# ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

WAZHINGTON BLADIMIR PROAÑO RIVERA

UNIVERSIDAD DEL AZUAY
Casa Editora

# ESTADÍSTICA DESCRIPTIVA E INFERENCIAL

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod

UNIVERSIDAD DEL AZUAY

Casa Editora

**UNIVERSIDAD DEL AZUAY**

ESTADÍSTICA DESCRIPTIVA

Francisco Salgado Arteaga
**RECTOR**

Martha Cobos Cali
**VICERRECTORA ACADÉMICA**

Jacinto Guillén García
**VICERRECTOR DE INVESTIGACIONES**

Toa Tripaldi Proaño
**DIRECTORA DE COMUNICACIÓN Y PUBLICACIONES**

Oswaldo Encalda
**CORRECCIÓN DE ESTILO**

Verónica Neira Ruiz y Sebastián Carrasco Hermida
**REVISORES DE APA**

Daniela Durán Pozo
**DISEÑO Y DIAGRAMACIÓN**

Román Idrovo Daza
Jorge Antonio Pérez Torres
**PARES ACADÉMICOS**

**ISBN:** 978-9942-618-26-9
**e- ISBN:** 978-9942-822-69-7

Cuenca-Ecuador
2020

# CONTENIDO

## A

# B

# Índice de Tablas

# Índice de gráficos

# ESTADÍSTICA DESCRIPTIVA

## Introducción

La estadística está alrededor de nosotros todo el tiempo, como sustento de un estudio de factibilidad, como herramienta para tabular datos de una encuesta, como insumo imprescindible para interpretar información que nos facilite tomar una decisión y como una gran instrumento para la toma de decisiones, por ejemplo: el INEC informa mensualmente que los precios han subido 0.15%; el BCE nos dice que en el 2006 el ecuatoriano en promedio ganó $300 mensuales; etc. Estos números (estadísticas) toman protagonismo tanto para un estudiante de economía, como para un sociólogo, un investigador, o cualquier otro profesional que pretenda interpretar los datos que se le presentan y a partir de ellos concluir un fenómeno para una población mayor, con lo cuales ya no hace falta conocer toda la información para entender qué es lo que está pasando.

La estadística ayuda a la toma de decisiones que afectan la vida misma o el bienestar colectivo, a través de su gran aplicación en ciencias como la economía, la mercadotecnia, o en general en las ciencias sociales.

Por la importancia que reviste esta ciencia dedicaremos en la primera sección de esta nota a estudiar los conceptos básicos que rodean a la estadística descriptiva, para luego entender la inferencia estadística.

# 1. Definición de estadística

**La estadística es la ciencia que se ocupa de recolectar, organizar, analizar e interpretar datos para realizar una toma de decisiones más efectiva.**

La importancia de estudiar esta ciencia radica en que generalmente existen datos que suelen estar "sueltos" en el mundo real, la estadística provee de herramientas para su recolección y análisis, en la actualidad esta tarea puede optimizarse gracias a los diversos softwares estadísticos, los cuales ofrecen una mejora en el procesamiento y generación de los informes estadísticos permitiendo una mayor claridad en la toma de decisiones.

Un gobierno debe manejar estadísticas cuando decide sobre tal o cual política, por ejemplo, si considera –en base a la información recolectada- que el salario promedio mensual de un ecuatoriano no permite satisfacer sus necesidades, podría decidir incrementar los sueldos, y gracias al análisis estadístico, conocer si eso afectará o tendrá incidencia tanto en el trabajador como en la marcha de la economía en su conjunto. Con esto queremos indicar que la estadística, a la vez que ayuda a describir datos, también puede hacer inferencias o estimaciones de lo que podrá ser el comportamiento de ese conjunto de datos.



Empezaremos con el estudio de la estadística descriptiva, siguiendo un ordenamiento de los temas. Así en la primera sección abordaremos elementos introductorios básicos. En la segunda estudiaremos la descripción de los datos; para ello explicaremos las medidas de tendencia central, de dispersión y otras medidas importantes. Y en las dos últimas secciones se estudiarán la regresión simple y los conceptos básicos de probabilidades y distribuciones de probabilidad.

## Iniciamos entonces clasificando la estadística en dos grandes áreas de estudio:

1. La estadística descriptiva que comprende un conjunto de métodos para organizar, resumir y presentar los datos de manera informativa.

2. La estadística inferencial comprende un conjunto de métodos para SABER ALGO acerca de una población, basándose en una muestra. De aquí se desprenden dos conceptos muy importantes dentro del mundo de la estadística como son: *la población y la muestra:*

- **Población** es el conjunto o recopilación de datos, objetos o medidas de interés de todos los individuos de una población. Será representada por la letra N.

- **Muestra** es una parte o subconjunto de la población, que para ser considerada como tal, debe ser significativa, es decir, que permita inferir o estudiar el total de los datos (la población). Será representada por la letra n.

Ilustración 1. Muestras y población


Elaboración propia.

Después de definir que N es la muestra es importante recalcar que, generalmente no es posible acceder a la información de toda la población principalmente por factores como: el costo, tiempo, naturaleza destructiva y la imposibilidad. Ejemplo: acceder al ingreso percibido por cada uno de los 12 millones de trabajadores resultaría complicado, se justifica entonces trabajar con datos más reducidos, es decir, con muestras.

Existen diferentes tipos de muestreo; pero se los puede clasificar en dos grandes grupos: el muestreo probabilístico y el muestreo no probabilístico.

- **El muestreo probabilístico:** es aquel en el cual la muestra que se selecciona se conforma por un conjunto de datos o individuos de la población en estudio, que tienen la misma probabilidad de ser escogidos. Es decir, que tenga una probabilidad conocida ($pi \neq 0$).
- **El muestreo no probabilístico:** es aquel muestreo en el cual no todos los integrantes tienen la probabilidad de ser incluidos en la muestra. Los resultados pueden estar sesgados, es decir, no ser representativos de la población.



Una empresa que desee realizar un estudio de mercados, deberá tener en consideración que para que el estudio de mercados tenga validez, la muestra obtenida en dicho estudio, será obtenida mediante un muestreo probabilístico.

## Dentro del muestreo probabilístico es posible identificar los siguientes tipos:

En el muestreo aleatorio simple (MAS) la muestra se selecciona de tal manera que cada integrante de la población tenga la misma probabilidad de ser incluido. El procedimiento que tenemos para ello es el de seguir la tabla de números aleatorios que constituye un medio eficiente para seleccionar elementos de una muestra. Por ejemplo, supongamos que queremos estudiar los hábitos de consumo de los ecuatorianos, entonces se tratará de generar en una computadora 10.000 números aleatorios (entre 1 y 17 millones) que seleccionen a los individuos que serán entrevistados en este estudio, a partir del censo de los aproximadamente 17 millones de ecuatorianos.

El muestreo aleatorio sistemático (MASIS), en este caso, los elementos de la población se ordenan en alguna forma -por ejemplo, alfabéticamente- en un archivo según la fecha en que se reciben o por algún otro método. Se selecciona al azar el punto de partida y después se elige para la muestra[1] cada k-ésimo elemento de la población. Este ordenamiento sigue un sistema en función del ordenamiento. Para este sistema calculamos una razón entre el número de los elementos de la muestra y el total de elementos de la población; luego vamos sumando al elemento anterior.

Siguiendo el ejemplo anterior, si tenemos la población de 17 millones ordenados de acuerdo al apellido, y si la muestra es de 10.000 elementos, la razón será de 1700 (17.000.000/10.000). Tomamos un número de partida, por ejemplo, el individuo 1.250 entonces los demás elementos de la muestra serán el elemento 1.250+1.700=2.950, el siguiente será el 4.650 (2.950+1.270), el siguiente será el elemento 6.350 (4.650+1.700) y así sucesivamente. Se advierte que este tipo de muestreo no debería usarse con datos temporales de periodicidad; por ejemplo, datos del consumo, de viajes, etc.



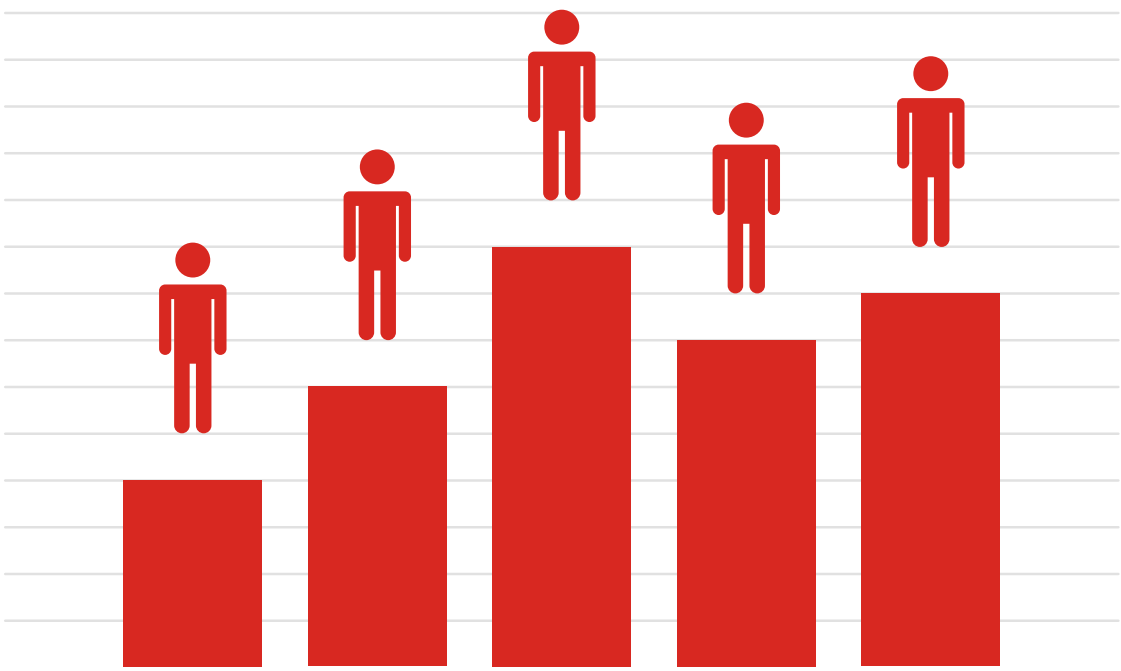
[1] Más adelante explicaremos cómo calcular el tamaño de la muestra. Pero podríamos anticiparnos en decir que para calcular el tamaño de la muestra se deben tomar en cuenta algunas de sus propiedades y el error máximo que se permitirá en los resultados. Para el cálculo de n (tamaño de la muestra) se puede emplear la siguiente fórmula:

$n = \delta ZE$

donde $\delta$ es la desviación estándar, Z es el nivel de confianza y E es el máximo error permitido.

El muestreo aleatorio estratificado (MAE) es aquel en el que la población se divide en subgrupos denominados *estratos* y se selecciona una muestra de cada uno de ellos siguiendo el MAS. En nuestro ejemplo dividimos la población en hombres y mujeres; por lo tanto, usando el muestro aleatorio simple se conforman subgrupos de 5.000 hombres y 5.000 mujeres.

El muestreo por conglomerados es aquel en el que la población se divide en grupos o conglomerados que se vuelven representativos de la población, se define como un muestreo combinado del aleatorio simple y el estratificado. En el ejemplo podríamos tomar una ciudad que pueda ser representativa y de allí trabajaríamos con una muestra.



# 2. Variables y datos estadísticos

**Cuando definíamos la población o la muestra, a través de los individuos, objetos o datos, estábamos en verdad intentando definir lo que es un dato estadístico, es hora entonces de precisar que:**

Datos son cada uno de los individuos, cosas, entes abstractos, que integran una población o universo determinado. Es cada valor de la variable observada.

Variable de una población que estamos conociendo es alguna *característica* que sea de nuestro interés. En otras palabras, entendemos por variable esta característica que estamos midiendo.

**Existen dos tipos de variables estadísticas:**

1. **Variables cualitativas:** aquellas que expresan un atributo o característica; por ejemplo, el color del cabello, la profesión de un individuo, los servicios de un hotel, etc.
2. **Variables cuantitativas:** aquellas características que podemos expresar numéricamente; ejemplo, edad, peso, tiempo de duración de un viaje, permanencia de un turista, etc.

**A su vez las variables cuantitativas pueden ser de dos tipos:**

1. **Variable discreta:** aquella que entre dos valores próximos puede tomar a lo sumo un número finito de valores; por ejemplo, el número de estrellas de un hotel, los trabajadores en una agencia de viajes, etc.
2. **Variable continua:** aquella que entre dos valores próximos puede tomar un número infinito de valores; por ejemplo, el tiempo que tarda un vuelo en avión entre Guayaquil y EE.UU., etc.

**De igual forma, los datos estadísticos también presentan algunas características especiales, por ejemplo, en los:**

- **Datos de corte transversal:** la observación de una o más variables para distintos individuos en un mismo momento del tiempo, suelen ser datos referentes a familias, empresas, sectores, etc. En este tipo de datos no hay problemas de interrelación. Por ejemplo, el consumo de una familia i, es indistinto del consumo de la familia j.

- **Datos temporales:** la observación de una o más variables a intervalos regulares de tiempo para un solo individuo. Por ejemplo, el consumo de la familia x en la temporada de navidad. Se denominan series temporales.

- **Datos de panel:** Se trata de una combinación de los dos anteriores, son diferentes variables para distintos individuos durante un cierto número de intervalos regulares de tiempo. Normalmente se usan en estudios de mercadotecnia. Por ejemplo, el consumo de bebidas gaseosas entre los jóvenes y adultos en épocas de verano.

# 3. Niveles de medición de los datos

El nivel de medición de los datos marca los cálculos que pueden realizarse para resumir y presentar la información y las pruebas estadísticas que pueden desarrollarse.

## Existen cuatro niveles de medición de los datos:

- **Nivel nominal**, en el cual los datos son clasificados destacando una cualidad que los identifica, en categorías sin un orden específico.

En la tabla 3.1 se evidencia cómo se agruparon los datos obtenidos entre los asistentes a un museo de la ciudad:

Tabla 3-1

Clasificación de datos en escala nominal

| Categoría | Total |
|---|---|
| Hombres | 30 |
| Mujeres | 33 |

Se clasifican los datos y realiza un conteo únicamente, los datos son mutuamente excluyentes (pertenecen a una sola categoría a la vez).

- **Nivel ordinal**, esta medición supone que los datos están clasificados en un orden lógico, es decir, una categoría definida como más alta o de mayor importancia que otra, por lo tanto, a más de ser las categorías de datos mutuamente excluyentes y siendo estas de naturaleza exhaustiva (analizada a detalle), se clasifican u ordenan de acuerdo con la característica particular que posean.

La tabla 3-2 ilustra la clasificación de información sobre el tipo de habitación de un hotel:

Tabla 3-2

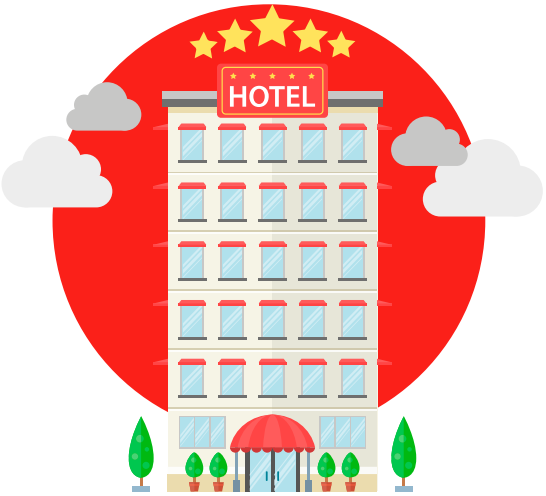
Clasificación de datos en escala ordinal

| Orden | Tipo de habitación (nominal) | Número de habitaciones |
|---|---|---|
| 1 | Suite | 3 |
| 2 | Triple | 5 |
| 3 | Doble | 12 |
| 4 | Simple | 13 |

- **Nivel de intervalo**, en este nivel la distancia entre los valores es importante. Las categorías de datos a más de ser mutuamente excluyentes, exhaustivas y tener un orden lógico, tienen las distancias iguales en las características, y son representadas por iguales diferencias en los números asignados a las categorías.

**Ejemplo:** el número de estrellas de un hotel.

La diferencia entre un hotel tres estrellas y otro de cinco estrellas en Cuenca es la misma que en Nueva York.




- **Nivel de razón**, este nivel incluye las características de los tres niveles anteriores, aquí el punto cero es reconocido como la falta de todas las cualidades que son medidas.

**Ejemplo:** los salarios de un vendedor o un ejecutivo de ventas. Puede ser que una persona no tenga sueldo por lo tanto el "Cero" será una unidad de medida.

# 4. Organización de los datos

Una vez que se dispone de los datos de la muestra seleccionada, el siguiente paso es organizarlos y tabularlos, se utilizará un tipo de organización por frecuencias con el objetivo de facilitar su manejo y su aplicabilidad; para ello es necesario conocer algunos conceptos básicos.

- **Frecuencia absoluta:** llamaremos así al número de repeticiones que presenta una observación dentro de un conjunto de datos; se representa por *fi.*
- **Frecuencia relativa:** es la frecuencia absoluta dividida entre el número total de datos, y se suele expresar en tanto por ciento. La remplazaremos como *fi%.*
- **Frecuencia absoluta acumulada:** es la suma de las frecuencias absolutas; la representaremos como *FAi.*
- **Frecuencia relativa acumulada:** es la frecuencia acumulada dividida entre el total de las observaciones o lo que es lo mismo, es la suma de las frecuencias relativas. La representaremos como *FAi%.*

Cuando los datos no son numerosos (datos sueltos) podremos organizarlos en una distribución denominada distribución unidimensional de frecuencias, tabla 4-1.



Tabla 4-1
Distribución unidimensional de frecuencias

| $x_i$ | *fi* | *fi%* | *FAi* | *FAi%* |
|---|---|---|---|---|
| $x_1$ | | | | |
| $x_2$ | | | | |
| : | | | | |
| : | | | | |
| $x_3$ | | | | |

**Ejemplo 4.1:**

Las horas trabajadas por usted, cada semana, durante los últimos dos meses son: 52, 48, 37, 54, 48, 15, 42,12. Organice los datos en una distribución de frecuencias.

| $x_i$ | $fi$ | $fi\%$ | $FAi$ | $FAi\%$ |
|---|---|---|---|---|
| 52 | 2 | 5,56% | 2 | 5,56% |
| 48 | 7 | 19,44% | 9 | 25,00% |
| 37 | 6 | 16,67% | 15 | 41,67% |
| 54 | 5 | 13,89% | 20 | 55,56% |
| 48 | 3 | 8,33% | 23 | 63,89% |
| 15 | 1 | 2,78% | 24 | 66,67% |
| 42 | 8 | 22,22% | 32 | 88,89% |
| 12 | 4 | 11,11% | 36 | 100,00% |
| | Σ=36 | Σ=100% | | |

$$n = 36 = \sum_{\square}^{\square} f_i$$

Cuando nos encontramos con un conjunto de datos numerosos se utiliza la agrupación de datos en las denominadas *clases* o intervalos, para facilitar la comprensión de los mismos; no obstante se puede correr el riesgo de perder alguna información valiosa.



Las clases comprenden un intervalo de datos que van desde un límite inferior ($L_{i-1}$) a un límite superior ($Li$). Por facilidad establecemos el siguiente procedimiento para agrupar datos:

- Determinaremos en primer lugar, el número de clases en las que se agruparán los datos. Para ello usamos cuatro posibles herramientas:

***1. Regla de Sturgess*** $k$=1+3,22 $log$ ($n$)

Cierta distribución de datos sobre los índices de alfabetización, fueron proporcionados por 57 países. ¿Cuántas clases se sugieren formar con estos datos?

**Solución:**

$k$=1+3,22$log$ ($n$)
$k$=1+3,22$log$ (57)
$k$=1+3,22 (1.7558)
$k$=6.83=7 clases

**2. Regla empírica** $k=\sqrt{n}$ esta regla se aplica cuando el número de observación es menor que 100.

La raíz del número de datos es una manera sencilla para determinar el número de clases.

**3. Regla empírica** $2^k \geq n$

El número de clases es la menor potencia a la a que se eleva 2 de tal manera que el resultado sea igual o se aproxime a $n$. Entonces, si existen 65 datos tenemos:

$2^5$=32
$2^6$=64
$2^7$=128

**4. Método Subjetivo** $5 \leq n \leq 20$ la determinación del número de clases depende del criterio del evaluador.

El valor resultante de $k$ siempre debe ser un número entero.



- En segundo lugar, establecemos el ancho de clase, el cual definimos como: la diferencia entre el límite inferior de la clase siguiente y el límite inferior de la clase anterior. Este valor del ancho de clase o amplitud del intervalo puede calcularse como:

$$\text{Ancho de clase} = C_i \text{ o } a_i = \frac{Re}{k}$$

- Re, es el recorrido o rango de datos o amplitud de variación de los datos, es igual al valor más grande o alto, menos el valor más bajo o más pequeño $Re = v_{max} - v_{min}$

Para construir la distribución de frecuencias hacemos uso de las marcas de clase o puntos medios, es decir, el límite superior menos el límite inferior dividido entre 2. Luego la elaboración de la tabla de frecuencias sigue como si se tratara de datos no agrupados.

**Ejemplo 4.2**

La agencia de viajes nacional Moore ofrece tarifas especiales en ciertas travesías por el Caribe a ciudadanos de la tercera edad. El presidente de la agencia quiere información adicional sobre las edades de las personas que viajan. Una muestra aleatoria de 40 clientes que hicieron un crucero el año pasado dio a conocer las siguientes edades:

Tabla 4-2
Ejemplo distribución de frecuencias

| Número de observaciones | Edad (según la muestra original) | Edad ordenada |
|---|---|---|
| 1 | 77 | 18 |
| 2 | 18 | 26 |
| 3 | 63 | 34 |
| 4 | 84 | 36 |
| 5 | 38 | 38 |
| 6 | 54 | 41 |
| 7 | 50 | 43 |
| 8 | 59 | 44 |
| 9 | 54 | 45 |
| 10 | 56 | 50 |
| 11 | 36 | 50 |
| 12 | 26 | 51 |
| 13 | 50 | 52 |
| 14 | 34 | 52 |

| | | |
|---|---|---|
| 15 | 44 | 53 |
| 16 | 41 | 53 |
| 17 | 58 | 54 |
| 18 | 58 | 54 |
| 19 | 53 | 56 |
| 20 | 51 | 58 |
| 21 | 62 | 58 |
| 22 | 43 | 58 |
| 23 | 52 | 59 |
| 24 | 53 | 60 |
| 25 | 63 | 60 |
| 26 | 62 | 61 |
| 27 | 62 | 61 |
| 28 | 65 | 62 |
| 29 | 61 | 62 |
| 30 | 52 | 62 |
| 31 | 60 | 63 |
| 32 | 60 | 63 |
| 33 | 45 | 63 |
| 34 | 66 | 65 |
| 35 | 83 | 66 |
| 36 | 71 | 71 |
| 37 | 63 | 71 |
| 38 | 58 | 77 |
| 39 | 61 | 83 |
| 40 | 71 | 84 |



Establecemos el número de clases:

**√n = √40 = 6,32** podríamos redondear en 6 clases.

**Recorrido (Re) =** Valor más alto menos valor más bajo = 84 - 18 = 66

**Ancho de clase** $C_i$ *o* $a_i = Re/k = 66/6 = 11$. El ancho de clase será 11, en caso de que el valor obtenido sea decimal deberá revisarse el ancho de clase obtenido para que facilite la presentación y análisis de los datos, considerando que la frontera no coincida con las observaciones.

Se recomienda que el límite inferior de la primera clase sea un poco más bajo que el valor más pequeño del conjunto de datos; en este caso puede ser 15, igualmente el límite superior de la última clase debe ser un poco mayor al valor más alto del conjunto de datos.

Entonces las clases quedan definidas así:

Tabla 4-3
Clasificación de clases

| Clases | f | f% | Marca de clase | FA | FA% |
|---|---|---|---|---|---|
| 15-26 | 2 | 5.0% | 20.5 | 2 | 5.0% |
| 26-37 | 2 | 5.0% | 31.5 | 4 | 10.0% |
| 37-48 | 5 | 12.5% | 42.5 | 9 | 22.5% |
| 48-59 | 14 | 35.0% | 53.5 | 23 | 57.5% |
| 59-70 | 12 | 30.0% | 64.5 | 35 | 87.5% |
| 70-81 | 3 | 7.5% | 75.5 | 38 | 95.0% |
| 81-92 | 2 | 5.0% | 86.5 | 40 | 100.0% |
| | 40 | 100.0% | | | |

## Diagrama de hoja y tallo

Esta es otra forma de organizar los datos o variables continúas creada por el estadístico John Tukey. Se refiere a una estructura en la que se trata de dividir los datos en el tallo y las hojas. El tallo representa el valor entero o múltiplo y la hoja los números adjuntos, unidades o decimales. El tallo y la hoja están colocados en series ordenadas. Por ejemplo, supongamos que tenemos los tiempos de vuelo medidos en horas de una aerolínea. Los tiempos son: 34.5; 34.6; 45.7; 45.8 y 56.2



Entonces organizamos en un diagrama de tallo y hoja:

| Tallo | Hoja |
|---|---|
| 34 | 5, 6 |
| 45 | 7, 8 |
| 56 | 2 |

# 5. Representación gráfica de los datos

Una vez que hemos organizado los datos en una tabla de frecuencias- distribución de frecuencias- podemos representarlos gráficamente. En estadística se usan generalmente las siguientes representaciones gráficas:

- **Diagrama de barras:** Comprende un gráfico en un eje cartesiano en donde las abscisas representan la variable de estudio y en las ordenadas la frecuencia o las observaciones de la variable. Se usa frecuentemente con variables discretas.

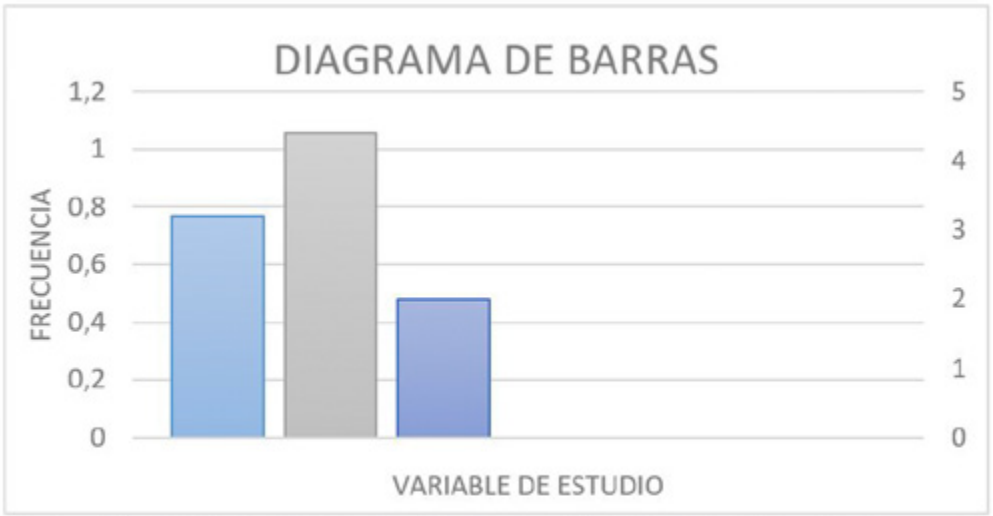



*Ilustración STYLEREF 1 \s 5 SEQ Ilustración * ARABIC \s 1 1 Diagrama de barras*

- **Histograma:** Es el gráfico más importante de la estadística, con grandes aplicaciones. Muestra, de manera similar al diagrama de barras, las categorías de la variable (las clases) en el eje de las abscisas y en el eje de las ordenadas la frecuencia absoluta. Puede también graficarse considerando la frecuencia relativa. Se usa generalmente con variables continuas.


*Ilustración STYLEREF 1 \s 5 SEQ Ilustración * ARABIC \s 1 2 Histograma de frecuencias*

- **Diagrama por sectores:** En ocasiones los datos (sobre todo en variables discretas) pueden ser representados mediante una estructura en forma de *pastel* en la que cada *sector* o pedazo constituye el porcentaje de representación dentro de la estructura. Cada sector será igual a 360 dividido para la frecuencia.


*Ilustración STYLEREF 1 \s 5 SEQ Ilustración * ARABIC \s 1 3 Gráfico de diagrama por sectores*

- **Pictograma para expresar un atributo de la variable:** Suelen utilizarse iconos que se identifican con la variable, existen dos formas de representar un pictograma:

a. Con un dibujo que representa a la variable y se representan tantas veces como aparezca la variable,
b. La representación gráfica y el tamaño guardan relación con la frecuencia.

- **Polígono de frecuencias:** Es un gráfico lineal que se construye al unir los puntos medios de cada clase –cuando los datos son agrupados- o los pares ordenados conformados por el valor de la variable y la frecuencia absoluta.

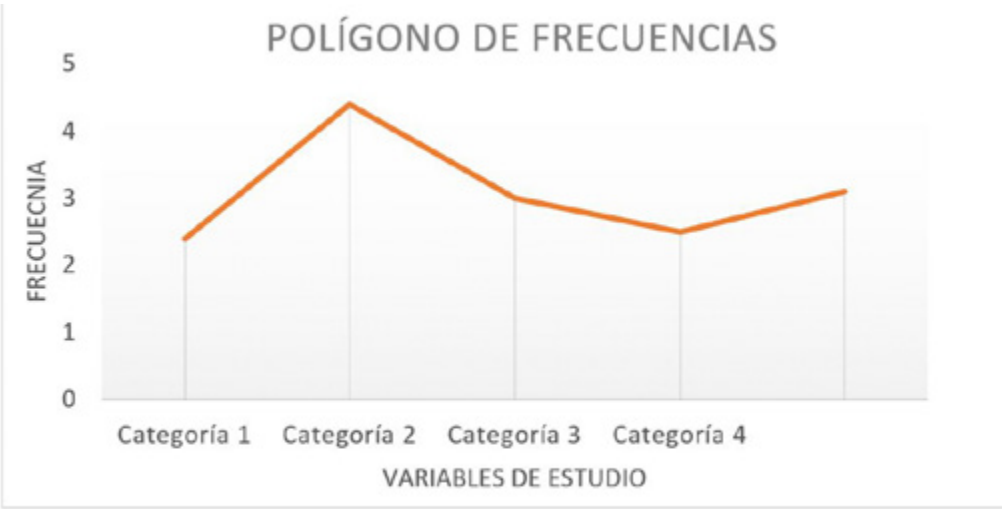





*Ilustración STYLEREF 1 \s 5 SEQ Ilustración * ARABIC \s 1 4 gráfico de un polígono de frecuencias*

- **Ojivas:** Una ojiva es un polígono de frecuencias que se construye con la distribución de frecuencias acumuladas-absolutas o relativas- estableciendo una relación en la variable “mayor que” o “menor que”.

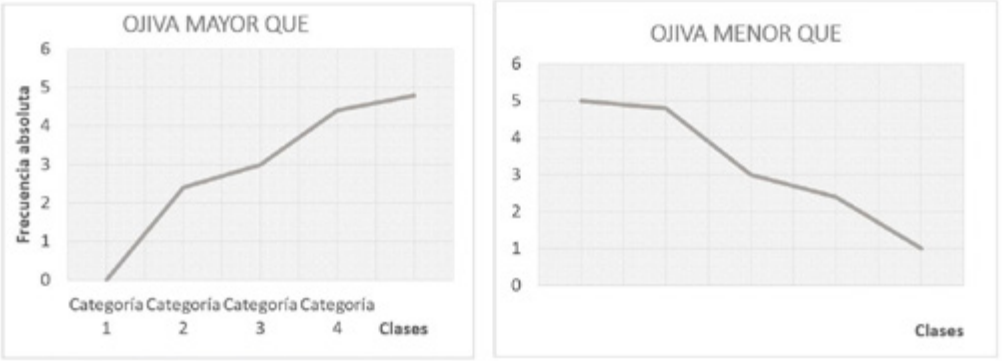



*Ilustración STYLEREF 1 \s 5 SEQ Ilustración **

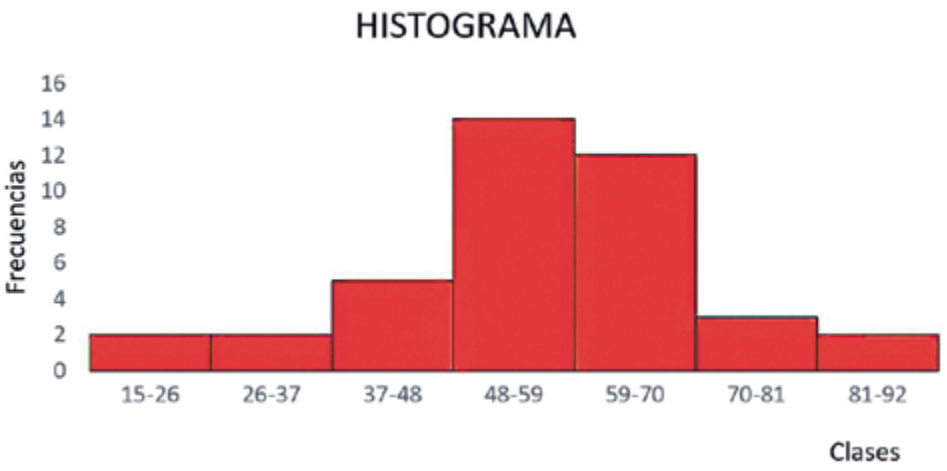
**Ejemplo 5.1**

Como una aplicación práctica a esta representación gráfica vamos a referirnos al conjunto de datos de la agencia de viajes Moore, que vimos antes; para ello tomamos la distribución de frecuencias calculada.

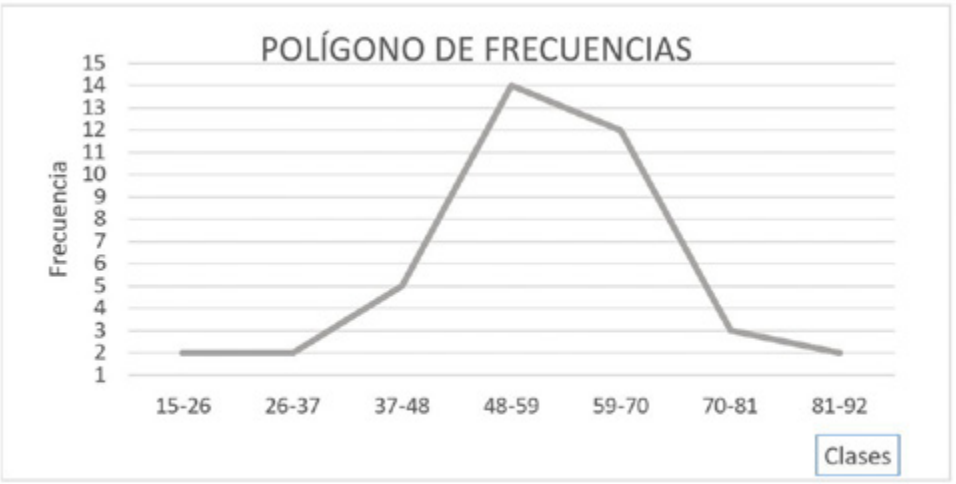
Tabla 5-1
Cálculo de distribución de frecuencias

| Clases | f | f% | Marca de clase | FA | FA% |
|---|---|---|---|---|---|
| 15-26 | 2 | 5.0% | 20.5 | 2 | 5.0% |
| 26-37 | 2 | 5.0% | 31.5 | 4 | 10.0% |
| 37-48 | 5 | 12.5% | 42.5 | 9 | 22.5% |
| 48-59 | 14 | 35.0% | 53.5 | 23 | 57.5% |
| 59-70 | 12 | 30.0% | 64.5 | 35 | 87.5% |
| 70-81 | 3 | 7.5% | 75.5 | 38 | 95.0% |
| 81-92 | 2 | 5.0% | 86.5 | 40 | 100.0% |
| | 40 | 100.0% | | | |

Construyamos entonces un histograma, un polígono de frecuencias y una ojiva "menor que":




SEQ Gráfico * ARABIC 1 Histograma de frecuencias Viajes Moore




*SEQ Gráfico * ARABIC 2* Polígono de frecuencias viajes Moore



Para construir la ojiva elaboramos la siguiente categorización como se ilustra en la tabla 5-2

## Tabla 5-2

Categorización de datos Viajes Moore

| Clases | FA |
|---|---|
| Menos de 26 | 2 |
| Menos de 37 | 4 |
| Menos de 48 | 9 |
| Menos de 59 | 23 |
| Menos de 70 | 35 |
| Menos de 81 | 38 |
| Menos de 92 | 40 |


*SEQ Gráfico * ARABIC 3* Ojiva menor que viajes Moore

# 6. Medidas de tendencia central

Una medida de tendencia central permite ubicar e identificar el punto alrededor del cual se centran los datos, entre las medidas más usadas tenemos:

**Media aritmética (X):** Es un promedio que resulta del cociente entre la suma de los valores de la variable y el número de observaciones. Si los datos son de la muestra o la población, las ecuaciones son las siguientes:

- Cuando los datos estudiados pertenecen a una muestra se está obteniendo un Estadístico, representado por:

$$\overline{X} = \frac{\sum \square X_i}{n}$$

- Si la media obtenida pertenece a la población se conoce como Parámetro y se representa por:

$$\mu = \frac{\sum \square X_i}{N}$$



- Si los datos son agrupados la suma de los valores de la variable será los puntos medios o marcas de clase ($x_i$) y el número de observaciones o datos será la suma de las frecuencias. La ecuación basándonos en datos obtenidos en una muestra es la siguiente:

$$\overline{X} = \frac{\sum \square f_i X_i}{n}$$

- Si utilizamos datos poblaciones la fórmula es:

$$\mu = \frac{\sum f_i X_i}{N}$$

En donde n es el tamaño de la muestra y N de la población.

Algunas características de la media son:

- Todo conjunto de datos de nivel de intervalo o nivel de razón tiene un valor medio.
- Un conjunto de datos solo tiene una media.
- Al evaluar la media se incluyen todos los valores.
- Es útil para comparar dos o más conjuntos de datos.
- Es la única medida de ubicación donde la suma de las desviaciones (distancias) de cada valor con respecto a la media siempre será cero. $\sum (X - \overline{X}) = 0.$
- La media puede verse afectada por valores extremos, que no son representativos de los datos.

La media no puede estimarse cuando se tienen datos agrupados con clases abiertas.

## Ejemplo en Microsoft Excel

| | A | B | C | D |
|---|---|---|---|---|
| 2 | Edad de los Estudiantes de Economía | | Media Aritmética | |
| 3 | 19 | 24 | =PROMEDIO(A3:B16) | |
| 4 | 28 | 30 | 21,8929 | |
| 5 | 19 | 21 | | |
| 6 | 20 | 19 | | |
| 7 | 18 | 19 | | |
| 8 | 22 | 24 | | |
| 9 | 18 | 22 | | |
| 10 | 25 | 21 | | |
| 11 | 19 | 19 | | |
| 12 | 26 | 19 | | |
| 13 | 30 | 20 | | |
| 14 | 20 | 21 | | |
| 15 | 23 | 19 | | |

Para obtener la media aritmética en un cuadro de Exel, se deberá usar la función (=PROMEDIO)

**Media ponderada ($X_w$):** Es una medida de tendencia central, muy parecida a la media aritmética; pero a diferencia de aquella esta considera que los valores de la variable tienen diferentes pesos o ponderaciones en el total de los datos. Su forma de cálculo es mediante la siguiente ecuación:

$$X_w = \frac{\sum (f)_i X_i}{\sum (f)_i}$$

En la tabla 6-1 se calcula la media ponderada con los datos a continuación:

El Hotel Hilton vendió 95 habitaciones para huéspedes selectos a un precio normal de $400. Para la venta de vacaciones las habitaciones vendidas bajaron a $200 y se vendieron 126. En la venta de fin de año el precio rebajó a $100 y se vendieron 79 habitaciones. ¿Cuál es el precio promedio ponderado de una habitación?:

Tabla 6-1
Cálculo de media ponderada

| Temporada | Precio ($x_i$) | Habitaciones ($w_i$) |
|---|---|---|
| Normal | 400 | 95 |
| Vacaciones | 200 | 126 |
| Fin de año | 100 | 79 |
| | | 300 |



$$X_w = \frac{\sum (f)_i X_i}{\sum (f)_i}$$

$$X_w = \frac{(95*400)+(126*200)+(79*100)}{95+126+79} = \$237$$

## Ejemplo en Microsoft Excel


**Media geométrica (MG):** Es una medida que permite conocer el promedio de un conjunto de datos cuando estos son tasas de crecimiento o porcentaje de rendimiento positivos. Su cálculo sigue la siguiente expresión:

$$MG = \sqrt[n]{X_1 * X_2 * X_3 * \ldots * X_n}$$

Esta es una medida muy representativa de datos que se relacionan o están expresados en porcentajes.

Las tasas de variación de la ocupación total de un hotel, en temporada baja en la ciudad de Cuenca se dan como lo muestra la tabla 6-2:



Tabla 6-2

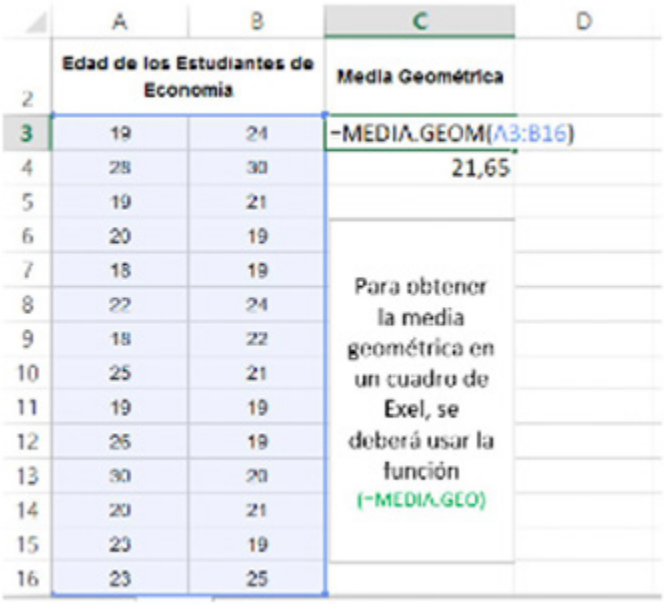
Datos de crecimiento de ocupación en hoteles

| Tasas de variación de ocupación total de un hotel | 5% | 6% | 8% | 2% |
|---|---|---|---|---|

$$MG = \sqrt[4]{5\% * 6\% * 8\% * 2\%} = 4{,}68\%$$

La tasa de ocupación creció en promedio un 4,68%.

## Ejemplo en Microsoft Excel



| | A | B | C | D |
|---|---|---|---|---|
| 2 | Edad de los Estudiantes de Economía | | Media Geométrica | |
| 3 | 19 | 24 | =MEDIA.GEOM(A3:B16) | |
| 4 | 28 | 30 | 21,65 | |
| 5 | 19 | 21 | | |
| 6 | 20 | 19 | Para obtener la media geométrica en un cuadro de Exel, se deberá usar la función (=MEDIA.GEO) | |
| 7 | 18 | 19 | | |
| 8 | 22 | 24 | | |
| 9 | 18 | 22 | | |
| 10 | 25 | 21 | | |
| 11 | 19 | 19 | | |
| 12 | 26 | 19 | | |
| 13 | 30 | 20 | | |
| 14 | 20 | 21 | | |
| 15 | 23 | 19 | | |
| 16 | 23 | 25 | | |

**Media armónica (H):** Es una medida que permite conocer el promedio de un conjunto de datos relacionados con velocidades, el tiempo, etc. Su cálculo es el siguiente:

$$H = \frac{n}{\sum \square \left(\frac{F_i}{X_i}\right)}$$



La tabla 6-3 nos provee información para ejemplificar el cálculo de la media armónica, con respecto al  número de huéspedes:

## Tabla 6-3

Tiempo de hospedaje en hotel

| Número de huéspedes ($x_i$) | Tiempo de hospedaje (días) ($f_i$) | $f_i/x_i$ |
|---|---|---|
| 100 | 10 | 10/100 |
| 120 | 5 | 5/120 |
| 125 | 4 | 4/125 |
| 140 | 3 | 3/140 |
| | n = Σfi | Σ= 0.195 |

$$H = \frac{n}{\sum \left(\frac{f_i}{x_i}\right)} = \frac{22}{0.195} = 112.82 \text{ huéspedes}$$

## Ejemplo en Microsoft Excel

| | A | B | C | D |
|---|---|---|---|---|
| 2 | Edad de los Estudiantes de Economía | | Media Armónica | |
| 3 | 19 | 24 | =MEDIA.ARMO(A3:B16) | |
| 4 | 28 | 30 | 21,42 | |
| 5 | 19 | 21 | | |
| 6 | 20 | 19 | | |
| 7 | 18 | 19 | | |
| 8 | 22 | 24 | | |
| 9 | 18 | 22 | | |
| 10 | 25 | 21 | | |
| 11 | 19 | 19 | | |
| 12 | 26 | 19 | | |
| 13 | 30 | 20 | | |
| 14 | 20 | 21 | | |
| 15 | 23 | 19 | | |
| 16 | 23 | 25 | | |

Para obtener la media armónica en un cuadro de Exel, se deberá usar la función (=MEDIA.ARMO)

**Mediana (Me):** Es otra medida de tendencia central, que divide al conjunto de datos ordenados en dos partes iguales; por lo tanto, el 50% está por encima de esta y el otro 50% estará por debajo. Para su cálculo es importante ordenar los datos de menor a mayor o de mayor a menor.

Si estos constituyen un *número par* la posición de la *Me* será (n+1)/2.

Si la información tiene un *número impar* la posición de la mediana será en la mitad de los datos. Cuando los datos están agrupados es importante indicar que la posición de la *Me* estará ubicada en aquella clase en la que la frecuencia acumulada (FA) sea mayor o igual que n/2 .

Su fórmula de cálculo es:

$$Me = Lmd + \left[\frac{\frac{n}{2} - F}{fmd}\right] * C_i$$

En donde:

*Lmd*: Límite inferior de la clase que contiene a la mediana.
*F*: Frecuencia acumulada de la clase que antecede a la mediana.
ƒmd: Frecuencia absoluta de la clase que contiene a la mediana.
$C_i$: ancho de clase

Utilizando los datos de la tabla 6-4, sobre los sueldos por semana del personal de un hotel:



Tabla 6-4
Sueldos del personal de cocina de un hotel

| **Sueldos / semana** | 45 | 52 | 56 | 67 | 67 |
|---|---|---|---|---|---|

Ordenamos los datos de menor a mayor: 45, 52, 56, 67, 67

La mediana en este caso (para datos no agrupados y número impar) será el valor central es decir Me = 56. Si aumentamos un sueldo, tenemos:

## Tabla 6-5

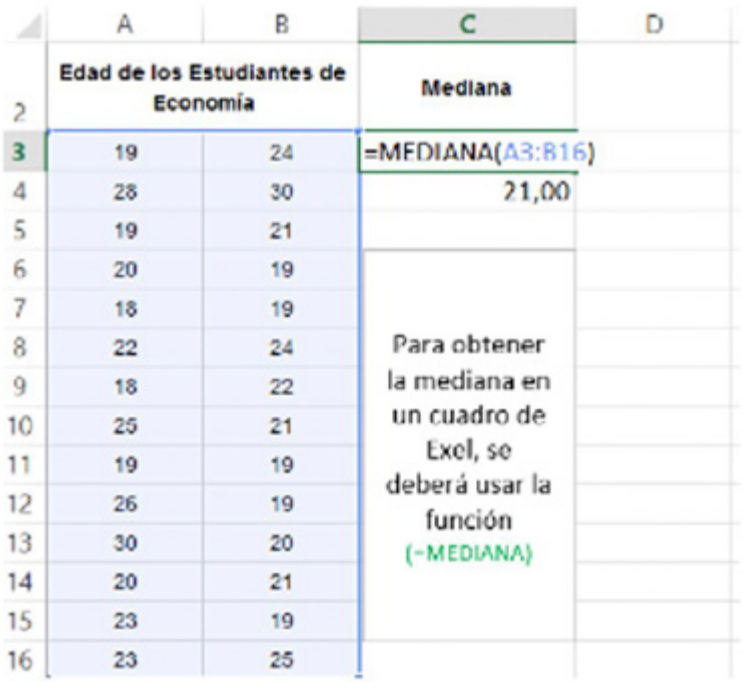
Sueldos del personal en un hotel 2

| Sueldos / semana | 45 | 52 | 56 | 67 | 67 | 35 |
|---|---|---|---|---|---|---|

Ordenamos los datos: 35, 45, 52, 56, 67, 67

La mediana en este caso (para datos no agrupados y número par) será el valor (n+1)/2, es decir, (6+1)/2 = 3.5 por lo tanto, la mediana estará entre el valor 3 y el valor 4 y su cálculo será (52+56)/2 = 54 dólares por semana.

## Ejemplo en Microsoft Excel



**Moda (Mo):** Es otra de las medidas de tendencia central aplicable sobre todo a variables discretas, y comprende el valor de la variable que más se repite. Cuando los datos están agrupados la Mo estará en la clase con la mayor frecuencia. Su cálculo es como sigue:

$$Mo = Lmo + \left[\frac{D_a}{D_b + D_b}\right] * C_i$$

En donde:

*Lmo*: Límite inferior de la clase que contiene a la moda o clase modal.
$D_a$: Diferencia entre la frecuencia de la clase modal y la clase que le antecede.
$D_b$: Diferencia entre la frecuencia de la clase modal y de la clase siguiente
$C_i$: Ancho de clase

Veamos con un ejemplo, usando los datos de la tala 6-6 calcular la mediana y la moda.

## Tabla 6-6

Pasajeros de aerolíneas clasificados por clases

| Número de pasajeros (Clases) | Días frecuencia ($f_i$) | FA |
|---|---|---|
| 50-59 | 3 | 3 |
| 60-69 | 7 | 10 |
| 70-79 | 18 | 28 |
| 80-89 | 12 | 40 |
| 90-99 | 8 | 48 |
| 100-109 | 2 | 50 |
| | 50 | |



Cálculo de la *Me*:

Posición de la mediana: FA> =n/2=28 en donde 28 >= 25, es decir, en la tercera clase se encuentra la mediana.

Aplicamos la fórmula: $\frac{n}{2} = \frac{50}{2} = 25$

$$Me = 70 + \left[25 - \frac{10}{18}\right] * 10 = 78.33\ pasajeros$$

Cálculo de la moda:

Determinamos la clase modal, que es la que tiene la mayor frecuencia absoluta. En este ejemplo será la tercera clase, ya que su frecuencia es 18.

$$Mo = 70 + \left[\frac{18 - 7}{(18 - 12) + (18 - 7)}\right] * 10 = 76.47\ pasajeros$$

## Ejemplo en Microsoft Excel

| | A | B | C | D |
|---|---|---|---|---|
| 2 | Edad de los Estudiantes de Economía | | MODA | |
| 3 | 19 | 24 | =MODA.UNO(A3:B16) | |
| 4 | 28 | 30 | 19,00 | |
| 5 | 19 | 21 | | |
| 6 | 20 | 19 | | |
| 7 | 18 | 19 | | |
| 8 | 22 | 24 | | |
| 9 | 18 | 22 | | |
| 10 | 20 | 21 | | |
| 11 | 19 | 19 | | |
| 12 | 26 | 19 | | |
| 13 | 30 | 20 | | |
| 14 | 20 | 21 | | |
| 15 | 23 | 19 | | |
| 16 | 23 | 25 | | |

Para obtener la moda en un cuadro de Exel, se deberá usar la función (=MODA.UNO)

Ahora pasemos a ver otro conjunto de medidas para describir datos; y nos referimos a las medidas de posición o ubicación:

Con la utilización de Excel la determinación de estas medidas de tendencia central se facilitan, para ello se debe activar: Complementos, complementos de Excel, Herramientas de análisis.

**Ejemplo:**

Tabla 6-7
Tabla aplicación Excel

| CALIFICACIONES /50 | | |
|---|---|---|
| 40 | 25 | 29 |
| 38 | 30 | 34 |
| 42 | 42 | 37 |
| 45 | 38 | 36 |
| 30 | 30 | 42 |
| 49 | 45 | 38 |
| 39 | 60 | 50 |

Se dispone de los datos de la tabla 6-7, que serán colocados y ordenados en la columna A, se activa Datos/Análisis de Datos/ Estadística Descriptiva, Rango de entrada A21:A21, Agrupado por columnas, Rango de Salida C10, Resumen de estadísticas √, Aceptar, con lo que se obtendrá una tabla similar a los mostrados en la tabla 6-8:

## Tabla 6-8

Resultados de aplicación análisis de datos Excel

| Calificaciones | |
|---|---|
| Media | 39 |
| Error típico | 1,78752289 |
| Mediana | 38 |
| Moda | 38 |
| Desviación estándar | 8,19145897 |
| Varianza de la muestra | 67,1 |
| Curtosis | 0,85345599 |
| Coeficiente de asimetría | 0,60446462 |
| Rango | 35 |
| Mínimo | 25 |
| Máximo | 60 |
| Suma | 819 |
| Cuenta | 21 |

# 7. Medidas de posición

Se denominan también medidas de localización de los datos, y permiten medir o informar el valor que la variable toma en una posición que sea de interés conocer respecto de todo el conjunto de datos, que se encuentran previamente ordenados de acuerdo a su magnitud.

Es posible identificar tres medidas de posición: los percentiles, los deciles y los cuartiles.

**Cuartiles (Q):** Es una medida de posición que divide al conjunto de datos en cuatro partes iguales, por lo tanto, ubicamos tres cuartiles:

$Q_1$ = primer cuartil
$Q_2$ = segundo cuartil
$Q_3$ = tercer cuartil. Gráficamente:

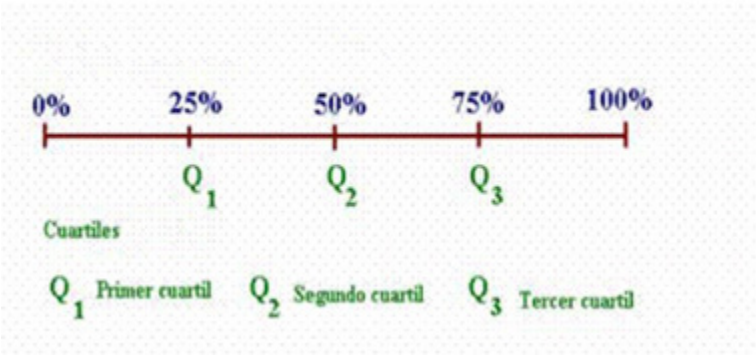



Figura 7-1. Cuartiles

Por debajo del primer cuartil ($Q_1$) se encuentra el 25% de los datos, en el $Q_3$ están el 25% superior, y entre $Q_3$y $Q_1$ se encuentra concentrado el 50% de los datos. La diferencia entre $Q_3$y $Q_1$ se denomina: Rango intercuartílico.

**Deciles (D):** Es una medida de posición que divide al conjunto de datos en diez partes iguales, por lo tanto, ubicamos 9 deciles.


Figura 7-2. Deciles

**Percentiles (P):** Es una medida de posición que divide al conjunto de datos en cien partes iguales, por lo tanto, ubicamos 99 percentiles.

Gráficamente:

Figura 7-3. Percentiles

La ubicación de los percentiles para datos no agrupados se localiza mediante la siguiente ecuación:

$$Lp = \frac{(n+1)}{(\frac{P}{100})}$$

*Lp*: es el sitio del percentil deseado en una serie ordenada.
*n*: es el número de observaciones.
*P*: es el percentil deseado.

Esta ecuación entonces permite ubicar un percentil, un decil o un cuartil, puesto que el decil 5 por ejemplo es el percentil 50, o el cuartil tres ($Q_3$) es el percentil 75, etc.

Para mejor comprensión, se utilizarán los datos de la tabla 7-1, en la que se muestran las horas trabajadas por un mesero en un restaurante cada semana durante los últimos dos meses:

Tabla 7-1
Horas trabajadas

| Horas trabajadas | 52 | 48 | 37 | 54 | 48 | 15 | 42 | 12 |
|---|---|---|---|---|---|---|---|---|

Pasamos a ordenar en primer lugar los datos: 12, 15, 37, 42, 48, 52, 54

Calculemos el percentil 40 ($P_{40}$) el rango intercuartílico:

$$L_{40} = (8+1)\frac{40}{100} = 3.6$$



Esto quiere decir que el P40 está entre la posición 3 y la posición 4, por tanto existe un 60% de distancia entre las dos posiciones:

$$P_{40} = 37 + 60\% (42 - 37) = 40$$

Para calcular el *RIQ* calculamos primero $Q_1$ y $Q_3$:

$Q_1 = P_{25}$ por lo tanto:

$$L_{25} = (8+1)\frac{25}{100} = 2.25$$

$$P_{25} = 15 + 25\%\,(37 - 15) = 20.5$$

$Q_3 = P_{75}$ por lo tanto:

$$L_{75} = (8 + 1)\frac{75}{100} = 6.75$$

$$P_{75} = 48 + 75\%(52 - 48) = 51$$

Por lo tanto $RIQ = Q_3 - Q_1 = 51 - 20.5 = 30.50$

La ubicación de los percentiles para *datos agrupados* se localiza mediante la siguiente ecuación:

$$P = Lp_{i-1} + \left[\frac{\%N - FA}{fp_i}\right] * c_i$$

En donde:

$P$: Es el percentil buscado.
$Lp_{i\text{-}1}$: Es el límite inferior de la clase que contiene el percentil buscado.
$\%N$: Es la localización del percentil buscado.
$FA$: Frecuencia acumulada de la clase que antecede a la clase del percentil buscado.
$fp_i$: Frecuencia absoluta de la clase que contiene al percentil buscado.
$c_i$: Ancho de clase.



Antes de aplicar esta fórmula debemos buscar:

La posición del percentil buscado es (P/100)$N$.
La frecuencia acumulada >= a Percentil buscado.

Se dispone de información referente al número de turistas que visitaron Galápagos el verano anterior.

## Tabla 7-2

Visitantes a las Islas Galápagos

| Turistas visitantes | Número/frecuencia | FA |
|---|---|---|
| 0-100 | 90 | 90 |
| 100-200 | 140 | 230 |
| 200-300 | 150 | 380 |
| 300-400 | 120 | 500 |
| | N = 500 | |

Supongamos que queremos buscar el percentil 25 ($P_{25}$) o el $Q_1$:

Posición del percentil buscado = $\left(\frac{25}{100}\right) * 500 = 125$

La frecuencia acumulada >= a 125 es 230, por lo tanto el percentil buscado está en la segunda clase.

Aplicamos la ecuación para el cálculo del percentil y obtenemos:

$$P_{25} = 100 + \left[\frac{125 - 90}{140}\right] * 100 = 125$$

Por ejemplo el decil 4.



Sabemos que el decil 4 es el percentil 40 es decir $D_4 = P_{40}$

Posición del percentil buscado =(40/100)*500 = 200

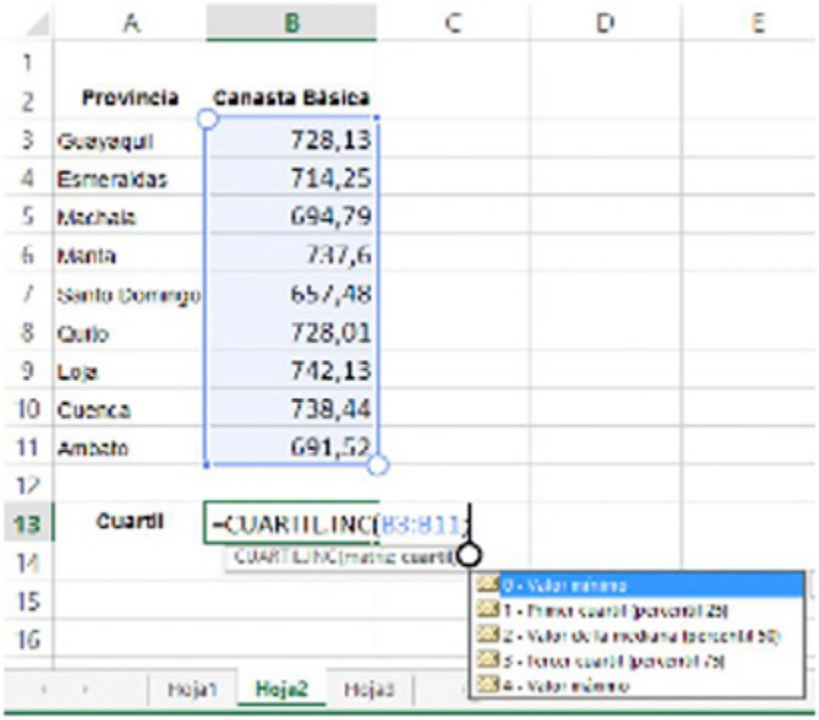
La frecuencia acumulada >= a 200 y es 230, por lo tanto el percentil buscado está en la segunda clase.

Aplicamos la ecuación para el cálculo del percentil y obtenemos:

$$P_{25} = 100 + \left[\frac{200 - 90}{140}\right] * 100 = 178.57$$

Con la utilización de Excel para el análisis de los datos de la tabla 7-3, se deben seguir los pasos que se detallan a continuación:

## Ejemplo en Microsoft Excel






Calcular $Q_1$, $Q_2$,$Q_3$, $P_{95}$

$Q_1$=PERCENTIL(A1:A10;0,25) O $Q_1$=CUARTIL.INC(B1:B10;1) =694,79

$Q_2$=PERCENTIL(A1:A10;0,50) O $Q_2$=CUARTIL.INC(B1:B10;2) = 728,01

$Q_3$=PERCENTIL(A1:A10;0,75) O $Q_3$=CUARTIL.INC(B1:B10;3) =737,60

$P_{95}$=PERCENTIL(A1:A10;0,95) = 740,65

Además de estudiar la ubicación de los datos (sin conocer la forma de la distribución de estos) podemos ayudarnos a ilustrar la simetría de ese conjunto con el *diagrama de caja*, que es una representación gráfica en forma de caja basada en cuartiles. Se construye con la siguiente información:

1. el valor mínimo
2. el $Q_1$
3. la Me
4. $Q_3$
5. el valor máximo

**Veamos un ejemplo:**

Supongamos que la media de un conjunto de datos es 27.5, la $Me$ = 26.8. El valor mínimo es 12.7, el valor máximo es 50.2, $Q_1$ = 17.95 y $Q_3$ =35.45. Con estos datos construya el diagrama de caja.

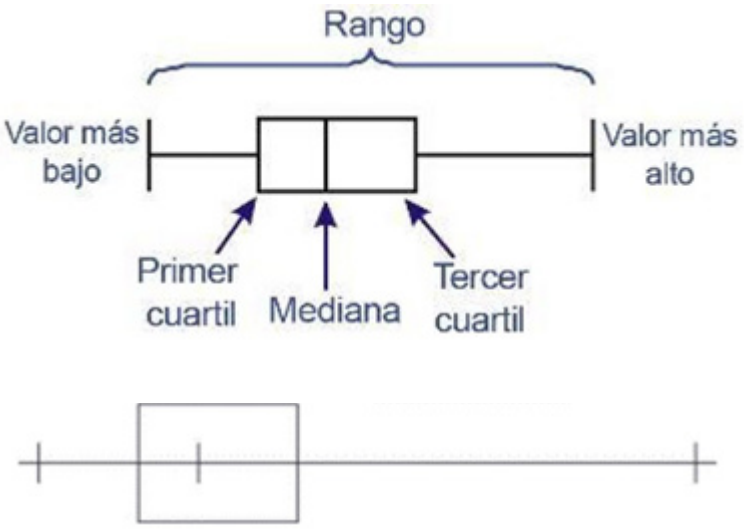





Figura 7-4. Diagrama de Caja

Como se aprecia, la $Me$ no divide la caja en dos partes iguales. Existe un sesgo hacia la derecha; por lo tanto, esta es una distribución en forma asimétrica.

# 8. Medidas de dispersión

Al describir los datos no solo es necesario conocer el valor central o el promedio de ese conjunto de datos, sino que también nos interesa estudiar cuán distantes, dispersos o alejados están los datos respecto de ese promedio que normalmente es la media aritmética.

Tenemos algunas medidas de dispersión, siendo las más relevantes las siguientes:

**a) Rango o amplitud de variación:**
Este factor estadístico mide la distancia entre el valor más alto y el valor más bajo del conjunto de datos, es decir:

$$R = Valor\ más\ alto - Valor\ más\ bajo$$

**b) Varianza:**
Es el promedio de las desviaciones de las observaciones respecto de su media aritmética, elevadas al cuadrado:

La varianza para *datos no agrupados* es la siguiente:

Si estamos trabajando con la población entonces diremos que la varianza de la población es como sigue:



**b.1) Varianza poblacional ($\delta^2$)**

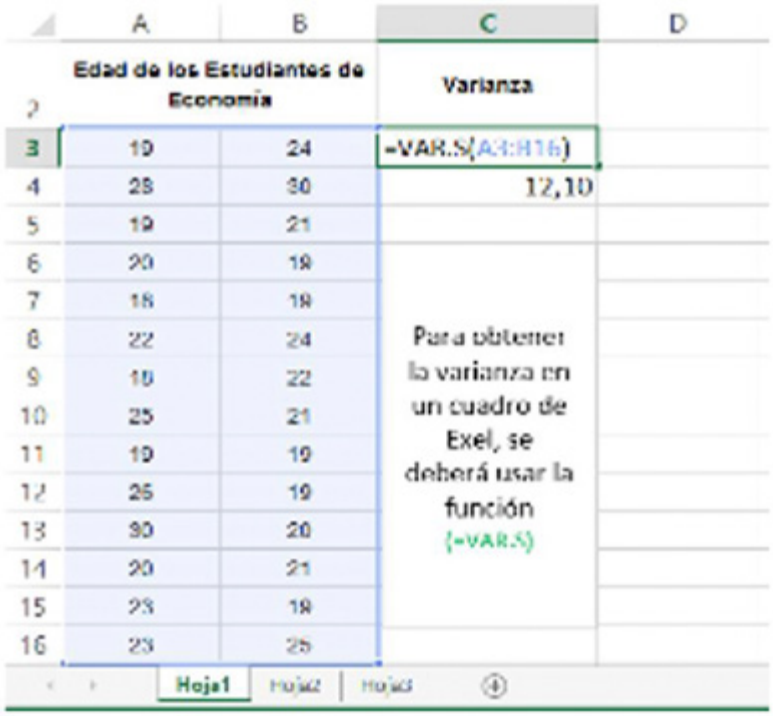
$$\delta^2 = \frac{\sum (X_i - \mu)^2}{N}$$

Si estamos trabajando con la muestra entonces nos interesa calcular la varianza muestral su cálculo es el siguiente:

**b.2) Varianza muestral ($S^2$)**

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

## Ejemplo en Microsoft Excel




**c) Desviación estándar:**

La desviación típica o desviación estándar es la raíz cuadrada de la varianza, y de igual manera podemos estimarla para datos de la población como para datos muestrales.

La desviación estándar para datos no agrupados y agrupados es igual a:

**c.1) Desviación estándar poblacional (δ):**

$$\delta = \sqrt{\left[\frac{\sum (X_i - \mu)^2}{N}\right]}$$

Si estamos trabajando con la muestra entonces nos interesa calcular la desviación muestral; y su cálculo es el siguiente:

**c.2) Desviación muestral (S)**

$$S = \sqrt{\left[\frac{\sum (X_i - \underline{X})^2}{n-1}\right]}$$

La varianza de datos muestrales para *datos agrupados* es la siguiente:

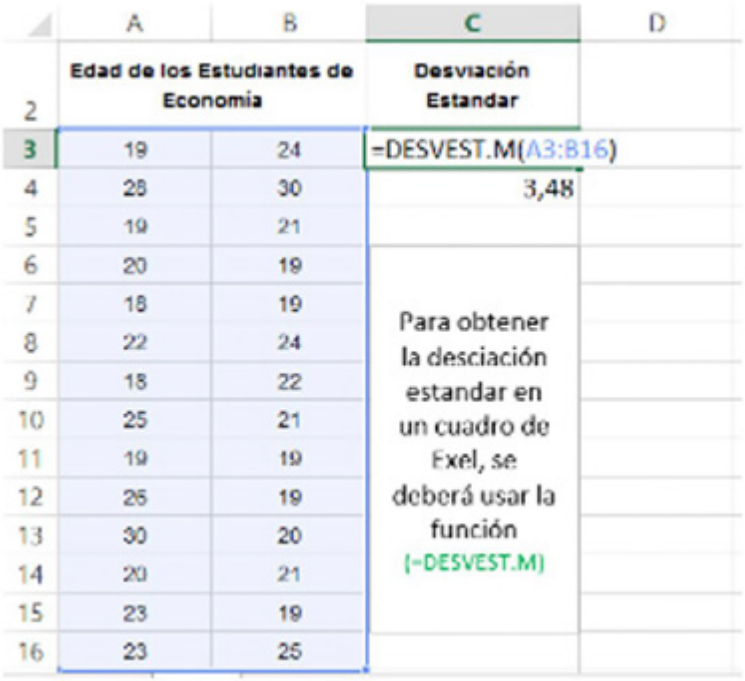
$$S^2 = \frac{\sum f Xi^2 - n\,\underline{X}^2}{n-1}$$

Por lo tanto la desviación estándar de la muestra en datos agrupados sigue siendo la raíz cuadrada de la varianza.



$$S = \sqrt{\frac{\sum f Xi^2 - n\,\underline{X}^2}{n-1}}$$

## Ejemplo en Microsoft Excel




### d) Coeficiente de variación (CV)

También es una medida de dispersión que mide el porcentaje de variabilidad de los datos, y se usa generalmente cuando se consideran dos o más distribuciones que tienen medias aritméticas significativamente distintas o cuando están medidas en unidades diferentes. Su fórmula de cálculo es como sigue

$$\mathrm{CV} = \frac{S}{X} * 100$$

**Ejercicio 1:**
Las horas trabajadas por usted cada semana, durante los últimos dos meses son: 52, 48, 37, 54, 48, 15, 42, 12. Calcule el rango, la varianza, la desviación estándar y el coeficiente de variación:

Solución:

Tabla 8-1
Ejercicio 1

| Observaciones | $X_i$ (horas trabajadas) | $(X_i - \bar{x})^2$ |
|---|---|---|
| 1 | 12 | 702.25 |
| 2 | 15 | 552.25 |
| 3 | 37 | 2.25 |
| 4 | 42 | 14.0625 |
| 5 | 48 | 95.0625 |
| 6 | 48 | 95.0625 |
| 7 | 52 | 182.25 |
| 8 | 54 | 240.25 |
| n = 8 | Σxi = 308 | Σ=1.883.7375 |

$$\overline{x} = \frac{\sum fx_i}{n} = \frac{393.5}{50} = 78.7$$

$$S^2 = \frac{316902.5 - 50 \cdot (78.7)^2}{49} = 174.31$$

$$S = 12.14$$

**Ejercicio 2:**
Los pasajeros de la aerolínea A&A, que viajaron el último mes fueron agrupados en las siguientes categorías. Calcule la media, varianza y la desviación estándar.

## Tabla 8-2
Ejercicio 2

| Número de pasajeros | ƒi (días) | Xi (Punto medio) | ƒ*Xi | $Xi^2$ | $ƒ*Xi^2$ |
|---|---|---|---|---|---|
| 50-59 | 3 | 54.5 | 163.5 | 2970.25 | 8910.75 |
| 60-69 | 7 | 64.5 | 451.5 | 4160.25 | 29121.75 |
| 70-79 | 18 | 74.5 | 1341.0 | 5550.25 | 99904.5 |
| 80-89 | 12 | 84.5 | 1014.0 | 7140.25 | 85683.0 |
| 90-99 | 8 | 94.5 | 756.0 | 8930.25 | 71442.0 |
| 100-109 | 2 | 104.5 | 209.0 | 10920.25 | 21840.5 |
| | N = 50 | | Σ=393.5 | | Σ=316902.5 |

$$\overline{x} = \frac{\Sigma fx_i}{n} = \frac{393.5}{50} = 78.7$$

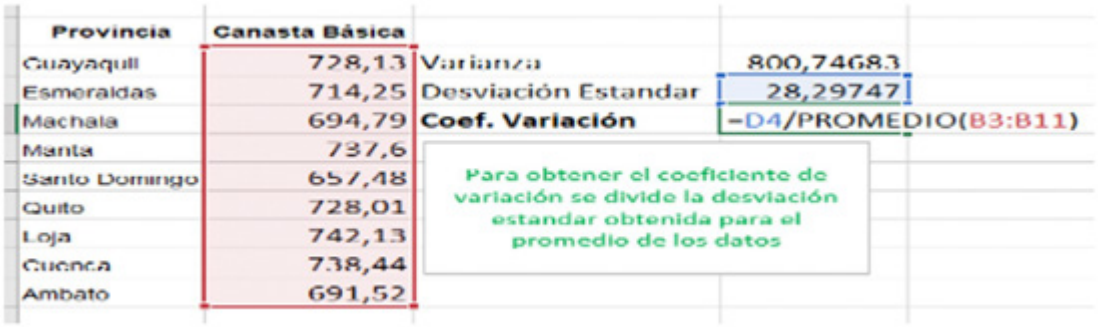
$$S^2 = \frac{316902.5 - 50*(78.7)^2}{49} = 174.31$$

$$S = 12.14$$



La desviación estándar es la medida de dispersión más útil para describir un conjunto de datos, midiendo el grado de dispersión de las observaciones individuales alrededor de la media y por ello existen dos aplicaciones básicas que son el teorema de Chebychev y la regla empírica:

## Ejemplo en Microsoft Excel



| Provincia | Canasta Básica | | |
|---|---|---|---|
| Guayaquil | 728,13 | Varianza | 800,74683 |
| Esmeraldas | 714,25 | Desviación Estandar | 28,29747 |
| Machala | 694,79 | **Coef. Variación** | =D4/PROMEDIO(B3:B11) |
| Manta | 737,6 | | |
| Santo Domingo | 657,48 | | |
| Quito | 728,01 | | |
| Loja | 742,13 | | |
| Cuenca | 738,44 | | |
| Ambato | 691,52 | | |

Para obtener el coeficiente de variación se divide la desviación estandar obtenida para el promedio de los datos

## 8.1 Teorema de Chebyshev (1821-1894)

Este teorema establece que:

Para todo conjunto de datos por lo menos $1-1/k^2$ por ciento de las observaciones está dentro de k desviaciones estándar de la media, en donde k es un número mayor que uno (k>1).

**Por ejemplo:**

Un conjunto de datos tiene una media de 5.000 y una desviación estándar de 400 ¿Qué porcentaje de las observaciones están entre 4500 y 5500?



$$X \pm ks = X + k\,(400) = 5500, \text{ es decir, } 5000 + 400k = 5500 \text{ de donde } K = 1.5$$

$$X - k(400) = 4500, \text{ es decir, } 5000 - 400k = 4500 \text{ de donde } k = 1.5$$

Por lo tanto siguiendo el teorema de Chebyshev:

$$(1 - \frac{1}{1.5^2}) * 100 = 55.56\% = 56\% \text{ de los datos está entre 4500 y 5500.}$$

## 8.2 La regla empírica

Esta regla se ajusta cuando los datos se distribuyen de una manera simétrica, y son datos continuos, no discretos, y afirma que:

Cuando nos alejamos ± una desviación estándar de la media, el porcentaje de datos que se aceptan en ese rango es del 68.3%; cuando nos alejamos ±2 desviaciones estándar, el rango de aceptación de los datos es del 95.5%; y cuando nos alejamos ±3 desviaciones estándar, el porcentaje es del 99.7%.

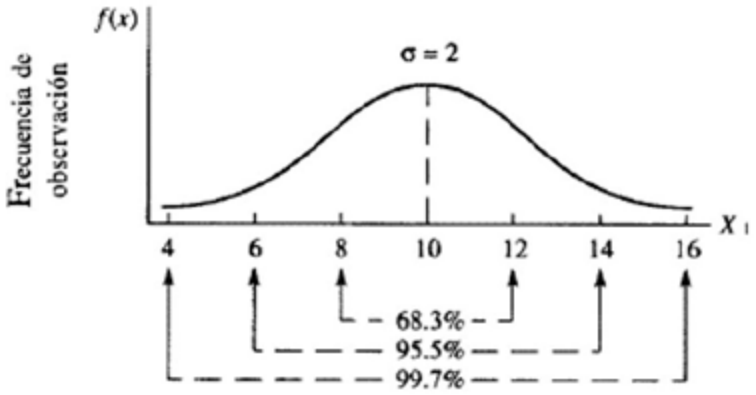





Ilustración 8.2-1- Regla empírica

**Por ejemplo:**

Un conjunto de datos distribuidos normalmente tiene una media de 5.000 y una desviación estándar de 450 ¿Qué porcentaje de las observaciones están por debajo de 4550?

Por definición el área dentro de la curva normal vale el 100% de las observaciones, si nos alejamos ± 1 entonces el área dentro de ese rango es de 68.3%, por lo tan-

to, de acuerdo con el gráfico, el resto será del 100% menos el 68.3%, es decir, el 31.70% de los datos estará por debajo de 4550.

Ahora no todas las distribuciones de los datos siguen una distribución normal; algunas están sesgadas ya sea a la derecha o a la izquierda, es decir, no hay simetría en los datos, por lo tanto para medir la simetría usamos otro grupo de medidas.

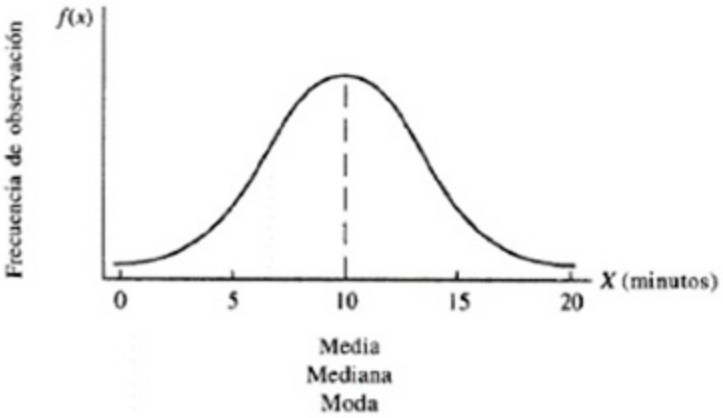



Figura 8.2-.2 Distribución normal

# 9. Medidas de simetría

Las medidas de simetría permiten medir la disposición de los datos o informar la forma de una distribución.

Si trazamos una línea vertical por el valor de la media de una variable en el diagrama de barras o en el histograma, esta línea vertical se convierte en el *eje de simetría* cuando a ambos lados de la media haya el mismo número de valores de la variable o sean equidistantes. Si tienen la misma frecuencia absoluta diremos que su distribución es simétrica; de lo contrario será asimétrica o sesgada, dependiendo del signo resultante podremos concluir si el sesgo es hacia la derecha o a la izquierda.

Usaremos la siguiente medida de simetría:

## 9.1 Coeficiente de sesgo de Pearson (P)

$$P = 3\left(\frac{\overline{X} - Me}{S}\right)$$

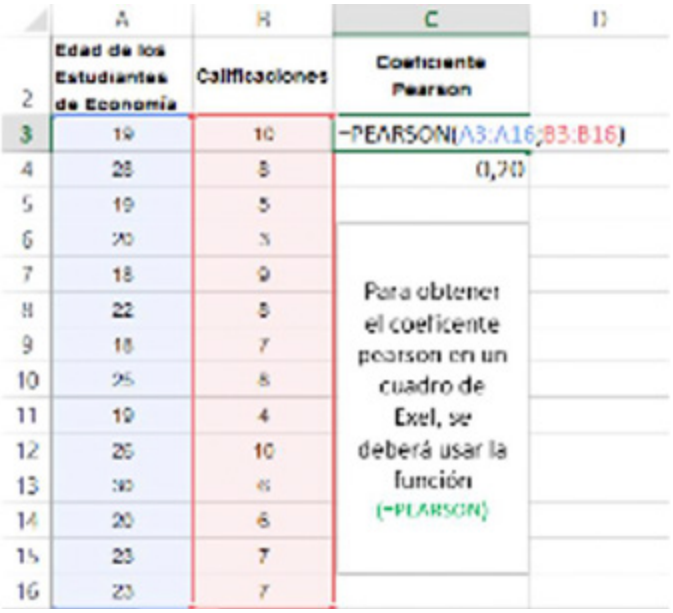


Si P>0 la distribución es sesgada a la derecha.

Si P<0 la distribución es sesgada a la izquierda.

Si P=0 la distribución no tiene sesgo es normal.

## Ejemplo en Microsoft Excel



| | A | B | C | D |
|---|---|---|---|---|
| 2 | Edad de los Estudiantes de Economía | Calificaciones | Coeficiente Pearson | |
| 3 | 19 | 10 | =PEARSON(A3:A16;B3:B16) | |
| 4 | 28 | 8 | 0,70 | |
| 5 | 19 | 5 | | |
| 6 | 20 | 8 | Para obtener el coeficiente pearson en un cuadro de Exel, se deberá usar la función (=PEARSON) | |
| 7 | 18 | 9 | | |
| 8 | 22 | 8 | | |
| 9 | 18 | 7 | | |
| 10 | 25 | 8 | | |
| 11 | 19 | 4 | | |
| 12 | 26 | 10 | | |
| 13 | 30 | 6 | | |
| 14 | 20 | 6 | | |
| 15 | 23 | 7 | | |
| 16 | 23 | 7 | | |

## 9.2 Coeficiente de Fisher (g1)

$$g1 = \frac{\sum (X_i - \overline{X})^3 * n}{n^s}$$



Si g1>0 la distribución es asimétrica positiva.
Si g1<0 la distribución es asimétrica negativa.
Si g1=0 la distribución es simétrica.

En la tabla 9.2-1 se listan las notas de un examen de estadística de la primera evaluación, valoradas sobre 100, y son: 80, 83, 87, 85, 90, 86, 84, 82, 88. Calcular el sesgo de Pearson y el de Fisher.

Solución:

Ordenamos los datos en la siguiente distribución:

## Tabla 9-2

Cálculo de coeficientes de Pearson y Fisher

| Notas | $N$ | $(x_i\text{-}\bar{X})^2$ | $(x_i\text{-}\bar{X})^3$ | $(x_i\text{-}\bar{X})^4$ |
|---|---|---|---|---|
| 80 | 1 | 25 | 125 | 625 |
| 82 | 2 | 9 | 27 | 12 |
| 83 | 3 | 4 | 8 | 16 |
| 84 | 4 | 1 | 1 | 1 |
| 85 | 5 | 0 | 0 | 0 |
| 86 | 6 | 1 | -1 | 1 |
| 87 | 7 | 4 | -8 | 16 |
| 88 | 8 | 9 | -27 | 12 |
| 90 | 9 | 25 | -125 | 625 |
| $\Sigma x_i = 765$ | | 78 | 0 | 1308 |

La media aritmética es 85:

$$\overline{X} = \frac{\Sigma \square x_i}{n} = \frac{765}{9} = 85$$



Posición de la mediana: $\frac{n+1}{2} = 5$, es decir, está en la quinta observación: por lo tanto, la $Me$ = 85

La desviación es como sigue:
Pearson es 0

S = 78/8 = 3.12

$$P=3\left(\frac{85-85}{3.12}\right) = 0$$

Y el coeficiente de Fisher es también 0

$$g1 = \frac{0 * 50}{50 * 30.37} = 0$$

En consecuencia, la distribución de estos datos es simétrica.

# 10. Medidas de apuntamiento

El grado de apuntamiento en la forma de una distribución lo vamos a medir con la curtosis. Para estudiar el grado de curtosis de una distribución de frecuencias se emplea un coeficiente denotado por $g_2$ con la siguiente expresión:

$$g_2 = \frac{\sum (x_i - \overline{X})^4 * n}{ns^4} - 3$$

- Si $g_2$>0 la distribución es apuntada o leptocúrtica, tiene tendencia a ser alta y alargada.
- Si $g_2$<0 la distribución es menos apuntada o platicúrtica, es baja y achatada.
- Si $g_2$=0 la distribución es normal o mesocúrtica, se eleva en el punto central.

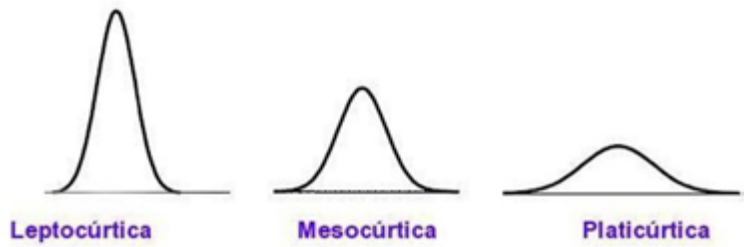



*Figura 10-1*, Tipos de curtosis en distribuciones

Del ejemplo 9.2, se puede decir que la distribución no es normal sino leptocúrtica o apuntada ya que g2>0.

$$g_2 = \frac{1308 * 9}{9 * 94.76} - 3 = 10.80$$

# 11. Medidas de concentración

Las medidas de desigualdad o concentración sintetizan el grado de equidad en el reparto de las observaciones de la variable. Generalmente el estudio de la concentración se realiza sobre variables como la renta o el ingreso.

Existen varias medidas de la concentración; pero para el presente estudio escogeremos el denominado índice de GINI, que lo denotaremos como *IG* y se calculará así:

$$IG = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

Si $IG = 0$ entonces $p_i = q_i$ y existe un reparto equitativo de los datos. La concentración es mínima.

Si $IG = 1$ entonces $q_i = 0$ y existe un reparto no equitativo de los datos. La concentración es máxima.



Si graficamos estos pares de datos $p_i$ y $q_i$ representados en un cuadrado de lado 100, se obtendrá una línea poligonal llamada curva de Lorenz.

La curva de Lorenz refleja cómo se reparte el total de los datos. Si la curva coincide con la recta de 45 grados o la diagonal del cuadrado, diríamos que no existe concentración o hay la máxima equidad en el reparto de los datos; y si cae o coincide dentro o con los lados del cuadrado, diríamos que existe concentración máxima o mínimo grado de equidad en el reparto de los datos.

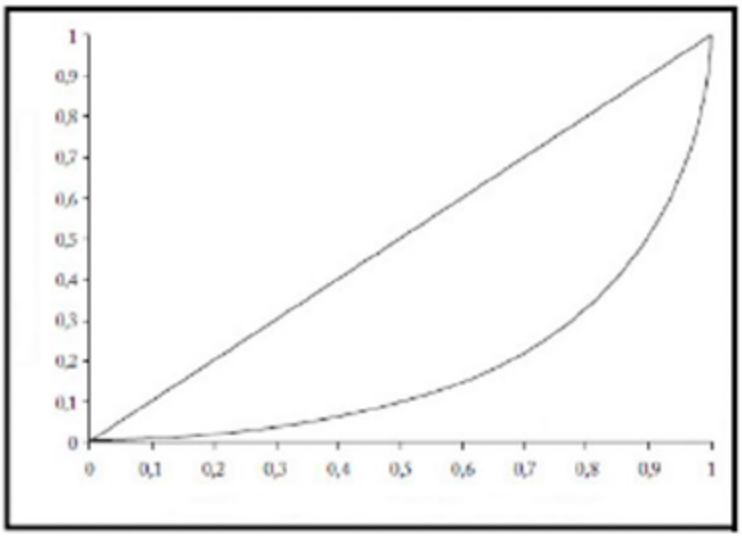



Figura 11-.1 Curva de Lorenz

Veamos con un ejemplo:

La siguiente tabla muestra la distribución de los ingresos en dólares de los espectadores que siguieron la gira en todo el mundo Ziggy Stardust de David Bowie, en 1972. Dicha gira, una de las más exitosas de todos los tiempos, sirve a las principales agencias mundiales como referencia para conocer el tipo de público que asiste a estos grandes acontecimientos como el tour <The rising> de Bruce Springsteen del año 2003, y de esta manera poner el precio a paquetes turísticos promocionales.



Calcule el índice de Gini y dibuje la curva.

## Tabla 11-1

Ingreso de gira David Bowie

| Rango | | Fi | Xi | FA | XiFi | FAXiFi | qi | pi | pi-qi |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1000 | 1 | 500 | 1 | 500 | 500 | 0,0002 | 0,0098 | 0,0096 |
| 1000 | 2000 | 2 | 2000 | 3 | 4000 | 4500 | 0,0021 | 0,0294 | 0,0273 |
| 2000 | 3000 | 3 | 3500 | 6 | 10500 | 15000 | 0,0070 | 0,0588 | 0,0518 |
| 3000 | 4000 | 4 | 5000 | 10 | 20000 | 35000 | 0,0164 | 0,0980 | 0,0816 |
| 4000 | 5000 | 5 | 6500 | 15 | 32500 | 67500 | 0,0317 | 0,1471 | 0,1154 |
| 5000 | 6000 | 5 | 8000 | 20 | 40000 | 107500 | 0,0505 | 0,1961 | 0,1456 |
| 6000 | 7000 | 5 | 9500 | 25 | 47500 | 155000 | 0,0728 | 0,2451 | 0,1723 |
| 7000 | 10000 | 15 | 12000 | 40 | 180000 | 335000 | 0,1573 | 0,3922 | 0,2349 |
| 10000 | 15000 | 26 | 17500 | 66 | 455000 | 790000 | 0,3709 | 0,6471 | 0,2762 |
| 15000 | 25000 | 26 | 27500 | 92 | 715000 | 1505000 | 0,7066 | 0,9020 | 0,1954 |
| 25000 | 50000 | 8 | 50000 | 100 | 400000 | 1905000 | 0,8944 | 0,9804 | 0,0860 |
| 50000 | 125000 | 2 | 112500 | 102 | 225000 | 2130000 | 1,0000 | 1,0000 | - |
| | | 102 | 254500 | 480 | 2130000 | 7050000 | 3,3099 | | 1,3960 |

De donde se obtiene que IG = 0,42, al ser $IG$ mayor que cero, diríamos que los ingresos están muy concentrados

Graficamos la curva de Lorenz:

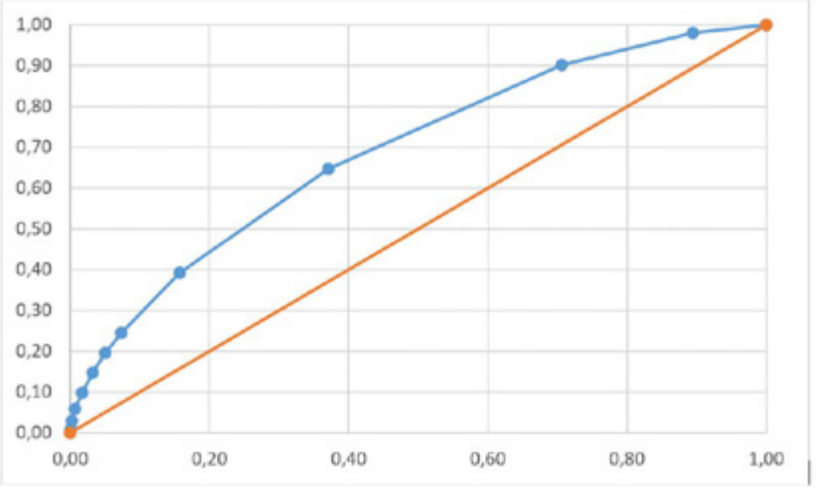



Figura 11-2. Curva de Lorenz Ingresos de Gira David Bowie

# 12. Análisis exploratorio de variables bidimensionales

Hasta el capítulo anterior estuvimos hablando únicamente de la observación de datos que corresponden a una sola variable, por ejemplo, el número de huéspedes de un hotel, los pasajeros de una aerolínea, etc. Ahora en este acápite vamos a describir la observación conjunta de dos variables normalmente conocidas como la variable X y la variable Y en las N unidades de una población.

Esta observación conjunta conduce a la obtención de pares de datos, tal que si $X_1$, $X_2$,.....$X_h$ son los valores de $X$; $Y_1, Y_2$,....$Y_k$ son los valores de $Y$. Estos valores podemos obtenerlos de una variable bidimensional que tiene una distribución de frecuencias, que estará representada en la tabla de contingencia o denominada también tabla de correlación.

La frecuencia absoluta conjunta se define como el número de veces que aparecen simultáneamente los valores de $X_i$ e $Y_j$ en las unidades de la población. Se representa por $n_{ij}$, y cumple la siguiente relación:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} n_{ij} = N$$



La frecuencia relativa conjunta la representaremos como $f_{ij}$; es la proporción de observaciones iguales a dicho valor con respecto al total de observaciones, es decir:

$$f_{ij} = \frac{n_{ij}}{N}$$

Entonces la suma de estas frecuencias es igual a 1, por tanto:

$$\sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = 1$$

Tabla de correlación, denominada también tabla de contingencia, es una forma sencilla de disponer de información proporcionada por una distribución BIDIMENSIONAL de frecuencias; así, si suponemos que tanto los valores de x como los de y están ordenados de menor a mayor tendremos:

## Tabla 12-1.

Tabla de Correlación

| $Y$ | $Y_1$ | $Y_2$ | ..... | ..... | $Y_j$ | $Y_k$ |
|---|---|---|---|---|---|---|
| $X$ | | | | | | |
| $X_1$ | $N_{1,1}$ | $N_{1,2}$ | ..... | ..... | $N_{1,j}$ | |
| : | : | : | | | | : |
| : | : | : | | | | : |
| $X_i$ | $N_{i,1}$ | | | | | : |
| : | : | : | ..... | ..... | | : |
| $X_h$ | $N_{h,1}$ | | | | | $N_{h,k}$ |

Uno de los aspectos fundamentales en el estudio conjunto de dos variables es el análisis de la posible relación existente entre ellas. La estadística permite, mediante procedimientos matemáticos, determinar si las variables tienen o no relación, y medir el grado de la misma.

Gráficamente la forma de una distribución bidimensional de frecuencias la podemos representar a través de un *diagrama de dispersión* o *nube de puntos* o con una *urbe de puntos* que muestra los pares de datos en un plano o eje cartesiano en el cual la variable $X$ graficamos en el eje de las abscisas y la variable $Y$ en el eje de las ordenadas.

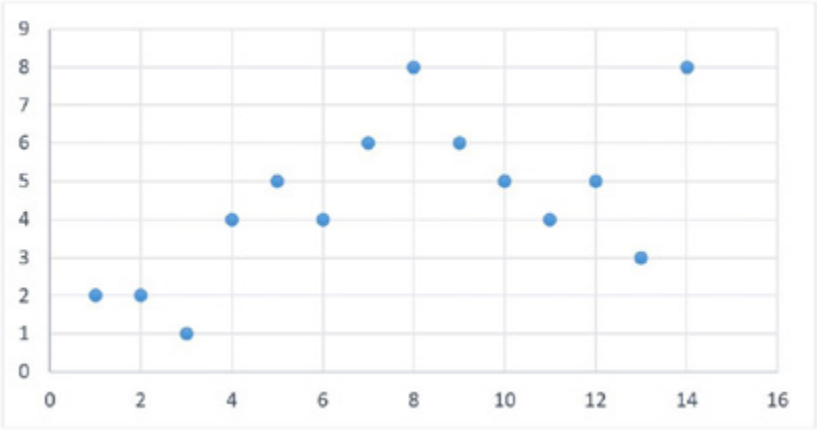



*Figura 12-1.* Diagrama de dispersión

Los factores estadísticos que nos ayudan a determinar el grado de relación o dependencia de las variables son la *covarianza* y la *correlación*, medidas a través del c*oeficiente de correlación o el coeficiente de determinación.*

La covarianza es una estadística que mide la interrelación entre dos variables ($X$ e $Y$) y se representa también por $S_{xy}$; y está dada por la siguiente relación:

$$S_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \bar{x}).(y_i - \bar{y}) f_{ij}$$

Pero como $\sum_{i=1}^{h} \sum_{j=1}^{k} f_{ij} = 1$

Entonces la covarianza estará dada por:

$$S_{xy} = \sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \bar{x}).(y_i - \bar{y})$$

Igualmente, el coeficiente de correlación es un factor estadístico que mide el grado de relación de una variable con otra. Se representa por r y está dado por la siguiente ecuación:

$$r = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2 * \sum (Y_i - \overline{Y})^2}$$

Aunque la forma de la relación puede ser muy diversa consideramos únicamente la existencia de una relación lineal o que puedan transformarse en relaciones lineales mediante sencillas operaciones matemáticas.

Así como el r mide el grado de relación o correlación que existe entre X e Y; el coeficiente de determinación ($r^2$), es el más utilizado, e indica qué tan correcto es el valor estimado de la ecuación de regresión. Mientras más alto sea $r^2$ más confianza se podrá tener en el valor estimado de la línea de regresión. Concretamente mide la proporción de la variación total, que se explica por la ecuación de regresión, asumiendo un valor entre 0 y 1. Se calcula por:

$$r^2 = \frac{[n\sum XY - (\sum X)(\sum Y)].^2}{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}$$

Esta relación lineal la denominamos recta de regresión o ecuación de regresión[2] y muestra la dependencia estadística de las variables. La podemos obtener mediante la siguiente expresión:

$$Y_i = a + bx_i$$

Utilizando el criterio de los mínimos cuadrados, esto es, haciendo mínimas las distancias al cuadrado entre los valores de la nube de puntos –valores observados- y los valores correspondientes a la ecuación de regresión, los coeficientes a y b se obtienen de la siguiente manera:



[2] El modelo de regresión se basa en tres supuestos básicos, los cuales si no se cumplen invalidan cualquier proyección: 1) los errores de la regresión tienen una distribución normal, con media = 0 y varianza constante.2) los errores no están correlacionados entre ellos (existe auto-correlación) y 3) todas las variables analizadas se comportan en forma de línea o son susceptibles de linealizarse.

$$a = \frac{\sum \square X.^2 \sum \square Y - \sum \square X \sum \square XY}{\left[n \sum \square X^2 - (\sum \square X)^2\right]}$$

$$b = \frac{n \sum \square XY - \sum \square X \sum \square Y}{\left[n \sum \square X^2 - (\sum \square X)^2\right]}$$

También $b$ puede calcularse de la siguiente manera:

$$b = \frac{\sum \square XY - \sum \square X \sum \square Y}{\left[\sum \square X^2 - nX^2\right]}$$

Gráficamente se aprecia la recta de regresión como:


Figura 12-2. Ecuación y gráfico de recta de regresión

Para medir el grado de relación lineal o grado de correlación entre las variables $X$ e $Y$, o lo que es lo mismo, la bondad de la regresión lineal llevada a cabo, se utiliza el coeficiente de determinación lineal $R^2$ y está dado por la siguiente relación:

$$R^2 = \frac{\sum (\hat{Y} - \overline{Y})^2}{\sum (Y_o - \overline{Y})^2}$$

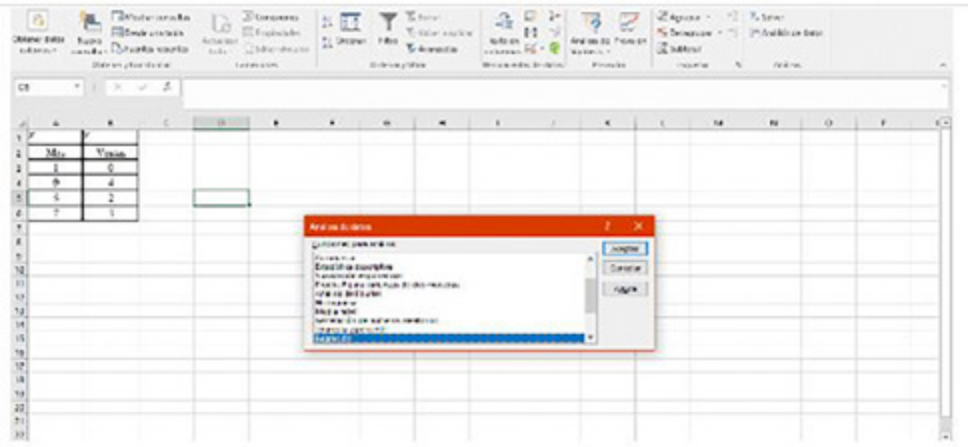
Veamos la aplicación de estos conceptos con los datos de la tabla 12-2:

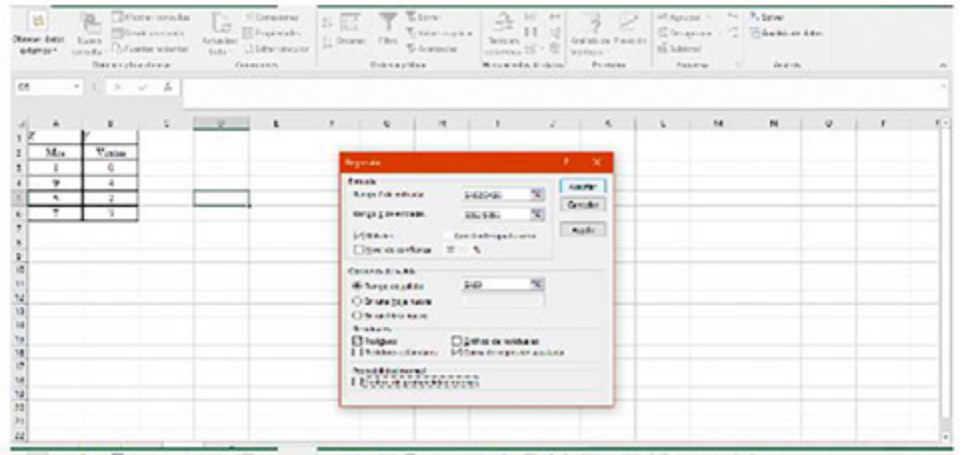
## Tabla 12-2.

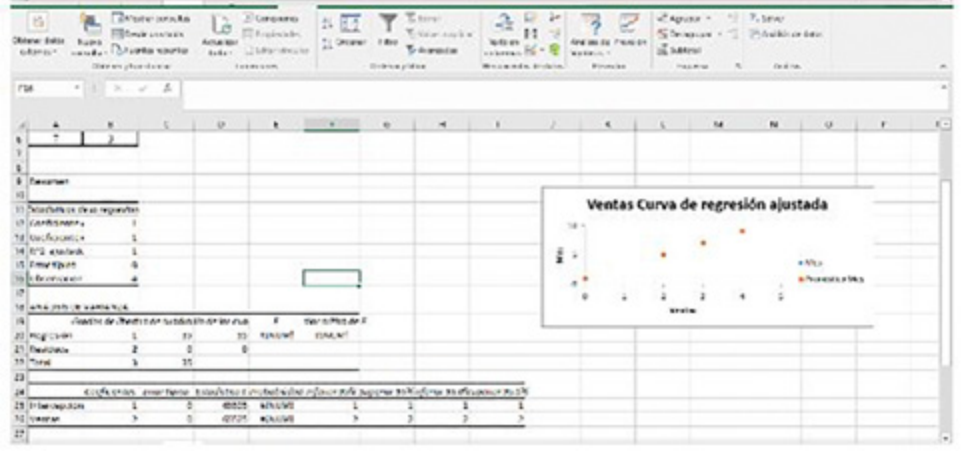
Datos de ventas por mes

| X | Y |
|---|---|
| Mes | Ventas |
| 1 | 0 |
| 9 | 4 |
| 5 | 2 |
| 7 | 3 |

## Ejemplo en Microsoft Excel







# 13. Introducción al cálculo de probabilidades

Debemos anticiparnos en aclarar que los conceptos que aquí discutamos, no constituyen un estudio del cálculo de probabilidades pormenorizado. En realidad, estudiaremos los principales conceptos y aplicaciones de estas. Hoy en día en el mundo de los negocios, de la medicina, etc., la teoría de la probabilidad reviste un lugar importante; por ejemplo: en el mundo de los seguros, la estimación de productos defectuosos y la llegada de pasajeros, entre otros.

## 13.1 Principales conceptos:

- Probabilidad (*pi*): es la posibilidad numérica de la ocurrencia de un evento. La representamos por *pi* y es una medida numérica comprendida entre 0 y 1, es decir, $0< p_i<1$.

- Experimento: es toda acción bien definida que lleva a un resultado único bien definido; por ejemplo: lanzar un dado. El experimento será *aleatorio* cuando al repetirse en las mismas condiciones, no da lugar al mismo resultado



- Espacio muestral (Ω): es el conjunto de resultados posibles de un experimento aleatorio. Cada resultado ω es un *punto muestral*. Por ejemplo: el espacio muestral de lanzar un dado está dado por:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Cada resultado de un número es el punto muestral entonces ω = 1,2....6.

**Un suceso A:** es un subconjunto del espacio muestral formado por puntos muestrales. Un suceso elemental consta de un único punto muestral, mientras que un suceso compuesto está formado por más de un punto muestral.

## 13.2 Enfoques de la teoría de la probabilidad

Hay tres enfoques teóricos: a) el modelo de frecuencia relativa; b) el modelo subjetivo; y, c) el modelo clásico (o *a priori*).

a. La probabilidad de un evento bajo *el modelo de frecuencia relativa* está dada por la frecuencia con que ha ocurrido algún evento en el pasado y la probabilidad de que el evento vuelva a ocurrir nuevamente con base en los datos históricos.

$$P_i = \frac{\text{número de veces que ha ocurrido el evento en el pasado}}{\text{número total de observaciones}}$$

**Por ejemplo:** durante los últimos 10 meses, 2 vuelos de la aerolínea TAME salieron con destino a EEUU, retrasados, asumiendo que TAME tiene un viaje por mes, podríamos afirmar que la probabilidad de que el vuelo del próximo mes sea retrasado, es del 20% o de 2/10.

b. El modelo subjetivo se define como la asignación en base a nuestro criterio, es decir, al no tener observaciones de los eventos en el pasado, nos toca signar una probabilidad a un evento que nunca ha ocurrido.

**Por ejemplo:** la probabilidad de que una mujer sea electa presidenta del Ecuador es un evento que nunca ha ocurrido.



c. El modelo clásico en cambio es el que se relaciona con mayor frecuencia en el mundo de la incertidumbre, y por ello la *Pi* se determina por:

$$P_i = \frac{\text{número de formas en las que puede ocurrir un evento}}{\text{número total de posibles resultados}}$$

**Veamos el siguiente ejemplo:**

Las ventas detalladas en la tabla 13-1 hacen referencia a las ventas semanales en la agencia de viajes "El Mundo" ¿Cuál es la probabilidad de que las ventas de esta semana sean: bajas, altas o por lo menos considerables?
Graficamos la distribución de probabilidad:

## Tabla 13-1

Ventas agencia "El mundo"

| Ventas | Frecuencia | FA |
|---|---|---|
| Bajas | 16 | 16 |
| Considerables | 27 | 43 |
| Altas | 9 | 52 |
| Total: | 52 | |

$P_i$ (considerables) = $\frac{27}{52}$ = 51.92%.

$P_i$ (bajas) = $\frac{16}{52}$ = 30.77%.

$P_i$ (Altas) = $\frac{9}{52}$ = 17.31%.

$P_i$ (por lo menos considerables) = $\frac{(52-16)}{52} = \frac{(36)}{52}$ = 69. 23%.

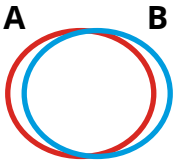


Dado que los sucesos son en realidad subconjuntos del espacio muestral, que están formados por los resultados de experimentos aleatorios, las operaciones (complementariedad, unión, intersección, diferencia y diferencia simétrica) y las relaciones (inclusión, igualdad e incompatibilidad) entre conjuntos son igualmente válidas para sucesos.

Así, las diferentes operaciones entre sucesos conducen a las siguientes definiciones:

- **Suceso complementario** de un suceso A:

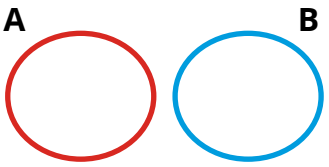
$\bar{A} = \{\omega \in \Omega / \omega \notin A\}$




Implica que si se da A también se da B

- **Suceso único** de los sucesos A y B:

$A \cup B = \{\omega \in \Omega / \omega \in A \text{ o bien } \omega \in B\}$

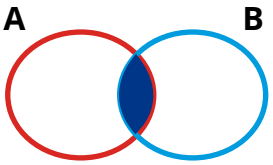



Si de da A no se da B

- **Suceso intersección** de los sucesos A y B:

$A \cap B = \{\omega \in \Omega / \omega \in A \text{ y } \omega \in B\}$

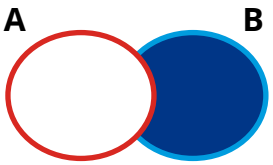





Formado por la parte común de los sucesos A y B

- **Suceso diferencia** de los sucesos A y B:

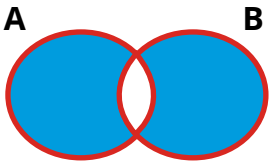
$A - B = \{\omega \in \Omega / \omega \in A \text{ y } \omega \in B\}$




Formado por los elementos de A pero no de B

- **Suceso diferencia simétrica** de los sucesos A y B:

$A \Delta B = \{\omega \in \Omega / (\omega \in A \text{ o bien } \omega \in B) \text{ y } \omega \in A \cap B\}$




Formado por los elementos que son exclusivos del suceso A y del suceso B

De igual manera se tienen las siguientes relaciones entre sucesos:

- El suceso A *está contenido* en el suceso B, si cualquier punto muestral que pertenece al suceso A, también pertenece al suceso B:



$A \subset B \text{ si } \omega \in A \rightarrow \omega \in B$

- Los sucesos A y B *son iguales*, si cualquier punto muestral de A está en B y viceversa:

$A = B \text{ si } \omega \in A \rightarrow \omega \in B$

- Los sucesos A y B son *incompatibles, disjuntos o mutuamente excluyentes*, si no tienen puntos muestrales en común, por lo tanto:

$A \cap B = \Phi$

**La probabilidad condicional** es la probabilidad de que el evento A *ocurra dado que, o a condición de que*, el evento B haya ocurrido. Se calcula de la siguiente manera:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B|A)}{P(B)}$$

## 13.3 Las dos reglas de la probabilidad

1. La regla de la multiplicación consiste en determinar la probabilidad del evento conjunto P(A∩B), es decir, que para encontrar la probabilidad de A y B multiplicamos sus respectivas probabilidades que son: P(A) * P(B). El procedimiento exacto depende de si A y B son dependientes o independientes.

    a. Los eventos A y B son independientes: P(A) = P(A/B), es decir, que la probabilidad de A debe ser igual a la de B, se considere o no ese evento. De igual forma, si A y B son independientes P(B) = P(B/A).

    b. Los eventos A y B son dependientes: cuando la P(B) depende de la condición de que el evento A haya ocurrido. Por lo tanto P(A∩B) = P(A)* P(B A). [3]

2. La regla de la adición consiste en determinar la probabilidad del evento conjunto P(A U B), es decir, que, para encontrar la probabilidad de A o B, cuando los eventos no son mutuamente excluyentes, sumamos sus respectivas probabilidades. Esto es: P(A) + P(B) - P(A ∩ B). En cambio, si los eventos son mutuamente excluyentes, es decir, P(A ∩ B) = 0, entonces la P(A U B) = P(A) + P(B).



[3] Cuando se saca de un conjunto finito de elementos, dos eventos son independientes si y sólo si se realiza el reemplazo. Sin embargo, si el primer elemento no se reemplaza antes de sacar el segundo elemento, los dos eventos son dependientes.

# 14. Diagramas de árbol (arborigramas)

Un diagrama de árbol es una representación gráfica útil para organizar cálculos que abarcan varias etapas. Cada segmento en el árbol es una etapa del problema. Las probabilidades escritas cerca de las ramas son las probabilidades condicionales del experimento.

Para ilustrar veamos el siguiente ejemplo:

**Ejemplo:** Lealtad de los profesores y años de servicio en la Universidad

Tabla 14-1
Profesores y sus años de servicio

| Lealtad | Tiempo de servicio | | | | |
|---|---|---|---|---|---|
| | Más de 1 año | 1 a 5 años | 6 a 10 años | Más de 10 años | Total |
| Se quedaría | 10 | 30 | 5 | 75 | 120 |
| No se quedaría | 25 | 15 | 10 | 30 | 80 |
| Total | | | | | 200 |



**Pasos a seguir:**

- Trazamos un pequeño punto a la izquierda que representa el punto central de un tronco de árbol.
- Para este problema salen dos ramas principales del tronco: la superior “Se quedaría” y la inferior “No se quedaría” sus probabilidades se explican en las ramas en este caso 120/200 y 80/200. Se simboliza P(A) y P(~A).

- Cuatro ramas secundarias se desprenden de cada rama principal y corresponden a los tiempos de servicio. Las probabilidades condicionales para la rama superior del árbol están en las ramas adecuadas. Se trata de las probabilidades: P(B1∩A) ; P(B2∩A); P(B3∩A) y P(B4∩A) donde B se refiere a los tiempos de servicio. Igualmente, en la rama inferior se colocan las probabilidades condicionales P(A1∩B) P(A2∩B) P(A3∩B) y P(A4∩B).
- Por último, las probabilidades conjuntas de que A y B ocurran al mismo tiempo se muestran al lado derecho, como sigue: P(A y B1) = P(A) x P(B1 ∩ A).

Veamos el gráfico:

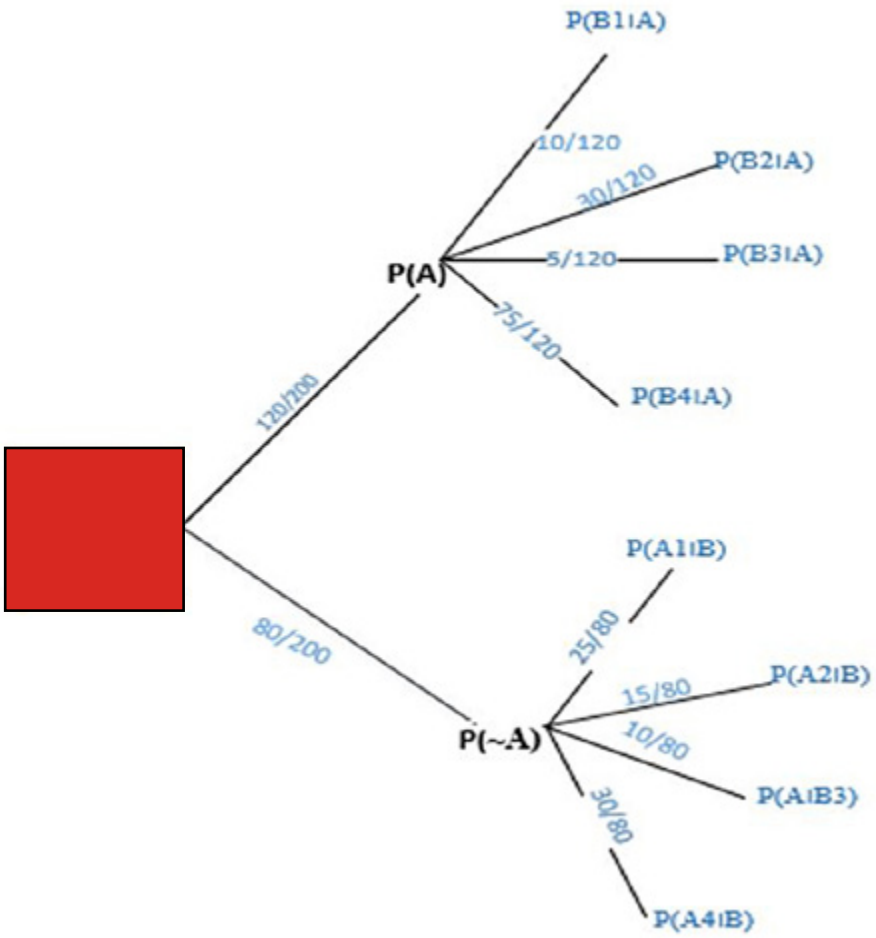



Figura 14-13-1. Diagrama de Árbol

# 15. Teorema de Bayes[4]

Este teorema parte del análisis del *diagrama de árbol*, por lo tanto, considera tanto la probabilidad condicionada como la adición de probabilidades.

El teorema queda expresado en:

$$p\left(\frac{A_i}{B}\right) = \frac{p(A_i).p\left(\frac{B}{A_i}\right)}{p(B)}$$

P(A∩D) = P(A ∩ D) / P(A ∩ D ) + P(B ∩ D ) =

= P(A) x P(D∩A) / P(A) x P(D∩A) + P(B) x P(D ∩ B)



[4] Reverendo Thomas Bayes (1702-1761)

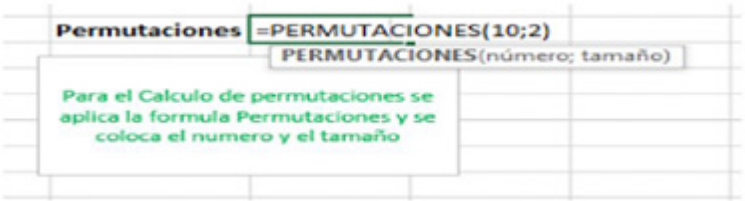
# 16. Técnicas de conteo

También llamado análisis combinatorio, es el análisis de las permutaciones y combinaciones que surgen para cuantificar los puntos muestrales en un espacio estudiado. Es pertinente hacer la distinción que si hablamos de permutaciones el orden de los factores u objetivos es de suma importancia, al contrario de las combinaciones donde el orden no presenta importancia alguna, Por ello dado un conjunto de n elementos el número de permutaciones, cada uno de tamaño r, se determina como:

- El número de permutaciones de n elementos tomados r a la vez (nPr), es igual a n factorial sobre (n-r) factorial, donde factorial (!) significa el producto para todos los números de 1 a n. En la ecuación siguiente se expresa la permutación:

$$nPr = \frac{n!}{(n-r)!}$$

## Ejemplo en Microsoft Excel






- El número de combinaciones de n elementos tomados r a la vez (nCr), es igual n factorial sobre r factorial por (n-r) factorial, donde factorial (!) significa el producto para todos los números de 1 a n[5]. En la ecuación siguiente se expresa la combinación:

$$nPr = \frac{n!}{r!\,(n-r)!}$$

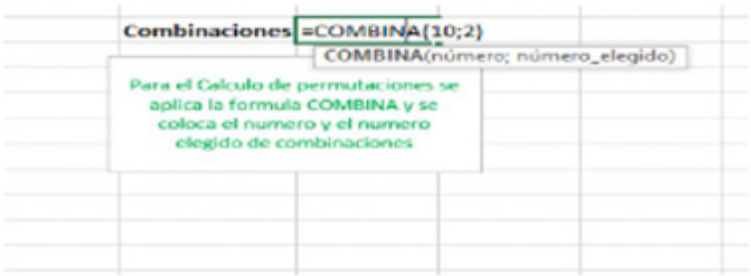
[5] por definición 0! es igual a 1

Las combinaciones y las permutaciones no permiten que se seleccione un elemento más de una vez, si se admite la duplicación se utilizará el método de la *escogencia múltiple* de conteo, el número de arreglos de escogencia múltiple de n elementos tomados r a la vez es:

$$nMr = n^r$$

Igualmente, cuando debemos escoger un elemento de dos o más conjuntos, es adecuado el proceso de *multiplicación*. Este principio requiere que simplemente se multiplique el número de elementos en cada conjunto.

## Ejemplo en Microsoft Excel




## Ejercicios de aplicación



**16.1** Dell Publishing tiene 75 títulos de libros clasificados por tipo y costo, de la siguiente manera:

Tabla 16-1.
Ejercicio

| Tipo | Costo | | | |
|---|---|---|---|---|
| | $10 | $15 | $20 | Total |
| Ficción | 10 | 8 | 3 | 21 |
| Biografía | 12 | 10 | 9 | 31 |
| Histórico | 4 | 17 | 2 | 23 |
| Total | 26 | 35 | 14 | 75 |

Halle la probabilidad de que un libro seleccionado aleatoriamente sea:

- Ficción o cueste $10.
- Histórico y cueste $20.
- Histórico y cueste $10 o $15.
- Biografía o cueste más de $10

P (Ficción o cueste $. 10) = 21/75 + 26/75 - 10/75 = 37/75

P (histórico y cueste $. 20) = 2/75

P (histórico y cueste o 10 o 15) = 23/75 - 2/75 = 21/75

P (Biografía o cueste más de 10) = 12/75 + 49/75 = 61/75

**16.2** Un guía turístico sabe, por experiencias anteriores, que la probabilidad que un turista compre paquetes turísticos es del 65%. La probabilidad de que el turista compre un ticket aéreo si ya tiene reservado el paquete turístico es del 35%.

a. ¿Cuál es la probabilidad de que el turista posea ambas cosas?
b. ¿Son los tickets aéreos y las reservas de hotel independientes?

a. P(A) = Paquete turístico

P(B) = Ticket aéreo

P(A ∩ B) = P(A) * P(B) = 0.65 * 0.35 = 0.2275

b. A y B no son independientes.



**16.3** De 1000 estudiantes de 18 años, 600 tienen empleo y 800 son bachilleres. De los 800 bachilleres, 500 tienen trabajo. ¿Cuál es la probabilidad de que un joven de 18 años tomado aleatoriamente sea:

a. Un bachiller empleado.
b. Empleado, pero no bachiller.

c. desempleado o un bachiller.
d. desempleado o no bachiller.

a. P(bachiller empleado) = 500/1000 = 50%

b. P(empleado, pero no bachiller) = 600/1000 - 500/1000 = 100/1000 = 10%

c. P(desempleado o un bachiller) = 400/1000 + 500/1000 = 90%

d. P(desempleado o no bachiller) = 400/1000 + 500/1000 - 200/1000 = 700/1000 = 70%

**16.4** En una agencia de viajes se planea contratar tres nuevos empleados. Había ocho candidatos para los cargos, seis de los cuales eran hombres. Los tres que consiguieron el puesto eran de sexo masculino. Un cargo por discriminación de sexo se impuso contra la agencia. ¿Cómo decidiría usted?

$$\text{P(todos 3 hombres)} = \frac{\text{número de formas de que los 3 sean hombres}}{\text{número total de posibles resultados}}$$

El número de formas en las cuales 3 de los 6 hombres y ninguna de las 2 mujeres pueda ser seleccionada es: $6C3 \times 2C0 = 20*1 = 20$.

El número total de formas en las que 3 de todos los 8 candidatos pueden ser contratados es: 8C3 =56

Entonces:

$$\text{P(todos 3 hombres)} = \frac{20}{56}$$

# 17. Distribuciones de probabilidades

En la sección anterior se estudió los conceptos y las reglas básicas de probabilidades, en este apartado vamos a estudiar más detenidamente la variable aleatoria y para ello usaremos las leyes de probabilidad.

**Principales conceptos:**

a. Una variable aleatoria (variable estocástica): es una variable cuyo valor es el resultado de un evento aleatorio, suele denotarse con letras mayúsculas como X o Y.

Las variables aleatorias son de dos tipos: discretas o continuas.

- Una variable aleatoria es discreta si puede asumir sólo ciertos valores, con frecuencia números enteros, y resulta del conteo.

- Una variable aleatoria es continua si resulta principalmente de la medición y puede tomar cualquier valor, al menos dentro de un rango de datos dado.



b. Una distribución de probabilidad es una tabla que muestra todos los posibles resultados de un experimento junto con sus respectivas probabilidades. Vale recordar que la suma de las probabilidades es igual a 1 o al 100%.

**Media y varianza de las distribuciones discretas**

- El Valor esperado de una distribución de probabilidad discreta es una media ponderada que resulta de la suma del producto de las probabilidades y los resultados correspondientes. Es decir:

$$E(X) = \sum [(X_i)\{P(X_i)\}]$$

$$E(X) = \sum [(X_i)(P_i)]$$

Donde cada $X_i$ es el resultado correspondiente del experimento y $P_i$ es la probabilidad de ese resultado.

- La varianza de una distribución de probabilidad discreta es el promedio de las desviaciones al cuadrado con respecto a la media, y se escribe de la siguiente manera:

$$\sigma^2 = \sum [(X_i - \mu)^2 * \{P(X_i)\}]$$

- La desviación estándar seguirá siendo la raíz cuadrada de la varianza, es decir



$$\sigma = \sqrt{\sigma^2}$$

**Veamos algunos ejemplos:**

17.1 El número de quejas de los huéspedes del Hotel Sheraton oscila entre 0 y 6 cada día, como se muestra en la siguiente tabla. Calcule e interprete el valor esperado, la varianza y la desviación estándar.

## Tabla 17-1.

Ejercicio

| Quejas | Número de días |
|---|---|
| 0 | 3 |
| 1 | 4 |
| 2 | 3 |
| 3 | 6 |
| 4 | 2 |
| 5 | 1 |
| 6 | 4 |

Calculamos la probabilidad de ocurrencia de las quejas según el registro de los días

## Tabla 17-2.

Cálculos de media y varianza Ejemplo 17.1

| Quejas ($X_i$) | Número de días | $P_i$ | $X_iP_i$ | $[X_i - E(X)]^2 * P_i$ |
|---|---|---|---|---|
| 0 | 3 | $\frac{3}{23}$ | 0 | $\left[0 - \frac{65}{23}\right]^2 * \frac{3}{23}$ |
| 1 | 4 | $\frac{4}{23}$ | $\frac{4}{23}$ | $\left[1 - \frac{65}{23}\right]^2 * \frac{4}{23}$ |
| 2 | 3 | $\frac{3}{23}$ | $\frac{6}{23}$ | $\left[2 - \frac{65}{23}\right]^2 * \frac{3}{23}$ |
| 3 | 6 | $\frac{6}{23}$ | $\frac{18}{23}$ | $\left[3 - \frac{65}{23}\right]^2 * \frac{6}{23}$ |
| 4 | 2 | $\frac{2}{23}$ | $\frac{8}{23}$ | $\left[4 - \frac{65}{23}\right]^2 * \frac{8}{23}$ |
| 5 | 1 | $\frac{1}{23}$ | $\frac{5}{23}$ | $\left[5 - \frac{65}{23}\right]^2 * \frac{5}{23}$ |
| 6 | 4 | $\frac{4}{23}$ | $\frac{24}{23}$ | $\left[6 - \frac{65}{23}\right]^2 * \frac{24}{23}$ |
| **TOTAL** | 23 | 1 | $\frac{65}{23}$ | |

E(X) =65/23 = 2.82
$\sigma^2$ = 3.76
σ = 1.95

## 17.1 Distribución probabilística binomial

Se dice que una distribución es binomial cuando la probabilidad de un experimento es conocida, constante, y además, esta se repite varias veces. Siguiendo el proceso conocido como Bernoulli[6], una distribución normal presenta las siguientes propiedades:

- Solo debe haber dos resultados posibles. Uno identificado como probabilidad de éxito, π , y el otro como probabilidad de fracaso,1-π
- La probabilidad de éxito sigue siendo constante de un experimento al otro, al igual que lo hace con la probabilidad de fracaso.
- La probabilidad de éxito en un experimento es totalmente independiente de cualquier otro experimento.
- El experimento puede repetirse muchas veces

Si se conoce la probabilidad de que un experimento determinado sea exitoso, será posible estimar cuántos éxitos habrá en un número dado de experimentos. Este cálculo se lo puede obtener usando la fórmula binomial dada por la siguiente ecuación:



$$P(x)= \frac{n!}{x!(n-x)!\ \pi(1-\pi)^{n-x}}$$

O también puede calcularse como:

$$P(x) = {}_nC_x\ (\pi)^x(1-\pi)^{n-x}$$

Los resultados de P(x) para diferentes valores de π,n y x están dados en las tablas estadísticas que se acompañan normalmente en los apéndices de dichos libros.

[6] Jacobo Bernoulli (1654-1705) fue un matemático suizo que lo descubrió

**Ejemplo:**

17.1.1 ¿Cuál es la probabilidad de que, de 20 pasajes aéreos seleccionados de manera aleatoria, cinco de los tickets no sean pagados? La aerolínea ha observado que el 10% de los pasajeros no paga el monto completo del ticket, durante un mes dado. Esto puede expresarse como:

$$\pi = 10\%;\ n = 20 \text{ y } x = 5, \text{es igual a } 0.0319 \text{ o } 3.19\%.$$

$$P(x) = {}_{20}C_5\ (10\%)^5\ (1-\ 10\%)^{20-5}$$

$$P(x) = 3.19\%$$

Si existe un 10% de probabilidad de que no se pague un pasaje en su totalidad; entonces, existe un 3.19% de probabilidad de que exactamente 5 de los 20 pasajes seleccionados de manera aleatoria tengan un pago completo.

a. Media y varianza de las distribuciones binomiales

Si sólo hay dos posibles resultados como en la distribución binomial, la media y la varianza pueden estimarse de la siguiente manera:

$$E(X) = \mu = n\ \pi$$

$$\sigma^2\ = n\ \pi(1-\pi)$$

## 17.2 Distribución hipergeométrica



Este tipo de distribución tiene lugar cuando la probabilidad de éxito no es constante, es decir, cuando la población es pequeña y ocurre el muestreo sin reemplazo; aquí la probabilidad de éxito variará.

La función de probabilidad para una distribución hipergeométrica es:

$$P(x) = \frac{rCx\ \ N\text{-}rCn\text{-}x}{NCn}$$

En donde:

N = tamaño de la población
r = número de éxitos en la población
n = es el tamaño de la muestra
x = número de éxitos en la muestra

Si se selecciona una muestra sin reemplazo de una población finita, conocida, y contiene una proporción relativamente grande de la población, se debe utilizar *la distribución hipergeométrica*.

**Ejemplo**

17.2.1. Se puede ilustrar de mejor manera con los datos de los quince profesores del Programa de Turismo y Gastronomía, se seleccionan doce para ser enviados al Japón a estudiar un nuevo concepto de turismo; ocho de los profesores ya tienen algo de entrenamiento en el concepto. ¿Cuál es la probabilidad de que cinco de los enviados tengan algo de conocimiento sobre el concepto antes de partir a ese país?

N = 15
n = 12
r = 8
x = 5

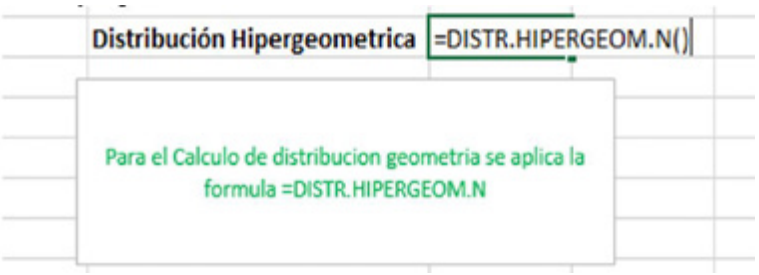
$$P(X) = \frac{rCx\ \ N - rCn - x}{NCn}$$

$$P(5) = \frac{8C5\ \ 15 - 8C12 - 5}{15C12}$$

$$P(5) = \frac{{}_8C_5\ {}_7C_7}{{}_{15}C_{12}}$$

$$P(5) = 0.1231\ o\ 12.31\%$$

### Ejemplo en Microsoft Excel




## 17.3 Distribución probabilística de Poisson

La distribución binomial trabaja con probabilidades de éxito relativamente grandes, al igual que tamaños de muestra pequeñas. Cuando la probabilidad de éxito ($\pi$) es muy pequeña y el tamaño de la muestra ($n$) es grande, se denomina *distribución probabilística de Poisson*. Generalmente se la conoce como la ley de eventos improbables[7], lo cual significa que la probabilidad de éxito de que suceda un evento específico es muy pequeña. La distribución Poisson es del tipo probabilístico discreto, porque se construye contando datos.

Esta distribución se usa mucho en los servicios, como, por ejemplo: el número de clientes en espera del servicio en un restaurante, o los que aguardan a entrar en un centro de recreación, etc.



Matemáticamente podemos escribir esta distribución de la siguiente forma:

$$P(x) = \frac{\mu^x \ e^{-\mu}}{X!}$$

[7] Existe elaborada una tabla de valores para las probabilidades con diferentes $\lambda(\pi)$

Donde:

$\mu$ (mu) = la media aritmética del número de ocurrencias (de éxitos) en un intervalo de tiempo específico.

e = base del logaritmo neperiano y es una constante igual al número 2.71828.

x = número de ocurrencias (o éxitos).

P(x) = es la probabilidad que se va a calcular para un valor dado de x.

**a. Media y varianza de las distribuciones de Poisson**

- El promedio o media de éxitos en la distribución de Poisson puede determinarse por medio del producto entre el número de observaciones (n) y la probabilidad de éxito ($\pi$), matemáticamente está dado por:

$$\mu = n\,\pi$$

- La varianza de una distribución de Poisson es igual a $n\,\pi$

**Ejemplo:**

17.3.1 Una aerolínea tiene problemas en sus viajes con el equipaje. Una muestra aleatoria de 5.000 equipajes reveló estos datos: muchas de ellas no contenían armas cortopunzantes, otras tenían solo una; algunas cuantas tenían dos y así sucesivamente. La distribución del número de armas corto punzantes se aproxima a la distribución de Poisson. El agente contó 3.500 armas en los 5.000 equipajes. ¿Cuál es la probabilidad de que un equipaje seleccionado al azar no contenga armas?



$$x = 0; n = 5.000; \mu = \frac{3500}{5000} = 0.70$$

$$P(X) = \frac{\mu^x \; e^{-\mu}}{X!}$$

$$P(0) = \frac{0.70^0 \; 2.71828^{-0}}{0!}$$

$$P(0) = 0.4966$$

## 17.4 La distribución exponencial

La distribución de Poisson es una distribución discreta, que mide el número de ocurrencias sobre algún intervalo de tiempo o espacio. En cambio, la distribución exponencial es una distribución continua que mide el paso del tiempo en tales ocurrencias.
Por ejemplo: en una distribución de Poisson se estudia el número de clientes que llegan a recibir algún servicio (la cola de clientes en un banco); mientras que en una distribución exponencial se estudia el tiempo entre la atención de un cliente y otro. La probabilidad de que el tiempo de atención sea menor que o igual a cierta cantidad x está dado por:

$$P(X < x) = 1 - e^{-ut}$$

En donde:
*t* = es el lapso
e = es la base del logaritmo natural 2.71828
μ = es la tasa promedio de ocurrencia.

Por lo tanto, el gráfico de una distribución exponencial de una variable aleatoria continua se muestra en forma descendente.

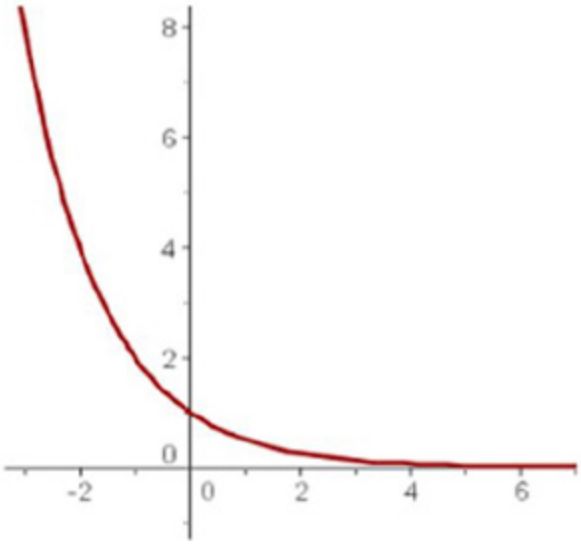



Figura 17.4-1. Distribución exponencial variable continua

**Ejemplo 17.4.1** El Hotel Marriot programa sus taxis para que lleguen al aeropuerto en una distribución de Poisson, con una tasa promedio de llegada de 12 por hora. Usted acaba de aterrizar en el aeropuerto y debe llegar al centro a cerrar un gran negocio. ¿Cuál es la probabilidad de que usted tenga que esperar un máximo de 5 minutos? El gerente es estricto, debido a esta razón no toleraría la falla, de manera que si la probabilidad de que pase otro taxi dentro de 5 minutos es menor al 50%, usted alquilará un carro para el viaje a la oficina

$\mu$ = 12, debido a que 5 minutos de 60 es 1/12, entonces t = 1/12.

$P(X < 5min) = 1 - (2.71828)^{(-12* 1/12)}$

$P(X < 5min) = 63.21\%$

## 17.5 La distribución uniforme

Es una distribución en la cual la probabilidad de todos los resultados posibles es la misma.

La media o el valor esperado de una distribución uniforme está a mitad de camino entre sus dos puntos extremos. Así:

$$E(x) = \mu = \frac{a+b}{2}$$



En donde a es el valor más bajo y b el valor más alto.

El gráfico de una distribución uniforme se muestra en la siguiente figura:

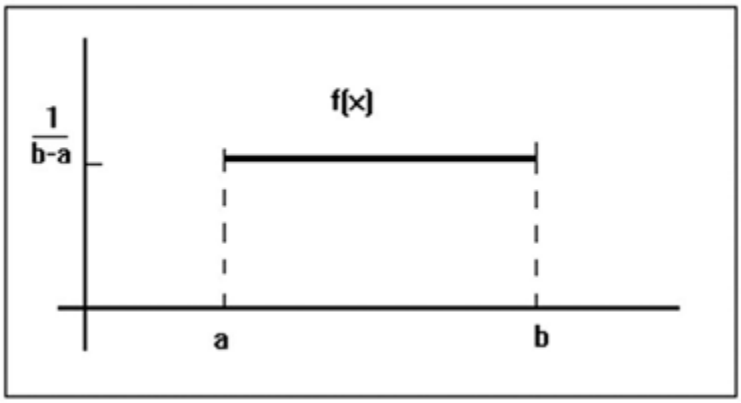



Figura 17.5-1. Distribución Uniforme

La varianza en cambio está dada por:

$$\sigma^2 = \frac{(b-a)^2}{12}$$

Al ser un gráfico en forma de un rectángulo, la probabilidad no es sino el área bajo la curva en este; entonces:

- $\textbf{Altura} = \frac{Área}{Ancho} = \frac{1}{(b-a)}$



En donde b-a es el ancho o rango de la distribución.

- La probabilidad de que una observación caiga entre dos valores está dada por:

$$P(X1>X>X2) = \frac{X_2 - X_2}{\text{Rango}}$$

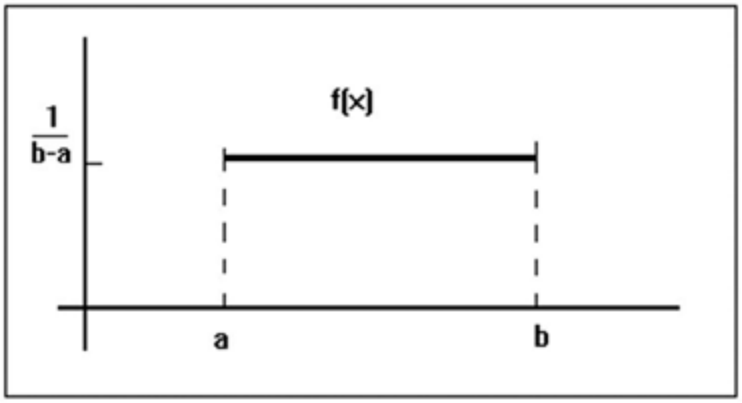
**Ejemplo**

17.5.1 Suponga que los contenidos de los equipajes de los 16Kg permitidos por Continental oscila entre 14.5 y 17.5 Kg y se ajusta a una distribución uniforme. Continental desea saber la probabilidad de que un solo equipaje pese entre 16 y 17.2 Kg.

$$\mu = \frac{a+b}{2} = \frac{14.5+17.5}{2} = \mathbf{16\ Kg}$$

$$\text{Altura} = \frac{1}{(b-a)} = \frac{1}{17.5-14.5} = \frac{1}{3}$$

La distribución es como sigue:






$$P(X1 > X > X2) = \frac{X_2 - X_2}{Rango}$$

$$P(16 > X > 17.2) = \frac{17.2 - 16}{17.5 - 14.5}$$

$$P(161 < X < 17.2) = 0.40$$

## 17.6 Distribución probabilística normal

Cuando se mencionan las aplicaciones de la desviación estándar a través de la regla empírica, se refiere a la distribución normal y se identifican algunas características como: la de ser una distribución simétrica, que tiene una forma de campana y que usa con variables continuas; tal como se aprecia en el siguiente gráfico:

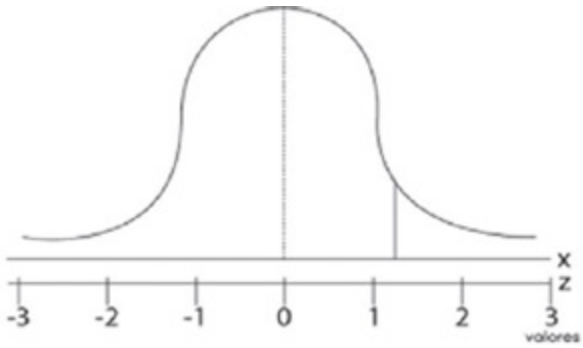



Figura 17.6-1. Distribución Normal

La forma y posición de una distribución normal está determinada por dos factores estadísticos: la media (μ) y la desviación estándar (σ). Se mencionó antes que el área bajo la curva contiene el 100% de las observaciones, es decir que todas caen o están dentro de la curva. Hoy podemos extender este concepto y afirmar que la probabilidad dentro del área de la curva es del 100% o que existe una probabilidad del 100% de que las observaciones ocurran dentro de la curva.

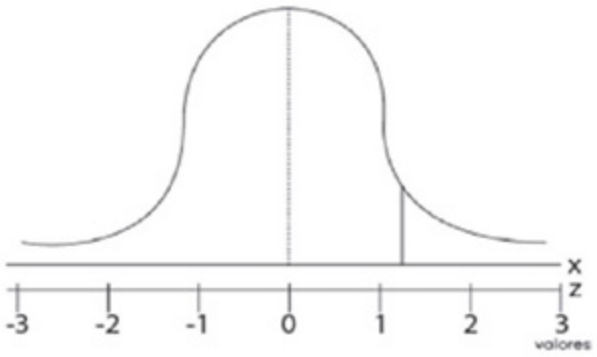


a. Distribución probabilística normal estándar

Puede existir un número infinito de distribuciones normales posibles, cada una con su propia media y su desviación estándar; y al no poder analizar un número tan grande de posibilidades, es necesario convertir todas estas distribuciones normales a una forma estándar. Este proceso se conoce como *estandarizar una distribución normal* y se efectúa con la siguiente fórmula (denominada fórmula Z):

$$Z = \frac{X - \mu}{\sigma}$$

En donde Z, es el número de desviaciones estándar de una observación que está por encima o por debajo de la media, y X es algún valor específico de la variable aleatoria; después de este proceso de conversión la media de la distribución es 0 y la desviación estándar es 1. Por lo tanto, el gráfico de la distribución normal tiene la siguiente interpretación:




Estandarizar una distribución normal permite determinar más fácilmente la probabilidad de ocurrencia de un evento, simplemente encontrando el área dentro de la curva entre el valor observado y el valor de la media. Si se conoce el área se conoce la probabilidad.

El área relacionada con un valor de Z dado, puede encontrarse en la tabla denominada *Área bajo la curva normal con Z desviaciones estándar* y que se muestra en los textos de estadística. Se adjunta a esta nota técnica.

**Ejemplo**



17.6.1 El Ministerio de Turismo en un estudio reciente sobre lugares turísticos ha detectado que el tiempo promedio de estadía de un turista en ciertos lugares del Azuay está distribuido en forma normal, con una media de 2.2 días durante el periodo de vacaciones. Se determinó que la desviación estándar era de 0.8 días. ¿Cuál es la probabilidad de que un turista se hospede más de 3.3 días en una época de vacaciones?

X = 3.3 días
μ= 2.2
σ= 0.8

$$Z = \frac{3.3 - 2.8}{0.8} = 1.38$$

Este valor lo ubicamos en el gráfico y determinamos el área bajo la curva de la siguiente manera:

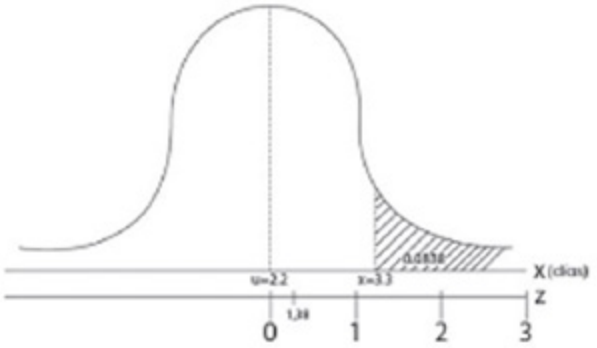



Figura 17.6-2. Ejercicio 17.6.1 Área Z

El área sombreada es la probabilidad buscada, es decir, 0.0838 o 8.38% que es el 0.5 menos el área comprendida entre la media y 1.38 desviaciones estándar que se encuentra en la tabla y que en este caso es de 0.4162.

Ahora se puede calcular un valor X a partir de una probabilidad conocida. En este caso con Z conocida y de la ecuación de la fórmula Z se despeja X, que resulta:



$$X = Z\sigma + \mu$$

B

# ESTADÍSTICA INFERENCIAL

## Introducción

La estadística inferencial comprende un conjunto de métodos que ayudan al investigador a obtener datos de una población, basándose en la información que le proporciona la muestra. Como se expuso anteriormente el tamaño de las poblaciones suele ser una de las dificultades con las que se encuentra el investigador en el desarrollo de su trabajo, por ello es necesario seleccionar una muestra representativa de un tamaño manejable, que le facilite el manejo de datos y proporcione conclusiones extrapolables a la población.

Esta nota técnica comprende el estudio de los siguientes temas: las distribuciones muestrales, la estimación mediante intervalos, la prueba de hipótesis para una y más de dos poblaciones, estadística inferencial, la regresión múltiple, las series de tiempo, los números índices, y finalmente las pruebas no paramétricas.

Se han enunciado anteriormente los métodos de muestreo de los que se puede disponer:

**El muestreo probabilístico:**

a. Muestreo aleatorio simple (MAS)
b. Muestreo aleatorio sistemático (MASIS)
c. Muestreo aleatorio estratificado (MAE)
d. Muestreo por conglomerados

# El muestreo no probabilístico

## a) Determinación del tamaño apropiado de la muestra

El tamaño de la muestra juega un papel importante al determinar la probabilidad de error, así como en la precisión de la estimación.

Para la selección de la muestra es necesario definir un nivel de confianza, además de revisar la varianza de la población y el error tolerable que se está dispuesto a aceptar.

El error que el estadístico está dispuesto a tolerar depende de varios aspectos, por ejemplo: de qué tan crítico es el trabajo, la variabilidad de la población, entre otros.

El cálculo del tamaño de la muestra para estimar la media poblacional (μ) o la proporción poblacional (π), puede determinarse con las siguientes expresiones:

$$n = \frac{Z^2 \sigma^2}{(X - \mu)}$$

*n*= tamaño muestral para intervalos de la media poblacional.

(X – μ) es el error muestral, es decir, la diferencia entre la media de la muestra y la media de la población.



El tamaño muestral para la proporción poblacional está dado por la siguiente expresión:

$$n = \frac{Z^x (\pi)(1 - \pi)}{(p - \pi)}$$

En donde, (p – π) es el error muestral o la diferencia entre la proporción muestral y la proporción poblacional.

Además debemos adelantarnos en definir:

$$Z = \frac{(p - \pi)}{\sigma p}$$

$$\sigma p = \frac{(\pi)(1 - \pi)}{n}$$

## b) Distribuciones muestrales

Una distribución muestral es una lista de todos los valores posibles para un estadístico y la probabilidad relacionada con cada valor. Es decir, la distribución muestral es simplemente una lista de todas las medias muestrales posibles. Estas medias muestrales, al igual que cualquier lista de números, tienen una media denominada media de las medias muestrales o la gran media.

### i. Varianza y desviación estándar de una distribución muestral

La distribución muestral también tiene una varianza y es similar a cualquier otra; de esta manera, es la medida de la dispersión de las medias muestrales alrededor de la gran media. Al sacar la raíz cuadrada obtenemos el error estándar de la distribución muestral, por lo tanto:

$$Varianza\ (\sigma) = \sum \frac{(X_i - \underline{X})^2}{k} = \sum \frac{(X_i - \mu)^2}{k}$$

### ii. Media de una distribución muestral

Esta gran media se calcula, sumando las observaciones individuales (que son las medias muestrales) y el resultado se divide por el número de observaciones (número de muestras). Su cálculo se presenta en la siguiente expresión:

$$\overline{X} = \frac{\sum \square X}{k}$$

En donde, k es el número de muestras y la media muestral siempre es igual a la media de la población (μ).

### iii. Error estándar (σ):

Este error estándar es la misma desviación estándar, por lo tanto, mide la tendencia a sufrir del error de muestreo en el esfuerzo por estimar la media de la población (μ).

Si se llegase a conocer la varianza poblacional y el muestreo fuese con reemplazo o si la muestra se tomara de una población muy grande (virtualmente infinita), una aproximación cercana para calcular la varianza y el error estándar es como sigue:

$$\sigma = \frac{\sigma}{\sqrt{n}}$$



Si el muestreo se realiza con reemplazo y si el tamaño de la muestra es más del 5% de la población, debe aplicarse el factor de corrección para poblaciones finitas, entonces, la expresión apropiada para calcular el error estándar es:

$$\sigma = \frac{\sigma}{\sqrt{n}} \frac{\sqrt{N-n}}{\sqrt{N-1}}$$

## c) Teorema del límite central

A medida que se vuelve más grande, la distribución de las medias muestrales se aproxima a una distribución normal con una media $\mu = X$ y un error estándar $\sigma_x = \frac{\sigma}{\sqrt{n}}$

### i. Uso de la distribución muestral

Como se ha hecho hincapié anteriormente, las decisiones se toman con base en los resultados muestrales, y dado que la distribución muestral estará distribuida normalmente, (la muestra se toma de una población normal y ≥ 30 elementos); el teorema del límite central garantiza la normalidad en el proceso de muestreo. La desviación normal puede utilizarse para ganar información esencial para el proceso de toma de decisiones. Por esta razón el valor de *Z* estaría dado por:

$$Z = \frac{X - \mu}{\sigma_x}$$

**Ejemplo i.1:**



Telcom planea instalar nuevos equipos que mejorarían la eficiencia de sus operaciones, sin embargo, antes que los ejecutivos puedan decidir si dicha inversión será eficaz en función de los costos, deben determinar la probabilidad de que la media de una muestra de n = 35 esté entre 145 y 150, cuando saben que la desviación estándar es de 15 segundos.

$$P\,(145 < X < 150) = \frac{145-150}{\frac{15}{35}} = -1.97 \text{ o un área de } 0.4756$$

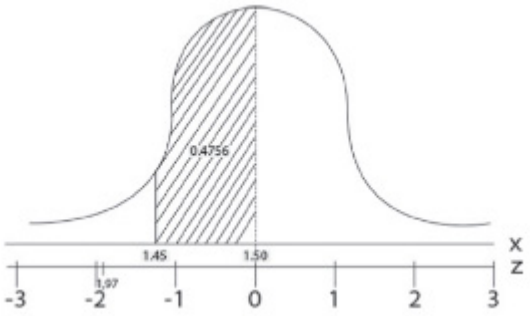



*Figura c-1* Ejemplo 3.1 Telcom

Telcom sí puede tomar la decisión con respecto al nuevo equipo.

## Conclusiones

El proceso inferencial es extremadamente importante en muchos análisis estadísticos. En una posterior nota técnica se estudiarán a profundidad aspectos como la estimación lineal múltiple y la prueba de hipótesis.

La presente nota técnica pretende hacer un análisis de la estadística descriptiva, considerada como la base para el cálculo de los datos estadísticos que se usan en la inferencia estadística, fueron consideradas las medidas de tendencia central y dispersión que constituyen los pilares de la inferencia estadística; sin embargo han sido tomadas en cuenta también otras medidas de descripción de datos que ayudan a interpretar y procesar información básica para la toma de decisiones.



No obstante, nos adelantamos en estudiar la estimación limitándonos a la regresión lineal simple como un paso previo para entender la estimación lineal múltiple.

Finalmente, lo que se intenta con esta nota técnica es facilitar el estudio de la estadística sin que constituya un sustituto del texto base.

# 1. Distribuciones muestrales

En la parte introductoria de esta nota se dijo que las poblaciones generalmente son grandes para estudiarlas y por ello la necesidad de seleccionar una muestra para sacar conclusiones (inferir) acerca de la población; por ello obtener una muestra puede ser útil para utilizarla como un estimador de la población, a través de obtener un valor estadístico que permita inferir el parámetro poblacional. Una *distribución muestral* es una *tabla de valores* que corresponde a la lista de los valores posibles para un estadístico y la probabilidad relacionada con cada valor.

En definitiva, si se obtienen muestras representativas y de ellas se calcula su media aritmética, entonces el uso de este valor estadístico permitirá sacar una conclusión o inferencia sobre el parámetro correspondiente, con una clara relación entre la media aritmética o cualquier dato estadístico y la muestra tomada.

Una población N de la que se extraen un conjunto de muestras $n_1, n_2, \ldots n_j$ y de las que se obtiene su respectiva media $n_1, n_2, \ldots n_j$, genera la posibilidad de que podamos sacar una conclusión, de la media $x$ de esta muestra , que refleje el comportamiento del parámetro $\mu$ de la población $N$. Con la probabilidad de cometer un error denominado "error de muestreo", ya que si $\mu \neq X$ entonces el error muestral existe y es igual a $\mu$-$X$.

**Ejemplo1.1:**

Con una población de $N$= 500 observaciones y se obtienen un conjunto de $n$= 50 muestras, la combinación posible de muestras resulta demasiado grande; ya que $500C50$ posibles es un número también grande. Para simplificar, suponemos que la población está compuesta de 4 elementos y tomamos muestras conformadas de dos elementos. (El número posible de muestras que se pueden extraer de esa población es de 6).



Para ejercicio se suponen los siguientes como elementos de la población: 100, 200, 300 y 400. La media de esta población es de 250.

$$\mu = \frac{100+200+300+400}{4} = 250$$

La tabla siguiente contiene los elementos de cada muestra[1]:

## Tabla 1-1

Ejercicio 1.1 Distribución Muestral

| Muestra ($n$) | Elementos ($X_i$) | Medias muestrales ($x_i$) |
|---|---|---|
| 1 | 100,200 | 150 |
| 2 | 100,300 | 200 |
| 3 | 100,400 | 250 |
| 4 | 200,300 | 250 |
| 5 | 200,400 | 300 |
| 6 | 300,400 | 350 |

La distribución muestral es la siguiente:

| Medias muestrales ($X_i$) | Probabilidad ($P_i$) |
|---|---|
| 150 | 1/6 |
| 200 | 1/6 |
| 250 | 2/6 |
| 300 | 1/6 |
| 350 | 1/6 |



La probabilidad de que la media de una muestra sea igual al parámetro de la población es 2/6 ya que si $\mu$=500 existen dos resultados en las muestras que tiene la misma media, por lo tanto en este caso

$$P\ (\mu = x) = \frac{2}{6}$$

[1] La media muestral sigue la forma de cálculo normal (X=Xk ), por ejemplo: para la primera muestra la media es 150.

Cuando se tiene una distribución negativa (-) como la distribución de frecuencias - es posible calcular tanto la media como su varianza.

La media de las medias muestrales es igual a la media poblacional, en el ejemplo: $x = \frac{\sum x_i}{k}$ en donde k es el número de muestras.

A su vez la varianza de las medias muestrales es:

$$(\sigma) = \sum \frac{(X_i - \overline{X})^2}{k} = \sum \frac{(X_i - \mu)^2}{k}$$

- *La varianza* mide la dispersión de las observaciones individuales (medias muestrales) alrededor de la gran media; y la desviación estándar mide el error estándar, es decir, la tendencia a sufrir del error de muestreo en el esfuerzo para estimar μ.

De lo anterior puede decirse:

$$\sigma_x = \sqrt{\sigma_x^2}$$



- *El error estándar* de la distribución muestral es entonces una medida de la dispersión de las medias muestrales alrededor de μ.

- Cuando la varianza de la población es conocida, el muestreo se hace con reemplazo y es tomado de una población grande; entonces la varianza y el error estándar es como sigue:

$$\sigma_x^2 = \frac{\sigma^2}{\sqrt{n}}$$

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

- Cuando la varianza de la población es conocida, el muestreo se hace sin reemplazo y es tomado de una población con  de la población, entonces el error estándar se calcula de la siguiente manera:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Este último elemento, se denomina *factor de corrección de poblaciones finitas* y es igual a 1 cuando n < 5% de *N*.

Entonces, según sea el tamaño de la muestra podemos llegar a un teorema llamado

- **Teorema del límite central**

Si de todas las muestras de un determinado tamaño que se pueden obtener de una población se obtiene la media muestra, la distribución muestral de estas medias se aproximara a una distribución muestral. Al mejorar con muestras de gran tamaño se puede decir que a medida que n se vuelve más grande, la distribución de las medias muestrales se aproxima a una distribución normal con una media igual a la media poblacional y un error estándar $\sigma_x = \frac{\sigma}{\sqrt{n}}$.

- **Uso de la distribución muestral**



Como hemos insistido las decisiones se toman en base a los resultados muestrales y dado que la distribución muestral estará distribuida normalmente, ya que la muestra se toma de una población normal y la muestra es mayor o igual a 30 elementos, el teorema del límite central garantiza la normalidad en el proceso de muestreo. La desviación normal puede utilizarse para ganar información esencial para el proceso de toma de decisiones, por lo tanto el valor de  estará dado por:

$$Z = \frac{X - \mu}{\sigma_x}$$

En resumen:

**a)** Muchas decisiones se toman con base en los resultados muestrales.

**b)** Las muestras tienen un impacto muy directo sobre las decisiones que se toman.

**c)** Una aplicación muy común y de gran utilidad en una distribución muestral es determinar la probabilidad de que una media muestral clasifique dentro de un rango dado.

**d)** Puesto que la distribución muestral se aproxima a una distribución normal, la desviación normal puede utilizarse para ganar información esencial para el proceso de tomas de decisiones.

Cuando se toman decisiones no solo interesa un valor único, sino se parte de una media de varias observaciones; por lo tanto, en lugar de determinar la probabilidad de un valor único, se puede calcular la probabilidad ($Pi$) de que la media de observaciones se de.

**Veamos el siguiente ejemplo:**

Los tickets aéreos vendidos en Metropolitan Touring tienen una venta promedio de 16.1 ticket, con una desviación estándar de 1.2 tickets. Si se toma una muestra de $n$ = 200, ¿cuál es la probabilidad de que la media sea menor que 16.27?

Datos



$$n = 200;\ \sigma_x = 1.2;\ x = 16.27;\ \mu = 16.1$$

$$Z = \frac{X - \mu}{\sigma_x}$$

$$Z = \frac{16.27 - 16.1}{\frac{1.2}{\sqrt{200}}} = 2.0024$$

*Con este valor nos vamos a la tabla Z y obtenemos el área bajo la curva que es de 0.4772; entonces la probabilidad es de P(x<16.27) = 0.5 + 0.4772 = 0.9772*

# 2. Distribución de proporciones muestrales

Es una distribución que no trata con medias, sino con proporciones  proporción de la población.

La proporción muestral  ($p$) es un estimador de la proporción poblacional (π); entonces el valor esperado de la distribución muestral de las proporciones muestrales será igual a la proporción de éxitos en la población.

De esta manera, los cálculos del valor esperado y el error estándar de las distribuciones muestrales es similar al de las medias muestrales, en definitiva:

El valor esperado (es decir la media) de las proporciones muestrales es igual:

$$E(p) = \frac{\Sigma \square p}{k}$$

En donde, $p$ es la proporción de la muestra y $k$ es el número de muestras.

A su vez el error estándar de las proporciones muestrales  es:



$$\sigma p = \sqrt{\frac{\pi(1-\pi)}{n}}$$ ; o también

Cuando la varianza de la población es conocida, el muestreo se hace sin reemplazo y es tomado de una población con n > 5%; entonces el error estándar se calcula de la siguiente manera:

$$\sigma p = \sqrt{\frac{\pi(1-\pi)}{n}} \sqrt{\frac{N-n}{N-1}}$$

**Ejemplo 2.1:**

El 30% de todos los empleados del Ministerio de Turismo tienen capacitación avanzada. Si en una muestra de 500 empleados menos del 27% estaba preparado de forma adecuada, todos los nuevos contratados necesitarán registrarse en un programa de capacitación. ¿Cuál es la probabilidad de que se inicie el programa?

Datos:

$$n = 500;\ p = 27\%;\ \pi = 30\%$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.3(1-0.3)}{500}} = 0.02$$

$$Z = \frac{p-\pi}{\sigma_p} = \frac{27\% - 30\%}{0.02} = 1.5^2$$

Con este valor nos vamos a la tabla y obtenemos un área de 0.4332; por lo tanto:

$$\mathrm{P}\ (\mathrm{P} < 27\%) = 0.5 - 0.4332 = 0.0668$$

# 3. Estimación con intervalos de confianza

En la sección anterior el propósito de la estadística inferencial fue estimar o inferir alguna conclusión de la población a partir de la muestra, por esa razón, en este apartado explicaremos por lo menos dos tipos de estimadores para este propósito.

**a.** Un *estimador puntual* se utiliza para estimar un valor único de la población o de parámetro poblacional.

**b.** Un estimador por intervalo se utiliza para estimar un rango dentro del cual está el parámetro poblacional desconocido, un intervalo de confianza denotará este rango en el cual puede encontrarse el parámetro, y el nivel de confianza es un coeficiente que mide el nivel de aceptación de que el intervalo contiene el parámetro y normalmente comprende los coeficientes de 90%, 95% y 99%.

Un intervalo de confianza es un rango conformado por un límite inferior de confianza (LIC) y un límite superior de confianza (LSC), dentro del cual se ubica el parámetro. Cuando se trata de estimar μ es posible partir de una desviación estándar poblacional conocida y un dato estadístico (la media muestral); μ se estima de la siguiente manera:

$$\mu = X \pm z\sigma_x$$



Cuando no es posible conocer la desviación estándar de la población se puede estimar el parámetro poblacional a partir de la muestra tal como:

$$\mu = X \pm zs_x$$

Para la cual, la desviación estándar de la muestra se calcula de la siguiente manera:

$$S_x = \frac{s}{\sqrt{n}}$$

Los niveles de confianza nos limitan el intervalo o rango de aceptación de la estimación, lo que significa que lo que esta fuera del intervalo es el error o la probabilidad de error y se denomina con el valor alfa ().

**Ejemplo 3.1:**

Un promotor turístico que intenta construir un gran centro hotelero puede estimar, en la zona donde va a llevar a cabo su proyecto, que el ingreso promedio por familia como indicador de las ventas esperadas. Una muestra de 100 familias da una media de $35.500. Se asume que la desviación estándar poblacional es de $7.200 y acepta un nivel de confianza del 95%, entonces ¿cuál será la media poblacional?

Datos:

n = 100; X = 35.500; σ = 7.200; α = 1-95%,

La media poblacional estimada se ubicará en el intervalo siguiente:

$$\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{7200}{\sqrt{100}}$$

$$\mu = X \pm z\sigma_x$$



$$LIC = x\,(720) = 34.088{,}80$$
$$LSC = x - 1.96\,(720) = 36.911{,}20$$

El promotor del proyecto tiene un 95% de confianza de que la media poblacional real (desconocida) esté entre 34 mil y 37 mil. Es decir, existe una probabilidad de error del 5% de que este intervalo no contenga la media poblacional.

**Ejemplo 3.2 (Caso en que la varianza no sea conocida)**

Rootours planea comprar una flota de nuevos taxis para sus operaciones en Guayaquil. La decisión depende de si el rendimiento del vehiculo en consideración es por lo menos 27.5 Km/gln. Los 36 vehículos que prueba la compañía reportan una media de 25.6 Kilómetros por galón, con una desviación estándar de 3.5 Km/gln. ¿A un nivel de confianza del 99% qué decisión debería tomarse?

Datos:

$$n = 36;\ X = 25.6;\ s = 3.5;\ \alpha = 1\text{-}99\%,$$

La media poblacional estimada se ubicará en el intervalo siguiente:

$$s_x = \frac{s}{\sqrt{n}} = \frac{3.5}{\sqrt{36}} = 0.5833$$

Con un nivel de confianza de 99%, $Z$ es 2.58.

$$\text{LIC} = 25.6 - 2.58\ (0.5833) = 27.11$$

$$\text{LSC} = 25.6 + 2.58\ (0.5833) = 24.10$$



Hasta este momento se ha trabajado con muestras de tamaño y por ello se utiliza la distribución como factor estadístico de estimación; sin embargo, para muestras se utilizará la distribución de probabilidad continua t para realizar la estimación.

# 4. La distribución t student:

Más conocida como distribución t, descubierta por William Gosset, es una familia de distribuciones cada una con su propia varianza. Es una distribución más plana que la distribución normal, con una desviación mayor que uno (σ > 1), también simétrica y con una varianza igual a $\frac{n-1}{n-3}$.

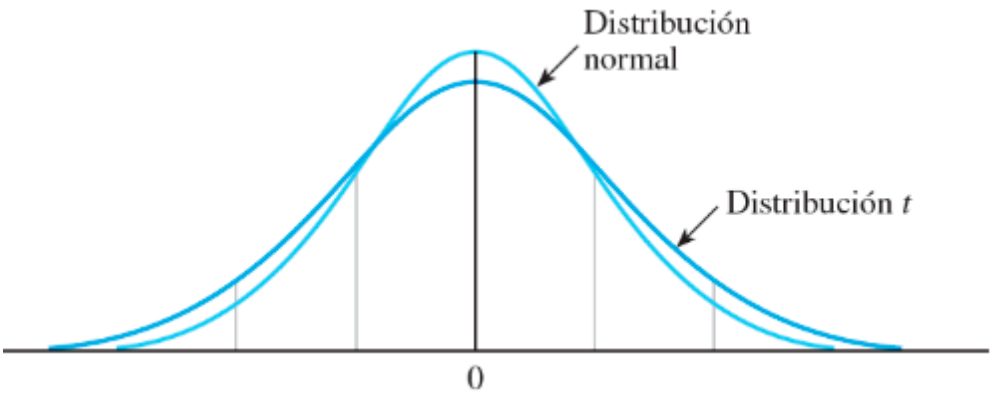



*Figura 4-1.* Distribución t de Student y distribución normal

La varianza de esta distribución depende de los grados de libertad (g.l.). Los grados de libertad son determinados por el número de observaciones.



d.1 Características de la distribución ***t***:

- Sirve para estudiar muestras pequeñas.
- Es continua
- Es simétrica
- Existe un valor t para cada tamaño de muestra
- La población es normal o casi normal.

La estimación a través del factor estadístico se calcula de la siguiente manera:

$$t = \frac{\overline{X} - \mu}{s_x}$$

A su vez el error estándar se calcula:

$$s_x = \frac{s}{\sqrt{n}}$$

El intervalo de confianza para el estimador de la media poblacional es:

$$\mu = X \pm t s_x$$

Cuando se trata de proporciones, este error se obtiene a partir de la siguiente fórmula:

$$s_p = \sqrt{\frac{p(1-p)}{n}}$$

Y para estimar (la proporción poblacional) es posible partir de una desviación estándar muestral y un valor estadístico p (la proporción muestral); se estima de la siguiente manera:

$$\pi = p \pm z s_p$$

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$



Para los cálculos anteriores se ha contado con el tamaño de la muestra como un dato, a continuación se mostrará el método para establecer el tamaño muestral a partir de seleccionar el nivel de confianza con el que se quiere trabajar en la estimación y la variabilidad.

# 5. Cálculo del tamaño de la muestra

Para calcular el tamaño de la muestra se debe tomar en cuanta varias consideraciones:

**a.** Margen de error que tolerará el investigador

**b.** Nivel de confianza deseado

**c.** Variación o dispersión de la población estudiada.

Es decir si:

$$Z = \frac{X-\mu}{\sigma_x} \qquad \text{Y} \qquad \sigma_x = \frac{\sigma}{\sqrt{n}}$$

Entonces:

$$Z = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}}$$

Por lo tanto despejando , el tamaño de la muestra, resulta:



$$n = \frac{z^2\sigma^2}{(X-\mu)^2}$$

Cuando no se dispone de la media poblacional, sino la proporción de la población, el cálculo del tamaño de la muestra estará dado por:

$$n = \frac{z^2(\pi)(1-\pi)}{(p-\pi)^2}$$

En ambos casos si la varianza es desconocida, es posible utilizar la desviación estándar de la muestra, tomando en consideración tres sugerencias: realizar una muestra piloto, utilizar un estudio comparativo, emplear un enfoque basado en el intervalo.

### i. Características de un buen estimador

Para que un estimador sea considerado apropiado y sea utilizado para estimar la población debe cumplir las siguientes características:

**Insesgado:** Significa que la media de la distribución muestral es igual al parámetro poblacional.

**Eficiente:** El estimador eficiente es aquel que tenga la varianza más pequeña.

**Consistente:** Significa que a medida que el tamaño de la muestra aumenta, el valor del factor estadístico se aproxima al parámetro poblacional.

**Suficiente:** Un estimador se dice que es suficiente si ningún otro estimador puede proporcionar más información sobre el parámetro.

# 6. Prueba de hipótesis

Una prueba de hipótesis es un procedimiento estadístico que permite verificar o determinar si existen diferencias estadísticamente significativas entre el valor muestral y el valor del parámetro poblacional, es decir, qué tan significativo es el error del muestreo.

La prueba de hipótesis es un procedimiento basado en evidencia de la muestra que se utiliza como herramienta estadística para la toma de decisiones a partir de probar una afirmación, con el uso de teoría de probabilidad.

### i. Procedimiento para prueba de hipótesis:

**a.** Plantear la hipótesis nula ($H_0$) y la hipótesis alternativa ($H_1$): la $H_0$ es la afirmación o enunciado acerca del valor del parámetro poblacional. La $H_1$ es la afirmación que se aceptará si los datos muestrales proporcionan amplia evidencia de que es falsa.

**b.** Determinar Z o t con los datos muestrales: Dependiendo si la muestra es de tamaño $n \geq 30$ o si $n \leq 30$ respectivamente.

**c.** Definir una regla de decisión: Se definirán los valores críticos y el valor de la prueba para localizar la zona de aceptación o rechazo.

**d.** Obtener la conclusión o interpretación

Al probar una hipótesis es posible que se cometan errores ($\mu \ \ x \ o \ p$) que pueden ser estadísticamente insignificantes, es decir, obedecer a un error de muestreo o ser estadísticamente significativa la diferencia entre el valor muestral y el valor poblacional.

Estos errores normalmente se agrupan o son de dos tipos: el error Tipo I o “α” el  y el error Tipo II o  “β”.

**a.** Error Tipo I significa rechazar la hipótesis nula cuando esta es verdadera

**b.** Error Tipo II significa no rechazar la hipótesis nula cuando esta es falsa.

La probabilidad de cometer un error tipo I es igual al nivel de significancia o es el valor de  en el que se prueba la hipótesis.

### ii. Tipos de pruebas de hipótesis:

**a.** Prueba de dos colas

**b.** Prueba de una cola derecha

**c.** Prueba de una sola cola izquierda.

**Gráficamente:**

- Cola hacia la derecha


*Figura 6-1.* Distribución muestral prueba de una cola a la derecha

- Cola hacia la izquierda


*Figura 6-2.* Distribución muestral prueba de una cola a la izquierda

- Con dos colas

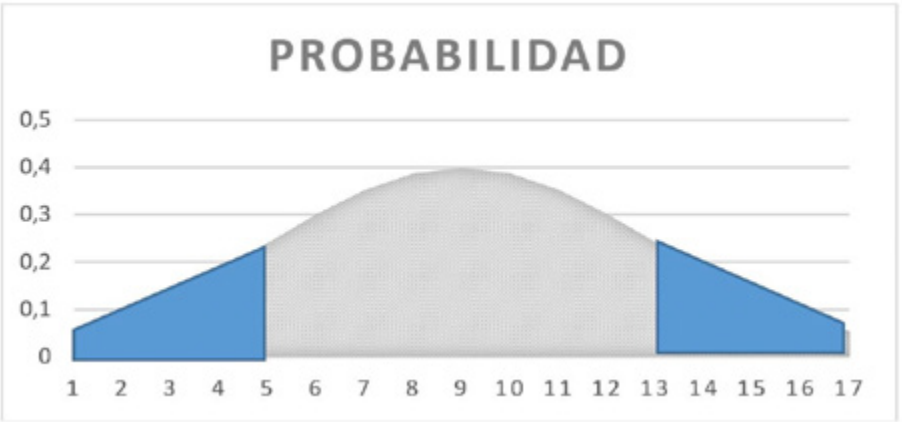





*Figura 6-3.* Distribución muestral prueba de dos colas

Para delimitar la zona de aceptación de la de rechazo, debemos calcular los *valores críticos* de la prueba, que son justamente los valores que se obtienen ya sea con *Z* o *t*, dependiendo del tamaño de la muestra.

La prueba de hipótesis debe medirse seleccionando un nivel de significancia denominado “valor alfa” o simplemente probarse con un  de 1%; 5% o 10%. La selección de este valor  depende del tipo de error que se quiere evitar.

- Es preferible un $\alpha$ más bajo para minimizar la probabilidad de cometer un error tipo I. Es decir, si rechazar una Ho verdadera es más serio que no rechazar una Ho que es falsa, es preferible un $\alpha$ bajo.
- De igual manera si rechazar la Ho que es falsa es más serio que no rechazar una Ho que es verdadera, en este caso es preferible un $\alpha$ alto.

**Ejemplos:**

**Ejemplo 6.2.1:**

La aerolínea AEROGAL afirma que en promedio sus vuelos no se adelantan ni se atrasan durante una semana. Una muestra de 18 vuelos presentó los siguientes adelantos (+) o (-) atrasos en segundos por semana.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| -0.38 | -0.20 | -0.38 | -0.32 | +0.32 | -0.23 | +0.30 | +0.25 | -0.10 |
| -0.37 | -0.61 | -0.48 | -0.47 | -0.64 | -0.04 | -0.20 | -0.68 | +0.05 |



¿Sería razonable llegar a la conclusión de que los adelantos o atrasos medios para los vuelos son cero (0)? Utilice el nivel de significancia 0.05. Calcule el valor p.

1. Plantear la  hipótesis nula ($H_o$) y la hipótesis alternativa ($H_1$).

$$H_o: \mu = 0$$

$$H_1 =$$

2. Determinar $Z$ o con los datos muestrales. Dependiendo si la muestra es de tamaño $n \geq 30$ o si $n \leq 30$ respectivamente. En este caso distribución t:

   Datos:

$$X = -0.2322;\ s = 0.3120;\ n = 18$$

$$t = \frac{X - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{-0.2322 - 0}{\frac{0.3120}{\sqrt{18}}}$$

$t$ = -3.157

3. Definir una regla de decisión. Se definirán los valores críticos y el valor de la prueba para localizar la zona de aceptación o rechazo.

   Si $\alpha$ = 0.01

$$g.l. = n - 1 = 18\text{-}1 = 17 \rightarrow t = 2.110$$



4. Obtener la conclusión o interpretación.

El estadístico cae en la zona de rechazo, eso significa que no se acepta $H_o$. Se rechazan los vuelos de AEROGAL porque no son puntuales.

**Ejemplo 6.2.2:**

En la actualidad, la mayoría de las personas que viajan en avión utilizan boletos electrónicos. Estos evitan a los pasajeros la preocupación de cuidar un boleto en papel, y su manejo es más económico para las líneas aéreas. Sin embargo, en fechas recientes las líneas aéreas han recibido quejas acerca de los boletos electrónicos sobre todo cuando es necesario hacer alguna conexión y cambiar de línea. Para investigar el problema, una agencia de investigación independiente tomó una muestra aleatoria de 20 aeropuertos y recopiló información sobre el número de quejas que tuvieron debido a los boletos electrónicos durante el mes de marzo. La información se reporta a continuación:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 14 | 14 | 16 | 12 | 12 | 14 | 13 | 16 | 15 | 14 |
| 12 | 15 | 15 | 14 | 13 | 13 | 12 | 13 | 10 | 13 |

Con un nivel de significancia 0.05 ¿la agencia de investigación puede llegar a la conclusión de que el número medio de quejas por aeropuerto es menor de 15 al mes?

**a.** ¿Qué suposición es necesaria antes de realizar una prueba de hipótesis?

**b.** Ilustre el número de quejas por aeropuerto en una distribución de la frecuencia o un diagrama de puntos. ¿sería razonable llegar a la conclusión de que la población sigue una distribución normal?

**c.** Realice una prueba de hipótesis e interprete los resultados.



1. Plantear la hipótesis nula ($H_o$) y la hipótesis alternativa ($H_1$)

$$H_o : \mu \geq 15.$$

$$H_1 : \mu < 15$$

2. *Determinar  o  con los datos muestrales*. Dependiendo si la muestra es de tamaño $n \geq 30$ o si $n \leq 30$ respectivamente. En este caso distribución :

$$t = \frac{X - \mu}{\frac{s}{\sqrt{n}}}$$

Datos:

$$X = 13.5;\ s = 1.50\ n = 20$$

$$t = \frac{13.5 - 15}{\frac{1.50}{\sqrt{20}}}$$

$$t = -31.09$$

3. Definir una regla de decisión. Se definirán los valores críticos y el valor de la prueba para localizar la zona de aceptación o rechazo.

$$\text{gl} = 19 = 20 - 1$$

$\alpha = 0.05 \rightarrow t$



4. Obtener la conclusión o interpretación.

Se debe rechazar la $H_o$.

**Ejemplo 6.2.3:**

Karina Dennos es contralora del Hotel Hilton Colon y cree que el problema actual con el flujo de efectivo en el hotel, se debe a la tardanza para cobrar las cuentas por cobrar. Karina cree que más del 60% de las cuentas se tardan en cubrir más de tres meses .Una muestra aleatoria de 200 cuentas reveló que 140 tenían más de tres meses de antigüedad. En el nivel de significancia de 0.01 ¿puede llegar a la conclusión de que más del 60% de las cuentas permanecen sin cobrarse durante tres meses?

1. Plantear la hipótesis nula ($H_0$) y la hipótesis alternativa ($H_1$)

$$H0: \pi \leq 60$$

$$H1: \pi > 0.60$$

2. Determinar $Z$ o $t$ con los datos muestrales. Dependiendo si la muestra es de tamaño $n \geq 30$ o si $n \leq 30$ respectivamente. En este caso distribución $Z$:

Datos:

$$p = 140/200 = 0.70; \; s = 1.50 \; n = 200$$

$$\sigma_p = \sqrt{\frac{0.6(1-0.6)}{200}}$$

$$\sigma_p = \sqrt{\frac{0.6(0.4)}{200}} = 0.035$$

$$Z = \frac{p - \pi}{\sigma_p}$$

$$Z = \frac{0.7 - 0.6}{0.035}$$

$$Z = 2.89$$

3. Definir una regla de decisión. Se definirán los valores críticos y el valor de la prueba para localizar la zona de aceptación o rechazo.

$$\alpha = 0.01 \rightarrow z = 2.58$$

4. Obtener la conclusión o interpretación.

El valor estadístico cae en la zona de rechazo, por lo tanto no se acepta $H_o$. Entonces, se confirma que el 60% de las cuentas tienen más de 3 meses de antigüedad.

# 7. Análisis de varianza

El análisis de varianza, o simplemente la prueba ANOVA, es una prueba estadística para comparar más de dos poblaciones. Está diseñada para probar si dos o más poblaciones tienen las mismas medias.

Esta prueba tiene ciertos supuestos para el cálculo del valor estadístico de prueba -el estadístico F o "prueba F" [2] y la ciencia estadística presenta una tabla para encontrar los valores críticos dado el nivel de significancia y los grados de libertad – y determinar comparando las medias muestrales si estas provienen de poblaciones iguales. Estos supuestos son:

- Todas las poblaciones siguen una distribución normal.
- Todas las poblaciones tienen la misma varianza.
- Las muestras se seleccionan de manera aleatoria.

Bajo estos supuestos se trata de probar las siguientes hipótesis:

$$H0: \ \mu_1 = \mu_2 = ..... = \mu_c$$

$$H1: \ \mu_1 \neq \mu_2 \neq ..... \neq \mu_c$$



En donde c significa el número de tratamientos o muestras.

[2] En esta prueba usamos la denominada distribución F en honor a su descubridor el estadístico Sir Ronald Fisher.

# Prueba ANOVA

La prueba ANOVA se basa en una comparación de la cantidad de variación en cada uno de los tratamientos. Se trata de estimar la variación total que incluye la variación del tratamiento o dentro de la muestra más la variación aleatoria o del error.

Existen dos tipos de prueba ANOVA que los definimos así:

**a)** ANOVA de una vía.

**b)** ANOVA de dos vías.

En la prueba de una vía se compara varias medias de muestras para ver si provienen de la misma población o de poblaciones iguales. En esta prueba existe solamente una variable que influencia en los elementos de la muestra. En cambio, la prueba de dos vías se caracteriza porque los elementos de la muestra son influenciados por más de una variable.

Vale precisar –siguiendo el texto de Webster- que en una prueba ANOVA existen los siguientes conceptos:

- Unidades experimentales, que son los objetos que reciben el tratamiento.
- Tratamiento, que comprende los niveles del factor.
- Factor, que es la variable cuyo impacto se quiere medir.



Cuando hablamos de tratamientos es posible identificar dos modelos de análisis:

1. El modelo de efectos fijos, en el cual se seleccionan tratamientos específicos o los tratamientos se fijan antes del estudio.
2. El modelo de efectos aleatorios, en el cual los niveles o tratamientos se seleccionan aleatoriamente.

Para el cálculo de la prueba F, se requiere hacer el siguiente proceso:

1. Se plantea la hipótesis nula y alternativa que se quiere probar.

2. Se calcula la variación total, es decir:
    - La suma de los cuadrados totales (*SCT*)
    - La suma de los cuadrados de los tratamientos (*SCTR*)
    - La suma de los cuadrados del error (*SCE*)

3. Se encuentra los grados de libertad de los *SCTR* (g.l. del numerador, $c$-1) y de los *SCE* (g.l. del denominador, $n$-$c$), por lo tanto los g.l. totales son igual a la suma de los g.l. de *SCTR* y de *SCE* en símbolos: $n$-1 = $c$-1 + $n$-$c$.

4. Se calcula el valor de la prueba F.

5. Elaboramos el gráfico de la distribución F y a través de la tabla calculamos los valores críticos.

6. Tomamos la decisión de aceptación o rechazo.

Antes de proceder a aplicar este tipo de prueba es importante anotar que entre las principales características de la distribución F, tenemos las siguientes:



1. Existe una familia de curvas porque se determina a través de dos parámetros: los grados de libertad en el numerador y los grados de libertad en el denominador.
2. El valor F no puede ser negativo (< 0) y se trata de una distribución continua.
3. La distribución F tiene sesgo positivo.
4. Sus valores varían de cero a infinito.
5. Es asintótica al eje X.

**Ejemplo 7.1:**

El Vicepresidente de Investigación de Mercados de la agencia METROPOLITAN TOURING estudia paquetes promocionales para atraer nuevos clientes, paquetes que incluyen algunos juegos y premios en cuatro sucursales del país. Está convencido de que diferentes tipos de premios atraerían a diferentes grupos de ingresos. Las personas de un cierto tipo de ingreso prefieren los regalos, mientras que las de otro grupo de ingresos prefieren o se sienten más atraídos por viajes gratuitos a sitios favoritos para pasar vacaciones. El Vicepresidente decide utilizar el monto de las compras como una medida representativa del ingreso. Él desea determinar si existe una diferencia en el nivel promedio de compras entre las cuatro sucursales del país. Si se halla alguna diferencia el Vicepresidente ofrecerá una diversidad de premios o paquetes promocionales.

A continuación se presenta el nivel de compras en miles de dólares en las cuatro sucursales de la agencia Metropolitan Touring:

| Compra (r) | Sucursal 1 | Sucursal 2 | Sucursal 3 | Sucursal 4 |
|---|---|---|---|---|
| 1 | 5.1 | 1.9 | 3.6 | 1.3 |
| 2 | 4.9 | 1.9 | 4.2 | 1.5 |
| 3 | 5.6 | 2.1 | 4.5 | 0.9 |
| 4 | 4.8 | 2.4 | 4.8 | 1.0 |
| 5 | 3.8 | 2.1 | 3.9 | 1.9 |
| 6 | 5.1 | 3.1 | 4.1 | 1.5 |
| 7 | 4.8 | 2.5 | 5.1 | 2.1 |
| Media de cada tratamiento ($x_j$) | 4.87 | 2.29 | 4.31 | 1.46 |



Solución:

- En este caso tenemos cuatro tratamientos (c=4); es decir, cada sucursal es una muestra o un tratamiento que deberá estudiarse.
- El número de observaciones o unidades experimentales ($r_j$) en cada tratamiento es de 7 ($r_j$=7).

- El número total ($n$) de observaciones es igual a 28 ($n$ = r*c)
- El nivel de factor, en este caso la variable que se estudia, es el nivel de compras como medida del grupo de ingreso.
- Con estos datos procedemos según el procedimiento -ya estudiado- de la prueba de hipótesis:

1. Plantear la hipótesis nula ($H_o$) y la hipótesis alternativa ($H_1$)

$$Ho: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Hipótesis Nula: todas las medias son iguales

$$H1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$$

Hipótesis alternativa: no toda*s* las medias son iguales

2. Determinar F con los datos muestrales:

En este ejemplo:

$$c = 4; n = 28; r = 7$$

$$x = \frac{(5.1 + 4.9 + 5.6 + \ldots\ldots\ldots + 2.1)}{28} = \frac{90.44}{28} = 3.23$$

$$SCT = \sum_{\square}^{\square} \square (x_{ij} - x)^2 = (5,1 - 3,23)^2 + \cdots \ldots . + (2,1 - 3,23)^2 = 61,00$$

$$SCTR = \sum_{r}^{j} \square (\underline{x} - \bar{\bar{x}})^2 = 7(4,87 - 3,23)^2 + \cdots \ldots . + 7(1,46 - 3,23)^2 = 55.33$$

$$SCE = = (5.1 - 4.87) + \ldots\ldots\ldots + (2.1 - 1.46) = 5.67$$

$CMTR$ = = 55.33/3 = 18.44

$CME$ = = 5.67/24 = 0.236

$F$ = = 18.44/0.236 = 78.14

3. Definir una regla de decisión. Se definirán los valores críticos y el valor de la prueba para localizar la zona de aceptación o rechazo.

$$\alpha = 0.05$$

$$gl\ numerador = 3$$

$$gl\ denominador = 24 \rightarrow F = 3.01$$

4. Obtener la conclusión o interpretación

El valor estadístico cae en la zona de rechazo por lo tanto no se acepta $H_o$ . Es decir no todas las medias que representan el promedio de compras en la agencia de viajes son iguales.

# 8. La regresión múltiple

En la nota técnica anterior se explicó la regresión simple, mediante la cual llegamos a establecer (estimar), con la utilización de la técnica de los mínimos cuadrados ordinarios (MCO), una ecuación (y = ) que muestra la relación existente entre la variable dependiente y una variable independiente.

Bajo este acápite se amplía la explicación para analizar cómo calcular una ecuación que recoja la relación entre una variable dependiente y más de una variable independiente.

En general, una ecuación de la forma $Y$ = es una ecuación que explica la relación de una variable dependiente ($Y$), explicada por n variables independientes ($X$). Por lo tanto, estamos frente a una regresión múltiple.

Para encontrar los valores de $\beta_0, \beta_{11}, \beta_2 ..... \beta_n$, su cálculo es a través de un sistema de ecuaciones o acudimos a uso de distinto *software* estadístico que ayuda con estos cálculos.

**Ejemplo 8.1**



Bajo la suposición que las ventas en un hotel dependen de la publicidad, de la fuerza de ventas y de la ciudad, nuestra ecuación explicativa de esta relación es la siguiente:

$$Ventas\ del\ hotel = \beta_0 + \beta_1\ publicidad + \beta_2\ Fuerzadeventas + \beta_3\ Ciudad$$

Supongamos que el último mes se obtuvieron de una muestra aleatoria a 12 hoteles de los cuales se obtuvo los datos de las habitaciones vendidas, los gastos de publicidad realizados en ese mes, igualmente el número de vendedores de tiempo completo contratados y si el hotel está ubicado en la ciudad. Los datos aparecen en la siguiente tabla:

## Tabla 8-1

Datos obtenidos de hoteles

| Habitaciones vendidas ($Y$) | Publicidad ($X$) | Fuerza de ventas ($X$) | Ciudad ($X$) |
|---|---|---|---|
| 127 | 18 | 10 | Si |
| 138 | 15 | 15 | No |
| 159 | 22 | 14 | Si |
| 144 | 23 | 12 | Si |
| 139 | 17 | 12 | No |
| 128 | 16 | 12 | Si |
| 161 | 25 | 14 | Si |
| 180 | 26 | 17 | Si |
| 102 | 15 | 7 | No |
| 163 | 24 | 16 | Si |
| 106 | 18 | 10 | No |
| 149 | 25 | 11 | Si |

## Ejemplo en Microsoft Excel



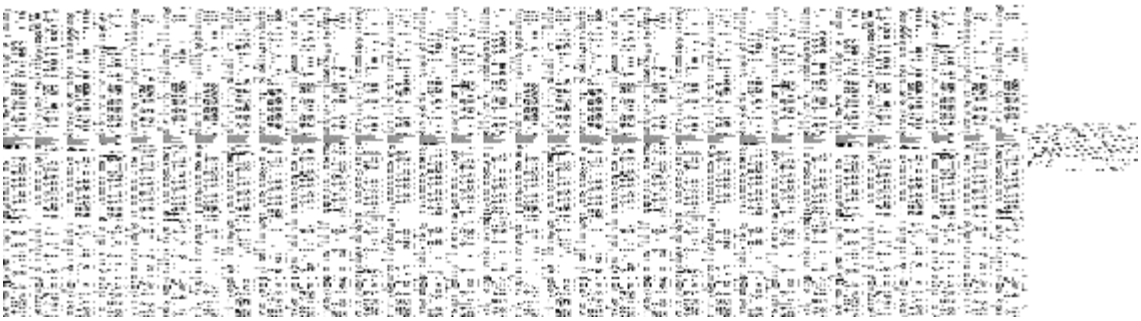


*Figura 8-1.* Regresión ejemplo 8.1

Que, al ser analizada nos muestra:

- Existe un 96,02% de relación entre la variables
- El 90,45% de las ventas son influenciadas por las ventas
- Nos provee la ecuación:

$$\hat{y}=25,2952+2,6187x_1+5,0233x_2$$

La que representa matemáticamente la relación entre las ventas del hotel y los factores publicidad y fuerza de ventas, de acuerdo al análisis de los datos provistos.

**Ejemplo 8.2:**

Con este ejemplo recordemos también la regresión simple. Por el momento supongamos que las habitaciones vendidas en estos hoteles dependen únicamente de los gastos de publicidad, es decir qué es las habitaciones y son los gastos de publicidad.

## Tabla 8-2

Ejemplo Regresión simple

| Habitaciones vendidas ($Y$) | Publicidad ($X$) |
|---|---|
| 127 | 18 |
| 138 | 15 |
| 159 | 22 |
| 144 | 23 |
| 139 | 17 |
| 128 | 16 |
| 161 | 25 |
| 180 | 26 |
| 102 | 15 |
| 163 | 24 |
| 106 | 18 |
| 149 | 25 |

## Ejemplo en Microsoft Excel

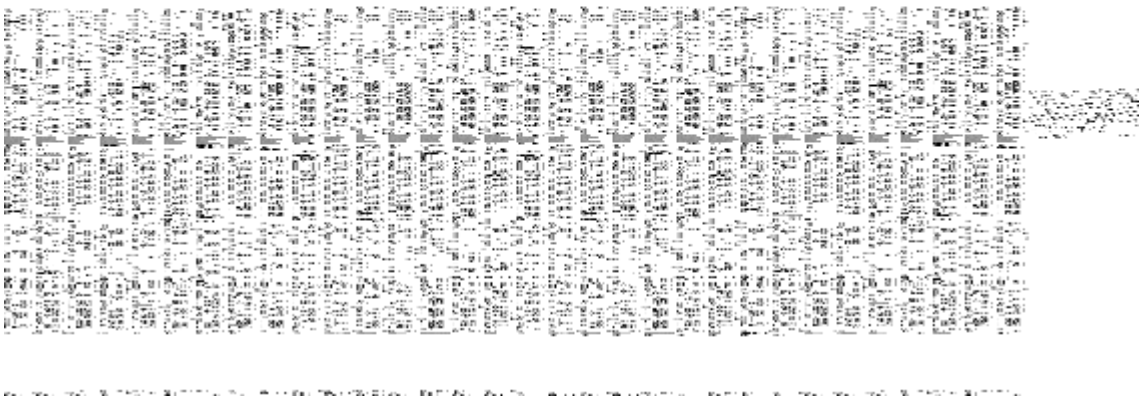


*Figura 8-2.* Regresión ejemplo 8.2

Entonces, la ecuación queda expresada como:

$$Y = 51.21 + 4.43X$$

$$Hab = 51.21 + 4.43\ (Pub).$$



Diríamos entonces que existe una relación de 81% de las dos variables y la publicidad explica el 65% del comportamiento de las ventas de las habitaciones de un hotel. Esto nos lleva a pensar que aún existe un 35% que está explicado por otras variables y que para este ejemplo podría ser la fuerza de ventas y la ubicación.

Algunos supuestos del análisis de regresión como:

- La homoscedasticidad: entendida como las varianzas en los valores de  son las mismas en todos los valores de $X$.

- La autocorrelación: ocurre cuando los términos de error (la diferencia entre el valor observado y el valor estimado) no son independientes.
- La multicolinealidad: existe si dos o más variables independientes están relacionadas linealmente.

Dentro de la regresión múltiple es posible calcular un dato estadístico muy importante conocido como el Durbin Watson ($d$), que es un factor estadístico que prueba la autocorrelación. El  se utiliza para probar la hipótesis de que no existe correlación entre términos de error sucesivos.

# 9. Análisis de series de tiempo

Un tema muy relacionado con la estimación simple o múltiple lo constituyen las series de tiempo.

## Definición:

Una serie de tiempo es un conjunto de observaciones medidas en puntos sucesivos, a lo largo del tiempo o en periodos sucesivos de tiempo.

Ejemplo 9.1 Los datos que representan 15 trimestres de ocupación de la infraestructura hotelera en el país (%).

### Tabla 9-1

*Ocupación de infraestructura hotelera*

| Trimestre/Año | Utilización |
|---|---|
| I/94 | 82.5 |
| II/94 | 81.3 |
| III/94 | 81.3 |
| IV/94 | 79.0 |
| I/95 | 76.6 |
| II/95 | 78.0 |
| III/95 | 78.4 |
| I/96 | 78.8 |
| II/96 | 78.7 |
| III/96 | 78.4 |
| IV/96 | 80.0 |
| I/97 | 80.7 |
| II/97 | 80.7 |
| III/97 | 80.8 |

Los modelos de series de tiempo se refieren a la medición de valores de una variable en el tiempo a intervalos espaciados uniformemente.

El objetivo de la identificación de la información histórica es determinar un patrón básico en su comportamiento, que posibilite la proyección futura de la variable deseada (extrapolar dicho patrón hacia el futuro).

Un método de proyección es el *método intuitivo,* el cual presume que el mejor predictor del valor de la variable en el siguiente periodo, es su valor en el periodo corriente, es decir, $y$ es el estimado del valor de la serie de tiempo en el siguiente periodo $_{t+1}$ e $y_t$ es el valor real en el periodo corriente.

## 1. Componentes de la serie de tiempo

En un análisis de serie de tiempo pueden distinguirse cuatro componentes básicos:

a. El componente tendencia se refiere al crecimiento o declinación en el largo plazo del valor promedio de la variable estudiada, por ejemplo: el porcentaje de ocupación hotelera.

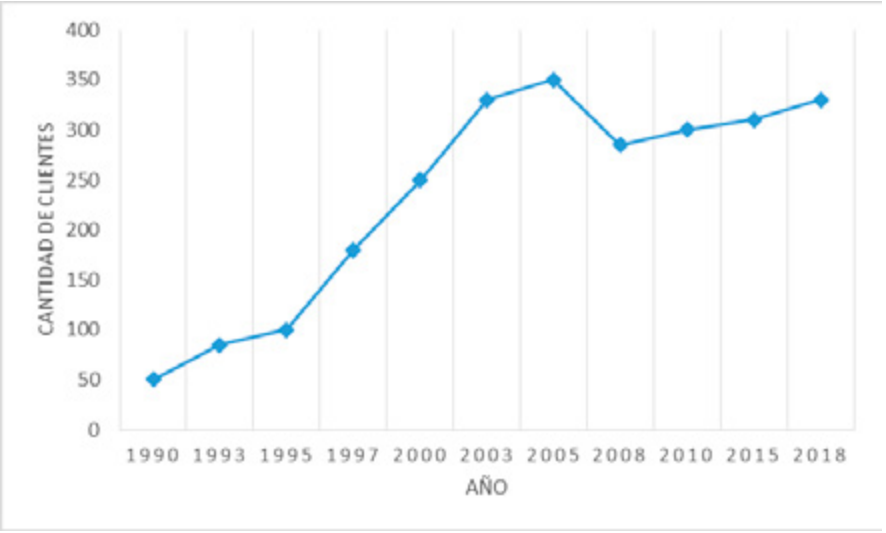



*Figura 9-1.* Tendencia de una serie de tiempo

b. El componente cíclico son las variaciones que se producen entre la línea de tendencia proyectada o estimada y el valor que realmente exhibe la variable. Estas variaciones admiten que son causadas por un efecto combinado de fuerzas económicas, sociales, políticas, ideológicas y tecnológicas.

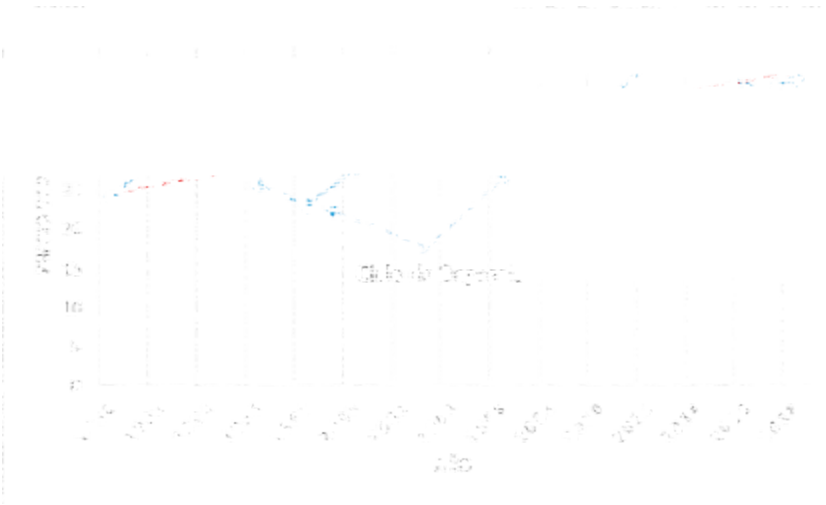


*Figura 9-2.* Componente Cíclico de una serie de tiempo

c. El componente estacional consiste en fluctuaciones de la variable, que se repiten periódicamente y que normalmente dependen de factores como: la costumbre, el clima, las tradiciones, etc.



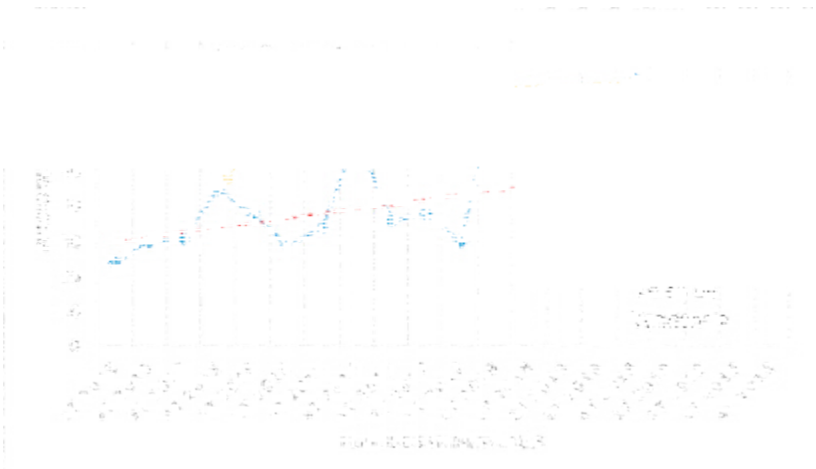


*Figura 9-3.* Componente estacional de una serie de tiempo

d. El componente aleatorio se refiere a las variaciones irregulares producidas por sucesos inusuales que provocan movimientos en la variable, sin un patrón discernible.

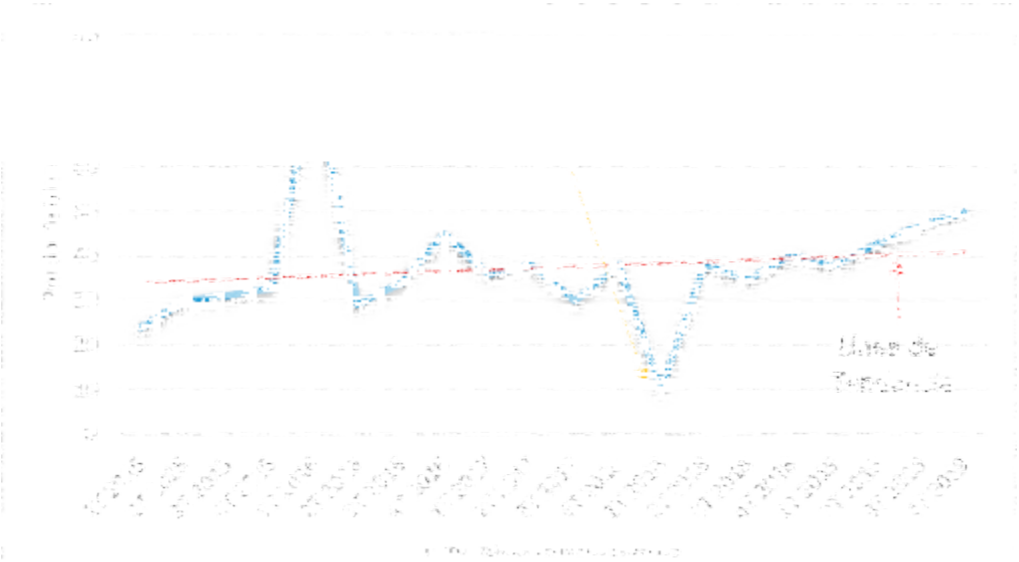


*Figura 9-4.* Componente aleatorio de una serie de tiempo

## 2. Modelos de series de tiempo



Un modelo puede expresarse como una ecuación que combina los cuatro componentes.

a. El modelo aditivo: supone que los componentes actúan de manera independiente, y se expresa con la siguiente ecuación:

$$Y^t = T_t + S_t + C_t + I_t$$

Lo que se traduce como: Valor de la serie de tiempo para el periodo t es igual al valor tendencial más valor estacional más variación cíclica más variación irregular.

**Ejemplo 2.1:**

Supongamos que se desea estimar el número de turistas que llegan al Hotel Galápagos en San Cristóbal. En el registro histórico del hotel se refleja que el hotel normalmente recibe a 100 huéspedes por año en la temporada baja ($T_t$ = 100). Por la estacionalidad o la temporada alta que muestra Galápagos registra 150 huéspedes adicionales ($S_t$ =150). Además, la economía este año muestra una mayor actividad y por ello hay un dinamismo del turismo interior y se espera que lleguen a Galápagos 100 turistas ($C_t$ =100) y finalmente, por la declaratoria de parque nacional en peligro se espera una caída de la demanda de 50 turistas ($I_t$ = 50); por lo tanto la demanda de huéspedes del hotel podría ser estimada de la siguiente manera:

$$YT$$

$$Y100 + 150 + 100 - 50 = 300$$

9.ii.2 El modelo multiplicativo: supone que los componentes interactúan entre sí, la ecuación que describe este modelo es:

$$YT$$

A diferencia del modelo aditivo, en este modelo los componentes se multiplican dejando al factor tendencial explicarse en la unidad original y el resto de componentes expresarse en porcentajes.



**Ejemplo 2.2:**

En una agencia de viajes los *ticket*s fallidos suman $10 mil. El componente estacional equivale a 1.7 veces el comportamiento normal, el factor cíclico incide en la venta del ticket en un 91% y el componente irregular incide en un 87%. Entonces la venta de ticket en esta agencia de viajes puede estimarse en:

$$YT$$

$$Y10.000 * 1.7*0.91*0.87 = 13.458 \text{ dólares.}$$

## 3. Técnicas de suavizamiento

Si la serie de tiempo contiene demasiadas variaciones o cambios estacionales a corto plazo la tendencia puede ser algo confusa y difícil de observar. Es posible eliminar muchos de estos factores usando ciertas técnicas de suavizamiento o de afinamiento proporcionando por tanto una conducta real de la serie, se mencionan las más relevantes:

### 3.1 Promedios móviles (*PM*):

Se calcula promediando los valores en la serie de tiempo sobre un número fijo de periodos, el cual se mantiene para cada promedio eliminando la observación más antigua y recogiendo la más reciente.

Los promedios móviles pueden utilizarse para eliminar variaciones irregulares y estacionales.

El PM promedia toda variación estacional que puede ocurrir dentro del año, eliminándolas de manera efectiva y dejando solo la tendencia y las variaciones cíclicas.

$$PM^{i} = \frac{\sum_{i=1}^{n} T_i}{n}$$



$\boldsymbol{T}_i =$ valor que adopta la variable en cada periodo $\boldsymbol{i}$, y $\boldsymbol{n}$ el número de periodos observados.

Utilizando los datos del ejemplo 9.1, sobre los porcentajes de ocupación hotelera podemos construir una serie de promedios móviles de 3 periodos (en este caso de trimestres), de la siguiente manera:

## Tabla 9-2

*Cálculo de promedios móviles PM3*

| Tiempo | Datos originales | PM3 |
|---|---|---|
| Trimestre/Año | Utilización | |
| I/94 | 82.5 | |
| II/94 | 81.3 | |
| III/94 | 81.3 | 81.70 |
| IV/94 | 79.0 | 80.53 |
| I/95 | 76.6 | 78.97 |
| II/95 | 78.0 | 77.87 |
| III/95 | 78.4 | 77.67 |
| IV/95 | 78.0 | 78.13 |
| I/96 | 78.8 | 78.40 |
| II/96 | 78.7 | 78.50 |
| III/96 | 78.4 | 78.63 |
| IV/96 | 80.0 | 79.03 |
| I/97 | 80.7 | 79.70 |
| II/97 | 80.7 | 80.47 |
| III/97 | 80.8 | 80.73 |

Para el $PM1$ o promedio móvil del primer periodo, utilizamos los tres primeros datos:



$PM1$ = 82.5+81.3+81.3/3 = 245.1/3 = 81.7. y así sucesivamente.

Esto podemos graficarlo de la siguiente manera:

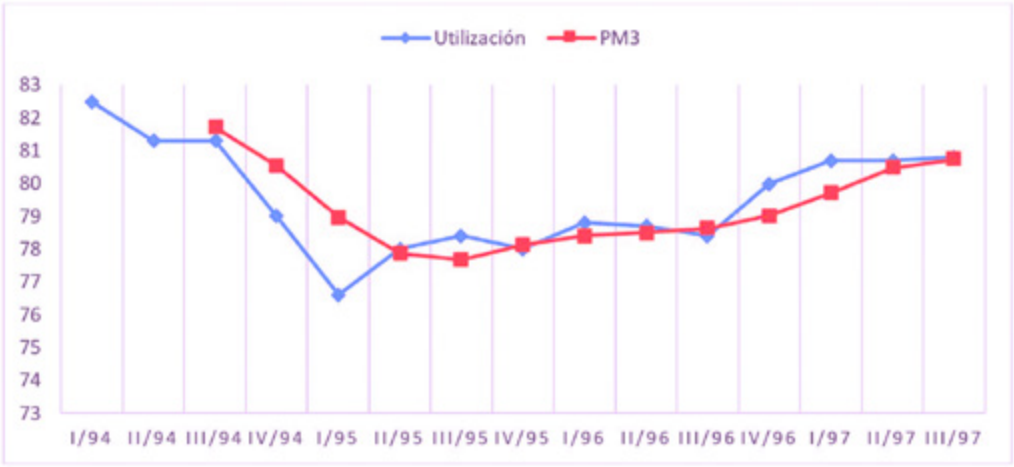



*Figura 9-5.* Promedio móvil

### 3.2 Suavizamiento exponencial:

Esta técnica tiene el efecto de suavizar una serie y por lo tanto proporciona un medio efectivo de predicción. Se basa en un promedio ponderado de los valores actuales y anteriores de la variable en estudio. Se calcula a través del siguiente modelo:

$$F_{t+1} = \alpha A_t + (1 - \alpha) F_t$$



$F$ es el pronóstico para el siguiente periodo.

$A_t =$ es el valor real observado para el periodo corriente.

$F_t$ = es la proyección hecha previamente para el periodo corriente.

$\alpha$ = constante de afinamiento a la cual se le da un valor $0 \leq \alpha \leq 1$

Siendo $\alpha$ lo crucial y puesto que se desea pronosticar con el error más pequeño posible, el valor de $\alpha$ que minimiza el cuadrado medio del error (*CMF*) debe ser el óptimo. Es decir, $\alpha$ será óptimo cuando se minimice el error calculado de la siguiente manera:

$$CMF = \frac{\sum (F_t - A_t)^2}{n-1}$$

Aplicación de esta técnica con el ejemplo 9.3.1:

## Tabla 9-3

*Aplicación del suavizamiento exponencial 1*

| Tiempo | Datos originales | $\alpha$=0.3 | $\alpha$=0.5 |
|---|---|---|---|
| Trimestre/Año | Utilización | | |
| I/94 | 82.5 | | |
| II/94 | 81.3 | 82.50 | 82.50 |
| III/94 | 81.3 | 82.14 | 81.90 |
| IV/94 | 79.0 | 81.20 | 80.45 |
| I/95 | 76.6 | 79.82 | 78.53 |
| II/95 | 78.0 | 79.27 | 78.26 |
| III/95 | 78.4 | 79.01 | 78.33 |
| IV/95 | 78.0 | 78.71 | 78.17 |
| I/96 | 78.8 | 78.74 | 78.48 |
| II/96 | 78.7 | 78.72 | 78.59 |
| III/96 | 78.4 | 78.63 | 78.50 |
| IV/96 | 80.0 | 79.04 | 79.25 |
| I/97 | 80.7 | 79.54 | 79.97 |
| II/97 | 80.7 | 79.89 | 80.34 |
| III/97 | 80.8 | 80.16 | 80.57 |

En este caso la proyección de febrero de 1994 (segundo periodo) es la observación del periodo anterior (proyección intuitiva), para el siguiente periodo, marzo del 94, se la hará aplicando la fórmula:

$$F_{t+1} = \alpha A_t + (1-\alpha)F_t$$

Proyección de marzo = 0.3 * 81.3 + (1 - 0.3)* 82.50 = 82.14

Si cambiamos la constante de afinamiento a α de 0.5, el pronóstico se muestra en la tabla adjunta. Ahora para mirar el mejor valor de α calculamos el *CME*.

| | | α | 0.5 | 1-α | 0.5 |
|---|---|---|---|---|---|
| Período | Pronóstico | PM3 | SE | | CME |
| 1 | 82.50 | | | | |
| 2 | 81.30 | | 82.50 | | 1.44 |
| 3 | 81.30 | 81.70 | 81.90 | | 0.36 |
| 4 | 79.00 | 80.53 | 80.45 | | 2.10 |
| 5 | 76.60 | 78.97 | 78.53 | | 3.71 |
| 6 | 78.00 | 77.87 | 78.26 | | 0.07 |
| 7 | 78.40 | 77.67 | 78.33 | | 0.00 |
| 8 | 78.00 | 78.13 | 78.17 | | 0.03 |
| 9 | 78.80 | 78.40 | 78.48 | | 0.10 |
| 10 | 78.70 | 78.50 | 78.59 | | 0.01 |
| 11 | 78.40 | 78.63 | 78.50 | | 0.01 |
| 12 | 80.00 | 79.03 | 79.25 | | 0.57 |
| 13 | 80.70 | 79.70 | 79.97 | | 0.53 |
| 14 | 80.70 | 80.47 | 80.34 | | 0.13 |
| 15 | 80.80 | 80.73 | 80.57 | | 0.05 |
| | | | | | 9.11 |
| | | | | CME | 0.65 |

Con el $\alpha = 0.3$ *CME* se calcula así:

| | | | α=0.3 | 1-α | 0,7 |
|---|---|---|---|---|---|
| | | | | | CME |
| 1 | 82.50 | | | | |
| 2 | 81.30 | | 82.50 | | 1.44 |
| 3 | 81.30 | 81.70 | 82.14 | | 0.71 |
| 4 | 79.00 | 80.53 | 81.20 | | 4.83 |
| 5 | 76.60 | 78.97 | 79.82 | | 10.36 |
| 6 | 78.00 | 77.87 | 79.27 | | 1.62 |
| 7 | 78.40 | 77.67 | 79.01 | | 0.37 |
| 8 | 78.00 | 78.13 | 78.71 | | 0.50 |
| 9 | 78.80 | 78.40 | 78.74 | | 0.00 |
| 10 | 78.70 | 78.50 | 78.72 | | 0.00 |
| 11 | 78.40 | 78.63 | 78.63 | | 0.05 |
| 12 | 80.00 | 79.03 | 79.04 | | 0.92 |
| 13 | 80.70 | 79.70 | 79.54 | | 1.35 |
| 14 | 80.70 | 80.47 | 79.89 | | 0.66 |
| 15 | 80.80 | 80.73 | 80.16 | | 0.41 |
| | | | | | 23.23 |
| | | | | CME | 1.66 |



En este caso la mejor constante de afinamiento es α de 0.5, puesto que me refleja el menor error (1.66 versus 0.65).

**Ejercicio propuesto:**

a. La siguiente serie de tiempo muestra las ventas de un producto en particular a lo largo de los últimos 12 meses.

| Mes | Ventas | Mes | Ventas |
|---|---|---|---|
| 1 | 105 | 7 | 145 |
| 2 | 135 | 8 | 140 |
| 3 | 120 | 9 | 100 |
| 4 | 105 | 10 | 80 |
| 5 | 90 | 11 | 100 |
| 6 | 120 | 12 | 110 |

1. Utilice una constante de afinamiento de 0.3 para calcular los valores de suavización exponencial de la serie de tiempo.
2. Utilice una constante de suavización de 0.5 para calcular los valores de suavización exponencial ¿Dará un mejor pronóstico la constante de 0.3 o de 0.5?
3. Realice el gráfico correspondiente.

# 10. Números índices

Un número índice relaciona un valor en un periodo de tiempo denominado "periodo base" con el valor en otro periodo denominado periodo corriente o de referencia. Podría ampliarse esta definición para indicar que se trata de un número que muestra la variación o el peso de una variable entre otra, en un intervalo de tiempo.

Existen varios números índices que se pueden construir partiendo de su definición, por ejemplo: el porcentaje de ocupación de un hotel puede expresarse a través de construir un número índice.

Vamos a referirnos a los siguientes:

**a) Índice de precios simple:**

Indica el cambio relativo en el precio de un bien o servicio en el periodo de referencia, con respecto al periodo base. Su cálculo es como sigue:

$$IP_R = \frac{P_R}{P_B} \times 100$$



$IP_R$=Índice de precios simple

$P_R$ = Precio del año de referencia

$P_B$ = Precio del año base

### b) Índice de precios agregativo

Calcula el índice de precios para varios bienes/servicios. Se usa en empresas que proveen 2 o más bienes/servicios. Su cálculo es como sigue:

$$P_A = \frac{\sum P_R}{\sum P_B} \times 100$$

$IP_A$=Índice de precios agregativo

Este índice no toma en cuenta, para su cálculo, la unidad de medida ni el volumen de venta.

### c) Índice de precios agregativos ponderados:

Calcula el índice de precios para varios bienes/servicios considerando el volumen de venta y la unidad de medida. Se usa en general dos tipos de índices, el índice de LASPEYRES ($L$) y, el índice de PAASCHE ($P$),su cálculo es como sigue:

- El índice de LASPEYRES ($L$): Utiliza las cantidades vendidas en el periodo base como factor de ponderación

$$L = \frac{\sum (P_R \times Q_B)}{\sum (P_B \times Q_B)} \times 100$$



- El índice de PAASCHE (P): Utiliza como factor de ponderación las cantidades vendidas en el periodo de referencia.

$$P = \frac{\sum (P_R \times Q_R)}{\sum (P_B \times Q_R)} \times 100$$

Ejemplo 10.1 Cálculo de números índices:

Como pasante en la Escuela de Economía de la Universidad de Cuenca le piden que desarrolle un índice para propósitos especiales. Tres series económicas parecen ser adecuadas para la base de un índice. Estos datos son el precio del algodón (por libra), el número de autos nuevos vendidos en Cuenca y los movimientos de dinero (publicados por la banca local). Después de discutir el proyecto con su profesor de estadística y con el decano decide que la recuperación monetaria debe tener una ponderación de 0.60, el número de autos nuevos vendidos de 0.30 y el precio del algodón de 0.10. El periodo base es 1995.

## Tabla 10-1

*Datos cálculo de índice 10.1*

| Año | Precio del algodón ($) | Automóviles vendidos | Movimientos de dinero |
|---|---|---|---|
| 1995 | 0.20 | 1.000 | 80 |
| 2000 | 0.25 | 1.200 | 90 |
| 2004 | 0.50 | 900 | 75 |

Elabore el índice para 2000



| Artículo | Ponderación |
|---|---|
| Algodón | (0.25/0.20)(100)(0.10) = 12.5 |
| Autos | (1200/1000)(100)(0.30) =36.0 |
| Movimiento de dinero | (90/80)(100)(0.60) = 67.5 |
| | 116.0 |

En el año base el número índice vale 100 por definición.

Entonces la variación entre 1995 y 2000 es de 16% .

### d) Índice ideal de Fisher (F)

Es un índice especial que combina a Laspeyres y Paasche encontrando la raíz cuadrada de su producto, es decir:

$$F=\sqrt{L}xP$$

### e) El índice de precios al consumidor (IPC)3

Mide el cambio en el precio de los bienes/servicios de la canasta básica, de un periodo a otro. Se lo conoce como tasa de inflación o índice de inflación; y su cálculo es como sigue:

$$IPC = \frac{IPC_t - IPC_{t-1}}{IPC_{t-1}} \times 100$$

El IPC es un índice de precios agregativo tanto del periodo o referencial con el periodo o periodo base.

Cuando se desea medir el poder real de compra de un individuo hay que tomar en cuenta dos datos el IPC y el ingreso monetario que recibe ese individuo, la relación entre los dos proporciona el ingreso real, es decir:

$$Ingreso\ real = \frac{Ingreso\ monetario}{IPC} \times 100$$



**Ejemplo 10.2:**

Suponemos que el sueldo nominal de un obrero es 150 dólares/mes y el IPC de ese mes es 120, entonces su ingreso real o poder de compra en ese mes es de 125

(150/120*100 = 125).

---

3 Además del IPC el Instituto Nacional de Estadísticas y Censos (INEC) calcula el IPP el índice de precios al productor, que mide el cambio de un periodo a otro de los insumos de la producción.

# 11. Pruebas no paramétricas

Cuando el objeto de nuestro estudio forma una población cuyos datos no se comportan como un distribución normal, nos proporciona datos con picos muy pronunciados, sesgados ya sea a la derecha o a la izquierda, o información de tipo cualitativa, no será posible aplicar la prueba t o F que suponen normalidad en los datos para poder inferir sobre la población de estudio, existe la posibilidad de analizarlos estadísticamente con pruebas no paramétricas, conocidas también como libres de distribución, pues no dependen de supuestos relativos a la distribución.

*Las pruebas no paramétricas* son procedimientos estadísticos que pueden utilizarse para contrastar hipótesis cuando no son posibles los supuestos respecto de los parámetros o de las distribuciones poblacionales.

Dentro de estas pruebas destacamos las siguientes:

### a) La prueba chi cuadrada ($\chi$)

Esta prueba estadística tiene la característica de una distribución sesgada. Comprende una familia de curvas que aumenta a medida que se incrementan los grados de libertad (gl). Sus aplicaciones más comunes son las pruebas de bondad de ajuste y las pruebas de independencia.



La prueba de bondad de ajuste se trata de un procedimiento para determinar si la distribución de los valores en la población se ajustan a una forma particular planteada como hipótesis. Los datos muestrales se toman de la población y estos constituyen la base de los hallazgos. En resumen, la prueba $X$ determina si las observaciones muestrales se ajustan a las expectativas.

Siguiendo los pasos normales de una prueba de hipótesis, en el caso de esta, la hipótesis nula que se trata de verificar es como sigue:

$H$: la distribución poblacional es uniforme.

$H_1$: la distribución poblacional no es uniforme.

Luego el factor estadístico de prueba se calcula de la siguiente manera:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

En donde:

$O_i$: Es la frecuencia de los eventos observados.

$E_i$: Es la frecuencia de los eventos esperados si la  es correcta.

$K$: Es el número de categorías o clases.

La prueba tiene $k$ - $m$ - 1 grados de libertad, en donde $m$ es el número de parámetros a estimar

Con el valor calculado de chi cuadrada, vamos a la tabla de chi cuadrada (como las que hemos visto prueba $Z$, $t$ $y$ $F$ ), elaboramos el gráfico de la prueba, ubicamos los valores críticos y el valor observado, que nos sirve para decidir si se acepta o rechaza la $H_o$.

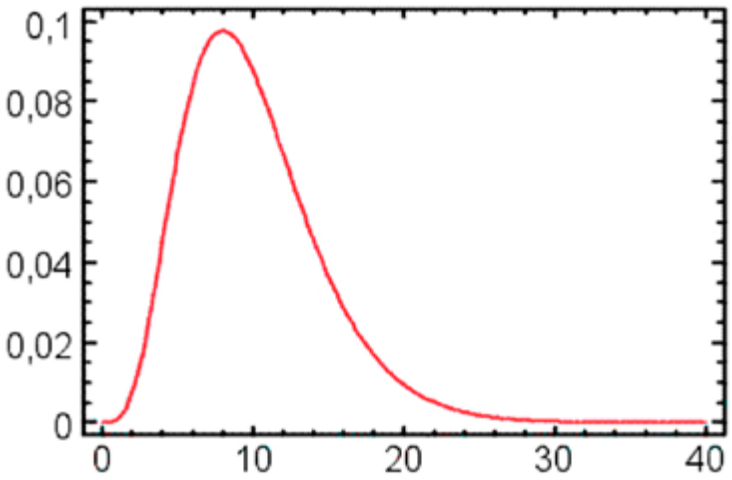



*Figura 11-1*. Gráfico Chi Cuadrada

Existen tres aplicaciones de esta prueba de bondad de ajuste:

I. para un ajuste uniforme,

II. un patrón específico y

III. una normalidad.

## i. Prueba para un ajuste uniforme

**Ejemplo 11.1**

Supongamos que tenemos que probar la hipótesis de que la distribución de las ventas de un hotel son uniformes, sabiendo que tenemos varios tipos de habitaciones, tal como lo muestra la siguiente tabla, y esperamos que se puedan vender de manera uniforme 12 habitaciones por periodo. Se trata de probar esta hipótesis a un nivel de significancia del 10%.

Tabla 11-1

*Datos ventas- habitación*

| Tipo de habitación | Ventas mensuales observadas del último trimestre ($O_i$) | Ventas mensuales esperadas del último trimestre ($E_i$) | ($O_i$-$E_i$) |
|---|---|---|---|
| Sencilla | 15 | 12 | 9 |
| Doble | 11 | 12 | 1 |
| Triple | 10 | 12 | 4 |
| Suite | 12 | 12 | 0 |



El procedimiento de la prueba de hipótesis sería:

1. **Planteamiento de las hipótesis $H_o$ y $H_1$:**

$H_o$: la distribución poblacional es uniforme.

$H_1$: la distribución poblacional no es uniforme.

**2. Cálculo del factor estadístico:**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Con los datos calculados tenemos:

$$\chi^2 = \frac{9}{12} + \frac{-1}{12} + \frac{4}{12} + \frac{0}{12} = 14/12 = 1.17$$

**3. Graficamos y calculamos los valores críticos:**

Para el cálculo del valor crítico necesitamos ir a la tabla de chi cuadrado y necesitamos los grados de libertad. Para este ejemplo tenemos $K$ = 4 categorías; $m$ = 0 parámetros a estimar; por lo tanto los g.l. es 3 (g.l. = $k$-$m$-1= 4 - 0 - 1)

$\chi^2$ 0.10; con 3 grados de libertad en la tabla el valor de chi cuadrado es 6,25

El valor $\chi^2$ calculado de 1.17 cae en la zona de rechazo, por lo tanto

**4. Decidimos rechazar $H_o$.**



### b) Prueba para un patrón específico

Dentro de esta prueba es importante considerar que las frecuencias esperadas ($E_i$) son iguales al tamaño de la muestra, $n$, por la probabilidad de cada categoría, es decir:

$$E_i = n\ p_i$$

En donde:

$n$ : tamaño de la muestra

$p_i$: probabilidad de cada categoría, como se especificó en la $H_o$.

**Ejemplo 11.2**

El gerente de un hotel trata de seguir una política de hospedaje de 60% de sus habitaciones a ejecutivos, 10% a nacionales y 30% a extranjeros. Para determinar si la política se estaba cumpliendo se seleccionó una muestra de 85 huéspedes, encontrando que 62 de ellos eran ejecutivos, 10 eran nacionales y 13 eran extranjeros. ¿A un nivel de significancia del 10% parece que se está cumpliendo el patrón de hospedaje deseado?

Sigamos el procedimiento de la prueba de hipótesis:

1. **Planteamiento de las hipótesis $H_o$ y $H_1$:**

$H_o$: Se mantuvo el patrón de hospedaje de 60% ejecutivos,10% nacionales y 30% extranjeros

$H_1$: No se mantuvo el patrón deseado.



2. **Cálculo del factor estadístico:**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Con los datos calculados tenemos:

## Tabla 11-2

Datos procedencia de huéspedes

| Tipo de huésped | Frecuencias observadas ($O_i$) | Frecuencias esperadas trimestre ($E_i = np_i$) | ($O_i$-$E_i$) |
|---|---|---|---|
| Ejecutivos | 62 | 85* 0.60 = 51.0 | $(62-51)^2$ |
| Nacionales | 10 | 85*0.10 = 8.50 | $(10-8.5)^2$ |
| Extranjeros | 13 | 85*0.30 = 25.50 | $(13-25.5)^2$ |
| | | | |
| Muestra $n$ = 85 | 85 | | |

$$\chi^2 = \frac{121}{51} + \frac{2.25}{8.5} + \frac{156.25}{25.5} = 8.76$$

**3. Graficamos y calculamos los valores críticos:**



Para el cálculo del valor crítico necesitamos ir a la tabla de chi cuadrado y necesitamos los grados de libertad. Para este ejemplo tenemos $K$ = 3 categorías; $m$ = 0 parámetros a estimar; por lo tanto los g.l. es 2 (g.l.= $k$-$m$-1= 3-0-1)

$\chi^2$ = 0.10; 2 en la tabla el valor de chi cuadrado es 4,61

El valor $\chi^2$ calculado de 8,76 cae en la zona de rechazo, por lo tanto

4. **Decidimos rechazar $H_o$.**

### c) Prueba para una normalidad

Dentro de esta prueba es importante recordar lo que habíamos visto en la distribución normal. Justamente en esta prueba la  es la probabilidad o el área bajo la curva.

**Ejemplo 11.3:**

Supongamos que las frecuencias de los turistas que han llegado al país en el último año se encuentran agrupadas en las siguientes clases, además el promedio de la población indica que existen $\mu = 600$ turistas que han llegado, con una desviación estándar de $\sigma = 10$ . Probemos esta hipótesis con el 5% de significancia.

Tabla 11-3

Información de arribos de turistas

| Número de turistas | Frecuencias observadas ($O_i$) | Probabilidades ($P_i$) | Frecuencias esperadas trimestre ($E_i = np_i$) | ($O_i$-$E_i$) |
|---|---|---|---|---|
| 0-580 | 20 | 0.0228 | 22.8 | $(20-22.8)^2$ |
| 580-590 | 142 | 0.1359 | 135.9 | $(142-135.9)^2$ |
| 590-600 | 310 | 0.3413 | 341.3 | $(310-341.3)^2$ |
| 600-610 | 370 | 0.3413 | 341.3 | $(370-341.3)^2$ |
| 610-620 | 128 | 0.1359 | 135.9 | $(128-135.9)^2$ |
| 620-mas | 30 | 0.0228 | 22.8 | $(30-22.8)^2$ |
| | | | | |
| Muestra | 1.000 | | | |

$$Z = \frac{X - \mu}{\sigma}$$

La probabilidad para la primera clase la calculamos de la siguiente manera:

$$X = 580 \; ; \; \sigma = 10 \text{ y } \mu = 600$$

$$Z = \frac{580\text{-}600}{10} = \text{-2 en la tabla obtenemos 0.4772}$$

**Gráficamente:**

Entonces P(0<X<580) = 0.50 - 0.4772 = 0.0228

Para la siguiente categoría:

$$X = 590; \; \sigma = 10 \text{ y } \mu = 600$$

$$Z = \frac{590\text{-}600}{10} = 1 \text{ en la tabla obtenemos } 0.3413$$

**Gráficamente:**

Entonces P (580 < X < 590) = 0.4772 - 0.3413 = 0.1359

Con este criterio y forma de cálculo completamos la tercera columna del cuadro anterior.



Sigamos el procedimiento de la prueba de hipótesis:

1. **Planteamiento de las hipótesis y :**

$H_o$: la llegada de turistas se distribuye normalmente.

$H_1$: la llegada de turistas se distribuye normalmente.

2. **Cálculo del factor estadístico:**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{121}{\quad} + \frac{2.\quad}{8.5} + \frac{156.\quad}{\quad .5} = 8.76$$

3. **Graficamos y calculamos los valores críticos:**

Para el cálculo del valor crítico necesitamos ir a la tabla de chi cuadrado y necesitamos los grados de libertad. Para este ejemplo tenemos $K$ = 6 categorías; $m$ = 0 parámetros a estimar; por lo tanto los g.l. es 5 (g.l.= $k$-$m$-1= 6-0-1)

$\chi^2$ 0.05; 5 en la tabla el valor de chi cuadrado es 10,64

El valor $\chi^2$ calculado de 8,76 cae en la zona de rechazo, por lo tanto

4. **Decidimos rechazar .**



### d) Prueba del signo

Esta prueba se utiliza comúnmente para tomar decisiones comerciales. Su propósito es contrastar la hipótesis comparando dos distribuciones poblacionales y que por lo general implica el uso de pares correspondientes. Se entiende que se tiene un conjunto de datos antes y después para una muestra, y se desea comparar estos datos.

Su cálculo se hace restando las observaciones por pares en un conjunto de datos de las del segundo conjunto, y se anota el signo algebraico que resulta. No se tiene interés en la magnitud de la diferencia sino solo en que su signo sea positivo (+) o negativo (-).

La $H_0$ establece que no existe diferencia en los conjuntos de datos. Si esto es cierto entonces un signo + o – son igualmente probables. La $p_i$ de obtener el signo + es de 0.50 al igual que la $p_i$ de obtener el signo - es de 0.50.

La hipótesis nula y alternativa para una prueba de dos colas queda expresada de la siguiente manera:

$$H_o: m = p$$
$$H_1: m \neq p$$

$m$ es el número de signos negativos y $p$ el número de signos positivos.

La hipótesis nula y alternativa para una prueba de una cola derecha e izquierda respectivamente queda expresada de la siguiente manera:

$$H_o: m \leq p$$
$$H_1: m > p$$

$m$ es el número de signos negativos y $p$ el número de signos positivos.



$$H_o: m \geq p$$
$$H_1: m < p$$

$m$ es el número de signos negativos y  el número de signos positivos.

**Ejemplo 11.3**

Supongamos que una agencia de viajes desea medir al 5% de significancia, que el impacto de sus paquetes promocionales es el mismo antes y después de la promociones. Para ello toma una muestra en doce sucursales con el nivel de ventas. Los datos se presentan en miles de dólares.

## Tabla 11-4

Ventas por sucursal

| Sucursal | Antes de la promoción | Después de la promoción | Signo |
|---|---|---|---|
| 1 | 42 | 40 | + |
| 2 | 57 | 60 | - |
| 3 | 38 | 38 | 0 |
| 4 | 49 | 47 | + |
| 5 | 63 | 65 | - |
| 6 | 36 | 39 | - |
| 7 | 48 | 49 | - |
| 8 | 58 | 50 | + |
| 9 | 47 | 47 | 0 |
| 10 | 51 | 52 | - |
| 11 | 83 | 72 | + |
| 12 | 27 | 33 | - |



Sigamos el procedimiento de la prueba de hipótesis:

1. **Planteamiento de las hipótesis $H_o$ y $H_1$ :**

$$H_o: m \leq p$$

$$H_1: m > p$$

En el ejemplo tenemos que $m = 6$ y $p = 4$ por lo tanto esta es la hipótesis a probar.

2. **Cálculo del factor estadístico:**

Para este cálculo usamos la distribución binomial acumulada y su respectiva tabla para facilitar el cálculo.

$n$ = 10 signos (los ceros no se consideran); $m$ = 6 y $p$ = 4

$p(m \geq 6 | n = 10; \pi = 0.5) = 1 - = p$ ( $x \leq 5$) = 0.3770

3. **Graficamos y calculamos los valores críticos:**

**Si** $p(m \geq 6 | n = 10; \pi = 0.5)$ > 5% entonces aceptamos $H_o$

37.70% > 5% entonces se acepta $H_o$

4. **Decidimos no rechazar $H_o$**

### e) Prueba de Mann-Whitney (o simplemente la prueba U)



El propósito de esta prueba es contrastar la igualdad de dos distribuciones poblacionales. Se basa en la suposición de que dos muestras aleatorias que se sacan independientemente de variables continuas tienen parámetros idénticos.

La $H_o$ establece que la distribución de dos poblaciones es idéntica.

Se puede realizar esta prueba para analizar la igualdad de las dos medias o medianas poblacionales. Se usan las medias si las poblaciones son simétricas y si tienen la misma varianza. Si se elimina este supuesto de simetría se pueden usar las medianas. Para probar la hipótesis es necesario calcular la prueba U de cada muestra, de la siguiente manera:

Luego procedemos a calcular la media de las muestras, como sigue:

$$\mu_u = \frac{n_1 n_2}{2}$$

Y también la desviación estándar, como sigue:

Al mantener el supuesto de normalidad es necesario calcular el factor estadístico Z como sigue:

$$z = \frac{u_i - \mu_u}{\sigma_u}$$

**Ejemplo 11.4**



Supongamos que tenemos el registro de huéspedes en dos hoteles de la ciudad y se trata de probar a un 10% de significancia que los dos hoteles registran el mismo número de huéspedes. El registro de las dos muestras es de 12 y 10 respectivamente y aparece en la siguiente tabla:

Tabla 11-5
Registro de ingreso de huéspedes

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Hotel 1** | 27 | 31 | 28 | 29 | 39 | 40 | 35 | 33 | 32 | 36 | 37 | 43 |
| **Hotel 2** | 34 | 24 | 38 | 28 | 30 | 34 | 37 | 42 | 41 | 44 | | |

Procedemos a ordenar del más bajo al más alto siguiendo un orden para el ranking del hotel:

Tabla 11-6

*Clasifiación de la información por rangos*

| Hotel 1 | Rango | Hotel 2 | Rango |
|---|---|---|---|
| | | 24 | 1 |
| 27 | 2 | | |
| 28 | 3.5 | 28 | 3.5 |
| 29 | 5 | | |
| | | 30 | 6 |
| 31 | 7 | | |
| 32 | 8 | | |
| 33 | 9 | | |
| | | 34 | 10.5 |
| | | 34 | 10.5 |
| 35 | 12 | | |
| 36 | 13 | | |
| 37 | 14.5 | 37 | 14.5 |
| | | 38 | 16 |
| 39 | 17 | | |
| 40 | 18 | | |
| | | 41 | 19 |
| | | 42 | 20 |
| 43 | 21 | | |
| | | 44 | 22 |
| ΣR1 | 130 | ΣR2 | 123 |



1. **Plantear la hipótesis nula ($H_o$) y la hipótesis alternativa ($H_1$)**

$$H_o: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

2. **Cálculo del factor estadístico:**

$$U_{n1} = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

Para el hotel 1:

$$U_{n1} = 12 * 10 + \frac{12(12+1)}{2} - 130 = 68$$

Para el hotel 2:

$$U_{n2} = 12 * 10 + \frac{10(10+1)}{2} - 123 = 52$$

Luego procedemos a calcular la media de las muestras, como sigue:

$$\mu_u = \frac{n_1 n_2}{2} = 12*10/2 = 60$$

Y también la desviación estándar, como sigue:

$$\sigma_u = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 15.17$$

Al mantener el supuesto de normalidad es necesario calcular el factor estadístico Z, como sigue:



$$z = \frac{u_i - \mu_u}{\sigma_u} = 52\text{-}60/15.17 = -0.53$$

3. **Graficamos y obtenemos el valor crítico:**

$\alpha = 0.10$ y Z = 1.65 y el valor de la prueba cae en la zona de aceptación

4. **Aceptar** $H_o$

### f) Prueba de correlación de rangos de Spearman ($r_s$)

Esta prueba debe utilizarse cuando no se cumple el supuesto de normalidad en las distribuciones. En este caso se puede clasificar sistemáticamente u ordenar las observaciones. Esta clasificación ordinal permite medir los grados de correlación entre dos variables utilizando este coeficiente de Spearman[4] que se calcula así:

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

En donde:

$d_i =$ es la diferencia entre las clasificaciones para cada observación.

$n =$ es el tamaño de la muestra.

**Ejemplo 11.5**



Supongamos que registramos el puntaje de un examen y el desempeño de 7 funcionarios de una empresa hotelera, tal como se muestra en la siguiente tabla. Se trata de probar la hipótesis de que no existe correlación entre el examen y el desempeño a un 10% de significancia.

4 Este factor estadístico de correlación de Spearman tiene su propia tabla.

## Tabla 11-7
Resultados pruebas de desempeño

| Ejecutivo | Puntaje examen | Evaluación de desempeño | Clasificación según prueba (*X*) | Evaluación de desempeño (*Y*) | $d_i = X - Y$ | $d_i^2$ |
|---|---|---|---|---|---|---|
| JS | 82 | 4 | 3 | 4 | -1 | 1 |
| AJ | 73 | 7 | 5 | 7 | -2 | 4 |
| DB | 60 | 6 | 7 | 6 | 1 | 1 |
| ML | 80 | 3 | 4 | 3 | 1 | 1 |
| GC | 67 | 5 | 6 | 5 | 1 | 1 |
| AL | 94 | 1 | 1 | 1 | 0 | 0 |
| GW | 89 | 2 | 2 | 2 | 0 | 0 |

Sigamos el procedimiento de la prueba de hipótesis:

1. **Planteamiento de las hipótesis $H_o$ y $H_1$:**

$$Ho : \rho_s = 0$$
$$H1: \rho_s \neq 0$$



2. **Cálculo del factor estadístico:**

El cálculo del coeficiente de correlación de Spearman es como sigue:

$$r_s = 1 - \frac{6\sum d_i^2}{n\left(n^2 - 1\right)}$$

$$r_s = 1 - \frac{6 * 8}{7\left(7^2 - 1\right)} = 0.857$$

3. **Graficamos y calculamos los valores críticos:**

Para el cálculo del valor crítico necesitamos ir a la tabla de Spearman y necesitamos $n$ = 7 y $\alpha$ = 0.10

$\rho_S$ 0.10; 7 en la tabla el valor de $\rho_S$ es: 0.6786

El valor calculado de 0.857 cae en la zona de rechazo, por lo tanto

4. **Decidimos rechazar**

## g) La prueba de Kruskal-Wallis (k)

Esta es una prueba que compara tres o más poblaciones para determinar si existe una diferencia entre la distribución de las mismas. Es análoga a la prueba ANOVA o prueba F.

Su cálculo es el siguiente:

$$k = \frac{12}{n(n+1)}\left[\sum \frac{R_i^2}{n_i}\right] - 3(n+1)$$



En donde:

$n_i =$ es el número de observaciones en la i-esima muestra.

$\eta =$ es el número total de observaciones de todas las muestras.

$R_i =$ es la suma de los rangos de la i-esima muestra.

**Ejemplo 11.6:**

Suponemos que la aerolínea TAME tiene tres clientes importantes, que con frecuencia compran sus ticket en los últimos 7 meses, tal como se muestra en la siguiente tabla:

Tabla 11-8

*Frecuencia de compra tickets*

| | Cliente | | |
|---|---|---|---|
| **Compra** | **1** | **2** | **3** |
| 1 | 28 | 26 | 37 |
| 2 | 19 | 20 | 28 |
| 3 | 13 | 11 | 26 |
| 4 | 28 | 14 | 35 |
| 5 | 29 | 22 | 31 |
| 6 | 22 | 21 | |
| 7 | 21 | | |

Se sigue el procedimiento de la prueba de hipótesis:

1. **Planteamiento de las hipótesis $H_o$ y $H_1$:**



$H_o$ **:** todas las k poblaciones tienen la misma distribución.

$H_1$**:** no todas las k poblaciones tienen la misma distribución.

**2. Cálculo del factor estadístico:**

| Cliente 1 | | Cliente 2 | | Cliente 3 | |
|---|---|---|---|---|---|
| Días | Rango | Días | Rango | Días | Rango |
| | | 11 | 1 | | |
| 13 | 2 | | | | |
| | | 14 | 3 | | |
| 19 | 4 | | | | |
| | | 20 | 5 | | |
| 21 | 6.5 | 21 | 6.5 | | |
| 22 | 8.5 | 22 | 8.5 | | |
| | | 26 | 10.5 | 26 | 10.5 |
| 28 | 13.5 | | | 28 | 13.5 |
| 28 | 13.5 | | | | |
| 29 | 15 | | | | |
| | | | | 31 | 16 |
| | | | | 35 | 17 |
| | | | | 37 | 18 |
| ... | | ... | | ... | |
| | 63 | | 34.5 | | 75 |



$$k = \frac{12}{n(n+1)}\left[\sum \frac{R_i^2}{n_i}\right] - 3(n+1)$$

$$k = \frac{12}{18(18+1)}\left[\frac{(63)^2}{7} + \frac{(34.5)^2}{6} + \frac{(75)^2}{5}\right] - 3(18+1)$$

$$k = 18.62$$

3. **Graficamos y calculamos los valores críticos:**

Para el cálculo del valor crítico necesitamos $\alpha$ = 0.05; g.l.= 2 = $n$-1 ; entonces:

$\chi^2$ 0.05; 2 en la tabla el valor de chi cuadrado es 5.99

El valor calculado de 18.62 cae en la zona de rechazo, por lo tanto

4. **Decidimos rechazar**

## Conclusiones

En esta nota técnica se ha tratado de presentar de manera resumida aspectos relevantes de temas y aplicaciones de la estadística descriptiva e inferencial, en todas las ramas del quehacer empresarial. Se ha dejado para una siguiente nota el análisis multivariable, por ser un tema de mayor profundidad.

Podemos concluir con lo siguiente:

- El cálculo del parámetro poblacional, a partir del dato muestral con los niveles de confianza, permite tomar decisiones bastante sólidas.
- La prueba de hipótesis, de igual manera, es una herramienta de toma de decisiones que implica la aceptación o no de una afirmación que se haga de la población.
- La estimación a través del análisis de regresión –simple y múltiple- es otra herramienta de toma de decisiones con el apoyo de programas de Excel, minitab o SPSS para mirar el comportamiento futuro de una variable.
- El análisis de series de tiempo permite estudiar los movimientos de la variable tomando en cuenta los componentes estacionales, cíclicos, etc. y ajustar la tendencia para entender el comportamiento de aquella variable estudiada.

Los números índices y las pruebas no paramétricas también contribuyen a la toma de decisiones cuando la información con la que se cuenta no se ajusta a una distribución conocida, permitiendo estudiarla y apoyar a la toma de decisiones.

Bladimir Proaño Rivera, economista, docente universitario y analista financiero y de riesgos. Nacido en la ciudad de Azogues, el 9 de octubre de 1968. Hijo de padre Cotacacheño, Guillermo Proaño; y de Madre azogueña, Toha Rivera. Luego de terminar sus estudios universitarios, entró a trabajar en la banca, registrando casi 30 años de vida profesional y experiencia en la industria bancaria, inicialmente como ejecutivo de negocios, tanto de personas como de pequeñas y grandes empresas. Y en los últimos 15 años, como analista de riesgos financieros, sobre todo el de crédito y el de mercado. Y durante 25 años ha vinculado su vida laboral (práctica) con la docencia universitaria (teoría).

ISBN: 978-9942-822-69-7

*Estadística descriptiva e inferencial* es un libro básico de estadística pensado para estudiantes de pregrado y como un texto paralelo o introductorio al complejo campo del análisis estadístico. La obra se ha condensado en dos grandes partes. La primera aborda la estadística descriptiva, es decir incluye definiciones de variables, sus formas de medir. La importancia del muestreo y los tipos de muestreo. También incluye la ilustración de las formas en que se puede organizar y manejar conjuntos de datos para proporcionar una interpretación. Asimismo, medidas para describir datos para el análisis estadístico, medidas de tendencia central y de dispersión. Se incluye unos capítulos intermedios para el análisis probabilístico y sus tipos de distribuciones más usadas. Estos capítulos anteceden a la segunda parte del libro que aborda la parte estadística inferencial, con el análisis de pruebas de hipótesis, análisis de varianza. También incluye el análisis de regresión simple y múltiple, análisis de series de tiempo y construcción de números índices. En cada tema se presenta ejercicios aplicados, y algunos de ellos con aplicaciones en Microsoft Excel.

UNIVERSIDAD DEL AZUAY | Casa Editora