.3	74.0	74.2	83.3

   a) Hacer el diagrama de dispersión.
   b) Indicar si un modelo de regresión lineal es adecuado para los datos de la tabla. Agregar un análisis de los residuales.
   c) Aplicar una transformación logarítmica a la variable que indica el número de periódicos vendidos y luego aplicar un modelo de regresión lineal a las variables ln(*Y*) y *X*. Analizar este modelo. Comparar este modelo con el modelo obtenido en b).
   d) Estimar el promedio de periódicos que se venden en las ciudades que tienen 50,000 familias.

9. La empresa Arroyo selecciona a sus empleados mediante una prueba de aptitud *X*. Un año después le aplica a los mismos empleados una prueba de evaluación de su desempeño, *Y*. Los resultados que se obtuvieron y que corresponden a una muestra de 15 empleados aparecen en la siguiente tabla.

Prueba de aptitud	*Prueba de desempeño*
83	76
76	75
92	81
75	86
76	68
82	88
69	62
87	79
75	79
88	79
76	80
68	76
88	80
77	79
80	76

a) Hacer el gráfico de dispersión. Comentar el resultado.

b) Hallar el índice de determinación. Indicar si un modelo de regresión lineal simple es adecuado para ajustar los valores de la muestra.

c) Estimar la recta de regresión lineal de *Y* en *X*.

d) Interpretar la pendiente de la recta de regresión.

e) Indicar si el modelo estimado es adecuado para la predicción.

f) Hallar el intervalo de confianza al 95% del puntaje de la prueba de desempeño que obtendría un empleado que al ingresar obtuvo un puntaje de 78 en la prueba de aptitud.

10. a) Ajustar un modelo de regresión lineal al siguiente conjunto de datos considerando que *Y* es la variable dependiente y las otras, las variables independientes.

X_1	X_2	*Y*
8.10	1.68	51.53
11.0	1.64	73.75
4.60	1.58	26.38
10.00	1.21	55.54
3.30	1.89	25.52
5.60	1.00	38.55
1.20	1.37	15.65
6.40	1.11	35.02
9.20	1.87	62.49
8.10	1.00	48.09
5.70	1.76	40.76
6.40	1.87	49.69
2.70	1.52	27.59
2.90	1.31	23.22
10.20	1.19	60.98

b) Estudiar la validez del modelo para predecir los valores de *Y* cuando se conocen los valores de las variables independientes.

c) Contrastar la hipótesis nula, relativa coeficiente de la variable x_2, $H_0 : \beta_2 = 0$ vs. la hipótesis $H_0 : \beta_2 \neq 0$.

d) Estimar puntualmente al promedio de la variable dependiente cuando $x_1 = 7$ y $x_2 = 2$.

11. Los datos muestrales que siguen corresponden a los pesos X_1 de ocho equipos de cocina que fueron trasladados una distancia X_2, en km, y a un costo *Y*, en dólares:

X_1	X_2	Y
4000	3.0	200
3000	5.0	150
1600	2.0	100
1200	4.0	130
3400	1.6	160
4800	3.2	220
3600	2.2	200
2000	3.0	320

a) Ajustar un modelo de regresión lineal tomando como variable dependiente al costo.

b) ¿Es adecuado el modelo para predecir el costo?

c) Analizar la significancia de cada uno de los coeficientes de las variables independientes.

d) Determinar el valor que predice el costo cuando las variables X_1 y X_2 tienen los valores 1,300 y 3.5, respectivamente.

12. Se cree que el ingreso, X_1, de los empleados y la antigüedad en el trabajo, X_2, influyen en la satisfacción laboral, *Y,* evaluada en puntajes del 1 al 10. Para ello se recabó la información correspondiente a 10 empleados. Los resultados fueron como sigue:

Ingreso	*Antigüedad*	*Satisfacción laboral*
2700	8	6.50
2200	4	6.30
3400	12	6.80
2800	9	6.70
3600	16	7.10
3900	14	7.09
3300	10	8.12
4200	15	7.85
4600	22	8.00
3300	10	6.00

a) Ajustar un modelo de regresión lineal a los datos, tomando como variable dependiente a la satisfacción laboral.

b) ¿Es adecuado el modelo para predecir la satisfacción laboral cuando se conoce el ingreso y la antigüedad?

c) Usando el nivel de significación 0.05, decir si el coeficiente individual de las variables independientes es significativo.

13. Con la finalidad de planear el ingreso de los estudiantes a una universidad se desea predecir el puntaje obtenido en el examen de ingreso en términos de los puntajes en los exámenes de los cursos de lenguaje y de matemáticas. Para ello se recogió la siguiente información.

Lenguaje	*Matem.*	*Ingreso*	*Lenguaje*	*Matem.*	*Ingreso*
16.20	17.40	17.50	12.80	15.20	13.50
13.60	19.80	14.50	13.20	11.80	11.00
11.40	17.20	13.50	16.00	13.00	16.50
20.00	9.80	7.75	20.00	17.00	19.00
10.80	16.60	13.00	16.60	15.20	13.50
16.40	17.20	17.00	12.80	13.20	16.00
15.00	14.80	18.00	16.60	14.40	11.50
11.60	19.60	14.50	18.60	10.80	14.50
11.00	10.80	7.50	14.80	11.80	12.50
9.80	16.20	10.50	10.20	15.00	14.00

a) Elegir el modelo de regresión lineal múltiple más adecuado y estimarlo.

b) Indicar el estimador del promedio del examen de ingreso cuando los puntajes de lenguaje y matemáticas son 15, 15, respectivamente.

14. Se llevó a cabo un plan de aprendizaje práctico individual para posteriormente realizar una determinada tarea. Se desea saber si el tiempo dedicado al aprendizaje influye en el tiempo utilizado para terminar la tarea. Los datos recolectados fueron como sigue.

Tiempo de aprendizaje: x (días)	*Tiempo para terminar una tarea: y (minutos)*
24	10
10	20
15	15
17	11
6	11
20	19
9	11
3	13
10	17
7	18
9	16
7	16
5	17
20	20

a) Hacer un diagrama de dispersión. ¿Sugiere el gráfico un modelo que ajuste a los datos?

b) Ajustar un modelo de la forma $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$.

c) ¿Contribuye significativamente el término cuadrático en el modelo? Si este término no contribuye, ¿es el modelo $y = \beta_0 + \beta_1 x + \varepsilon$ adecuado?

d) Graficar los residuales. ¿Respalda esta gráfica las asunciones del modelo?

15. Se llevó a cabo un estudio para determinar los efectos del tamaño de la empresa, medido por el número (X_1) de empleados y de la aplicación o no de un programa educacional (X_2) sobre el número de horas perdidas (Y), durante un determinado periodo. Los datos recogidos fueron los siguientes.

x_1	$x_2 = 0$	y	x_1	$x_2 = 1$	y
600	0	120	307	1	36
700	0	170	660	1	44
650	0	170	270	1	73
313	0	120	700	1	8
870	0	50	830	1	90
210	0	170	241	1	71
410	0	30	960	1	37
325	0	95	930	1	111
880	0	70	680	1	89
200	0	170	480	1	72
900	0	23	960	1	35
950	0	170	290	1	86
950	0	170	630	1	40
270	0	60	190	1	44
844	0	140	850	1	36
630	0	120	475	1	78
230	0	40	605	1	47
790	0	156	705	1	56
700	0	110	780	1	75
552	0	120	170	1	46

a) Plantear el modelo adecuado para este estudio, estimar los coeficientes y hacer las interpretaciones que correspondan.

b) Probar la utilidad del modelo.

c) Probar la significancia de cada uno de los coeficientes de las variables independientes.

d) Graficar los residuales. ¿Respalda esta grafica las asunciones del modelo?

e) Usar un modelo de regresión lineal para estudiar si las variables independientes interaccionan.

16. El administrador de un banco cree que sus clientes de mayor edad tienden a ahorrar más. Se consideró una muestra de 100 empleados y se ajustó un modelo de la forma $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ para relacionar los ahorros con la edad. Usando el modelo indicado, decir si la suposición del administrador concuerda con los datos si el estimador del coeficiente del cuadrado de la edad fue 0.0015 y su error estándar, 0.00712. Use $\alpha = 0.05$.

RESPUESTAS A LOS EJERCICIOS

1. b) $y = 66.340 - 0.417x$. $R = 0.974$. La significación o valor p del coeficiente de x es 0.000. c) 37.15 **2.** a) Observando el gráfico y el índice de correlación entre X e Y no se puede decir que existe una relación entre estas variables. $y = 70.050 + 0.333x$. $R = 0.156$. La significación o valor p del coeficiente de x es 0.667. b) Aun cuando existiera una relación lineal entre X e Y no se puede decir que la causa del aumento o disminución de los sueldos es la estatura. **3.** $y = 5.636 + 0.111x$. $R = 0.759$. La significación o valor p del coeficiente de x es 0.018. **4.** a) el índice de correlación es 0.911 b) $y = 33.466 + 0.945x$. La significación del coeficiente de x es 0.002. **5.** a) $y = 1251.73 - 47.351x$. $R = 0.747$. La significación o valor p del coeficiente de x es 0.033. **6.** a) $y = 1.33 + 0.931x$. $R = 0.994$. La significación o valor p del coeficiente de x es 0.000 b) el índice de determinación es igual a 0.988. Los resultados indican que al parecer el instrumento no está muy bien calibrado. La recta de regresión debería ser aproximadamente la recta $y = x$. **7.** a) $y = 34.360 + 0.1x$. $R = 0.812$. La significación o valor p del coeficiente de x es 0.014. **8.** b) $y = 4943 - 0.120x$. $R = 0.131$. La significación o valor p del coeficiente de x es 0.641. **9.** c) $y = 41.375 + 0.359x$. $R = 0.394$. La significación o valor p del coeficiente de x es 0.078 0.253 resulta el estimador de *b.* **10.** a) $y = -4.791 + 5.0409X_1 + 8.665X_2$ b) el modelo es adecuado a nivel de muestra, $R^2 = 0.951$. El modelo también es adecuado a nivel de población, sirve para la predicción c) el coeficiente de X_2 es significativamente diferente de 0 d) 47.82. **11.** a) $y = 139.62 + 0.015X_1 + 0.679X_2$ b) el modelo no es adecuado. $R^2 = 0.198$ (el valor p para probar la significancia del modelo es igual a 0.827). **12.** a) $y = 4.492 + 0.001X_1 + (-0.015) X_2$. El modelo no es adecuado. $R^2 = 0.695$ (el valor p para probar la significancia del modelo es igual a 0.100). **13.** a) $y = -2.334 + 0.428X_1 + 0.672X_2$. El modelo es adecuado. b) 14.1646.

CAPÍTULO 10

Análisis de la varianza

Ronald Fisher

Ronald Fisher nació en 1890, en Londres, Inglaterra. Se graduó en Cambridge en 1912, en la especialidad de Matemáticas y Física.

Después de enseñar en una escuela secundaria, en 1919 comenzó a trabajar como estadístico en la estación experimental de Harpenden, donde organizó la información sobre el tiempo acumulada durante muchos años. Fue considerado un líder de la estadística y genetista de primer nivel.

Fisher escribió un gran número de artículos de investigación e hizo muchas contribuciones al campo de los diseños experimentales, al análisis de la varianza, al análisis de pequeñas muestras, etcétera.

Fisher fue profesor de Cambridge y de University Collage, en Londres.

En 1952 le concedieron el título de Sir.

Murió en Australia, en 1962.

CONTENIDO

10.1 Introducción

Se han presentado métodos para comparar las medias de los valores de una variable que resultan al aplicar dos tratamientos diferentes; sin embargo, son muchas las ocasiones en las que es necesario comparar más de dos medias, por ejemplo, cuando se requiere comparar el efecto en las ventas de un producto de tres tipos de publicidad realizados en la TV, en el periódico y en la radio. En este caso se busca analizar de qué manera la variable independiente o *factor propaganda* impacta en la variable "ventas". Al factor "propaganda" le corresponden tres métodos o niveles que en este contexto son los *tratamientos*. La experiencia que permita el análisis deberá prepararse de tal manera que otros factores diferentes al factor "propaganda" no incidan en las ventas.

10.2 Análisis de la varianza de un solo factor

Con esta metodología se analiza si los niveles o tratamientos de un factor determinado influyen de manera diferente en los resultados que se obtienen al realizar una experiencia en donde no existe influencia de otros factores, excepto el que se estudia. La experiencia debe ser diseñada de tal manera que se reduzca al máximo la influencia de los factores que no son de interés en el análisis. Al diseño de estas experiencias, en donde se permite al investigador la manipulación de los diferentes niveles del factor, con la finalidad de controlar la influencia de otros factores, se le llama *diseño experimental*, y la metodología que se usa para el análisis se le llama *análisis de varianza*.

Para cada tratamiento, aun cuando las unidades en donde se aplica la experiencia son tratadas de "manera idéntica e independiente", por razones que tienen que ver con su naturaleza, con la imposibilidad de reproducir las mismas condiciones u otras circunstancias, se produce una variabilidad entre las mediciones realizadas. A la variabilidad que se produce dentro de cada tratamiento se le llama *error experimental*. A menudo se usa la distribución normal para modelar esta variabilidad.

Llevar a cabo un diseño de este tipo se asemeja a la recepción de una señal acústica. Al recibir la señal interesa su volumen y la cantidad de ruido que interfiere con ella. En un diseño experimental, el ruido corresponde al error experimental, y de lo que se trata es de planear la experiencia de tal modo que disminuya el ruido para que así la señal (los valores de la variable dependiente) tenga el mayor volumen posible y se pueda captar mejor.

El análisis de los resultados de un diseño experimental mediante el análisis de la varianza, en buena cuenta, se reduce a la descomposición de la variabilidad de los valores de la variable dependiente en la variabilidad que produce el factor en estudio

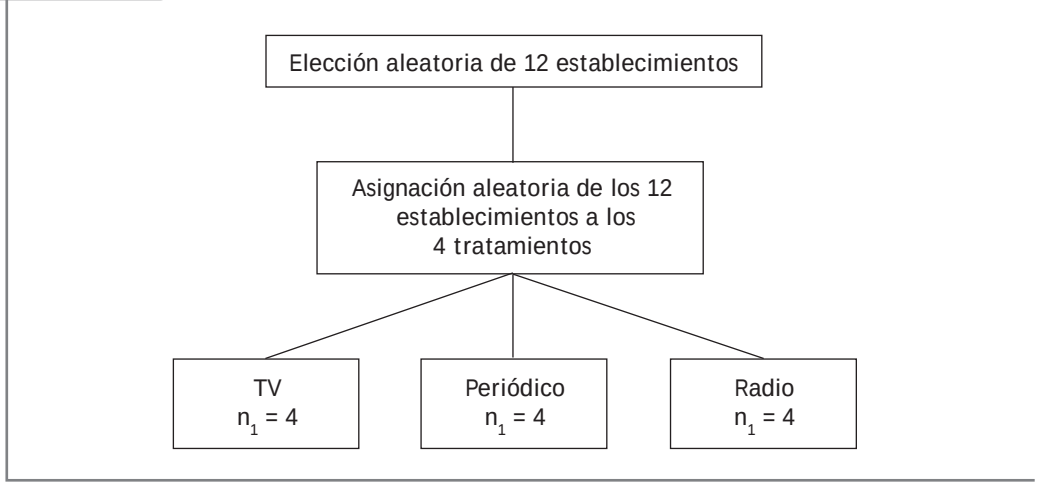
en los diferentes niveles y el error experimental. Si la variabilidad que producen los niveles o tratamientos es significativa respecto del error experimental, es razonable indicar que los tratamientos influyen de manera diferente en los resultados.

Para analizar si las medias de las ventas, correspondientes a los tipos de propaganda, son diferentes o no; es decir, para analizar si el tipo de propaganda influye en las ventas, se han elegido al azar 12 establecimientos (*unidades estadísticas*) que venden el mismo producto y que no existen otros factores (distintos a los tipos de propaganda) que pudieran influir en las ventas. Sin embargo, para neutralizar la influencia de algún factor no identificado en las ventas, se asignó de manera *aleatoria* a 4 de los establecimientos la propaganda del producto por televisión, a otros 4 la propaganda por periódico y al resto la propaganda por radio (los números de unidades asignados a cada tratamiento pueden ser diferentes, lo que importa es la aleatoriedad).

Las ventas obtenidas, en unidades monetarias, fueron como sigue.

Ventas		
TV	***Periódico***	***Radio***
26,000.00	32,000.00	26,000.00
25,000.00	31,000.00	25,000.00
31,000.00	35,000.00	27,000.00
26,000.00	34,000.00	30,000.00

FIGURA 10.1 *Diseño*



Las medias y la desviación estándar de las ventas para cada tratamiento, así como los gráficos de puntos y cajas para las ventas, aparecen a continuación (Tabla 10.1 y Figura 10.2).

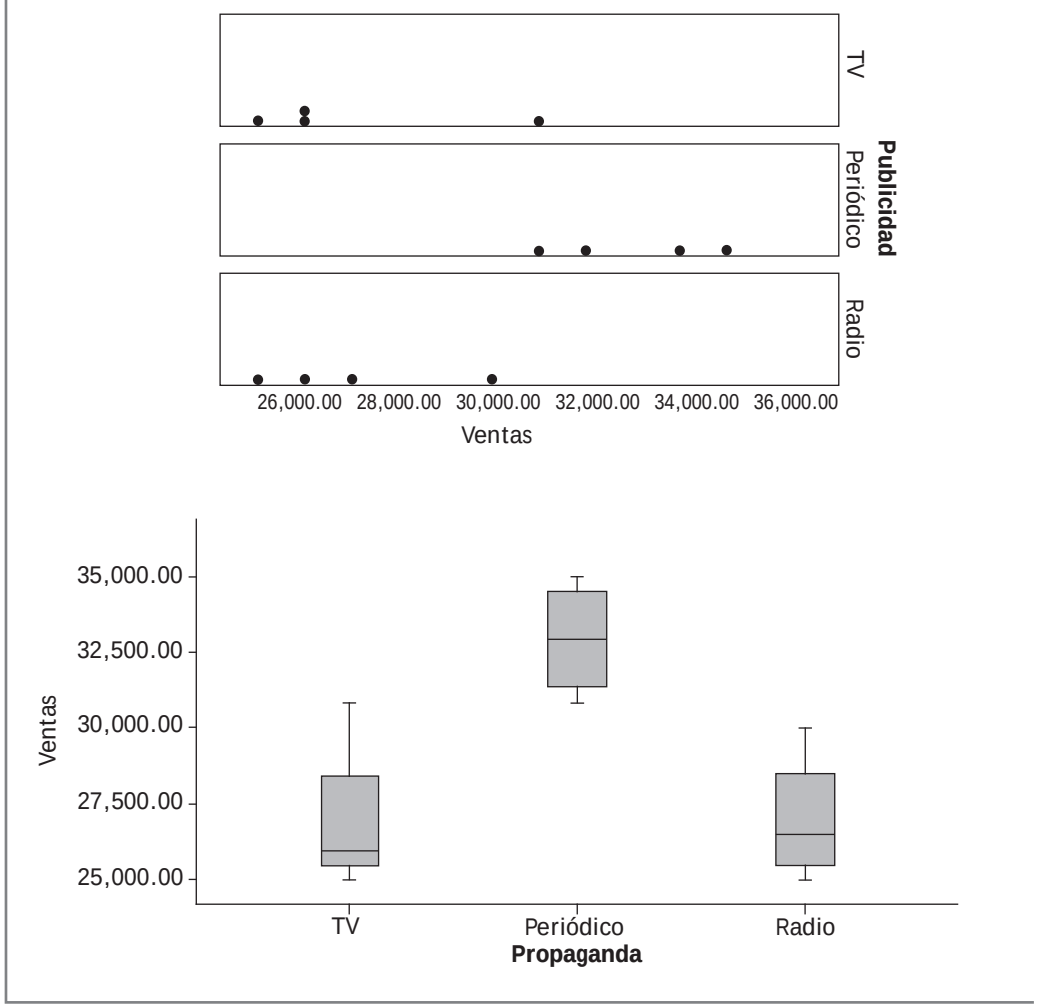
TABLA 10.1 *Resumen numérico*

Ventas

	N	Media	Desviación típica
1	4	27,000.00	2708.01280
2	4	33,000.00	1825.74186
3	4	27,000.00	2160.24690
Total	12	29,000.00	3592.92234

1: TV, 2: Periódico 3: Radio

FIGURA 10.2 *Gráfico de puntos y de cajas*



Al observar los resultados estaríamos tentados a decir que los diversos tipos de propaganda influyen en las ventas; sin embargo, debemos analizar con mayor detalle estos resultados, pues podrían ser producto del azar.

Si suponemos que las medias de las ventas son iguales, cualquiera que sea el tipo de publicidad utilizada, podremos escribir el siguiente modelo:

$$y_{ij} = U + \varepsilon_{ij}$$

en donde:

y_{ij} es la *i-ésim*a venta en el *j-ésimo* establecimiento.

U es la media total que recoge el efecto de todos los factores constantes.

ε_{ij} es el efecto residual aleatorio (error experimental) con media igual a 0 y varianza constante.

Sin embargo, si la idea es realizar la experiencia para analizar si los diferentes tipos de publicidad influyen en las ventas resultantes debemos usar un modelo que considere esta particularidad. Un modelo adecuado es el modelo llamado *modelo de una sola vía o de un solo factor*, que indica que cada observación y_{ij} puede escribirse como:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},\ i = 1, 2, 3;\ j = 1, 2, 3, 4$$

en donde:

μ = media constante o media total que resume el efecto de todos los factores constantes diferentes al factor publicidad.

τ_1= es el efecto que se obtiene al hacer publicidad por TV.

τ_2= es el efecto que se obtiene al hacer publicidad por periódico.

τ_3= es el efecto que se obtiene al hacer publicidad por radio.

$\mu_i = \mu + \tau_i$ es la media de las ventas para el tratamiento (tipo de propaganda) *i*.

Sobre los efectos de los niveles del factor, se asume que $\sum_{i=1}^{3} \tau_i = 0$.

ε_{ij} = es el error aleatorio con valor esperado igual a 0 y con varianza constante igual a σ^2.

Si para cada tratamiento *i*, las observaciones y_{ij} provienen de poblaciones normales con la misma varianza y todas ellas independientes, es posible llevar a cabo pruebas respecto a la igualdad de las medias de las ventas para los diferentes tipos de publicidad.

Se puede probar la hipótesis nula $H_0 : \mu_1 = \mu_2 = \mu_3$ versus la hipótesis alternativa H_A: *la hipótesis nula es falsa* (al menos dos medias son diferentes).

La hipótesis nula equivale a suponer que el efecto de los tratamientos es igual a 0. Por ello la hipótesis nula podría escribirse en términos del efecto de los diferentes tipos de publicidad:

$$H_0: \tau_1 = \tau_2 = \tau_3 = 0$$

El método para la prueba de hipótesis es el *análisis de la varianza*. Este procedimiento, ideado por Ronald A. Fisher, consiste en descomponer la variabilidad total de las observaciones $SCT = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\bullet\bullet})^2$ en dos partes: una que corresponda a la variabilidad que se produce al cambiar los tratamientos, que se mide con $SCTr = \sum_{i=1}^{a} n_i(\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2$, y la otra que corresponda al error experimental (el ruido), que se mide con $SCE = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{\bullet})^2$:

$$SCT = SCTr + SCE$$

Las notaciones corresponden a:

$$\sum_{j=1}^{n} y_{ij} = y_{i\bullet} \text{ o } 1/n \sum_{j=1}^{n} y_{ij} = y_{i\bullet}/n = \bar{y}_{i\bullet} \text{ o } 1/mn \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij} = \bar{y}_{\bullet\bullet}$$

Cada una de las tres sumas de cuadrados tiene distribución ji-cuadrado, *SCT* con $N - 1$ grados de libertad, *SCTr* con $a - 1$ y *SCE* con $N - a$.

Un valor de la suma *SCTr* significativamente mayor que la suma *SCE* constituye evidencia en contra de la hipótesis nula que indica que las medias son iguales. Para analizar si esto ocurre se comparan las sumas mediante el estadístico de prueba:

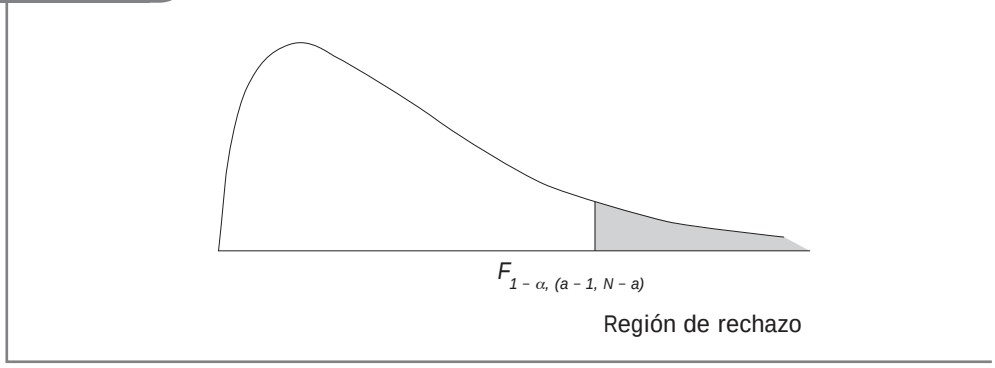
$$F_0 = \frac{SCTr/(a-1)}{SCE/(N-a)}$$

Si la hipótesis nula es verdadera y se cumplen los supuestos del modelo, este estadístico corresponde a una distribución *F* con $a - 1$ grados de libertad para el numerador y $N - a$ grados de libertad para el denominador.

Al nivel de significación α, para contrastar la hipótesis nula frente a la alternativa, la regla de decisión es la siguiente:

Rechazar la hipótesis nula si F_0, el valor del estadístico de prueba, es mayor que el cuantil de orden $1 - \alpha$ de la distribución *F* con $a - 1$ grados de libertad para el numerador y $N - a$ grados de libertad para el denominador. A este cuantil se le denota con $F_{1-\alpha,\,(a-1,\,N-a)}$.

FIGURA 10.3 *Distribución F. Región de rechazo*



La hipótesis nula se rechaza, al nivel de significación α, si el valor del estadístico de prueba es mayor que el cuantil $F_{1-\alpha,\,(a-1,\,N-a)}$.

Los resultados del análisis se suelen escribir en una tabla llamada *tabla Anova* (análisis de la varianza).

TABLA 10.2 *Tabla ANOVA*

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Media cuadrática*	*F*
Entre tratamientos o intergrupos	*SCTr*	$a-1$	*MCTr*	$F_0 = MCTr/MCE$
Dentro de los tratamientos o intragrupos	*SCE*	$N-a$	*MCE*	
Total	*SCT*	$N-1$	*MCT*	

En la tabla se indican las sumas de cuadrados de la descomposición de la varianza, los grados de libertad de cada suma de cuadrados, la media cuadrática (cociente de la suma de cuadrados por sus grados de libertad) para cada fuente de variabilidad y el valor del estadístico de prueba F_0.

TABLA 10.3 *Tabla ANOVA*

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Media cuadrática*	*F*	*Sig*
Entre tratamientos o intergrupos	96000000	2	48000000	9.391	0.006
Dentro de los tratamientos o intragrupos	46000000	9	5111111.1		
Total	142000000	11			

La suma de cuadrados entre tratamientos (*SCTr*) es igual a 96000000 y el número de grados de libertad es igual a 2.

La suma de cuadrados dentro de los tratamientos (*SCE*) es igual a 46000000 y sus grados de libertad son 9.

La suma de cuadrados total (*SCT*) es 142000000.

En la columna titulada con **sig** aparece el valor – *p* de la prueba. Este valor indica la probabilidad de encontrar un valor mayor o igual al valor calculado de *F cuando la hipótesis nula es verdadera*.

La hipótesis nula se rechaza cuando el valor – *p* es "pequeño". Si se indica el nivel de significación α de la prueba, la hipótesis nula se rechaza si el valor – *p* es menor o igual que el nivel de significación.

Según los resultados, si la hipótesis nula es verdadera, la probabilidad de encontrar un valor de *F* mayor o igual que 9.391 es 0.006. Hay una probabilidad muy pequeña de encontrar un valor del estadístico como el que se ha hallado; sin embargo se ha encontrado. Lo que queda es rechazar la hipótesis nula. Es decir, los diferentes tipos de propaganda producen diferentes efectos en las ventas.

Si en la tabla no aparece el *p–value*, habrá que calcular el valor crítico de la prueba. El cuantil 0.95 de la distribución *F* con 2 grados de libertad para el numerador y 9 para el denominador es 4.26 (valor crítico de la prueba).

Como el valor del estadístico (9.391) es mayor que 4.26, se rechaza la hipótesis nula, al nivel de significación 0.05.

Observaciones

1. Las condiciones que se deberían verificar para tener confianza en los resultados del análisis de varianza son las que se refieren a los supuestos del modelo. El modelo deberá describir de manera adecuada las observaciones y los errores deberán seguir la distribución normal con media 0 y varianza constante para todos los niveles.

2. Un análisis de varianza que muestra que existe diferencia entre las medias para los diferentes tratamientos no indica cuáles son las medias diferentes. Sin embargo, es un paso previo para llevar a cabo un análisis más profundo sobre la manera como influye la variable independiente. Para estudiar el efecto de los diferentes niveles de la variable independiente se realizan *comparaciones múltiples a posteriori o no planificadas*; para ello se han desarrollado una serie de pruebas de este tipo, entre las que están los métodos de Duncan, Bonferroni, Scheffé, Tukey, de Student-Newman-Keuls (S-N-K), Dunnett, etcétera.

10.3 Diseño de bloques aleatorizados

Cuando el error experimental (el ruido) que se genera al usar un modelo de una sola vía es bastante alto, la consigna es reducir este error experimental con la finalidad de aumentar la precisión de las estimaciones. Varias son las recomendaciones al respecto; generalmente se indican las siguientes:

a) Aumentar el número de réplicas en cada muestra. De este modo, el número de grados de libertad aumenta, la media cuadrática del error experimental disminuye y por lo tanto la señal aumenta.
b) Neutralizar o eliminar cualquier "factor perturbador" que no es de interés para el experimentador y que, por el contrario, pueda incrementar el ruido.

La segunda recomendación a menudo no es posible de aplicar, sobre todo cuando las influencias son externas al laboratorio en donde se lleva a cabo la prueba. Cuando la fuente de variabilidad perturbadora es conocida y controlable, el análisis de la influencia del factor de interés puede realizarse con la *técnica del bloqueo*.

La técnica del bloqueo se basa en la formación de grupos de unidades experimentales homogéneas respecto de los valores de la variable que se considera perturbadora en la experiencia.

Cada bloque se forma antes de realizar la experiencia, con tantas observaciones como tratamientos existan, y cada uno de ellos se considera que es una réplica completa del experimento.

La idea es controlar la influencia del bloque, no medir su efecto.

EJEMPLO. *Comparando tres tipos de gasolina*

Se realiza un estudio para comparar el rendimiento de un vehículo para tres marcas de gasolina. Se seleccionaron cuatro modelos diferentes de automóviles de tamaño variable. El estudio podría realizarse usando el análisis de varianza de una sola vía; sin embargo, los resultados que se obtengan podrían estar influenciados por los diferentes tipos de automóviles. Por esta razón, y con la finalidad de controlar a esta variable "perturbadora", se forman cuatro bloques, uno por cada modelo de vehículo. En cada bloque se repite la experiencia, una vez por cada tipo de gasolina.

Los datos en millas por galón son los siguientes. La prueba para cada modelo de automóvil se realiza al azar.

TABLA 10.4 *Rendimientos de tres tipos de gasolina*

Gasolina/Modelo	1	2	3	4
1	32.4	28.8	36.5	38.7
2	35.6	28.6	37.6	29.9
3	38.7	29.9	36.2	37.9

El modelo general para tratar esta situación se llama *diseño de bloques aleatorizados* y tiene la siguiente forma:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij} = \quad i = 1, 2,..., a;\ j = 1, 2,..., b$$

en donde:

μ es la media constante que resume el efecto de todos los factores constantes.

τ_i es el efecto debido al factor en estudio. Se supone que $\sum_{i=1}^{a} \tau_i = 0$.

β_j es el efecto debido a la variable de bloqueo. Supondremos además que $\sum_{j=1}^{b} \beta_j = 0$.

ε_{ij} es el error aleatorio. Se supone que la distribución del error aleatorio es normal con media 0 y varianza constante e igual a σ^2.

Con este modelo se logra aislar el efecto del factor perturbador, y como en el caso del diseño de un solo factor, para comprobar la influencia que podrían ejercer los niveles del factor de interés, se contrasta la hipótesis nula $H_0 : \tau_1 = \tau_2 = \tau_a ... = 0$ con la hipótesis alternativa: $H_A = H_0$ no es verdad. Si se rechaza la hipótesis nula, al menos un par de tratamientos son diferentes.

Como en el caso de una vía, la prueba consiste en:

1. Descomponer la suma total de cuadrados *SCT* en las sumas de cuadrados: *SCTr*, *SCB* y *SCE*, debidas a los tratamientos, a los bloques y al error experimental, respectivamente. Los grados de libertad para estas sumas de cuadrados son: $an - 1$, $a - 1$, $b - 1$ y $an - a - b + 1$, respectivamente. Se cumple que:

$$SCT = SCTr + SCB + SCE$$

2. Descompuesta la varianza, se calcula el estadístico de prueba:

$$F_0 = \frac{SCTr/(a - 1)}{SCE/(an - a - b + 1)}$$

Este valor corresponde a una variable aleatoria que tiene distribución F con (a − 1) grados de libertad para el numerador y (*an* − *a* − *b* + 1) grados de libertad para el denominador.

3. Al nivel de significación α, la hipótesis se rechaza si el valor del estadístico F_0 es mayor que el cuantil de orden 1 − α para la distribución $F_{1-\alpha,\,(a-1,\,an-a-b+1)}$.

Los resultados del análisis de varianza (Anova) se escriben en una tabla (10.5) de la siguiente manera:

TABLA 10.5 *ANOVA para diseños de dos vías*

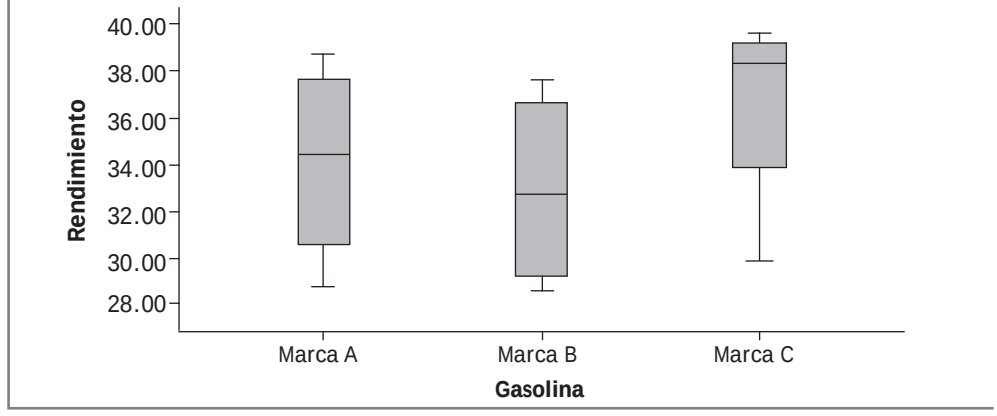
Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Media cuadrática*	*F*
Entre tratamientos	*SCTr*	$a - 1$	*MCTr*	F_0
Entre bloques	*SCB*	$b - 1$	*MCB*	
Dentro de los tratamientos	*SCE*	$an - a - b + 1$	*MCE*	
Total	*SCT*	$an - 1$	*MCT*	

En esta tabla se indican las sumas de cuadrados, sus grados de libertad correspondientes, las medias cuadráticas (sumas de cuadrados divididas entre los grados de libertad) y el valor del estadístico F_0.

Volviendo al ejemplo de las ventas, se grafican los *box plots* de los rendimientos para cada gasolina.

A nivel de muestra, el rendimiento que produce la gasolina 3 es mejor; sin embargo, será el siguiente análisis el que confirme esta presunción.

FIGURA 10.4 *Box plots para rendimiento de cada tipo de gasolina*



En efecto, los resultados que aparecen en el siguiente cuadro (Tabla 10.6) no permiten indicar que las medias de los rendimientos de los tres tipos de gasolina son significativamente diferentes. Al nivel de significación 0.05, no se rechaza la igualdad de tratamientos, pues el valor del estadístico de prueba es menor que el cuantil $F_{1-0.05}$, que según la tabla del apéndice D es 4.256.

TABLA 10.6 *Tabla ANOVA para tres tipos diferentes de gasolina*

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Media cuadrática*	*F*	*Sig*
Entre tratamientos	26.962	2	13.481	0.694	0.524
Error	174.735	9	19.415		
Total	201.697	11			

Sin embargo, los resultados obtenidos podrían estar influenciados por el tipo de automóvil; por ello, y como los datos obtenidos lo permiten, se hace posteriormente un análisis usando un modelo de bloques aleatorios. La tabla Anova que aparece a continuación (Tabla 10.7) resume los resultados. El efecto que se produce al considerar el tipo de auto se observa en la reducción del error (de 174.735 se redujo a 46.165).

TABLA 10.7 *Tabla ANOVA para bloques aleatorios*

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Media cuadrática*	*F*
Gasolina	26.962	2	13.481	0.252
Automóvil	128.570	3	42.857	0.036
Error	46.165	6		
Total	201.697	11		

Se observa que el tipo de gasolina (factor de interés) no influye en los resultados (*valor – p* = 0.252), aun cuando el tipo de auto sí afecta a los resultados (*valor – p* = 0.036).

(Al nivel de significación 0.05, el valor del estadístico de prueba para el tipo de gasolina no es mayor que el cuantil de orden $1 - \alpha$ de la distribución *F* con 2 grados de libertad para el numerador y 6 para el denominador (5.1433). La hipótesis de la igualdad de los tipos de gasolina no se rechaza.)

EJERCICIOS

1. Indicar el objetivo de los diseños experimentales.
2. Como resultado de aplicar un diseño experimental de un solo factor, se encontraron los siguientes resultados para cada tratamiento.

Tratamiento A	*Tratamiento B*	*Tratamiento C*
42	58	50
36	55	47
41	57	46
42	54	43
40	53	46

   Al nivel de significación 0.05, indicar si la influencia de los diferentes tratamientos es diferente en las mediciones.

3. Para estudiar si los tiempos de aprendizaje de tres programas computacionales de contabilidad son diferentes o no, se asignaron de manera aleatoria a 15 operadores al aprendizaje de tales programas. Los resultados, en horas, fueron como sigue.

Programa A	*Programa B*	*Programa C*
14	14	20
17	15	17
12	14	18
11	13	21
17	16	25

   Al nivel de significación 0.05, ¿se podría decir que las medias de los tiempos de aprendizaje difieren de acuerdo a los programas?

4. Una compañía textil utiliza tres telares para la producción de telas. Para probar si estos telares producen telas con igual resistencia, el ingeniero de procesos fabricó, con cada telar y con el mismo material, telas cuya resistencia fue como sigue.

Telar	*Resistencias*			
1	98	97	99	96
2	91	90	93	92
3	96	95	97	95

   Escribir el modelo adecuado para analizar si existe diferencia en la resistencia medias de las telas producidas entre los telares. Usar el nivel de significación 0.05.

5. Indicar un procedimiento para probar si las medias de los impuestos que pagan los empleados de tres sectores empresariales son iguales o diferentes.
6. Un estudio ha sido dirigido para estudiar si tres tipos diferentes de ocupación influyen de manera diferente en el estado de ansiedad de los habitantes de una región. Para ello se aplicó un test de ansiedad a 21 personas que tenían la misma edad y buenas condiciones de salud. Indicar el procedimiento para analizar los resultados.

7. Como es sabido, la fricción es la causa del desgaste de las piezas de un motor. Para evitar en parte este inconveniente se usan lubricantes de diferente tipo. En una compañía ensambladora de motores se prueban tres tipos de lubricantes A, B y C, habiéndose controlado cualquier otro factor que pudiera intervenir en el desgaste. Las siguientes mediciones corresponden a los pesos en gramos de los residuos que se obtuvieron cuando se probaron los tres tipos de lubricante.

Lubricante A	*Lubricante B*	*Lubricante C*
12.2	10.1	18.1
11.8	7.5	14.3
13.1	13.5	11.7
11.0	11.4	12.7
6.9	10.3	13.2
4.5	8.4	10.9
8.4	13.2	11.2

¿Se puede decir que el lubricante A es mejor que el lubricante B y que este es mejor que el lubricante C, al nivel de significación del 95%?

8. Para estudiar los retrasos en el tránsito de vehículos en una ciudad se usaron tres tipos de semáforos en una ciudad: S1, S2 y S3. Los semáforos se colocaron en tres intersecciones diferentes y se midió el tiempo que cada vehículo permanece detenido en cada intersección. Los tiempos en segundos por vehículo aparecen a continuación.

Semáforo	*Retraso*				
S1	36.6	39.2	30.4	37.1	34.1
S2	17.5	20.6	18.7	25.7	22.0
S3	15.0	10.4	18.9	10.5	15.2

a) Utilizando un modelo adecuado, estimar e interpretar el efecto de utilizar un semáforo de tipo S2 en el valor promedio de los retrasos. Usar el nivel de significación 0.05.

b) Al nivel de significación 0.05, ¿se puede decir que existen diferencias significativas en las medias de los retrasos para los diferentes tipos de semáforo?

9. Para determinar si los precios influyen en las ventas de un determinado producto se visitaron tres centros comerciales con análogas características, obteniéndose la siguiente información.

Precios	*Venta*		
100	200	0	200
105	90	150	150
110	300	40	40
115	150	150	75

a) ¿Se podría decir, al nivel de significación 0.05, que los diferentes niveles de precios afectan de manera diferente la venta del producto? Usar el nivel de significación de 0.05.

b) ¿Cuál es la respuesta a la pregunta en a) si se considera que los datos corresponden a un modelo de bloques aleatorios en donde cada bloque corresponde a cada uno de los tres centros comerciales en la ciudades A, B, C, respectivamente? Usar el nivel de significación de 0.05.

10. Para estudiar si la intensidad de tránsito de vehículos en 4 lugares de la ciudad es diferente o no, se colocó en cada lugar un contador en 5 días de la semana. Los resultados fueron como sigue.

Lugar	*Día de la semana*				
	Lunes	*Martes*	*Miércoles*	*Jueves*	*Viernes*
A	453	500	392	441	427
B	482	605	400	450	431
C	444	505	383	429	440
D	395	490	390	405	430

a) Analizar los resultados como si fuera un diseño de una sola vía. Use $\alpha = 0.05$.

b) Analizar los resultados como si fuera un diseño de bloques (días). Use $\alpha = 0.05$.

Indicar el modelo más adecuado para el análisis. Explicar.

¿Existe un punto de mayor tránsito? Si existe, ¿cuál es?

11. Para determinar la dureza de un metal se utiliza una máquina que tiene una punta. La punta es presionada sobre la probeta de metal y se mide la profundidad de la marca. Existe la posibilidad de utilizar diferentes tipos de puntas, pero antes de hacerlo se necesita probar si los diferentes tipos influyen en la profundidad de las marcas. Para controlar el "efecto probeta" se diseño la experiencia formando bloques con tres probetas. Así se anuló la variabilidad entre las probetas. Los resultados fueron los siguientes.

Tipo de punta	*Bloques*		
	Probeta 1	*Probeta 2*	*Probeta 3*
1	3.4	3.8	3.7
2	3.5	3.7	3.6
3	4.1	4.2	3.9

Al nivel de significación 0.05, ¿se puede indicar que los diferentes tipos de punta son distintos? ¿Fue necesario realizar el bloqueo? Usar el nivel de significación 0.05.

12. Para investigar los conocimientos de un grupo de secretarias con 5 años de servicio o más acerca de ciertos procesadores científicos se diseño un cuestionario. Para validar el cuestionario se seleccionaron a 20 secretarias a las que previamente se les solicitó que explicaran sus conocimientos y experiencias en dichos procesadores. De acuerdo a la información que proporcionaron se les clasificó en nivel bajo (A), nivel medio (B) y nivel alto (C). Las calificaciones obtenidas en el cuestionario diseñado así como el nivel obtenido aparecen a continuación.

Identificación de la secretaria	Nivel de habilidad	Calificación cuestionario	Identificación de la secretaria	Nivel de habilidad	Calificación cuestionario
1	A	82	11	A	80
2	A	114	12	A	105
3	A	90	13	B	110
4	A	80	14	B	133
5	B	128	15	C	128
6	B	90	16	B	130
7	C	156	17	B	104
8	A	88	18	C	151
9	A	93	19	C	140
10	B	130	20	C	155

¿Se puede indicar, al nivel de significación 0.05, que existen diferencias entre las medias de las calificaciones entre las secretarias de los tres niveles?

13. Los datos que se presentan a continuación corresponden a la productividad por hora en el montaje de cierto tipo de motores para tres procedimientos A, B y C. Los datos se han aleatorizado de manera conveniente, y nada hace suponer que exista algún factor que no ejerza el mismo tipo de influencia para todos los resultados obtenidos.
    a) ¿Se puede indicar que los tres procedimientos dan diferente productividad? Use $\alpha = 0.05$.
    b) Si la respuesta en a) es verdadera, ¿cuál o cuáles procedimientos difieren en productividad?

	Procedimiento		
	A	B	C
Productividad	2.6	3.2	2.6
	2.5	3.1	2.5
	3.1	3.5	2.7
	2.6	3.4	2.7

14. Para analizar la influencia que tres formas diferentes de promoción publicitaria (TV, radio y periódico) podrían tener en las ventas, una gran tienda de artefactos realizó un análisis de varianza de donde obtuvo los siguientes resultados.

Fuente de variación	Suma de cuadrados	Grados de libertad
Entre tratamientos (tipos de propaganda)	250.34	2
Dentro de los tratamientos (error)	201.66	9
Total	452.00	11

a) Usando un nivel de significación del 5%, indicar si los diferentes tipos de promoción publicitaria impactan de manera diferente en las ventas.

b) Al revisar los resultados, se dedujo que cada valor de las ventas correspondía a un tipo de propaganda y a cada uno de los días: jueves viernes y sábado. Por ello, se convino en hacer un análisis de varianza para bloques aleatorios. Si la suma de cuadrados para los bloques fue 91.66, indicar si los diferentes tipos de promoción publicitaria impactan de manera diferente en las ventas. Usar el nivel de significación 0.05.

RESPUESTAS A LOS EJERCICIOS

2.

ANOVA

Mediciones

	Suma de cuadrados	*gl*	*Media cuadrática*	*F*	*Sig.*
Intergrupos	584.133	2	292.067	52.155	0.000
Intragrupos	67.200	12	5.600		
Total	651.333	14			

3.

ANOVA

Tiempo

	Suma de cuadrados	*gl*	*Media cuadrática*	*F*	*Sig.*
Intergrupos	116.133	2	58.067	9.316	0.004
Intragrupos	74.800	12	6.233		
Total	190.933	14			

7.

ANOVA

Residuos

	Suma de cuadrados	*gl*	*Media cuadrática*	*F*	*Sig.*
Intergrupos	44.818	2	22.409	3.150	0.067
Intragrupos	128.071	18	7.115		
Total	172.890	20			

10. **Modelo de bloques aleatorios**

Variable dependiente: frecuencia

Fuente	*Suma de cuadrados*	*gl*	*Media cuadrática*	*F*	*Significación*
Lugar	6875.600	3	2291.867	3.884	0.038
Día	38594.300	4	9648.575	16.351	0.000
Error	7080.900	12	590.075		
Total	52550.800	19			

11. Usar diseño de bloques aleatorios.

12. **ANOVA**

Productividad

	Suma de cuadrados	*gl*	*Media cuadrática*	*F*	*Sig.*
Intergrupos	1.095	2	0.548	14.180	0.002
Intragrupos	0.348	9	0.039		
Total	1.443	11			

CAPÍTULO 11

Cartas de control. Análisis de la capacidad de un proceso. Aceptación por muestreo

William Edwards Deming

William Edwards Deming, considerado como el padre de la calidad moderna, nació en 1900, en Sioux City, Iowa, EE. UU. Obtuvo su bachillerato en la especialidad de Física en 1921 y su grado de Máster en Física y Matemáticas en la Universidad de Colorado en 1924. Se doctoró en Física Matemática en 1928 en la Universidad de Yale.

Trabajando en diversas agencias federales, Deming se transformó en un experto en muestreo y control de calidad. Posteriormente, Deming comandó el equipo de la Oficina de Censos de Estados Unidos y a la vez profundizó los trabajos de Walter A. Shewhart, aplicando los métodos estadísticos para el mejoramiento de la calidad.

Dejó la Oficina de Censos en 1946 para dedicarse a la consultoría relacionada con el control de la calidad, siendo su trabajo más exitoso la transformación de la industria japonesa. El comienzo de su relación con Japón sucedió cuando el gobierno americano lo envió a este país para instruir a los industriales en los métodos del control de la calidad.

Los métodos de Deming fueron utilizados por la gente de negocios de EE. UU. después de 30 años de ser aplicados con mucho éxito en Japón.

La obra de Deming se resume en una serie de trabajos de consultoría internacional, enseñanza en la universidad, el entrenamiento de estadísticos, etc. Murió en 1993.

CONTENIDO

11.1 Introducción

Desde sus orígenes, en la década de 1920, el concepto de calidad de un producto o servicio ha ido evolucionando, desde una actividad de plena medición para la detección de errores hasta una actividad que tiene por objeto la satisfacción total del usuario. En este sentido, la estadística ha jugado un rol muy importante, proporcionando técnicas poderosas para este objetivo. Estas técnicas se inician con Walter A. Shewhart, quien en 1924 creó las gráficas de control. Posteriormente, estas técnicas se robustecieron con los métodos de muestreo para inspeccionar lotes de artículos. Harold Dodge y Harold Roming desarrollaron el muestreo por aceptación como alternativa de las revisiones al ciento por ciento. Después de la Segunda Guerra Mundial y siguiendo los principios de Shewhart, Edwards Deming, en 1947, inició en el Japón el entrenamiento de cientos de ingenieros y administradores, dándole un gran impulso a la aplicación de las técnicas del control de calidad. Tomando como punto de partida los conceptos introducidos por estos investigadores se han sucedido una serie de ideas y procedimientos, como los introducidos por Genichi Taguchi en 1960, o los empleados por Armand Feigenbaum, relacionados con el control de calidad total para integrar los sistemas responsables de la calidad y el Seis Sigma, creado en 1988 por la compañía Motorola, con el objetivo de lograr la "satisfacción total del cliente".

La calidad de un producto

La calidad de un producto o servicio está directamente relacionada con el buen desempeño de las funciones básicas para las cuales fue creado: la confiabilidad (alta probabilidad de que funcione sin desperfectos y bajo normas establecidas), la facilidad y el bajo costo de mantenimiento. Estas características son atributos de carácter tangible; sin embargo, deben considerarse también las características no tangibles, como la estética, que le dan al producto o servicio "un carácter de estatus, prestigio y la ilusión de poseer el bien".

Con el control de calidad se intenta asegurar el cumplimiento de las exigencias que determinan la calidad. Para ello es necesario un proceso de fabricación o servicio que opere con baja variabilidad.

En todo este proceso de control y mejora de la calidad, la utilización de herramientas que proporciona la estadística es primordial. Estas herramientas parten de mediciones o datos del proceso que es preciso analizar. Estos datos presentan variabilidad, y su estudio permite conocer el funcionamiento y eficacia del proceso para de este modo obtener las mejoras, si fuera necesario, que satisfagan las exigencias de los clientes. Existen muchas técnicas para la detección de la variabilidad, entre las más sencillas y que permiten conocer un alto porcentaje de las causas de esta variabilidad están: las hojas de recogida de datos o de registro, los histogramas, el diagrama de Pareto, los diagramas de dispersión, las cartas de control, los diagramas de causa-efecto y los diagramas de flujo.

Las causas comunes y asignables

Se considera que las causas que originan la variabilidad de un proceso son de dos tipos: las *causas comunes* o *fortuitas* y las *causas asignables*.

Las ***causas comunes o fortuitas***, llamadas también *ruido de fondo*, son las causas inherentes al proceso. Son las que siempre existen, aun cuando el proceso sea adecuado. La eliminación de estas causas es difícil y depende de la dirección que lleva a cabo el proceso. Estas causas son el resultado de un cúmulo de muchas causas pequeñas. Muchas veces la variabilidad que estas originan tiene una representación estadística, permitiendo su predicción.

Las ***causas asignables o específicas*** son causas que aparecen de manera esporádica en el proceso, permitiendo su identificación y fácil eliminación. A menudo este tipo de causas no son muchas, pero sus efectos son importantes en el proceso. La eliminación de estas causas dependerá del operador o de la administración.

Los procesos bajo control o estables

Shewart estableció el concepto de ***proceso bajo control*** para describir los procesos cuya variabilidad se debe solo a causas comunes. Un proceso que está bajo control también se llama *proceso estable*.

Específicamente, un proceso es estable si los parámetros que caracterizan a las variables que se usan para medir el desempeño del proceso son constantes por un periodo de tiempo suficientemente largo.

Si un proceso opera bajo causas asignables se dice que el proceso está fuera de control.

11.2 Las cartas o gráficas de control

Una ***carta o gráfica de control*** es un gráfico en donde se registran las realizaciones muestrales de una característica de calidad en función del tiempo, permitiendo *detectar de manera rápida la presencia de causas asignables del proceso*. Estas herramientas, fáciles de usar e interpretar, fueron propuestas en 1920 por W. Shewart.

El establecimiento de una carta de control contempla la designación previa del modelo estadístico que gobierna el proceso. Generalmente el modelo subyacente es la distribución normal cuando se analizan características continuas de un producto o la distribución binomial cuando se estudian propiedades de carácter cualitativo.

Una carta de control consta de tres líneas paralelas, separadas a igual distancia:

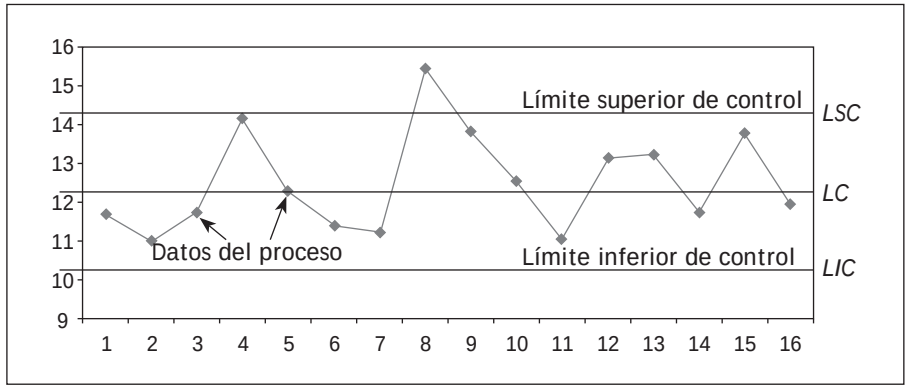
La *línea central* (*LC*), que representa a la media del proceso en estado de control.

La *línea superior de control* (*LSC*), por encima de la línea central.

La *línea inferior de control* (*LIC*), por debajo de la línea central.

A estas dos últimas líneas se les llama ***límites de control o límites de tolerancia natural***.

FIGURA 11.1 *Carta de control*



En el eje vertical de la carta se representa el indicador de la variable cuya calidad se está midiendo, por ejemplo, medias, rangos y proporciones.

Se considera que el proceso es estable o está bajo control si los puntos están situados dentro de los límites superior e inferior de control. Cuanto más cerca están los puntos de la línea central o media, el proceso será más estable.

Es posible, sin embargo, tener procesos que están fuera de control aun cuando todos los puntos están dentro de los límites de control. Esto sucede cuando los puntos presentan patrones de comportamiento no aleatorios que describen: trayectorias cíclicas, estratos o agrupamientos alrededor de la línea de control, desplazamientos en cierta dirección, cambios en el nivel del proceso y cercanías a los límites de control.

Las diversas causas que originan estos patrones pueden ser: cambios ambientales, desgaste de herramientas, cambios en los horarios, ajuste inadecuado de las máquinas, introducción de nuevos trabajadores, cambios en las materias primas, falta de motivación de los operarios, cansancio de los trabajadores, cálculos incorrectos de los parámetros, diversos modelos subyacentes en el proceso, etcétera.

El reconocimiento de las causas de los patrones en un proceso no es una tarea fácil; ello requiere de experiencia y conocimiento del proceso, y no solo de los principios estadísticos de las cartas de control.

El análisis para determinar si un proceso está bajo control está relacionado con la prueba de la hipótesis nula: el *proceso está bajo control* versus la hipótesis alternativa: *el proceso no está bajo control*. El error de tipo I de la prueba (indicar que el proceso no está bajo control cuando realmente lo está) se reduce alejando los límites de control de la línea central de la carta de control, mientras que el error de tipo II (indicar que el proceso está bajo control cuando realmente no lo está) se reduce si los límites se acercan a la línea central.

Teniendo en cuenta lo anterior, la obtención de gráficos de control adecuados dependerá de la elección equilibrada de la probabilidad de cometer el error de tipo I y la probabilidad de cometer el error de tipo II. Los límites de control se fijan de tal manera que la mayoría de los puntos de la muestra estén dentro de ellos cuando el proceso está bajo control. Generalmente, se exige que, cuando el proceso está bajo control, la probabilidad de que los valores de las características que se miden estén fuera de los límites sea menor o igual a 0.0027. Con este criterio, y si las medidas tienen distribución normal, los límites de control se encontrarán *a tres desviaciones estándar de la media muestral*, a partir de la media del proceso.

En resumen, las cartas de control son útiles para mejorar la productividad (la aplicación exitosa de las cartas de control reducen los reprocesamientos y los desechos), previenen los defectos y el ajuste innecesario del proceso, permiten la implementación de cambios en el proceso –al comportarse como elementos de diagnóstico– y proporcionan información de gran utilidad para el diseño del producto.

Dependiendo del tipo de variable o característica que se controla, hay dos tipos de gráficas de control: gráficas de control *para variables* y gráficas de control *para atributos*.

Cartas de control para variables

Estas gráficas se utilizan para controlar valores de una variable continua, considerados como características de calidad. Por ejemplo, para controlar la cantidad de líquido vertido en un depósito, el tiempo de realización de una tarea, la temperatura necesaria para un proceso químico, etcétera.

Si se supone que el modelo para describir los valores de la variable es la distribución normal, bastará con estudiar el cambio de la media y de la varianza del proceso. El cambio de la media del proceso se controla graficando las medias de las muestras con tamaño, generalmente, entre 4 y 6. Estas gráficas, que se refieren a la media, se llaman *cartas de control de medias*. El cambio de la varianza del proceso se controla graficando *los rangos*, obteniéndose las *cartas de control de rangos*.

Cartas de control de medias y rangos

Cuando los parámetros de la distribución de las medidas que se controlan se conocen, los límites de control de la carta de control de las medias se rigen por la variabilidad natural del proceso. Estos valores son los siguientes:

$$LSC: \quad \mu + 3\frac{\sigma}{\sqrt{n}}$$

$$LC: \quad \mu$$

$$LSC: \quad \mu - 3\frac{\sigma}{\sqrt{n}}$$

donde μ y σ son, respectivamente, la media y la desviación estándar de la variable que describe el proceso, y n es el tamaño de las muestras que se toman durante el control.

Una carta con los límites indicados se llama carta a 3 *sigmas*. Si un proceso está funcionando bajo control, se espera que de 1,000 medias graficadas en esta carta, 3 de ellas estarán fuera de los límites de control, aproximadamente.

Cuando los parámetros del proceso no se conocen, estos deben estimarse y así resultan los límites:

$$LSC: \quad \bar{\bar{x}} + A_2\bar{R}$$

$$LC: \quad \bar{\bar{x}}$$

$$LIC: \quad \bar{\bar{x}} - A_2\bar{R}$$

Los límites de una carta de control para los rangos R son:

$$LSC: \quad D_4\bar{R}$$

$$LC: \quad \bar{R}$$

$$LIC: \quad D_3\bar{R}$$

Las constantes A_2, D_3 y D_4 dependen del tamaño de la muestra que se utilice; sus valores se encuentran en la Tabla 11.1

TABLA 11.1 *Factores para la construcción de gráficos de control*

Número de observaciones en cada muestra	*Factor para la media*	*Factores para el rango*		
n	A_2	D_3	D_4	d_2
2	1.880	0	3.267	1.128
3	1.023	0	2.575	1.693
4	0.729	0	2.282	2.059
5	0.577	0	2.115	2.326
6	0.483	0	2.004	2.534
7	0.419	0.076	1.924	2.704
8	0.373	0.136	1.864	2.847
9	0.337	0.184	1.816	2.970
10	0.308	0.223	1.777	3.078

Los valores $\bar{x}$ y $\bar{R}$ se estiman como se indica en el procedimiento que aparece a continuación.

1. Tomar k muestras de tamaño n de la variable que corresponde a la característica en estudio. El número k debe ser mayor o igual a 20 y el tamaño de cada muestra n debe estar entre 2 y 6.

2. Calcular la media $\bar{x}_i$ y el rango R_i de cada muestra.

Muestras	*Medias*	*Rangos*
$x_{11}, x_{12},\ldots, x_{1n}$	$\bar{x}_1$	$R_1 = \max(x_{1j}) - \min(x_{1j})$ $j = 1, 2,\ldots, n$
...	...	...
$x_{k1}, x_{k2},\ldots, x_{kn}$	$\bar{x}_k$	$R_k = \max(x_{kj}) - \min(x_{kj})$ $j = 1, 2,\ldots, n$

3. Estimar μ y R, respectivamente, con $\bar{\bar{x}} = \dfrac{\sum_{i=1}^{k} \bar{x}_i}{k}$ y $\bar{R} = \dfrac{\sum_{i=1}^{k} R_i}{k}$.

4. Calcular los límites de control, tal como antes se indicó.

La media μ del proceso se estima con $\bar{\bar{x}}$, mientras que la desviación estándar σ puede estimarse con $\dfrac{\bar{R}}{d_2}$, donde d_2 aparece en la Tabla 11.1.

Este procedimiento, que proporciona los límites de control, indica la manera de analizar la estabilidad del proceso. Si el proceso está bajo control, estos límites se pueden tomar para monitorear y controlar el proceso durante el periodo llamado de *vigilancia*.

Si existieran observaciones fuera de los límites de control, se eliminan las muestras relacionadas a estas observaciones y se recalculan los límites de control a partir de los nuevos estimadores de la media y el rango. Si hubiera que eliminar muchas observaciones se tendría que tomar un nuevo conjunto de muestras e iniciar el proceso descrito.

EJEMPLO. *Control de calidad de servicios hospitalarios*

Una empresa de servicios hospitalarios ha proyectado atender a los pacientes después de un tiempo de espera comprendido en el intervalo 13 ± 1 minutos. Para controlar si esto se mantiene a lo largo del tiempo se miden los tiempos de espera de 20 grupos de 4 pacientes cada uno. Las medias y los rangos calculados se indican a continuación.

Grupo	*Media*	*Rango*	*Grupo*	*Media*	*Rango*
1	11.68	2.34	2	13.83	1.07
3	11.01	1.48	4	12.56	2.87
5	11.72	1.59	6	11.03	2.86
7	14.18	2.81	8	13.16	1.01
9	12.27	2.36	10	13.21	2.12
11	11.38	2.52	12	11.75	2.11
13	11.22	1.94	14	13.78	2.71
15	15.45	3.20	16	11.95	1.10
17	12.70	1.80	18	13.50	1.15
19	13.40	2.10	20	13.24	1.50

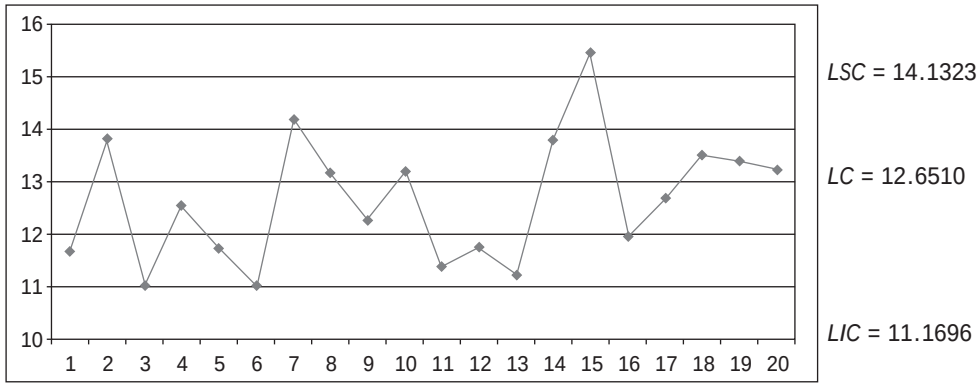
Los límites de control y la carta de control aparecen a continuación.

LSC: $\bar{\bar{x}} + A_2\bar{R} = 12.6510 + 0.729(2.0320) = 14.1323$

LC: $\bar{\bar{x}} = 12.6510$

LIC: $\bar{\bar{x}} - A_2\bar{R} = 12.6510 - 0.729(2.0320) = 11.1696$

FIGURA 11.2 *Carta de control para la media*



El proceso no está bajo control, las medias de las muestras 7 y 15 están fuera de los límites de control. Si se deseara establecer la vigilancia del proceso habría que eliminar las muestras 7 y 15 y recalcular los límites, estableciendo previamente las causas asignables que han dado origen a las observaciones fuera de los límites de control. Se observa que los límites de control inherentes al proceso no coinciden con los límites (13 ± 1), impuestos desde el exterior del proceso. En general, los límites impuestos no necesariamente coinciden con los límites del proceso aun cuando el proceso esté bajo control.

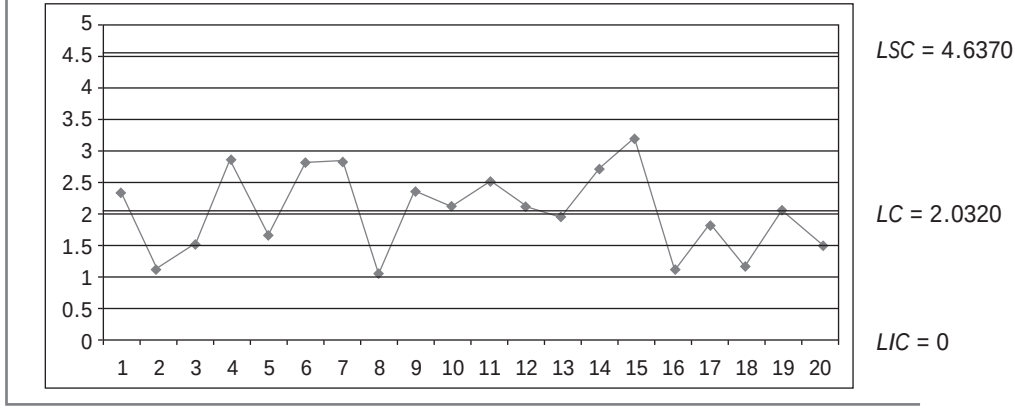
Los límites de control para *R* son los siguientes.

LSC: $D_4\overline{R} = 2.282(2.0320) = 4.6370$

LC: $\overline{R} = 2.0320$

LIC: $D_3\overline{R} = 0(2.0320) = 0$

FIGURA 11.3 *Carta de control para R*



Cartas de control para atributos

Estas cartas sirven para registrar valores (llamados *atributos*) de variables aleatorias discretas, como por ejemplo el número de defectos, el porcentaje de artículos defectuosos.

Los tipos de cartas o gráficas de control para atributos son:

1. Las cartas de control *p*, que registran el porcentaje de unidades defectuosas.
2. Las cartas de control *np*, que registran el número de unidades defectuosas.
3. Las cartas de control *c*, que registran el número de defectos observados.
4. Las cartas de control *u*, que registran el número de defectos por unidad.

Cartas de control para la proporción de productos defectuosos, *p*

En estas cartas se registran a lo largo del tiempo las proporciones de productos que resultan defectuosos o disconformes, al no tener la cualidad que define su calidad.

Al igual que para la carta de las medias, los límites de control de una carta de control para *p* se construyen considerando que el proceso está bajo control y que debe existir, aproximadamente, el 0.3% de las proporciones fuera de estos límites. El procedimiento es el siguiente.

1. Considerar *k* muestras de tamaño *n* cada una. Si se conoce históricamente la proporción *p* de productos defectuosos, los límites de control son:

   $LSC: p + 3\sqrt{\frac{p(1-p)}{n}}$

   $LC: p$

   $LIC: p - 3\sqrt{\frac{p(1-p)}{n}}$

   Los límites se obtienen suponiendo que la distribución de las proporciones se aproxima a la distribución normal.

2. Si no se tiene información histórica de la proporción de productos defectuosos, se calcula la proporción p_i de defectuosos en cada muestra *i* y

3. Los límites de control se estiman con:

   $LSC: \bar{p} + 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$

$LC: \bar{p}$

$$LIC: \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

$$\text{donde } \bar{p} = \frac{\sum_{i=1}^{k} np_i}{kn} = \left(\frac{\textit{Total de productos defectuosos}}{\textit{Total de productos en la muestra}}\right).$$

Establecidos los límites de control, las proporciones p_i, para cada muestra, son registradas en la carta para comprobar que no hubo causas asignables que influyeron en el proceso. Si fuera así, se establecen los límites de control y posteriormente se monitorea las proporciones generadas en el proceso.

Ocasionalmente una característica medible puede transformarse en un atributo. Según esta operación, un producto se considera defectuoso si las medidas salen fuera del intervalo de tolerancia que define su calidad. Usando este procedimiento la operación de control es fácil; sin embargo, con este procedimiento se pierde la información que se refiere a la característica continua del producto.

EJEMPLO. *Control de calidad de computadoras*

En la Tabla 11.2 se registra el número de computadoras con algún defecto en 12 días de producción.

Construir un gráfico de control para el porcentaje de computadoras defectuosas. Indicar si el proceso está bajo control.

TABLA 11.2 *Computadoras con defecto*

Día	*Inspecciones*	*Con algún defecto*
1	1,000	2
2	1,000	1
3	1,000	3
4	1,000	4
5	1,000	2
6	1,000	3
7	1,000	5
8	1,000	3
9	1,000	6
10	1,000	3
11	1,000	5
12	1,000	4

Solución

El estimador de p es $\bar{p} = \dfrac{\sum_{i=1}^{12} x_i}{12(1000)} = \dfrac{41}{12000} = 0.003416$ ($LC = 0.003416$).

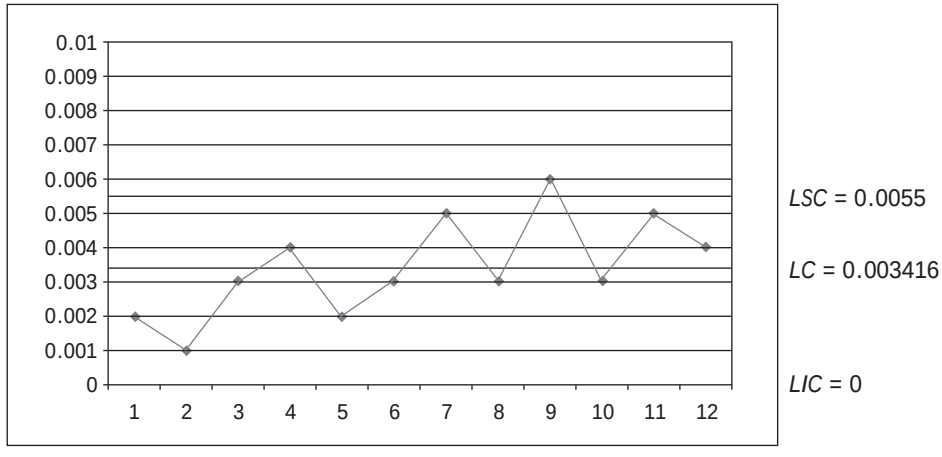
Los límites de control superior e inferior son:

$LSC = 0.003416 + 3.\sqrt{\dfrac{0.003416(1 - 0.003416)}{1000}} = 0.0055$ y

$LIC = 0.003416 - 3.\sqrt{\dfrac{0.003416(1 - 0.003416)}{1000}} = -0.0055$ (se considera que LIC es igual a 0, por ser negativo).

La gráfica de control de proporciones de productos defectuosos es la Figura 11.4.

FIGURA 11.4 *Carta de control de proporciones. La tendencia es creciente*



El proceso está fuera de control (la proporción de productos defectuosos de la muestra 9 está fuera de los límites de control). Además deberá ponerse atención a la tendencia creciente del proceso.

Cartas de control para el número de productos defectuosos, *np*

Estos gráficos de control sirven para controlar el número de artículos defectuosos utilizando muestras de tamaño fijo.

El modelo que describe el número de artículos defectuosos en cada muestra de tamaño n, constante, es la distribución binomial de parámetros n y p. Esta distribución puede aproximarse a la distribución normal cuando el tamaño de la muestra es grande. Usando esta propiedad, se tienen los límites de control:

LSC: $np + 3\sqrt{np(1 - p)}$

LC: np

LIC: $np - 3\sqrt{np(1 - p)}$, cuando p se conoce

Si p no se conoce, el número de artículos defectuosos np en cada muestra se estima con $\bar{d} = n\bar{p}$, donde $\bar{p} = \dfrac{\sum_{i=1}^{k} p_i}{k}$ y k es el número de muestras.

De esta manera, los límites de control para el número de artículos defectuosos son:

LSC: $n\bar{p} + 3\sqrt{n\bar{p}(1 - \bar{p})}$

LC: $n\bar{p}$

LIC: $n\bar{p} - 3\sqrt{n\bar{p}(1 - \bar{p})}$

Si el proceso está bajo control y no existen causas asignables que afecten el proceso, los límites de control quedan establecidos de la manera indicada.

Si existe la evidencia de causas asignables, estas deberán identificarse y hacer las correcciones necesarias y cambiar los límites de control.

Como en el caso de los gráficos de la media y rango, con estos gráficos se trata de mantener el proceso bajo control sin que esto asegure que el proceso satisface las especificaciones o tolerancias que se exigen.

EJEMPLO. *Control de calidad de computadoras*

Usando los datos del problema anterior, se tiene que los límites de control son:

LSC: $1000(0.003416) + 3\sqrt{1000(0.003416)(1 - 0.003416)} = 8.9512$

LC: $1000(0.003416) = 3.416$

LIC: $1000(0.003416) - 3\sqrt{1000(0.003416)(1 - 0.003416)} = -2.1192$

El lector puede construir el gráfico de control para np.

Cartas de control para el número de disconformidades por unidad de inspección: *u*

Para tener esta carta de control, se considera en primer lugar muestras de tamaño constante *n*. El número *C* de disconformidades en cada muestra es una variable aleatoria que tiene distribución de Poisson de parámetro *u = c/n por unidad*, en donde *c* es el número esperado de disconformidades en la muestra.

El valor esperado del número de disconformidades por unidad es *u* y su varianza es *u/n*.

Usando estos resultados se tiene que los límites de control son:

LSC: $u + 3\sqrt{u/n}$

LC: u

LIC: $u - 3\sqrt{u/n}$

Cuando *u* no se conoce, se puede estimar con $\bar{u} = \frac{1}{n.k}\sum_{i=1}^{k} c_i$, en donde *k* es el número de muestras y *n* es el tamaño de cada muestra.

EJEMPLO. *Gráfica de control para disconformidades por impresora*

Con la finalidad de obtener un gráfico de control para el número de disconformidades por impresora que se fabrica en la empresa CCC, se seleccionaron 20 muestras de 4 impresoras cada una. Los resultados fueron los siguientes.

TABLA 11.3 *Número de disconformidades*

Muestra i	*Tamaño de muestra n*	*Número de disconformidades por muestra* c_i	*Número de disconformidades por unidad* u_i
1	4	7	1.75
2	4	3	0.75
3	4	7	1.75
4	4	7	1.75
5	4	3	0.75
6	4	8	2.00
7	4	8	2.00
8	4	9	2.25
9	4	8	2.00
10	4	4	1.00
11	4	9	2.25
12	4	8	2.00
13	4	8	2.00
14	4	14	3.50
15	4	8	2.00
16	4	11	2.75
17	4	7	1.75
18	4	9	2.25
19	4	8	2.00
20	4	7	1.75
	80	153	

El estimador del número de disconformidades por unidad es $\bar{u} = \frac{\sum_{i=1}^{k} c_i}{n.k} = \frac{153}{80} = 1.91$ y los límites de control son:

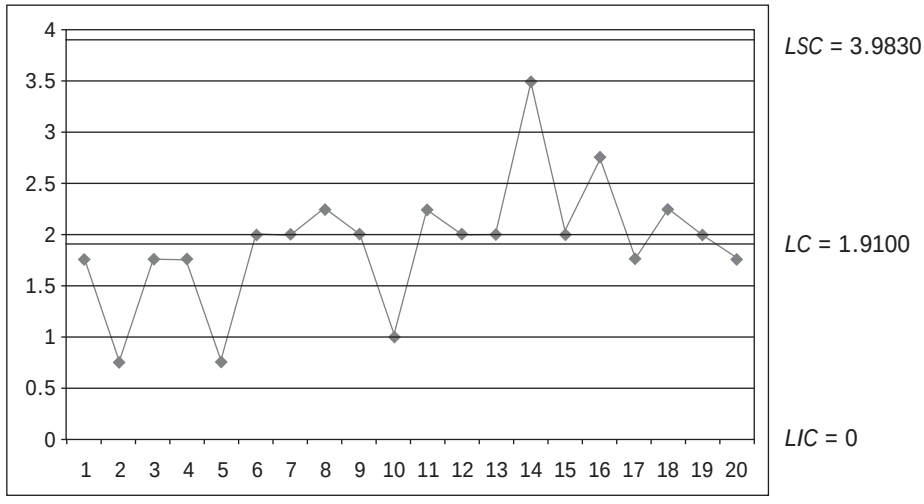
LSC: $\bar{u} + 3\sqrt{\bar{u}/n} = 1.91 + 3\sqrt{1.91/4} = 3.9830$

LC: $\bar{u} = 1.91$

LIC: $\bar{u} - 3\sqrt{\bar{u}/n} = 1.91 - 3\sqrt{1.91/4} = -0.1630$

El límite de control inferior, *LIC*, se considera que es 0. La carta de control aparece a continuación. En ella se observa que el proceso está bajo control.

FIGURA 11.5 *Carta de control de disconformidades*



Límites de prevención

Los límites de prevención son los que se establecen dentro de los límites de control con la finalidad de utilizarlos como aviso o advertencia de que el proceso podría estar fuera de control. Estos límites son importantes, pues es posible que un proceso no muestre señales que indiquen que el proceso está fuera de control aun cuando todos los puntos están dentro de los límites de control establecidos. Los límites de prevención que generalmente se establecen están a 1 y 2 desviaciones estándar del límite central.

Interpretación de las cartas de control

En general, las cartas de control se utilizan para detectar, con alta probabilidad, anomalías en los procesos. Una señal que indica que el proceso no está bajo control es *cualquier observación que está fuera de los límites de control*. Sin embargo, aun cuando esto no suceda, es posible que el proceso no esté bajo control cuando las observaciones forman patrones dentro de los límites establecidos. Algunos de estos patrones pueden ser:

a) Catorce puntos alternados alrededor del límite central.

b) Seis o más puntos consecutivos con tendencia creciente o decreciente.

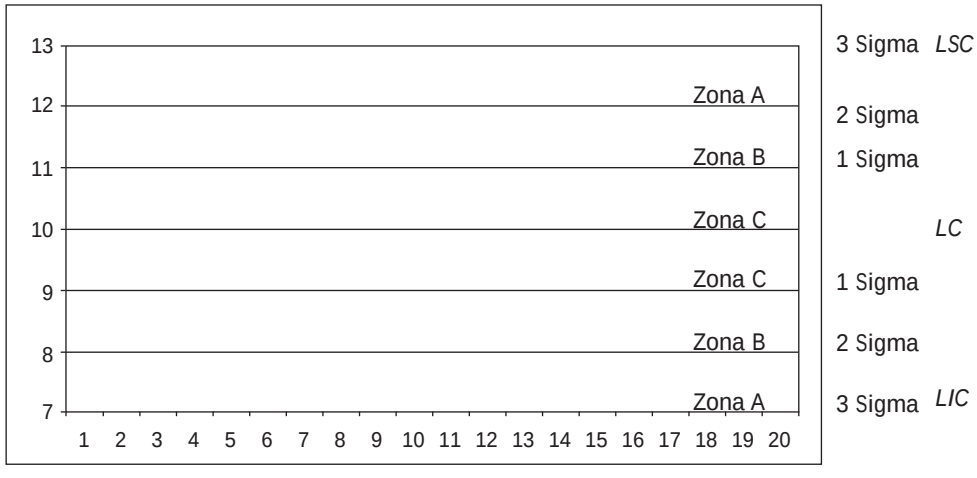
c) Siete u ocho puntos consecutivos al mismo lado del límite central.

d) Ocho puntos sucesivos dentro de los límites de prevención que están a 1 desviación estándar del límite central (zona C) (Figura 11.6).

e) Cuatro de cinco puntos sucesivos en la región comprendida entre los límites de prevención que están a 1 y 2 desviaciones estándar del límite central (zona B) (Figura 11.6).

f) Dos de tres puntos consecutivos en la región comprendida entre los límites de prevención que están a 2 y 3 desviaciones estándar del límite central (zona A) (Figura 11.6).

FIGURA 11.6 *Zonas de prevención*



11.3 Análisis de la capacidad de un proceso

Se ha indicado que una manera de medir la perfomance de un proceso en marcha es a través de las cartas de control. Con las cartas de control se pueden encontrar las causas asignables de la variabilidad para luego aplicar las acciones correctivas.

Sin embargo, una carta de control no permite conocer si el proceso satisface los requerimientos o exigencias del cliente. Un proceso puede estar bajo control, pero no necesariamente satisface los requerimientos del cliente.

Los requerimientos de un cliente se establecen a partir de un intervalo [*LIE*, *LSE*], llamado *intervalo de tolerancia o de especificación*, donde *LIE* es el límite inferior de especificación, mientras que *LSE* es el límite superior de especificación. Se considera que esta es la *voz del cliente*, mientras que los límites de las cartas de control corresponden a la *voz del proceso*.

Con el *análisis de la capacidad del proceso* se trata de comparar la variabilidad de un proceso estable o bajo control con las especificaciones de los requerimientos del consumidor.

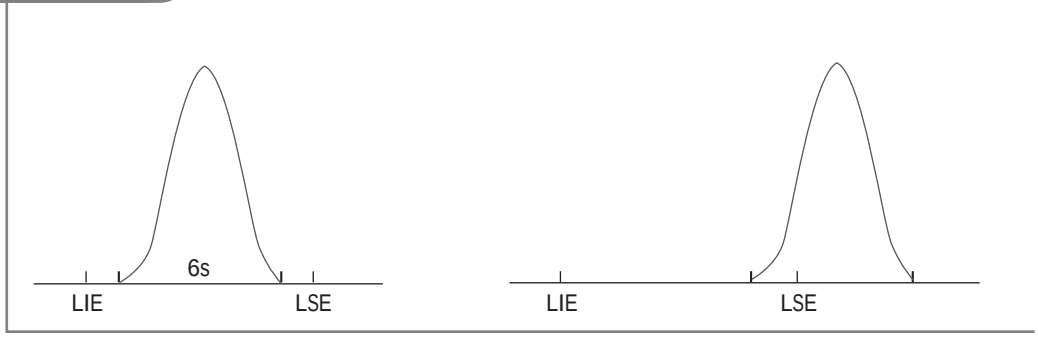
Un proceso *es* ***capaz*** si tiene la aptitud de producir unidades cuya característica de calidad está dentro del intervalo de tolerancia o especificación. Es decir, si está centrado y la longitud del intervalo de especificación es mayor que la longitud del intervalo de control, [*LSC, LIC*].

Existen diversos índices para medir la capacidad del proceso; el más simple consiste en calcular el porcentaje de productos que satisfacen las especificaciones. A mayor porcentaje mayor capacidad del proceso.

Otro de los índices de la capacidad de un proceso se define como $C_p = \dfrac{LSE - LIE}{6\sigma}$.

El índice C_p compara la longitud del intervalo de especificación, [*LIE, LSE*], con el rango de los valores del proceso (si el proceso está bajo control y el modelo subyacente es el de la distribución normal, el rango mide aproximadamente 6σ).

FIGURA 11.7 *Proceso centrado y proceso no centrado*



Cuando el proceso está centrado, es decir, si la media del proceso coincide con la media del intervalo de especificación, se puede hacer las siguientes interpretaciones:

Si $C_p > 1$, $((LSE - LIE) > 6\sigma)$ se considera que el proceso es capaz.

Si $C_p < 1$, $((LSE - LIE) < 6\sigma)$ se considera que el proceso no es capaz.

Si $C_p = 1$, $((LSE - LIE) = 6\sigma)$ se considera que el proceso es marginalmente capaz.

Si el parámetro σ no se conoce se estima con $\hat{\sigma} = \bar{R}/d_2$ (d_2 *aparece en la Tabla* 11.1)

Dependiendo del valor de C_p se consideran cinco categorías para los procesos centrados. Estas se encuentran en la Tabla 11.4.

TABLA 11.4 *Clasificación de los procesos centrados*

Valores de C_p	***Categoría del proceso***	
$C_p \geq 2$	Clase mundial	Calidad excelente.
$1.33 \leq C_p < 2$	1	Calidad adecuada.
$1 < C_p < 1.33$	2	Parcialmente adecuado. Requiere de control estricto.
$0.67 < C_p < 1$	3	No adecuado. Debe ser modificado.
$C_p < 0.67$	4	No adecuado. Requiere de modificaciones.

EJEMPLO. *Capacidad para atender a los clientes en una ventanilla de atención al cliente*

Después de estudiar los tiempos X de atención en una de las ventanillas de un banco se ha determinado que el proceso está bajo control. La gerencia de operaciones ha indicado que las especificaciones del tiempo de servicio en minutos son: $LIE = 8$ y $LSE = 12$. Si la desviación estándar de los tiempos de atención en la ventanilla se ha estimado con $\hat{\sigma} = 3.4$ y la media del proceso es 10 minutos (el proceso está centrado), entonces:

El índice de capacidad del proceso es $C_p = \dfrac{12 - 8}{6(3.4)} = 0.5882$.

El proceso no es capaz. No es adecuado y tendrá que realizarse serias modificaciones.

También, la probabilidad de que un tiempo de servicio no caiga en el intervalo de especificaciones es:

$$1 - P(8 \leq X \leq 12) = 1 - P\left(\frac{8 - 10}{3.4} \leq Z \leq \frac{12 - 10}{3.4}\right) = 0.7218 - 0.2781 = 0.4437$$

El porcentaje de los servicios que se espera que no cumplan las especificaciones es 44.37%.

Cuando el proceso no está centrado puede ocurrir que, aun cuando la magnitud natural del proceso sea pequeña, se esté produciendo un gran porcentaje de productos que no satisfacen las especificaciones. Por ello se define el ***índice de funcionamiento*** con:

$$C_{pk} = mínimo\left(\frac{LSE - \mu}{3\sigma}\ ,\ \frac{\mu - LIE}{3\sigma}\right)$$

Los parámetros μ y σ tendrán que ser estimados si no se conocen.

El índice *k* se refiere al valor $k = \dfrac{\dfrac{LSE + LIE}{2} - \mu}{\dfrac{LSE - LIE}{2}} \times 100$, que mide la distancia de la media del proceso al centro del intervalo de especificación.

El índice *k* mide qué tan descentrado está el proceso en función de la mitad de la amplitud de las especificaciones.

Valores de *k* menores que 20%, en términos absolutos, se pueden considerar como aceptables. A medida que *k* crece en valor absoluto el proceso será más descentrado.

EJEMPLO. *Índice de funcionamiento*

Las especificaciones indicadas para el llenado de bebidas gaseosas en una planta embotelladora es 1 ± 0.02 litros. Si los estimadores de la media y la desviación estándar del volumen de líquido depositado, cuando el proceso está bajo control, son: 0.995 y 0.001, respectivamente, hallar el índice de funcionamiento del proceso.

Solución

El índice de funcionamiento es:

$$C_{pk} = min\left(\frac{1.02 - 0.995}{3(0.001)}\ ,\ \frac{0.995 - 0.98}{3(0.001)}\right) = min(8.3333,\ 5) = 5$$

El proceso es capaz, pero debemos medir su descentralización.

$$k = \frac{\dfrac{1.02 + 0.98}{2} - 0.995}{\dfrac{1.02 - 0.98}{2}} \times 100 = 25\%$$

Este resultado indica que el proceso es descentrado, lo que contribuye de manera considerable a la baja capacidad del proceso.

11.4 Planes de muestreo. Aceptación por muestreo

La aceptación por muestreo es otra técnica para auditar la calidad de un producto terminado. Utilizando esta técnica se aceptan o se rechazan lotes de elementos a partir de muestras extraídas de estos. Los lotes aceptados se incorporan al proceso de producción; los lotes rechazados pueden devolverse al productor o someterse a otra acción, como reemplazar los elementos defectuosos, previa revisión del 100% del lote.

Este tipo de muestreo, conocido también como ***control de recepción***, se desarrolló en los laboratorios Bell, alcanzando su avance pleno en la Segunda Guerra Mundial, en donde se incorporó a los estándares militares. Este procedimiento se aplica en ocasiones en las cuales se sospecha que el proveedor del producto (aun cuando tiene certificación de buena calidad) ha disminuido la calidad o cuando se han producido cambios de proveedores.

De acuerdo al plan de muestreo, al recibir un lote de artículos el cliente puede actuar de la siguiente manera:

a) Aceptar el lote sin ninguna inspección, situación que ocurre cuando existe plena confianza en el proveedor, al estar respaldado por algún tipo de certificación.

b) Aceptar el lote, previa la inspección al 100%. Esta acción se usa sobre todo cuando se produce un costo muy alto al dejar pasar un artículo defectuoso o se ha perdido la confianza en el proveedor.

c) Aceptar o rechazar el lote, realizando previamente un muestreo por aceptación. Esta situación se presenta cuando el costo y el tiempo a dedicar por inspeccionar el 100% del lote es muy elevado.

La aceptación de lotes por muestreo es útil cuando:

a) Es necesario destruir los productos.

b) La inspección del 100% no es tecnológicamente factible o costosa o requiere de mucho tiempo.

c) Los lotes tienen muchos elementos, permitiendo que la tasa de errores por inspección se eleve.

d) El proveedor tiene un buen historial de calidad excelente.

e) Sea necesario un monitoreo continuo de la producción.

Es necesario recalcar que con esta técnica no se estima ni se controla directamente la calidad, es solo para asegurarse de que la salida de un proceso cumple con los requerimientos y que existe el riesgo de aceptar lotes malos y de rechazar lotes buenos, aun cuando el muestreo se diseñe adecuadamente.

Tipos de aceptación por muestreo

La aceptación por muestreo puede ser: *por atributos y por variables*.

La aceptación por muestreo *por atributos* se realiza extrayendo una muestra aleatoria del lote, y cada elemento de la muestra se clasifica, de acuerdo a ciertas características, en "defectuoso" y "no defectuoso". Si el número de unidades defectuosas es menor o igual a un cierto número antes fijado, *c*, el lote es aceptado, de otra manera el lote es rechazado.

La aceptación por muestreo *por variables* se realiza sacando una muestra aleatoria del lote y midiendo una característica de calidad de tipo continuo (peso, longitud, etcétera). Posteriormente, y a partir de la muestra, se calcula algún estadístico de la muestra y se compara con el valor convenido por el productor y el consumidor. Dependiendo de esta operación, se aceptará o se rechazará el lote.

De acuerdo a las fases en que se realiza el muestreo de aceptación puede clasificarse en muestreo de aceptación simple (una sola fase), muestreo de aceptación doble (dos fases) y así sucesivamente.

Se desarrolla el muestreo por atributos simple por ser el de más fácil aplicación.

El convenio entre el consumidor y el productor consiste en la aceptación por parte del consumidor de una parte no excesiva de productos defectuosos. Es así que el productor indica que sus productos tienen una proporción de defectuosos p_0. El consumidor acepta esta proporción; sin embargo, por la variabilidad que puede existir en la producción, está dispuesto aceptar una proporción igual a p_1 (mayor a p_0).

Al valor p_0 se le llama *nivel de calidad aceptable* (AQL, *Aceptante Quality Level*).

A p_1, proporción de productos defectuosos por lote que un consumidor está dispuesto a tolerar, se le llama *tolerancia del porcentaje de defectuosos en el lote* (LTPD, *Lot Tolerante Percentaje Defective*). El LTPD es el nivel de calidad más pobre que el consumidor está dispuesto a aceptar en un lote individual.

Plan de muestreo por aceptación simple para atributos

Llevar a cabo ***un plan de muestreo*** por aceptación simple para atributos es sencillo. Básicamente consiste en elegir los enteros positivos *n* y *c* y seguir los siguientes pasos:

1) Se toma una muestra aleatoria de *n* unidades del lote de *N* unidades terminadas.
2) Si hay *c* o menos artículos defectuosos en la muestra se acepta el lote. De otro modo, se rechaza.

Esto es, si *X* = *Número de elementos no conformes en la muestra, el lote se rechaza si X > c*.

Así por ejemplo, un plan de muestreo con N = 1,000, n = 50 y c = 4, significa que del lote de 1,000 artículos se sacarán 50 artículos al azar; si el número de defectuosos X es menor o igual a 4, entonces se acepta el lote.

A c se le llama ***número de aceptación***.

Con este procedimiento se trata de contrastar la hipótesis del productor $H_0 : p = p_0$ versus la hipótesis del consumidor $H_1 : p > p_0$.

En este contexto, existen dos tipos de errores:

Error de tipo I, que sucede si se rechaza un lote bueno.

Error de tipo II, que sucede si se acepta un lote malo.

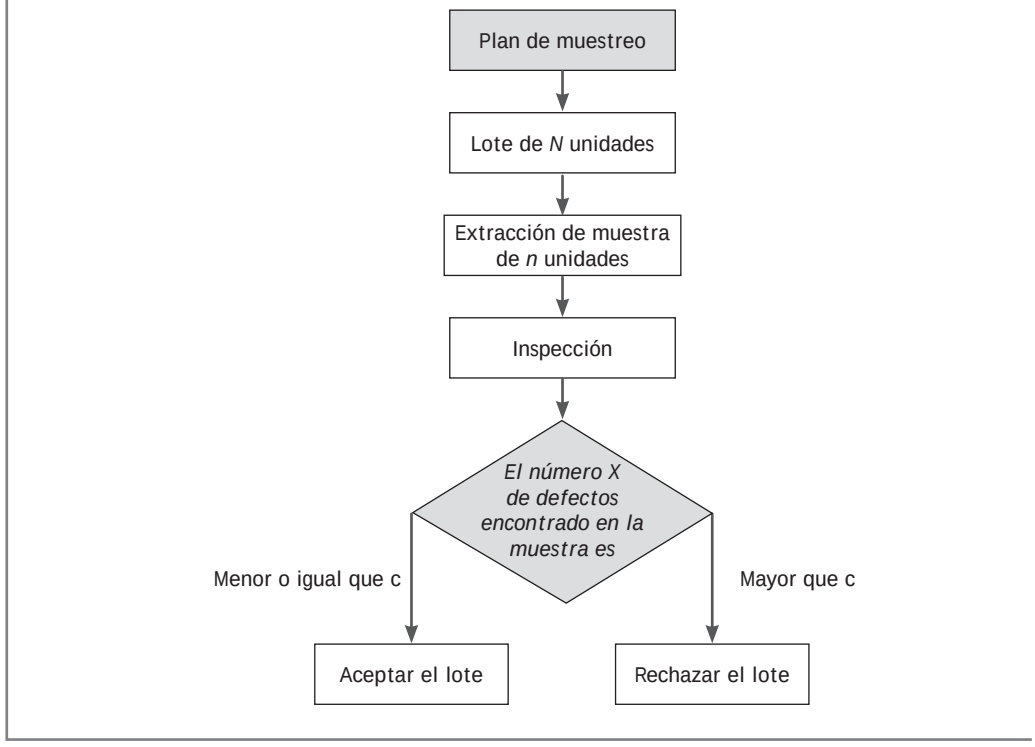
Al nivel de significación α (probabilidad de cometer el error de tipo I) se le llama *riesgo del productor*. Este valor corresponde al porcentaje aproximado de lotes buenos que serán rechazados por el consumidor.

A β (probabilidad de cometer el error de tipo II) se le llama *riesgo del consumidor*. Este valor corresponde al porcentaje aproximado de lotes malos que serán aceptados por el consumidor.

La idea es determinar n y c ("fijar el plan de muestreo") para valores de α y β, previamente fijados. Generalmente los valores que se fijan para α están alrededor de 0.05, mientras que para β están alrededor de 0.10.

Como en muchos de los procedimientos estadísticos, es necesario tener en cuenta algunas condiciones para que el muestreo por aceptación tenga buenos resultados. Por ejemplo, que las unidades que lo conforman hayan sido fabricadas en las mismas condiciones, que los lotes sean manipulables en el momento de la extracción de la muestra y que el tamaño del lote sea lo más grande posible para de este modo inspeccionar menor cantidad de elementos y así originar menores costos.

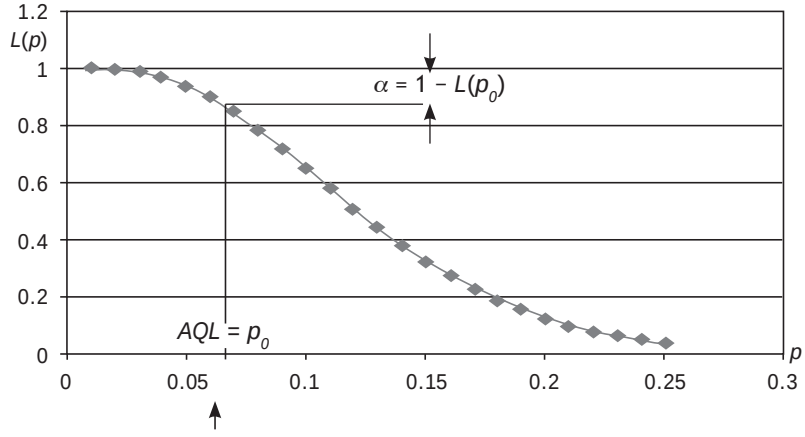
FIGURA 11.8 *Muestreo de aceptación simple*



Evaluación de un plan de muestreo

El desempeño de un plan de muestreo se evalúa calculando la probabilidad de aceptar el lote para distintos valores de la proporción p ($0 \leq p \leq 1$) de artículos defectuosos. Si denotamos con $L(p)$ a la probabilidad de aceptar el lote para p y graficamos los puntos $(p, L(p))$ se tendrá una gráfica llamada ***curva característica de operación*** (*curva OC*). Esta curva permite observar y evaluar la sensibilidad del plan frente a los diferentes niveles de elementos no conformes. Nótese que cuando $p = 0$ (el lote no contiene elementos defectuosos) la probabilidad de aceptar el lote es 1; es decir $L(0) = 1$; en cambio, si el lote contiene todos sus artículos defectuosos, $L(1)$ es 0.

FIGURA 11.9 *Curva característica de operación L(p). Curva OC*

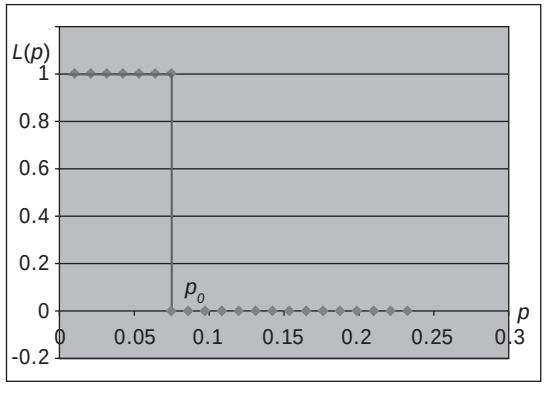


Por otro lado, el riesgo α del productor es igual a $1 - L(p_0)$, mientras que el riesgo β del consumidor es igual a $L(p_1)$, donde p_1 es un valor mayor que p_0.

El grado de control deseado sobre el lote se obtiene al observar el gráfico de la curva *OC*. Un plan de muestreo es inadecuado si la probabilidad de aceptar el lote es grande para lotes de baja calidad (*p* grande), en cambio es adecuado si la probabilidad de aceptar el lote es pequeña cuando el lote es de baja calidad.

Una situación ideal es la que se ilustra en el gráfico siguiente (Figura 11.10). Si el porcentaje de artículos defectuosos es menor o igual a p_0, el lote se acepta con probabilidad 1; mientras que si el porcentaje de artículos defectuosos es mayor que p_0, el lote no se acepta.

FIGURA 11.10 *Curva OC*



Encontrar un procedimiento que permita distinguir perfectamente los lotes malos de los lotes buenos, es decir, obtener una curva OC ideal es imposible, salvo que la inspección sea al 100% y sin errores. Sin embargo, se debe tratar de construir planes de muestreo con los cuales se tenga alta probabilidad de aceptar lotes buenos y baja probabilidad de aceptar lotes malos.

Aproximaciones del valor de $L(p)$

En general, si se tiene un lote de N artículos con una proporción p de defectuosos y un plan de *muestreo* n, c, entonces la probabilidad de aceptar el lote, $L(p)$, se calcula usando la distribución hipergeométrica, resultando:

$$L(p) = P(X \leq c) = \sum_{x=0}^{c} \frac{C_x^{N(p)} C_{n-x}^{N(1-p)}}{C_n^N}$$

En donde X indica el número de artículos defectuosos en la muestra.

Si los lotes son grandes respecto del tamaño de la muestra, la distribución hipergeométrica de X puede aproximarse con la distribución binomial con parámetros n y p, y así se tiene:

$$L(p) = P(X \leq c) = \sum_{k=0}^{c} P(X = k) = \sum_{k=0}^{c} \binom{n}{k} p^k (1-p)^{n-k}$$

Si el tamaño n de la muestra es mayor o igual que 30, usando el teorema del límite central, la distribución binomial puede aproximarse con la distribución normal, y así $L(p)$ puede escribirse como:

$$L(p) = P(X \leq c) \approx F_z\left(\frac{c + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

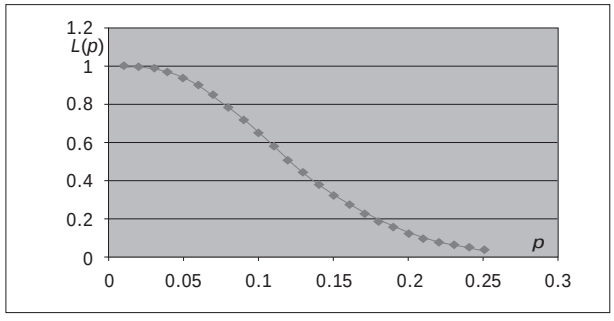
En la Tabla 11.5 se tienen valores de $L(p)$ para algunos valores de p, cuando $n = 30$ y $c = 3$. En el cálculo se han utilizado aproximaciones con la distribución binomial y con la distribución normal. Los valores en ambos casos son iguales aproximadamente.

TABLA 11.5 *Valores de L(p) para diferentes valores de p, n = 30 y c = 3*

p	*Cálculo de L(p) con la binomial*	*Cálculo de L(p) con la normal*
0.01	0.999777402	1
0.02	0.997106519	0.99992218
0.03	0.988095357	0.99730456
...	...	...
...	...	...
0.18	0.18563557	0.18328389
0.19	0.151616832	0.1529495
0.2	0.122710806	0.12691651
0.21	0.098440038	0.10472339
0.22	0.078289494	0.08592421
0.23	0.061738086	0.07009852
0.24	0.048281766	0.05685714
0.25	0.037449326	0.04584514

Usando la aproximación normal de *L*(*p*) se ha construido la gráfica *OC* que aparece a continuación (Figura 11.11).

FIGURA 11.11 *Gráfica de la función L(p). Curva OC*



Inspección con rectificación

A la inspección del 100% de los lotes rechazados para luego reemplazar los elementos defectuosos por elementos buenos se le llama ***inspección con rectificación***.

Si se lleva a cabo la inspección con rectificación y el lote tiene *M* elementos defectuosos, mientras que el número de elementos defectuosos en la muestra es *X*, la proporción de elementos defectuosos recibidos es igual a $P = (M - X)/N$.

P es una variable aleatoria cuyo valor esperado es igual a:

$$E(P) = \sum_{x=0}^{c} \left(\frac{M - x}{N} \right) P[X = x]$$

Como M no se conoce, $(M - X)/N$ se aproxima con p, y así se tiene que:

$$E(P) \approx p \sum_{x=0}^{c} P[X = x] = pL(p)$$

El valor esperado $E(P)$ se denota con $AOQ(p)$ y a la curva que la representa se le llama *curva de calidad media de salida o curva AOQ* (*Average Outgoing Quality*). Se tiene de este modo que:

$$AOQ(p) = pL(p)$$

$AOQ(p)$ es la proporción esperada de productos defectuosos por lote a la salida del control.

Usando los resultados indicados en la Tabla 11.5, si el porcentaje de artículos defectuosos fuera 18%, el número de artículos defectuosos que se espera tener por lote después de la inspección es:

$$AOQ(0.18) = 0.18(0.1856) = 0.0334, \text{ aproximadamente.}$$

Número total esperado de elementos a inspeccionar

Otra de las medidas de interés cuando se realiza un muestreo de rectificación es el número esperado de elementos a ser inspeccionados por lote. Si el lote no se rechaza el número de elementos inspeccionados es igual al tamaño de la muestra n; pero si el lote es rechazado, el número total de elementos inspeccionados es igual al número N de elementos que tiene el lote. El valor esperado del número de elementos a inspeccionar por lote se denota con ATI (*Average Total Inspection*) y es igual a:

$$ATI = nL(p) + N(1 - L(p))$$

Esta cantidad está relacionada directamente con la calidad de lote. Cuanto mayor es la proporción p de artículos defectuosos, mayor es la cantidad de lotes que deben ser inspeccionados en su totalidad. El conocimiento de esta medida proporciona información sobre el costo adicional en que se incurre con la inspección de rectificación.

Usando los resultados de la Tabla 11.5, el número esperado de artículos a inspeccionar por lote es:

$ATI = 30L(0.02) + 1000(1 - L(0.02)) = 30(0.9971) + 1000(1 - 0.997) \approx 33$, aproximadamente.

11.5 Planes de muestreo según el estándar MIL STD 105E

Para fines de aplicación práctica, existen tablas preparadas que permiten establecer el número de unidades que será preciso inspeccionar teniendo en cuenta el tamaño del lote y diferentes niveles de inspección. Entre estas tablas se encuentra la tabla llamada MIL STD 105E. Al respecto, el lector puede consultar sobre este tema en los textos especializados de control de calidad.

LA ESTADÍSTICA EN LA EMPRESA

La empresa Lavason

La empresa Lavason opera en el país hace 50 años. Fue creada por un grupo de trabajadores que egresaron del importante centro de estudios técnicos CETP, con apoyo económico del gobierno de entonces. Lavason es una empresa que fabrica y comercializa productos de línea blanca de alta calidad y de gran aceptación en el mercado nacional. Los productos que fabrica son lavadoras, cocinas eléctricas, estufas y refrigeradoras.

La inversión constante en la compra de tecnología de punta para la fabricación ha permitido la evolución de Lavason y ha hecho que se consolide como líder en la categoría.

Los productos que Lavason fabrica están sometidos a procesos cuidadosos y controles estrictos de calidad que van desde el diseño hasta los servicios de posventa. Las estrategias de comercialización adecuada que esta empresa realiza han sido el producto de una serie de análisis estadísticos del mercado, entre los que destaca la segmentación del mercado a nivel nacional, permitiendo además la cobertura de un alto porcentaje del mercado de la línea blanca.

Las características de los productos que la empresa Lavason fabrica y comercializa han hecho que esta sea reconocida como una marca confiable y de alto servicio técnico.

El equipo de investigación estadística para el control de calidad de los productos tiene muy en cuenta los gustos y preferencias del público en el diseño de los productos, lo que conjuntamente con la red de servicios técnicos diseminados en todo el país y con un gran stock de repuestos garantiza la alta confiabilidad de los productos vendidos.

EJERCICIOS

1. Una empresa de alimentos envasa arroz en bolsas de papel. Para controlar el proceso de pesado se toman muestras de 5 bolsas cada una y se registra el promedio y el rango de cada muestra. Los resultados fueron como sigue.

Muestra	*Media*	*Rango*	*Muestra*	*Media*	*Rango*	*Muestra*	*Media*	*Rango*
1	33.0	4.1	2	30.9	4.8	3	29.8	5.1
4	32.0	4.2	5	34.1	3.2	6	32.0	5.4
7	33.0	3.8	8	31.7	4.5	9	34.8	4.7
10	34.7	4.5	11	34.5	3.1	12	32.5	4.8
13	34.8	3.2	14	33.4	2.7	15	30.6	3.9
16	32.9	4.3	17	31.9	3.4	18	33.2	4.1
19	34.1	4.1	20	38.2	3.9			

Indicar si el proceso está bajo control.

2. Una empresa de productos de limpieza controla el volumen de llenado de detergente en cajas usando 15 lotes de 5 cajas cada uno. Los datos recogidos fueron los siguientes.

	Muestras					Media	Rango
1.	5.07	4.92	5.05	5.15	4.96	5.030	0.230
2.	5.09	4.99	5.12	4.89	5.10	5.038	0.230
3.	4.87	4.95	5.05	5.07	5.00	4.988	0.200
4.	4.90	5.09	5.10	5.02	5.15	5.052	0.250
5.	5.00	5.07	5.12	4.93	5.12	5.048	0.190
6.	4.95	5.13	5.15	5.17	5.10	5.100	0.220
7.	5.00	4.90	4.92	4.89	4.78	4.898	0.220
8.	4.85	4.86	4.87	4.88	4.95	4.882	0.100
9.	4.95	4.96	4.97	4.99	4.89	4.952	0.100
10.	4.90	4.91	4.87	4.86	4.93	4.894	0.070
11.	5.00	5.01	5.02	5.01	5.04	5.016	0.040
12.	4.95	5.04	5.05	5.02	5.04	5.020	0.100
13.	5.01	5.04	5.03	5.06	5.02	5.032	0.050
14.	5.00	4.92	4.95	4.99	5.03	4.978	0.110
15.	4.99	4.97	4.98	5.01	5.05	5.000	0.080

Indicar si el proceso está bajo control.

3. Los siguientes datos corresponden a las desviaciones del diámetro de los cojinetes esféricos de la longitud que deben tener para su venta. Las unidades han sido medidas en 0.001 pulgadas.

1.	2.1	0.4	1.9	1.4	–1.9
2.	0.8	0.3	0.5	1.1	1.3
3.	1.5	1.2	1.2	0.9	0.6
4.	0.2	–0.6	2.0	0.0	–1.8
5.	0.8	–1.5	0.9	1.4	1.3
6.	1.1	0.8	1.5	–1.5	1.2
7.	0.5	1.3	1.2	0.7	–2.1
8.	1.1	1.0	1.5	1.3	0.3
9.	0.1	1.5	0.3	0.5	–1.0
10.	–2.0	–1.5	–0.5	1.6	2.1

Indicar si el proceso está bajo control.

4. Con la finalidad de mejorar la efectividad de un proceso de entregas se construye una carta de control para monitorear el porcentaje de entregas que no se hacen a tiempo. Durante 20 días se observaron 100 entregas diarias y se anotaron las entregas realizadas en un tiempo mayor al previsto. Los resultados fueron como sigue.

Día	*Entregas a destiempo*
1	4
3	3
5	4
7	3
8	5
11	2
13	2
15	5
17	1
19	2

Día	*Entregas a destiempo*
2	4
4	10
6	11
8	9
9	10
12	5
14	1
16	14
18	10
20	9

Indicar si el proceso está bajo control.

5. El siguiente cuadro contiene las cantidades de refresco, en litros, depositado en botellas de vidrio para ser vendidas al público.

Muestras					*Media*	*Rango*
0.93	0.91	1.06	1.04	1.09	**1.0068**	**0.18**
0.99	0.95	1.02	1.03	1.05	**1.0070**	**0.10**
1.02	1.00	1.10	1.06	1.02	**1.0378**	**0.09**
1.02	1.00	1.10	1.06	1.02	**1.0378**	**0.09**
0.98	0.93	1.00	1.03	1.09	**1.0035**	**0.16**
1.06	1.10	0.92	1.00	1.00	**1.0160**	**0.19**
1.02	1.01	0.97	0.96	1.12	**1.0139**	**0.16**
1.02	0.98	0.96	1.00	1.05	**1.0011**	**0.09**
1.08	0.98	0.96	1.01	0.98	**1.0019**	**0.12**
1.00	0.96	0.98	1.00	1.03	**0.9942**	**0.07**
0.96	1.00	1.02	1.00	1.00	**0.9941**	**0.07**
1.06	1.05	1.01	0.91	0.92	**0.9909**	**0.15**
1.01	1.07	0.95	0.97	0.98	**0.9951**	**0.12**
1.01	0.94	1.07	1.02	0.971	**1.0023**	**0.11**
0.97	1.03	1.08	1.00	1.031	**1.0204**	**0.13**
1.01	0.98	0.98	0.98	1.121	**1.0121**	**0.14**
1.08	0.96	0.90	0.98	0.98	**0.9803**	**0.18**
1.04	1.05	0.85	0.99	0.97	**0.9817**	**0.20**
1.05	1.02	1.05	0.98	0.931	**0.0043**	**0.13**
1.04	0.96	0.89	0.98	1.01	**0.9758**	**0.15**

Indicar si el proceso está bajo control.

6. Mediante un proceso de fabricación se obtienen piezas de rodamiento. El peso en gramos de cada pieza es controlado antes de que salga a la venta. Se registraron las medias y rangos de 20 muestras de 5 rodamientos cada una.

Grupo muestral	*Media*	*Rango*
1	1569	34
2	1557	24
3	1548	23
4	1559	38
5	1572	28
6	1559	20
7	1558	36
8	1539	36
9	1559	27
10	1554	18

Grupo muestral	*Media*	*Rango*
11	1567	35
12	1552	35
13	1572	28
14	1574	21
15	1556	23
16	1568	40
17	1548	35
18	1563	30
19	1557	24
20	1547	29

Indicar si el proceso está bajo control.

7. Después del proceso de pintado de automóviles en una empresa automotriz se seleccionan 20 automóviles con la finalidad de inspeccionar y registrar el número de rasguños, manchas y otros defectos en cada uno de ellos. Con los datos registrados, construir una gráfica de control del número de defectos por unidad. Indicar si el proceso está bajo control.

Auto	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Número de defectos	2	4	1	3	4	6	5	4	5	8	3	5	6	2	10	6	1	5	9	4

8. Para controlar el número de defectos por unidad de una tela fabricada, una empresa textil anota el número de defectos (D) que aparecen en rollos de tela de diferente superfície (S), en metros cuadrados. Los datos registrados fueron como sigue.

Rollo	***S***	***D***
1	30	8
2	30	10
3	30	6
4	30	7
5	30	4
6	30	5
7	30	14
8	30	6
9	30	15
10	30	12
11	30	20
12	30	7
13	30	8

Rollo	***S***	***D***
14	30	11
15	30	13
16	30	7
17	30	11
18	30	9
19	30	14
20	30	9
21	30	8
22	30	6
23	30	10
24	30	9
25	30	8

Considerando que la unidad es el metro cuadrado de tela, construir un gráfico de control de unidades defectuosas por unidad.

9. Con la finalidad de estandarizar el tiempo de espera previa a la atención en un hospital, la dirección ha indicado que este debería ser menor o igual a 15 minutos. Para vigilar el proceso se han recogido los tiempos de espera que corresponden a 20 muestras conformadas por 100 pacientes cada una. El número de pacientes *X* por muestra cuyos tiempos de atención eran mayores que 15 minutos se indican a continuación.

Muestra	***X***
1	4
3	3
5	5
7	8
8	2
11	3
13	6
15	7
17	9
19	5

Muestra	***X***
2	5
4	10
6	6
8	7
9	3
12	7
14	4
16	14
18	12
20	8

Construir el gráfico de control apropiado e indicar si el proceso está bajo control.

10. Un proceso para fabricar objetos cuyo peso debe estar entre 1.95 y 2.05 kilogramos está bajo control. Si la media y la varianza del proceso son 2.00 y 0.04, respectivamente, y el modelo que subyace es el de la distribución normal, indicar si el proceso es capaz.

11. Si en el proceso anterior la media es 2.04, indicar si el proceso está bajo control. Calcular el porcentaje de objetos que no satisfacen las especificaciones.

12. Las estimaciones de la media y la desviación estándar de un proceso son, respectivamente, 13.60 y 0.091, y los límites de especificación son *LIE* = 13.50 y *LSE* = 14.00. Hallar la capacidad del proceso.

13. Calcular los índices C_p y C_{pk} que corresponden a un proceso bajo control, cuyas estimaciones de la media y la desviación estándar son 3.99 y 0.008, respectivamente. Los límites de especificación son: *LIE* = 3.95 y *LSE* = 4.05.

14. Si el índice C_p de un proceso centrado es 0.2, hallar el porcentaje de unidades que no satisfacen las especificaciones. Indicar las unidades por millón que no satisfacen las condiciones.

15. Construir una tabla que indique el porcentaje de unidades que no satisfacen las especificaciones y el número de productos por millón que no satisfacen las especificaciones cuando los valores de C_p son:

    0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9 2.0

16. El porcentaje de productos defectuosos en un proceso centrado es 2.5%. Estimar el índice C_p del proceso.

17. Para decidir si se acepta o no un lote de 10,000 tarjetas de sonido, se acordó utilizar una inspección por muestreo, tomando 100 tarjetas al azar, y rechazar el lote si el número de tarjetas defectuosas en la muestra es mayor que 3. Hallar el riesgo del productor y el riesgo del consumidor si se considera que el leal *AQL* es 3% y el *LTPD* es 5%.

18. Para un plan de muestreo en donde $n = 10$ y $c = 2$, hallar el riesgo del receptor si se considera que la calidad es aceptable cuando la proporción de artículos defectuosos contenida en el lote es del 3%. Se considera que el lote es grande.

19. A través de su departamento de compras, la empresa de telefonía Movilsa adquiere de su proveedor lotes de 10,000 unidades de pantallas LCD para teléfonos móviles. Si el nivel de calidad aceptable fijado es de 1% y el plan de muestreo se considera que corresponde a la distribución normal, hallar el número de pantallas a inspeccionar y el número de pantallas defectuosas que deberán ser considerados para rechazar el lote si se fijan como riesgos $\alpha = 0.05$ y $\beta = 0.01$.

20. a) En el problema anterior, usando la aproximación normal, graficar la curva OC para determinar la bondad del plan de muestreo.
    b) Si la inspección se realiza con rectificación, dibujar la curva que indica la calidad media de los lotes que se aceptan cuando se aplica este plan de muestreo.
    c) Si la inspección se realiza con rectificación, ¿cuál es el promedio de tarjetas que se deben inspeccionar por lote?

RESPUESTAS A LOS EJERCICIOS

1. Para la media: *LIC* = 30.7450, *LC* = 33.1050, *LSC* = 35.4649 Para el rango: *LIC* = 0, *LC* = 4.09, *LSC* = 8.6504.
2. Para la media: *LIC* = 5.2326, *LC* = 5.3267, *LSC* = 5.4088 Para el rango: *LIC* = 0, *LC* = 0.1527, *LSC* = 0.3229.
3. Para la media: *LIC* = −1.0043, *LC* = 0.54, *LSC* = 2.0863 Para el rango: *LIC* = 0, *LC* = 2.68, *LSC* = 5.6682.
4. Para la proporción: *LIC* = 0, *LC* = 0.057, *LSC* = 0.1265 **7.** *LIC* = 0, *LC* = 4.65, *LSC* = 11.1192 **8.** *LIC* = 0.0081, *LC* = 0.3160, *LSC* = 0.6239. **10.** C_p = 0.0833. **11.** C_{pk} = 0.0166. **12.** C_{pk} = 0.3663. **14.** 54.86%

15.

Cp	*Partes por millón que no satisfacen*
0.3	368120.2507
0.4	230139.3404
0.5	133614.4025
0.6	71860.63823
0.7	35728.84113
0.8	16395.07185
0.9	6933.947606
1	2699.796063
1.1	966.8482848
1.2	318.2171803
1.3	96.19268804
1.4	26.69149805
1.5	6.795346267
1.6	1.586656083
1.7	0.339653481
1.8	0.066640897
1.9	0.011980743
2	0.001973175

17. El riesgo del productor es 0.2789. **18.** El riesgo del receptor es 0.9972.

CAPÍTULO 12

Introducción a la teoría de decisiones

El empresario y la toma de decisiones

Una de las tareas primordiales del empresario es la toma efectiva de decisiones. Este proceso se hace cada vez más difícil y delicado a medida que el empresario avanza dentro de la empresa. Los tiempos en que el empresario solo utilizaba su olfato para la toma de decisiones están pasando. El entorno en que se desarrolla la empresa es cada vez más complicado; la cantidad de información que se genera hoy día y que es imprescindible para tomar una decisión que lleve al éxito a una empresa debe ser analizada con técnicas avanzadas como las que proporciona la estadística.

Precisamente el pensamiento estadístico indica la ruta que Peter Drucker señaló en su obra *La decisión efectiva para la toma efectiva de decisiones*: a) definición, importancia y delimitación del problema, b) análisis de las soluciones, c) la toma de la decisión, d) ejecución y control de la decisión tomada y e) retroalimentación de la información.

En este capítulo se presentan, de forma introductoria y a manera de ejemplos, principios aplicables a situaciones en donde deba tomarse una decisión, sin pretender indicar que se utilicen en todas las situaciones.

CONTENIDO

12.1 Introducción

En este capítulo se estudian algunas técnicas que tienen que ver con la elección de una de varias estrategias para resolver situaciones que involucran eventos de carácter incierto. Un ejemplo que se desarrolla en este marco se refiere al problema que enfrenta una persona que debe decidir si invierte o no en bonos de largo plazo cuyo rendimiento en el futuro es incierto. Otro ejemplo está referido a la decisión que debe tomar un consultor que ha sido invitado a presentar una propuesta para la realización de un proyecto de transporte. El consultor debe presentar además una oferta económica para obtener el contrato respectivo. Esta acción se desarrolla en un ambiente incierto, al no saber el consultor cuál debe ser la cuantía de la oferta a presentar. Se trata de situaciones en donde el resultado de una decisión individual depende de situaciones que suelen llamarse *estados de la naturaleza*, y sobre los cuales no se tiene control.

A diferencia de los problemas clásicos de optimización, los problemas que se tratan en la teoría de decisión no contemplan una función objetivo de carácter formal que deba ser optimizada. A cambio de ello se conocen las retribuciones asociadas (o pérdidas asociadas) a las posibles acciones y estados de la naturaleza.

EJEMPLO. *¿Habrá concierto mañana?*

Para decidir si se lleva a cabo o no (decisiones: *llevar a cabo, no llevar a cabo*) un concierto al aire libre el día de mañana, un empresario de espectáculos se ha visto precisado a considerar la posibilidad de que llueva o no llueva (estados de la naturaleza: *llueve, no llueve*). El servicio meteorológico ha informado que la probabilidad de que llueva mañana es 0.3 (*distribución de los distintos estados de la naturaleza*).

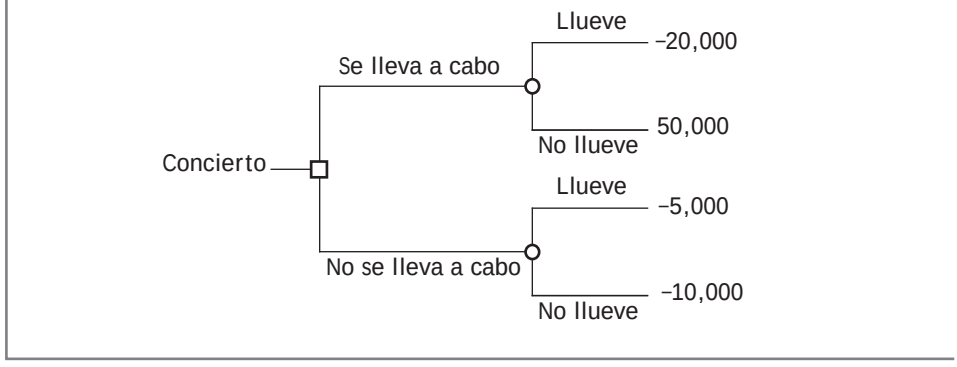
Las retribuciones o ganancias, en dólares, que podría tener la empresa aparecen en la Tabla 12.1, en donde las retribuciones con signo negativo significan pérdidas.

TABLA 12.1 *Tabla de retribuciones en dólares*

Decisión	*Estados de la naturaleza*	
	Llueve	*No llueve*
Se lleva a cabo	−20,000	50,000
No se lleva a cabo	−5,000	−10,000

Para la fácil comprensión del problema y los cálculos respectivos, se acostumbra a usar los *árboles de decisión*. Un árbol de decisión está formado por ramas que unen nodos de decisión (estos indican que se toma una decisión) o nodos de probabilidad (estos indican la ocurrencia de un evento aleatorio). Los nodos de decisión se representan con rectángulos, mientras que los nodos de probabilidad se representan con círculos

FIGURA 12.1 *Árbol de decisión*



El proceso de decisión es como sigue:

1. El decisor toma una de las alternativas.
2. Habiéndose tomado la decisión de elegir una alternativa, ocurre un estado de la naturaleza que está fuera de control del decisor.
3. Habiéndose tomado la decisión y ocurrido un estado de la naturaleza, se obtendrá un rendimiento o retribución.

Al decisor le gustaría obtener el mayor rendimiento, y para ello deberá elegir una de las alternativas. Por ejemplo, llevar a cabo el concierto. Dependiendo del estado de la naturaleza obtendrá una retribución o ganancia. Si el estado de la naturaleza es *llueve*, el decisor tendrá una pérdida de 20,000 dólares, pero si *no llueve*, el decisor tendrá una ganancia de 50,000 dólares. Si la decisión es no llevar a cabo el concierto y llueve, el decisor tendrá una pérdida de 5,000 dólares, pero si no llueve, el decisor tendrá una pérdida de 10,000 dólares.

El problema ahora es ¿cuál decisión tomar? Los procedimientos que siguen ayudan a la toma de la mejor decisión.

12.2 Toma de decisiones bajo incertidumbre

Cuando el decisor no tiene conocimiento de la distribución de los estados de la naturaleza se usan los siguientes criterios.

El ***criterio maximin*** consiste en determinar el valor mínimo en la tabla de retribuciones que resulta de la ejecución de cada acción, para luego elegir como *mejor acción* aquella cuya resultante es la máxima entre todas las retribuciones mínimas.

Este método es típico de los conservadores, que se preocupan "por lo peor que puede pasar" con respecto a cada acción y que no se permiten un error.

En el ejemplo relacionado con el concierto, para elegir la decisión a tomar, según este criterio, se considera:

1. La retribución mínima que se obtendría si se toma la acción de llevar a cabo el concierto. Esta es –20,000 dólares.
2. La retribución mínima que se obtendría si se toma la acción de no llevar a cabo el concierto. Esta es –10,000 dólares.
3. De las dos retribuciones mínimas, la mayor es –10,000 dólares, que se obtiene cuando se toma la decisión de no llevar a cabo el concierto.
4. La decisión óptima, según el criterio maximin, es *no llevar a cabo el concierto*.

El ***criterio máximax*** consiste en determinar el valor máximo en la tabla de retribuciones que resulta de la ejecución de cada acción para luego elegir como *mejor acción* aquella cuya resultante es máxima entre todas las retribuciones máximas.

En oposición al criterio del maximin, los que usan el criterio máximax se preocupan "por lo mejor que puede suceder".

En el ejemplo relacionado con el concierto, para elegir la decisión a tomar, según este criterio, consideramos:

1. La retribución máxima que se obtendría si se toma la acción de llevar a cabo el concierto. Esta es 50,000 dólares.
2. La retribución máxima que se obtendría si se toma la acción de no llevar a cabo el concierto. Esta es –5,000 dólares.
3. De las dos retribuciones máximas, la mayor es 50,000 dólares, que se obtiene cuando se toma la decisión de llevar *a cabo el concierto*.
4. La decisión óptima, según el criterio del máximax, es llevar a cabo el concierto.

12.3 Toma de decisiones bajo riesgo

El tipo de decisiones al que este título se refiere se aplica a situaciones en donde el decisor puede estimar la probabilidad de la ocurrencia de los distintos estados de la naturaleza.

La estrategia que se usa en estos casos es *la del mejor valor esperado*. Según esta estrategia, la mejor decisión se obtiene calculando el *valor esperado de la retribución* para cada una de las decisiones a_k, para luego tomar la decisión que produce el valor óptimo entre los valores monetarios esperados. Al valor óptimo que se obtiene se le denota con *VME**.

Para el ejemplo del concierto, si la probabilidad de que llueva es 0.3 (la probabilidad de que no llueva es 0.7), entonces:

para la decisión a_1: *llevar a cabo el concierto*, el valor esperado, $VE(a_1)$, es:

$$(-20{,}000.00)(0.3) + (50{,}000.00)(0.7) = 29{,}000.00 \text{ dólares, y}$$

para la decisión a_2: *no llevar a cabo el concierto*, el valor esperado, $VE(a_2)$, es:

$$(-5{,}000.00)(0.3) + (-10{,}000.00)(0.7) = -15{,}000.00 \text{ dólares}$$

El óptimo valor esperado sucede para la decisión a_1: llevar a cabo el concierto. Se tiene de este modo que *VME** = 29,000.00 dólares.

EJEMPLO. *¿Fabricar o comprar?*

Un fabricante tiene la oportunidad de producir 100,000 piezas para computadora con un costo de preparación de $ 60,000 más $ 20 por costos fijos por cada una, o bien comprarlas en $ 25 cada una. Las piezas compradas siempre son buenas, en cambio existe la posibilidad de que las piezas hechas por el fabricante sean defectuosas. La información que se tiene al respecto es que puede existir: el 0%, el 10% o el 15% de piezas defectuosas con probabilidades: 0.1, 0.6 y 0.3, respectivamente. Cuando se instala una pieza fabricada y se descubre que es defectuosa, esta deberá ser corregida, produciéndose un costo adicional de $ 10. ¿Cuál es la mejor decisión, fabricar o comprar las piezas?

Solución

Si, por ejemplo, el 15% de piezas fabricadas son defectuosas, entonces, el costo por fabricar 100,000 piezas es:

$$60{,}000 + 20(100{,}000) + (10)(15/100)\ (100{,}000) = 2{,}210{,}000 \text{ dólares}$$

El costo por comprar las 100,000 piezas es 25(100,000) = 2,500,000 dólares, cualquiera que sea el estado de la naturaleza.

Para cada decisión y cada porcentaje de artículos defectuosos, el costo para producir 100,000 piezas, cualquiera que sea el estado de la naturaleza, aparece en la siguiente tabla (12.2) de retribuciones.

TABLA 12.2 *Tabla de costos*

Decisión	*Estados de la naturaleza: porcentaje de artículos defectuosos*		
	s_1: 0%	s_2: 10%	s_3: 15%
a_1: Fabricar	1,260,000	2,160,000	2,210,000
a_2: Comprar	2,500,000	2,500,000	2,500,000

El valor esperado del costo por fabricar las 100,000 piezas es:

(1,260,000)(0.1) + (2,160,000)(0.6) + (2,210,000)(0.3) = 2,085,000 dólares

La decisión óptima corresponde a la decisión "fabricar", pues con esta decisión se obtiene el menor costo esperado. Se tiene que *VME** = 2,085,000 dólares.

EJEMPLO. *¿Se automatiza o no la fabricación?*

En un proceso industrial se fabrican 1,000 motores. Para una mejor performance de estos aparatos deben hacerse ajustes finales. Cuando los ajustes se realizan manualmente la proporción de motores defectuosos que resultan puede ser 0.01, 0.05, 0.10 o 0.20. Experiencias anteriores indican que la distribución de estas proporciones es 0.40, 0.30, 0.20 y 0.10, respectivamente. Existe la posibilidad de realizar los ajustes de manera electrónica, a un costo adicional de $ 10,000 por cada 1,000 motores; en tal caso la calidad se mantendría a un nivel del 1% de defectuosos.

Estudiar la conveniencia o no de establecer la automatización de la fabricación si cuando un motor resulta defectuoso este debe ser reparado con un costo de $ 150.00.

Solución

Estudiaremos el problema a partir de la tabla de retribuciones (12.3) que aparece a continuación, en donde también se indican las distribuciones de los diferentes estados.

TABLA 12.3 *Tabla de costos*

	Estados naturales: proporción de artículos defectuosos	s_1: 0.01	s_2: 0.05	s_3: 0.10	s_4: 0.20
	Probabilidades	0.40	0.30	0.20	0.10
Decisión	a_1: Ajuste manual	1,500	7,500	15,000	30,000
	a_2: Ajuste electrónico	11,500	11,500	11,500	11,500

Bajo la decisión a_1: ajuste manual, el costo esperado es:

(1,500)(0.40) + (7,500)(0.30) + (15,000)(0.20) + (30,000)(0.10) = 8,850 dólares

Bajo la decisión a_2: ajuste electrónico, el costo esperado es 11,500 dólares.

La decisión óptima es la decisión a_2: ajuste electrónico, pues a esta corresponde el menor costo esperado.

12.4 El valor esperado de la información perfecta

Supongamos ahora que el decisor tiene la opción de comprar información que le indicará cuál es el real estado de la naturaleza. La pregunta es ¿cuál es la cuota más alta que el decisor está dispuesto a pagar para obtener esa información?

La respuesta se obtiene comparando el mejor valor esperado (*VME**), obtenido "sin información perfecta", con la retribución esperada que el decisor obtendría al recibir la información "perfecta" (*VECIP*).

El valor a pagar se llama valor esperado de la información perfecta, se denota con *VEIP* y se calcula con la diferencia *VECIP* − *VME**. Se tiene que:

$$VEIP = VECIP - VME^*$$

(Nótese que si se conoce el estado de la naturaleza, es obvio que la decisión a tomar deberá ser la que produzca la mayor retribución. Sin embargo, al no saberse cuál es el estado real, hasta no recibirse el pago por la información recibida, el valor *VECIP* será igual al valor esperado de las máximas retribuciones.)

Para ilustrar este concepto se considera que existe la posibilidad de invertir en el sector hotelería o en el sector de transporte. Existe también la posibilidad de que el gobierno dicte una ley que influirá en el desarrollo de los dos sectores. Los inversores han calculado las retribuciones que se producirían si se promulga o no la ley. Estas retribuciones, en millones de dólares, así como la distribución de los diferentes estados, aparecen en la Tabla 12.4.

TABLA 12.4 *Tabla de retribuciones*

	Estados de la naturaleza	
	s_1: Se da la ley	s_2: No se da la ley
Probabilidades de los estados naturales	0.60	0.40
a_1: Inversión en hotelería	4	2
a_2: Inversión en transporte	5	−1

Si se invirtiera en el sector hotelería, la retribución esperada sería:

$$(4)(0.60) + (2)(0.40) = 3.20 \text{ millones de dólares.}$$

Si se invirtiera en el sector transporte, la retribución esperada sería:

$$(5)(0.60) + (-1)(0.40) = 2.60 \text{ millones de dólares.}$$

Usando el criterio del mejor valor esperado, se encuentra que la decisión óptima será invertir en el sector hotelería. En este caso se obtiene $VME^* = 3.20$ millones de dólares.

Se supone ahora que existe la posibilidad de conseguir un informante, quien le podrá decir al inversor si se dará o no la ley, pero al cual deberá pagársele una cierta cantidad por la información. La pregunta es ¿hasta cuánto estaría dispuesto a pagar el inversor? ¿Valdrá la pena pagar por la información?

Es claro que el informante solo será tal, no podrá cambiar el estado natural. Si se diera la ley, él no podrá hacer nada para que no se dé. Visto esto, si el informante indicara que se dará la ley, el decisor invertirá en transporte, porque de esta manera obtendrá la mayor retribución, mientras que si se indica que no se dará la ley, la inversión será en hotelería.

Como el informante no proporcionará la información a menos que le paguen, al decisor solo le queda calcular el valor esperado de la retribución que obtendría a partir de la información. Esto es:

$$VECIP = (5)(0.60) + (2)(0.40) = 3.80 \text{ millones de dólares}$$

Luego, el mayor valor a pagar por la información será:

$$VEIP = VECIP - VME^* = 3.80 - 3.20 = 0.60 \text{ millones de dólares}$$

12.5 Toma de decisiones usando información muestral

La calidad de las decisiones puede mejorarse si se actualizan las probabilidades de los estados naturales que a priori se tienen. Esta actualización puede hacerse con muestras aleatorias o con pruebas adicionales. A la luz de la información que las muestras o las pruebas proporcionan, y usando el teorema de Bayes, se calculan las nuevas probabilidades asociadas a los estados naturales (probabilidades a posteriori).

Usando las probabilidades a posteriori se calcula la retribución esperada que el decisor obtendría, y que proviene de las muestras o de las pruebas. Para diferenciarla del valor esperado de la información sin muestreo, se usa la notación $VECIP_+$.

La diferencia $VECIP_+ = VECIP_+ - VME^*$ corresponde a la cantidad máxima que en promedio debería pagar el agente decisor para obtener la información muestral o la prueba adicional.

EJEMPLO. *El caso de la tienda Riley*

La empresa ADMR es la encargada de administrar una cadena de tiendas de ropa para caballeros, entre las que está la tienda Riley. Esta tienda ha venido teniendo un éxito en las ventas; sin embargo, a raíz del anuncio de la crisis financiera, sus ventas se han debilitado últimamente a tal punto que hay incertidumbre acerca de lo que sucederá en la próxima temporada veraniega.

El gerente de ventas tiene que decidir el número de lotes de pantalones de verano que debe solicitar para la venta, lo que dependerá de si la demanda es alta o baja. El administrador puede ordenar 1 o 2 lotes al comienzo de la estación. Las proyecciones de las ganancias, en miles de dólares, se muestran en la Tabla 12.5.

A priori, con la experiencia del administrador y la información de los economistas, se ha determinado que las probabilidades para la demanda son: 0.3 para la demanda alta y 0.7 para la demanda baja. Sin embargo, con la finalidad de obtener una buena decisión, existe la posibilidad de llevar a cabo una investigación de mercado para conocer el estado de la demanda. La investigación sería realizada por una empresa de investigación de mercados que cobra 4,000 dólares y que tiene suficiente experiencia, a tal punto que en situaciones parecidas acertó en sus resultados en el 90% cuando la demanda fue alta y en 75% cuando la demanda fue baja.

La tabla de retribuciones, de acuerdo a la información descrita, es como sigue.

TABLA 12.5 *Tabla de retribuciones en miles de dólares y con probabilidades a priori*

	Estados de la naturaleza	
	s_1: Alta demanda	***s_2: Baja demanda***
Probabilidad de los estados de la naturaleza	0.30	0.70
Decisión a_1: Ordenar un lote	70	40
Decisión a_2: Ordenar dos lotes	90	35

1. Si el administrador de ADMR solo tomara en cuenta las probabilidades a priori y ordenara un lote, el valor esperado de la retribución sería:

$$(70)(0.3) + (40)(0.7) = 49.00 \text{ miles de dólares}$$

Si el administrador de ADMR solo tomara en cuenta las probabilidades a priori y ordenara dos lotes, el valor esperado de la retribución sería:

$$(90)(0.3) + (35)(0.7) = 51.50 \text{ miles de dólares}$$

De acuerdo a la estrategia del mejor valor esperado, el administrador deberá ordenar dos lotes. En este caso, el valor *VME** = 51.50 miles de dólares.

2. Para analizar si conviene o no utilizar la información de la investigación de mercado, se puede proceder de la siguiente manera. De acuerdo a los resultados y para simplificar las notaciones, se escribe:

   a_1 = Ordenar un lote

   a_2 = Ordenar dos lotes

   s_1 = Demanda alta

   s_2 = Demanda baja

   F = Informe acertado de la investigación de mercado

   D = Informe no acertado de la investigación de mercado

A partir de la información que proporciona la investigación de mercado y usando el teorema de Bayes, calcular las probabilidades revisadas o a posteriori de los estados de la naturaleza.

Si el informe de la investigación es acertado (F), entonces las probabilidades a posteriori son:

$$P(s_1|F) = \frac{P(F|s_1)P(s_1)}{P(F)} = \frac{(0.90)(0.30)}{0.795} = 0.3396 \text{ y}$$

$$P(s_2|F) = \frac{P(F|s_2)P(s_2)}{P(F)} = \frac{(0.75)(0.70)}{0.795} = 0.6604$$

En donde la probabilidad de que el informe sea acertado se calcula como:

$$P(F) = P(F|s_1)P(s_1) + P(F|s_2)P(s_2) = (0.90)(0.30) + (0.75)(0.70) = 0.795$$

TABLA 12.6 *Tabla de retribuciones en miles de dólares y con probabilidades* ***a posteriori***

Si el informe es acertado	***Estados de la naturaleza***	
	s_1: Alta demanda	***s_2: Baja demanda***
Probabilidad ***a posteriori*** de los estados de la naturaleza	0.3396	0.6604
Decisión a_1: Ordenar un lote	70	40
Decisión a_2: Ordenar dos lotes	90	35

El valor esperado para la decisión a_1, ordenar un lote, es:

$$(70)(0.3396) + (40)(0.6604) = 50.19 \text{ miles de dólares}$$

El valor esperado para la decisión a_2, ordenar dos lotes, es:

$$(90)(0.3396) + (35)(0.6604) = 56.68 \text{ miles de dólares}$$

Se observa que el mejor valor esperado (56.68 miles de dólares) se logra ordenando dos lotes.

Si el informe de la investigación no es acertado (D), entonces las probabilidades a posteriori son:

$$P(s_1|D) = \frac{P(D|s_1)P(s_1)}{P(D)} = \frac{(0.10)(0.30)}{0.205} = 0.1463 \text{ y}$$

$$P(s_2|D) = \frac{P(D|s_2)P(s_2)}{P(D)} = \frac{(0.25)(0.70)}{0.205} = 0.8537$$

La probabilidad de que el informe de la investigación no sea acertado se calcula con:

$$P(D) = P(D|s_1)P(s_1) + P(D|s_2)P(s_2) = (0.10)(0.30) + (0.25)(0.70) = 0.205$$

TABLA 12.7 *Tabla de retribuciones en miles de dólares y con probabilidades a posteriori*

Si el informe es acertado	***Estados de la naturaleza***	
	s_1: Alta demanda	***s_2: Baja demanda***
Probabilidad ***a posteriori*** de los estados de la naturaleza	0.1463	0.8537
a_1: Ordenar un lote	70	40
a_2: Ordenar dos lotes	90	35

El valor esperado para la decisión a_1, ordenar un lote, es:

$$(70)(0.1463) + (40)(0.8537) = 44.39 \text{ miles de dólares.}$$

El valor esperado para la decisión a_2, ordenar dos lotes, es:

$$(90)(0.1463) + (35)(0.8537) = 43.05 \text{ miles de dólares.}$$

Se observa que el mejor valor esperado (44.39 miles de dólares) se logra ordenando un lote.

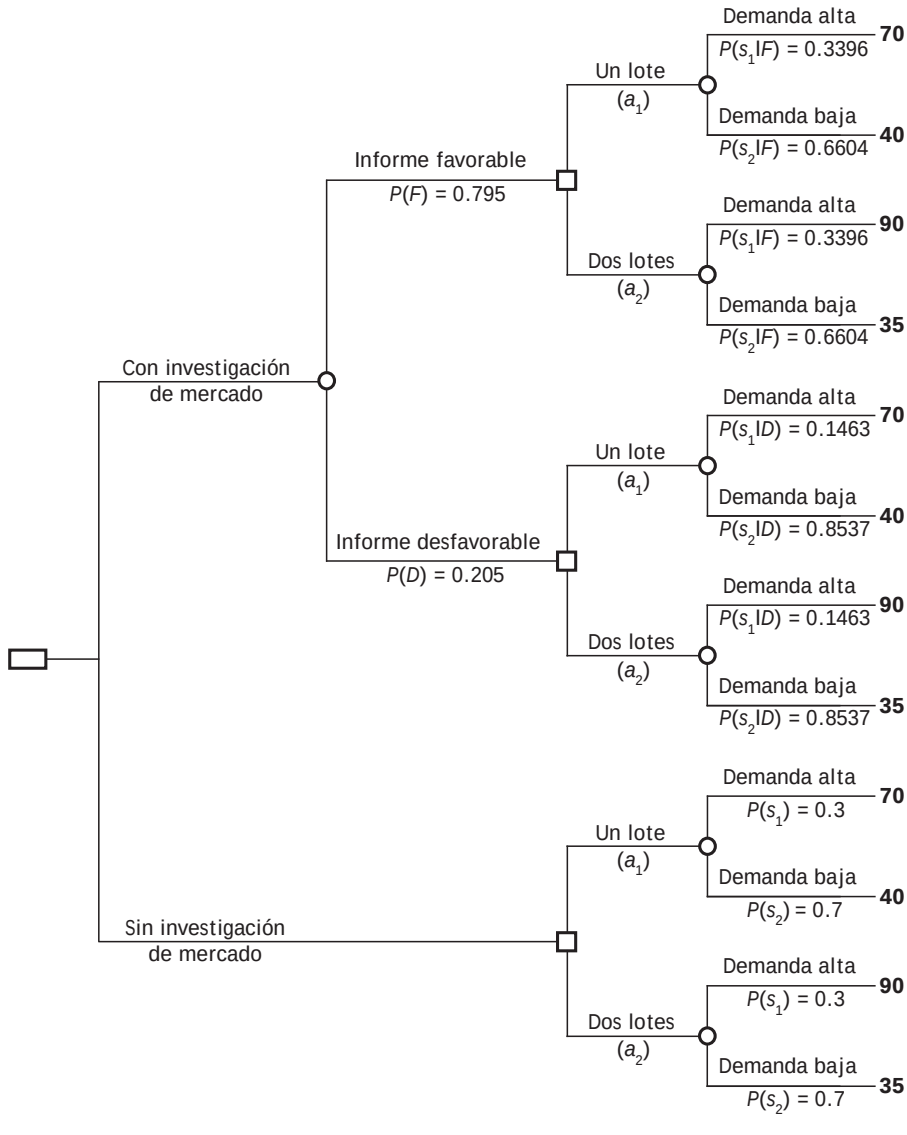
Así se tiene que, cuando se utiliza la información que brinda la encuestadora, la rentabilidad neta esperada es:

$$\text{VECIP}_+ - 4 = (0.795)(56.68) + (0.205)(44.39) - 4 = 50.16 \text{ miles de dólares}$$

Como resultado, la administradora no deberá usar el estudio de mercado, pues sin este obtendrá una mayor rentabilidad esperada (51.50 miles de dólares).

También se puede construir un árbol de decisión indicando las probabilidades en las ramas que corresponden a los estados de la naturaleza (demanda alta o demanda baja). Esta gráfica (Figura 12.2) servirá como un resumen de lo desarrollado.

FIGURA 12.2 *Árbol de decisión con las probabilidades en las ramas de estado*



APLICACIÓN: El caso de la empresa editora Editesa

La empresa editora Editesa tiene a su cargo la edición de textos escolares para la región sur, y viene operando desde hace algunos años con una máquina impresora grande. Ante la posibilidad de expandir su negocio a otros lugares, quisiera cambiar la impresora por otra nueva de mayor velocidad de impresión. Sin embargo, el papel que se consigue en el mercado no siempre es de buena calidad. Si el papel que se use es de mala calidad, mayor será la pérdida por rotura del papel. También existe la posibilidad de reparar la impresora existente o de seguir usándola sin repararla. El gerente de la empresa ha calculado las retribuciones, en miles de dólares, que se obtendrían al usar una máquina nueva o la máquina vieja reparada o la máquina vieja sin reparar para dos calidades de papel (buena y mala). Estas retribuciones se indican en la Tabla 12.8.

Por otro lado, existe un 70% de probabilidad de conseguir papel de buena calidad en el mercado, y la empresa tiene la posibilidad de realizar una prueba preliminar para determinar la calidad del papel, pero esta no es totalmente infalible.

TABLA 12.8 *Tabla de retribuciones en miles de dólares*

	Calidad del papel	
	Buena	***Mala***
Máquina nueva	40	−12
Máquina reparada	15	5
Máquina vieja	10	8

La prueba cuesta 100 dólares y los registros anteriores indican que:

Cuando la calidad del papel era buena realmente, la prueba indicó que la calidad era buena en el 80% de las veces (en el 20% indicó que era mala).

Cuando la calidad del papel era mala realmente, la prueba indicó que la calidad era buena en el 10% de las veces y que era mala en el 90% de las veces.

El administrador de la empresa quisiera, con ayuda del lector, determinar una estrategia que le permita la mejor ganancia esperada por su parte.

EJERCICIOS

1. Un inversor debe elegir entre invertir $ 10,000 durante un año al 12% de interés garantizado o invertir la misma cantidad en una cartera de acciones ordinarias. El rendimiento al invertir en la cartera de acciones dependerá del comportamiento del mercado. Si el mercado está boyante, se espera un beneficio de $ 2,500; si el mercado se mantiene estable el beneficio esperado es $ 500; y si está deprimido, se espera una pérdida de $ 1,000. Usando el criterio del maximin encontrar la mejor acción a tomar.

2. Un inversionista tiene un capital de $ 100,000, los que podría invertir en agricultura, minería o industria metalmecánica. Los resultados de la inversión dependen del estado de la economía, los que pueden tener las siguientes probabilidades de realización: para bueno, 0.3, para regular, 0.3, y para malo, 0.4. La tabla de las retribuciones que el inversionista podría obtener es como se observa en el cuadro de la página siguiente.

Estado de la economía	***Inversión***		
	Agricultura	***Minería***	***Metalmecánica***
Bueno	+15%	+25%	+15%
Regular	+6%	+10%	–20%
Malo	–5%	–30%	–60%

   ¿Cuál es la mejor inversión que se puede realizar?

3. Un tipo de chip se produce en lotes de 1,000. Si la máquina en los que estos se fabrican está bien regulada la proporción de elementos defectuosos p tiene distribución de probabilidad como sigue:

p	0.03	0.06	0.10	0.20
P(p)	0.50	0.30	0.15	0.05

   Cuando un chip defectuoso se instala y debe ser reemplazado cuando la unidad completa es probada, se incurrirá en un costo de $ 2 por cada corrección.

   Es posible contratar un inspector de calidad integral pagándole $ 40 por lote, de tal manera que la calidad se mantenga siempre a un nivel del 3% de chips defectuosos.

   Hacer un análisis sobre la conveniencia de contratar al inspector de calidad.

4. La demanda de los vestidos que se venden en una tienda por departamentos tiene la siguiente distribución:

Unidades: u	14	16	18	20
P(u)	0.10	0.25	0.30	0.35

   La tienda adquiere cada vestido a $ 60 al inicio de la temporada y lo vende a $ 100. Una vez que la estación comienza, se pueden reemplazar los faltantes a un costo de $ 75 cada uno. Encontrar el número óptimo de vestidos que la tienda debe adquirir inicialmente.

5. La empresa Petagra está relacionada con la venta de combustible, y tiene un terreno dentro del cual podría existir petróleo o no. La empresa tiene la posibilidad de hacer la perforación de un pozo para la búsqueda de petróleo o de sembrar caña de azúcar para la obtención de alcohol. Si la empresa lleva a cabo la perforación gastará $ 100,000, pero de hallar petróleo ganará $ 800,000, mientras que si siembra caña ganará $ 90,000. Los informes técnicos indican que la probabilidad de que exista petróleo es 0.20; sin embargo, la empresa podría obtener mayor información llevando a cabo un sondeo sísmico con una técnica que cuando existe petróleo indica resultados desfavorables en el 40% de las veces y en el 80% si no lo hay.
   a) Sin tomar en cuenta el sondeo sísmico, indicar la mejor decisión de parte de la empresa Petagra.
   b) ¿Cuál debe ser la probabilidad de que exista petróleo para que sea indiferente el campo de inversión?
   c) Calcular la probabilidad de que el sondeo sísmico indique resultado desfavorable.
   d) A la luz del sondeo sísmico, calcular las probabilidades a posteriori de que exista o no petróleo.
   e) Si el sondeo sísmico indicara un resultado favorable, ¿cuál sería la decisión a tomar?
   f) Si el sondeo sísmico indicara un resultado desfavorable, ¿cuál sería la decisión a tomar?
   g) Indicar el máximo valor que deberá pagarse por la prueba de sondeo sísmico.

6. Una empresa desea introducir en el mercado un nuevo producto. El gerente de comercialización considera que hay dos clases de resultados de la decisión a tomar; estos son: un mercado bueno y un mercado malo. Se estima que las probabilidades y sus correspondientes rentabilidades, en millones de dólares, son como los que aparecen en la siguiente tabla.

Resultados	Bueno	Malo
***P*(*Resultados*)**	0.35	0.65
Rentabilidad	15	−5

   Es posible realizar una encuesta con un costo de $ 700,000. Los registros de los resultados de esta encuesta que ha sido aplicada en varias oportunidades anteriores señalan que cuando el mercado fue realmente bueno esta indicó que el mercado era bueno en el 65% de las oportunidades, mientras que cuando el mercado fue malo, la encuesta señaló que el mercado era malo en el 75% de las oportunidades. Indicar la mejor decisión a tomar y la rentabilidad que se espera con esta decisión.

7. Una empresa necesita 12 vendedores, y se propone entrenarlos mediante un curso con costo fijo (profesor y alquiler de salón) de $ 5,000. No todos los entrenados terminan el curso, y el número *n* que se requiere para obtener 12 vendedores graduados tiene las siguientes probabilidades:

n	12	13	14	15
***P*(*n*)**	0.20	0.30	0.25	0.25

   El curso durará 6 semanas y a cada candidato se le paga para manutención 100 dólares por semana. Si no aprueban los suficientes alumnos, el curso deberá repetirse. Determinar el tamaño óptimo de la clase inicial y el costo de esta política.

RESPUESTAS A LOS EJERCICIOS

1. La mejor estrategia: invertir al 12% anual. **2.** La mejor estrategia: invertir en agricultura. **3.** La mejor estrategia: contratar al inspector de calidad. **4.** La ganancia esperada si se piden 14 vestidos es igual a: $560 \times 0.10 + 610 \times 0.25 + 660 \times 0.30 + 710 \times 0.35 = 655$ dólares. La ganancia esperada si se piden 16 vestidos es igual a: $560 \times 0.10 + 640 \times 0.25 + 690 \times 0.30 + 740 \times 0.35 = 682$ dólares. La ganancia esperada si se piden 18 vestidos es igual a 701.50 dólares. La ganancia esperada si se piden 20 vestidos es igual a 712 dólares. La mejor estrategia: pedir 20 vestidos. **7.** Se indica la matriz de costos.

		Tamaño escogido			
		12	13	14	15
Alumnos entrantes	12	12,200	12,800	13,400	14,000
	13	17,800	12,800	13,400	14,000
	14	18,400	18,400	13,400	14,000
	15	19,000	19,000	19,000	14,000

CAPÍTULO 13

Introducción a la simulación: el método de Montecarlo

La simulación del universo

Una de las grandes aplicaciones de la simulación ha sido llevada a cabo por el Instituto Max Plank de Alemania. Se trata de la simulación por computador del universo, publicada en 2005, que ha permitido la prueba de una serie de teorías relacionadas con la cosmología.

El programa, llamado Millennium Run, ha logrado simular una región del espacio de 2,000 millones de años luz de longitud, divide al espacio en pequeñas regiones llamadas partículas y permite simular un universo con 20 millones de galaxias.

La cantidad de datos que esta simulación generó durante el tiempo que duró fue de 25 *terabyts*.

La simulación permitió la detección de agujeros negros, galaxias y quasars aparecidos inmediatamente después del Big Bang.

CONTENIDO

13.1 Introducción

La simulación es la construcción de dispositivos experimentales que permiten la recreación simple de procesos que suceden en entornos complejos para su mejor comprensión.

La simulación se usa, por ejemplo:

- Para predecir los efectos al aplicar diferentes estrategias de negocios.
- Para estudiar el proceso de atención a los clientes en las ventanillas de un banco.
- Para tratar de predecir los efectos del calentamiento global en las ciudades.
- Para resolver problemas de las aerolíneas relacionados con la sobreventa de los pasajes o de las empresas hoteleras en la sobreventa de las habitaciones.

Para llevar a cabo el proceso de la simulación existe una amplia variedad de programas computacionales, para los cuales la simulación de valores de una variable aleatoria es primordial. Un método muy conocido y popular para la simulación de valores aleatorios es el *método de Montecarlo* o *algoritmo de la función inversa*. Este método, llamado así en alusión al casino de Montecarlo, fue creado alrededor de 1945, habiéndose perfeccionado con el avance de las computadores.

13.2 El método de Montecarlo o algoritmo de la transformada inversa: método para generar valores de una variable aleatoria

Para variables aleatorias continuas

La siguiente proposición se conoce como el algoritmo de la transformada inversa. Esta permite la generación de valores de una variable aleatoria continua.

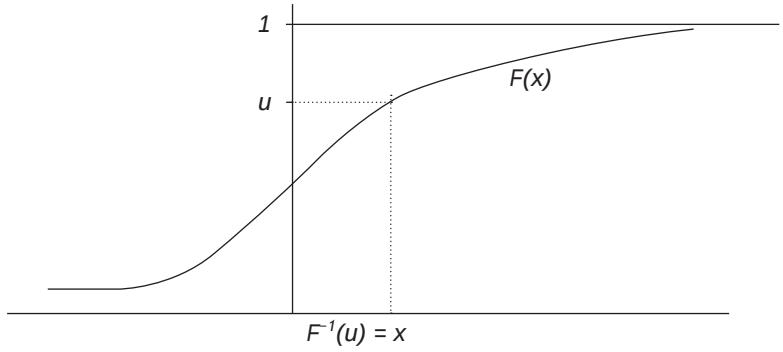
> *Si X es una variable aleatoria continua con distribución $F(x)$ y u es un número aleatorio entre* 0 *y* 1*, entonces $x = F^{-1}(u)$ es un valor de la variable aleatoria X.*

De acuerdo a esta proposición, para generar o simular el valor de una variable aleatoria X continua con función de acumulación $F(x)$ se sigue el siguiente procedimiento:

1. Se genera mediante una tabla de números aleatorios o un programa computacional un número aleatorio u entre 0 y 1, el que se ubica en el eje Y.
2. Se encuentra el valor de x para el que se cumple $F(x) = u$.

El valor de x que resuelve la ecuación $F(x) = u$ corresponde al valor simulado de la variable X.

FIGURA 13.1 *Función de distribución o de acumulación de X*



EJEMPLO. *Simulando valores de una variable con distribución exponencial*

Usaremos la proposición para simular un valor de la variable aleatoria X cuya distribución es la exponencial con parámetro 1, y cuya función de distribución es $F(x) = 1 - e^{-x}$.

Primero, se genera un número aleatorio u entre 0 y 1. Supongamos que este es $u = 0.2345$.

En segundo lugar, en la ecuación $u = F(x)$ se reemplaza $F(x) = 1 - e^{-x}$ y $u = 0.2345$.

Despejando x se obtiene el valor simulado de X: $x = -\ln(1 - u)$. Esto es:

$$x = -\ln(1 - 0.2345) = 0.2672$$

Siguiendo este procedimiento se puede simular muchos valores de X. En realidad, se obtiene así una muestra aleatoria de valores de la variable aleatoria.

EJEMPLO. *Simulando valores de una variable aleatoria con distribución normal*

Para generar valores de una variable aleatoria normal con media 0 y varianza 1 se puede usar simplemente la tabla de Z, del apéndice A. El número aleatorio u generado es igual o aproximadamente igual a uno de los que aparecen en el cuerpo de la tabla. El valor simulado de la variable aleatoria normal estándar se obtiene recorriendo de manera inversa la fila y la columna correspondiente de la tabla Z.

Simularemos ahora el valor de la variable aleatoria que corresponde al tiempo *X* de realización de una tarea, cuya distribución es normal de media 12 y desviación estándar 2.

La variable *X* se puede escribir como $X = 12 + 2Z$, donde *Z* tiene distribución $N(0, 1)$.

Suponiendo que el número aleatorio generado entre 0 y 1 es $u = 0.876$, la inversa de la función de acumulación de la normal estándar se encuentra usando la tabla de la distribución normal estándar, del apéndice A. Este valor es z = 1.16, y el tiempo simulado de la realización de la tarea es:

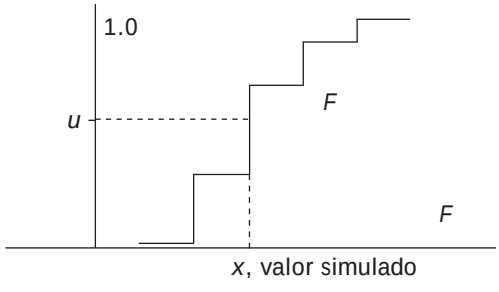
$$x = 12 + 2(1.16) = 14.32$$

Para variables aleatorias discretas

Cuando la variable *X* es discreta, el procedimiento para simular valores a seguir es análogo al caso continuo:

1. Se considera la función de acumulación *F* de la variable.
2. Se genera al azar un número *u* entre 0 y 1, el que se ubica en el eje *Y.*
3. Desde *u* avanzar horizontalmente hasta la función de acumulación y luego, verticalmente, avanzar hasta el eje *X.* El valor *x* encontrado en el eje horizontal es el valor de *X* simulado.

FIGURA 13.2 *Función de distribución o de acumulación de X*



EJEMPLO. *Simulando el número de defectos en tabletas de circuitos*

Se ha determinado que el número *X* de defectos que aparecen en una tableta de circuitos y que se usan en el ensamblaje de una computadora tienen la siguiente distribución.

TABLA 13.1 *Distribución del número de defectos*

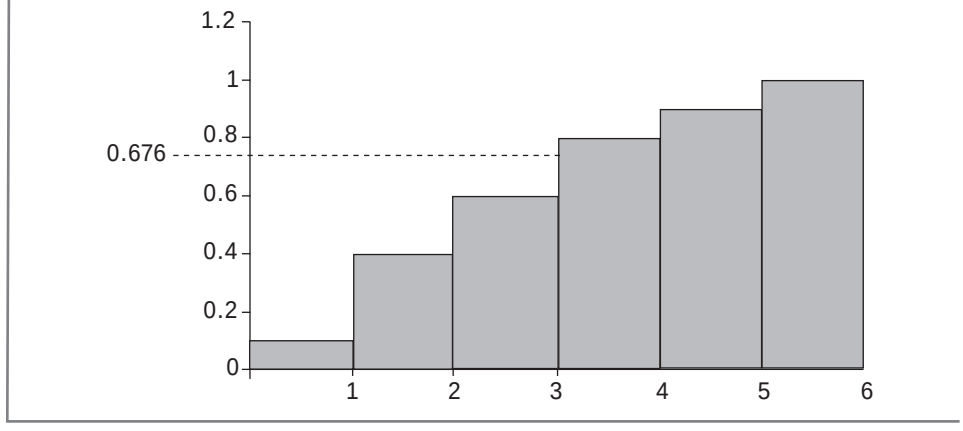
***X* = Defect**	**0**	**1**	**2**	**3**	**4**	**5**
$P[X = x]$	0.10	0.30	0.20	0.20	0.10	0.10

Simularemos el número de defectos que puede tener una tableta a fabricar.

La gráfica de la función de acumulación de *X* aparece a continuación.

Un número aleatorio generado entre 0 y 1 es 0.676, entonces el valor simulado de *X* es 3.

FIGURA 13.3 *Función de acumulación de X*



EJEMPLO. *Simulando la ganancia de un vendedor*

Un vendedor de artículos perecibles se encuentra ante el dilema de solicitar diariamente para la venta un número determinado de estos. Para el vendedor es un dilema porque la demanda *D* de este tipo de artículos es aleatoria, y si esta fuera baja tendría una pérdida ocasionada por tener que vender los artículos sobrantes a un precio menor al que los compró. La historia que se tiene indica que la demanda diaria puede ser, con igual probabilidad, 40, 50, 60, 70, 80 o 90 artículos (la probabilidad en cada caso es 1/6). El costo de cada artículo es 10 dólares y se vende a 12 dólares; sin embargo, si al finalizar el día quedan artículos sin vender, el vendedor tendrá que rematarlos a 6 dólares cada uno. Al vendedor le interesa, en buena cuenta, conocer el número *Q* de artículos que debe pedir al iniciar el día para la venta, así como el riesgo (*el riesgo se define como la probabilidad de ocurrencia de un evento no deseado*) de obtener cada día una ganancia negativa.

El criterio que el vendedor utilizará para solicitar el número de artículos es el del *valor esperado de la ganancia*. Según este criterio, el vendedor pedirá un número de artículos que maximice el valor esperado de la ganancia.

Si el vendedor solicita Q artículos, la ganancia G del vendedor *depende de la demanda* y puede expresarse de la siguiente manera:

$$G(D) = \begin{cases} 12D - 10Q + 6(Q - D) & si\ D \leq Q \\ 12Q - 10Q & si\ D > Q \end{cases}$$

En este caso, se puede calcular analíticamente el valor esperado de la ganancia; sin embargo, simularemos un conjunto de valores de G, encontrando luego su promedio, para así tener un valor aproximado de su valor esperado.

Para simular los valores de la ganancia se debe generar valores de la demanda usando tablas de números al azar o programas computacionales. A falta de estos, y por la forma que tiene la distribución de la demanda, se puede utilizar un dado equilibrado, asociando primero a cada uno de los resultados los valores de la demanda.

TABLA 13.2

1	2	3	4	5	6
40	50	60	70	80	90

Suponiendo que se han pedido $Q = 40$ artículos:

1. Lanzar el dado (esto equivale a simular valores de una variable que toma los valores 1, 2, 3, 4, 5 y 6, cada uno con probabilidad 1/6).
2. De acuerdo al resultado en 1, y usando la tabla anterior, determinar el valor de la demanda.
3. Usando el valor $Q = 40$ y el de la demanda generada, calcular la ganancia.
4. Registrar la ganancia.

Repetir, por ejemplo, 1,000 veces el procedimiento, y así se tendrá 1,000 valores simulados de la ganancia G. La distribución de estos valores será una aproximación de la distribución teórica de la ganancia. Tomando el promedio de los 1,000 valores de la ganancia se tendrá la ganancia esperada aproximada cuando se solicitan 40 artículos.

Nótese que cuando se piden 40 artículos, cualquiera que sea la demanda, al final del día se habrán vendido todos los artículos, por lo cual la ganancia será $ 80.00.

De manera análoga se estima la esperanza de la ganancia cuando se solicitan 50, 60, 70, 80 y 90 artículos.

La distribución de 1,000 valores simulados de la ganancia *G* cuando la cantidad pedida es *Q* = 50 es como se indica en la Tabla 13.3, y el valor promedio es igual a $ 88.60.

TABLA 13.3 *Distribución de la ganancia cuando Q* = 50. *El valor promedio es $* 88.60

G	*Frecuencia*	*Frecuencia relativa en %*
40.00	190	19.00
100.00	810	81.00
Total	1,000	100.00

Las distribuciones para los otros valores de *Q,* así como los promedios de la ganancia *G* correspondientes, se indican a continuación.

TABLA 13.4 *Distribución de la ganancia cuando Q* = 60. *El valor promedio es $* 89.58

G	*Frecuencia*	*Frecuencia relativa en %*
00.00	166	16.6
60.00	175	17.5
120.00	659	65.9
Total	1,000	100.00

TABLA 13.5 *Distribución de la ganancia cuando Q* = 70. *El valor promedio es $* 81.74

G	*Frecuencia*	*Frecuencia relativa en %*
–40.00	159	15.90
20.00	170	17.00
80.00	154	15.40
140.00	517	51.70
Total	1,000	100.00

TABLA 13.6 *Distribución de la ganancia cuando Q* = 80. *El valor promedio es $* 58.66

G	*Frecuencia*	*Frecuencia relativa en %*
–80.00	176	17.60
–20.00	161	16.10
40.00	167	16.70
100.00	168	16.80
160.00	328	32.80
Total	1,000	100.00

TABLA 13.7 *Distribución de la ganancia cuando Q = 90. El valor promedio es $ 24.42*

G	*Frecuencia*	*Frecuencia relativa en %*
–120.00	191	19.10
–60.00	172	17.20
00.00	147	14.70
60.00	177	17.70
120.00	155	15.50
180.00	158	15.80
Total	1,000	100.00

El vendedor puede tomar la decisión de solicitar 60 artículos, pues de esta manera obtiene la mayor ganancia esperada.

Notemos que cuando se pide 40, 50 o 60 artículos el vendedor no corre ningún riesgo, pues la ganancia es no negativa; sin embargo, si solicita 70 artículos el riesgo (la probabilidad de que la ganancia sea negativa) es 0.159. Si solicita 80, el riesgo es 0.337 y si solicita 90 el riesgo es 0.363.

La simplicidad de la distribución de la demanda permitió simular sus valores usando un dado equilibrado; sin embargo, no siempre esto es posible, por lo que será necesario algún programa computacional para generar los números al azar entre 0 y 1 para luego obtener los valores de las variables que definen el modelo.

En este caso, el valor esperado de la demanda se puede encontrar directamente aplicando las propiedades de la esperanza de una variable; sin embargo, el método desarrollado ilustra la manera como se puede aplicar la simulación.

EJEMPLO. *Simulando un inventario semanal*

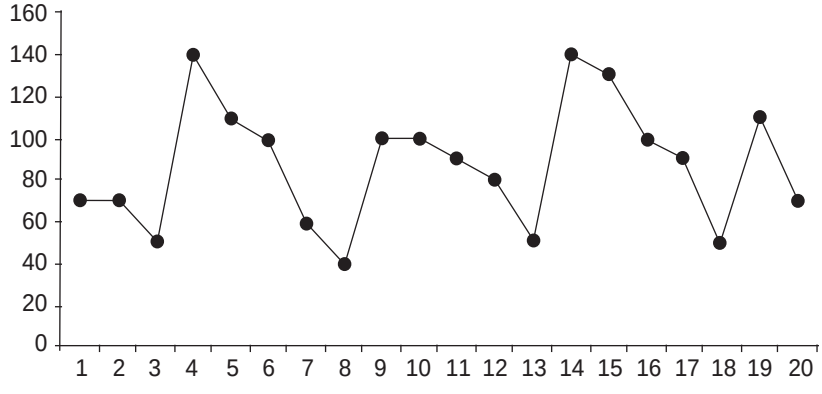
En este caso se simula un inventario para un productor de baterías que por diversas razones no puede mantener un número indiscriminado de estas en su almacén. Teniendo en cuenta esta limitación, el productor decide fabricar 200 baterías si después de registrar la demanda el inventario es 50 o menos, y en otro caso fabricará solo 100. Considerando que los valores de la demanda de baterías pueden ser 100, 110, 120, 130 o 140 con igual probabilidad (1/5), se simula el inventario para 20 semanas consecutivas. El modelo para el inventario al final de cada semana puede escribirse como:

Inventario al final de la semana = Inventario al inicio de la semana + Producción

Los valores de la demanda generados aparecen en la siguiente tabla. En la misma tabla se han escrito los valores simulados del inventario final de la semana. La gráfica del inventario para 20 semanas consecutivas aparece en la Figura 13.4.

Demanda simulada	*Inventario inicial + producción*	*Inventario final*
130	200	70
100	170	70
120	170	50
110	250	140
130	240	110
110	210	100
140	200	60
120	160	40
140	240	100
100	200	100
110	200	90
110	190	80
130	180	50
110	250	140
110	240	130
130	230	100
110	200	90
140	190	50
140	250	110

FIGURA 13.4 *Inventario final*



El paquete computacional Excel trae consigo procedimientos para simular valores de variables aleatorias con determinadas distribuciones. Paquetes estadísticos como el SPSS son más extensos en este punto. Existen paquetes computacionales como el Crystal Ball especialmente preparados para la simulación.

APLICACIÓN: El caso de la sobreventa de habitaciones en el hotel Melodía

El caso del hotel Melodía, que tiene 500 habitaciones, ha sido presentado cuando se trataron las series de tiempo. Ahora se trata de usar la simulación para contestar la eterna pregunta de la administración del hotel: ¿cuántas habitaciones se pueden sobrevender? Si un día se acepta 500 reservaciones para el día siguiente, es muy posible que queden habitaciones vacías, pues existirán personas que no hacen efectivas las reservaciones. Por otro lado, si se acepta más de 500 reservaciones, se corre el riesgo de que, aun contando las reservas no llevadas a cabo, se tenga clientes que se queden sin habitación. El costo de cada habitación es $ 150. Si un cliente que ha hecho reserva no encuentra habitación desocupada el hotel le busca habitación en otro hotel y le envía una botella de vino, lo cual le acarrea un costo de $ 250.

La experiencia indica que el porcentaje de personas que no hacen efectivas las reservaciones es del 4%. Usando esta información, simular la ganancia promedio que el hotel Melodía tendrá cuando venda 510, 520 o 530 habitaciones.

Si el administrador quisiera evaluar la confiabilidad de la simulación, a partir de la muestra obtenida, calculará el intervalo de confianza y el margen de error le indicará lo que él desea.

EJERCICIOS

1. Simular 3 valores de una variable aleatoria cuya distribución de probabilidad se indica en la siguiente tabla.

x	1	2	3	4	5	6
$P[X = x]$	0.1	0.4	0.3	0.1	0.05	0.05

2. Usando algún paquete computacional, simular:
   a) 10 valores de la distribución uniforme en el intervalo [0, 1].
   b) 50 valores de la distribución normal de media 50 y desviación estándar 10.
3. Simular 2 valores de la distribución binomial con parámetros $n = 10$, $p = 1/5$.
4. Usando la distribución exponencial simular 2 valores de una variable aleatoria con distribución de Poisson con parámetro 5 por hora.
5. A una estación de servicio de gasolina, en donde existe un solo surtidor, llegan vehículos de acuerdo a una distribución de Poisson con tasa 10 en un intervalo de 30 minutos. El servicio que se realiza a cada vehículo tiene distribución exponencial con media 5 minutos. Simular 10 llegadas y los tiempos de servicios correspondientes. Indicar si al cabo de la décima llegada existirá una cola de espera.
6. La probabilidad de que una tarjeta de sonido resulte defectuosa en una línea de producción es 0.001. Las tarjetas se empacan en paquetes de 5 tarjetas cada uno. Usando los números aleatorios 0.632, 0.245, 0.762 y 0.8112, simular el número de tarjetas defectuosas que resultarán en 4 paquetes de 5 tarjetas cada uno.
7. La compañía manufacturera ACME ha recibido de una tienda distribuidora el encargo de fabricar bombas de agua para automóviles. El número de bombas que serán pedidas semanalmente dependerá de la demanda que tenga la tienda. La capacidad de producción planeada

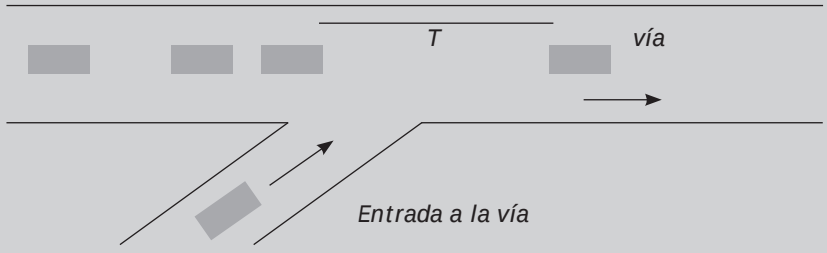
por ACME es de 100 bombas por turno; sin embargo, para mantener un inventario suficiente ha considerado que si el inventario cae por debajo de 60 unidades se establecerá un segundo turno de producción con 100 unidades adicionales.

Suponiendo que ACME cuenta al inicio con 100 unidades y que la demanda de bombas en la tienda tiene la siguiente distribución:

Demanda	**80**	**90**	**100**	**110**	**120**	**130**
Probabilidad	1/6	1/6	1/6	1/6	1/6	1/6

Simular el inventario de ACME durante 50 semanas.

8. Con la finalidad de ordenar el tránsito se estudia cómo ingresan los vehículos a una vía principal donde fluye el tránsito. Se ha determinado que los conductores ingresan a la vía si el tiempo T entre dos vehículos consecutivos es no menor de 25 segundos. Se supone que los vehículos que transitan la vía se acercan a la intersección según un proceso de Poisson de tasa 2 cada minuto.



Simular el ingreso o no de un vehículo a la vía principal después que han pasado 5 vehículos por la vía.

9. El tiempo T para que una componente falle tiene una distribución exponencial con media 100 horas. Una máquina usa esta componente para operar. Cuando la máquina está funcionando la entrada que se obtiene es de $ 10.00 la hora, pero para operar la máquina se contrata un operador por 100 horas pagándole $ 5.00 la hora. Simular 10 valores de la utilidad.

10. Un fabricante decide producir 3 toneladas diarias de cierto producto. El investigador de mercado le indica que la demanda diaria del producto es una variable X continua cuyos valores están entre 0 y 5 toneladas, y cuya función de densidad es $f(x) = 2x/25$. El costo de fabricación y el precio de venta por cada tonelada del producto son 1,000 y 1,200 dólares, respectivamente, y el analista de mercado le cobra por el estudio 500 dólares.

    Simular valores de la utilidad para dos días, tomando como valores aleatorios 0.12 y 0.50, respectivamente.

11. A la ventanilla de pago de un comedor de estudiantes llegan los alumnos a razón de 10 cada 2 minutos. El servicio de caja atiende a cada alumno en un tiempo promedio de 0.5 minutos, luego de lo cual los alumnos pasan al servicio de comida, en donde son atendidos en un tiempo promedio de 1 minuto. Suponiendo que el comedor comienza a atender a las 12 m, que existe una sola caja y un solo puesto de atención de comida, simular llegadas al comedor desde las 12 m hasta la 1 de la tarde y:

a) Indicar una aproximación del valor esperado del tiempo que el estudiante gasta en pagar y ser atendido.

b) ¿Se recomendaría que se abra otra caja para agilizar la atención?

c) Si cada estudiante consume un plato común con probabilidad 0.8 y a la carta con probabilidad 0.2, indicar el valor total consumido aproximado si el plato común cuesta $ 4 y el valor de consumo a la carta tiene una distribución normal de media $ 6 y desviación estándar 2.

12. Se supone que una determinada estación de servicio de venta de gasolina tiene una sola "isla" en donde existen tres surtidores de gasolina, de 84, de 90 y de 95 octanos.

De las 8 a las 11 de la mañana llegan λ vehículos por hora (se indicará el valor de λ suponiendo que se tiene alguna fuente de información).

Suponer que haciendo una averiguación estadística se ha encontrado que:

a) *a*% de los vehículos son autos pequeños.

b) *b*% de los vehículos son autos medianos.

c) *c*% de los vehículos son autos grandes.

(Indicar valores aproximados para *a, b* y *c*, de acuerdo a la realidad.)

Una encuesta realizada en varios grifos muestra que los autos pequeños y medianos prefieren la gasolina de 84 octanos en el *d*% de las veces y el resto prefieren la de 90 octanos. Los autos grandes prefieren la de 90 octanos el *e*% de las veces y en el resto de las veces prefieren la de 95 octanos (se indicará, usando los conocimientos que se tiene respecto de este mercado, los porcentajes para *d* y *e*).

Otro estudio estadístico muestra que la distribución de la cantidad de gasolina que se compra para cualquiera de los autos es normal, pero las medias y varianzas respectivas varían de acuerdo al tamaño del auto (se indicarán las medias y varianzas usando los conocimientos que se tiene respecto de este mercado).

Usando la simulación, estimar el valor de la media del total de ventas del grifo entre las 8 y las 11 de la mañana (para cada tipo de gasolina, usar el precio promedio en el mercado actual).

SOLUCIÓN DE LOS EJERCICIOS

1. Aplicar el método de Montecarlo para el caso de las variables aleatorias discretas. **2.** Se puede usar el EXCEL, el SPSS, el Minitab, etcétera. **3.** Escribir, usando por ejemplo el EXCEL, la distribución de la binomial y luego aplicar el método de Montecarlo para el caso de las variables aleatorias discretas. **4.** Simular valores de una distribución exponencial de media 12 minutos. El número de valores simulados que caen en el intervalo de 60 minutos es el valor simulado de la distribución de Poisson de parámetro 5. **6.** Con los números aleatorios dados simular 4 valores de una variable aleatoria con distribución binomial de parámetros 5, 0.001. **8.** Como los vehículos llegan a la intersección a una tasa de 2 por minuto, el tiempo que transcurre entre llegada y llegada tiene distribución exponencial de media 0.5 minutos. Luego, debe simularse 5 valores de la distribución exponencial indicada, y si alguno de estos es mayor que 25 segundos el vehículo puede entrar a la vía. **10.** Usando la función de densidad de la demanda, encontrar su función de acumulación. Con la función de acumulación simular dos valores de la demanda *D*. Los valores simulados de la utilidad *U* serán obtenidos de $U = 1{,}200D - 1{,}000(3) - 500$.

CAPÍTULO 14

Introducción al muestreo aleatorio estratificado para encuestas

Algunos hechos en la historia del muestreo

En 1802, P. S. Laplace usó el muestreo para estimar la población de Francia.

En 1934, Neyman Pearson presentó un trabajo múltiple que comprendía:

- La "asignación de Neyman" para el muestreo estratificado.
- La estimación de las varianzas de los estimadores que se obtienen con el muestreo estratificado.
- La importancia de los números aleatorios para seleccionar las muestras.
- La demostración de que el esquema probabilístico es más preciso que el muestreo por "conveniencia".

En 1938, El U. S. Census Bureau usó una muestra nacional para estimar el desempleo.

En 1951, J. Cornfield introdujo el concepto del efecto del diseño que compara la varianza muestral de un diseño con la varianza muestral de un muestreo aleatorio simple.

En 1952, Horvitz y Thompson presentaron la teoría general de muestreo para probabilidades desiguales.

CONTENIDO

14.1 Introducción

El desarrollo realizado y que correspondía a la estadística inferencial se refería al análisis de los elementos de una población basándose en la información extraída de muestras obtenidas usando el *muestreo aleatorio simple*. La aplicación de este tipo de muestreo supone que la población de donde se extrae la muestra es homogénea; sin embargo, en muchas ocasiones no es así, por lo cual es necesario dividir la población en subconjuntos, para luego a partir de ellos formar la muestra. Las diversas maneras como se forman los subconjuntos originan diferentes tipos de muestreo, como el muestreo estratificado, el muestreo por conglomerados, el muestreo en diferentes etapas, etc. En este capítulo, y como una introducción al estudio de la teoría del muestreo, se presenta el muestreo aleatorio estratificado.

14.2 El muestreo aleatorio estratificado

El *muestreo aleatorio estratificado* se usa para estimar parámetros de poblaciones muy heterogéneas; consiste en la separación de las unidades de la población en grupos, de tal manera que respecto de la variable en estudio sean lo más heterogéneos posibles, pero que sean homogéneos en su interior. Estos grupos se llaman *estratos*. De cada estrato se obtiene una muestra aleatoria simple, y los estimadores de los parámetros de la población se estiman como combinaciones de los estimadores obtenidos en cada estrato.

La precisión de los estimadores que con este tipo de muestreo se obtiene depende de la variabilidad dentro de los estratos y entre los estratos. Esta precisión es cada vez más grande cuando la variabilidad dentro de cada estrato es menor que la variabilidad entre los estratos.

¿Cómo formar los estratos?

En la práctica, el modo como se forman los estratos depende de los objetivos, del conocimiento de la población y de las variables en estudio. Dependiendo del estudio que se aplique, los estratos se pueden formar, por ejemplo, a partir de regiones naturales, de grupos de edades, de grupos socioeconométricos, etcétera.

Para formar los estratos es necesario conocer la distribución de la variable en estudio en cada estrato; sin embargo, como esto generalmente no sucede, se puede usar otra variable cuya distribución es conocida y que está correlacionada con la primera. Para estudiar la producción de las empresas resulta conveniente estratificar de acuerdo al número de empleados. Las variables que se usan para la estratificación se llaman *variables de estratificación*.

Algunas veces, y con el fin de conseguir una buena estratificación, se usa más de una variable de estratificación, llegando a lo que se llama *estratificación cruzada*. Por citar un caso, para estudiar el consumo de un producto, la estratificación puede hacerse, por ejemplo, cruzando las variables "edad" y "nivel de renta".

EJEMPLO. *Para conocer la opinión de los jefes de familia*

Para conocer las opiniones de los jefes de familia de una ciudad respecto de los colegios laicos, en lugar de seleccionar una muestra aleatoria simple, tomada de una lista de todos los jefes de familia, se dividió a la ciudad en sectores o estratos de acuerdo al nivel socioeconómico y se tomó una muestra en cada uno de ellos por separado. Los resultados encontrados en cada muestra se combinaron y así se obtuvo información de todos los jefes de familia.

Estimaciones a partir del muestreo aleatorio estratificado

Para este tipo de muestreo se indican a continuación las notaciones usuales, así como las estimaciones para la media poblacional, para el total de la población y para una proporción poblacional.

Nota

L = Número de estratos

μ = Media de la población

T = Total de las observaciones en la población

p = Proporción en la población de elementos con el atributo A

N_i = Número de elementos en el estrato i, $i = 1, 2, \ldots, L$

N = Número total de elementos en la población, $(N = \sum_{i=1}^{L} N_i)$

n_i = Número de elementos en el estrato i de la muestra seleccionada

n = Tamaño total de la muestra

$\overline{x}_i$ = Media de las mediciones en la muestra seleccionada en el estrato i

p_i = Proporción muestral en el estrato i

s_i^2 = Varianza de las mediciones en la muestra tomada en el estrato i

Estimador puntual de la media poblacional

Habiendo realizado una estratificación de la población en L estratos de tamaños: N_1, N_2,..., N_L, para estimar la media μ de la población (media de la variable en estudio) se toma una muestra aleatoria en cada estrato i de tamaño n_i, y con las medias muestrales de cada estrato se define como *estimador de la media de la población* a:

$$\bar{x} = \frac{1}{N}(N_1\bar{x}_1 + \ldots + N_L\bar{x}_L)$$

Cada expresión N_i/N es una *ponderación del estrato i.*

Nótese que el estimador indicado es diferente a $\frac{1}{n}(n_1\bar{x}_1 + \ldots + n_L\bar{x}_L)$, salvo que $\frac{N_i}{N} = \frac{n_i}{n}$. Cuando esto sucede se dice que la *muestra es autoponderada.*

Si el muestreo en cada estrato i es aleatorio simple con restitución, el error estándar del estimador de μ es $\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \frac{s_i^2}{n_i}}$.

Si el muestreo en cada estrato i es aleatorio simple sin restitución, el error estándar es:

$$\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}}.$$

El intervalo de confianza aproximado al 95% para μ cuando el muestreo en cada estrato es aleatorio simple sin restitución es:

$$\bar{x} \pm 1.96\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{s_i^2}{n_i}}$$

Los factores de corrección $\left(1 - \frac{n_i}{N_i}\right)$ se reducen a la unidad cuando el muestreo es aleatorio simple con sustitución o si n_i es pequeño respecto de N_i (en la práctica, si es menor o igual al 10% de N_i).

El *estimador puntual del total poblacional T* es:

$$\tau = N \cdot \bar{x}$$

El error estándar de este estimador cuando el muestreo en cada estrato es simple y sin restitución es:

$$\sqrt{\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}}$$

El intervalo de confianza al 95% para el total poblacional cuando el muestreo en cada estrato es simple y sin restitución es:

$$\tau \pm 1.96 \sqrt{\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}}$$

El *estimador puntual de la proporción p* poblacional es:

$$\hat{p} = \frac{1}{N} (N_1 \hat{p}_1 + \ldots + N_L \hat{p}_L)$$

El *error estándar aproximado* de este estimador cuando el muestreo en cada estrato es aleatorio simple y sin restitución es:

$$\sqrt{\frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{p}_i \hat{q}_i}{n_i - 1}}$$

El *intervalo de confianza* al 95% para la proporción cuando el muestreo en cada estrato es simple y sin restitución es:

$$\hat{p} \pm 1/96 \sqrt{\frac{1}{N^2} \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{\hat{p}_i \hat{q}_i}{n_i - 1}}$$

Los factores de corrección $\left(1 - \frac{n_i}{N_i}\right)$ se reducen a la unidad cuando el muestreo es aleatorio simple con sustitución o si n_i es pequeño respecto de N_i (en la práctica menor o igual al 10% de N_i).

EJEMPLO. *Para conocer el tiempo de dedicación a la TV*

Un estudio para conocer el tiempo promedio que las familias dedican a ver televisión diariamente fue llevado a cabo en una ciudad. La ciudad se dividió en tres zonas o estratos. En cada estrato se seleccionó de manera aleatoria el 1% del número total de familias. Un resumen de los resultados de las operaciones aparece a continuación.

TABLA 14.1 *Resumen de los resultados*

Zona	*Número de familias en el estrato i,* N_i	*Tamaño muestral en el estrato i* n_i	*Media muestral en el estrato i* x_i	*Varianza muestral en el estrato i* s_i^2
Zona 1	12,000	120	2.8	1.44
Zona 2	25,000	250	3.5	1.96
Zona 3	30,000	300	2.1	3.20
Total =	67,000			

El estimador de la media poblacional es:

$$\overline{x} = \frac{1}{67{,}000}\ (1{,}200 \times 2.8 + 25{,}000 \times 3.5 + 30{,}000 \times 2.1) = 2.74$$

Considerando que para cualquier estrato $\left(1 - \frac{n_i}{N_i}\right) \approx 1$, el error estándar del estimador es igual a:

$$\sqrt{\frac{1}{(67000)^2}\left[(12000)^2\,\frac{1.44}{120} + (25000)^2\,\frac{1.96}{250} + (30000)^2\,\frac{3.2}{300}\right]}\text{, aproximadamente.}$$

El intervalo, al nivel de confianza del 95%, para la media poblacional es:

$$2.74 \pm 1.96\sqrt{\frac{1}{(67{,}000)^2}\left[(12{,}000)^2\,\frac{1.44}{120} + (25{,}000)^2\,\frac{1.96}{250} + (30{,}000)^2\,\frac{3.2}{300}\right]} = 2.74 \pm 0.12$$

El número promedio de horas que una familia pasa frente al televisor por día está en el intervalo [2.62, 2.86].

El estimador puntual del total del tiempo en horas que las familias de toda la población dedican a ver televisión es:

$$\tau = 67{,}000(2.74) = 183{,}580$$

El tiempo total *t* que las personas en toda la población dedican a ver televisión está en el intervalo [175,540, 191,620], con un nivel de confianza del 95%.

Para completar la ilustración, se desarrolla a continuación el procedimiento para estimar la proporción de personas que ven televisión más de 4 horas

diarias. Se supone que las proporciones de personas que ven televisión más de 4 horas diarias en las zonas 1, 2 y 3 son: 0.10, 0.15 y 0.12, respectivamente.

El estimador puntual de la proporción de los que ven televisión más de 4 horas diarias en toda la población es:

$$\hat{p} = \frac{1}{67{,}000}\ (12{,}000(0.10) + 25{,}000(0.15) + 30{,}000(0.12)) = 0.1276$$

Y el intervalo al 95% de confianza para estimar la proporción de todas las personas que diariamente ven más de 4 horas de televisión es:

$$0.1276 \pm 1.96\sqrt{\frac{1}{67{,}000^2}\left(12{,}000^2\left(\frac{0.10(0.90)}{120-1}\right)+25{,}000^2\left(\frac{0.15(0.85)}{250-1}\right)+30{,}000^2\left(\frac{0.12(0.88)}{300-1}\right)\right)}$$

en donde las expresiones de la forma $\left(1 - \frac{n_i}{N_i}\right)$ se han aproximado con 1.

Como resultado se tiene que la proporción de todos los que ven televisión más de 4 horas diarias en toda la población está entre 0.1023 y 0.1529, con un nivel de confianza del 95%.

Tamaño de muestra en el muestreo estratificado

Para el muestreo estratificado, el proceso de la determinación del tamaño de muestra comprende dos pasos:

Primer paso: se determina el tamaño n de la muestra total.

Segundo paso: se reparte la muestra total en todos los estratos. A esta operación se le llama ***afijación de la muestra***.

Hay varias formas de afijar la muestra; se indican dos maneras que a menudo se usan.

1. *Tomando en cuenta el número de elementos en cada estrato*. En este caso, la repartición de la muestra total n se realiza de manera proporcional al número de unidades muestrales en cada estrato, multiplicando la proporción que representa el estrato en la población por el tamaño de la muestra total.

   Si se tienen tres estratos con poblaciones N_1, N_2 y N_3, los tamaños de muestra que le corresponden a cada estrato son:

   $n_1 = n.\ \frac{N_1}{N_1 + N_2 + N_3}$, $n_2 = n.\ \frac{N_2}{N_1 + N_2 + N_3}$ y $n_3 = n.\ \frac{N_3}{N_1 + N_2 + N_3}$, respectivamente.

2. *Tomando en cuenta la varianza dentro de cada estrato*. En este caso, la repartición se lleva a cabo tomando como peso relativo de cada estrato a la variabilidad que tiene cada uno de ellos. Esta repartición se conoce como *asignación de Neyman*, y distribuye el total *n* de la muestra a los diversos estratos de la siguiente manera:

$$n_h = n\left(\frac{N_h\sigma_h}{\sum_{k=1}^{L} N_k\sigma_k}\right)$$

Si no se conoce σ_h, se aproxima con la desviación estándar muestral, s_h, calculada a partir de una muestra piloto.

La determinación del tamaño de la muestra total *n* dependerá del nivel de precisión que se desee y del parámetro que se requiera estimar. A continuación se presentan las expresiones para calcular el tamaño de muestra necesario para estimar la media, la proporción y total de la población para el caso en que la afijación de la muestra es proporcional. El lector interesado puede encontrar en textos especializados la información requerida cuando la afijación es óptima.

El ***tamaño de muestra para obtener un intervalo de confianza al 95% para la media poblacional y para el total poblacional cuando la afijación es proporcional:***

$$n = \frac{\sum \frac{N_i^2\sigma_i^2}{w_i}}{N^2D + \sum N_i\sigma_i^2}$$

en donde:

w_i = Importancia del estrato *i*, bien puede ser la proporción de unidades en el estrato (si no se tiene esta información se usa el peso 1 para todos los estratos).

σ_i^2 = Varianza en el estrato *i*

$D = \begin{cases} d^2/4 \text{ para estimar la media} \\ d^2/(4N^2) \text{ para estimar el total} \end{cases}$

d = Margen de error del estimador

Tamaño de muestra para obtener un intervalo de confianza al 95% para la proporción poblacional cuando la afijación es proporcional:

$$n = \frac{\sum \frac{N_i^2\pi_i(1-\pi_i)}{w_i}}{N^2D + \sum N_i\pi_i(1-\pi_i)}$$

en donde:

π_i = Proporción en el estrato *i*

Si no se conoce este dato, la expresión $\pi_i (1 - \pi_i)$ puede reemplazarse con el mayor valor que esta puede tomar: (0.25).

w_i = Importancia del estrato *i*, bien puede ser la proporción de unidades en el estrato *i* (si no se tiene esta información se usa el peso 1 para todos los estratos).

$D = \dfrac{d^2}{4}$

d = Margen de error del estimador

EJEMPLO. *Para conocer a los internautas*

Para estimar la proporción *p* de estudiantes de una universidad que usan Internet se asumió como estratos las facultades que ahí funcionan: Ciencias, Letras y Arte. Se supone que se tomarán tamaños de muestra proporcionales al número de estudiantes en cada facultad. La información referida a la población y las proporciones para cada estrato aparecen a continuación.

TABLA 14.2 *Resumen de los resultados*

Facultades	*Población en cada facultad,* N_i	*Proporción para cada estrato*
Ciencias	2,400	3/11
Letras	4,800	6/11
Artes	1,600	2/11
Total	8,800	1

Se desea hallar el tamaño necesario de muestra en cada estrato para estimar la proporción *p* mediante un intervalo al 95% de confianza y con un margen de error de 0.015.

Solución

Suponiendo que no se tiene ninguna información previa de las proporciones en cada facultad, se usará como valores conservativos: $p_1 = p_2 = p_3 = 0.5$. El tamaño total de la muestra es:

$$n = \frac{\sum \dfrac{N_i^2 p_i (1 - p_i)}{w_i}}{N^2 D + \sum N_i p_i (1 - p_i)} =$$

$$\frac{\frac{(2{,}400)^2 0.5 \times 0.5}{3/11} + \frac{(4{,}800)^2 0.5 \times 0.5}{6/11} + \frac{(1{,}600)^2 0.5 \times 0.5}{2/11}}{(8{,}800)^2(0.015^2/4) + (2{,}400)0.5 \times 0.5 + (4{,}800)0.5 \times 0.5 + (1{,}600)0.5 \times 0.5} = 2{,}953$$

Aplicando la afijación proporcional se obtienen los siguientes tamaños de muestra para las facultades de Ciencias, Letras y Arte, respectivamente:

$$2{,}953(3/11) \approx 805,\ 2{,}953(6/11) \approx 1{,}611,\ 2{,}953(2/11) \approx 537$$

EJEMPLO. *Para conocer la edad promedio se ha estratificado*

Para estimar la edad promedio de una población se ha dividido a esta en dos estratos adecuados de tamaños $N_1 = 10{,}000$ y $N_2 = 40{,}000$, respectivamente. Usando información anterior, se determinó que las varianzas de los estratos son: $\sigma_1^2 = 50$ y $\sigma_2^2 = 500$, respectivamente. Usando la afijación proporcional y la asignación de Neyman, calcular el error estándar del estimador de la media de toda la población cuando el total de la muestra que se toma es $n = 2{,}000$.

Solución

Si la afijación es proporcional, los tamaños de la muestra en cada estrato son:

$$n_1 = \frac{10{,}000}{50{,}000}\,2{,}000 = 400,\ n_1 = \frac{40{,}000}{50{,}000}\,2{,}000 = 1{,}600$$

La varianza del estimador de la media muestral es:

$$Var(\bar{x}) = \sum_i \frac{N_i^2}{N^2}\left(1 - \frac{n_i}{N_i}\right)\frac{\sigma_i^2}{n_i} =$$

$$(10{,}000/50{,}000)^2\left(1 - \frac{400}{10{,}000}\right)\frac{50}{400} + (40{,}000/50{,}000)^2\left(1 - \frac{1{,}600}{40{,}000}\right)\frac{500}{1{,}600} = 0.2544$$

El error estándar de la media muestral es $\sqrt{0.2544} = 0.5043$.

Si los tamaños de la muestra en cada estrato son:

$$n_1 = \frac{N_1\sigma_i^2}{N_1\sigma_i^2 + N_2\sigma_2^2}\ n = \frac{10{,}000\ (50)}{10{,}000\ (50) + 40{,}000\ (500)}\,2{,}000 \approx 49$$

$$n_2 = \frac{N_2\sigma_2^2}{N_1\sigma_i^2 + N_2\sigma_2^2}\ n = \frac{40{,}000\ (500)}{10{,}000\ (50) + 40{,}000\ (500)}\ 2{,}000 \approx 1{,}951$$

En este caso, la varianza del estimador de la media muestral es:

$$Var(\bar{x}) = \sum_i \frac{N_i^2}{N^2}\left(1 - \frac{n_i}{N_i}\right)\frac{\sigma_i^2}{n_i} =$$

$$(10{,}000/50{,}000)^2\left(1 - \frac{49}{10{,}000}\right)\frac{50}{49} + (40{,}000/50{,}000)^2\left(1 - \frac{1{,}951}{40{,}000}\right)\frac{500}{1{,}951} = 0.2030$$

El error estándar es $\sqrt{0.2030} = 0.4505$.

Puede notarse que el error estándar del estimador es menor en el caso de la afijación de Neyman.

EJEMPLO. *Para repartir mejor los refrescos*

Un fabricante de bebidas gaseosas desea introducir un nuevo sistema de reparto para entregar los pedidos en menos tiempo que en el sistema tradicional.

Para lograr a cabo su objetivo, el fabricante debe conocer previamente si el cliente está de acuerdo con el nuevo sistema de reparto. Para ello confeccionó una lista de todos los establecimientos que eran clientes, y así resultaron 10,000 establecimientos, de los cuales 9,000 eran pequeñas tiendas y 1,000 eran establecimientos que en conjunto representaban un alto volumen de ventas, sin llegar a ser mayor que el de los establecimientos menores. Teniendo en cuenta esta distribución se formaron dos estratos: uno con los 9,000 clientes menores y otro con los 1,000 clientes mayores.

El cuestionario aplicado tuvo varias preguntas, pero la más importante para el estudio consideraba si el cliente estaba de acuerdo o no con el nuevo sistema de reparto; de ahí que se considerara que debía estimarse una proporción.

La afijación, como era de esperar, se realizó de manera proporcional al volumen de ventas. Para el estrato de los clientes menores, el peso resultó ser $w_1 = 0.6$ y para los clientes mayores el peso fue $w_2 = 0.4$.

El tamaño de muestra se calculó considerando que el margen de error que se podía cometer al estimar la proporción de los que estaban de acuerdo con el nuevo sistema era de 3%, con un nivel de confianza del 95%. Así se obtuvo:

$$n = \frac{(9{,}000/10{,}000)^2(1/0.6)(1/2)(1/2) + (1{,}000/10{,}000)^2(1/0.4)(1/2)(1/2)}{(0.03/1.96)^2 + (1/10{,}000)[(9{,}000/10{,}000)(1/2)(1/2) + (1{,}000/10{,}000)(1/2)(1/2)]} = 1{,}375$$

(La expresión que indica el tamaño de muestra ha sido dividida entre el cuadrado de la población total para simplificar los cálculos.) Como la afijación fue proporcional al volumen de ventas, de los estratos debería extraerse muestras con los siguientes tamaños respectivos:

$$n_1 = 1{,}375 \times 0.6 = 825 \qquad n_2 = 1{,}375 \times 0.4 = 550$$

En el primer estrato, 578 estuvieron de acuerdo con el nuevo sistema y en el segundo, 303.

La estimación de la proporción de los establecimientos que estaban de acuerdo con el nuevo sistema fue como sigue:

$p_1 = \frac{578}{825} = 0.70$, para el primer estrato y $p_2 = \frac{303}{550} = 0.55$ para el segundo estrato.

La estimación de la proporción de todos los establecimientos a favor del nuevo sistema de reparto fue:

$$\hat{p} = (9{,}000/10{,}000)(0.70) + (1{,}000/10{,}000)(0.55) = 0.685\ (68.50\%)$$

El error estándar del estimador en el estrato 1 resultó igual a:

$$\sqrt{\left(1 - \frac{825}{9{,}000}\right)\frac{0.7 \times 0.3}{825}} = 0.0152$$

El error estándar del estimador en el estrato 2 resultó igual a:

$$\sqrt{\left(1 - \frac{550}{1{,}000}\right)\frac{0.55 \times 0.45}{550}} = 0.01423$$

El error estándar del estimador de la proporción en la población resultó:

$$\sqrt{\frac{N_1^2}{N^2}\left(1 - \frac{n_1}{N_1}\right)\frac{p_1(1-p_1)}{n_1 - 1} + \frac{N_2^2}{N^2}\left(1 - \frac{n_2}{N_2}\right)\frac{p_2(1-p_2)}{n_2 - 1}} = 0.0137$$

valor que no supera al 3%.

El muestreo aleatorio estratificado tiene ciertas ventajas sobre el muestreo aleatorio simple:

- La información que se obtiene de los estratos no solo permite obtener datos de toda la población. Cuando las muestras en cada estrato son suficientemente grandes, estas se pueden usar para obtener información de cada uno de ellos.
- Los estimadores que se obtienen con este tipo de muestreo son más precisos que los que se consiguen con el muestreo simple.
- Los costos para recoger la muestra (por ejemplo los de transporte) son menores que los que se ocasionan cuando se realiza un muestreo simple.

¿Cómo elegir las unidades muestrales en cada estrato?

Establecidos los estratos, debe seleccionarse una muestra por estrato. La elección de elementos en cada estrato generalmente se hace usando el muestreo aleatorio simple; sin embargo, también se usan otras maneras, como las siguientes:

a) *La selección sistemática.* Esta manera de selección se usa como alternativa de la anterior cuando se dispone de listas o ficheros de los estratos. Según este método, para seleccionar una muestra de tamaño n de una población de tamaño N debe calcularse un intervalo de muestreo de longitud $k = N/n$, entero. En este intervalo se selecciona un número r al azar entre 1 y k, que se llama arranque aleatorio. Los elementos que se seleccionan en el estrato son los que tienen la siguiente numeración:

$$r,\ r+k,\ r+2k,\ldots,\ r+(n-1)k$$

Así, para seleccionar una muestra de tamaño 200 en un estrato que tiene 1,000 elementos se calcula el intervalo de longitud entera 1,000/200 = 5 (= k). Se selecciona al azar un número, por ejemplo 4 (arranque aleatorio). Los elementos seleccionados son los que tienen la siguiente numeración: 4, 9, 14,..., 999.

b) *La selección de números telefónicos* de una guía telefónica, cambiando el último dígito aleatoriamente. Esta selección se realiza cuando se desea recabar información de personas que tienen teléfono.

c) *La selección por rutas o itinerarios aleatorios*. Esta selección se realiza cuando las unidades muestrales son las viviendas. Se seleccionan puntos aleatorios en la ciudad como partida de cada ruta. Las casas o viviendas a encuestar por la persona encargada del diseño de la muestra serán las situadas en la ruta.

En muchas ocasiones es posible aplicar la teoría descrita; sin embargo, se debe considerar las ocasiones en donde es imposible emplearla por una serie de dificultades que pueden presentarse, como por ejemplo las relacionadas con:

- La localización de las unidades muestrales.
- El cálculo de las varianzas poblacionales.
- La necesidad de estudiar varios parámetros a la vez.
- El costo de la muestra.
- La afijación de la muestra.

La dificultad de localizar unidades muestrales aparece cuando no se tienen marcos muestrales para cada uno de los estratos. En tal caso se deja al encuestador la libertad de que elija, en cada estrato, a los individuos que debe entrevistar, cubriendo los tamaños de muestra asignados; por ejemplo, debe seleccionarse para la entrevista

35 hombres mayores de 25 años de la clase media, y así para los diferentes grupos. A este tipo de muestreo no probabilístico se le llama muestreo por cuotas. El número de elementos seleccionados se denomina cuota. Generalmente la cuota que se asigna para cada estrato está relacionada con información ya existente, como por ejemplo los censos nacionales.

Cuando no se conocen las varianzas poblacionales puede recurrirse a información anterior o a pequeñas muestras piloto para aproximarlas.

El desarrollo realizado acerca de los errores y del tamaño de muestra está relacionado con la estimación de un solo parámetro; sin embargo, en la mayoría de encuestas se trata de estimar más de un parámetro poblacional, sea para aprovechar la encuesta o para consolidar los resultados. Cuando esto sucede se elige el parámetro más relevante, y de acuerdo a este se escoge el tamaño de la muestra y el error estándar que le corresponda.

La ponderación de la muestra es uno de los remedios a las dificultades que pueden presentarse por la falta de representación de la población en la muestra. Por ser un procedimiento que se usa a menudo, se desarrolla a continuación.

Ponderación en el muestreo estratificado después de la selección de la muestra

Muchas veces la ubicación de las unidades muestrales en sus estratos correctos solo es posible después de que la muestra ha sido seleccionada. Por ejemplo, para estimar el promedio del tiempo por día que una persona ocupa el teléfono se estratifica a la población en hombres y mujeres, y luego se realiza una encuesta por teléfono. Como es de suponer, la afijación que teóricamente se hubiera realizado generalmente no se cumple en la práctica, pues solo se conoce si el encuestado es hombre o mujer cuando este responde. Otras veces, la *afijación no proporcional* es realizada obligatoriamente, sobre todo si se desea analizar con mayor detalle algunos estratos a los cuales les hubiera correspondido un tamaño de muestra inferior.

La falta de proporcionalidad no reviste ningún problema si solo se requieren análisis individuales y/o comparativos entre los estratos. Sin embargo, si se desean estimaciones muestrales de toda la población es necesario *ponderar la muestra*. Esto significa asignar pesos o ponderaciones a cada estrato de tal manera que se logre compensar la desigual probabilidad de selección dada a cada unidad de la población que compone el estrato.

EJEMPLO. *Para conocer a los lectores*

Se desea estimar en una población, en donde el 60% son varones, el promedio de horas por semana que las personas se dedican a la lectura. Para llevar a cabo esta tarea se tomó una muestra total de $n = 150$, de los cuales $n_1 = 40$ eran hombres y $n_2 = 110$, mujeres. Los resultados fueron como sigue.

TABLA 14.3 *Resumen de los resultados*

Hombres	*Mujeres*
$\bar{x}_1 = 7$ horas	$\bar{x}_1 = 4$ horas
$n_1 = 40$	$n_1 = 110$

Tomando en cuenta tan solo las proporciones muestrales, el estimador de la media poblacional sería:

$$\frac{40}{150}\ 7 + \frac{110}{150}\ 4 = 4.08 \text{ horas}$$

Sin embargo, la muestra no es autoponderada (la proporcionalidad en la muestra no es igual a la de la población), y por lo tanto el estimador no es adecuado. Para corregir este inconveniente se usan las ponderaciones de la población, y así se tiene que un estimador de la media poblacional es:

$$0.6 \times 7 + 0.4 \times 4 = 5.8 \text{ horas}$$

En la práctica, con esta técnica los estimadores que se obtienen tienen casi la misma precisión que los conseguidos mediante el muestreo estratificado con afijación proporcional cuando los tamaños de las muestras encontradas en cada estrato son mayores o iguales que 20.

EJEMPLO. *Para saber el porcentaje de los estudiantes que concurren al cine*

En un estudio realizado en la universidad para conocer la proporción de estudiantes que van al cine se tomó una muestra en la que se incluía una sobrerrepresentación de estudiantes de ciertas facultades. Este hecho daba lugar a que se tuviera una muestra no proporcional al tamaño de cada una de las facultades, y por tanto, a que la probabilidad de que un estudiante elegido en la misma muestra fuera diferente según su facultad de procedencia. A continuación el resumen de los resultados.

TABLA 14.4 *Resumen de los resultados*

Facultad	*% que van al cine*	*% de muestra*	*Tamaño de la muestra*	*% de la población*	*Tamaño de la población*
CC. II.	40	50	250	30	2,400
Letras	20	30	150	60	4,800
Artes y otros	35	20	100	10	1,600
Total			500		8,800

Según estos resultados, el 40% de los estudiantes de CC. II., el 20% de estudiantes de Letras y el 35% de los estudiantes de Arte van al cine. Se observa que en CC. II. se han hecho más entrevistas que en Letras, lo que no se justifica, pues en Letras hay más estudiantes que en Ingeniería.

El problema aparece cuando se desea estimar el porcentaje global de los que van al cine, independientemente de la facultad de procedencia, no así cuando se comparan los porcentajes de estudiantes que van al cine en las diversas facultades.

Un primer intento para estimar el porcentaje de los que van al cine en toda la universidad corresponde al cálculo de la media ponderada de los porcentajes encontrados, tomando como ponderaciones a los porcentajes de las muestras. Así se tiene:

$$\bar{y}_w = 40 \times 0.5 + 20 \times 0.3 + 35 \times 0.2 = 33.00\%$$

La media que se acaba de calcular otorga a cada facultad el peso que tiene en los datos con los que se está trabajando, y que constituyen la muestra. Pero la información que se tiene en la Tabla 14.5 indica que esta media no refleja la realidad: la facultad de CC. II. tiene el 30% del total de la población, mientras que en la muestra acapara el 50% del total.

Para acercarnos a la realidad es necesario ponderar la muestra según el peso que tiene en la población, lo que significa calcular otra proporción, dando a cada facultad su peso en la población. De este modo se tiene:

$$\bar{y} = 40 \times 0.30 + 20 \times 0.60 + 35 \times 0.10 = 27.5\%$$

valor menor al que obteníamos cuando se ponderaba según el peso en la muestra.

En el ejemplo anterior, la ponderación toma en cuenta el peso que cada facultad tiene en la población; sin embargo, puede resultar útil hacer la ponderación dando pesos a *cada individuo* y no a las facultades.

En el ejemplo, para la facultad CC. II., un estudiante de la muestra tomada representa a 2,400/250 = 9.6 estudiantes de la población correspondiente, mientras que para Letras un estudiante de la muestra representa a 32 de la población respectiva. Para las otras facultades un estudiante de la muestra representa a 16 de la población respectiva. Esto significa que para cada individuo de CC. II., Letras y de las otras facultades, los pesos serán 9.6/57.6, 36/57.6 y 16/57.6, respectivamente.

LA ESTADÍSTICA EN LA EMPRESA

La empresa de investigación de mercados Imaginim

La empresa de investigación de mercados Imaginim tiene su origen en la unión de la empresa IMAG, que opera en más de 100 países en el mundo, con la empresa nacional NIM, reconocida por su actitud innovadora y de servicio al cliente.

Imaginim reúne una cantidad respetable de profesionales de la estadística dedicados a la investigación para mejor aconsejar a los decisores del mercado en la óptima atención de las necesidades de ciudadanos y consumidores.

Imaginim opera en la región sur de América Latina y sus actividades se desarrollan en grandes áreas del marketing y de la inteligencia de mercados. La estadística y la minería de datos son las herramientas que esta empresa emplea a menudo para desarrollar trabajos de muestreo, segmentaciones del mercado, presentación de nuevos productos, la elección de mejores clientes, mejoras en las ventas, prevención de fugas de clientes y problemas generales de clasificación. La preparación que tienen sus profesionales en la estadística es la garantía para brindar a los clientes soluciones altamente confiables.

Como es de suponer, la estadística es la principal herramienta de trabajo de esta empresa.

EJERCICIOS

1. Se desea conocer el pago promedio mensual por servicios de agua potable que realizan los pobladores de un distrito de la ciudad. El distrito está formado por 2,000 hogares y está dividido en tres barrios, A, B y C, de 800, 700 y 500 hogares, respectivamente. Una muestra aleatoria estratificada de los hogares del distrito de tamaño 400 está formada de 160 hogares de A, 140 de B y 100 de C. Las medias muestrales del consumo de agua facturado fueron como sigue: $ 80 para A, $ 60 para B y $ 50 para C, y las respectivas desviaciones estándar fueron: $ 10, $ 15 y $ 20, respectivamente.

   a) Usar el método de estimación estratificado para estimar el promedio de pagos mensuales por servicio de agua de todos los hogares del distrito.

   b) Hallar el intervalo de confianza al 95% para el promedio de pagos mensuales por servicio de agua de todos los hogares del distrito.

2. Una empresa dedicada a los restaurantes tiene 40 en el norte del país, 30 en el centro y 25 en el sur. Con la finalidad de averiguar la satisfacción del servicio por los clientes, la empresa realiza un muestreo estratificado tomando una muestra de 15 restaurantes del norte, 10 del centro y 8 del sur. Durante una semana a los clientes que concurren a los restaurantes se les solicita que califiquen los servicios en general utilizando una escala del 1 al 100. Las medias y las desviaciones estándar muestrales obtenidas fueron como sigue:

   Para el norte: $\bar{x}_N = 68.00$, $s_N = 10$

   Para el centro: $\bar{x}_C = 55.00$, $s_C = 15$

   Para el sur: $\bar{x}_S = 75.00$, $s_S = 20$

   Usando un intervalo de confianza del 95% estimar la media de las calificaciones del servicio de todos los restaurantes.

3. De las 100 escuelas de negocios que hay en el país 60 son estatales y 40 son particulares. Con la finalidad de estimar el promedio de la edad de los matriculados en las 100 escuelas, el Ministerio de Educación debe seleccionar una muestra para luego aplicar la afijación proporcional en ambos grupos de escuelas. Se sabe que la varianza de las edades es 100 para las escuelas estatales y 144 para las escuelas particulares. Indicar el tamaño de muestra a tomar si se desea que el margen de error del estimador muestral de la media sea 11, con un nivel de confianza de 95%. Indicar el tamaño de muestra total para el caso en que la afijación muestral sea proporcional y el tamaño de muestra que se tomaría para cada grupo de escuelas.

4. En el contexto del ejercicio anterior, hallar el tamaño de muestra a tomar en cada grupo de escuelas si se desea estimar el porcentaje de mujeres que estudian en las 100 escuelas.

5. Se desea conocer la opinión de los ciudadanos (personas de 18 años a más) de un país sobre las elecciones realizadas en el mes pasado. El parámetro más importante a estimar se refiere al porcentaje de personas de una determinada zona que están de acuerdo con que las elecciones se realizaron limpiamente.

   Para efectuar la encuesta, considerar que la zona deberá estratificarse de acuerdo al tipo de hábitat (urbano y rural) y a su tamaño poblacional. Para que una persona que vive en una ciudad grande tenga igual posibilidad de ser elegida que una que vive en una ciudad pequeña la muestra será repartida de manera proporcional al tamaño de la población.

Se ha considerado la elección de 150 ciudades repartidas proporcionalmente en cada región, teniendo en cuenta el hábitat y el tamaño de la población. Las poblaciones serán elegidas de manera sistemática.

Para tomar una muestra de *n* ciudades de manera sistemática de un grupo de *k* ciudades cuya población total es *N*, se seguirá el siguiente procedimiento:

a. Listar las ciudades y sus respectivas poblaciones.

b. Calcular el intervalo de muestreo: $N/n = c$.

c. Todas las ciudades cuya población es mayor o igual a *c* entrarán a formar parte de la muestra.

d. Las ciudades elegidas en c) se eliminan manteniéndose las restantes, acumulando su población.

e. Se selecciona un número *r* al azar entre 1 y *c*. Las ciudades cuyas poblaciones acumuladas contengan a los números r, $r + c$, $r + 2c$... completarán la muestra.

Por ejemplo, para elegir 2 ciudades de las 6 que aparecen a continuación:

Ciudades	*Población*
A	200,000
B	50,000
C	100,000
D	20,000
E	800,000
F	500,000
Total	1,670,000

el intervalo de muestreo es 1,170,000/3 = 390,000. La ciudad E entra en la muestra.

Listar las ciudades restantes acumulando las poblaciones.

Ciudades	*Población*
A	200,000
B	250,000
C	350,000
D	370,000
Total	370,000

Seleccionar un número al azar entre 1 y 390,000, por ejemplo, 180,500. Se elige a la ciudad A, pues esta contiene al número 180,500.

La entrevista se hará en el domicilio de los encuestados usando las rutas aleatorias y el método de Kish. Las revisitas y las llamadas telefónicas se llevarán a cabo si fuera necesario.

Se desea diseñar un muestreo para estudiar las actitudes de las personas mayores de 18 años de toda la región norte acerca del proyecto de la ley de aguas.

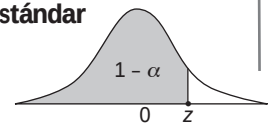
La encuesta deberá realizarse personalmente en el domicilio del encuestado. Las respuestas, se cree, pueden variar según la edad y las subregiones. La persona que diseña la encuesta deberá considerar la población total de la región y de cada una de las cuatro subregiones que conforman la región, así como la distribución de las edades en cada una de las subregiones.

La muestra a tomar deberá permitir un error máximo de 3% para toda la región (dado el tamaño de la población, se puede tomar como tamaño de muestra el que le corresponde al muestreo aleatorio simple) y un nivel de confianza del 95%. La afijación de la muestra en cada una de las subregiones se hará proporcional a la población, pero de tal manera que el error máximo no supere el 4.5% (si después de hacer la afijación el error fuera mayor, recalcular el tamaño de la muestra con la fórmula correspondiente).

Apéndices

APÉNDICE A Tabla de áreas acumuladas de la distribución normal estándar

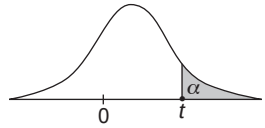
Ejemplo: Para $z = 1.14$, el área acumulada $1 - \alpha$ es igual a 0.8729



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9762	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998

APÉNDICE B Valores críticos de la distribución *t*

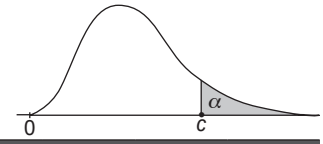
Ejemplo: Para 6 grados de libertad y un área $\alpha = 0.05$ en la cola superior, $t = 1.943$



	ÁREA α DE LA COLA DERECHA					
Grados de libertad	0.005	0.01	0.025	0.05	0.10	0.15
1	63.657	31.821	12.706	6.314	3.078	1.963
2	9.925	6.965	4.303	2.92	1.886	1.386
3	5.841	4.541	3.182	2.353	1.638	1.25
4	4.604	3.747	2.776	2.132	1.533	1.19
5	4.032	3.365	2.571	2.015	1.476	1.156
6	3.707	3.143	2.447	1.943	1.44	1.134
7	3.499	2.998	2.365	1.895	1.415	1.119
8	3.355	2.896	2.306	1.86	1.397	1.108
9	3.25	2.821	2.262	1.833	1.383	1.1
10	3.169	2.764	2.228	1.812	1.372	1.093
11	3.106	2.718	2.201	1.796	1.363	1.088
12	3.055	2.681	2.179	1.782	1.356	1.083
13	3.012	2.65	2.16	1.771	1.35	1.079
14	2.977	2.624	2.145	1.761	1.345	1.076
15	2.947	2.602	2.131	1.753	1.341	1.074
16	2.921	2.583	2.12	1.746	1.337	1.071
17	2.898	2.567	2.11	1.74	1.333	1.069
18	2.878	2.552	2.101	1.734	1.33	1.067
19	2.861	2.539	2.093	1.729	1.328	1.066
20	2.845	2.528	2.086	1.725	1.325	1.064
21	2.831	2.518	2.08	1.721	1.323	1.063
22	2.819	2.508	2.074	1.717	1.321	1.061
23	2.807	2.5	2.069	1.714	1.319	1.06
24	2.797	2.492	2.064	1.711	1.318	1.059
25	2.787	2.485	2.06	1.708	1.316	1.058
26	2.779	2.479	2.056	1.706	1.315	1.058
27	2.771	2.473	2.052	1.703	1.314	1.057
28	2.763	2.467	2.048	1.701	1.313	1.056
29	2.756	2.462	2.045	1.699	1.311	1.055
30	2.75	2.457	2.042	1.697	1.31	1.055
40	2.704	2.423	2.021	1.684	1.303	1.05
60	2.66	2.39	2	1.671	1.296	1.046
120	2.617	2.358	1.98	1.658	1.289	1.041
Grande	2.576	2.326	1.96	1.645	1.282	1.036

APÉNDICE C Tabla de valores críticos de la distribución ji-cuadrado

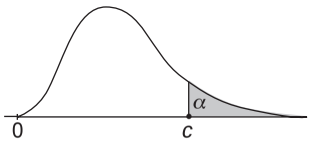
Ejemplo: Para 6 grados de libertad y un área a la derecha de α = 0.05 en la cola superior, c = 12.59



	ÁREA α									
Grados de libertad	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.64	7.88
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.12	0.22	0.35	0.58	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	2.7	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.6	3.05	3.82	4.58	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.4	5.23	6.3	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.08	4.66	5.63	6.57	7.79	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.87	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.2	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.9	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.2	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.4	13.85	15.66	33.2	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.2	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.57	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.6	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.4	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.40	104.20
80	51.17	53.54	57.15	60.39	64.28	96.58	101.90	106.60	112.30	116.30
90	59.2	61.75	65.65	69.13	73.29	107.60	113.10	118.10	124.10	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.30	129.60	135.80	140.20

APÉNDICE D Distribución F

Ejemplo: Para 5 grados en el numerador y 3 grados en el denominador y un área de α = 0.025 en la cola superior el valor de *c* es 14.90



		GRADOS DE LIBERTAD EN EL NUMERADOR													
Grados de libertad del denominador	α	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**120**
1	0.050	161.00	200.00	216.00	225.00	230.00	234.00	237.00	239.00	241.00	242.00	244.00	246.00	248.00	253.00
	0.025	648.00	800.00	864.00	900.00	922.00	937.00	948.00	957.00	963.00	969.00	977.00	985.00	993.00	1,014.00
2	0.050	18.50	19.00	19.20	19.20	19.30	19.30	19.40	19.40	19.40	19.40	19.40	19.40	19.40	19.50
	0.025	38.50	39.00	39.20	39.20	39.30	39.30	39.40	39.40	39.40	39.40	39.40	39.40	39.40	39.50
	0.010	98.50	99.00	99.20	99.20	99.30	99.30	99.40	99.40	99.40	99.40	99.40	99.40	99.40	99.50
	0.005	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00	199.00
3	0.050	10.10	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.55
	0.025	17.40	16.00	15.40	15.10	14.90	14.70	14.60	14.50	14.50	14.40	14.30	14.30	14.20	13.90
	0.010	34.10	30.80	29.50	28.70	28.20	27.90	27.70	27.50	27.30	27.20	27.10	26.90	26.70	26.20
	0.005	55.60	49.80	47.50	46.20	45.40	44.80	44.40	44.10	43.90	43.70	43.40	43.10	42.80	42.00
4	0.050	6.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.66
	0.025	12.20	10.60	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.31
	0.010	21.20	18.00	16.70	16.00	15.50	15.20	15.00	14.80	14.70	14.50	14.40	14.20	14.00	13.60
	0.005	31.30	26.90	24.30	23.20	22.50	22.00	21.60	21.40	21.10	21.00	20.70	20.40	20.20	19.50
5	0.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.40
	0.025	10.00	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.07
	0.010	16.30	13.30	12.10	11.40	11.00	10.70	10.50	10.30	10.20	10.10	9.89	9.72	9.55	9.11
	0.005	22.80	18.30	16.50	15.60	14.90	14.50	14.20	14.00	13.80	13.60	13.40	13.10	12.90	12.30
6	0.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.70
	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	4.90
	0.010	13.70	10.90	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	6.97
	0.005	18.60	14.50	12.90	12.00	11.50	11.10	10.80	10.60	10.40	10.20	10.00	9.81	9.59	9.00

7	0.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.27
	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.20
	0.010	12.20	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	5.74
	0.005	16.20	12.40	10.90	10.10	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.19
8	0.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	2.97
	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.73
	0.010	11.30	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	4.95
	0.005	14.70	11.00	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.06
9	0.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.75
	0.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.39
	0.010	10.60	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.40
	0.005	13.60	10.10	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.23	6.03	5.83	5.30
10	0.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.84	2.77	2.58
	0.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.14
	0.010	10.00	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.00
	0.005	12.80	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.66	5.47	5.27	4.75
12	0.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.34
	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	2.79
	0.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.45
	0.005	11.80	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20	5.09	4.91	4.72	4.53	4.01
15	0.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.11
	0.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.46
	0.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	2.96
	0.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.25	4.07	3.88	3.37
20	0.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	1.90
	0.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.16
	0.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.52
	0.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.68	3.50	3.32	2.81
30	0.050	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.68
	0.025	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	1.87
	0.010	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.11
	0.005	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.18	3.01	2.82	2.30

Bibliografía

Agresti, A. (2002). *A Introduction to Categorical Data Analysis*. Hoboken, N. J.: John Wiley and Sons.

Ash, R. B. (1970). *Basic Probability Theory*. New York: John Wiley and Sons.

Barret, R. (1991). *Handbook of Statistical Methods in Manufacturing*. Englewood Cliffs, N. J.: Prentice Hall.

Barry, R. J. (1981). *Probabilidade: um curso em nivel intermédio*. Río de Janeiro: IMPA.

Bass, I. (2007). *Six Sigma Statistics with EXCEL and MINITAB.* New York: McGraw-Hill.

Besterfield, D. (2009). *Quality Control*. New Jersey: Pearson Prentice Hall.

Box, H. (1993). *Estadística para investigadores.* Barcelona: Reverté.

Brockwell, P. J. & Davis, R. A. (1987). *Time Series: Theory and Methods*. New York: Springer Verlag.

Calot, G. (1973). *Cours de statistique descriptive*. París: Dunod.

Casella, G. & Berger, R. (1990). *Statistical Inference.* Belmont, California: Duxbury Press.

Chamber, J., Cleveland, M., Kleiner, B. & Tukey, P. A. (1975). *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth International Group.

Dagnelie, P. (1976). *Theorie et methods statistiques* (2 vols.). Gembloux: Presses Agronomiques de Gembloux.

Grande, E. I. (2009). *Fundamentos y técnicas de la investigación comercial*. Madrid: ESIC.

Gutiérrez Pulido, H. y De la Vara Salazar, R. (2004). *Control estadístico de la calidad y seis sigma*. México D. F.: McGraw-Hill.

Hanke, J. y Wichern, D. (2006). *Predicción en los negocios* (8ª ed.). México: Pearson Educación.

Harvard Business Review (2001). *La toma de decisiones*. Barcelona: Deusto.

Hoaglin, D. C., Mosteller, F. & Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley and Sons.

Howard, R. (1978). *Análisis de la decisión empresarial*. Bilbao: Fondo Educativo Interamericano S. A.

Iglewicz, D. & Hoagling, D. (1993). *How to Detect and Handle Outliers*. Milwauke: ASQ Quality Press.

Kenett, R. y Zacks, S. (1998). *Estadística industrial moderna*. México D. F.: Internacional Thomson Editores.

Kuehl, R. (2001). *Diseño de experimentos* (2ª ed.). México D. F.: Thomson-Learning.

Lebart, L. y Fenelon, J. P. (1973). *Statistique et informatique appliquées*. París: Dunod.

Ledolter, A. (1983). *Statistical Methods for Forecasting*. New York: John Wiley and Sons.

Lohr, S. (2000). *Muestreo: diseño y análisis*. México D. F.: Thomson.

Makridakis, S. & Heelwright, S. (1978). *Forecasting. Methods and Applications*. New York: John Wiley and Sons.

Montgomery, D. (2002). *Diseño y análisis de experimentos* (2ª ed.). México D. F.: Limusa Wiley.

Montgomery. D., Peck, Vinny, (2002). *Introducción al Análisis de la Regresión Lineal*. CECSA. (3ª ed.).

Montserrat, F. y otros (1992). *Análisis exploratorio de datos*. Barcelona: Colección LCT.

Morgan, J. J. (1977). *Introducción a la teoría de decisiones*. México D. F.: Representaciones y Servicios de Ingeniería S. A.

Myatt, G. (2007). *Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining*. Hoboken: John Wiley and Sons.

Pardo, L. y Valdés, T. (1987). *Simulación. Aplicación práctica en la empresa.* Madrid: Díaz de Santos.

Peña, D. (1987). *Estadística* (2 tomos). Madrid: Alianza Editorial.

Peña, D. (2001). *Fundamentos de estadística* (2 tomos). Madrid: Alianza Editorial.

Powers, D. & Xie, Y. (1999). *Methods for Categorical Data Analysis*. San Diego: Academic Press.

Pulido, R., Rodríguez, A. y Olmo, J. (2005). *Control estadístico de la calidad*. Granada: Grupo Editorial Universitario.

Ripley, B. (1987). *Simulación estocástica.* New York: John Wiley and Son, Inc.

Ross, S. (2005). *Simulation*. Harcourt, India: PVT Ltd.

Ross, S. (2000). *Probabilidad y estadística* (2ª ed.). México D. F.: McGraw-Hill.

Salinas, J. (2008). *Análisis de decisiones estratégicas.* Buenos Aires: Cengage.

Sánchez Carrión, J. J. (1989). *Análisis de tablas categóricas*. Madrid: CIS.

Scheaffer, R. L., Mendenhall, W. y Lyman, O. (2007). *Elementos de muestreo* (6ª ed.). Madrid: Thomson.

Tucker, H. G. (1966). *Introducción a la teoría matemática de las probabilidades y la estadística*. Barcelona: Vicens-Vives.

Ullman, J. (1979). *Métodos cuantitativos en administración*. México D. F.: McGraw-Hill.

Véliz, C. (2000). *Estadística. Aplicaciones.* Lima: SCGSA.

Weisberg, S. (2005). *Applied Linear Regression.* Hoboken, New Jersey: Wiley-Interscience.