INSTITUTO TECNOLÓGICO SUPERIOR "JAPÓN"

Guía Metodológica De Estadística I




Compilado por:
Ing. José Luis Herrera

Carrera: Administración de Empresas

2019

1. IDENTIFICACIÓN DE

Nombre de la Asignatura: **ESTADÍSTICA I**	**Componentes del Aprendizaje**	

Resultado del Aprendizaje:

COMPETENCIAS Y OBJETIVOS

- Usa de manera apropiada los conceptos básicos de estadística**.**
- Produce la información Estadística requerida de acuerdo a las necesidades de los usuarios para la planificación y gestión administrativa.
- Calcula e interpreta las Medidas de Resumen aplicando el procedimiento pertinente según se trate de datos agrupados y no agrupados.
- Analiza de dependencia ya asociación de variables estadísticas mediante el uso de modelos lineales y de medidas de asociación en el estudio de los fenómenos económico-financieros.

Docente de Implementación:

JOSE LUIS HERRERA	**Duración:** 30 horas

Unidades	Competencia	Resultados de Aprendizaje	Actividades	Tiempo de Ejecución
Elemento, población, caracteres. Variables estadísticas.	Conoce y analiza las características que definen el razonamiento estadístico	Valora la importancia de los conceptos básicos de la estadística.	Desarrolla ejercicios prácticos.	**5**

Medidas de Tendencia Central: medias, mediana y moda. Propiedades de la media aritmética. Cuartiles, deciles y percentiles. Características comparativas de las medidas de tendencia central.	Aplicación de las formulas., cálculos e interpretaciones. Usos.	Valora la importancia de la información estadística a recopilar. Comparte la importancia	Desarrollo de trabajo experimental exposición.	**5**
Medidas de dispersión: absolutas y relativas. Propiedades de la varianza. Otras medidas de dispersión: momentos.	Aplicación de fórmulas Cálculos e interpretaciones. Usos	Resalta la importancia de cada medida de dispersión y el uso peculiar correspondiente en la Economía.	Desarrollo de trabajo experimental. Exposición	**5**

Tabla de doble entrada. Distribuciones marginales. Distribuciones condicionadas. Independencia estadística.	Valora la importancia de las distribuciones bidimensionales	Valora la importancia de las distribuciones bidimensionales Valora la importancia del uso de distribuciones bidimensionales.	Desarrollo de trabajo experimental. Exposición.	**5**
Idea de ajuste. Ajuste lineal. Modelo de regresión lineal. Otros modelos. Coeficientes de correlación	Confecciona el diagrama de dispersión que genera una muestra de pares ordenados. Aplica formulas. Realiza los cálculos, interpreta los resultados e indica el uso correspondiente	Juzga de mucha importancia el conocimiento, interpretación y uso de las medidas de regresión.	Desarrollo de trabajo experimental. Exposición.	**5**
Distribución de frecuencias con intervalos cerrados y semi abiertos.	Clasifica los datos empleando tablas de distribución con intervalos cerrados o semi abiertos.	Demuestra habilidad, conocimiento y destreza en la aplicación de las formas de cálculo, mediante resolución de casos en clase y en grupo.	Desarrollo de trabajo experimental. Exposición.	**5**

2. CONOCIMIENTOS PREVIOS Y RELACIONAD

Co-requisitos

3. UNIDADES TEÓRICAS

- **Desarrollo de las Unidades de Aprendizaje (contenidos)**

  A. Base Teórica

1. Elemento, población, caracteres. Variables estadísticas

Establecemos a continuación algunas definiciones de conceptos básicos y fundamentales básicas como son: elemento, población, muestra, caracteres, variables, etc.

- Elementos o unidades: personas u objetos que contienen cierta información que se desea estudiar.
- Población: conjunto de individuos o elementos que cumplen ciertas propiedades comunes.

En relación al tamaño de la población, esta puede ser:

  - Finita, como es el caso del número de personas que llegan al servicio de urgencia de un hospital en un día;
  - Infinita, si por ejemplo estudiamos el mecanismo aleatorio que describe la secuencia de caras y cruces obtenida en el lanzamiento repetido de una moneda al aire.

- Muestra: subconjunto representativo de una población.
- Parámetro: función definida sobre los valores numéricos de características medibles de una población.
- Estadístico: función definida sobre los valores numéricos de una muestra.
- Caracteres: propiedades, rasgos o cualidades de los elementos de la población. Estos caracteres pueden dividirse en cualitativos y cuantitativos.

- Modalidades: diferentes situaciones posibles de un carácter. Las modalidades deben ser a la vez exhaustivas y mutuamente excluyentes --cada elemento posee una y solo una de las modalidades posibles.

- Clases: conjunto de una o más modalidades en el que se verifica que cada modalidad pertenece a una y solo una de las clases.

- VARIABLE ESTADÍSTICA: Es una representación, a través de números u otros símbolos, de una variable. Esta representación se obtiene mediante algún procedimiento de medición. Ejemplos: 1) Cantidad de palabras recordadas de una lista de 12.

Las variables estadísticas se clasifican de acuerdo con el tipo de valores que pueden tomar en:

- Variable cualitativa. Es aquella cuyos valores expresan atributos. Ejemplo: Tipo de trastorno que presentan los pacientes de un servicio de salud mental (de ansiedad, de atención, de sueño, etc.)
- Variable cuasi-cuantitativa. Es aquella cuyos valores indican un orden o jerarquía. Ejemplo: Nivel de deserción escolar (bajo, medio, alto).
- Variable cuantitativa. Es aquella cuyos valores expresan cantidades numéricas. Dentro de las variables cuantitativas se diferencian las llamadas discretas de las continuas. Se consideran discretas aquellas cuyos valores son puntos aislados; esto es, cuando todo valor tiene un consecutivo. Se dice que dos valores son consecutivos cuando no puede existir un valor de la variable entre ellos. Ejemplo: Cantidad de palabras recordadas. Se consideran continuas a las variables que, al menos teóricamente, pueden tomar cualquier valor dentro de un intervalo numérico. Ejemplo: Tiempo de reacción ante un estímulo. Hay variables que no son discretas ni continuas pero no se tratarán en este curso.

2. Medidas de Tendencia Central: medias, mediana y moda.

Al describir grupos de diferentes observaciones, con frecuencia es conveniente resumir la información con un solo número. Este número que, para tal fin, suele situarse hacia el centro de la distribución de datos se denomina medida o parámetro de tendencia central o de centralización. Cuando se hace referencia únicamente a la posición de estos parámetros dentro de la distribución, independientemente de que esté más o menos centrada, se habla de estas medidas como medidas de posición. En este caso se incluyen también los cuantiles entre estas medidas.
Entre las medidas de tendencia central de las que las más importantes en el ámbito estadístico, matemático y contable son:

- Media aritmética

En matemáticas y estadística, la media aritmética, también llamada promedio o media, de un conjunto finito de números es el valor característico de una serie de datos cuantitativos, objeto de estudio que parte del principio de la esperanza matemática o valor esperado, se obtiene a partir de la suma de todos sus valores dividida entre el número de sumandos.

Cuando el conjunto es una muestra aleatoria recibe el nombre de media muestral siendo uno de los principales estadísticos muéstrales.

- Mediana (Me)

Dato que se encuentra en el lugar central de todos los datos de un estudio cuando estos están ordenados de menor a mayor.

- Moda (Mo)

La moda es el dato más repetido de la encuesta, el valor de la variable con mayor frecuencia absoluta. En cierto sentido la definición matemática corresponde con la locución "estar de moda", esto es, ser lo que más se lleva.

Su cálculo es extremadamente sencillo, pues solo necesita un recuento. En variables continuas, expresadas en intervalos, existe el denominado intervalo modal o, en su defecto, si es necesario obtener un valor concreto de la variable, se recurre a la interpolación.

Por ejemplo, el número de personas en distintos vehículos en una carretera: 5-7-4-6-9-5-6-1-5-3-7. El número que más se repite es 5, entonces la moda es 5.

Calificaciones	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Número de alumnos	2	2	4	5	8	9	3	4	2

Hablaremos de una distribución bimodal de los datos, cuando encontremos dos modas, es decir, dos datos que tengan la misma frecuencia absoluta máxima. Cuando en una distribución de datos se encuentran tres o más modas, entonces es multimodal. Por último, si todas las variables tienen la misma frecuencia diremos que no hay moda.

Cuando tratamos con datos agrupados en intervalos, antes de calcular la moda, se ha de definir el intervalo modal. El intervalo modal es el de mayor frecuencia absoluta.

3. Propiedades de la media aritmética.

- Cálculo sencillo.
- Interpretación muy clara.
- Al depender solo de las frecuencias, puede calcularse para variables cualitativas. Es por ello el parámetro más utilizado cuando al resumir una población no es posible realizar otros cálculos, por ejemplo, cuando se enumeran en medios periodísticos las características más frecuentes de determinado sector social. Esto se conoce informalmente como "retrato robot".
- Se puede calcular en relación con un conjunto cualquiera de datos numéricos, de manera que siempre existe.
- Un conjunto de datos numéricos tiene una y sólo una media, de modo que siempre es única.
- Se presta a un tratamiento estadístico más profundo (por ejemplo, las medias de varios conjuntos de datos se pueden combinar el total de los datos).
- Es relativamente confiable en el sentido de que las medias de muchas muestras tomadas de la misma población por lo general no fluctúan, o varían, tan ampliamente como otras medidas estadísticas que se emplean para estimar la media de una población m
- Toma en cuenta todos y cada uno de los elementos de los datos
- La media no es confiable en reportes.

Las muestras contienen a veces valores muy pequeños o muy grandes, que se apartan tanto del cuerpo principal de los datos que es cuestionable su inclusión en las muestras. Cuando estos valores, llamados valores aberrantes, se promedian con otros, pueden afectar a la media, a tal grado que se vuelve debatible su valor como una descripción razonable respecto del "centro" de los datos.

4. Cuartiles, deciles y percentiles.

Aunque la varianza y la desviación estándar son la medidas de dispersión más útiles en análisis estadístico, existen otras técnicas con las cuales puede medirse la dispersión de un conjunto de datos. Estas medidas adicionales de dispersión son los cuartiles, los deciles y los percentiles.

- Cuartiles: Son valores de la variable que dividen los datos ordenados en cuartos; cada Conjunto de datos tiene tres cuartiles. El primer cuartil, $1Q_1$, es un número tal que a lo sumo 25% de los datos son menores en valor que 1 Q1 y a lo sumo 75% son mayores. El segundo cuartil es la mediana (50%). El tercer cuartil, $3Q_3$, es un número tal que a lo sumo 75% de los datos son menores en valor que 3 Q_3 y a lo sumo 25% son mayores.

- Deciles. Son valores de la variable que dividen los datos ordenados en diez partes iguales (9 divisiones).
- Percentiles. Son los valores de la variable que dividen un conjunto de datos clasificados en 100 subconjuntos iguales; cada conjunto de datos tiene 99 percentiles. El k-ésimo percentil, kP, es un valor que a lo sumo k% de los datos son menores en valor que k P y a lo sumo (100 - k)% de los datos son mayores.

Ejemplo 1 Para la siguiente colección de datos 1, 1, 1, 2, 3, 3, 4, 4, 5, 9, 9, 19, 20 y 20 calcule:

a) El primero y el tercer cuartil. Ubíquelos en un diagrama de caja.
b) El octavo decil.
c) El percentil 42, el 50 y el 87.

Solución Cuartiles: Para el cálculo de los cuartiles debemos determinar la posición del dato que ocupa cada cuartil con la condición de que dividan a la colección de datos en cuatro partes iguales. De esta forma encontramos con que el segundo cuartil coincide con la mediana dado que divide a la colección en dos partes iguales, por lo que su posición es

$$\frac{n+1}{2} \quad \frac{14+1}{2}=7.5$$

1 1 1 2 3 3 4 4 ↑ 5 9 9 19 20 20

Posición 7.5

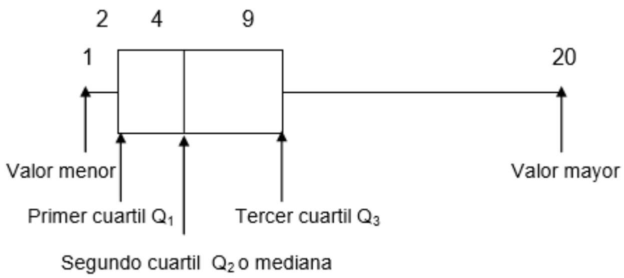
Valor de Q2 = 4

Esto significa que la mediana es 4. Ahora el primer cuartil es la mediana de los datos que se encuentran a la izquierda de la mediana o segundo cuartil. Así, la posición del primer cuartil

$$\frac{n+1}{2}=\frac{7+1}{2}=4.$$

Esto significa que el primer cuartil es el valor que está en la cuarta posición, es decir, 2. Del mismo modo, el tercer cuartil es el valor que está en la cuarta posición desde el otro extremo, es decir, 9.

Como vemos en el diagrama de caja siguiente, al ubicar la caja entre el primero y el tercer cuartil, se puede tener una idea de la distribución de los datos, es decir, se observa que hay una mayor concentración de datos hacia los valores pequeños puesto que la caja está desplazada a la izquierda.



Solución Deciles: En cuanto al octavo decil, bastaría con ubicar la posición en que se encuentra a través de la fórmula.

$$8\frac{n}{10} = 8\frac{14}{10} = 11.2.$$

Esto quiere decir que entre el dato que se encuentra en la posición 11 y la 12 está el octavo decil, pero más cerca de la 11 que de la 12 puesto que la posición es la 11.2. El resultado sería 9.2 porque entre el 9 y el 19 (que son los datos cuyas posiciones son 11 y 12 respectivamente) hay exactamente 10 unidades.

Solución Percentiles: Con relación a los percentiles pedidos, tendríamos que ubicar las posiciones correspondientes como lo hicimos con los deciles. Para la posición del percentil 42 tendríamos la siguiente fórmula

$$42\frac{n}{100} = 42\frac{14}{100} = 5.88$$

Esto quiere decir que el percentil 42 se encuentra entre los datos que ocupan la posición 5 y la 6. Afortunadamente en este caso ambos datos son 3 por lo que el percentil 42 es 3.

5. Características comparativas de las medidas de tendencia central.

Media Aritmética

1.- Es una medida totalmente numérica o sea sólo puede calcularse en datos de características cuantitativas.

2.- En su cálculo se toman en cuenta todos los valores de la variable.

3.- Es lógica desde el punto de vista algebraico.

4.- La media aritmética es altamente afectada por valores extremos.

5.- No puede ser calculada en distribuciones de frecuencia que tengan clases abiertas.

6.- La media aritmética es única, o sea, un conjunto de datos numéricos tiene una y solo una media aritmética.

Mediana

1.- En su cálculo no se incluyen todos los valores de la variable.

2.- La Mediana no es afectada por valores extremos.

3.- Puede ser calculada en distribuciones de frecuencia con clases abiertas.

4.- No es lógica desde el punto de vista algebraico.

Moda

1.- En su cálculo no se incluyen todos los valores de la variable.

2.- El valor de la moda puede ser afectado grandemente por el método de designación de los intervalos de clases.

3.- No está definida algebraicamente.

4.- Puede ser calculada en distribuciones de frecuencia que tengan clases abiertas.

5.- No es afectada por valores extremos.

Media Geométrica

1.- Se toman en cuenta todos los valores de la variable

2.- Es afectada por valores extremos, aunque en menor medida que la media aritmética.

3.- La media geométrica de un número y su recíproco será siempre igual a uno.

4.- No puede ser calculada en distribuciones con clase abiertas.

5.- Es mayormente usada para promediar tazas de cambio, razones y valores que muestren una progresión geométrica.

6. Medidas de dispersión: absolutas y relativas.

Las medidas de dispersión están encaminadas a cuantificar lo próximos o alejados que están los datos de la muestra de un punto central. Estas medidas indicaran por un lado el grado de variabilidad que hay en la muestra y, por otro, la representatividad de dicho punto central, ya que, si se obtiene un valor pequeño, eso significara que los valores se concentran en torno a ese centro (por lo que habrá poca variabilidad y el centro representará bien a todos). En cambio, si se obtiene un valor grande, significará que los valores no están concentrados, sino dispersos (por lo que habrá mucha variabilidad y el centro no será muy representativo).

Medidas de dispersión absolutas: Se tomará como punto central de referencia la media aritmética, aunque de forma análoga podría utilizarse la mediana.

En primer lugar, se buscará una forma de medir la dispersión en términos absolutos (i.e., dependiendo de las unidades en las que se mida la variable). Matemáticamente la distancia de un valor a otro es la diferencia entre esos valores sin signo, es decir, en valor absoluto. Sin embargo, el valor absoluto tiene muy malas propiedades matemáticas. En cambio, si se elevan las distancias al cuadrado, también "se quita" el signo y eso sí que tiene buenas propiedades matemáticas. Así surge la varianza, que sirve igualmente para medir la dispersión conjunta en torno a la media, pero es más operativa. El resultado se mide en unidades al cuadrado y es difícil de interpretar, por lo que lo habitual es calcular su raíz cuadrada (y surge así lo que se llama desviación típica).

matemáticamente, para calcular la varianza y la desviación típica en general, que se denotan respectivamente S 2 x y Sx, se puede operar como sigue a partir de la tabla de frecuencias de las distancias al cuadrado:

Distancias^2	n_i	Distancias$^2 \times n_i$
$(x_1 - \overline{x})^2$	n_1	$(x_1 - \overline{x})^2 \times n_1$
$\vdots$	$\vdots$	$\vdots$
$(x_k - \overline{x})^2$	n_k	$(x_k - \overline{x})^2 \times n_k$
Total	N	$\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i$

La varianza es la media de las distancias al cuadrado y la desviación es su raíz, es decir:

$$S_x^2 = \frac{\text{suma de todas las distancias}^2}{\text{número de individuos}} = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N}.$$

$$S_x = \sqrt{S_x^2}.$$

Puede comprobarse que la varianza se puede calcular alternativamente como la media de los valores de la variable al cuadrado menos la media al cuadrado de la variable, es decir, S 2 x = x 2 − x 2. Esta segunda fórmula suele resultar más cómoda a la hora de hacer los cálculos.

Medidas de dispersión relativas: A la hora de establecer comparaciones no es aconsejable manejar magnitudes absolutas, ya que las unidades no son siempre comparables. Cuando se pretende comparar la dispersión de variables medidas en distintas unidades o variables con distinto orden de magnitud, es necesario relativizar. Una forma de relativizar es considerar la dispersión en relación al valor absoluto de la media, consiguiendo así el coeficiente de variación, que suele ser interpretado en términos de proporción o porcentaje:

$$CV_x = \frac{S_x}{|\overline{x}|}.$$

El coeficiente de variación se suele utilizar con variables que toman valores positivos (precios, rentas, etc.) y cuya media no esta próxima al 0, ya que en caso contrario al dividir por un valor muy cercano a 0, el resultado puede desvirtuarse. Tampoco se debe utilizar la desviación para comparar la dispersión de dos variables medidas en las mismas unidades (a no ser que el valor de la media sea el mismo), porque, por ejemplo, no es lo mismo 1 cm de dispersión en relación a una media de 170 cm que 1 cm de dispersión en relación a una media de 2 cm. Como la dispersión también se utiliza para medir la representatividad de la media, el CV también se puede utilizar para comparar la representatividad de dos medias. Por otro lado, como el valor de CV es una cantidad relativa, ayuda a valorar la dispersión/representatividad de la media, en el sentido de establecer si es grande o pequeña. Valores bajos del CV indicaran poca dispersión/mucha representatividad y valores altos indicaran mucha dispersión/poca representatividad. No hay criterios universales para decir que un valor del CV es "bajo" o "alto", aunque en la práctica se suelen considerar bajos los valores inferiores al 30 o 40 %, moderados entre esas cantidades y aproximadamente el 80 % y cuando se superan el 120 o 140 % ya se considera que la dispersión es bastante elevada. En resumen, se utiliza el CV para variables positivas cuya media no está próxima a 0 con el objetivo de medir la dispersión entorno a la media y/o su representatividad cuando se deben establecer comparaciones o valoraciones.

7. Propiedades de la varianza. Otras medidas de dispersión: momentos.

En teoría de probabilidad, la **varianza** o **variancia** (que suele representarse como σ^2) de una variable aleatoria es una medida de dispersión definida como la esperanza del cuadrado de la desviación de dicha variable respecto a su media.

Su unidad de medida corresponde al cuadrado de la unidad de medida de la variable: por ejemplo, si la variable mide una distancia en metros, la varianza se expresa en metros al cuadrado. La varianza tiene como valor mínimo 0. La desviación estándar (raíz cuadrada de la varianza) es una medida de dispersión alternativa, expresada en las mismas unidades que los datos de la variable objeto de estudio.

Hay que tener en cuenta que la varianza puede verse muy influida por los valores atípicos y no se aconseja su uso cuando las distribuciones de las variables aleatorias tienen colas pesadas. En tales casos se recomienda el uso de otras medidas de dispersión más robustas.

El término *varianza* fue acuñado por Ronald Fisher en un artículo publicado en enero de 1919 con el título *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*.

Algunas propiedades de la varianza son:

- V(X) mayor o igual a 0
- V(aX+b)=a^2 V(X) siendo *a* y *b*números reales cualesquiera. De esta propiedad se deduce que la

  varianza de una constante es cero, es decir,
- V(X+Y)=V(X)+ V(Y) + 2Cov(X,Y), donde Cov(*X*,*Y*) es la covarianza de *X* e *Y*.
- V(X-Y)= V(X) +V(Y) – 2Cov(X;Y), donde Cov(*X*,*Y*) es la covarianza de *X* e *Y*.

8. Tabla de doble entrada.

Una tabla de doble entrada o cuadro de doble entrada, también denominadas de contingencias, son tablas de datos que hacen referencia dos variables. En la cabecera de las filas establecemos las categorías o valores variables mientras que en la columna principal se añaden las otras variables. En la confluencia entre la primera fila y la primera columna encontramos los datos que corresponden a ambas variables.

¿Para qué sirve una tabla de doble entrada?

Una tabla de doble entrada nos ofrece información estadística de dos eventos relacionados entre sí para contrastar los diferentes valores que obtenemos. Se llaman cuadros o tablas de doble entrada porque organiza los temas en dos direcciones hacia donde debemos llevar la mirada para saber que es, que hacer o qué valor representa. Los cuadros de doble entrada nos permiten organizar la información en columnas horizontales y verticales concentrado en un mismo lugar toda la información obtenida a partir de una lectura.

Lo primero que debemos tener en cuenta son los ejes que se irán cruzando en la tabla y posteriormente añadir los datos que queremos obtener según indique el encabezado de la columna donde se encuentre. Lo primero que debemos hacer **antes de confeccionar un cuadro de doble entrada es**:

- Leer detenidamente el texto que nos ofrecerá los datos para confeccionar el cuadro.
- Determinar las categorías que se cruzarán en cada uno de los ejes.
- Finalmente tenemos que introducir los datos que se cruzan entre los ejes con los datos que extraigamos de la lectura.

9. Distribuciones marginales.

Al analizar una distribución bidimensional, uno puede centrar su estudio en el comportamiento de una de las variables, con independencia de cómo se comporta la otra. Estaríamos así en el análisis de una distribución marginal.
De cada distribución bidimensional se pueden deducir dos distribuciones marginales: una correspondiente a la variable x, y otra correspondiente a la variable y.

Distribución marginal de X

X	ni.
x	x
x1	n1.
x2	n2.
.....	...
xn-1	nn-1.
xn	nn.

Distribución marginal de Y

Y	n.j
x	x
y1	n.1
y2	n.2
.....	...
ym-1	n.m-1
ym	n.m

10. Distribuciones condicionadas.

Partiendo de una distribución bidimensional de frecuencias (xi, yj; nij), se denomina distribución condicionada de la variable X a un valor dado yj de la variable Y a la distribución unidimensional definida por el conjunto de valores tomados por X y de las frecuencias condicionadas de dichos valores de X a qué Y tome el valor yj. Análogamente, se denomina distribución de la variable y condicionada a un valor dado xi de la variable X a la distribución unidimensional definida por el conjunto de valores tomados por Y y de las frecuencias de dichos valores de Y condicionadas a que X tome el valor xi.

11. Independencia estadística.

Cuando no se da ningún tipo de relación entre 2 variables o atributos, diremos que son independientes.

Dos variables X e Y, son independientes entre sí, cuando una de ellas no influye en la distribución de la otra condicionada por el valor que adopte la primera. Por el contrario, existirá dependencia cuando los valores de una distribución condicionan a los de la otra

Dada dos variables estadísticas X e Y, la condición necesaria y suficiente para que sean independientes es:

Propiedades:

1. Si X es independiente de Y, las distribuciones condicionadas de X/Yj son idénticas a la distribución marginal de X
2. Si X es independiente de Y, Y es independiente de X
3. Si X e Y son 2 variables estadísticamente independientes, su covarianza es cero. La recíproca de esta propiedad no es cierta, es decir, la covarianza de 2 variables puede tomar valor cero, y no ser independientes.

12. Idea de ajuste, Ajuste lineal, Modelo de regresión lineal.

En estadística la **regresión lineal** o **ajuste lineal** es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde:

Y_t: variable dependiente, explicada o regresando.

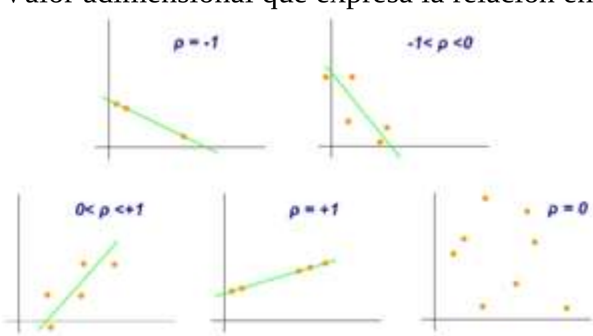
$X_1, X_2, \cdots, X_p$: variables explicativas, independientes o regresores.

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$: parámetros, miden la influencia que las variables explicativas tienen sobre el regrediendo.

donde β_0 es la intersección o término "constante", las β_i $(i > 0)$ son los parámetros respectivos a cada variable independiente, y p es el número de parámetros independientes a tener en cuenta en la regresión.

13. Coeficientes de correlación

Valor adimensional que expresa la relación entre dos variables estadísticas.



Ejemplos de diagramas de dispersión con diferentes valores del coeficiente de correlación (ρ)

En estadística, el coeficiente de correlación de Pearson es una medida de la relación lineal entre dos variables aleatorias-cuantitativas. A diferencia de la covarianza, la correlación de Pearson es independiente de la escala de medida de las variables.

De manera menos formal, podemos definir el coeficiente de correlación de Pearson como un índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas.

Propiedades:

- El **coeficiente de correlación** no varía al hacerlo la escala de medición.
- Es decir, si expresamos la altura en metros o en centímetros el coeficiente de correlación no varía.
- El signo del **coeficiente de correlacion**es el mismo que el de la **covarianza**.
- Si la covarianza es positiva, la correlación es directa.
- Si la covarianza es negativa, la correlación es inversa.
- Si la covarianza es nula, no existe correlación.
- El **coeficiente de correlación lineal** es un número real comprendido entre −1 y 1.
- $\mathbf{-1 \leq r \leq 1}$
- Si el **coeficiente de correlación lineal**toma valores cercanos a −1 la correlación es **fuerte e inversa**, y será tanto más fuerte cuanto más se aproxime r a −1.
- Si el **coeficiente de correlación lineal**toma valores cercanos a 1 la correlación es **fuerte y directa**, y será tanto más fuerte cuanto más se aproxime r a 1.
- Si el **coeficiente de correlación lineal**toma valores cercanos a 0, la correlación es **débil**.
- Si r = 1 ó −1, los puntos de la nube están sobre la recta creciente o decreciente. Entre ambas variables hay **dependencia funcional**.

Ejemplo

Las notas de 12 alumnos de una clase en Matemáticas y Física son las siguientes:

Matemáticas	Física
2	1
3	3
4	2
4	4
5	4
6	4
6	6
7	4
7	6
8	7
10	9

10	**10**

Hallar el **coeficiente de correlación** de la distribución e interpretarlo.

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
2	**1**	**2**	**4**	**1**
3	**3**	**9**	**9**	**9**
4	**2**	**8**	**16**	**4**
4	**4**	**16**	**16**	**16**
5	**4**	**20**	**25**	**16**
6	**4**	**24**	**36**	**16**
6	**6**	**36**	**36**	**36**
7	**4**	**28**	**49**	**16**
7	**6**	**42**	**49**	**36**

8	**7**	**56**	**64**	**49**
10	**9**	**90**	**100**	**81**
10	**10**	**100**	**100**	**100**
72	**60**	**431**	**504**	**380**

1º Hallamos las **medias aritméticas**.

$$\bar{x} = \frac{72}{12} = 6 \qquad \bar{y} = \frac{60}{12} = 5$$

2º Calculamos la **covarianza**.

$$\sigma_{xy} = \frac{431}{12} - 6 \cdot 5 = 5.92$$

3º Calculamos las **desviaciones típicas**.

$$\sigma_x = \sqrt{\frac{504}{12} - 6^2} = 2.45 \qquad \sigma_y = \sqrt{\frac{380}{12} - 25} = 2.58$$

4º Aplicamos la fórmula del **coeficiente de correlación lineal**.

$$r = \frac{5.92}{2.45 \cdot 2.58} = 0.94$$

Al ser el **coeficiente de correlación**positivo, la correlación es directa.

Como **coeficiente de correlación** está muy próximo a 1 la correlación es muy fuerte.

Los valores de dos variables X e Y se distribuyen según la tabla siguiente:

Y/X	0	2	4
1	2	1	3
2	1	4	2
3	2	5	0

Determinar el **coeficiente de correlación**.

Convertimos la tabla de doble entrada en tabla simple.

x_i	y_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i \cdot f_i$	$y_i^2 \cdot f_i$	$x_i \cdot y_i \cdot f_i$
0	1	2	0	0	2	2	0
0	2	1	0	0	2	4	0
0	3	2	0	0	6	18	0
2	1	1	2	4	1	1	2
2	2	4	8	16	8	16	16
2	3	5	10	20	15	45	30
4	1	3	12	48	3	3	12

4	**2**	**2**	**8**	**32**	**4**	**8**	**16**
		20	**40**	**120**	**41**	**97**	**76**

$$\bar{x} = \frac{40}{20} = 2 \qquad \bar{y} = \frac{41}{20} = 2.05$$

$$\sigma_x^2 = \frac{120}{20} - 2^2 = 2 \qquad \sigma_y^2 = \frac{97}{20} - 2.05^2 = 0.65$$

$$\sigma_x = \sqrt{2} = 1.41 \qquad \sigma_y = \sqrt{0.65} = 0.81$$

$$\sigma_{xy} = \frac{76}{20} - 2 \cdot 2.05 = -0.3$$

$$r = \frac{-0.3}{1.41 \cdot 0.81} = -0.26$$

Al ser el **coeficiente de correlación**negativo, la correlación es inversa.

Como **coeficiente de correlación** está muy próximo a 0 la correlación es muy débil.

14. Distribución de frecuencias con intervalos cerrados y semi abiertos.

Las distribuciones de frecuencias son tablas en que se dispone las modalidades de la variable por filas. En las columnas se dispone el número de ocurrencias por cada valor, porcentajes, etc. La finalidad de las agrupaciones en frecuencias es facilitar la obtención de la información que contienen los datos.

Ejemplo: Quieren conocer si un grupo de individuos está a favor o en contra de la exhibición de imágenes violentas por televisión, para lo cual han recogido los siguientes datos:

X: 2,1,5,3,3,2,3,1,4,2,4,2,3,2,3,4,3,3,1,2

(Regla de codificación:
1= En contra.
2= Bastante en contra.
3= Indiferente.
4= Bastante a favor.
5= A favor.)

X: Símbolo genérico de la variable.

f: Frecuencia (también se simboliza como ni).

La distribución de frecuencias de los datos del ejemplo muestra que la actitud mayoritaria de los individuos del grupo estudiado es indiferente.

La interpretación de los datos ha sido facilitada porque se ha reducido el número de números a examinar (en vez de los 20 datos originales, la tabla contiene 5 valores de la variable y 5 frecuencias).

Generalmente las tablas incluyen varías columnas con las frecuencias relativas (son el número de ocurrencias dividido por el total de datos, y se simbolizan "fr" o "pi"), frecuencias acumuladas (la frecuencia acumulada es el total de frecuencias de los valores iguales o inferiores al de referencia, y se simbolizan "fa" o "na". No obstante la frecuencia acumulada también es definida incluyendo al valor de referencia), frecuencias acumuladas relativas (la frecuencia acumulada relativa es el total de frecuencias relativas de los valores iguales o inferiores al de referencia, y se simbolizan "fr" o "pa")

Elementos básicos de las tablas de intervalos:

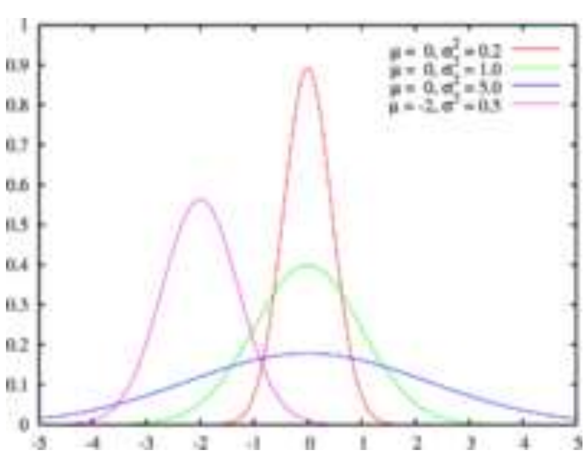
- Intervalo: Cada uno de los grupos de valores de la variable que ocupan una fila en una distribución de frecuencias
- Límites aparentes: Valores mayor y menor del intervalo que son observados en la tabla. Dependen de la precisión del instrumento de medida. En el ejemplo, los límites aparentes del intervalo con mayor número de frecuencias son 34 y 39.
- Límites exactos: Valores máximo y mínimo del intervalo que podrían medirse si se contara con un instrumento de precisión perfecta. En el intervalo 34-39, estos límites son 33.5 y 39.5

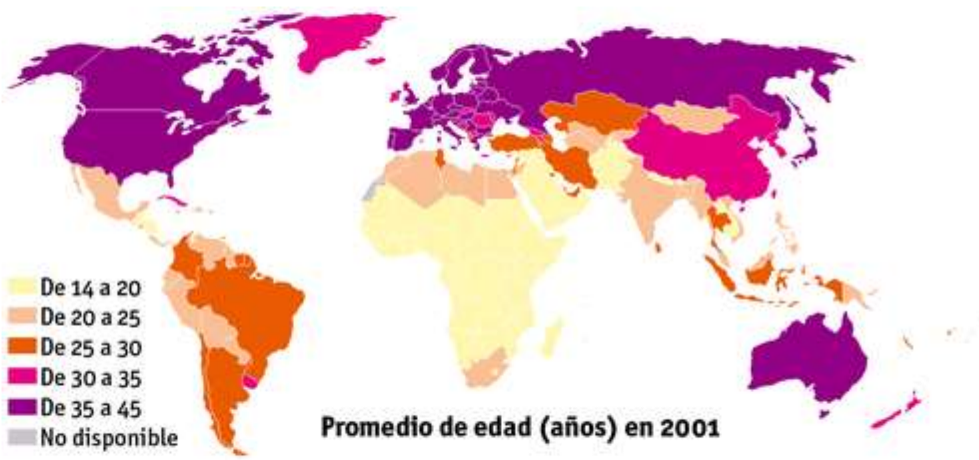
- Punto medio del intervalo (Mco Marca de clase): Suma de los límites dividido por dos. Mc del intervalo del ejemplo= 36.5
- Amplitud del intervalo: Diferencia entre el límite exacto superior y el límite exacto inferior. En el ejemplo es igual a 6.

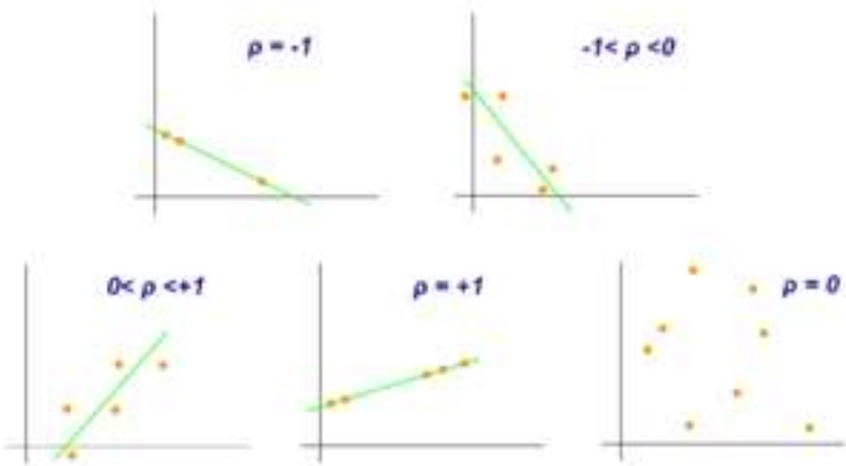
B. Base de Consulta

TÍTULO	AUTOR	EDICIÓN	AÑO	IDIOMA	EDITORIAL
"Elementos Básicos De Estadística Económica y Empresarial"	MONTIEL, A. M. ; RIUS, F. Y BARON, F. J.		1997	Español	Pretince Hall
"Estadística para los Negocios y la Economía”	NEWBOLD, PAUL		1997	Español	Pretince Hall
Estadística para las ciencias administrativas.	CHAO, Lincoln L.	3	2003	Español	México: Mc-Graw Hill
Estadística elemental.	FREUND, Jhon E.	8	2002	Español	México: P.H.H.

C. Base práctica con ilustraciones




De 14 a 20
De 20 a 25
De 25 a 30
De 30 a 35
De 35 a 45
No disponible
Promedio de edad (años) en 2001


ρ = -1
-1< ρ <0
0< ρ <+1
ρ = +1
ρ = 0

X: 2,1,5,3,3,2,3,1,4,2,4,2,3,2,3,4,3,3,1,2

(Regla de codificación:
- 1= En contra.
- 2= Bastante en contra.
- 3= Indiferente.
- 4= Bastante a favor.
- 5= A favor.)

Distribució de freqüencies agrupades en intervals

X_i	f_i	fr_i	f_a	fr_a
64-69	2	0.02	100	1.00
58-63	8	0.08	98	0.98
52-57	7	0.07	90	0.90
46-51	11	0.11	83	0.83
40-45	16	0.16	72	0.72
34-39	22	0.22	56	0.56
28-33	21	0.21	34	0.34
22-27	9	0.09	13	0.13
16-21	4	0.04	4	0.04
	100	1.00		

4. ESTRATEGIAS DE APRENDIZAJE

ESTRATEGIA DE APRENDIZAJE 1: Análisis y Planeación
Descripción: • Lectura comprensiva del material proporcionado. • Trabajos grupales. • Equipos de investigación. • Desarrollo de ejercicios prácticos. • Participación de los estudiantes en el aula de clases.
Ambiente(s) requerido: Aula amplia con buena iluminación.
Material (es) requerido: Pizarrón, marcadores, materiales de apoyo para los estudiantes.
Docente: Con conocimiento de la materia.

5. ACTIVIDADES

- Controles de lectura
- Exposiciones
- Desarrollo de talleres y actividades grupales en el aula
- Tareas en plataforma
- Presentación del Trabajo final

Se presenta evidencia física y digital con el fin de evidenciar en el portafolio de cada aprendiz su resultado de aprendizaje. Este será evaluable y socializable

6. EVIDENCIAS Y EVALUACIÓN

Tipo de Evidencia	**Descripción (de la evidencia)**
De conocimiento:	Ensayo expositivo grupal de lecturas Definición del tema de investigación Lecturas que permitan el resumen y aplicación de Estadística.
Desempeño:	Trabajo grupal y colaborativo para presentación sobre estadística socializadas en clase.
De Producto:	Trabajo final donde los estudiantes aplican lo conocido sobre los temas de estadística trabajadas en el módulo.
Criterios de Evaluación (Mínimo 5 Actividades por asignatura)	Actividad N.- 1 AULA Exposición de la clase Elaboración de trabajo académico desarrollado por los estudiantes Taller de ejercicios para refuerzo y entendimiento dela clase proporcionada. PLATAFORMA Evaluación sobre los ejercicios tratados en la clase
	Actividad n.- 2 AULA Exposición de la clase Taller sobre medidas de tendencia central: media, mediana, moda. PLATAFORMA Observación de videos relacionados con la clase para refuerzo de los estudiantes.
	Actividad N.-3 AULA Exposición del tema

	Trabajo cooperativo para aplicación de ajuste lineal, modelo de regresión lineal, otros. PLATAFORMA Investigación sobre temas de modelos coeficientes de correlación
	Actividad N.- 4 AULA Exposición de temas. Desarrollo de trabajos grupales sobre temas distribuvion de frecuencias con intervalos semi-abiertos y cerrados. PLATAFORMA Desarrollo de ejercicios asociados con los temas tratados en clase.
	Actividad N.- 5 AULA Presentación del trabajo final. PLATAFORMA Evaluación de refuerzo de ejercicios tratados en el modulo.

Elaborado por: **Ing.: José Luis Herrera**	**Revisado Por:** **Msc.: Daniel Shauri**	**Reportado Por:** **Dr.: Milton Altamirano**

INSTITUTO TECNOLÓGICO SUPERIOR "JAPÓN"

www.itsjapon.edu.ec

Calle Mariete de Veintimilla y
Cuarta Transversal
2 356 368