Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 5
ESTADÍSTICA DESCRIPTIVA Y ANÁLISIS EXPLORATORIO

Estadística Descriptiva

- Se utiliza cuando los resultados del análisis estadístico no pretende ir más allá del conjunto de datos investigados.
- Describe numéricamente, analiza y representa un conjunto de datos ordenados mediante la utilización de métodos numéricos, tablas y gráficas, simplificando y resumiendo la información.

Una vez que hemos recogido los datos, tenemos que:

1. Confeccionar tablas acompañadas de gráficos para una mejor visión de los datos.
2. Hacer una recopilación y reducción de dichos datos a unas pocas medidas representativas.
3. Interpretar los resultados y obtener conclusiones para predecir y tomar decisiones estadísticas.

Podríamos pensar en analizar a todos los individuos de la población. Sin embargo, esto puede ser inviable por su costo o por el tiempo que requiere. Entonces nos conformamos con extraer una muestra. La muestra proporciona información sobre el objeto de estudio. Lo habitual en nuestro contexto es que en el procedimiento de extracción intervenga el azar.

Ejemplo 1: Se quiere analizar el número de horas de estudio semanal que dedican los estudiantes de Ingeniería en Informática de esta Facultad. Para ello se consulta a 50 alumnos de esta ingeniería.
Población: Todos los estudiantes de Ingeniería en Informática de esta Facultad.
Variable: Número de horas de estudio semanal.
Muestra: 50 alumnos encuestados.

Más formalmente se denomina muestra de una población original con función F(x), a la sucesión $x_i\, x_2 \ldots\ x_n$ de los valores observables de la variable aleatoria x, que corresponden a n repeticiones independientes de un experimento aleatorio. Se define de forma análoga a la muestra en el caso que el experimento aleatorio esté relacionado con varias variables aleatorias (variable bidimensional, por ejemplo).

Uno de los tipos de muestreo más utilizado es el *muestreo aleatorio simple* (m.a.s.) en el que cada individuo de la población tiene la misma probabilidad de ser incluido en la muestra.

No siempre es necesario tomar una muestra, ya que si queremos estudiar el fracaso de un curso determinado, deberíamos analizar todos los alumnos de dicho curso, y no una muestra de ellos.

Antes de hacer un estudio estadístico, tenemos que plantearnos bien que problema vamos a estudiar, y cuáles van a ser los objetivos de nuestra investigación o trabajo, fijando los pasos a seguir, las clasificaciones que se van a realizar, las variables que debemos observar y cómo medirlas, los gráficos que vamos a realizar.

Los datos que constituyen la muestra son llamados observaciones porque representan lo que se observa actualmente.

Caracteres estadísticos: es una propiedad que permite clasificar a los individuos de una población. Se distinguen dos tipos:

> *a) Cualitativos.* Son aquellos cuya variación se recoge por la presentación de distintas cualidades, es decir, los que no se pueden medir. Ejemplo: estado civil, color de ojos, sexo, profesión de una persona, carrera que piensa elegir un alumno.
> Las modalidades son las diferentes situaciones de un carácter, por ejemplo, las modalidades del carácter profesión podrían ser: ingeniero, economista, psicólogo, informático, periodista …
> *b) Cuantitativos.* Son aquellos que se pueden medir o contar y están formadas por cantidades numéricas

Una observación puede ser numérica o no numérica; las primeras se denominan cuantitativas y las segundas cualitativas. Las cuantitativas se refieren en general a medidas, por ejemplo precios, cantidad de PC en una oficina, etc.; las cualitativas se refieren a clasificaciones; por ejemplo, mediciones buenas o malas, datos confiables o anómalos etc.

También se hará la distinción entre datos de tipo continuo (por ejemplo ganancias, tiempo de ejecución de un programa, km recorridos) y datos de tipo discreto (número de días que falta un empleado de una empresa, número de sectores en un hipermercado).

Las observaciones que se obtienen de forma aleatoria (uno de los métodos de obtención de muestras) y que no se han ordenado de ninguna forma constituyen los datos crudos. Es necesario por lo tanto ordenar, presentar, agrupar y resumir los datos para cumplir con el objetivo de la investigación, estudio, trabajo etc. Desde que se dispone de computadoras, y cada vez más avanzadas, es posible manejar una gran cantidad de datos, pero de todas maneras la organización es siempre necesaria.

Cuando los datos se ordenan de acuerdo a su magnitud lo que se obtiene es una distribución de frecuencias; si se tiene en cuenta el tiempo en que ocurrió ese dato, lo que se obtiene es una serie cronológica, y si lo que se toma en consideración es la ubicación geográfica se obtiene una distribución espacial.

En el ordenamiento de los datos se deberá hacer la distinción entre datos o variables de tipo continuo y discreto.

Si los valores de la muestra se han presentado sólo una vez, se los ordena de acuerdo a su magnitud y la serie se denomina variacional o serie simple. Ahora supóngase que un elemento se encuentra más de una vez en la muestra, ese número de veces que se repite ese valor se denomina frecuencia f_i del elemento x_i. Se denomina serie estadística o serie de datos agrupados a la sucesión de pares $x_i \sim f_i$.

La forma de la distribución de los datos (de una variable) se denomina *distribución de frecuencias*.

El estudio de las distribuciones de frecuencias tiene por objeto la construcción de tablas de frecuencias que podrán utilizarse para una mejor presentación e interpretación de la información contenida en los

datos observados en la muestra. En este apartado, nos referimos a las distribuciones unidimensionales de frecuencias, que son aquellas utilizadas para describir una variable individual sin tener en cuenta la información de otras variables que pudieran haberse incluido en el estudio.

Para poder obtener la forma general de una distribución de frecuencias unidimensional, es necesario introducir algunos conceptos previos.

Consideremos una población estadística de N individuos, descrita según una variable o carácter X, cuyas modalidades han sido agrupadas en un número n de clases, para cada una de esas clases i=1, . . ., n, vamos a definir:

Frecuencia absoluta de la clase: Es el número de observaciones que existen en dicha clase o sea es el número de veces que se repite dicho valor (f_i).

Frecuencia absoluta acumulada de la clase: Es el número de elementos de la muestra cuya modalidad es inferior o equivalente a las de la clase considerada (F_i).

Además se cumple que:

$$F_i = f_1 + f_2 + ... + f_i = \sum_{j=1}^{i} f_j$$

Frecuencia relativa de la clase: Es el cociente entre las frecuencias absolutas de dicha clase y el número total de observaciones o datos que denotamos por N:

$$h_i = \frac{f_i}{N}$$

Si estamos interesados en trabajar con porcentajes, sólo tenemos que multiplicar la frecuencia relativa por 100 y así representamos el porcentaje (%) de la muestra que comprende a esa clase.

Frecuencia relativa acumulada de la clase: es el número de elementos de la población que están en alguna de las clases inferior o igual a la clase.

$$H_i = \frac{F_i}{N}$$

Como normalmente el conjunto de datos que se recolecta suele ser muy grande, es necesario disponer de alguna herramienta mediante la cual podamos visualizarlos. Para ello, una vez ordenados, hacemos un recuento de dichos datos y realizamos tablas estadísticas. En estas tablas, deberán figurar los valores de la variable en estudio, y sus frecuencias correspondientes. Si bien este ordenamiento puede evitarse al trabajar con programas específicos o alguno que posea este tipo de análisis, es útil para la realización de algunos gráficos.

La principal dificultad para la obtención de una distribución de frecuencias, reside en la construcción de las modalidades, ya que ésta variará de acuerdo con el tipo de variable que se pretende describir: si la variable es cualitativa, se tomarán como modalidades las distintas respuestas observadas de la muestra; si la variable es discreta (que tome pocos valores distintos),

las modalidades coincidirán con los distintos valores medidos en la muestra; si la variable es continua (o bien discreta, pero toma muchos valores distintos), se tomarán como modalidades intervalos de clase. Son los intervalos donde se encuentran los datos agrupados, se simbolizan por $[L_{i-1}, L_i)$.

Gráficos

Una de las herramientas más populares y utilizada dentro de la estadística descriptiva es, sin lugar a dudas, el análisis gráfico de los datos. Las tablas estadísticas, resumen los datos que disponemos sobre una muestra y dan toda la información necesaria, pero como se suele decir, *"Una imagen vale más que mil palabras"*, es conveniente expresar la información que disponemos mediante un gráfico o diagrama, con el fin de hacerla más clara y captar de un solo vistazo las características de los datos.

Gracias a la informática y los programas que se han desarrollado se pueden realizar fácilmente todo tipo de representaciones gráficas y de gran calidad.

Gráficos para variables cualitativas o atributos

Diagrama de barras o bastones

En este tipo de gráficos se representan en el eje de abcisas (X) las diferentes modalidades de la variable y en el eje de ordenadas (Y) la frecuencia relativa o absoluta.

Este tipo de gráficos también se puede hacer en el espacio, incorporando una nueva variable (Z) y realizando un dibujo tridimensional.

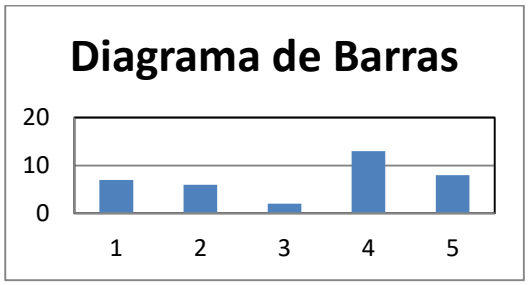


Figura Nº 1 – Diagrama de Barras

Diagramas de sectores

Se utilizan para hacer comparaciones de las distintas modalidades de un carácter mediante sectores circulares. Para construirlos se divide un círculo en tantas porciones como modalidades existan de manera que el ángulo central de cada sector ha de ser proporcional a la frecuencia absoluta o relativa correspondiente.

Este tipo de diagramas recibe también el nombre de *tartas* o *tortas*, por la forma que tiene su representación.

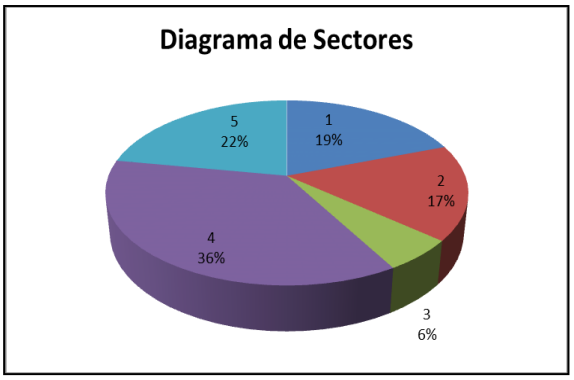


Figura Nº 2 – Diagrama de Sectores

Pictogramas

Para realizarlos se representan a diferentes escalas un mismo dibujo teniendo en cuenta que el perímetro del dibujo tiene que ser proporcional a la frecuencia, pero esto puede incurrir en un efecto visual engañoso ya que a frecuencia doble corresponde un dibujo de área cuádruple, con lo cual tiene un inconveniente debido a la falta de precisión.

A pesar de este inconveniente este tipo de dibujos son muy utilizados por los medios de comunicación a la hora de hacer que el público no especializado comprenda temas complejos sin necesidad de dar una explicación complicada.

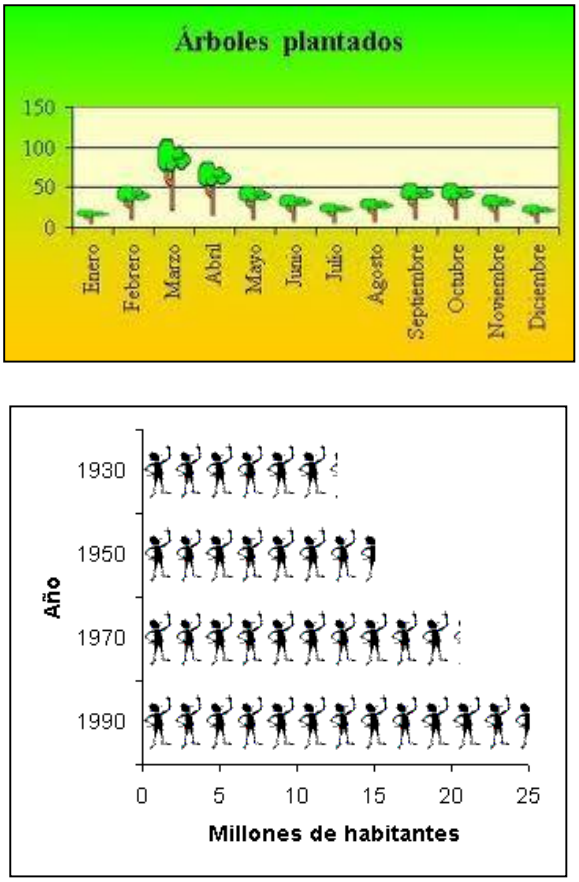


Figura Nº 3 – Pictogramas

Gráficos para Variables Cuantitativas

Para este tipo de variables, tenemos diferentes gráficos según el tipo de frecuencia que usemos y además tenemos que tener en cuenta si la variable es discreta o continua.

Gráficos para variables cuantitativas discretas

Diagrama de barras

Su representación es idéntica a la explicada para variables cualitativas, las barras deben de ser estrechas para mostrar que los valores que toma la variable son discretos.

Gráficos para variables cuantitativas continuas

Histograma

Es una manera sencilla de representar una gran masa de datos y debe ser el comienzo de cualquier estudio más sofisticado y en el que pueden observarse tres propiedades esenciales de una distribución: forma, tendencia central o acumulación y dispersión o variabilidad.

Este se obtiene por graficar en el eje x las clases y en el eje y las frecuencias. La altura de las barras del histograma tienen distinta significación según el ancho de clase sea constante o no. En el primer caso se representan frecuencias, o sea la cantidad de valores en cada clase; en el segundo caso densidad de frecuencias, o sea es el promedio, en cada clase, de cuántos valores hay por unidad de ancho de clase: $f_i / c_i = h$ **.** En este caso el área de cada rectángulo es proporcional a la frecuencia.

A diferencia del diagrama de barras, los rectángulos verticales, se representan contiguos para reflejar la idea de que la variable es continua. La forma del histograma refleja propiedades importantes de la variable estadística a la que se refiere.

El número de clases o intervalos y la longitud que se consideran, depende de cada problema y de la utilización que se quiera dar a las tablas estadísticas. Lo normal es que todos los intervalos sean de la misma amplitud ($L_i - L_{i-1}$), aunque pueden existir múltiples razones donde se aconseje tomar intervalos de amplitud variable, como puede ser el caso en el que existan uno o dos intervalos donde se concentren la mayoría de los datos.

La construcción de los intervalos de clase, introduce algunas cuestiones subjetivas, como son:
1) *¿Cuántos intervalos construir?*
Aunque no existe una regla general para usar, es evidente que el número de intervalos debe ser mayor al aumentar el tamaño muestral, lo ideal entre 5 y 20.
2) *¿Qué valor se elige como extremo inferior del primer intervalo L_0 ?*
Se toma como L_0 un valor "un poco menor" que el mínimo de la muestra (o el mínimo).

Es muy importante hacer una buena elección de la cantidad de clases a utilizar. Para este fin se utilizan distintas reglas, una de ellas consiste en tomar el número de clases igual al entero más próximo a la raíz cuadrada del número de observaciones que se estudian, $\sqrt{N}$ y no ser inferior a 5 ni superior a 20, ya que en el primer caso se produciría una concentración de datos que no sería representativa

de la muestra, y en el segundo caso podrían quedar intervalos vacíos, en los cuales no habría ningún valor.

Consejos:

1. Usar intervalos de la misma longitud
2. Los intervalos no pueden solaparse
3. Cada observación sólo puede pertenecer a un intervalo
4. Todos los datos deben pertenecer a algún intervalo
5. La forma del histograma depende de la amplitud del intervalo que se elija.

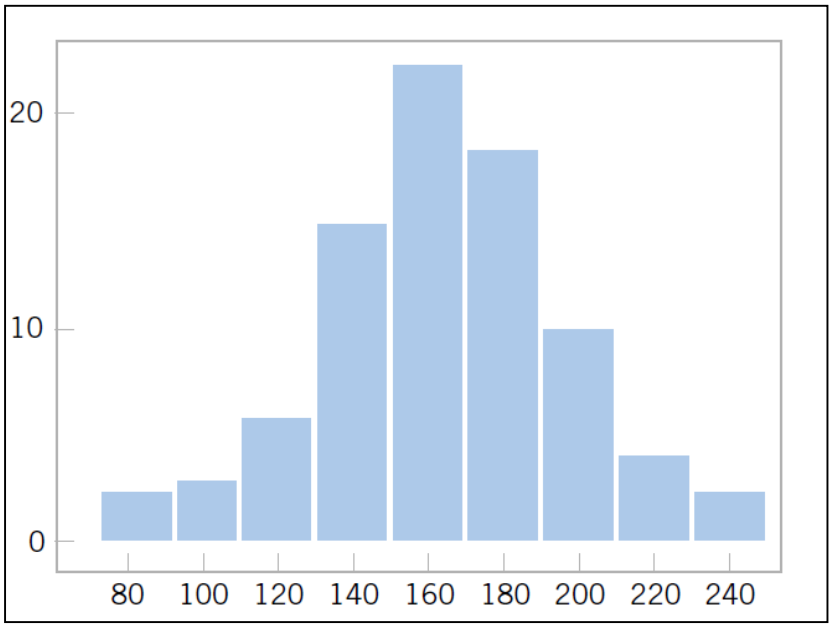


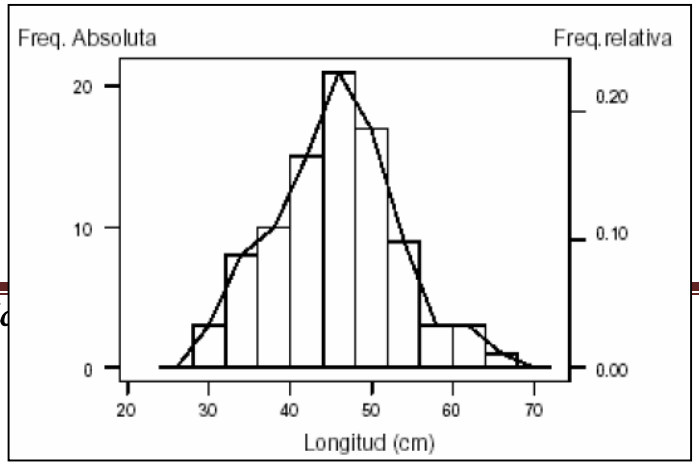
Figura Nº 4 - Histograma

Polígono de frecuencias

Se construye fácilmente una vez representado el histograma, y consiste en unir los puntos del histograma que corresponden a las marcas de clase de cada intervalo mediante una recta.

El diagrama, para variables continuas, se denomina *polígono de frecuencias acumulado u ojiva*. En estos polígonos obtenidos se aprecian con claridad propiedades importantes, da idea aproximada de qué curva teórica le correspondería a la población de la cual se obtuvo la muestra.

Si las frecuencias se expresan como proporciones - es decir, divididas por el total de observaciones en la muestra lo que se obtiene es una distribución de frecuencias relativas. Cuando se realice el histograma en este caso**,** el área total de las barras será igual a 1.



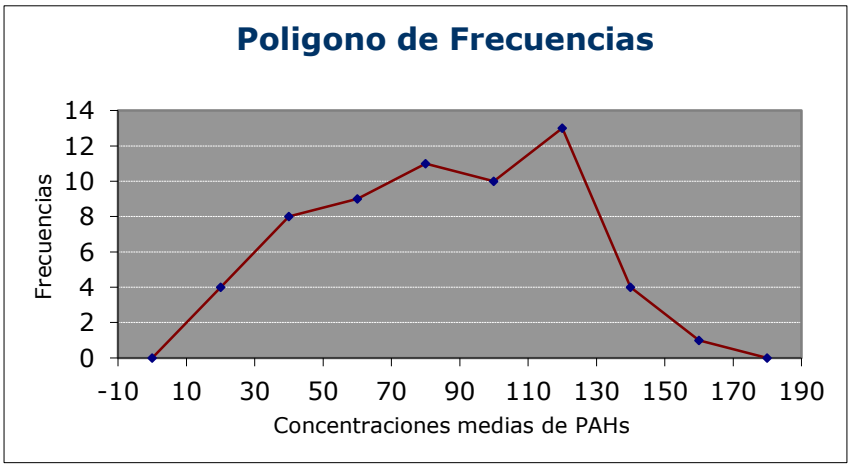
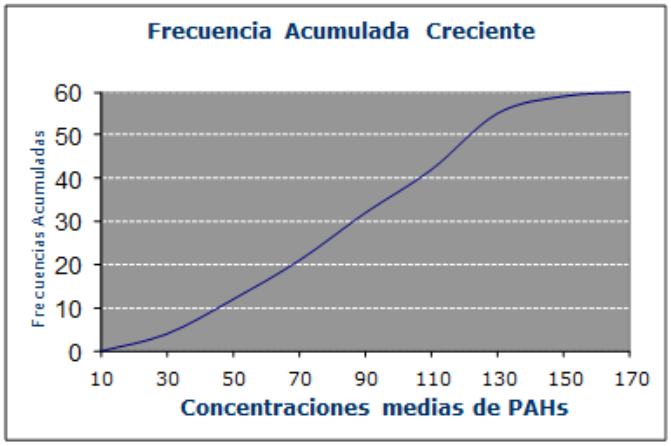
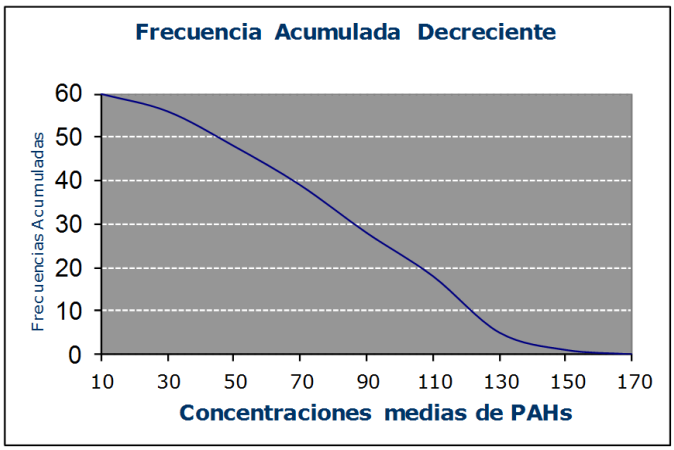


Figura Nº 5 – Polígono de Frecuencias

Polígono de frecuencias acumuladas: se utilizan en variables continuas. El eje de abscisas se construye igual que en los histogramas, pero en el de ordenadas se incluyen las frecuencias acumuladas, ya sean absolutas o relativas. Sobre cada límite se levanta una perpendicular cuya longitud sea idéntica a la frecuencia acumulada y se unen los extremos superiores de dichas perpendiculares.





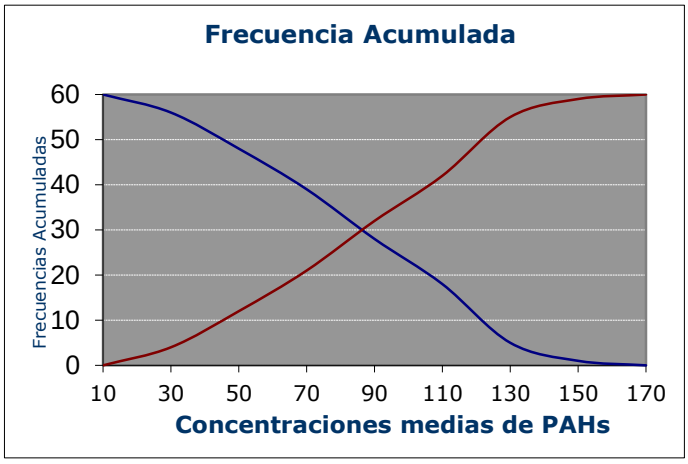


Figura Nº 6 – Curvas de Frecuencias Acumuladas

Todos estos gráficos y tablas se pueden realizar de una manera rápida y sencilla con ayuda de los softwares disponibles tanto planillas de cálculo como específicos de estadística.

ANÁLISIS EXPLORATORIO DE DATOS

Un análisis reciente que ha impulsado la estadística se debe al esfuerzo de John Tukey, quién ha producido una gran cantidad de métodos innovadores para el análisis de datos. En su libro "Exploratory Data Analysis" de 1977 y en otras publicaciones recientes (Hoaglin, Mosteller y Tukey, 1983) , Tukey ha expuesto una filosofía práctica para el análisis de datos. La escuela de Tukey se ha extendido en los últimos años (Breckenridge, 1983) haciendo énfasis en la exploración de los datos por métodos gráficos previos al clásico análisis estadístico tradicional. La visualización de los datos permite al investigador penetrar en su estructura, minimizando los supuestos probabilísticos que tradicionalmente se asumen con respecto a su comportamiento y

distribución. Lo anterior equivale a proporcionarle al investigador "una lente" de aumento que le permite:

- Exhibir características o patrones ocultos dentro de los datos.
- Resaltar con claridad la tendencia que conforman los datos.
- Proporcionar hipótesis o modelos acerca del comportamiento de los datos

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas.

El AED proporciona métodos sencillos para organizar y preparar los datos, detectar fallos en el diseño y recogida de datos, tratamiento y evaluación de datos ausentes, identificación de casos atípicos y comprobación de los supuestos subyacentes en la mayor parte de las técnicas

El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

Cabe mencionar que esta parte de la Estadística se ha robustecido con la reciente aparición de diversos programas como por ejemplo Statgraphics, Statistica, SPLUS, etc

Algunas de las herramientas más importantes son:

-El diagrama de tallo y hoja.

-El diagrama de caja.

- Las profundidades.

-El diagrama de letras.

-Las transformaciones matemáticas. -Las suavizaciones.

-Las series de tiempo.

Diagrama de Tallo y Hoja

El objetivo del **diagrama de tallo y hojas** es mostrar la frecuencia con la que ocurren los valores dentro de un conjunto de datos lo cual es muy parecido a lo que hace un histograma pero la diferencia es que en el diagrama de tallo y hojas no se observan barras sólidas sino que son los mismos números los que dan forma al diagrama.

Puede definirse como un híbrido que combina los aspectos visuales del histograma con la información numérica que proporciona una tabla de distribución de frecuencias.

Este diagrama se construye colocando en una columna todos los números que conforman los datos eliminando la última cifra, es decir las unidades. Esta columna debe ordenarse de menor a mayor.

A la derecha de cada número se escribe la última cifra o unidad de cada dato que comienza con ese número. Luego se ordenan de menor a mayor los números de cada fila.

Cada valor se subdivide en tres componentes el más significativo, o sea el situado más a la derecha, se usa para formar el tallo, el segundo en significación forma la hoja, que servirá para generar un histograma y con ello proporcionar una idea de la forma de la variable, y el tercero, si existe, que es el menos significativo, se puede despreciar.

Cuando existen valores muy separados del conjunto, se puede simplificar el gráfico eliminando las filas sin hojas e indicando los valores altos o bajos completos precedidos de esas palabras, ALTOS, si son muy elevados, o BAJOS si se da la circunstancia contraria.

La elaboración de un gráfico de tallo y hojas es muy sencilla, y se puede considerar como la técnica de representación gráfica recomendable para variables cuantitativas, por encima de otra forma muy usual como el histograma.

Se construye de la siguiente manera:

1. Ordenar el lote de datos en magnitud creciente.

2. Seleccionar un par conveniente de dígitos que permita fraccionar en dos partes el lote de datos según la característica de los datos o lo que se quiere mostrar.

3. Formar el tallo y las hojas con las fracciones respectivas.

4. Construir el tallo escribiendo verticalmente los dígitos enteros entre el 22 y 31, asociando a cada uno su hoja respectiva. Los dígitos del tallo están separados de los dígitos de la hoja por medio de una línea vertical.

En términos generales un diagrama de esta naturaleza hace visibles las siguientes características:

1. Muestra el rango de valores que los datos cubren.

2. Determina donde se concentran la mayoría de los datos

3. Describen la simetría del conjunto de datos.

4. Identifica si existen huecos en la distribución de los datos.

5. Señala aquellos valores que claramente se desvían del conjunto de datos.

Otra opción que presenta el diagrama de tallo y hoja es la comparación entre dos lotes de datos, aspecto que no considera el histograma. A esta derivación se le llama diagrama de tallo y hoja en espejo.

La observación de cualquiera de estos gráficos, el histograma o el diagrama de tallo y hoja, permite extraer ideas de las características generales de la variable representada.

```
  1   0|5
  7   0|666777
 26   0|8889999999999999999
(20)  1|00000000001111111111
 43   1|2233333
 36   1|44444444444555555
 19   1|6666666677
  9   1|889
  6   2|000
  3   2|22
```

7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

Figura Nº 7 – Diagrama de Tallo y Hojas

Gráfico de Caja y Bigote - Box Plot

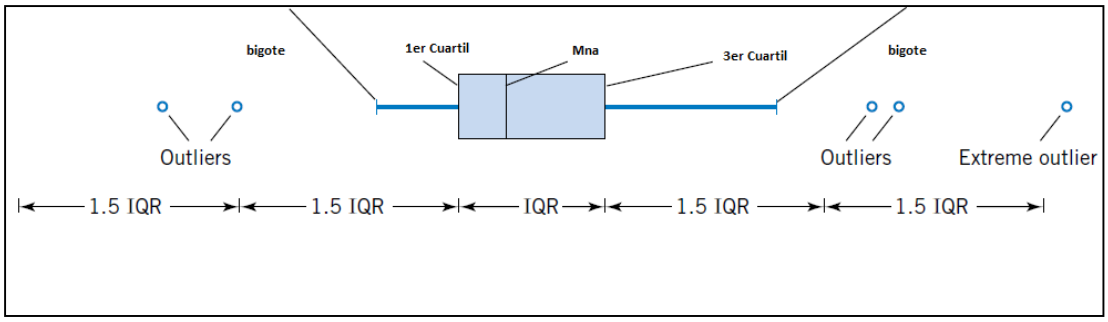
Un diagrama de caja y bigotes es una gráfica basada en cuartiles, que ayuda a visualizar un conjunto de datos.

Para construir un diagrama de caja se necesita el valor mínimo, el primer cuartil, la mediana, el tercer cuartil, y el valor máximo.

El cuadro encierra el rango intercuartil con el borde izquierdo (o más bajo) en el primer cuartil, Q_1, y el borde derecho (o superior) en el tercer cuartil, Q_3.

Se traza una línea a través de la caja en el segundo cuartil (que es el porcentil 50 o la mediana); una línea, o el bigote, se extiende desde cada extremo de la caja. El bigote inferior es una línea desde el primer cuartil hasta el punto de datos más pequeño dentro de 1,5 rango intercuartil. El bigote superior es una línea desde el tercer cuartil hacia los valores más grandes a 1,5 rango intercuartil. Datos que estén más lejos de estos bigotes se representan como puntos individuales. Un punto más allá de un bigote, pero menos de 3 rango intercuartil desde el borde de la caja, es

llamado un caso atípico. Un punto más allá de 3 rangos intercuartil del borde de la caja se llama un extremo atípico. Diferentes símbolos, tales como círculos abiertos y llenos, se utilizan en ocasiones para identificar los dos tipos de valores atípicos.



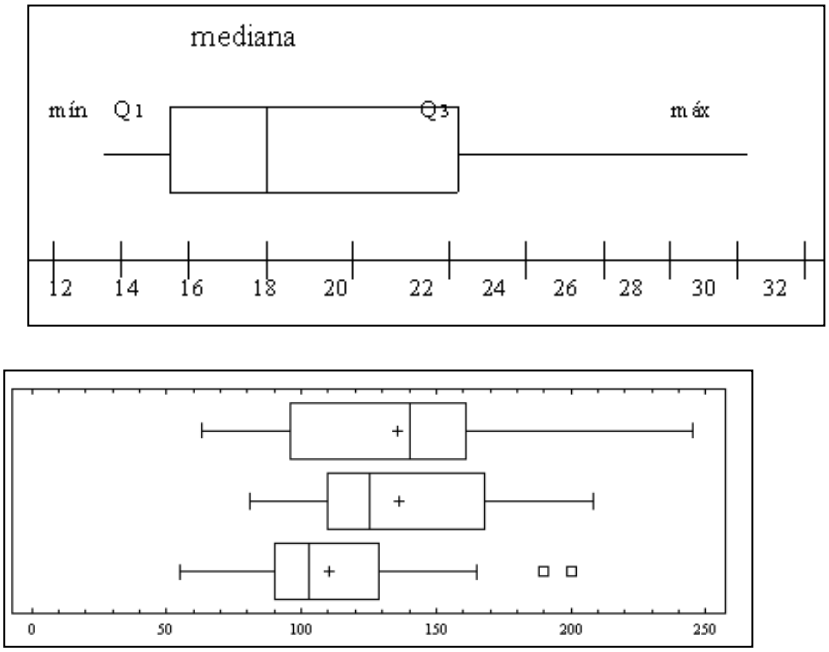
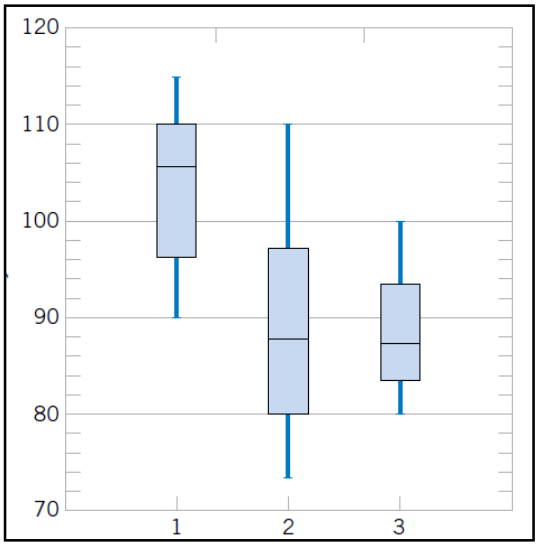


Figura Nº 8 – Diagrama de Caja y Bigote

Es posible introducir algunas variaciones en la construcción de estos diagramas, dependiendo del tipo de estudio y de la información disponible. La caja o rectángulo contiene un porcentaje de la muestra y puede construirse con diferentes rangos de variación.

Los diagramas de caja son muy útiles para realizar comparaciones gráficas entre los conjuntos de datos, debido a que tienen alto impacto visual y son fáciles de entender como se muestra en la figura siguiente.



Si la mediana está ubicada como sigue

entonces la distribución es asimétrica negativa. (izquierda)

Si la mediana está de esta manera

Entonces la distribución es asimétrica positiva. (derecha)

Todo este análisis se puede realizar también con la ayuda de softwares muchos de los cuales tiene incorporado sino todo parte de este análisis.

CARACTERÍSTICAS de muestra

Además de organizar los datos y mostrarlos en gráficos, se necesita de ciertas medidas representativas que puedan resumir una gran cantidad de ellos.

Estos números que sirven para caracterizar las **distribuciones de frecuencias** de datos univariados pueden resumirse en aquellos que tienen en cuenta las cuatro propiedades básicas:

1-Ubicación del centro de la distribución, comúnmente llamadas **MEDIDAS DE IA TENDENCIA CENTRAL**

2-Variación de las observaciones alrededor del punto central. Estas son conocidas como **MEDIDAS DE DISPERSIÓN**

3-Grado de asimetría conocida como **MEDIDAS DE ASIMETRIA**

4-Grado de variación en altura de una distribución respecto de un modelo o patrón **MEDIDAS DE CURTOSIS**

MEDIDAS DE LA TENDENCIA CENTRAL

Entre estas medidas se diferencian los llamados *promedios* y las *medidas de ubicación.* El más comúnmente usado es la **media aritmética**; otros menos usados y útiles en algunas circunstancias especiales son la media geométrica y la media armónica.

Entre las medidas de ubicación se consideran la **mediana,** el **modo**, los **cuantiles**: **deciles, cuartiles, porcentiles.**

Promedios

La **media aritmética** se define como el promedio de todos los valores de la muestra:

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

Es uno de los valores de la tendencia central más usados y el más simple de calcular.

La media aritmética tiene algunas propiedades matemáticas:

Es el *centro de gravedad, un punto de equilibrio.* De su expresión se puede ver que:

$$n.\bar{x} = \sum x_i$$

Esto significa que concentra en su valor toda la información que hay en la muestra.

La suma de las desviaciones respecto a este valor es igual a cero.

$$\sum (x_i - \bar{x}) = 0$$

$$\sum x_i - n\bar{x} = 0$$

La suma de las desviaciones cuadradas de los datos con respecto a la media es menor que si estas desviaciones se toman con respecto a cualquier otro valor. Se demuestra que:

$$\sum (x_i - \bar{x})^2 = mínimo$$

Debido a esta propiedad la media aritmética se emplea como base de las medidas de dispersión.

La media queda fuertemente afectada por los valores extremos, y por esto puede que en algunos casos no sea representativa.

La media puede tratarse algebraicamente, esto es: si se tiene la media de subgrupos puede obtenerse la media general promediando estas medias; si el número de elementos de cada subgrupo no es el mismo se efectuará un promedio ponderado por la cantidad de elementos en cada grupo:

$$\overline{X} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \bar{x}_n N_n}{N} \qquad (2)$$

$\bar{x}_1, \bar{x}_2 \ldots\ldots \bar{x}_n$ s on las medias de cada grupo,
N_i, **...,** N_n es la cantidad de elementos de cada grupo; $N = N_i$ + **...** + N_n es la longitud total de la muestra.

Media Geométrica: se define como la raíz n-ésima del producto de los elementos que conforman la muestra.

$$Gm = \sqrt[n]{\Pi\, x_i} \qquad (3)$$

Es apropiada para promediar razones, porcentajes o velocidades de cambio. Para facilitar su cálculo se pueden aplicar logaritmos a la expresión anterior.

$$\log G_m = \frac{1}{n}(\log x_1 + \log x_2 + \ldots\ldots + \log x_n)$$

Puede verse de esta expresión, que el logaritmo de la media geométrica es igual al promedio del logaritmo de los valores de las observaciones.

Si la variable presenta frecuencia:

$$G_m = \sqrt[n]{\prod x_i^{fi}}$$

$$\log G_m = \frac{1}{n}(\log x_1 \ f_1 + \log x_2 \ f_2 + ... + \log x_n \ f_n)$$

Tiene las siguientes propiedades:

-está menos afectada por valores extremos.

-para cualquier serie es siempre menor que la media aritmética.

-es muy útil en el cálculo de *números índice.*

-se puede manipular algebraicamente.

-no es muy conocida y no puede evaluarse cuando hay datos negativos o ceros.

Media armónica: Se define como la inversa de la media aritmética de las inversas de los valores muestrales. Es apropiada para el procesamiento de datos de razones que tienen dimensiones fisicas como *km/I, producción/hora,* etc. Sus expresiones para los diferentes casos son:

$$\frac{1}{Hm} = \frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{N} \qquad (4)$$

Observaciones sobre la media Geométrica y la media Armónica

El empleo de la media geométrica o de la armónica equivale a una transformación de la variable en log x ó 1/x , respectivamente, y el cálculo de la media aritmética de la nueva variable; por ejemplo, si la variable abarca un campo de variación muy grande, tal como el porcentaje de impureza de un producto químico, por lo general alrededor del 0.1%, pero que en ocasiones llega incluso al 1% o más, puede ser ventajoso el empleo de log x en lugar de x para obtener una distribución más simétrica.

Relación entre las medias:

$$\boxed{H \leq G \leq \overline{X}}$$

Medidas de ubicación

Modo: Es un valor muy útil en la descripción de la muestra. Se lo define como el valor de la variable que aparece más veces que otro; se puede decir que es el valor con mayor frecuencia.

Si la variable analizada es continua se puede obtener con alguna de las siguientes expresiones:

$$Mo = L_{iMo} + \frac{d_1}{d_1 + d_2} c \qquad (5)$$

Li_{Mo} = límite inferior del intervalo que contiene al modo.
d_1:diferencia (sin tener en cuenta el signo) entre las frecuencias del intervalo que contiene el modo y el intervalo anterior.
d_2:diferencia entre las frecuencias del intervalo que contiene al modo y a del intervalo posterior.
c = ancho de clase.

$$Mo = L_{iMo} + \frac{f_1}{f_1 + f_2} c \qquad (6)$$

f_1=frecuencia del intervalo de clase anterior al modal.
f_2=frecuencia del intervalo de clase posterior al modal.

Es un valor muy inestable, ya que puede cambiar con el método de redondeo de los datos. Puede determinárselo gráficamente a partir del histograma interpolando en la barra más alta.

Mediana: Es el valor de la serie para el cual el 50% de los valores son menores y el 50% mayores o iguales. Se la denomina valor medio de la serie.

Tenemos que diferenciar entre:

Variables discretas:

- Si el número de datos es impar Mna. es el valor central.
- Si el número de datos es par Mna. es la semisuma de los valores centrales.

si hay frecuencias:

- Se calcula N/2 y se obtiene Ni (frecuencias acumuladas)
- Se observa cual es la primera Ni que supera o iguala a N/2, distinguiéndose dos casos:
  - Si existe un valor de x_i tal que $N_{i-1} < N/2 < N_i$, entonces se toma como

$Mna. = x_i$

- Si existe un valor i tal que N_i=N/2 entonces la mediana será $Mna. = \frac{x_i + x_{i+1}}{2}$

Variables continuas

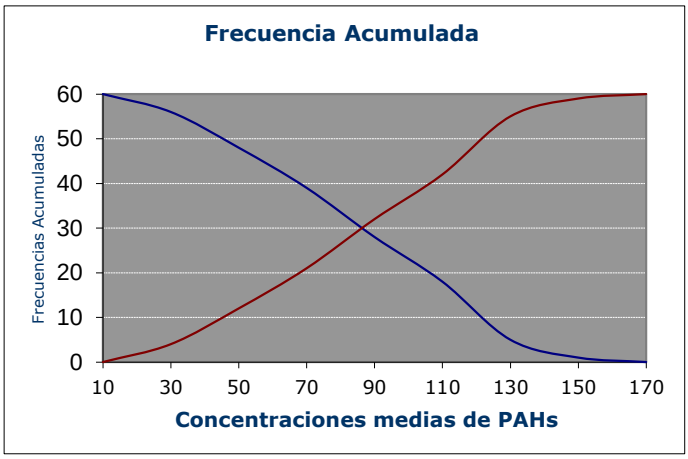
$$Mediana = L_i + \frac{N/2 - FL_i}{f_i} c \qquad (7)$$

siendo:

Li=límite inferior del intervalo que contendrá a la mediana.
FLi=frecuencia acumulada hasta el intervalo anterior a la clase mediana
fi=frecuencia del intervalo mediano.

También se la puede obtener a través de las ojivas, en su intersección o en el valor correspondiente al 50%.

Gráficamente en la intersección de las ojivas.



Propiedades

-No está influenciada por valores extremos. Por lo tanto, es una medida conveniente de la ubicación central.

-Un valor seleccionado a azar se ubicará por arriba o por debajo de ella con igual probabilidad; por esto suele llamársela valor probable.

-Su cálculo es fácil.

Algunas desventajas son:

-los datos deben ordenarse para su cálculo.

-No se la puede manipular algebraicamente.

-No es tan usada como la media aritmética, y tiene mayor error que ella.

Cuantiles (cuartiles, deciles, percentiles): son también medidas de ubicación. Como la mediana divide a la distribución de datos en dos partes, los cuartiles la dividen en cuatro, los deciles en diez y los percentiles en cien. Estas medidas posibilitan un análisis más minucioso de la distribución.

Los cuartiles dividen a la distribución en cuatro partes; por lo tanto hay tres cuartiles. El caso de los percentiles se usa cuando existen muchas observaciones.

Se calculan de la misma forma que la mediana, sólo que cambia come se determina el orden del cuantil. Por ejemplo para ubicar el primer cuartil se hace N/4, para el segundo 2N/4 y así para el resto, el cálculo es luego similar al realizado para la mediana. De la misma forma se precede para calcular los deciles y percentiles.

Se define el cuantil p como el número que deja a su izquierda una frecuencia relativa p. Lo que es lo mismo, la frecuencia relativa acumulada hasta el cuantil p es p. Claro está que los cuantiles sólo se podrán calcular con variables ordinales. Nótese que la mediana es el cuantil 0o5. Para calcular los cuantiles seguiremos las siguientes indicaciones.
Si la variable es discreta, o si es continua y disponemos de todos los datos:
Ordenamos la muestra. Tomamos el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor o igual que p. Si se supera p estrictamente, este dato ya es el cuantil p; mientras que si se alcanza con igualdad, el cuantil p es la media de este dato con el siguiente.
Ejemplo 8: Muestra ordenada: 1, 2, 3'5, 6, 7, 9, 12, 13, 14'5, 15'2.
Cuantil 0'10=1'5; Cuantil 0'43=7.

Si la variable es continua y se encuentra agrupada en intervalos de clase:
Buscamos sobre la tabla de frecuencias el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que p (pensemos que es el intervalo $[L_i ; L_{i+1})$) y, dentro de ese intervalo, calculamos el cuantil p por interpolación lineal, esto es:

$$\text{Cuantil } p = L_i + \frac{p \cdot n - N_{i-1}}{n_i}(L_{i+1} - L_i)$$

$$Cuantil_i = L_{i-1} + \frac{i\left(\frac{N}{p}\right) - F_{i-1}}{f_i} c_i$$

con

p=4 cuartil, p=5 quintil, p=10 decil, p=100 percentil o centil.

Primero se calcula i.(N/p), y se mira a ver en qué intervalo cae el número que salga, y es con este intervalo en el que nos tenemos que fijar para usar F_{i-1}, f_i,… c_i es la longitud de ese intervalo).

Algunos órdenes de los cuantiles tienen nombres específicos. Así los cuartiles son los cuantiles de orden (0.25, 0.5, 0.75) y se representan por Q_1, Q_2, Q_3. Los deciles son los cuantiles de orden (0.1, 0.2,..., 0.9). Los percentiles son los cuantiles de orden j/100 donde j=1,2,...,99.

MEDIDAS DE DISPERSION

Permiten *calcular la representatividad de una medida de posición*, para lo cual es preciso cuantificar la distancia entre los diferentes valores de la distribución respecto a dicha medida. A esta distancia es a lo que se denomina ***variabilidad o dispersión de la distribución***.

La finalidad de estas medidas es estudiar *hasta qué punto para una determinada distribución de frecuencias, las medidas de tendencia central o de posición son representativas* como síntesis de toda la información de la distribución.

La medida más simple de variación o dispersión es la amplitud que se considera como la diferencia entre los valores mínimo y máximo de la serie, también conocida come *rango*. También puede obtenerse la amplitud media como el promedio de los valores extremos.

Puede obtenerse además la *amplitud intercuantílica*, es considerar las distancias entre cuartiles, deciles y percentiles. Por ejemplo para los cuartiles se calculan el primero y el tercero y se restan sus valores, de la misma forma se precede con los deciles y percentiles.

Interesa fundamentalmente poder determinar una medida de variabilidad que involucre a todos los valores contenidos en la muestra, para esto se puede considerar el promedio de las desviaciones de los valores respecto a la media aritmética

$$\frac{\sum(x_i - \bar{x})}{n}$$

pero como se vió:

$$\sum(x_i - \bar{x}) = 0$$

para evitar esto es que se toman las desviaciones sin considerar su signo y se obtiene la desviación media:

$$\frac{\sum|(x_i - \bar{x})|}{n}$$

En vez de usar valores absolutos, se pueden tomar cuadrados, ya que es más sencillo trabajar con ellos. El promedio de estos desvíos respecto a la media se denomina Varianza de la muestra y se simboliza S^2

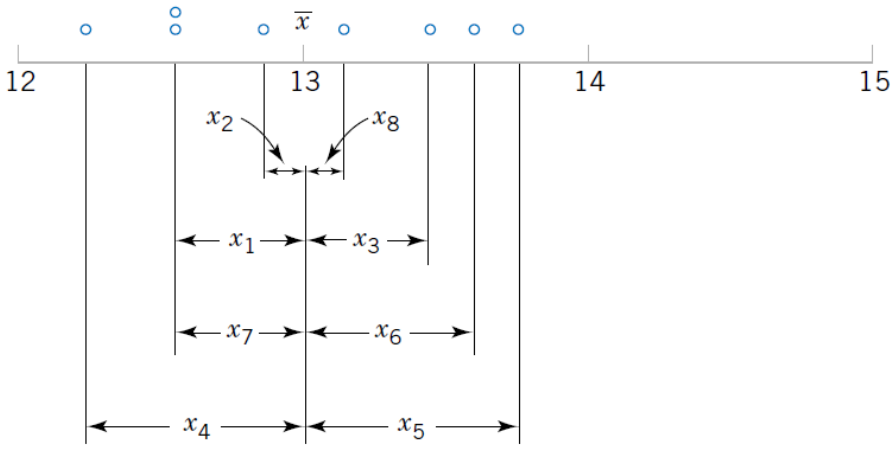
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n} \qquad (8)$$

también se la suele expresar como:

$$S = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad (9)$$

Cuando la muestra es grande la diferencia entre ambas expresiones es despreciable y se utiliza la primera.

El promedio de las diferencias al cuadrado respecto a la media se denomina momento centrado de orden 2: **m_2**



La raíz cuadrada de la varianza se denomina desvío estándar de la muestra y se lo suele preferir a la varianza ya que tiene las mismas unidades que la variable. Para simplificar su cálculo se desarrolla el cuadrado de su expresión y se consigue la expresión en función de los momentos muestrales respecto al origen (recuérdese población):

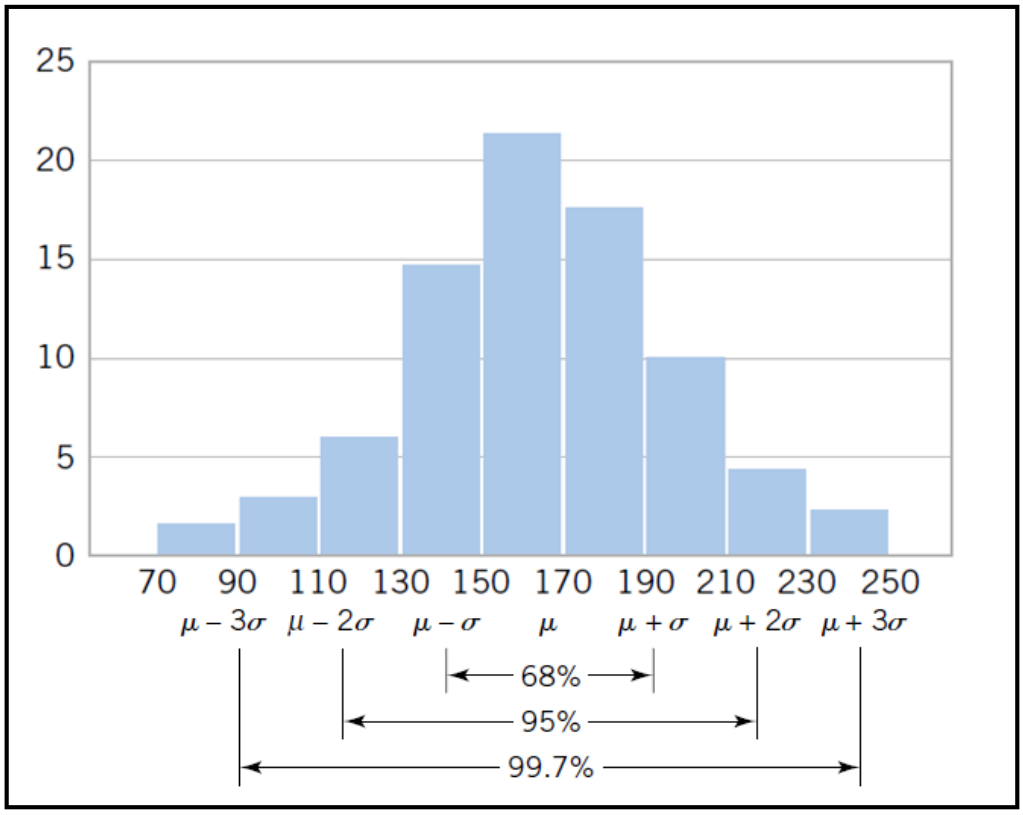
$$S^2 = \frac{\sum x_i^2}{n} - \bar{x}^2$$

$$S^2 = a_2 - a_1^2 \qquad (10)$$

Estas medidas (varianza y desvío) son las medidas de variabilidad más empleadas. Pero si se halla

sólo el desvío o varianza de un conjunto de valores, no se puede asegurar si representa un valor alto, medio o bajo de variabilidad.

Se utiliza por esto una regla empírica para interpretar los valores de la varianza o desvío, se usará cuando la muestra sea grande y la forma de la muestra sea aproximadamente de campana; esta regla considera que si se miden en el eje x y hacia ambos lados de la media una distancia igual al desvío, en ese intervalo quedarán comprendidos el 68% de las observaciones. Si se traza dos veces el desvío hacia ambos lados de la media quedarán comprendidos el 95% de las observaciones en ese intervalo, y si se trazan tres veces el desvío quedarán comprendidos el 99% de las observaciones entre esos límites.



Estas medidas descriptas son en valor absoluto. Si se comparan las variabilidades de dos conjuntos de datos en base a estas medidas anteriores, puede darse una respuesta cierta sólo si las medias (respecto a las cuales se hallan los desvíos) son aproximadamente iguales. Cuando esto no sucede, o bien tienen unidades diferentes, se recurre a medidas de variabilidad relativas para independizarlas de las unidades de medida. Una muy usada es el coeficiente de variabilidad o coeficiente de Pearson, que para la muestra es:

$$Cv = \frac{S}{\bar{x}} * 100 \qquad (11)$$

normalmente expresado en porcentaje. Por ejemplo si se quiere comparar la variabilidad de los caudales del río Paraná y los del arroyo Saladillo, la desviación o varianza no sirven ya que los valores medios son muy diferentes; en este caso es conveniente calcular este coeficiente y

comparar los porcentajes obtenidos.

MEDIDAS DE FORMA

ASIMETRÍA

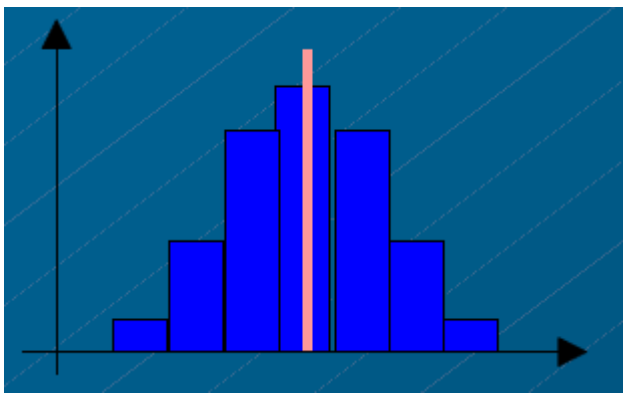
Cuando la distribución es simétrica la media, mediana y el modo coinciden. Cuando es asimétrica esos valores difieren. Come se ha visto la media aritmética es el valor de la tendencia central más afectado por los valores extremos, es por esto que cuanto mayor sea la distancia entre la media y el modo mayor será el grado de asimetría. Esta diferencia entre media y modo se suele usar como medida de asimetría ya que cuanto mayor sea esta distancia mayor será la asimetría. A los fines de comparar la asimetría entre dos distribuciones y salvar la diferencia de unidades y la diferencia en las dispersiones es que se divide por el desvío. Debido a que el modo se encuentra aproximadamente algunas veces, se prefiere trabajar con la mediana que se encuentra mejor y teniendo en cuenta la relación vista entre la media, mediana y modo, se pueden obtener las siguientes expresiones para la asimetría:

$$As = \frac{(\bar{x} - Modo)}{S} \qquad As = \frac{3(\bar{x} - Mediana)}{S} \qquad (12)$$

Este valor seria aproximadamente igual a 0 para una distribución simétrica, positivo para una distribución asimétrica hacia la izquierda y negativo para una distribución asimétrica hacia la derecha.

El valor exacto de la Asimetría está dado por el momento centrado de tercer orden adimensionalizado:

$$m_3 = \frac{\sum (x_i - \bar{x})^3}{n} \qquad As = \frac{m_3}{S^3} \qquad (13)$$



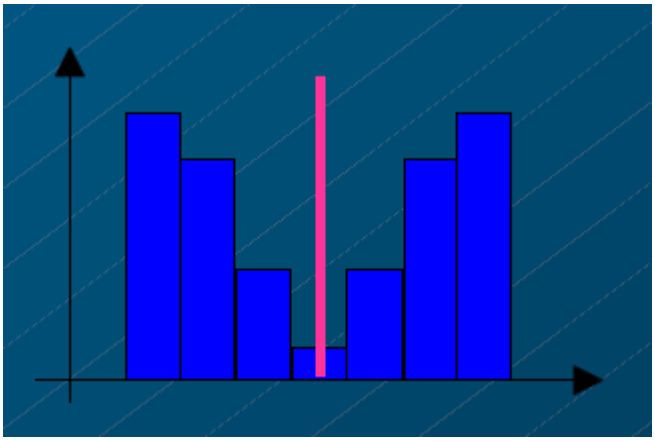
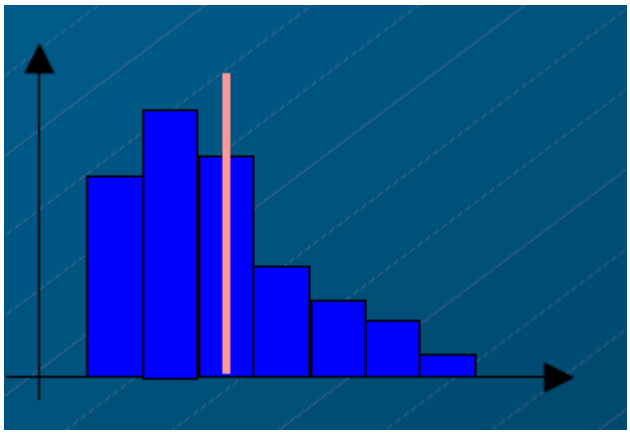


Figura Nº 9 – Distribución simétrica



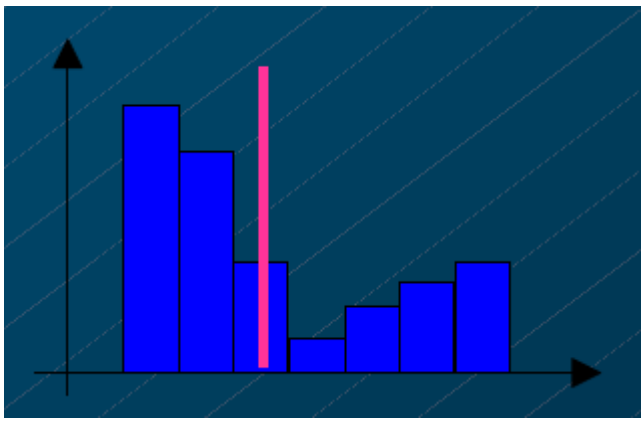


Figura Nº 10 – Distribución asimétrica positiva o a la derecha

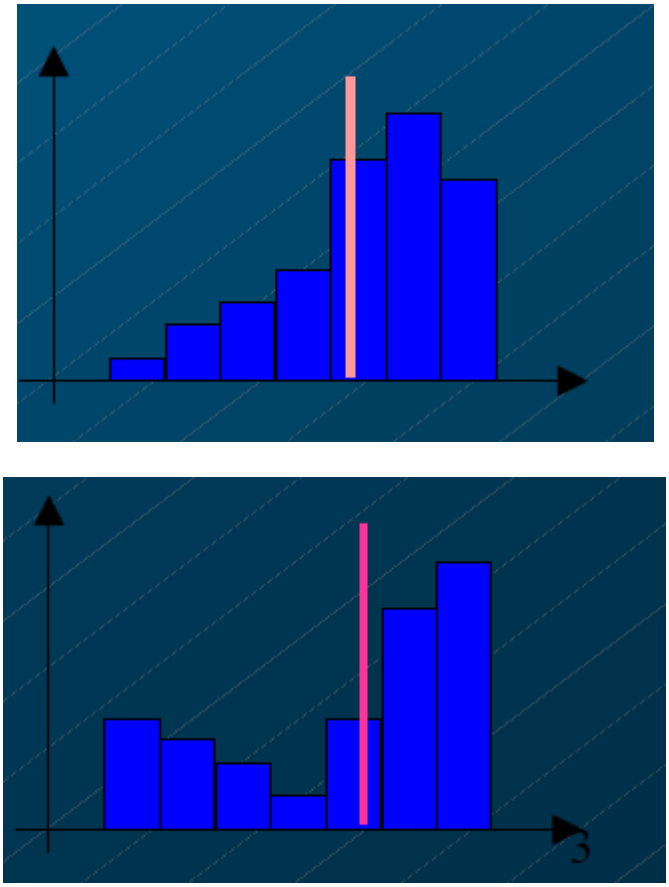


Figura Nº 11 – Distribución asimétrica negativa o a la izquierda

CURTOSIS

Esta medida dada por el coeficiente de Curtosis mide la diferencia en elevación respecto a una curva tomada come patrón e modelo que es la curva normal.

Se la puede definir aproximadamente como una razón de la amplitud semiintercuartil y la amplitud 90-10 porcentil:

$$K = \frac{1}{2}\frac{(Q_3 - Q_1)}{(P_{90} - P_{10})} \qquad (14)$$

este coeficiente clasifica distintos tipos de agudeza:

K>0	leptocúrtica o más empinada que la curva normal
K=0	mesocúrtica igual a la curva normal
K <0	platicúrtica o menos empinada que la curva normal

La forma exacta de calcular es a través del momento centrado de orden 4:

$$m_4 = \frac{\sum (x_i - \bar{x})^4}{n} \qquad K = \frac{m_4}{S^4} \qquad (15)$$

RESUMEN:

DIAGRAMAS SEGÚN EL TIPO DE VARIABLES

Tipo de variable	Diagrama o gráfico
Cualitativa	Barras, sectores, pictogramas
Cuantitativa (discreta)	Barras Escalera
Cuantitativa (continua)	Histograma, polígono de frecuencias diagramas acumulativos

Medidas Descriptivas Numéricas y Representaciones Graficas aconsejadas en función de la escala de medida de la variable

Escala de medida	Representaciones gráficas	Medidas de tendencia central	Medidas de dispersión
Nominal	Diagrama de barras Diagrama de líneas Diagrama de sectores	Moda	
Ordinal	Boxplot	Mediana	Rango Intercuartílico
Intervalo	Histogramas Polígono de frecuencias	Media	Desviación Típica
Razón		Media Geométrica	Coeficiente de Variación

Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 6

INFERENCIA ESTADÍSTICA DISTRIBUCIÓN EN EL MUESTREO -ESTIMACIÓN

DISTRIBUCIÓN EN EL MUESTREO

Hemos desarrollado en la unidad anterior lo referido al análisis de datos, ahora bien cuando se enfrenta un trabajo, generalmente el objetivo es conocer más acerca de la población de referencia, entonces las características muestrales serán el comienzo para este proceso.

Si la obtención de características muestrales se repite un determinado número de veces, es decir se sacan muestras de igual extensión, de la misma población que tiene una distribución dada, y en todas ellas se obtiene la misma función (la misma característica), los valores variarán de muestra a muestra. Esto permite considerar a las características muestrales como variables aleatorias. Como variables aleatorias tienen una distribución de probabilidad que les es propia. Generalmente se las conoce como ***distribución del estadístico por muestreo***. Estas distribuciones tienen propiedades bien definidas.

El proceso de analizar los datos tratando de traducir lo que ellos dicen en términos de probabilidad, con el fin de obtener conclusiones respecto a la población es lo que se denomina ***Inferencia Estadística***

ESTADÍSTICOS TRATADOS COMO VARIABLES ALEATORIAS

1- La media **y** la varianza muestral son dos de los estadísticos más importantes que serán estudiadas.

La media muestral será estudiada respondiendo a las siguientes preguntas:

- ¿Cuál es su valor medio?
- ¿Cuál es su varianza?
- ¿Cuál es su distribución?

Para empezar a responder se considera que los valores muestrales x_i son **independientes e idénticamente distribuidos**, con esperanza común μ y varianza σ^2 ya que provienen de la misma población.

$$E(\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\, n\, E(x) = \mu$$

Este resultado indica que el valor medio de la variable aleatoria media muestral es igual al valor medio de la población.

La varianza de la variable aleatoria media muestral se obtiene por hallar la varianza del promedio de n valores independientes o idénticamente distribuidos:

$$\mathrm{Var}(\bar{x}) = \mathrm{Var}\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{1}{n^2}\cdot \mathrm{Var}\sum_{i=1}^{n} x_i = \frac{1}{n^2}\cdot \sum_{i=1}^{n} \mathrm{Var}(x_i) = \frac{1}{n^2}\cdot n \cdot \sigma^2$$

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

El error estándar de la media muestral, que mide la variabilidad casual en medias de muestras es:

$$\sigma(\bar{x}) = \sqrt{\mathrm{Var}(x)} = \frac{\sigma}{\sqrt{n}}$$

La expresión anterior nuestra que el desvío de la media muestral es menor que el desvío de la población. Además cuando n tiende a infinito el desvío de la media muestral tiende a cero, esto significa que cuanto mayor es la extensión de la muestra, menor será el error o fluctuación de las medias de una muestra a otra.

Si las muestras son extraídas de una población finita y el muestreo se realiza sin reposición, se debe introducir un factor de corrección por población finita en el error de la media:

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

N extensión de la población y n: extensión de la muestra.

Llegados hasta aquí es conveniente presentar un teorema básico para el desarrollo de la Teoría de Inferencia, ***Teorema del Limite Central.***

Teorema del Límite Central.

Este teorema es muy útil, ya que es importante saber más acerca de la distribución de una suma de variables aleatorias. Su enunciado es el siguiente:

Sí se considera la suma de n variables* aleatorias x *independientes e idénticamente distribuidas, cada una con media y varianza finita, cuando el número de variables involucradas es mayor, la distribución de la suma se aproxima a una distribución Normal.

El valor de este teorema es que no requiere condiciones para las distribuciones de las variables aleatorias que se suman, sólo es necesario que cada una tenga un efecto insignificante sobre la distribución de la suma. Además brinda un método práctico para calcular valores de probabilidad aproximados asociados con sumas de variables aleatorias independientes distribuidas arbitrariamente

Este teorema es muy usado, ya que muchas variables aleatorias pueden considerarse como la suma de efectos independientes.

$$S = x_1 + x_2 + .. + x_n$$

$$S = \sum_{i=1}^{n} x_i = n\bar{x} \qquad E(x_i) = \mu \qquad \sigma^2(x_i) = \sigma^2$$

$$E[S] = E\left[\sum_{i=1}^{n} x_i\right] = E\left[n \cdot \bar{x}\right] = nE\left[\bar{x}\right]$$

$$\mathrm{Var}[S] = \mathrm{Var}\left[\sum_{i=1}^{n} x_i\right] = \mathrm{Var}\left[n\bar{x}\right] = n^2.\mathrm{Var}\left[\bar{x}\right]$$

$$\sigma(S) = n.\sigma\left(\bar{x}\right)$$

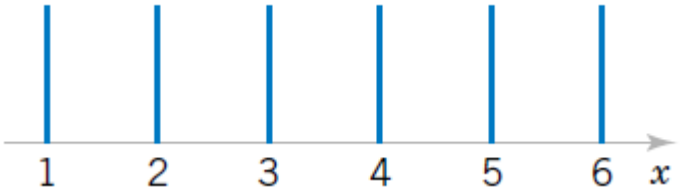
$$P\left(\frac{\bar{x} - E\left(\bar{x}\right)}{\sigma\left(\bar{x}\right)} \leq x\right) \rightarrow_{n\to\infty} N(0,1)$$

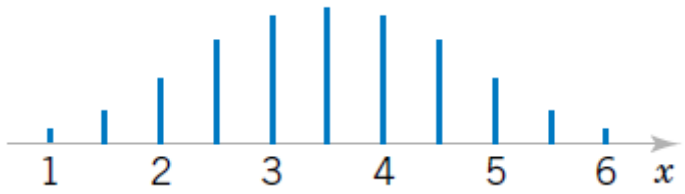
Se concluye, por lo tanto, que la variable aleatoria *media muestral* se distribuye normalmente con parámetros $E\left(\bar{x}\right)$ y $\sigma\left(\bar{x}\right)$

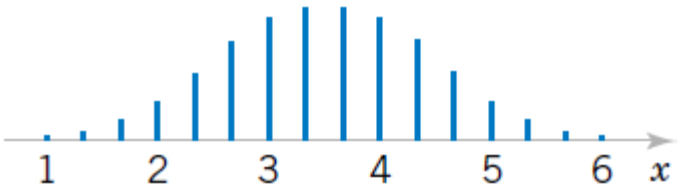
La conclusión es muy importante, ya que la mayor parte de los procedimientos de inferencia se basan en $\bar{x}$ .Si las variables que conforman la muestra se distribuyen normalmente, entonces la $\bar{x}$ también será distribuida normalmente, y así se puede

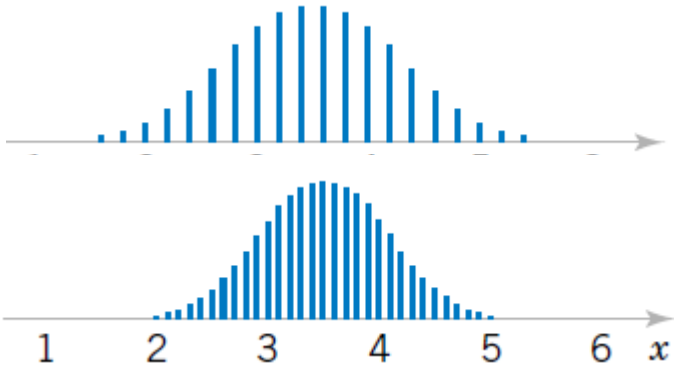
aplicar la teoría sobre variables distribuidas normalmente. En cambio, si las variables que conforman la muestra no son normales, entonces, para aplicar este teorema, es necesario que la extensión de la muestra n sea grande, y así $\bar{x}$ puede considerarse como distribuida normalmente.

Aunque el Teorema del Límite Central va a funcionar bien para muestras pequeñas (n= 4, 5) en la mayoría de los casos, sobre todo cuando la población es continua, unimodal y simétrica, se requiere muestras más grandes en otras situaciones, dependiendo de la forma de la población. En muchos casos de interés práctico, si n≥ 30, la aproximación normal será satisfactoria independientemente de la forma de la distribución de la población.









Se dijo anteriormente que la varianza muestral es, junto a la media, uno de los estimadores más importantes. Tratada como una variable aleatoria, es necesario obtener sus momentos:

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

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \mu) - (\bar{x} - \mu)\right]^2$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{2}{n}\left(\sum(x_i - \mu)(\bar{x} - \mu)\right) + \frac{1}{n}\sum(\bar{x} - \mu)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - 2\left(\frac{\sum x_i}{n} - \frac{n.\mu}{n}\right)(\bar{x} - \mu) + \frac{n}{n}(\bar{x} - \mu)^2$$

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

luego $E[S^2] = \frac{1}{n}E\left[\sum_{i=1}^{n}(x_i - \mu)^2\right] - E[(\bar{x} - \mu)^2]$

$$E[S^2] = \sigma_x^2 - \sigma_x^2 = \sigma_x^2 - \frac{\sigma_x^2}{n} = \sigma_x^2 \cdot \frac{n-1}{n}$$

$$E[S^2] \neq \sigma^2$$

La diferencia (n-1)/n *se denomina sesgo* y tiene realmente importancia cuando n es pequeño, ya que en caso contrario, el sesgo tiende a 1.

Un estimador insesgado de σ^2 es la varianza muestral corregida:

$$S'^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Con lo cual:

$$E[S'^2] = E\left(\frac{n.S^2}{n-1}\right) = E\left(\frac{n}{n-1} \cdot \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}\right)$$

$$E[S'^2] = \sigma^2$$

Para seguir el mismo razonamiento que con la media haría falta encontrar la varianza de la varianza muestral pero es muy extensa esta demostración.

Decimos únicamente que para poblaciones Normales la varianza del estimador S^2 es:

$$Var(S^2) = \frac{2\sigma^4}{\upsilon}$$

Para obtener la distribución por muestreo que le corresponderá a S^2 es necesario recordar la variable χ^2 que surge como la suma de cuadrados de variables aleatorias normales estandarizada:

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

Si μ se desconoce se lo estima a través de la media muestral con lo cual la expresión anterior se transforma en:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^2 = \frac{n S^2}{\sigma^2} \sim \chi^2_{n-1}$$

O bien utilizando la varianza muestral corregida:

$$\frac{(n-1).S'^2}{\sigma^2} = \chi^2_{n-1}$$

En los problemas que ocurren frecuentemente en ingeniería se necesita hacer estimación, relacionada generalmente a:

- ✓ **La media de la población**
- ✓ **La varianza de la población**
- ✓ **La proporción p de elementos en una población que pertenecen a una clase de interés**
- ✓ **La diferencia entre las medias de dos poblaciones**
- ✓ **La diferencia entre las proporciones de dos poblaciones**

Los estimadores razonables de esos parámetros son:

Para μ, la media muestral $\bar{x}$

Para σ^2, la varianza muestral S^2
Para π, la proporción muestral p
Para la diferencia de medias poblacionales μ_1- μ_2, la diferencia de medias muestrales $\bar{x}_1 - \bar{x}_2$
Para la diferencia de proporciones poblacionales π_1-π_2, la diferencia de proporciones muestrales p_1-p_2

Por lo tanto se debe analizar de todos estos estimadores su distribución muestral:

DISTINTOS CASOS DE DISTRIBUCIÓN POR MUESTREO

1.-Distribución por muestreo de medias

Población Normal con desvío σ conocido

Estandarizando la variable aleatoria media muestral se obtiene una variable Normal estándar:

$$X \sim N \quad (\mu, \sigma) \quad ; \quad \bar{x} \sim N \quad \left(\mu, \frac{\sigma}{\sqrt{n}}\right);$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N \quad (0,1)$$

En el caso de que la población tenga una distribución aproximadamente Normal, los resultados son similares sólo que le variable z será también distribuida en forma aproximadamente Normal.

Población Normal, σ desconocido, muestra chica (n < 30)

Si las variables que constituyen la muestra son independientes e idénticamente distribuidas con media y varianza finita, pero como sucede generalmente, la varianza se desconoce será reemplazada por la varianza muestral, entonces la variable resultante se distribuye como t de Student. Esto es debido a que como ya se demostró, una variable ***t*** se genera como el cociente de una variable Normal y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad; entonces:

$$t = \frac{z}{\sqrt{\frac{\chi^2}{v}}} \quad siendo \quad z \sim N \quad (0,1)$$

$$= \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\left(\frac{nS^2}{\sigma^2}\right)}{n-1}}} = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{\frac{\sqrt{n}.S}{\sigma\sqrt{n-1}}}$$

$$= \frac{\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\left(\frac{nS^2}{\sigma^2}\right)}{n-1}}} = \frac{\left(\bar{x} - \mu\right)\sqrt{n}}{\frac{\sqrt{n}.S}{\sigma\sqrt{n-1}}}$$

$$t_{n-1} = \frac{\bar{x} - 1}{\frac{S}{\sqrt{n-1}}} \quad o \quad bien \quad t_{n-1} = \frac{\bar{x} - \mu}{\frac{S'}{\sqrt{n}}}$$

Población Normal, σ desconocido, muestra grande (n ≥ 30)

Cuando la muestra es grande se puede considerar la varianza poblacional desconocida reemplazada por la varianza muestral y la distribución de la variable resultante sigue siendo Normal:

$$z = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

2.-Distribución muestral de la varianza

Provenientes los valores muestrales de una población con distribución Normal recordar, ya se ha mostrado que la varianza muestral tiene una distribución Chi-cuadrado.

$$\chi^2_{n-1} = \frac{n.S^2}{\sigma^2}$$

3.-Distribución muestral de proporciones

En la población, una proporción se define como:

$$\pi = \frac{K}{N}$$

Siendo K el número de elementos que tienen una característica deseada y N el total de elementos de la población. En la muestra, se define como:

$$p = \frac{x}{n}$$

siendo p la proporción muestral, x la cantidad de elementos que poseen la categoría deseada y n la extensión de la muestra. Suele considerarse a p como la proporción de éxitos, y por esto se la asocia a la distribución Binomial (recuérdese que E = n.p y Var = n.p.q).

Luego, las características de esta variable aleatoria son:

$$E[p] = E\left[\frac{x}{n}\right] = n\frac{\pi}{n} = \pi$$

$$Var[p] = Var\left[\frac{x}{n}\right] = \frac{n\,\pi\,(1-\pi)}{n^2} = \frac{\pi\,(1-\pi)}{n}$$

El error estándar de p mide las variaciones casuales de proporciones de muestra de una muestra a otra:

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

Este error debe ajustarse por un factor de corrección por población finita, si el muestreo se hace sin reposición:

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

Luego, la distribución muestral de es la siguiente:

$$p \sim N\left(\pi;\sqrt{\frac{\pi\ (1-\pi\ )}{n}}\right)$$

$$z = \frac{p-\pi}{\sqrt{\frac{\pi\ (1-\pi\ )}{n}}} \sim N(0,1)$$

4.-Distribución muestral de la diferencia de dos medias muestrales

Varianzas poblacionales conocidas

Cuando sea de interés comparar las medias de dos variables aleatorias, esto se hará sobre la base de dos muestras extraídas de las poblaciones cuyas medias se quiere comparar.

$x \sim N(\mu_x, \sigma_x)$ $\quad y \sim N(\mu_y, \sigma_y)$

$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n_x}}\right)$ $\quad \bar{y} \sim N\left(\mu_y, \frac{\sigma_y}{\sqrt{n_y}}\right)$

siendo $\bar{x}$ independiente de $\bar{y}$

Usando los resultados de combinaciones lineales de variables distribuidas normalmente puede decirse que la variable aleatoria *diferencia de medias* muestrales se distribuye normalmente. Aún sin saber si las poblaciones son normales, si las extensiones de muestras son suficientemente grandes, como cada media muestral se distribuye normalmente, es de esperar que la diferencia de medias muestrales sea también normalmente distribuida. Los parámetros de esta distribución normal son:

$$E(\bar{x}-\bar{y}) = E(\bar{x}) - E(\bar{y}) = \mu_x - \mu_y$$

$$\mathrm{Var}\left(\bar{x}-\bar{y}\right)=\mathrm{E}\left[\bar{x}-\bar{y}-\mathrm{E}\left(\bar{x}-\bar{y}\right)\right]^2=$$

$$=\mathrm{E}\left[\left(\bar{x}-\bar{y}\right)-\mathrm{E}\left(\bar{x}\right)+\mathrm{E}\left(\bar{y}\right)\right]^2=\mathrm{E}\left[\left(\bar{x}-\mathrm{E}\left(\bar{x}\right)\right)-\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]+\mathrm{E}\left[\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]^2=$$

$$=\mathrm{E}\left[\bar{x}-\mathrm{E}\left(\bar{x}\right)\right]^2-2\mathrm{E}\left[\left(\bar{x}-\mathrm{E}\left(\bar{x}\right)\right)\cdot\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]+\mathrm{E}\left[\left(\bar{y}-\mathrm{E}\left(\bar{y}\right)\right)\right]^2$$

Como la covarianza de variables aleatorias independientes es igual a cero, luego:

$$\mathrm{Var}\left(\bar{x}-\bar{y}\right)=\mathrm{Var}\left(\bar{x}\right)+\mathrm{Var}\left(\bar{y}\right)=\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}$$

Entonces, si las varianzas de ambas poblaciones se conocen, se obtiene:

$$\frac{\bar{x}-\bar{y}-\left(\mu_x-\mu_y\right)}{\sqrt{\left(\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}\right)}}=z\sim N\left(0,1\right)$$

Varianzas poblacionales desconocidas - Muestras grandes

La mayoría de las veces las varianzas poblacionales se desconocen y deben ser estimadas. En este caso, es decir muestras grandes, la distribución de la diferencia de medias muestrales sigue siendo Normal pero con la variable z con la siguiente forma:

$$z=\frac{\bar{x}-\bar{y}-\left(\mu_x-\mu_y\right)}{\sqrt{\left(\frac{S_x^2}{n_x}+\frac{S_y^2}{n_y}\right)}}\sim N\left(0,1\right)$$

Varianzas poblacionales desconocidas - Muestras chicas

En este caso debe hacerse la siguiente consideración respecto a las varianzas poblacionales desconocidas:

Varianzas poblacionales desconocidas pero supuestas iguales e iguales a un valor constante.

Recordar que la distribución chi-cuadrado está asociada a la varianza muestral, luego, usando las propiedades reproductivas de la distribución Chi-cuadrado se obtiene:

$$\frac{(n_x-1)S'^2_x}{\sigma^2}+\frac{(n_y-1)S'^2_y}{\sigma^2} \sim \chi^2_{n_x+n_y-2}$$

con lo cual la distribución deja de ser Normal para transformarse en t de Student. Recordar como surge una variable t de Student: como el cociente entre une variable Normal (0,1) y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad

$$= \frac{\dfrac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sigma\sqrt{\dfrac{1}{n_x}+\dfrac{1}{n_y}}}}{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{\sigma^2(n_x+n_y-2)}} =$$

$$= \frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{n_x+n_y-2}}\sqrt{\left(\dfrac{1}{n_x}+\dfrac{1}{n_y}\right)}} = t_{n_x+n_y-2}$$

siendo $S_w=\sqrt{\dfrac{(n_x-1)S'^2_x+(n_y-1)S'^2_y}{n_x+n_y-2}}$ el estimador ponderado de σ

Varianzas poblacionales desconocidas y distintas

Los desvíos poblacionales desconocidos son reemplazados por los desvíos muestrales pero se obtiene una variable t de Student cuyos grados de libertad deben ser calculados:

$$t_v=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\left(\dfrac{S'^2_x}{n_x}+\dfrac{S'^2_y}{n_y}\right)}}$$

5.-Distribución muestral de la diferencia de proporciones

Si de dos poblaciones independientes, cada una con distribución Binomial de parámetro π, se extrae una muestra, luego el estimador de la diferencia de proporciones poblacionales $\pi_1 - \pi_2$, será $p_1 - p_2$, de la cual se quiere determinar su distribución por muestreo:

$$P_1 \sim N\left(\pi_1 ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right)$$

$$P_2 \sim N\left(\pi_2 ; \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

$$E(p_1 - p_2) = \pi_1 - \pi_2$$

$$Var(p_1 - p_2) = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$$

Luego si los tamaños de muestra son suficientemente grandes, la distribución de Δp por muestreo es aproximadamente Normal (por el Teorema del límite Central).

$$z = \frac{(p_1 - p_2) - \pi_1 - \pi_2}{\sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}} \sim N(0,1)$$

6.-Distribución muestral del cociente de varianzas

Suele ser de interés comparar la variabilidad de dos poblaciones, esto se puede realizar a través de la razón de varianzas muestrales. Si esta razón es cercana a la unidad, las variabilidades se puede decir que son casi equivalentes; Si por el contrario se aleja de uno, se dice que son no equivalentes; pero para que esta decisión pueda ser correcta, se deberá analizar la distribución de la razón de varianzas muestrales. Para esto se extrae una muestra aleatoria de tamaño n_x, de la primer población, constituida por variables independientes y distribuidas Normalmente, cada una con media μ_x y varianza σ^2_x; lo mismo se hace con la población dos, se extrae una muestra de extensión n_y de variables aleatorias independientes, cada una con media μ_y y varianza σ^2_y siendo X e Y independientes. Luego la distribución de cada varianza se vincula a la distribución χ^2 de la siguiente forma:

$$\frac{(n_x - 1)S'^2_x}{\sigma^2_x} = \frac{\sum_{i=1}^{n_x}(x_i - \bar{x})^2}{\sigma^2_x} \sim \chi^2 \text{ con } n_x - 1 \text{ grados de libertad.}$$

$$\frac{(n_y - 1)S'^2_y}{\sigma_y^2} = \frac{\sum_{j=1}^{n_y}(y_j - \bar{y})^2}{\sigma_y^2} \sim \chi^2 \text{ con } n_y - 1 \text{ grados de libertad.}$$

Al ser X y Y variables aleatorias independientes, entonces estas dos variables χ^2 también son independientes. De esta manera, el cociente de estas variables χ^2 origina une variable F de Snedecor, con n_x-1 y n_y-1 grados de libertad.

ESTIMACIÓN

Frecuentemente, los parámetros de las distribuciones son valores que se desconocen. Se busca, entonces, a partir de valores observados, estimar el o los valores desconocidos. Este procedimiento se denomina ***estimación de parámetros***.

Un estimador es una función de valores observados (muestra) que no depende de ningún parámetro desconocido. ***Un estimador es un estadístico***, ***y*** **una** ***estimación es cualquiera*** **de sus** ***posibles valores*****.**

Para estimar un parámetro pueden utilizarse distintos estadísticos (características de muestra). Es evidente que en calidad de estimación conviene tomar estadísticos cuyos valores, para distintas muestras de la población sean, por término medio, próximos al valor real del parámetro. También es deseable que con el aumento del tamaño de la muestra crezca la fiabilidad de la estimación.

Si se ha obtenido un estimador puntual, es conveniente tener una medida de precisión atribuida al estimador. La precisión de un estimador se mide por el error estándar del estimador. Es decir, cuanto menor sea este error, tanto más preciso será el estimador. Es bueno, entonces, que cuando se de una estimación, también se brinde el error estándar de la estimación.

Hay varios métodos para realizar estimación de los cuales se va a desarrollar uno de ellos.

Si se quiere una expresión más formal de la estimación y su precisión, se puede obtener lo que se denomina ***estimación por intervalos***. Es la estimación de un parámetro por un intervalo al azar, que se denomina ***intervalo de confianza***, cuyos extremos son funciones de las variables aleatorias observadas.

Se llama ***intervalo de confianza para el parámetro θ*** al intervalo (θ_1, θ_2) que contiene el valor real del parámetro, con una probabilidad dada 1-α, siendo ésta la probabilidad confidencial. Las cotas del intervalo, como se dijo, son funciones de las observaciones y, por lo tanto, son variables aleatorias. Es por esto que se dice que el intervalo de confianza "cubre" al parámetro que se estima con una probabilidad l- α, o bien, en el 100 (1- α) % de los casos. La elección de la

probabilidad confidencial se determina por las condiciones concretas; por regla general se utilizan 0.90, 0.95 y 0,99.

Esta formulación puede expresarse de forma general como sigue:

$$P\left(\left|\theta - \hat{\theta}\right| \leq k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

1-α coeficiente de confianza; k constante no negativa que depende de la distribución por muestreo del estimador $\hat{\theta}$

Esta desigualdad puede escribirse de la siguiente manera:

$$P\left(-k\,\sigma_{\hat{\theta}} \leq \theta - \hat{\theta} \leq k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

$$P\left(\hat{\theta} - k\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

Con esto puede obtenerse una expresión general de un estimador por intervalo de confianza simétrico, para un parámetro:

$$P\left(\hat{\theta} - k\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + k\,\sigma_{\hat{\theta}}\right) = 1 - \alpha$$

Siendo $\hat{\theta} - k\,\sigma_{\hat{\theta}}$ = L límite inferior

$\hat{\theta} + k\,\sigma_{\hat{\theta}}$ = U límite superior

Esta expresión permite obtener intervalos de confianza para cualquier parámetro, sea la distribución del estimador simétrica o no.

La constante k depende de la distribución muestral del estimador y del valor de 1-α.

En la estimación por intervalos se desea obtener intervalos de poca amplitud, ya que esto hará más precisa la estimación. El ancho real de un intervalo es dictado por el coeficiente de confianza y por el tamaño de la muestra, entre otras cosas. Dados la extensión de la muestra y el error estándar del estimador, cuanto más corto es el intervalo, tanto menor es el nivel de confianza.

Es posible obtener el tamaño de muestra adecuado a la precisión y a la confianza con la cual se quiere trabajar:

$$\text{Error de estimación} = |z_{\left(\frac{\alpha}{2}\right)}| \frac{\sigma}{\sqrt{n}}$$

$$n = \frac{z^2_{\frac{\alpha}{2}} \cdot \sigma^2}{\text{Error}^2}$$

Esto permite variar el nivel de confianza sin aumentar el error de estimación, sólo variando el tamaño de muestra; o bien reducir el error de estimación sin variar el nivel de confianza.

INTERVALOS PARA PARÁMETROS

Intervalos para la media poblacional

a -- Población Normal con desvío parámetro conocido

El intervalo para la media poblacional se basa en el estimador media muestral. En este caso su distribución muestral es la siguiente:

$$x \sim N(\mu;\sigma)$$

$$\bar{x} \sim N\left(\mu;\frac{\sigma}{\sqrt{n-1}}\right)$$

Luego, estandarizando la variable media muestral se obtiene:

$$\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = z \sim N(0,1)$$

De la expresión general de intervalos se deduce el intervalo correspondiente:

$$\theta = \bar{x} \qquad \sigma_\theta = \frac{\sigma}{\sqrt{n}} \qquad k = |z_{\left(\frac{\alpha}{2}\right)}|$$

$$\left(\bar{x} \pm |z_{\left(\frac{\alpha}{2}\right)}| \frac{\sigma}{\sqrt{n}} \right)$$

b - Población Normal con desvío parámetro desconocido Muestra grande(n > 30)

EL desvío poblacional que se desconoce se estima por S, con lo cual la distribución del estimador media muestral se transforma en:

$$\bar{x} \sim N\left(\frac{S}{\sqrt{n}} \right)$$

con lo cual el intervalo de acuerde a la expresión general es:

$$P\left(\bar{x} - z_{\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\left(\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n}} \right) = 1 - \alpha$$

Recordar que si el muestreo es sin reposición de una población finita debe hacerse la corrección del error estándar de la media muestral.

c - Población Normal con desvío parámetro desconocido. Muestra chica (n < 30)

$$x \sim N(\mu; \sigma)$$

$$\bar{x} \sim N\left(\mu; \frac{S}{\sqrt{n-1}} \right)$$

$$\text{con lo cual } \frac{\bar{x} - 1}{\frac{S}{\sqrt{n-1}}} = t_{n-1}$$

y el intervalo partiendo de la expresión general será:

$$p\left(\bar{x} - t_{\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{\left(1-\frac{\alpha}{2}\right)} \frac{S}{\sqrt{n-1}}\right) = 1-\alpha \quad \text{ó bien}$$

$$p\left(\bar{x} - t_{\left(1-\frac{\alpha}{2}\right)} \frac{S'}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\left(1-\frac{\alpha}{2}\right)} \frac{S'}{\sqrt{n}}\right) = 1-\alpha$$

$$\left(\bar{x} \pm |t_{\left(1-\frac{\alpha}{2}\right)}| \frac{S'}{\sqrt{n}}\right)$$

Intervalo para la varianza poblacional

Para obtener un intervalo para el desvío poblacional se toma como estimador la varianza o desvío muestral. Recordar la distribución muestral de la varianza muestral:

$$\frac{(n-1)\, S'^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{ó bien}$$

$$\frac{n\, S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Luego el intervalo buscado será;

$$P\left(\frac{n S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \leq \sigma^2 \leq \frac{n S^2}{\chi^2_{\frac{\alpha}{2};n-1}}\right) = 1-\alpha$$

Un intervalo para el desvío poblacional se deduce del anterior por obtener la raíz cuadrada de todos los términos de le desigualdad.

Intervalo para la proporción poblacional

La estimación de la proporción poblacional π se basa en la proporción muestral p. Recordando su distribución muestral y suponiendo una extensión de muestra suficientemente grande, entonces:

$$p \sim N\left(\pi \; ; \; \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

$$z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$$

El intervalo buscado, de acuerde e le expresión general, será

$$\left(p \pm |Z_{\left(1-\frac{\alpha}{2}\right)}| \sqrt{\frac{\pi(1-\pi)}{n}} \right)$$

En la expresión del desvío de p, se deberá sustituir el parámetro desconocido π por su estimador puntual p para poder obtener un valor. Además, si el muestreo es sin reposición y la población finita, deberá corregirse este desvío de acuerde a la siguiente expresión:

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

Intervalo para la diferencia de medias poblacionales

a – Poblaciones Normales. Desvíos parámetros conocidos

$$X \sim N(\mu_x; \sigma_x) \qquad Y \sim N(\mu_y; \sigma_y)$$

$$\bar{x} \sim N\left(\mu_x; \frac{\sigma_x}{\sqrt{n_x}}\right) \qquad \bar{y} \sim N\left(\mu_y; \frac{\sigma_y}{\sqrt{n_y}}\right)$$

El estimador, es en este caso, la diferencia de medias muestrales. Recordar su distribución muestral.

$$\bar{x} - \bar{y} \sim N\left(\mu_x - \mu_y \; ; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right)$$

$$Z = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = N(0,1)$$

$$\theta = \mu_x - \mu_y \qquad \hat{\theta} = \bar{x} - \bar{y} \qquad \sigma_\theta = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

Esto permite obtener con la expresión general de intervalo, el buscado para este caso:

$$\left(\left(\bar{x} - \bar{y}\right) \pm \; | Z_{\left(1-\frac{\alpha}{2}\right)} | \; \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}} \right) = 1 - \alpha$$

b- Poblaciones normales, desvíos poblacionales desconocidos pero supuestos iguales

Las consideraciones son similares el caso anterior, solo que ahora σ_x y σ_y se desconocen, pero se consideran iguales (esto debe ser verificado previamente). Para este caso, la distribución por muestreo de la diferencia de medias muestrales es la siguiente:

$$\frac{\bar{x} - \bar{y} - \left(\mu_x - \mu_y\right)}{S_w \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \; t_{n_x + n_y - 2}$$

Con $S_w = \sqrt{\frac{\left(n_x - 1\right) S_x'^2 + \left(n_y - 1\right) S_y'^2}{n_x + n_y - 2}}$ ó $S_w = \sqrt{\frac{n_x S_x^2 + n_y S_y^2}{n_x + n_y - 2}}$

Por lo tanto el intervalo, de acuerde a la expresión general, es el siguiente:

$$\left(\left(\bar{x} - \bar{y}\right) \; \pm \; | t_{\left(1-\frac{\alpha}{2}\right)} | \; S_w \; \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right)$$

c- Poblaciones normales, desvíos poblacionales desconocidos pero distintos

Si los desvíos no pueden considerarse iguales, esto lleva a una nueva expresión de la variable:

$$\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}}} = t_\nu \quad , \quad \nu = \frac{\left(\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}\right)^2}{\frac{\left(\frac{S_x'^2}{n_x}\right)^2}{n_x-1}+\frac{\left(\frac{S_y'^2}{n_y}\right)^2}{n_y-1}}-2$$

Con lo cual, el intervalo es el siguiente:

$$\left((\bar{x}-\bar{y}) \pm |t_{\left(1-\frac{\alpha}{2}\ ,\ \nu\right)}| \sqrt{\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}} \right)$$

Intervalo para la diferencia de proporciones poblacionales

El estimador de $\Delta\pi$ es $\Delta p = p_1$ y p_2, siendo p_1 y p_2 proporciones muestrales obtenidas de muestras al azar independientes de cada una de las poblaciones, con n_1 y n_2 suficientemente grandes:

$$p_1 \sim N\left(\pi_1 \ ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}}\right)$$

$$p_2 \sim N\left(\pi_2 \ ; \sqrt{\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

$$\Delta p \sim N\left(\pi_1-\pi_2 \ ; \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}\right)$$

Con lo cual el intervalo será:

$$\Delta p \pm |Z_{\left(1-\frac{\alpha}{2}\right)}| \sqrt{\frac{\pi_1(1-\pi_1)}{n_1}+\frac{\pi_2(1-\pi_2)}{n_2}}$$

Notar que en el desvío de Δp es necesario reemplazar los valores desconocidos de π_1 y π_2 por sus estimadores puntuales p1, y p_2.

Intervalo para la razón de varianzas poblacionales

Sean X y Y dos variables aleatorias independientes, con distribución Normal. Si interesa obtener un intervalo para la razón de las varianzas poblacionales, esto se obtiene a partir de la razón de varianzas muestrales de la siguiente manera:

$$X \sim N(\mu_x, \sigma_x) \qquad Y \sim N(\mu_y, \sigma_y)$$

conocidos $\bar{x},\ \bar{y},\ S_x^2 \ \ y \ \ S_y^2,\ n_x \ \ y \ \ n_y,$ luego

$$F_{n_x-1;n_y-1} = \frac{\frac{S_x^2}{\sigma_x^2}}{\frac{S_y^2}{\sigma_y^2}}$$

Entonces, el intervalo para la razón de varianzas es:

$$F\left(\frac{S_x^2}{S_y^2} \ F_{1-\frac{\alpha}{2};\ n_y-1;\ n_x-1} \ \leq \ \frac{\sigma_x^2}{\sigma_y^2} \ \leq \ \frac{S_x^2}{S_y^2} \ F_{\frac{\alpha}{2};\ n_y-1;\ n_x-1} \right) = 1-\alpha$$

El intervalo para la razón de desvíos se obtiene directamente por hallar la raíz cuadrada a todos los términos de la desigualdad anterior.

$$P\left(\frac{S_x}{S_y} \ \sqrt{\frac{1}{F_{\frac{\alpha}{2};\ n_x-1;\ n_y-1}}} \ \leq \ \frac{\sigma_x}{\sigma_y} \ \leq \ \frac{S_x}{S_y} \ \sqrt{F_{\frac{\alpha}{2};\ n_x-1;\ n_y-1}} \right) = 1-\alpha$$

En la tabla final de esta unidad puede observarse, en forma resumida, todo lo expresado anteriormente.

PROCEDIMIENTO GENERAL PARA DETERMINAR LAS COTAS DE LOS INTERVALOS

1-Identificar el estimador apropiado para el parámetro que se desea estimar

2-Determinar su distribución por muestreo.

3-De acuerdo a la expresión general, plantear el intervalo.

4-Sustituir en la desigualdad los valores obtenidos de la muestra.

5-Una vez obtenido el intervalo, se concluye diciendo que el intervalo hallado cubre el valor del parámetro desconocido con una confianza de (1-α) %.

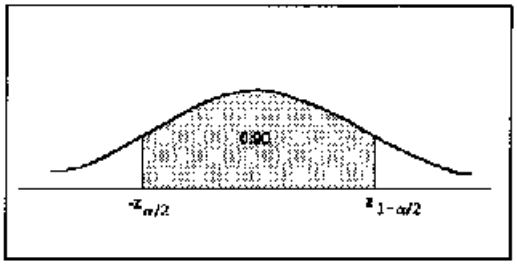
Ejemplo

-El tiempo de funcionamiento sin fallas de una máquina se sabe que es distribuido normalmente. Se realizaron 100 observaciones del tiempo sin fallas y se obtuvo un valor medio de 500 horas, conociendo que el desvío parámetro es de 10 horas. Se desea estimar, con 90% de confianza, el valor medio del tiempo de funcionamiento sin fallas.

$\bar{x} = 500 \ \ hs \ \ ; \ \ n = 100 \ \ ; \ \ \sigma = 10 \ \ hs$

$$\bar{x} \cong N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad ; \quad z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \quad ; \quad N(0,1)$$

$$P\left(\bar{x} - z_{\left(\frac{\alpha}{2}\right)} \cdot \frac{\sigma}{\sqrt{n}} \ \leq \ \mu \ \leq \ \bar{x} + z_{\left(1-\frac{\alpha}{2}\right)} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$



$$P\left(500 - 1.65 \cdot \frac{10}{\sqrt{100}} \ \leq \ \mu \ \leq \ 500 + 1.65 \frac{10}{\sqrt{100}}\right) = 0.90$$

$(498.35; 501.65)$ es posible decir que con 90% de confianza el valor media del tiempo de funcionamiento sin fallas se encuentra en este intervalo hallado

UNIVERSIDAD NACIONAL DEL LITORAL

Facultad de Ingeniería y Ciencias Hídricas

CÁTEDRA ESTADÍSTICA

Profesor Titular: Ing. Susana Vanlesberg

PRUEBAS DE HIPÓTESIS

Introducción

En la parte anterior hemos visto cómo construir una estimación a partir de un intervalo de confianza para un parámetro a partir de datos de la muestra. Sin embargo, muchos problemas de ingeniería requieren que decidamos si aceptar o rechazar un supuesto acerca de algún parámetro. El supuesto se denomina hipótesis, y el procedimiento de toma de decisiones acerca de la hipótesis se denomina pruebas de hipótesis. Este es uno de los aspectos más útiles de la inferencia estadística, ya que muchos problemas de toma de decisiones, pruebas o experimentos en el mundo de la ingeniería se pueden formular como problemas de prueba de hipótesis. Además, como veremos más adelante, hay una conexión muy estrecha entre pruebas de hipótesis e intervalos de confianza.

Prueba de hipótesis estadística y estimación de intervalos de confianza de los parámetros son los métodos fundamentales utilizados en la etapa de análisis de los datos de un experimento en el que el ingeniero está interesado, por ejemplo, en la comparación de la media de una población con un valor especificado.

Una hipótesis estadística es una declaración o supuesto acerca de los parámetros de una o más poblaciones.

En muchas circunstancias, las decisiones se deben tomar con base solo en la información de la muestra. Un gerente de control de calidad debe determinar si su

proceso funciona correctamente. Un director de una repartición provincial debe determinar si una nueva estrategia de manejo de recursos es adecuada para su provincia. Un ingeniero proyectista saber cuál es el valor de la precipitación media mensual en un lugar en el que se diseñará un desagüe.

El tomador de decisiones querría cerciorarse, hasta donde sea posible, de que ha llegado a la conclusión correcta.

Conceptos principales

En muchos casos, los resultados de observaciones se utilizan para verificar suposiciones, ***Hipótesis***, respecto a algunas propiedades de la distribución de una población.

Generalmente la distribución de la variable x se conoce y, basándose en la muestra de observaciones, es necesario comprobar la hipótesis sobre los valores de los parámetros de esta distribución. ***Estas hipótesis se denominan paramétricas.***

La hipótesis sujeta a verificación se denomina ***hipótesis nula*** H_0. Paralelamente con esta, se analiza la hipótesis denominada ***hipótesis alternativa*** H_1, que suele ser la opuesta a la hipótesis nula.

Por ejemplo, si se comprueba la hipótesis de que el parámetro θ es igual a cierto valor θ_0 (H_0: $\theta = \theta_0$) puede analizarse como hipótesis alternativa cualquiera de las siguientes:

a) H_1: $\theta > \theta_0$ **b)** H_1: $\theta < \theta_0$ **c)** H_1: $\theta \neq \theta_0$

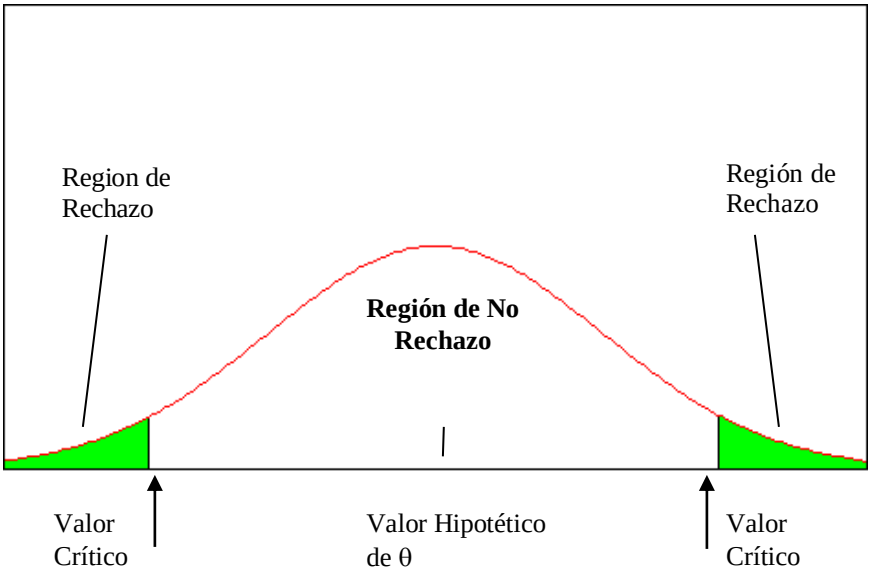
dependiendo de la formulación del problema.

Región de Rechazo y Región de Aceptación

Para cada tipo de procedimiento de prueba de hipótesis, se puede calcular una ***prueba estadística*** apropiada. Esta prueba estadística mide el acercamiento del valor de la muestra a la hipótesis nula.

La distribución apropiada de la prueba estadística se divide en dos regiones: una ***región de rechazo*** y una ***región de aceptación***.

Al tomar la decisión con respecto a la hipótesis nula, en la distribución estadística se debe determinar el ***valor crítico*** que divide la región de aceptación de la región de rechazo; este valor crítico depende del tamaño de la región de rechazo. El gráfico siguiente muestra las dos zonas.



ERRORES DE TIPOS I Y II

Este procedimiento de decisión puede conducir a una de dos conclusiones erróneas. Por ejemplo, dado el supuesto como verdad del valor del parámetro y formulado en la hipótesis nula, sin embargo, para los valores muestrales seleccionados al azar, podríamos observar un valor de la estadística de prueba que cae en la región crítica. Entonces podríamos rechazar la hipótesis nula hipótesis H_0 en favor de la alternativa cuando, de hecho, H_0 es realmente cierta. Este tipo de mal conclusión se llama un error de tipo I.

Rechazar la hipótesis nula H_0 cuando es verdadera se define como un error de tipo I.

Supongamos ahora que el verdadero valor del parámetro es diferente del valor plateado en H_0, sin embargo, el valor de la característica de muestra cae en la región de aceptación. En este caso queremos dejar de rechazar H_0 cuando es falsa.

Aceptar la hipótesis nula cuando en realidad es falsa se denomina error de tipo II.

Las probabilidades de cometer estos errores pueden considerarse como los riesgos de tomar decisiones incorrectas. La probabilidad máxima de cometer un error del tipo I se llama nivel de significación, y generalmente se representa con α:

$$\boldsymbol{\alpha} = \text{Máx P (I)} = \text{Máx P } (H_1/H_0) = \text{Máx P(rechazar } H_0/H_0 \text{ es verdadera)}$$

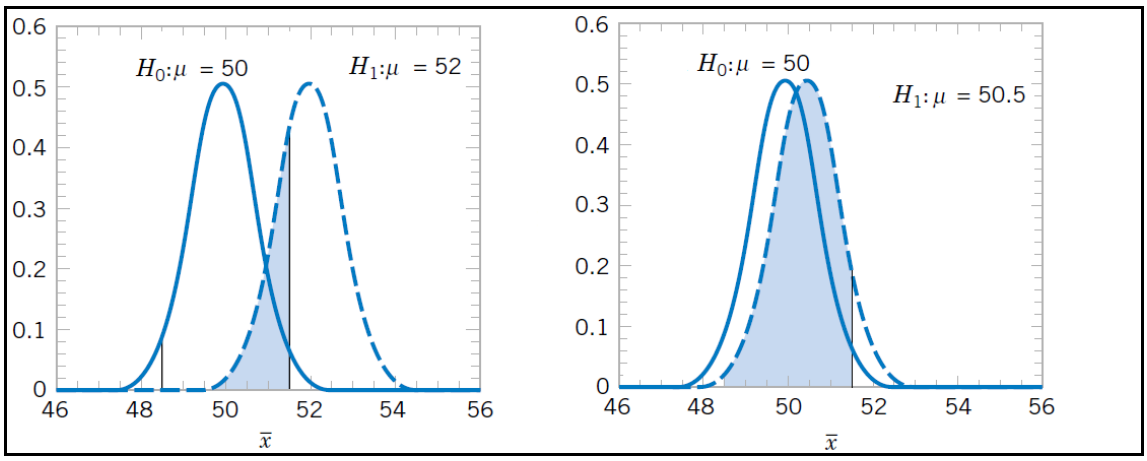
La probabilidad máxima de cometer un error del tipo II se denomina β:

$$\boldsymbol{\beta} = \text{Máx P (II)} = \text{Máx P } (H_0/H_1) = \text{Máx P(aceptar } H_0/H_0 \text{ es falsa)}$$

$$P(\alpha) = \int_{-\infty}^{k_c} f \left(k \ / \ H_0 \right) dk$$

$$P(\beta) = \int_{-\infty}^{k_c} f \left(k \ / \ H_1 \right) dk$$

Una buena prueba estadística es aquella en donde tanto α como β son pequeñas, porque permitirá tomar una decisión correcta, con menor riesgo de equivocarse.



Probabilidad de error tipo II para dos H_1 propuestas

1.- El tamaño de la región crítica, y en consecuencia la probabilidad de un error de tipo I, siempre se puede reducir mediante la selección apropiada de los valores críticos.
2.- Errores tipo I y tipo II están relacionados. Una disminución en la probabilidad de un tipo de error siempre resulta en un aumento en la probabilidad de la otra, siempre que el tamaño de la muestra no cambie.
3.- Un aumento del tamaño de la muestra generalmente reduce tanto α como β siempre que los valores críticos se mantengan constantes.
4.- Cuando la hipótesis nula es falsa β aumenta a medida que el verdadero valor del parámetro se acerca al valor propuesto en la hipótesis nula.

En general, el analista controla la probabilidad de error tipo I. Dado que el analista puede controlar directamente la probabilidad de rechazar erróneamente H_0, siempre pensamos en el rechazo de la hipótesis nula H_0 como una fuerte conclusión.

Por otra parte, la probabilidad de error tipo II β no es una constante, sino que depende del verdadero valor del parámetro. También depende del tamaño de la muestra que hemos seleccionado.

Debido a que la probabilidad de error de tipo II, β, es una función tanto del tamaño de la muestra y la medida en que la hipótesis nula H_0 es falsa, se acostumbra a pensar en la decisión de aceptar H_0 como débil conclusión, a menos que sepamos que β es aceptablemente pequeña. Por lo tanto, en lugar de decir que "Aceptamos H_0", preferimos la terminología " No se rechaza H_0." No rechazar H_0 implica que no se ha

encontrado evidencia suficiente para rechazar H_0, es decir, para hacer una declaración fuerte. El no poder rechazar H_0 no significa necesariamente que hay una alta probabilidad de que H_0 sea cierta. Se puede simplemente significar que se necesitan más datos para llegar a una conclusión firme. Esto puede tener importantes implicaciones para la formulación de hipótesis.

Un concepto importante es la potencia de una prueba estadística.

La potencia de una prueba estadística es la probabilidad de rechazar la hipótesis nula H_0 cuando la hipótesis alternativa es verdadera.

La potencia se calcula como **1-β**, y puede ser interpretada como la probabilidad de rechazar correctamente una hipótesis nula falsa.

Para una probabilidad dada del error del primer género α, la del error de segundo género β puede disminuirse aumentando el volumen de la muestra.

El error α está bajo control del investigador y se elige o establece antes de realizar la prueba de hipótesis, es el nivel de significación para la prueba de hipótesis, entonces como se puede controlar α , también 1-α está controlada.

Nivel de significación de la prueba

La probabilidad α es el nivel de significación de la prueba, es el riesgo o la probabilidad que el investigador asume de manera voluntaria de equivocarse al rechazar la hipótesis nula, cuando en realidad es verdadera.

Es también la confiabilidad de decidir si se rechaza o no la hipótesis nula. Los niveles de significación más usados son: 0,01, 0,05 y 0,10.

Cuando se rechaza la hipótesis nula, se dice que hay significancia estadística, pero cuando no se rechaza la hipótesis nula significa que "no existe suficiente información como para rechazarla", es errado afirmar que se acepta la hipótesis nula. No se puede aceptar algo que no sabemos si es verdadero o falso.

Que una prueba sea estadísticamente significativa, es decir, rechazar la hipótesis nula no asegura que la hipótesis alternativa sea cierta ante la evidencia de datos muestrales, sino que los datos muestrales discrepan con el supuesto bajo la hipótesis nula.

Recordar que la muestra es aleatoria, también lo son los estadísticos que se usan para someter a prueba hipótesis estadísticas.

Por tanto se recomienda no ser mecanicistas y estar dependiendo del valor α porque lo estadísticamente significativo no siempre es relevante para la investigación.

Ahora se usan los softwares estadísticos como SPSS, MINITAB, SAS, entre otros y es preocupante ver como se los usa de manera indiscriminada, sin sustento, se cree que es solo poner los datos y ver el resultado si es o no estadísticamente significativo. No hay que contentarse con que sea estadísticamente significativo sino que sea relevante para la investigación.

Estadístico de prueba

Para rechazar o no la hipótesis nula se toma una muestra aleatoria de la población bajo estudio y los resultados contenida en ella se usa en expresiones llamadas estadísticos o estadísticas de prueba e indican el grado de discrepancia entre la hipótesis nula y los datos muestrales que están resumidos en las estadísticas.

Cuando la discrepancia es "grande", es decir la evidencia de la muestra (datos muestrales) difiere del valor supuesto para el parámetro bajo la hipótesis nula; se rechaza la hipótesis nula en caso contrario no se rechaza.

La regla de acuerdo a la cual se toma la decisión de aceptar o rechazar la hipótesis H_0 se llama ***criterio*** *k.* Ya que la decisión se adopta basándose en la muestra de observaciones de la variable aleatoria, es necesario elegir un estadístico adecuado que se denomina, en este caso, ***estadístico del criterio o estadístico de prueba***. Se trata, en general, de que el estadístico de prueba sea uno cuya distribución en el muestreo se conozca, en el supuesto de que H_0 sea cierta. Generalmente resulta el estimador del parámetro que se quiere probar en la hipótesis nula.

Región de rechazo

Al conjunto de valores de la estadística de prueba para los que la hipótesis nula se rechaza se llama "región de rechazo o región crítica".

El establecimiento de la región de rechazo depende de la distribución de probabilidad de la estadística de prueba, el punto de corte (punto o valor que divide a la región crítica de la no crítica) se llama también "valor crítico o punto crítico", cuyo su valor depende de la distribución de probabilidad de la estadística de prueba.

El conjunto de todos los valores del estadístico del criterio para los cuales se toma la decisión de rechazar H_0 se denomina **dominio crítico**. El conjunto contrario se llama **dominio de aceptación.**

El valor del estadístico de prueba que separa a ambos conjuntos se llama **valor crítico del estadístico de prueba θc.** Este valor depende del tamaño de la muestra, de α, de la forma de H_1 y de la distribución del estadístico.

Nivel crítico de una prueba de hipótesis (p-value)

Una forma de comunicar los resultados de una prueba de hipótesis consiste en afirmar que la hipótesis nula es aceptada o no a un valor determinado α o nivel de significación.

El valor P es el nivel mínimo de significación que conduciría a un rechazo de la hipótesis nula H_0 con los datos dados.

Es costumbre decir que la estadística de prueba es significativa cuando la hipótesis nula H_0 se rechaza, por lo tanto, podemos pensar en el valor **p** como el mínimo valor en el que los datos son significativos.

El valor *p* es la probabilidad de obtener un estadístico de prueba igual o más exacto que el resultado obtenido a partir de los datos de la muestra, dado que la hipótesis nula es verdadera.

A menudo, al valor *p* se lo conoce como *nivel de significación observado,* que es el mínimo nivel al cual H_0 puede ser rechazada.

- **Si el valor *p* es menor que α, H_0 es rechazada.**
- **Si el valor *p* es mayor o igual que α, H_0 no es rechazada.**

Decisiones posibles en un test de hipótesis

	Aceptar H0	**Rechazar H0**
θ En la región de aceptación	No hay error	Error del tipo I
θ En la región de rechazo	Error del tipo II	No hay error

PASOS PARA LA VERIFICACIÓN DE UNA PRUEBA DE HIPÓTESIS

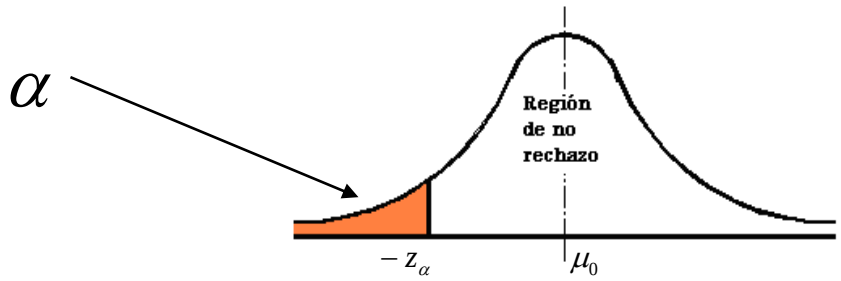
1) Expresar la hipótesis nula.
2) Expresar la hipótesis alternativa.
3) Especificar el nivel de significación, (α).
4) Determinar el tamaño de la muestra.
5) Establecer los valores críticos que dividen las regiones de rechazo y de no rechazo.
6) Seleccionar la prueba estadística según la distribución por muestreo del estadístico que se elige para llevar adelante la prueba referida al parámetro.
7) Coleccionar los datos y calcular el valor muestral de la prueba estadística.
8) Determinar si la prueba estadística ha caído en la región de rechazo o en la de aceptación.
9) Tomar la decisión.
10) Expresar la decisión estadística en términos del problema.

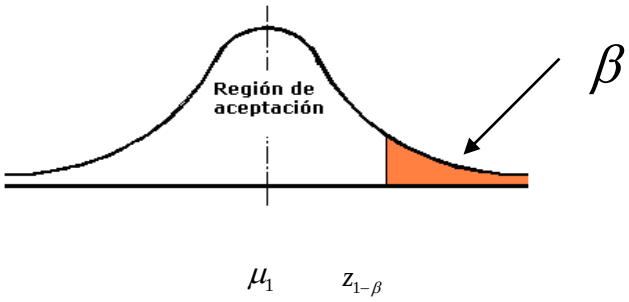
Observación: En las etapas 5 a 8 son utilizados estadísticos cuyos cuantiles se encuentran tabulados (Normal, Student, Chi-cuadrado, F de Snedecor).

La verificación de las hipótesis estadísticas puede realizarse basándose en los intervalos de confianza. En este caso, la hipótesis propuesta se aceptará si el valor θ_0 es cubierto por el intervalo respectivo.

Determinación del tamaño de muestra sin cambiar las probabilidades de los errores de Tipo I y II (válido para cualquiera de los casos que se presentarán)

Frente a α y β predeterminados, es posible encontrar un valor de **n** apropiado.
Supóngase una prueba unilateral por izquierda correspondiente al parámetro μ, con α prefijado y ubicado de acuerdo a H_1:





Luego, H_0 se rechazará si:

$$\bar{x} < \bar{x}_c = \mu_0 + Z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$$

Para el β prefijado, β es la probabilidad de aceptar H_0 cuando lo es H_1; entonces se aceptará H_0 con una probabilidad de (1 - α) sólo si:

$$\bar{x} > \bar{x}_c = \mu_1 + Z_{1-\beta} \cdot \frac{\sigma}{\sqrt{n}}$$

De esta dos últimas ecuaciones puede obtenerse **n**, y luego de cualquiera de ellas el $\bar{x}_c$.

Si se quiere verificar que con **n** y $\bar{x}_c$ determinados anteriormente se mantienen los valores de α y β prefijados, puede realizarse lo siguiente:

$$\alpha = P\ \ (rechazar\ \ \ H_0 / siendo\ H_0\ verdadera) = P\ \ (\bar{x} < \bar{x}_c / \mu = \mu_0) = P\ \left(\frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \leq \frac{\bar{x}_c - \mu_0}{\sigma_{\bar{x}}} \right)$$

$$\beta = P\ \ (aceptar\ \ \ H_0 / siendo\ H_1\ verdadera) = P\ \ (\bar{x} \geq \bar{x}_c / \mu = \mu_1) = P\ \left(\frac{\bar{x} - \mu_1}{\sigma_{\bar{x}}} \leq \frac{\bar{x}_c - \mu_1}{\sigma_{\bar{x}}} \right)$$

Los cuales deberían ser iguales a α y β prefijados.

DISTINTOS CASOS DE PRUEBAS PARAMÉTRICAS

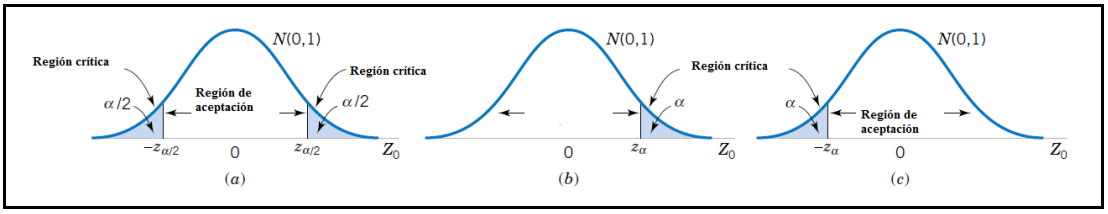
Pruebas de hipótesis referidas a la media de una distribución Normal

Desvío Poblacional Conocido

a. Formulación de las hipótesis.
b. Fijar el nivel de significación; dependiendo del problema tomar 5% o 1%
c. Elegir el estadístico de prueba: en este caso la media muestral.
d. Distribución muestral del estadístico. El estadístico de prueba será:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

e. Establecer las zonas de aceptación y rechazo de acuerdo a como fue formulada H_1.
f. Calcular el estadístico de prueba para la muestra en consideración. De acuerdo a su valor, se toma la decisión según haya caído en una u otra zona.



Desvío Poblacional Desconocido. Muestra Grande (n>30)

El planteo general es similar al caso anterior, solo que la distribución muestral del estadístico media muestral es ahora $N(\mu, S/n^{1/2})$ ya que el desvío es desconocido, pero como la muestra es grande es posible estimarlo por el desvío muestral S.

$$Z = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}$$

Desvío Poblacional Desconocido. Muestra Chica (n<30)

El planteo es nuevamente similar a los casos anteriores; la diferencia está en que el estadístico media muestral se distribuye ahora según una t de Student, ya que como la muestra es chica no se puede considerar al desvío poblacional desconocido igual a S.

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n-1}}}$$

En este caso los puntos críticos corresponden a la distribución muestral t con n-1 grados de libertad por tabla o programa.

La decisión se realiza de acuerdo a las zonas de aceptación y rechazo y al valor del estadístico de prueba.

Prueba acerca de la varianza de la población

Recordar que la distribución muestral de la varianza es chi-cuadrado:

$$\chi^2 = \frac{(n-1)S'^2}{\sigma_0^2}$$

Prueba referida a una proporción poblacional

La distribución muestral de la proporción es Normal:

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

Con:
p = x/n número de éxitos en la muestra / tamaño de la muestra
$\pi =$ Proporción de éxitos en la población.

PRUEBAS DE HIPÓTESIS RESPECTO A DOS PARÁMETROS

Prueba acerca de la igualdad de varianzas de dos poblaciones

Cuando es de interés comparar dos poblaciones, generalmente cuando las muestras son pequeñas, independientes, y provenientes de poblaciones normales o aproximadamente normales, se utiliza la razón de varianzas muestrales como base para esta comparación.

Este procedimiento tiene las siguientes ventajas: La razón es independiente de las unidades (siempre que ambas tengan las mismas unidades), la distribución de esta razón debido a la hipótesis de igualdad de varianza poblacionales, es independiente de los parámetros de las poblaciones.

La razón de varianzas muestrales tiene una distribución F de Snedecor ya que cada una de las varianzas esta relacionada a una variable chi-cuadrado:

$$F_{\nu_1;\nu_2} = \frac{\frac{(n_1-1)\hat{S'}_1^2}{\sigma_1^2}}{\frac{(n_2-1)\hat{S'}_2^2}{\sigma_2^2}}$$

es el estadístico de la prueba.

Generalmente las pruebas son unilaterales y por derecha ya que la hipótesis nula será rechazada para valores de F muy grandes debido a que en el cálculo se coloca la varianza mayor en el numerador.

Prueba respecto a las medias de dos poblaciones Normales. Desvíos parámetros conocidos

Las hipótesis a formular pueden ser las siguientes:

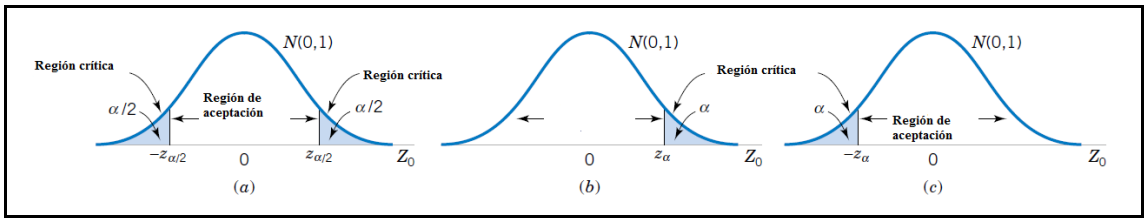
$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$H_0 : \mu_1 > \mu_2 \quad H_1 : \mu_1 \leq \mu_2$

$H_0 : \mu_1 < \mu_2 \quad H_1 : \mu_1 \geq \mu_2$

$H_0 : \mu_1 \leq \mu_2 \quad H_1 : \mu_1 > \mu_2$

$H_0 : \mu_1 \geq \mu_2 \quad H_1 : \mu_1 < \mu_2$



Se fijan luego las probabilidades de los errores de primero y segundo género.

El estadístico a utilizar es la diferencia de medias muestrales, cuya distribución por muestreo es para este caso:

$$\bar{x}-\bar{y} \sim N\left(\mu_x-\mu_y \; ; \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Con lo cual el estadístico de la prueba será:

$$z=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N\left(0 \;,1\right)$$

Este valor una vez calculado deberá compararse con el valor tabulado que surge de la ubicación de las zonas de aceptación y rechazo. Así se puede tomar la decisión de aceptar o rechazar lo propuesto.

Prueba respecto a las medias de dos poblaciones Normales
Desvío poblacionales desconocidos-Muestras grandes (n_x y/o $n_y \geq 30$)

Los pasos son similares al caso anterior, solo que la distribución muestral del estadístico diferencia de medias muestrales cambia al ser las varianzas poblacionales desconocidas y ser estimadas por las muestrales:

$$\bar{x}-\bar{y} \sim N\left(\mu_x-\mu_y \; ; \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

Luego el estadístico de la prueba en base al cual se tomará una decisión será:

$$z=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim N\left(0 \;,1\right)$$

Muestras pequeñas (n_x y/o $n_y < 30$)

Previamente debe verificarse si las varianzas que se desconocen pueden considerarse iguales o no.

Varianzas desconocidas e iguales

En este caso la distribución del estadístico diferencia de medias muestrales es una t de Student.

Debido a que las varianzas muestrales no son estimadores insesgados de las varianzas poblacionales por el tamaño de las muestras:

$$\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{S_w\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}} \sim \quad t_{n_x+n_y-2}$$

Con $S_w=\sqrt{\frac{(n_x-1)\ S_x'^2+(n_y-1)\ S_y'^2}{n_x+n_y-2}}$ *ó* $S_w=\sqrt{\frac{n_xS_x^2+n_yS_y^2}{n_x+n_y-2}}$

El resto de los pasos para obtener una conclusión correcta respecto a la hipótesis planteada son similares a los casos anteriores.

Varianzas desconocidas y distintas

En este caso la distribución muestral del estadístico diferencia de medias muestrales es una t de Student pero se calculan los grados de libertad:

$$t_\nu=\frac{\bar{x}-\bar{y}-(\mu_x-\mu_y)}{\sqrt{\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}}} \quad con \quad \nu=\frac{\left(\frac{S_x'^2}{n_x}+\frac{S_y'^2}{n_y}\right)^2}{\left(\frac{S_x'^2}{n_x}\right)^2\cdot\frac{1}{n_x-1}+\left(\frac{S_y'^2}{n_y}\right)^2\cdot\frac{1}{n_y-1}}-2$$

Se adopta como grados de libertad el entero más próximo al valor anterior.

El resto de los pasos para tomar una decisión respecto a la hipótesis planteada es similar a los casos anteriores.

Prueba respecto a proporciones de dos poblaciones Normales

Una vez establecidas las hipótesis nulas y alternativas que corresponden, fijados α y β se busca el estadístico y su distribución muestral:

$$p_x - p_y \sim N \left(\pi_x - \pi_y \;;\; \sqrt{\frac{\pi_x(1-\pi_x)}{n_x} + \frac{\pi_y(1-\pi_y)}{n_y}} \right)$$

Debido a que las proporciones poblacionales se desconocen se toma como estimador de ellas la media ponderada de las proporciones muestrales, resultando el siguiente estadístico de prueba:

$$z = \frac{p_x - p_y - (\pi_x - \pi_y)}{\sqrt{\hat{p}(1-\hat{p}) \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \quad ; \quad con \quad \hat{p} = \frac{n_x \cdot \hat{p}_x + n_y \cdot \hat{p}_y}{n_x + n_y}$$

Ejemplo

Para determinar que influencia ejerce la temperatura del medio ambiente en el error sistemático de un instrumento que mide ángulos, se han efectuado mediciones del ángulo horizontal de un objeto, durante la mañana (t = 10° C) y durante el resto del día (t = 26° C). Los resultados de las mediciones del ángulo (en segundos angulares) son:

De mañana	38.2	36.4	37.7	36.1	37.9	37.8		
De tarde	39.5	38.7	37.8	38.6	39.2	39.1	38.9	39.2

Si los valores provienen de una distribución normal y conociendo la distribución de los errores, ¿se puede considerar que la temperatura ambiente influye en el error sistemático del instrumento?

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$

$\sigma_1 \;\; y \;\; \sigma_2 \;\; desconocidos, \quad n_1 \;\; y \;\; n_2 \;\; < \;\; 30$

Corresponde previamente realizar el test de igualdad de varianzas, porque de este dependerá la prueba de igualdad de medias.

$H_0 : \sigma_1^2 = \sigma_2^2 \quad H_1 : \sigma_1^2 \neq \sigma_2^2$ **α=0,10**

$\overline{x_1} = 37.35 \quad \overline{x}_2 = 38.87 \quad S_1 = 0.636 \quad S_2 = 0.628 \quad S'^2_1 = 0.76 \quad S'^2_2 = 0.72$

$$F = \frac{S'^2 \;\; mayor}{S'^2 \;\; menor} \qquad F_c = \frac{0.76}{0.72} = 1.05$$

$F_{(5;7;0.05)} = 3.97 \quad F_c < F_{tabulado}$

Según estos valores obtenidos las varianzas pueden considerarse iguales. De acuerdo a esto se plantea el test de igualdad de medias y se verifica la misma.

Prueba respecto a las medias de dos poblaciones con muestras dependientes

En los casos anteriores, se compararon medias de dos poblaciones y se suponía que las muestras que se extraían eran de poblaciones Normales e independientes. Esto significa que cada una no se relaciona con algún elemento de la otra de forma significativa.

Dependencia significa que están emparejadas en el sentido que cada observación en una se asocia con alguna observación en la otra. **Si dos muestras son dependientes deberán tener el mismo número de elementos.**

Hay situaciones en las que puede haber necesidad de comparar dos procedimientos diferentes y en las que se sospecha un cambio de nivel en la variable de respuesta, que no se puede controlar. Ya que el nivel de la medición puede variar de muestra a muestra es posible que las dos observaciones no sean independientes, por el contrario que estén relacionadas. Por ejemplo: si se tienen dos instrumentos para medir precipitación, se quiere saber si son o no equivalentes los resultados de ambos, problema muy común en el caso de comparación de datos.

En esta situación la prueba de comparación de medias se realiza de forma similar a los casos anteriores. Para derivar el estadístico de prueba adecuado se obtienen la diferencia entre las observaciones emparejadas para cada elemento d_i, que pertenecen a una nueva variable aleatoria **D**: diferencia en la población.

Podría decirse si las dos muestras no son diferentes entre sí, entonces el valor medio de **D** sería igual a 0. Un supuesto importante a realizar es que las diferencias se distribuyen normalmente.

Es posible a partir de los datos obtener la media de los n pares formados d_i; el valor del error estándar de esta media será:

$$\frac{S_d}{\sqrt{n-1}}$$

Y la razón de $\bar{d}$ a su error estándar se distribuye según una t de Student con n – 1 grados de libertad:

$$t_{n-1} = \frac{\bar{d} - \bar{D}}{\frac{S_d}{\sqrt{n-1}}}$$

con $\overline{D} = 0$.

Este es el estadístico de la prueba. Los pasos posteriores son similares a los casos ya desarrollados; es decir una vez calculado el valor del estadístico de prueba se deberá ubicar en que región cae, si en la de aceptación o la de rechazo.

RELACIÓN ENTRE LOS INTERVALOS DE CONFIANZA Y LOS TEST DE HIPOTESIS

Los intervalos de confianza se utilizan para estimar parámetros, mientras que los test de hipótesis se utilizan para tomar decisiones en relación a parámetros de la población especificada.

En muchos casos los intervalos de confianza y los test de hipótesis se pueden usar en forma indistinta.

Si al calcular el intervalo para el parámetro, no se rechazaría la hipótesis nula planteada si el intervalo contiene el valor que se quiere probar.

TEST NO PARAMÉTRICOS

Usos de la distribución Chi-cuadrado

En lo presentado hasta este punto se desarrollaron métodos referidos a test de hipótesis paramétricos, es decir donde la hipótesis H_0 era referida al valor de un parámetro dado θ.

Existe una clase de problemas en la cual la proposición a testar se refiere a la forma de la distribución postulada F(x). Estas hipótesis se denominan *no paramétricas*.

La distribución chi-cuadrado que se ha utilizado como base en las inferencias relativas a la varianza poblacional, tiene otros usos:

- Comprobar si un conjunto de datos observados coincide o no con un conjunto de datos esperados (Prueba de frecuencia);
- Verificar la ley de distribución de una población (Prueba de la bondad del Ajuste)
- Comparar más de dos varianzas (Prueba de homogeneidad);
- Verificar la independencia de dos variables aleatorias (Pruebas e independencia)

En estos test no paramétricos la hipótesis alternativa H_1 no está especificada explícitamente. El nivel de significación α (probabilidad de cometer un error de tipo I), es elegido como una cantidad de decisión y depende de que la probabilidad de obtener el evento observado sea mayor o menor que α, para aceptar o no la hipótesis propuesta; por ésta razón es que son test unilaterales por derecha.

Un valor de α pequeño resultará en un test menos crítico o sea es muy probable que se acepte H_0.

PRUEBA DE FRECUENCIAS

En este caso se desea comprobar si un conjunto de datos coincide o no con un conjunto de datos esperados o frecuencias teóricas.

El estadístico de prueba será:

$$\chi^2_{n-k} = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

De igual forma que en otros ensayos se fijará un riesgo α y se comparará el estadístico calculado con un valor tabulado.

Siendo fo la frecuencia de realización de un acontecimiento determinado, de acuerdo a H_0, y fe la frecuencia esperada.

La suma de los cocientes sigue aproximadamente la distribución chi-cuadrado si es que no hay diferencias entre las frecuencias esperadas y las observadas.

Se establece que fe = n*pi, frecuencias esperadas, deben ser mayores o iguales que 5 para que el estadístico siga una distribución chi-cuadrado, si esto no sucede, se las puede combinar hasta lograrlo.

Cuando los grados de libertad que resultan luego del agrupamiento, que son iguales a n-1, son iguales a 1, se deberá realizar una corrección restando ½ a la diferencia del denominador. Se aplica cuando **n** es pequeño, ya que si **n** es grande su efecto es despreciable:

PRUEBA DE BONDAD DE AJUSTE

En este punto se combinarán algunas ideas presentadas previamente respecto a la naturaleza de los datos observados con las técnicas de test de hipótesis ya desarrolladas. En este caso la hipótesis nula postula que x tiene la ley de distribución F(x).

Generalmente lo que se postula es la *forma* de la distribución, pero se deben estimar los parámetros a partir de los datos; el efecto de esta estimación es reducir los grados de libertad del estadístico.

La aplicación del criterio χ^2 para verificar la hipótesis nula se compone de las siguientes etapas:

a- Basándose en la muestra de observaciones de la variable aleatoria x, es necesario hallar las estimaciones de los parámetros desconocidos de la ley de distribución hipotética F(x).

b- Si x es una variable aleatoria discreta, hay que determinar las frecuencias fi con la cuales cada valor o grupo de valores se encuentra en la muestra.

Si x es una variable aleatoria continua, es necesario partir el dominio de sus valores en r intervalos disjuntos: Δ_1, Δ_2, Δ_3,…. Δ_r y determinar el número de elementos de la muestra, fi, que pertenecen a cada intervalo.

c- Si x es una variable discreta utilizando la ley de distribución hipotética F(x), es necesario calcular las probabilidades Pk, con las cuales la variable aleatoria x toma cada valor.
Si x es una variable aleatoria continua conviene determinar la probabilidad P_k de tomar un valor perteneciente a cada intervalo Δ.

d – Calcular el valor muestral del estadístico:

$$\chi^2_m = \sum_{i=1}^{k} \frac{(fo - fe)^2}{fe}$$

e – Tomar la decisión estadística: La hipótesis nula no contradice a la muestra para el nivel de significación α, si $\chi^2_m < \chi^2_{1-\alpha/2\ (k-L-1)}$, siendo L el número de parámetros de la distribución F(x) que se estiman con ayuda de la muestra; en caso contrario se rechaza la hipótesis propuesta.

CASOS ESPECÍFICOS

Variable discreta

Modelo Binomial

Los pasos generales son:

a) H_0: f(x) = Binomial:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

Basados en la muestra, será necesario estimar el parámetro p del modelo.

b) Determinar, en base a los datos, las frecuencias observadas.
c) Utilizando el modelo postulado, se obtienen las probabilidades para cada valor de la variable.
d) Se obtienen las diferencias entre las frecuencias observadas y las frecuencias esperadas, siendo estas últimas $f_e = n.p_i$ para obtener el valor del estadístico del test.

e) Se obtiene de tablas el valor de χ^2 crítico, de acuerdo a $\nu = k - L - 1$ y α. Concluir en base a estos valores, si se acepta o no que los datos se ajustan al modelo postulado.

Modelo de Poisson

Con un ejemplo, se desarrollarán los pasos para desarrollar el test.

Considere la ocurrencia de tormentas con granizo que se presentan en una región de la provincia de Santa Fe, mostradas en la siguiente tabla:

N° de tormentas por año	0	1	2	3	4	5	6
N° de ocurrencias observadas	102	144	74	28	10	2	0

La hipótesis propuesta es que las tormentas que ocurren se consideran independientes, tienen un promedio de ocurrencia, y estas ocurrencias son de tipo Poisson, de parámetro **λ**

$$f(x)= P(X = x)= \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

N° de tormentas por año	N° de ocurrencias observadas		$P(X = x) = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$	Fe=n*P		$\frac{(fo - fe)^2}{fe}$
0	102		0.3073	110.62		0.67
1	144		0.3626	130.53		1.39
2	74		0.2139	77.01		0.12
3	28		0.0841	30.29		0.17
4	10	12	0.025	8.94	11.46	0.025
5	2		0.0059	2.11		
6	0		0.0011	0.41		
				359.91		χ^2_{calc} =2.375

Se acepta la hipótesis propuesta ya que el valor calculado es menor al de tabla que es $\chi^2_{5\text{-}1\text{-}1;\ 0,05}$=7.81

Variable contínua

Modelo Normal

Los pasos a seguir son similares a los casos anteriores, sólo que para obtener las frecuencias esperadas se deberá obtener la probabilidad en un intervalo, de acuerdo a lo propuesto en H_0, en base al modelo Normal.

A partir de los datos de precipitación en 55 días de un año, se quiere comprobar la hipótesis nula que pertenecen a una población con distribución Normal.

17	19	23	18	21	15	16	13	20	18	15
20	14	20	16	14	20	19	15	19	16	19
15	22	21	12	10	21	18	14	14	17	16
13	19	18	20	24	16	20	19	17	18	18
21	17	19	17	13	17	11	18	19	19	17

Ancho de intervalo: c=2; n=55 ; $\bar{x}$ =17.84 mm
S = 2.92 mm

Intervalo	f.obs.		Li est.	Ls est.	F(Li)=a	F(Ls)=b	P=b–a	Fe= n.P	$\frac{(fo-fe)^2}{fe}$
10-12	2	6	- ∞	-2	0	0.0228	0.0228	1.254	0.145
12-14	4		-2	-1.32	0.0228	0.0934	0.0706	3.883	
14-16	8		-1.32	-0.63	0.0934	0.2643	0.1709	9.39	0.21
16-18	12		-0.63	0.055	0.2643	0.5199	0.2556	14.06	0.30
18-20	16		0.055	0.74	0.5199	0.7704	0.2505	13.78	0.36
20-22	10	13	0.74	1.42	0.7704	0.9222	0.1518	8.35	0.011
22-24	3		1.42	∞	0.9222	1	0.0778	4.28	
									χ^2_{calc} =1.026

Ya que $\chi^2_m < \chi^2_t$, la hipótesis sobre distribución de los datos no contradice los resultados observados.

TEST DE BONDAD DE AJUSTE DE KOLMOGOROFF

Otro test cuantitativo que se utiliza para verificar si los datos se ajustan a un modelo propuesto es el de Kolmogoroff.

Primero se ordena la muestra:

$x(1) \leq x(2) \leq \ldots.. \leq x(n)$

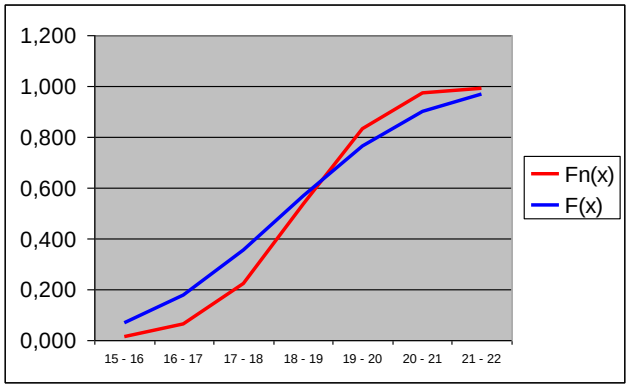
Luego puede definirse la función de cuantía muestral considerando la proporción de valores que no exceden a un valor dado x, $G_n(x)$: r/n.

Ya que r s el número de observaciones que son menores o iguales que x_r, r puede ser considerado como la suma de variables Bernoulli con $p = F(x_r)$.

El Teorema del Límite Central implica que la distribución límite de r/n es Normal, con valor esperado p.

Es así que el test para el ajuste puede ser construido sobre una medida apropiada de la desviación de $G_n(x)$ de F(x). Esta medida es el valor absoluto de la máxima diferencia entre $G_n(x)$ y F(x), y es lo que se define como el *estadístico de Kolmogoroff*.

$$D_n = máx\left|G_n(x) - F(x)\right|$$



La distribución de este estadístico no depende de F(x):

$$\lim_{n \to \infty} \qquad P\left(D_n > \frac{C}{\sqrt{n}}\right) = 2\sum (-1)^{m-1} e^{-2m^2C^2}$$

este test es independiente de la distribución.

Han sido tabuladas, para distintos valores de n, las probabilidades exactas
P ($D_n > C/n^{1/2}$) o su complemento, así como los valores críticos $\mathbf{D_n}$ para los cuales la probabilidad es 0.01 ó 0.05. Dos valores críticos asintóticos muy importantes son:

$$\alpha = 0.01 \qquad d_n = \frac{1.63}{\sqrt{n}}$$

$$\alpha = 0.05 \qquad d_n = \frac{1.36}{\sqrt{n}}$$

Los cálculos para desarrollar el test son muy simples, casi siempre un examen preliminar de los datos puede revelar áreas para las cuales las desviaciones son probablemente mayores. Una vez obtenido el valor absoluto de la máxima diferencia, se obtiene el valor crítico del estadístico tabulado para tomar una decisión: aceptar o no el modelo propuesto.

La comparación entre los dos test permite analizar los siguientes puntos:

- El test de Kolmogoroff es más sencillo en cuanto a cálculos.
- No pierde información por agrupamiento y puede ser aplicado a muestras pequeñas.
- El test de Chi-cuadrado puede ser aplicado a datos discretos y continuos, y también a hipótesis compuestas, es decir, cuando además de probar un modelo,

deban estimarse sus parámetros a partir de los datos, mientras que el de Kolmogoroff se aplica sólo a datos continuos.
- Los dos test son comparables en términos de potencia, y existe alguna evidencia de que el test de Kolmogoroff es algo más potente que el de Chi-cuadrado.

Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 7
REGRESIÓN Y CORRELACIÓN

Muchos problemas en la ingeniería y la ciencia implican la exploración de las relaciones entre dos o más variables.

Se va a considerar la diferencia entre relación funcional y relación estadística:

Relación funcional entre dos o más variables se expresa por una fórmula matemática, por ejemplo en el cálculo de la velocidad de caída de un cuerpo, $v = (2gh)^{1/2}$, conociendo la altura de caída h y la gravedad del lugar, es posible obtener un valor exacto de velocidad; es por lo tanto una relación **determinística**.

Relación estadística, si se ajusta una curva a observaciones, existe variación de los puntos en torno de la curva que relaciona a las variables. Generalmente es posible encontrar una relación media con cierto grado de precisión.

El estudio de la asociación entre variables se hace a través de dos aspectos:

Análisis de regresión: es el que permite encontrar el modelo que vincula a las variables en cuestión, brindando así un mecanismo de pronóstico.

Análisis de correlación: determina la medida del grado de exactitud de la relación entre variables.

Se analizará el caso de asociación lineal simple entre dos variables, pero pueden darse casos más complejos como por ejemplo relación no lineal, relación múltiple, etc.

El modelo de regresión es una manera de expresar formalmente los aspectos esenciales de la relación estadística entre las variables:

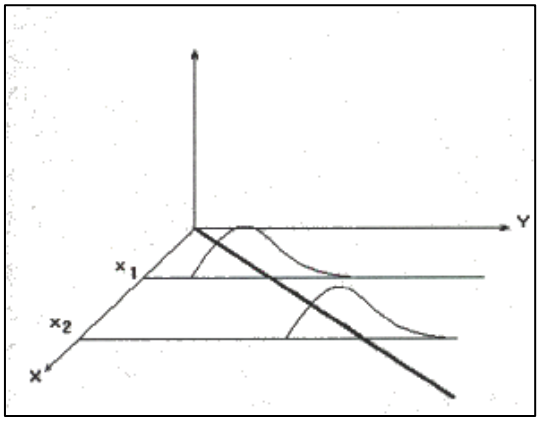
- La tendencia de la variable **y** (dependiente) a variar con la variable independiente de una manera sistemática.

- La dispersión de los puntos entorno a la curva que relaciona las variables.

Estos aspectos son contenidos en el modelo de regresión por enunciar lo siguiente:

- Existe una distribución de probabilidades de Y para cada valor de X: las variables X son fijas, es decir no aleatorias, mientras que las Y si lo son. Existen grupos de valores de Y para cada valor de la variable fija X, que se denominan **subpoblaciones**. Dependiendo del tipo de distribución de las variables aleatorias Y se clasifican en: -población tipo I es el caso en que la distribución de Y en cada subpoblación no está especificada; en el caso en que la distribución de Y en cada subpoblación es Normal se denomina población tipo II.

Las medias de estas distribuciones de probabilidad varían con la variación de x (variable fija).



Modelo de regresión bivariado

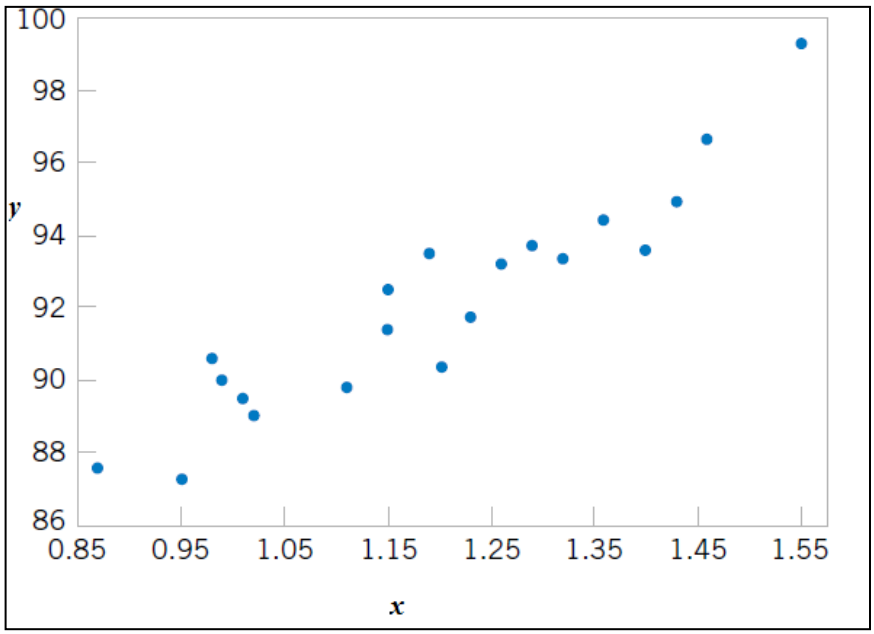


Figura N°1- Dispersiograma

La inspección del diagrama de dispersión indica que no hay una curva que pase exactamente por todos los puntos, pero es un fuerte indicio de que los puntos se encuentran dispersos al azar en torno a una línea recta. Por lo tanto, es razonable asumir que la media de la variable aleatoria Y está relacionada con x por la siguiente relación lineal:

$$E(Y/x) = \mu_{Y/x} = \alpha + \beta X_i \qquad (1)$$

α y **β** coeficientes de regresión.

Mientras que la media de Y es una función lineal de x, el valor observado y real no cae exactamente sobre una recta. La forma más adecuada para generalizar a un modelo lineal probabilístico es asumir que el valor esperado de Y es una función lineal de x, pero que para un valor fijo de x el real valor de Y está determinada por el valor medio de la función (el modelo lineal) más un término de error aleatorio, por ejemplo:

$$Y_i = \underbrace{\alpha + \beta X_i}_{I} + \underbrace{\varepsilon_i}_{II} \qquad (2)$$

donde ε_i es el término de error aleatorio.

siendo I la parte sistemática y II la parte estocástica, que hace que Y no pueda ser pronosticado exactamente como sucedería en un caso determinístico.

α y β son los parámetros del modelo.
Xi es la variable independiente, fija, conocida, variable explicativa.

Vamos a llamar a este modelo, **modelo de regresión lineal simple**, ya que cuenta con una sola variable independiente.

A veces, un modelo como este surgirá de una relación teórica. En otras ocasiones no vamos a tener ningún conocimiento teórico de la relación entre x e y, y la elección del modelo se basa en la inspección del diagrama de dispersión.

Para conocer más sobre este modelo, suponemos que podemos fijar el valor de x y observar el valor de la variable aleatoria Y. Ahora bien, si x es fijo, el componente aleatorio del lado derecho del modelo de la Ecuación (2) determina las propiedades de la variable dependiente Y.

Supongamos que la media y la varianza de ε son 0 y σ^2, respectivamente. Entonces:

$$E(Y/x) = E(\alpha + \beta x + \varepsilon) = \alpha + \beta x + E(\varepsilon) = \alpha + \beta x$$

Tener en cuenta que esta es la misma relación que en un principio escribimos empíricamente a partir de la inspección del diagrama de dispersión en la Figura Nº 1.

La varianza de Y dado x es:

$$Var(Y/x) = Var(\alpha + \beta x + \varepsilon) = Var(\alpha + \beta x) + Var(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

Por lo tanto, el modelo de regresión es una línea de valores medios, es decir, el valor de la línea de regresión en cualquier valor de x es sólo el valor esperado de Y para que x.

La pendiente, puede ser interpretada como el cambio en la media de Y para un cambio unitario en x. Por otra parte, la variabilidad de Y para un valor particular de x se determina por la varianza del error $\in$, σ^2. Esto implica que hay una distribución de los valores Y para cada x, y que la varianza de esta distribución es la misma en cada x.

Significado de los parámetros α y β

α: intercepción de la línea de regresión con en eje Y.

β: pendiente de la recta, proporción de cambio en la media de la distribución de probabilidades de Y por unidad de incremento de X.

Es de destacar que el sentido con que son utilizados los términos **dependiente** e **independiente**, no es el mismo que el de dependencia e independencia de variables aleatorias.

Se dice que el modelo es de regresión simple cuando hay dos variables asociadas. Si esto no ocurre, el modelo es de regresión múltiple.

Se dice que el modelo es lineal si los parámetros y la variable independiente están elevados a la primera potencia. Si no es así, el modelo será no lineal.

¿Cómo puede asegurarse que el término de error $\in_i$ sea normalmente distribuido? Invocando el Teorema del Límite Central. Cuando existe una gran cantidad de causas independientes contribuyendo cada una con un pequeño efecto, la distribución de su suma es normal. En muchos casos en los que se aplica el análisis de regresión, las variables están influenciadas por un gran número de pequeños efectos independientes, por esto puede invocarse este teorema y justificar así la normalidad del término de error.

El modelo asume que la distribución de Y tiene igual varianza que el término de error, independientemente del valor de X. Esta propiedad se denomina *homoscedasticidad*.

Se asume independencia entre los términos de error. Esto significa que el resultado en alguna prueba no tiene efecto sobre el término de error de alguna otra prueba. ε no correlacionado con ε implica que Y_i no está correlacionado con Yj.

La completa especificación del modelo de regresión no solo incluye la forma del modelo (ecuación de regresión), sino una expresión de cómo son determinados los valores de la variable independiente y una especificación de la distribución de ε.

Cambiando los supuestos referidos a $\in$ y a X se obtienen distintos modelos de regresión. Por ejemplo, decir:

a) **ε** es una variable aleatoria independiente.

b) **ε** es una variable aleatoria, pero no independiente.

c) La distribución de **ε** no está especificada.

d) La distribución de **ε** es normal.

e) X es un conjunto de números fijos.

f) X es una variable aleatoria, pero su distribución no está especificada.
g) X es una variable aleatoria con distribución normal.

El modelo que se desarrollará considerará los supuestos a), d) y e) y, por lo tanto, se denominará modelo de regresión que tiene una población de Tipo II.

El modelo correspondiente a población Tipo I se basa en los supuestos a), c) y e)

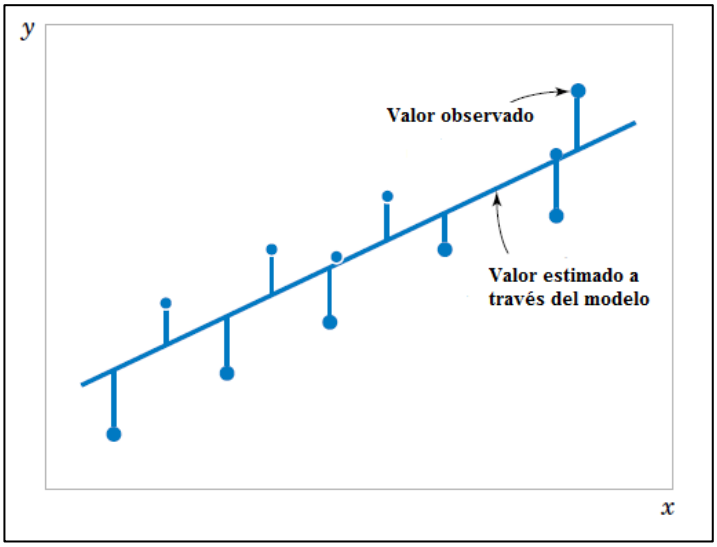
El modelo correspondiente a población Tipo III se basa en les supuestos a), c) y f).

El modelo correspondiente a población Tipo IV se basa en los supuestos a), d) y g).

La distribución de la variable Y, al ser una función lineal de $\in$, presenta su misma distribución.

Se trata de obtener el mejor estimador insesgado lineal del modelo planteado. El método empleado para esto es el de mínimos cuadrados. Una de las razones de su uso es la sencillez de su tratamiento matemático y, además, las estimaciones de α **y** β que produce son idénticas a las obtenidas por el método de máxima verosimilitud.

Estimación por el método de mínimos cuadrados



Se parte de considerar que la subpoblación de Y es normal, y que la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta verdadera sea mínima:

$$S = \sum_{i=1}^{n} \left[Y_i - (\alpha + \beta X_i) \right]^2$$

$$S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2, con\_\hat{Y}_i = a + bX_i$$

$$luego \quad S = \sum_{i=1}^{n} \left[Y_i - (a + bX_i) \right]^2$$

Los estimadores de α **y** β serán aquellos que minimicen el valor de S:

$$\frac{\partial S}{\partial \alpha} = 0 \quad \text{y} \quad \frac{\partial S}{\partial \beta} = 0$$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} X_i (Y_i - \alpha - \beta X_i)$$

Igualando a cero, aplicando sumatoria a todos los términos y reemplazando a α **y** β por a y b, se obtienen las siguientes ecuaciones normales, que conducen a obtener los estimadores a y b:

$$\begin{cases} \sum_{i=1}^{n} Y_i - na - b \sum_{i=1}^{n} X_i = 0 \\ \sum_{i=i}^{n} X_i Y_i - a \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} X_i^{\,2} = 0 \end{cases}$$

De la primera ecuación: $a = \bar{Y} - b\bar{X}$

Sustituyendo a en la segunda ecuación, se obtiene:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

$$b = \frac{\text{COV}}{S_x^2}$$

Luego la recta de regresión es $\hat{Y} = \hat{\alpha} + \hat{\beta}\, x$ (3)

Propiedades de los estimadores

Pueden considerarse a **a** y **b** como combinación lineal de las Y, que tienen distribución normal:

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

Las propiedades estadísticas de los estimadores de mínimos cuadrados $\hat{\alpha}\,;\hat{\beta}$ se pueden describir fácilmente.

Recordemos que hemos supuesto que el término de error $\in$ en el modelo Y es una variable aleatoria con media cero y varianza σ^2. Dado que los valores de x son fijos, Y es una variable aleatoria con media $\mu_{Y/x} = \alpha + \beta x$ y varianza σ^2. Por lo tanto, los valores de $\hat{\alpha}$ y $\hat{\beta}$ dependen de los valores observados, por lo que los estimadores de mínimos cuadrados de los coeficientes de regresión pueden ser considerados como variables aleatorias. Vamos a investigar las propiedades de los estimadores.
Debido a que β es una combinación lineal de las observaciones Yi, podemos utilizar las propiedades de la esperanza (Ver Anexo Cap. VII) para demostrar que el valor esperado de β es:

$$E(b) = \beta \qquad (4)$$

Por lo tanto **b** es un estimador insesgado de β, y su distribución es Normal ya que se consideró a b como combinación lineal de variables normales independientes.

Para obtener las varianza el análisis se basa en el hecho de que las Y_i son variables independientes, cada una con varianza σ^2, y que las K_i utilizadas antes son constantes, pues dependen de X_i y $\overline{X}$:

$$\sigma^2(b) = \sigma^2\left(\sum_{i=1}^{n} K_i Y_i\right) = \sum_{i=1}^{n} K_i{}^2 \sigma^2(Y_i)$$

$$como$$

$$\sigma^2(Y_i) = \sigma^2(\in) = \sigma^2, entonces$$

$$\sigma^2(b) = \sum_{i=1}^{n} K_i{}^2 \sigma^2$$

$$\sigma^2(b) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad (5)$$

Se realizará algo similar para obtener la esperanza del estimador de α:

$$E(a) = E(\overline{Y} - b\overline{X}) = E(\overline{Y}) - E(b\overline{X}) = \alpha + \beta\overline{X} - \beta\overline{X} = \alpha$$

$$E(a) = \alpha \qquad (6)$$

Se comprueba así la insesgabilidad del estimador **a**, siendo su distribución Normal por ser también combinación lineal de variables normales independientes.

Se demuestra (Ver Anexo Cap VII) que la Varianza del estimador a es:

$$\sigma^2(a) = E(a - \alpha)^2$$

$$\sigma^2(a) = \sigma^2\left[\frac{1}{n} + \overline{X}^2 \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}\right]$$

$$\sigma^2(a) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] \qquad (7)$$

Han sido así obtenidos los valores medios y varianzas de los estimadores puntuales de la pendiente y ordenada al origen del modelo de regresión lineal simple.

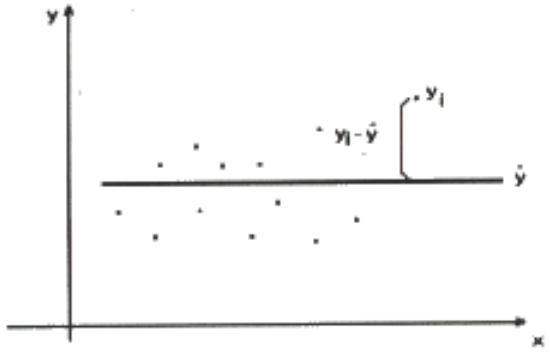
Varianza de la regresión

Se suele llamar ecuación predictiva a la ecuación de regresión muestral, ya que su principal objetivo es predecir valores medios de la variable dependiente asociados con un valor dado de la variable independiente. Pero para saber si es realmente conveniente utilizar esta ecuación para predicción, puede analizarse la variabilidad del valor pronosticado a través del modelo de regresión.

Una primera manera de analizar esta variabilidad puede ser a través de la inspección visual por trazar en el diagrama de puntos la recta obtenida. La medida numérica de la desviación de las observaciones respecto al modelo es el estimador de la varianza de la regresión de la población: $S^2_{y/x}$.

El análisis de la varianza de regresión se basa en la partición de la suma de cuadrados.

La variación de la variable dependiente Y_i generalmente se mide en términos de las desviaciones respecto al valor medio:

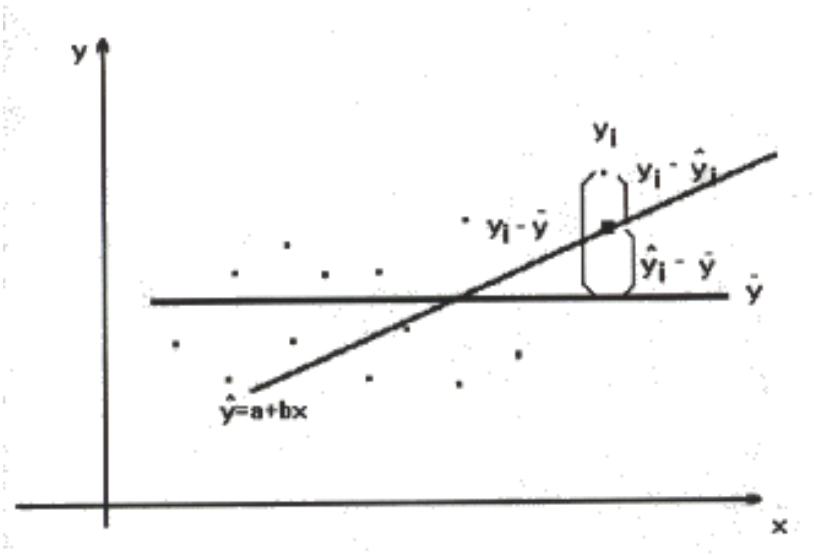


La medida de variación total es para todos los puntos:

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

Cuanto mayor es este valor, mayor es la variación de la curva ajustada respecto a las observaciones.

Utilizando el modelo ajustado, *la variación se da por la diferencia de los valores observados con los valores ajustados o estimados:*



$$Y_i-\hat{Y}_i$$

Por lo tanto la variación total será: **SSE** suma de desvíos cuadrados o suma de errores cuadrados:

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)^2$$

Si este valor es igual a cero todos los puntos caen sobre el modelo ajustado; cuanto mayor sea, mayor será la variación o dispersión alrededor de la recta.

Particionando la suma total, o sea la dispersión respecto al valor medio, se obtiene:

$$\left(Y_i-\bar{Y}\right)=\left(\hat{Y}_i-\bar{Y}\right)+\left(Y_i-\hat{Y}_i\right)$$

siendo:

$Y_i-\bar{Y}$ *: desviación total*

$\hat{Y}_i - \overline{Y}$ *: desviación de la recta respecto al valor medio*

$Y_i - \hat{Y}_i$ *: desviación respecto a la línea ajustada*

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \left(\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}i\right)^2\right)$$

desarrollando el último término de la suma:

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - \left(\hat{Y}_i - \overline{Y}\right)$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \left(a + b\overline{X}\right)\right) - \left(a + bX_i - \overline{Y}\right)$$

siendo $a = \overline{Y} - b\overline{X}$, $\overline{Y} = a + b\overline{X}$, $\hat{Y}_i = a + bX_i$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - \left(\left(a + bX_i\right) - \left(a + b\overline{X}\right)\right)$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - a - bX_i + a + b\overline{X}$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - b\left(X_i - \overline{X}\right)$$

Como se quiere encontrar una expresión para la varianza de la estimación, a la expresión anterior se le deberá aplicar el operador esperanza y elevarla al cuadrado, y esto para todos los puntos, es decir sumatoria. Luego de estos pasos matemáticos, se obtiene la expresión de la varianza de la regresión:

$$S^2_{y/x} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} \qquad (8)$$

La distribución en el muestreo de este estimador, recuérdese que la varianza estaba relacionada con la variable χ^2:

$$\chi^2_{n-2} = \frac{(n-2)S^2_{y/x}}{\sigma^2}$$

Esto sirve para realizar inferencias (construir intervalos de confianza o realizar test de hipótesis) respecto a los parámetros de la ecuación de regresión.

Predicción y pronóstico

Como ya se dijo, uno de los objetivos principales del análisis de regresión es la predicción. Pero es importante hacer notar claramente la diferencia entre predicción y pronóstico:

***Predicción*:** es la estimación del valor medio de Y dado un valor particular de X:

$$\hat{Y}_h = a + bX_h$$

a + bX es el estimador insesgado de α + βX, y su distribución es Normal, ya que es una combinación lineal de variables aleatorias normales. Por lo tanto, para poder encontrar un intervalo para cualquier punto de la recta de regresión poblacional, faltaría encontrar la varianza o error de la regresión. Para este caso, la variación depende de la variación en ambos estimadores, a y b:

$$\sigma^2\left(\hat{Y}_h\right) = \sigma^2\left(a + bX_h\right) = \sigma^2\left(\overline{Y} - b\overline{X} + bX_h\right) = \sigma^2\left(\overline{Y} + b(X_h - \overline{X}\right)$$

Como $\overline{Y}$ y b son variables independientes, y X_h y $\overline{X}$ son constantes, es posible hallar la varianza por términos:

$$\sigma^2\left(\hat{Y}_h\right) = \frac{\sigma^2}{n} + \sigma^2\left(b\left(X_h - \overline{X}\right)\right) =$$

$$= \frac{\sigma^2}{n} + \left(X_h - \overline{X}\right)^2 \sigma^2\left(b\right) =$$

$$= \frac{\sigma^2}{n} + \left(X_h - \overline{X}\right)^2 \sigma^2 \frac{1}{\sum_i \left(X_i - \overline{X}\right)^2} = \qquad (9)$$

$$\sigma^2\left(\hat{Y}_h\right) = \sigma^2\left(\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right)$$

Siendo σ^2 la varianza de la regresión.

***Pronóstico*:** es la proyección de un solo valor de Y correspondiente a un valor de X particular:

$$\tilde{Y} = a + bX_h \qquad (10)$$

Se ve que las expresiones son las mismas, pero existen diferencias:

σ^2 *(error de pronóstico): consta de dos partes:*

1- σ^2 *del error de predicción (ya analizada)*

2- σ^2 *debida a errores casuales, tenida en cuenta por* σ^2

$$\sigma^2\left(Y_i - \hat{Y}_h\right) = \sigma_p^2 + \sigma^2 = \sigma^2\left(\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right) + \sigma^2$$

$$\sigma^2\left(Y_i - \hat{Y}_h\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right) \qquad (11)$$

Regresión no lineal

La regresión lineal no siempre da buenos resultados ya que veces la relación entre *Y* e *X* no es lineal. La estimación directa de los parámetros de funciones no-lineales es un proceso bastante complicado. No obstante, a veces se pueden aplicar las técnicas de regresión lineal por medio de transformaciones de las variables originales.

Por ejemplo los resultados de este análisis pueden proporcionar una buena indicación sobre el comportamiento de los costos para un banco "típico", aunque la naturaleza misma de un estudio de este tipo no puede arrojar resultados estrictamente aplicables a cada uno de los bancos considerados individualmente. No obstante, a pesar de esto, un estudio de este tipo de todas maneras puede ser muy útil, porque los resultados pueden proporcionar una "norma" o "estándar" contra el cual se pueden comparar los costos administrativos en un banco particular. En ausencia de un estudio de este tipo, un banco no tiene realmente un criterio para determinar si sus costos son "muy elevados," "aceptables," o "normales," ya que los bancos difieren enormemente en cuanto a cantidad de activos, número de sucursales, etc., de modo que el único criterio objetivo sería el de compararse con un banco de similar tamaño y características. Sin embargo, si se pudiera obtener una fórmula empírica que permita calcular un valor "normal" o "promedio" para los costos administrativos en función de unas pocas variables que permitan una medición numérica, entonces se podría fácilmente determinar si el banco en cuestión está "mejor" o "peor" que el banco "típico" a ese respecto.

Una función no-lineal que tiene muchas aplicaciones es la *función potencial*:

$$Y = A * X^b$$

donde *A* y *b* son constantes desconocidas. Si se aplica logaritmos, esta función tambiénpuede ser expresada como:

log(Y) = log(A) + *b* .log(X)

Considerando ahora la siguiente regresión lineal:

$\log(Y) = b_0 + b_1 \log(X)$

En esta regresión (denominada *regresión doble-log*), en lugar de calcular la regresión de Y en X, calculamos la regresión del *logaritmo* de *Y* vs el *logaritmo* de X.

Comparando estas dos ecuaciones, podemos apreciar que el coeficiente b_0 es un estimador de log(A), mientras que b_1 es un estimador de *b* (el exponente de la función potencial). Este modelo es particularmente interesante en aplicaciones econométricas, porque el exponente *b* mide la *elasticidad* de Y respecto de X.

CASO APLICADO

Desempleo y Crecimiento Económico

En 1963 el economista norteamericano Arthur M. Okun planteó un modelo macroeconómico para explicar la relación entre el crecimiento económico y las variaciones en la tasa de desempleo. Según este modelo, que se conoce hoy en día como la "ley de Okun," existe una relación lineal entre el *cambio* en la tasa de desempleo y la tasa de crecimiento del Producto Interno Bruto (PIB) real. En el cuadro adjunto se muestran datos anuales para la tasa de desempleo y el cambio porcentual en el PIB real en Alemania Occidental durante el período 1960-1981. Usar estos datos para estimar el modelo de Okun, y explicar el significado de los resultados obtenidos.

Año	Crecimiento PIB Real (%)	Desempleo (%)
1960	4.6	1.2
1961	5.1	0.9
1962	4.4	0.7
1963	3.1	0.9
1964	6.7	0.8
1965	5.5	0.7
1966	2.6	0.7
1967	-0.1	2.1
1968	5.9	1.5
1969	7.5	0.8
1970	5.1	0.7
1971	3.1	0.8
1972	4.2	1.1
1973	4.6	1.2
1974	0.5	2.6
1975	-1.7	4.8
1976	5.5	4.7
1977	3.1	4.6
1978	3.1	4.4
1979	4.2	3.8
1980	1.8	3.8

1981	0.1	5.5

Fuente: Frank Wolter, "From Economic Miracle to Stagnation: On the German Disease," en A. C. Harberger, ed., *World Economic Growth* (San Francisco: ICS Press, 1984), Table A-3, p. 119.

Medida del grado de asociación entre las variables

La medida del grado de relación entre dos variables se denomina **Coeficiente de Correlación ρ**.

Consideraciones a tener en cuenta en este análisis:

1 - Las variables X e Y son variables aleatorias, esto significa que no es fijo decir variable dependiente o independiente, cualquiera de las dos puede ser la variable independiente o a la inversa.

2 - Las variables proceden de una población Normal bivariada, o sea X e Y están distribuidas conjuntamente como Normal.

3 - X e Y tienen cada una distribución Normal:

$$X \approx N(\mu_x;\sigma_x) \qquad Y \approx N(\mu_y;\sigma_y)$$

4 - La relación entre X e Y es lineal; este supuesto implica decir que las medias de y para valores de X caen sobre la recta $Y_i = \alpha + \beta X_i$ de la misma manera que para $X_i = \alpha + \beta Y_i$.

5 - Si las dos rectas de regresión (con X dependiente o con Y dependiente) son iguales, quiere decir que la relación es perfecta.

El coeficiente de correlación poblacional se define como:

$$\rho = \frac{E(X-\mu_x)E(Y-\mu_y)}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}} = \frac{\text{cov}arianza}{\sigma_x \sigma_y} \qquad (12)$$

Siendo μ_x , μ_y ,σ_x , σ_y , ρ, los parámetros de la distribución Normal bidimensional.

De la expresión de ρ puede decirse:

- se aprecia que un cambio en el orden de las variables no afecta su valor, luego es un número adimensional.

- la covarianza y de aquí ρ, serán positivos si grandes valores medios de X se asocian con grandes valores medios de Y (y pequeños valores medios con pequeños valores medios). Por el contrario si grandes valores medios se asocian con pequeños valores medios (o viceversa), la covarianza y por lo tanto ρ, serán negativos. En ambos casos puede decirse que existe al menos alguna vinculación o dependencia estocástica entre X e Y.

Específicamente el **coeficiente de correlación** es una medida de la dependencia lineal entre dos variables aleatorias. Dado solamente el valor de ρ, puede decirse que un alto valor implica dependencia estocástica alta y de esta manera se puede decir que existe entre X e Y una tendencia lineal conjunta. Lo cual no significa necesariamente relación de causa y efecto, mientras que un bajo valor implica que las variables no tienen un comportamiento lineal conjunto y esto no asegura que falte dependencia estocástica. Es por esto, que debe tenerse **CUIDADO EN SU INTERPRETACIÓN**.

Correlación espuria: suele aparecer cuando se busca normalizar los datos, dividiendo por algún factor, el cual es en sí mismo una variable aleatoria, las variables originales pueden ser independientes pero los pares formados por los cocientes pueden presentar alta correlación, cuando en realidad no existe.

Valores posibles de ρ

$$Cov(x,y) = E(x - \mu_x)E(y - \mu_y)$$

$$x \text{ e } y \text{ son } N(\mu, \sigma)$$

$Cov(x^*,y^*) = E(x^*)E(y^*)$, *las esperanzas son* 0 *y los desvìos* 1. *Luego:*

$$Cov(x,y) = E(xy) - E(x)E(y) = E(x^*y^*)$$

$$\rho = \frac{Cov(xy)}{\sigma_x \sigma_y}. \text{ Siendo } \sigma_x \text{ y } \sigma_y \text{ iguales a } 1, \text{ luego:}$$

$$Var(x^* \pm y^*) = Var(x^*) + Var(y^*) \pm 2Cov \geq 0$$

$$1 + 1 \pm 2\rho \geq 0$$

$$2 \pm 2\rho \geq 0$$

$$-1 \leq \rho \leq 1$$

Coeficiente de correlación muestral

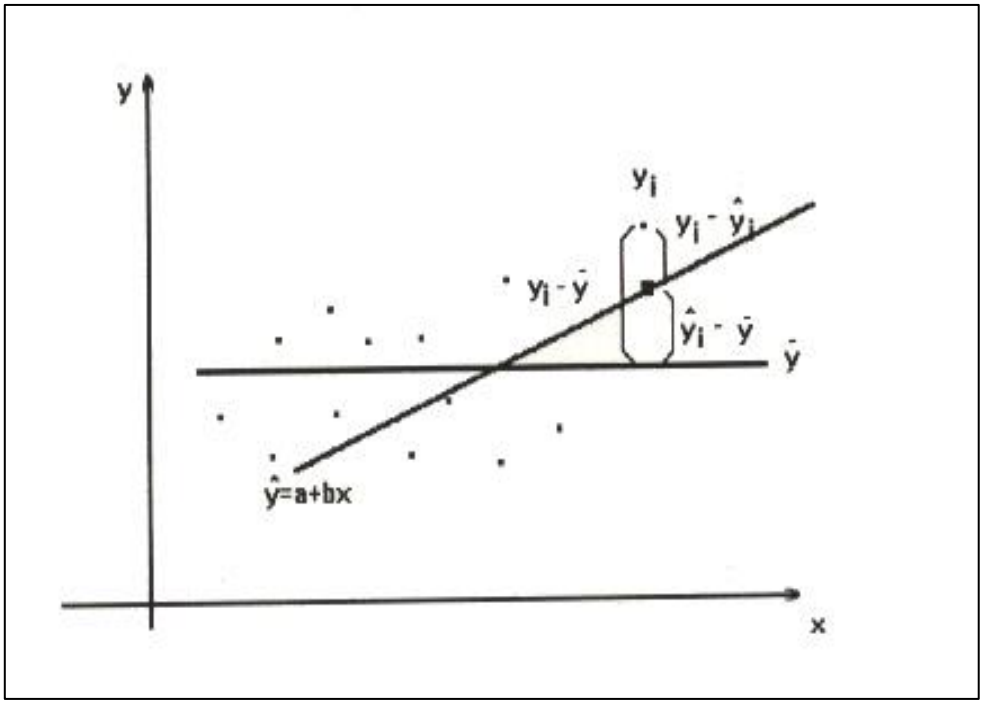
El estimador de ρ se obtiene considerando los momentos muestrales:

$$r = \hat{\rho} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{m_{1,1}}{S_x S_y} \qquad (13)$$

Su variación es la misma que la del coeficiente de correlación poblacional.

Otro coeficiente usado en este análisis es el de determinación, que está relacionado con el de correlación.

Recordando el análisis de varianza ya realizado, se partirá de estas expresiones para obtener el **coeficiente de determinación**.



$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\sum_i (y_i - \bar{y})^2 = \ (SCT)$$

$$\sum_i (\hat{y}_i - \bar{y})^2 = \ (SCR)$$

$$\sum_i (y_i - \hat{y}_i)^2 = \ (SCE)$$

$$SCR = \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (a + bx_i - \bar{y})^2 = \sum_i (\bar{y} - b\bar{x} + bx_i - \bar{y})^2 =$$

$$= \sum_i \left[b(x_i - \bar{x})\right]^2 = b^2 \sum_i (x_i - \bar{x})^2$$

$$SCT = \sum_i (y_i - \bar{y})^2$$

$$SCE = SCT - SCR$$

$$Dividiendo por\ SCT:$$

$$\frac{SCE}{SCT} = \frac{SCT}{SCT} - \frac{SCR}{SCT}$$

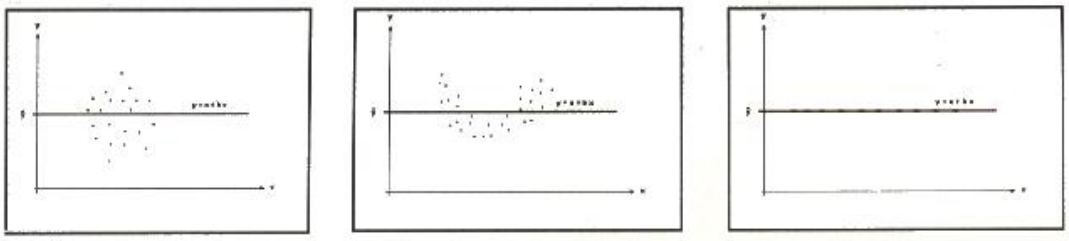
$$1 = \frac{SCR}{SCT} + \frac{SCE}{SCT}$$

$$r^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT} \qquad (14)$$

$$r^2 = \frac{suma\ de\ cuadrados\ \text{explicados}}{suma\ de\ cuadrados\ total}$$

r^2 varía entre 0 y 1, ya que SCR es menor o igual que SCT.
Pueden hacerse algunos comentarios respecto de su valor:
Si SCE = 0, esto implica que SCR = SCT, lo que lleva a decir que r^2 es igual a 1. Esto significa que todos los puntos están sobre la recta estimada.
Si SCR = 0, implica que SCE = SCT, con lo cual $r^2 = 0$. Esto significa que la pendiente de la recta es igual a cero. Esto puede deberse a que la línea de regresión sea horizontal, y esto ser debido a distintas causas:

a) las observaciones se dispersan alrededor del valor medio en forma aleatoria.
b) las observaciones se dispersan alrededor de una curva tal que la línea mejor ajustada es una línea recta horizontal.
c) todas las observaciones tienen el mismo valor, cualquiera sea el valor de x.



Este coeficiente es también denominado **índice de correlación**, y se utiliza para medir el grado de asociación entre las variables cuando la regresión es lineal y no lineal.

Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 8
SERIES CRONOLÓGICAS

SERIES CRONOLÓGICAS

Son aquellas que están formadas por valores de una variable observada a <u>intervalos regulares de tiempo</u>.

Pueden ser

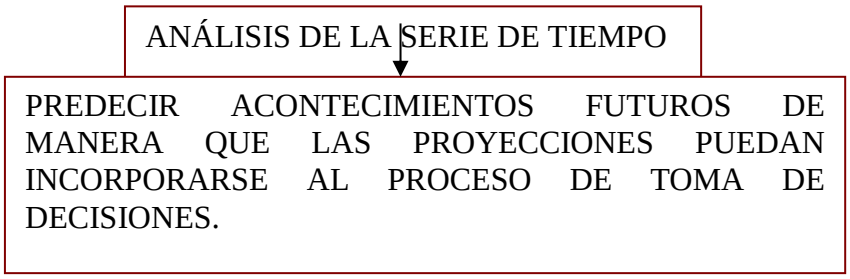
- ANUALES
- MENSUALES
- BIMESTRALES
- OTROS

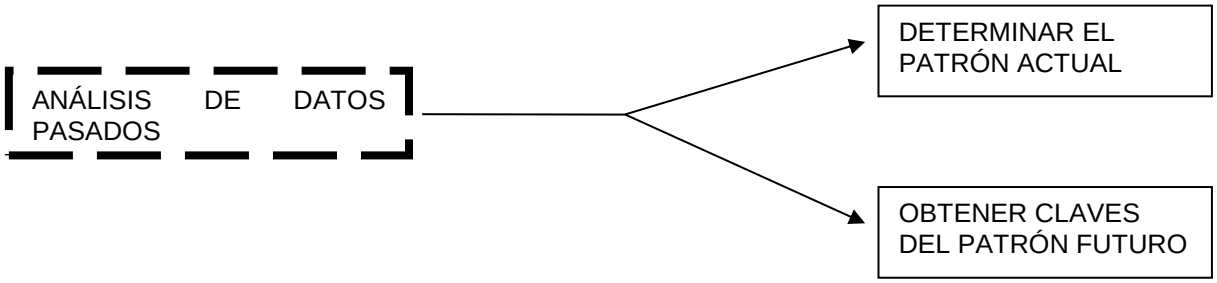
EJEMPLOS:

- Ventas semanales de un supermercado.
- Producción de una fábrica textil durante un mes.
- Informes anuales de un Municipio en cuanto a la recaudación de impuestos.
- Informes mensuales de un determinado Banco en cuanto al ingreso de depósitos.

EL ANÁLISIS DE SERIES DE TIEMPO

PERMITE → CONTROLAR LAS OPERACIONES PRESENTES Y LLEVAR ADELANTE EL PLANEAMIENTO FUTURO A TRAVÉS DE LA APLICACIÓN DE TÉCNICAS ESTADÍSTICAS





COMPONENTES DE UNA SERIE DE TIEMPO

El análisis de una serie de tiempo es un tema complicado; hay una variedad de opiniones en cuanto a como se tendrían que realizar los análisis.

Uno de los enfoques de mayor aceptación es considerar una serie de tiempo como una combinación de 4 elementos, los cuales superpuestos y actuando en forma conjunta contribuyen a los cambios que se observan en un período de tiempo. Estos elementos son:

a)- Tendencia a largo plazo

b)- Variación estacional

c)- Variación cíclica

d)- Variación aleatoria e irregular, impredecible

Estos componentes se aíslan y se ajustan utilizando algunos métodos y son:

a)- Tendencia a largo plazo: Es el movimiento de una serie de tiempo (creciente o decreciente) gradual en el tiempo de acuerdo a una curva .

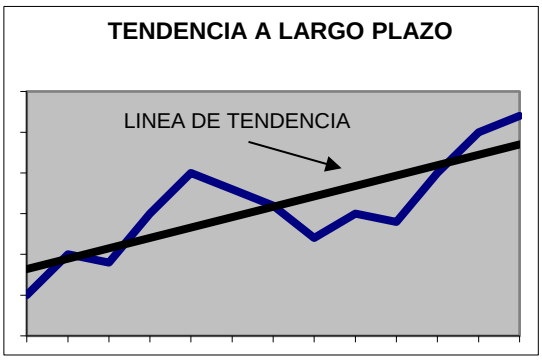
b)- Variación estacional: representa la tendencia de la serie de tiempo a variar hacia arriba y hacia abajo durante épocas especificas del año y más o menos con igual intensidad, pueden

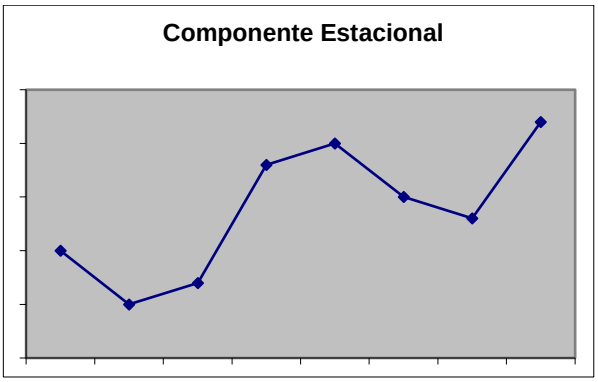
ser meses, bimestres o trimestres, además puede presentarse con movimientos periódicos por naturaleza. Es decir que esta componente se presenta cuando se trabaja con datos mensuales.

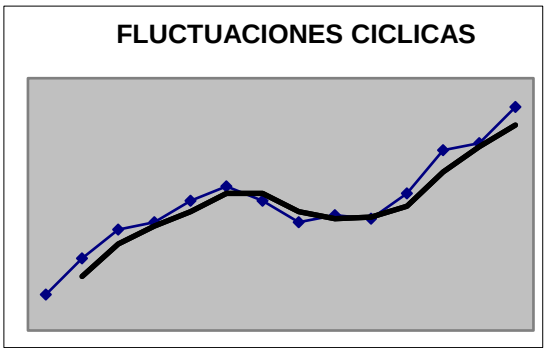
c)- Variación Cíclica: Los componentes cíclicos de una serie de tiempo tienden a subir o bajar según un patrón cíclico alrededor de la curva de la tendencia. Difieren de la componente anterior en que se extienden por períodos de tiempo más largos derivándose de causas diferentes.

d)- Variación aleatoria e irregular: Esta variación se añade a las anteriores. Se presentan de manera casual debido a efectos inexplicados como por ejemplo:

- Guerras
- Inundaciones
- Huelgas
- Etc.







Características de las Componentes

Componente	***Definición***	Razón	***Duración***
Tendencia	Patrón de movimientos ascendentes o descendentes, persistente a largo plazo.	Debido a cambios en la tecnología, riqueza, población.	Varios años.
Estacional	Fluctuaciones periódicas regulares que ocurren dentro de cada período de 12 meses, año a año.	Debido a condiciones de costumbres, tiempo, etc.	Dentro de los 12 meses.
Cíclica	Movimientos repetitivos ascendentes y descendentes mediante cuatro fases: -punto más alto: Prosperidad -contracción: Recesión -sima: Depresión -expansión: Recuperación.	Interactúan una serie de combinaciones de factores que influyen en la economía.	Generalmente de 2 a 10 años con diferente intensidad para cada ciclo completo.
Irregular	Fluctuaciones que existen en una serie luego de tomar en cuenta los efectos sistemáticos anteriores.	Se relacionan con acontecimientos imprevistos como huelgas, inundacio-nes, etc,.	Breves y no repetitivas.

Se supone además que estas cuatro componentes están ligadas por una relación, vamos a nombrar dos relaciones, si bien, hay que aclarar que no son los únicos esquemas de análisis.

Modelo o hipótesis *Aditiva*: supone que los cuatro componentes son independientes unos de otros:

$$Yi = Ti + Si + Ci + Ii$$

Modelo o hipótesis *multiplicativa:* supone que los cuatro componentes se deben a diferentes causas, y que se relacionan entre sí por un efecto multiplicador:

$$Yi = Ti * Si * Ci * Ii$$

El modelo clásico multiplicativo que se analizará considera que cualquier valor observado en una serie de tiempo es el producto de los factores componentes:

$$Y_i = T_i * S_i * C_i * I_i$$

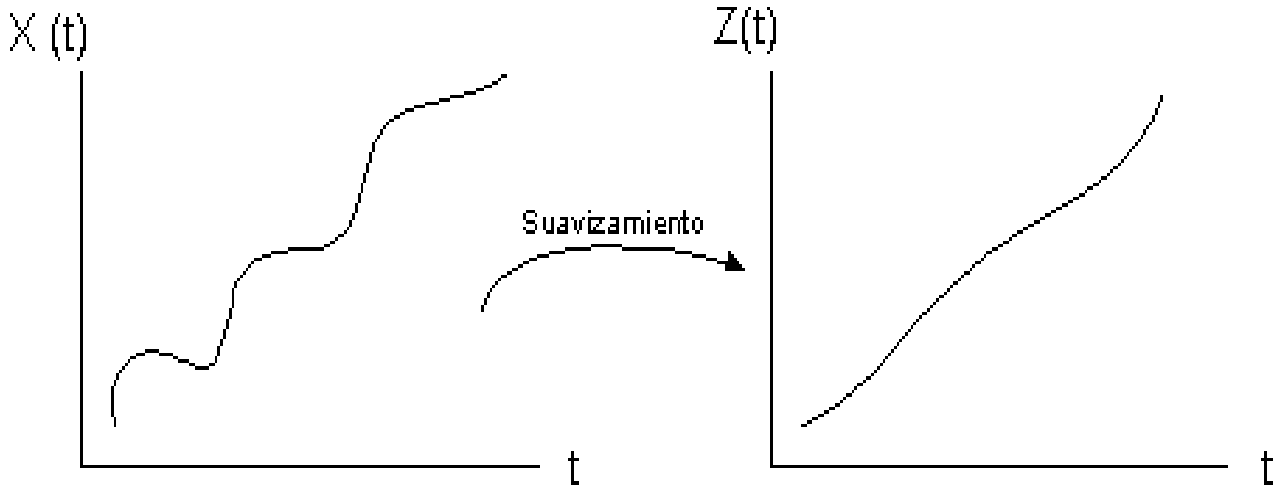
Donde: i es el año

T_i = Valor de la componente de la tendencia
S_i = Valor del componente estacional
C_i = Valor del componente cíclico
I_i = Valor del componente irregular o aleatorio

SUAVIZACIÓN DE SERIES DE TIEMPO

Muchas veces resulta difícil al examinar la información, decir si la **tendencia** es descendente o ascendente debido a que existen amplias fluctuaciones en sus componentes cíclicos e irregular. Entonces antes de tratar de modelar una serie de tiempo es útil graficarla para determinar la naturaleza de los componentes secular, cíclica y estacional si es que existen. Pueden utilizarse métodos para **suavizar o alisar la serie** y poder así distinguir los distintos movimientos libre de los efectos de la variación aleatoria.

La idea central es definir a partir de la serie observada una nueva serie en la que suavizan los efectos ajenos a la tendencia (estacionalidad, efectos aleatorios), de manera que se pueda determinar claramente la tendencia.



Lo que se hace es usar una expresión lineal que transforma la serie *X(t)* en una serie *suavizada Z(t): Z(t) = F(X(t)), t* = 1,...,n

de tal modo que F(*X(t)) = T(t).* La función F se denomina Filtro Lineal. El filtro lineal más usado es el ***promedio móvil***.

PROMEDIOS MÓVILES

Dado un conjunto de números

Y1, Y2, Y3 ...

se define un movimiento medio de orden N el que viene dado por la sucesión de medias aritméticas.

$$\frac{Y_1+Y_2+...+Y_N}{N}, \quad \frac{Y_2+Y_3+...+Y_N+_1}{N}, \quad \frac{Y_3+Y_4+...+Y_N+_1}{N}, \ ...$$

Las sumas de los numeradores se llaman movimientos totales de orden N.

Si los datos son anuales o mensuales, se llama movimiento medio de N años o un movimiento medio de N meses, respectivamente. Así, se habla de movimientos medios de cinco años, movimientos medios de doce meses, etc. Naturalmente que cualquier otra unidad de tiempo puede igualmente utilizarse.

Los movimientos medios tienen la propiedad de tender a reducir la cantidad de variación presente en un conjunto de datos. En el caso de series de tiempo, esta propiedad se utiliza a menudo para eliminar las fluctuaciones no deseadas y el proceso se llama **suavización de series de tiempo.**

Para aplicar el método del promedio móvil a una serie de tiempo, los datos deben tener una tendencia bastante lineal y un esquema de fluctuaciones rítmico definido (que se repite, por ejemplo cada tres años). Cuando no hay componente estacional o sea para datos anuales lo que se hace en el método de promedios móviles es en realidad promediar C e I. El residuo es la tendencia.

Si la duración de los ciclos es constante y si las amplitudes de tales ciclos son iguales, las fluctuaciones cíclica e irregular pueden eliminarse por completo usando el método del promedio móvil.

El problema principal en los promedios móviles es la elección apropiada de período para el promedio, esto depende de la naturaleza de los datos y el propósito que se persigue. Generalmente el **objetivo** de aplicar un promedio móvil es eliminar, tanto como sea posible, las variaciones indeseables de los datos, tratando de darle a la serie un aspecto más uniforme.

Si por ejemplo, a una serie temporal de 7 observaciones se le aplica este método, tomando medias aritméticas de tres observaciones (promedio móvil de orden tres), las operaciones a seguir son:

Ti	**Yi**	$\overline{Y}$
T1	Y1	
T2	Y2	$\overline{Y}2$
T3	Y3	$\overline{Y}3$
T4	Y4	$\overline{Y}4$
T5	Y5	$\overline{Y}5$
T6	Y6	$\overline{Y}6$
T7	Y7	

$$\text{donde } \overline{Y}2 = \frac{Y1 + Y2 + Y3}{3}$$

$$\text{donde } \overline{Y}3 = \frac{Y2 + Y3 + Y4}{3}$$

$$\text{donde } \overline{Y}4 = \frac{Y3 + Y4 + Y5}{3}$$

$$\text{donde } \overline{Y}5 = \frac{Y4 + Y5 + Y6}{3}$$

$$\text{donde } \overline{Y}6 = \frac{Y5 + Y6 + Y7}{3}$$

Como se observa cada media aritmética se obtiene de la anterior con solo eliminar el primer valor Yi y añadir el siguiente; de ahí su denominación de medias móviles. Si se tomase un número par de valores para obtener las medias móviles, la nueva serie de promedios quedaría descentrada en el sentido de que sus valores no corresponderían ya a los momentos originales de tiempo, sino a momentos intermedios; luego deberá calcularse una nueva serie de medias, promediando los valores medios obtenidos; o sea:

Ti	**Yi**	$\overline{Y}$	$\overline{\overline{Y}}$
T1	Y1	$\overline{Y}2$ →	$\overline{\overline{Y}}2$
T2	Y2	$\overline{Y}3$ →	$\overline{\overline{Y}}3$
T3	Y3	$\overline{Y}4$ →	$\overline{\overline{Y}}4$
T4	Y4	$\overline{Y}5$ →	$\overline{\overline{Y}}5$
T5	Y5	$\overline{Y}6$ →	$\overline{\overline{Y}}6$
T6	Y6	$\overline{Y}7$	
T7	Y7		

$$\text{donde } \overline{Y}2 = \frac{Y1 + Y2}{2}$$

$$\text{donde } \overline{Y}3 = \frac{Y2 + Y3}{2}$$

$$\text{donde } \overline{Y}4 = \frac{Y3 + Y4}{2}$$

$$\text{donde } \overline{Y}5 = \frac{Y4 + Y5}{2}$$

$$\text{donde } \overline{Y}6 = \frac{Y5 + Y6}{2}$$

$$\text{donde } \overline{Y}7 = \frac{Y6 + Y7}{2}$$

$$\text{donde } \overline{\overline{Y}}2 = \frac{\overline{Y}2 + \overline{Y}3}{2}$$

$$\text{donde } \overline{\overline{Y}}3 = \frac{\overline{Y}3 + \overline{Y}4}{2}$$

$$\text{donde } \overline{\overline{Y}}6 = \frac{\overline{Y}6 + \overline{Y}7}{2}$$

Los inconvenientes de este método son los siguientes:

1. Se pierden los períodos al comienzo y a la finalización de la serie.
2. Se pueden generar componentes que los datos originales no tenían
3. Estos promedios móviles están fuertemente afectados por los valores extremos

Sin embargo permite suavizar la serie original y es muy utilizado como base para el futuro análisis de la componente estacional.

PARA RESUMIR: la técnica de promedios móviles auxilia en la identificación de la tendencia a largo plazo en una serie de tiempo ya que amortigua las fluctuaciones a corto plazo. Sirve para revelar cualquiera de las fluctuaciones cíclicas y estacionales.

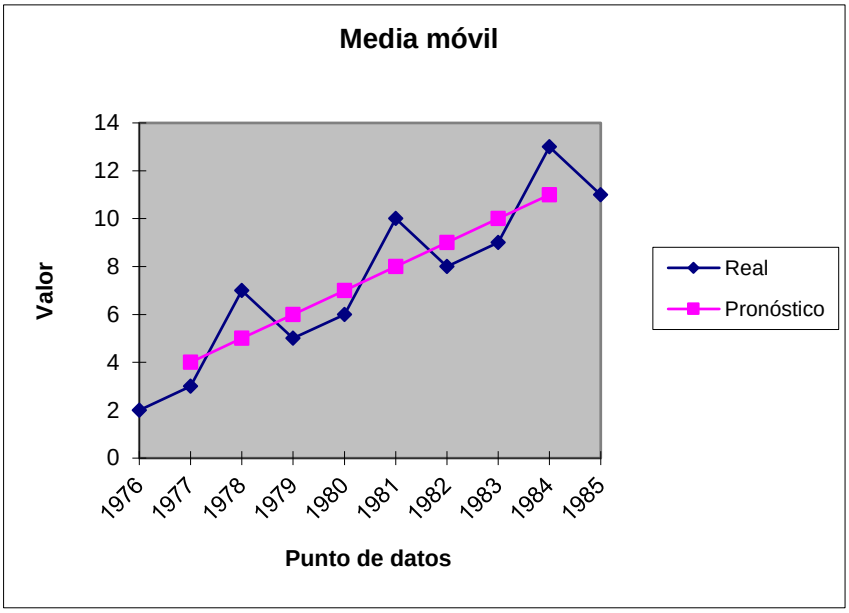
Ejemplo:

Sea la serie ventas de una determinada empresa comercial:

Año	**Ventas en miles**	**Prom. Móvil de Orden 3**	**Prom. Móvil de Orden 2 (1ª Media)**	**Prom. Móvil de Orden 2 (2ª Media)**
1976	**2**		**2,5**	
1977	**3**	**4**	**5**	**3,75**
1978	**7**	**5**	**6**	**5,50**
1979	**5**	**6**	**5,5**	**5,75**
1980	**6**	**7**	**8**	**6,75**
1981	**10**	**8**	**9**	**8,50**
1982	**8**	**9**	**8,50**	**8,75**

1983	**9**	**10**	**11**	**9,75**
1984	**13**	**11**	**12**	**11,50**
1985	**11**			

Promedio Móvil de orden 3



Como se puede ver en el gráfico anterior, los valores faltantes al comienzo y al final de la serie artificial son una característica de este tipo de suavizado: se pierde un valor en cada extremo de un promedio móvil de tres años, dos en un promedio de cinco, tres en uno de siete, etc. Esto puede o no tener consecuencias, pero puede ocasionar problemas cuando la serie es muy corta y se necesite realizar un pronóstico a largo plazo.

SUAVIZACIÓN EXPONENCIAL

Es otro método para suavizar los movimientos generales a largo plazo en la información. Puede ser utilizado también para obtener pronósticos a corto plazo para series de tiempo en las que resulta dudoso el tipo de efecto de tendencia a largo plazo.

La suavización exponencial es un tipo especial de promedio móvil, pero su naturaleza es muy diferente del cálculo del Promedio Móvil.

La *suavización exponencial*, una clase especial de *promedio móvil ponderado*, es útil en pronóstico a corto plazo proporciona una impresión de los movimientos globales a largo plazo de los datos.

Esta técnica posee una ventaja con respecto al promedio móvil porque proporciona un promedio móvil ***exponencialmente ponderado*** a través de la serie de tiempo, ya que cada valor suavizado depende de todos los valores anteriormente observados (en los promedios móviles no se toman en cuenta todos los valores observados).

Si se desea pronosticar el valor de una serie de tiempo para el período *t + 1* sobre la base de la información obtenida inmediatamente después del período t, el pronóstico se considera mejor como una función de dos componentes: el valor real de una serie para el período t, y el valor pronosticado para el mismo período hecho en el período anterior *t – 1*. El uso de valores observado y estimado disponibles ahora para predecir valores futuros es mejor que el uso de cualquiera de ellos sólo, porque el valor real en el periodo t podría haber sido indebidamente influido por factores aleatorios o porque puede que las condiciones que condujeron al pronóstico para el período t no se cumplan ya, o por todos estos factores juntos.

Un valor suavizado exponencialmente para una serie de tiempo está dado por:

$$S_i = W * Y_i + (1 - W) * S_i - 1$$

Donde: **S1 = Y1**

Siendo:

S_i: valor de la serie exponencialmente suavizada calculada en el período **i**
$S_i - 1$: valor de la serie exponencialmente suavizada ya calculada en el período **i – 1**
Y_i: valor observado de la serie de tiempo en el período **i**
W: ponderación o coeficiente de suavización asignado en forma subjetiva ($0 < W < 1$)

Como elegir W:

Si el valor de *W* es demasiado grande, se le dará un valor muy grande a los datos actuales y no se suavizan adecuadamente las variaciones irregulares; pero si *W* es demasiado chico, se le dará un valor muy pequeño a los datos actuales de la serie y el promedio móvil será insensible a las variaciones que se pueden dar en realidad. Esto podría considerarse como una desventaja en la utilización de este método, ya que el rango de valores que se le pueden asignar a W es muy grande, se debe tener mucho cuidado dependiendo de lo que se quiera reflejar con los resultados.

Si sólo se quiere suavizar una serie mediante la eliminación de variaciones cíclicas e irregulares que no se desean, debe seleccionarse un valor pequeño de W (cercano a 0); pero si se quieren hacer pronósticos se elegirá un valor grande de W (cercano a 1). En el primer caso las tendencias generales de la serie a largo plazo serán aparentes, y en el segundo caso tal vez se pronostiquen en forma adecuada las direcciones a corto plazo.

Importante:

Si se quieren utilizar los valores suavizados exponencialmente para realizar pronósticos, sólo se toma el valor suavizado en el período **i** como el valor pronosticado en el período **i + 1** por lo tanto:

$$Y_i + 1 = S_i$$

ANÁLISIS DE TENDENCIA

Tendencia

Es la componente de una serie que más se estudia, se puede hacer con fines de pronóstico a largo y mediano plazo.

Tendencia Lineal:

El método de mínimos cuadrados permite ajustar una línea recta de la forma:

$$Y = a + bx$$

a: Ordenada al origen, es el valor de Y cuando X=0
b: Pendiente de la Tendencia, indica el aumento o disminución medio en Y por cambio unitario en X.

Recordamos el análisis de regresión lineal que permite calcular los estimadores de los parámetros de la regresión.

Cuando utilizamos el método de mínimos cuadrados para ajustar tendencias en series de tiempo, nuestros esfuerzos de cálculo pueden minimizarse si se codifican apropiadamente los valores de X (con software estos pasos no se realizan)

Se debe distinguir además números de años par e impar:

- Si la serie tiene numero impar de años se elige como inicio de la sucesión al año que esta en el centro, asignándole valores crecientes a los posteriores 1,2,3,..... y enteros consecutivos decrecientes a los anteriores -1,-2,-3,.....
- Si la serie tiene un número par se toma como origen el primer año X = 0 y se le asigna números consecutivos crecientes a los años siguientes.

Una vez calculada la ecuación de la tendencia puede ser utilizada para calcular la tendencia en un año dado sustituyendo el valor X que corresponde a ese año en la ecuación obtenida. Esta ecuación puede utilizarse para pronóstico.

Tendencias no Lineales:

Si el modelo a ajustar a la tendencia no es una recta sino una curva, puede pensarse en un "*polinomio de segundo grado*"

$$\hat{y}_i = B_0 + B_1 X_i + B_2 X_i^2$$

B_0 = intersección con el eje Y
B_1 = efecto lineal estimado
B_2 = efecto curvilíneo estimado

Los coeficientes se estimarán a través del método de mínimos cuadrados resolviendo el sistema de ecuaciones correspondiente.

Los cálculos de los coeficientes B_0, B_1, B_2 pueden simplificarse usando la condición de codificación mencionada anteriormente.

Tendencia Potencial:

Cuando una serie parece estar incrementando con rapidez cada vez mayor tal que la diferencia porcentual de una observación a otra es constante se puede considerar una ecuación de tendencia potencial de la forma

$$Y_i = B_0 * B_1^{Xi}$$

En la que B_0 = intersección estimada con el eje Y

$(B_1 - 1)*100\%$ = Tasa de crecimiento compuesta estimada anual (en porcentaje)

Si se toma logaritmo (Base 10) de ambos lados de la ecuación se tiene:

$$\log \hat{y}i = \log B_0 + X_i \log B_1$$

Puesto que la ecuación tiene forma lineal, se puede utilizar el método de mínimos cuadrados, si se trabaja con el logaritmo de los valores de Y_i, en lugar de hacerlo con los valores de Y_i, y obtener la pendiente ($\log B_1$) y la intersección ($\log B_2$)

Aislamiento y eliminación de la tendencia en datos anuales y mensuales

Si los pronósticos que se desean elaborar son a *corto plazo*, debería ser eliminado el efecto del componente tendencia ya que esta es la componente a largo plazo.

Se estima la tendencia y para eliminarla es suficiente con dividir por ella en la expresión del modelo multiplicativo *sin considerar el efecto estacional*, se lo logra ya que los datos son anuales:

$$Y_i = T_i * C_i * I_i$$

$$\frac{Y_i}{\hat{y}_i} \text{ como } \hat{y}_i = T_i \quad \text{luego} \quad \frac{Y_i}{\hat{y}_i} = \frac{T_i * C_i * I_i}{T_i}$$

$$\frac{Y_i}{\hat{y}_i} = C_i * I_i$$

Este cociente se denomina Relativas cíclicas irregulares ya que muestran el desarrollo tanto cíclico como irregular de la serie de datos una vez eliminada la tendencia en los datos. Gráficamente estos valores fluctuarán alrededor -1 y 1.

Aislamiento y eliminación de la tendencia en datos mensuales

En la serie de datos mensuales existe la componente estacional, además de la tendencia, cíclica e irregular.

Al igual que para datos anuales, interesa pronosticar algunos movimientos futuros de los movimientos mensuales, sólo que, se agregarán los datos mensuales para llevarlos a anuales y luego de obtenida la ecuación de tendencia deberá tenerse en cuenta lo que significa cada parámetro de la regresión para su transformación.

a es la ordenada al origen, por lo tanto dará el valor de la variable serie de tiempo expresada en utilidades por año con lo cual será necesario dividir su valor por doce (12), para expresarlo en unidades por meses.

b es la pendiente que da el cambio medio de Y por cambio en X, entonces el valor b se dividirá por 144 (12^2) para expresarlo en unidades por mes.

VARIACIÓN CÍCLICA:

A la mayoría de las empresas les interesan los pronósticos a corto plazo, por ejemplo el próximo mes, bimestre o año, estas predicciones se usan para planificar y controlar los negocios día a día en un futuro próximo.

En las predicciones a corto plazo, el objetivo principal al analizar la componente cíclica es identificar la posición actual y poder predecir el comportamiento futuro.

Es muy difícil encontrar un modelo regular, en promedio, que permita la proyección mecánica hacia el futuro en este tipo de casos, esto se debe a que existe mucha variabilidad de un ciclo a otro en la mayor parte de las actividades económicas; por esto, es que no se puede obtener predicciones confiables de los movimientos cíclicos en la actividad industrial o de empresas para poder proyectarlos mecánicamente hacia el futuro.

Además la identificación del estado actual de los movimientos cíclicos se ve obstaculizada por la presencia de movimientos irregulares.

Por lo tanto, con datos anuales, con la descomposición se llega hasta la obtención de las relativas cíclicas-irregulares.

VARIACIÓN ESTACIONAL

Dada una serie cronológica con valores mensuales, nos proponemos aislar su componente estacional; es decir, medir la influencia que depende del calendario y, por lo tanto, de las estaciones del año. Las variaciones estacionales generalmente provienen, en economía, de causas climáticas que determinan los ciclos vegetativos, lo que a su vez, influyen en la producción, el consumo, la ocupación y otros factores económicos, de manera que esta variación se repite año tras año de manera parecida.

Existen también modelos estacionales que se repiten en periodos inferiores a un año como por ejemplo los modelos diarios de rendimiento de productividad por hora en una planta.

Un conjunto de números mostrando los valores relativos de una variable durante los meses del año se llama **índice estacional** de la variable. Aunque dichos índices suelen determinarse en forma mensual, se pueden elaborar para otras subdivisiones de un año, por ejemplo bimestre, trimestrales, semanales, etc.

Como se dijo anteriormente a dicha variación estacional se la denomina índice estacional por ser mensuales los datos, este índice consta de 12 valores, uno por cada mes y cada uno expresa la actividad de ese mes en particular como porcentaje de la actividad del mes promedio.

Cada índice es un porcentaje con el promedio del año igual a 100 o sea cada índice mensual indica el nivel de ventas de producción, o de otra variable, en relación con el promedio anual de 100. Por ejemplo decir que el índice para octubre es de 107 significa que la variable en forma característica está 7% por arriba del promedio anual.

El método más usado el llamado método de razón a promedio móvil que elimina las componentes de tendencia, cíclica e irregular de los datos originales. Los números que resultan se denominan *índice estacional.*

Por ejemplo:

Se sabe que la venta durante enero, febrero, marzo, etc., son de 50, 120, 90,... por ciento de la venta media mensual del año completo, los números 50, 120, 90,... suministran el índice estacional del año y a veces se conocen el número del índice estacional. El promedio (media) del índice estacional deberá ser 100% es decir, la suma de los números índice deberá ser 12.

Al elaborar un índice estacional, todos los esfuerzos se encaminan a la eliminación de las variaciones de tendencia, cíclicas e irregulares de la serie para que se lo quede el estacional. La manera de lógralo, en el método básico de la razón del promedio móvil es relativamente simple.

INDICE ESTACIONAL

Obtención:

1) Se comienza obteniendo una serie de promedio móviles de 12 meses para eliminar los movimientos estacionales de la serie. Como un promedio móvil de n periodos elimina por completo cualquier movimiento recurrente pero absolutamente uniforme en los n

períodos, el promedio móvil de 12 meses suprimirá todos los movimientos estacionales de la serie.

Dichos patrones estacionales varían año tras año con absoluta regularidad; de manera que no se puede eliminar por completo las variaciones estacionales; sin embargo eliminará la mayor parte de esta variación, por lo tanto el promedio móvil de 12 meses será una estimación de la componente de la tendencia y de la variación cíclica.

2) El resultado se centra entre los dos meses centrales que forman cada total móvil, por ejemplo el primer total móvil que consta de los meses de enero a diciembre del primer año se coloca entre junio y julio de ese año, el segundo total móvil que consta de los meses de febrero del primer año a enero del segundo año se coloca entre julio y agosto del primer año y sucesivamente.

3) Para centrar este resultado dentro de un mes en particular se obtiene totales móviles de dos meses de los totales de 12 meses. El primer resultado que consiste en el total registrado entre junio y julio más el de julio y agosto se centra en julio del primer año. Al dividir esto totales por 24 se obtiene promedio móviles centrados. Se dice que estos promedios móviles centrados constan de las componentes cíclicas y tendencia de la serie.

4) Los datos originales se dividen por este promedio y esto hace que únicamente se tengan los factores de la variación estacional e irregular ya que en el modelo que se está analizando se obtendría:

Yi / (promedio móvil centrado)= SI =TSCI / TC

Para elaborar el índice estacional los datos de estas razones se reordenan de acuerdo a los valores mensuales para cada año. Se obtienen así para cada mes un valor de índice que se repetirán año año.

Entonces ahora resta eliminar hasta donde sean posibles las variaciones irregulares para obtener solo la parte estacional. Por ejemplo una manera de reducir estas fluctuaciones es a través del uso de la mediana de los valores dados de cada mes. Estos valores de la mediana se ajustan de manera tal que el valor total de los índices estacionales durante el año sea 12 y el promedio de cada índice estacional (mensual) sea 1.Este factor de corrección será 12/total de las doce medias.

RESUMIENDO: los datos originales contienen las cuatro componentes T C S I. El objetivo es eliminar S de los datos originales. Al obtener los promedios móviles se han eliminado las fluctuaciones estacionales e irregulares solo queda T y C. A continuación al dividir los datos originales por los promedios móviles se obtiene los valores de estacionalidad específicos SI que se expresan en forma de índice multiplicándolos por 100. Por último se toma la media o la mediana de todos índices mensuales ordenados para eliminar la mayor parte de las fluctuaciones irregulares y los valores resultantes indican el patrón de la variación estacional.

DESESTACIONALIZACIÓN DE DATOS

Un conjunto de índices estacionales es muy útil para ajustar las series respecto a fluctuaciones estacionales. La serie resultante se denomina serie desestacionalizada. La razón para desestacionalizar las series es eliminar las fluctuaciones estacionales a fin de estudiar la tendencia y el ciclo. Se consigue dividiendo a cada dato original por e l índice obtenido para cada período, así los datos contienen las componentes de T C e I.

RESUMEN DE LOS PASOS EN EL ANÁLISIS DE SERIES DE TIEMPO

1 Coleccionar los datos de la serie de tiempo, procurando asegurarse de que estos datos sean dignos de confianza. En la colección de datos se debe siempre tener en cuenta el propósito que se persigue en cada caso con el análisis de la serie de tiempo.

2 Representar la serie de tiempo, anotando cualitativamente la presencia de la tendencia de larga duración, variaciones cíclicas y variaciones estacionales.

3 Construir la curva o recta de tendencia de larga duración y obtener los valores de tendencia apropiados mediante cualquiera de los métodos, de mínimos cuadrados, libre, movimientos medios o semimedias.

4 Si están presentes variaciones estacionales, obtener un índice estacional y ajustar los datos a estas variaciones estacionales, es decir, desestacionalizar los datos.

5 Ajustar los datos desestacionalizados a la tendencia. Los datos resultantes contienen solamente las variaciones cíclicas e irregulares. Un movimiento medio de 3, 5 o 7 meses sirve para eliminar las variaciones irregulares y poner de manifiesto las variaciones cíclicas.

6 Representar las variaciones cíclicas obtenidas anteriormente, anotando cualquier periodicidad que pueda aparecer.

7 Combinando los resultados con cualquier otro tipo de información útil, hacer una predicción (si se desea) y si es posible discutir las fuentes de error y su magnitud.