

Glosario básico de términos estadísticos

Lima, mayo de 2006

CREDITOS

Dirección y Supervisión

Lupe Berrocal de Montestruque
Directora Técnica del Centro de Investigación y Desarrollo

Responsable del documento

Herminia Asurza Olaechea

Apoyo en revisión

Santiago Alejandro Billón

Preparado : Por el Centro de Investigación y Desarrollo
Impreso : Talleres de la Oficina Técnica de Administración (OTA) del
Instituto Nacional de Estadística e Informática
Diagramación : Centro de Edición del INEI
Tiraje : ejemplares
Nº de Orden : ~~485~~-OI-OTA-INEI

Hecho el Depósito Legal en la Biblioteca Nacional del Perú N° : 2006-6441



Presentación

El Instituto Nacional de Estadística e Informática (INEI), a través del Centro de Investigación y Desarrollo, continuando con su política de difusión y fortalecimiento de la cultura estadística, pone a disposición de los usuarios interesados en conocer los conceptos básicos de la ciencia estadística el documento **Glosario básico de términos estadísticos**.

La estadística es la ciencia que se ocupa del estudio de fenómenos de tipo genérico, en el ámbito social y económico, normalmente complejos y enmarcados en un universo variable. Emplea modelos de reducción de la información y de análisis de validación de los resultados en términos de representatividad. La información puede ser numérica o alfabética.

Una de las ramas de la ciencia estadística es la estadística descriptiva, que se encarga desde la recolección, procesamiento, análisis y hasta la presentación de un conjunto de datos, mediante las denominadas medidas de posición, dispersión, forma y concentración, con el fin de describir, apropiadamente, ese conjunto de datos. La otra rama es la estadística inferencial que se refiere al método para lograr generalizaciones acerca de las propiedades del todo.

Usualmente el término estadística se utiliza como sinónimo de dato. Sin embargo una información numérica cualquiera puede no constituir una estadística. Para merecer esta denominación, los datos han de constituir un conjunto coherente, organizado de forma sistemática y siguiendo un criterio de ordenación.



El presente documento comprende los términos más usuales de la estadística. Los conceptos incluidos son de fácil comprensión y permiten conocer las definiciones elementales del argot estadístico, ordenadas alfabéticamente.

El INEI espera contribuir con esta publicación al manejo básico de los términos estadísticos incluidos.

Lima, mayo de 2006

FARID MATUK
Jefe

Instituto Nacional de
Estadística e Informática





Glosario básico de términos estadísticos

Este Glosario le permite acceder fácilmente a una definición sencilla de los principales términos utilizados en estadística ordenados alfabéticamente.

A

AFIJACIÓN DE UNA MUESTRA.- Es un método utilizado para establecer cómo debe distribuirse la muestra. En un muestreo estratificado, se refiere generalmente a la determinación del número de unidades en la muestra de cada estrato. En el muestreo por conglomerados, se refiere a la decisión sobre el número de conglomerados por seleccionar y el tamaño de la muestra en cada conglomerado.

AFIJACIÓN ÓPTIMA DE UNA MUESTRA.- Es la forma de seleccionar una muestra de manera tal que produzca un error estándar mínimo para un tamaño de muestra constante. Se utiliza en muestreo estratificado y en muestreo por conglomerados.

AMPLITUD DE UN INTERVALO.- Conocido también como amplitud de clase, es la diferencia entre los dos extremos de un intervalo.

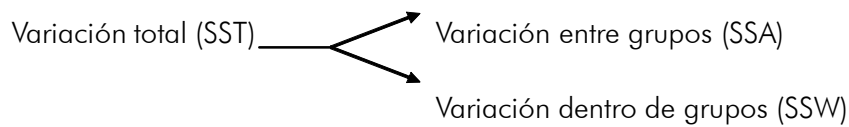
ANÁLISIS DE CONTINGENCIA.- Es el estudio que se realiza con las tablas de contingencia y consiste en analizar el grado de asociación o dependencia entre dos variables cualitativas; para medir el grado de dependencia se utiliza el coeficiente de contingencia. (Ver coeficiente de contingencia).

ANÁLISIS DE CORRELACIÓN.- Es el estudio que se realiza para medir la intensidad o grado de la asociación que existe entre variables numéricas.

ANÁLISIS DE REGRESIÓN.- Es el estudio que se realiza con el propósito de hacer predicciones. El objetivo es el desarrollo de un modelo estadístico que pueda ser utilizado para predecir valores de una variable dependiente, basado en los valores de la variable independiente.

ANÁLISIS DE VARIANZA.- Es un método para comparar dos o más medias (Ver media) de «n» grupos analizando la varianza de los datos, tanto entre «n» grupos como dentro de ellos.

En el análisis de varianza se subdivide la variación total de las mediciones resultantes (SST Sum of squares of the treatments) en lo que puede atribuir a diferencias entre los «n» grupos (SSA Sum of squares between(among)) y lo que se debe al azar o que se puede atribuir a una variación inherente dentro de los «n» grupos (SSW Sum of squares within). La variación dentro de grupos se considera error experimental, mientras que la variación entre grupos se atribuye a efectos de tratamiento.



ASIMETRÍA.- Es la falta de simetría entre los datos de una distribución. El concepto de asimetría se refiere a si la curva que forman los valores de la serie presenta la misma forma a la izquierda y derecha de un valor central (media aritmética).

AUTOCORRELACIÓN.- Se denomina así a la correlación de una variable consigo misma cuando se desfasa uno o más periodos de tiempo. Se determina calculando el coeficiente de autocorrelación. Se usa para tal efecto la siguiente fórmula:

$$r_k = \frac{\sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

Donde:

- r_k Es el coeficiente de autocorrelación para un desfase de k periodos.
- \bar{Y} Es la media de los valores de la serie
- Y_t Es la observación en el periodo de tiempo t
- Y_{t+k} Es la observación en k periodos posteriores o en el periodo t+k.

Por lo cual

- r_1 es el coeficiente de autocorrelación en el primer desfase,
- r_2 es el coeficiente de autocorrelación en el segundo desfase y así sucesivamente hasta un r_k desfase.

B

BASE DEL ÍNDICE.- Es la magnitud utilizada como unidad de referencia, contra la cual se hacen todas las comparaciones de la variable en estudio. Esta base puede corresponder a un año, un trimestre, un mes, etc. Al seleccionar el período base para un índice (Ver índice), debe tomarse en cuenta dos reglas:

1. El período base seleccionado, hasta donde sea posible, debe ser de normalidad o estabilidad económica.
2. El período base debe ser reciente a fin de que las comparaciones no se afecten por cambios en la tecnología, en la calidad del producto o por las actitudes e intereses de los consumidores. El valor del índice para el período base es 100.

BONDAD DE AJUSTE.- Es un indicador que permite discernir acerca de qué tan buena es la ecuación obtenida. Para determinar la bondad de un ajuste se utilizan diferentes criterios en la regresión lineal. Unos se refieren a los residuales como son el valor de la sumatoria de residuales al cuadrado, la varianza, la desviación estándar del ajuste y el coeficiente de correlación al cuadrado. Otro indicador de la bondad de ajuste es el realizado mediante el test de bondad de ajuste utilizando la prueba Ji-Cuadrada (X^2), Kolgomorov -Smirnov (K-S) entre otras.

BOXPLOT.- (Ver diagrama de caja).

C

CARTOGRAMAS.- Es un tipo de gráfico mediante el cual se muestra datos estadísticos sobre una base geográfica como mapas.

CENSO.- Es una investigación estadística que consiste en el recuento de la totalidad de los elementos que componen la población por investigar. Es necesario que se especifique el espacio y el tiempo al que se refiere el recuento.

CICLO.- (Ver variaciones o fluctuaciones cíclicas).

CLASE MEDIANA.- En una tabla de datos agrupados, es la clase o intervalo al que pertenece el valor de la mediana.

CLASE MODAL.- En una tabla de datos agrupados, es la clase o intervalo que tiene la mayor frecuencia.

CLASE O CATEGORÍA.- Se denomina así a la característica o a los intervalos contruidos convenientemente para agrupar la información. Está conformada por el número de particiones que se realiza al conjunto de información.

CODIFICACIÓN.- Es asignar números o claves a la información para facilitar el procesamiento. Generalmente se realiza sobre las respuestas de un cuestionario, para poder identificarlas con mayor eficacia al momento del procesamiento de datos.

COEFICIENTE DE ASIMETRÍA DE FISHER.- Es un valor que indica la asimetría. Simbólicamente se representa por γ_1 . Se obtiene mediante la siguiente fórmula:

$$\gamma_1 = \frac{\mu_3}{S^3}$$

$$\mu_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

Donde:

S es la desviación estándar

Los resultados pueden ser los siguientes:

$\gamma_1 = 0$ La distribución es simétrica: existe la misma concentración de valores a la derecha y a la izquierda de la media.

$\gamma_1 > 0$ La distribución es asimétrica positiva: existe mayor concentración de valores a la derecha de la media que a su izquierda. La cola derecha es más larga.

$\gamma_1 < 0$ La distribución es asimétrica negativa: existe mayor concentración de valores a la izquierda de la media que a su derecha. La cola izquierda es más larga.

COEFICIENTE DE ASIMETRÍA DE PEARSON.- Es un valor que indica la asimetría. Simbólicamente se representa por A_s , y se obtiene mediante la siguiente fórmula:

$$A_s = \frac{3(\bar{x} - Me)}{S}$$

$$A_s = \frac{\bar{x} - Mo}{S}$$

Donde:

$\bar{\chi}$	Es la media aritmética
Mo	Es la moda
S	Es la desviación estándar
Me	Es la mediana

$A_s = 0$ Entonces la distribución es simétrica.

$A_s > 0$ Entonces la distribución es asimétrica hacia la derecha o tiene sesgo positivo.

$A_s < 0$ Entonces la distribución es asimétrica hacia la izquierda o tiene sesgo negativo.

COEFICIENTE DE CONFIANZA.- Se representa por $(1 - \alpha)$ y es la probabilidad de que la hipótesis nula H_0 no sea rechazada cuando de hecho es verdadera y debería ser aceptada.

COEFICIENTE DE CONTINGENCIA Chi-Cuadrado (χ^2).- Es un número que mide el grado de asociación o dependencia de las clasificaciones en una tabla de contingencia ($h \times k$). Se obtiene mediante la siguiente fórmula:

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \quad 0 \leq \chi^2 \leq N[\min(h, k) - 1]$$

Donde:

$$e_{ij} = \frac{n_{i \cdot} \times n_{\cdot j}}{N}; \forall i, j$$

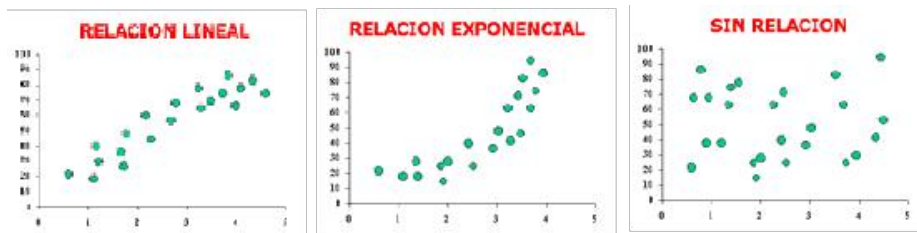
Cuanto más se acerque la Chi-Cuadrado a cero menos asociación hay (más independencia) entre los atributos.

Cuanto más se acerque la Chi-Cuadrado a su cota superior más asociación hay (menos independencia) entre los atributos.

Cuando la Chi-Cuadrado es igual a cero no hay asociación entre los atributos. Es decir los atributos son independientes.

COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON.- Es un número que mide la intensidad de la asociación lineal entre dos variables. El **coeficiente de correlación** se representa simbólicamente por "r".

Este coeficiente se aplica cuando la relación que puede existir entre las variables es lineal (es decir, si representáramos en un gráfico los pares de valores de las dos variables, la nube de puntos se aproximaría a una recta).



No obstante, puede que exista una relación que no sea lineal, sino exponencial, parabólica, etc. En estos casos, el coeficiente de correlación lineal mediría mal la intensidad de la relación de las variables, por lo que convendría utilizar un tipo de coeficiente más apropiado.

El coeficiente de correlación lineal se calcula aplicando la siguiente fórmula:

$$r = \frac{\frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$r = \frac{d[X, Y]}{s_x s_y}$$

Los valores que puede tomar el **coeficiente de correlación "r"** son: $-1 < r < 1$

Si "r" > 0 La correlación lineal es positiva (si sube el valor de una variable sube el de la otra). La correlación es tanto más fuerte cuanto más se aproxime a 1.

- Si $|r| < 0$ La correlación lineal es negativa (si sube el valor de una variable disminuye el de la otra). La correlación negativa es tanto más fuerte cuanto más se aproxime a -1.
- Si $|r| = 0$ No existe correlación lineal entre las variables, aunque podría existir otro tipo de correlación (parabólica, exponencial, etc.)

De todos modos, aunque el valor de "r" fuera próximo a 1 ó -1, tampoco esto quiere decir obligatoriamente que existe una relación de causa-efecto entre las dos variables, ya que este resultado podría haberse debido al puro azar.

COEFICIENTE DE CURTOSIS.- Es una medida de forma (Ver curtosis). Se conoce como coeficiente de curtosis de Fisher, en honor al matemático británico Ronald Fisher (1890-1962).

El valor se obtiene mediante la siguiente fórmula:

$$\gamma_2 = \frac{\mu_4}{s^4} - 3$$

Donde:

$$\mu_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$$

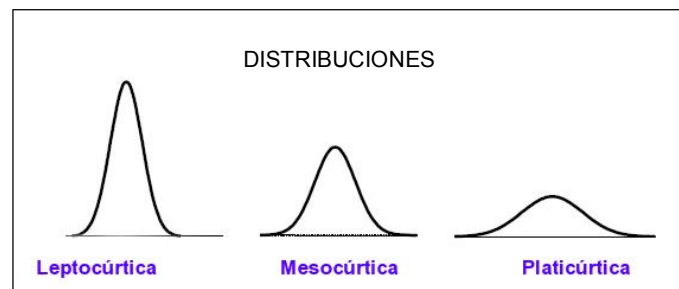
S es la desviación estándar

Los resultados pueden ser los siguientes:

$\gamma_2 > 0$ (distribución leptocúrtica).

$\gamma_2 = 0$ (distribución mesocúrtica).

$\gamma_2 < 0$ (distribución platicúrtica).



COEFICIENTE DE DETERMINACIÓN.- Es un valor que se obtiene elevando al cuadrado el coeficiente de correlación. Se representa simbólicamente por r^2 y puede tomar valores entre 0 y 1.

El coeficiente de determinación mide la proximidad del ajuste de la ecuación de regresión de la muestra a los valores observados de la variable dependiente.

COEFICIENTE DE GINI.- (Ver índice de concentración de Gini). Es una medida de la desigualdad. Mide la distribución o nivel de concentración del ingreso o renta. Su denominación es en honor al estadístico italiano Corrado Gini. El coeficiente de Gini es un número entre 0 y 1, en donde 0 se corresponde con la perfecta igualdad o distribución equitativa (todos tienen los mismos ingresos); y 1 se corresponde con la perfecta desigualdad (una persona tiene todos los ingresos y todos los demás ninguno).

COEFICIENTE DE VARIACIÓN DE PEARSON.- Es una medida de dispersión relativa y se calcula dividiendo la desviación típica entre la media aritmética:

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

La ventaja de este coeficiente es que no lleva asociado ninguna unidad de medida. Se interpreta como porcentaje, por lo que nos permitirá decidir entre dos muestras, cuál es la que presenta mayor dispersión. Simbólicamente se denota por CV.

COEFICIENTES DE REGRESIÓN.- Son los valores constantes de una ecuación de regresión lineal. En el modelo de regresión lineal siguiente los coeficientes son **a** y **b**.

$$y = a + bx$$

a —————> representa el punto de intersección con el eje

b —————> representa la pendiente de la recta

$$b = \frac{\sum XY - n \bar{Y} \bar{X}}{\sum X^2 - n \bar{X}^2}$$

$$a = \bar{Y} - b \bar{X}$$

COMBINACIONES.- Consiste en tomar diferentes agrupaciones de r elementos de un total de n objetos sin importar el orden, y el número de combinaciones se obtiene mediante la siguiente fórmula.

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

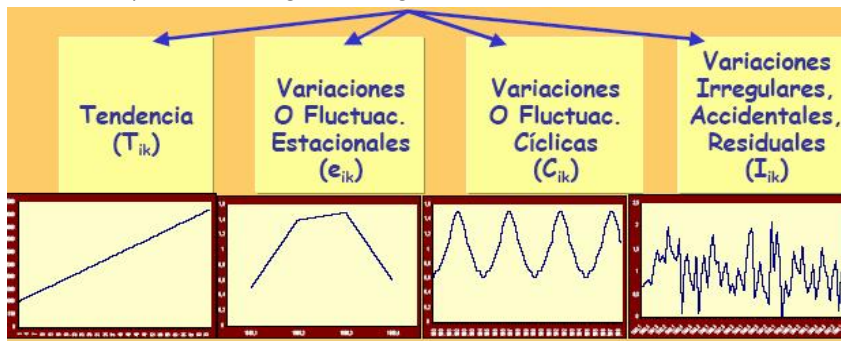
Donde:

n Representa el total de objetos

r Número de objetos agrupados

n! Representa Factorial del total de datos, se obtiene 1x2x3x...xn

COMPONENTES DE UNA SERIE TEMPORAL .- Los datos de un fenómeno se representan ordenados en el tiempo (Ver series temporales) Según el enfoque clásico una serie es el resultado de cuatro componentes: tendencia, variaciones o fluctuaciones estacionales, variaciones o fluctuaciones cíclicas y variaciones irregulares, accidentales, residuales, como se aprecia en el gráfico siguiente:



CONCENTRACIÓN.- Cuantifica el grado de equidistribución de la distribución de un fenómeno: salarios, rentas etc. Para medir el nivel de concentración de una distribución de frecuencia se puede utilizar distintos indicadores entre ellos el Índice de concentración de Gini.

CONGLOMERADO.- Es una subpoblación que reúne características presentes en la población. Los elementos que la componen poseen cierta característica que les hace ser propios de cierta cualidad o atributo, tal como lugar geográfico, grupo étnico, ideología, organización social, etc.

CONTRASTE DE HIPÓTESIS.- Conocido también como dócima o prueba de hipótesis, es el proceso estadístico que se sigue para la toma de decisiones a partir de la información de la muestra. Comparando el valor del estadístico experimental con el valor teórico, se rechaza o acepta la hipótesis nula (H_0). Lo contrario a la hipótesis nula se llama hipótesis alterna (H_1).

CORRELOGRAMA.- Es un gráfico que permite apreciar las autocorrelaciones r_1, r_2, \dots, r_k mediante el cual se identifican si los datos de una serie de tiempo tienen las siguientes características: estacionalidad, aleatoriedad, tendencia y estacionariedad.

COVARIANZA.- Es una medida de la asociación lineal entre dos variables.

$$c[X, Y] = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} = \frac{\sum_{i=1}^n X_i Y_i}{n} - \bar{X}\bar{Y}$$

Si $c[X, Y] > 0$ hay dependencia directa (positiva), es decir a grandes valores de X corresponden grandes valores de Y.

Si $c[X, Y] = 0$ las variables están incorrelacionadas, es decir no hay relación lineal.

Si $c[X, Y] < 0$ hay dependencia inversa o negativa, es decir a grandes valores de X corresponden pequeños valores de Y.

Una desventaja de la covarianza como medida de asociación es que su valor depende de las unidades en que se miden las variables de interés. Para evitar esta propiedad, se ha ideado una medida de asociación que es independiente de las unidades de medición, la cual recibe el nombre de correlación (Ver coeficiente de correlación lineal de Pearson).

CUARTIL.- Es una medida de posición no central o de localización. Los cuartiles son los tres valores que dividen la distribución en cuatro partes iguales, es decir, en cuatro intervalos dentro de cada cual están incluidos el 25% de los datos de la distribución:

- Q_1 Representa el primer cuartil y se interpreta como que el 25% de la distribución es menor que el Q_1 obtenido.
- Q_2 Representa el segundo cuartil y se interpreta como que el 50% de la distribución, es menor que el Q_2 obtenido. Este valor es igual a la mediana.
- Q_3 Representa el tercer cuartil y se interpreta como que el 75% de la distribución, es menor que el Q_3 obtenido.

FORMULA PARA DATOS AGRUPADOS
$Q_r = L_i + \frac{(rN/4) - N_{i-1}}{n_i} \times c$

$r = 1, 2, 3$

Donde:

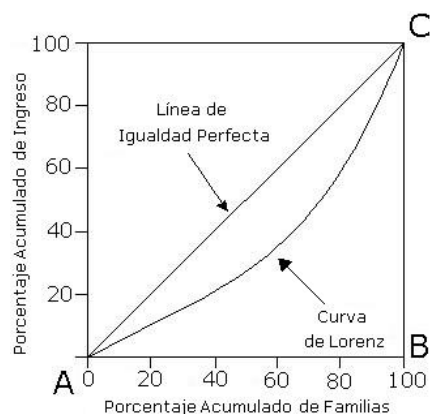
- r Es el número del cuartil que se desea calcular y puede tomar los valores de: 1, 2 y 3
- L_i Límite inferior de la clase cuartílica
- N Total de datos
- N_{i-1} Frecuencia absoluta acumulada menor o igual a $rN/4$.
- n_i Frecuencia absoluta de la clase cuartílica
- c Amplitud del intervalo

CUASIVARIANZA.- Es un valor que se obtiene de manera similar a la varianza pero dividiendo entre $n-1$ en lugar de n . La cuasivarianza cuantifica la dispersión o variabilidad de la muestra. La cuasivarianza muestral es un estimador centrado (no sesgado) de la varianza poblacional.

CUESTIONARIO.- Es el instrumento más utilizado para recolectar datos. Consiste en un conjunto de preguntas respecto a una o más variables a medir. La esencia de los cuestionarios son las preguntas que permiten alcanzar los objetivos de la investigación. Las respuestas a estas preguntas constituyen los datos estadísticos que serán utilizados para conocer las características de la población o muestra bajo estudio.

CURTOSIS.- Es una medida de forma. También se conoce como medida de apuntamiento mide si los valores de la distribución están más o menos concentrados alrededor de los valores medios de la muestra. Se definen 3 tipos de distribuciones según su grado de curtosis: Distribución mesocúrtica, distribución leptocúrtica y distribución platicúrtica. (Ver gráfico en coeficiente de curtosis).

CURVA DE LORENZ.- Es una gráfica de concentración acumulada de la distribución de la riqueza. Para elaborar una curva de Lorenz, se anotan los porcentajes acumulados del ingreso contra los porcentajes acumulados de las familias clasificadas, de las de ingresos más bajos a las de ingresos más altos. Los números requeridos se derivan de la información obtenida en la investigación. Esos pares de números determinan la curva de Lorenz. Se dibuja una línea diagonal perfecta a lo largo del cuadrante (por ejemplo el 20% del ingreso es recibido por el 20% de las familias). Mientras más cerca esté la curva de Lorenz de la línea diagonal, será más equitativa la distribución del ingreso. Por lo tanto, una medida de igualdad debe medir qué tan cerca se encuentra la curva de Lorenz de la diagonal. Una medida de este tipo es el coeficiente de Gini.



CURVA NORMAL.- También denominada curva o campana de Gauss, en honor al matemático alemán Karl Friedrich Gauss. La curva normal es una distribución simétrica de mediciones, con el mismo número de casos

a distancias específicas tanto por debajo como por encima de la media. Su media es el punto debajo del cual cae exactamente el 50% de los casos y sobre el que se encuentra el otro 50%. En estas distribuciones la media, mediana y la moda son valores idénticos. En una curva normal la mayoría de los casos se concentran alrededor de la media.



Donde:

- e es la constante 2,7182... (base de los logaritmos neperianos).
- p es 3,1415... (relación entre la longitud de la circunferencia y su diámetro).
- x es la abscisa, cualquier punto del intervalo.
- m es la mediana de la variable aleatoria.
- s es la desviación tipo de la variable aleatoria, y
- f(x) la ordenada de la curva.

D

DATO.- Conocido también como información, es el valor de la variable asociada a un elemento de una población o una muestra.

DATO CUALITATIVO.- Es aquel que representa alguna característica de los elementos de una muestra o una población que presentan, atributos, actitudes o son opiniones. Son datos NO NUMÉRICOS. (Ver variable cualitativa).

DATO CUANTITATIVO.- Es aquel dato numérico que representa aspectos de una muestra o una población que es medible o que se puede contar. (Ver variable cualitativa).

DATOS DE PANEL.- Son aquellos datos que son una combinación de series de tiempo y datos de sección cruzada o corte transversal que se obtienen sobre un mismo conjunto de unidades de análisis (individuos, familias o empresas) en distintos periodos de tiempo.

DATOS DE SECCIÓN CRUZADA O DE CORTE TRANSVERSAL.- Son aquellos que corresponden a distintas unidades de análisis (individuos, familias o empresas) pero referidos al mismo periodo de tiempo.

DECIL.- Es una medida de localización o posición no central. Los deciles son los nueve puntos que dividen la distribución en diez puntos de forma tal que dentro de cada una, están incluidos el 10% de los datos. Entonces, un decil es un valor que representa la décima parte de un conjunto de información. Se representa simbólicamente por D_r .

$$D_r = L_i + \frac{(rN/10) - N_{i-1}}{n_i} \times c \quad r = 1, 2, 3, \dots, 9$$

Donde:

- r Es el número del decil que se desea calcular. Puede tomar valores de 1,2,3,.....,9
- L_i Límite inferior de la clase decílica
- N Total de datos
- N_{i-1} Frecuencia absoluta acumulada anterior a la clase decílica
- n_i Frecuencia absoluta de la clase decílica
- c Amplitud o tamaño del intervalo

DEFLACTAR.- Es transformar valores expresados en precios corrientes (valor nominal) a valores en precios constantes (valor real). La deflactación se calcula usando la expresión siguiente:

$$\text{Valor real} = (\text{valor nominal} / \text{índice de precios}) \times 100$$

Lo cual indica el valor expresado en unidades monetarias de igual poder adquisitivo que el del año base.

DENSIDAD DE POBLACIÓN.- Es la medida más tradicional y usada con mucha frecuencia para expresar el número de habitantes por kilómetro cuadrado. Se calcula dividiendo el número de habitantes de una zona por la superficie total que tiene esa zona.

$$DN_i^z = \frac{N_i^z}{S_i}$$

Donde:

DN_i^z Representa la densidad de población del lugar "i" en el año "z".

N_i^z Representa la población total del lugar "i" en el año "z".

S_i Representa la superficie del lugar "i".

DESVIACIÓN ESTÁNDAR.- Conocida también como desviación típica, es una medida de dispersión que se obtiene como la raíz cuadrada de la varianza. (Ver varianza).

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2 n_i}{n}}$$

Datos agrupados

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

Datos simples o sin agrupar

Este estadístico se mide en la misma unidad que la variable por lo que se puede interpretar mejor que la varianza.

DESVIACIÓN MEDIA.- Es una medida de dispersión. Es un número que representa la media de los valores absolutos de las desviaciones respecto a su media aritmética. Se expresa en la misma unidad en la que se presentan los datos. Se la denota como **DM**.

$$DM = \frac{\sum_{i=1}^m |X_i - \bar{X}|n_i}{N}$$

Datos agrupados

$$DM = \frac{\sum_{i=1}^m |X_i - \bar{X}|}{N}$$

Datos simples o sin agrupar

DESVÍO TIPIFICADO (z).- Conocido también como estandarización de la distribución normal. Es la transformación de cualquier variable aleatoria normal x con media μ y una desviación estándar σ , en una variable aleatoria estandarizada de distribución normal, con media 0 y desviación típica 1.

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

DIAGRAMA.- Es un dibujo o representación gráfica que sirve para representar un objeto, indicar la relación entre elementos o mostrar el valor de una magnitud.

DIAGRAMA DE BARRAS.- Es un gráfico utilizado para representar la distribución de frecuencias de una variable cualitativa y cuantitativa discreta. Puede graficarse en forma horizontal o vertical.

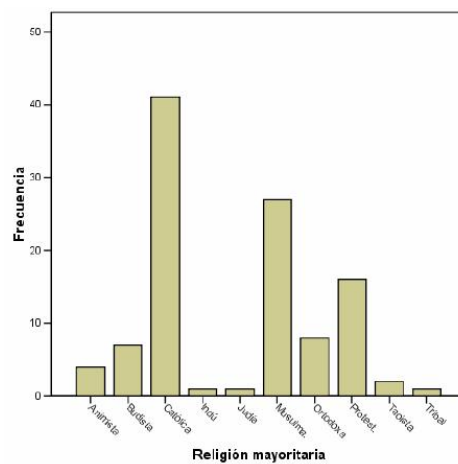
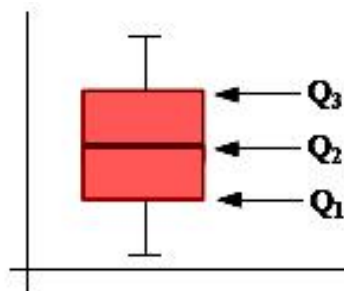


DIAGRAMA DE BASTONES (ESPECTRO).- Es un gráfico utilizado para representar una distribución de frecuencias o frecuencias relativas de una variable numérica (en general discreta) sin agrupar.



DIAGRAMA DE CAJAS.- Conocido también como **BOXPLOT**. Es un importante gráfico del análisis exploratorio de datos. Al igual que el histograma, permite tener una idea visual de la distribución de los datos. Permite determinar si hay simetría, ver el grado de variabilidad existente y detectar los "outliers" (datos muy diferentes al conjunto de información), es decir la existencia de posibles datos discordantes. Además, el Boxplot es bien útil para comparar grupos. Es un diagrama que muestra la distancia en que se encuentran los datos y cómo están distribuidos equitativamente.



Recorrido intercuartílico $RI = Q_3 - Q_1$

DIAGRAMA DE DISPERSIÓN.- Es un gráfico utilizado para representar la relación entre los valores observados de dos variables numéricas. También se conoce como nube de puntos.

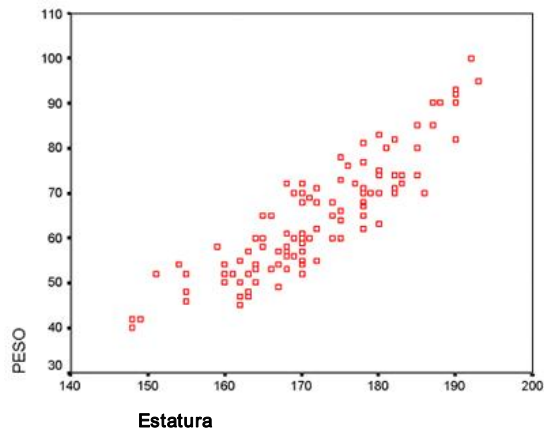


DIAGRAMA DE FLUJO.- Es una representación gráfica de los pasos en un proceso. Útil para determinar cómo funciona realmente el proceso para producir un resultado. El resultado puede ser un producto, un servicio, información o una combinación de los tres. Los diagramas de flujo se pueden aplicar a cualquier aspecto del proceso desde el flujo de materiales hasta los pasos para realizar la venta u ofrecer un producto.

DIAGRAMA DE FLUJO DE LA DIVISIÓN DE DOS NÚMEROS

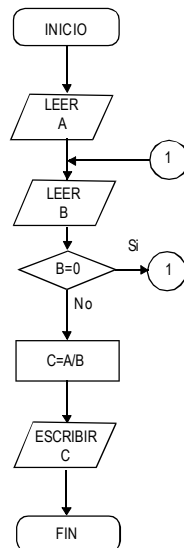


DIAGRAMA DE PARETO.- Es una forma especial de gráfico de barras verticales donde se organizan diversas clasificaciones de datos por orden descendente, de izquierda a derecha, por medio de barras sencillas después de haber reunido los datos para calificar las causas. De modo que se pueda asignar un orden de prioridades, separa los problemas muy importantes de los menos importantes, estableciendo un orden de prioridades.

El nombre de Pareto fue dado por Joseph Juran en honor del economista italiano Vilfredo Pareto (1848-1923) quien realizó un estudio sobre la distribución de la riqueza, en el cual descubrió que la minoría de la población poseía la mayor parte de la riqueza y la mayoría de la población poseía la menor parte de la riqueza. Con esto estableció la llamada "Ley de Pareto" según la cual la desigualdad económica es inevitable en cualquier sociedad. Juran aplicó este concepto a la calidad, obteniéndose lo que hoy se conoce como la regla 80/20.

Según este concepto, si se tiene un problema con muchas causas, podemos decir que el 20% de las causas resuelven el 80% del problema y el 80% de las causas sólo resuelven el 20% del problema.

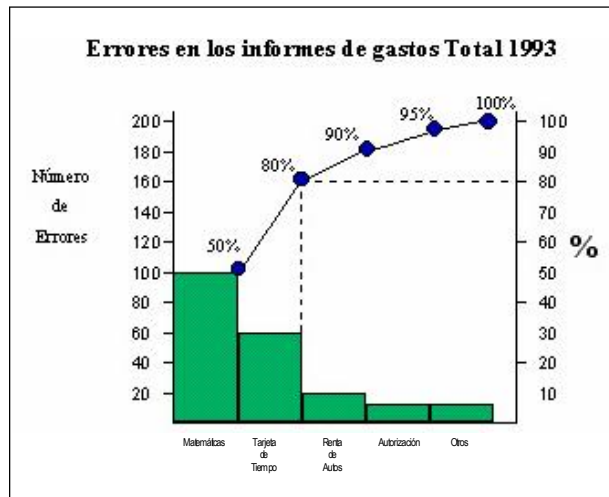


DIAGRAMA DE SECTORES.- Es un gráfico utilizado para representar la distribución de frecuencias relativas de una variable cualitativa. (Ver gráfico circular).

Hábitos de fumar

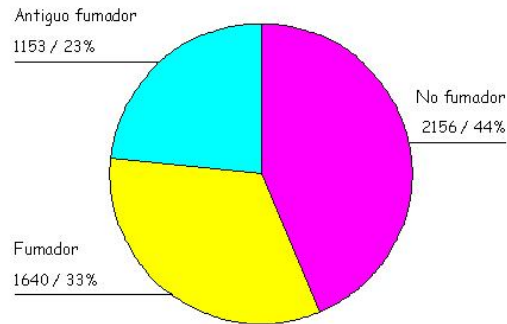


DIAGRAMA DE TALLOS Y HOJAS.- Es una forma semigráfica de representar una distribución de frecuencias de una variable numérica.

Vamos a construir un diagrama de tallo y hojas para el siguiente conjunto de 20 puntajes de ingreso a la universidad:

62 68 72 92 86 76 52 76 82 78 82 74 88 66 58 74 78 84 96 76

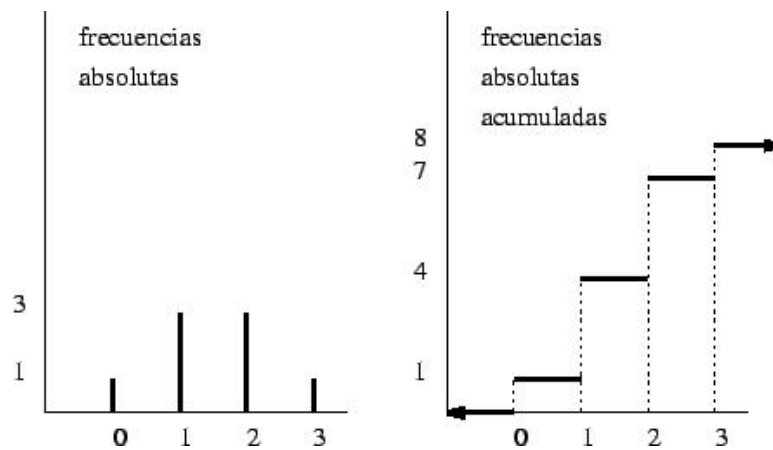
Siendo los datos números de dos cifras, vemos que hay datos en los grupos del 50, 60, 70, 80 y 90. El primer dígito de cada dato debe utilizarse como tallo y el segundo como hoja. Se traza una línea vertical y se colocan los tallos a su izquierda, en columna. Luego se coloca cada hoja junto a su tallo hasta completar la lectura de todos los datos. La presentación de tallo y hojas es la siguiente.

Frecuencia	Tallo	Hojas
2	5	8 2
3	6	6 2 8
8	7	4 4 8 6 2 6 6 8
5	8	2 8 4 6 2
2	9	6 2
	N = 20	
	Unidad = 1	

Proporciona una idea de la distribución de la variable en estudio.

Si los datos tuviesen cifras decimales, al construir el diagrama de tallo y hojas, el punto decimal se pierde por tal razón se acostumbra indicar las unidades que los datos del tallo representan. Así, si los datos de arriba fuesen decimales 6.2 6.87.6, debajo del diagrama se pondría "Unidad = 0.1".

DIAGRAMA EN ESCALERA.- Es un gráfico utilizado para representar la distribución de frecuencias acumuladas de una variable discreta numérica.



DISTRIBUCIÓN BIDIMENSIONAL.- Es la disposición de la frecuencia de dos variables de cada elemento de la población. Por ejemplo: peso y altura de un grupo de estudiantes, superficie y precio de las viviendas de una ciudad, potencia y velocidad de una gama de autos deportivos etc.

Sea una población donde se estudia simultáneamente dos características X e Y, se representa genéricamente como (x_i, y_j, n_{ij}) , donde x_i, y_j , son dos valores cualesquiera y n_{ij} es la frecuencia absoluta conjunta del valor i-ésimo de X con el j-ésimo de Y.

Una forma de disponer estos resultados es la conocida como tabla de doble entrada o tabla de contingencia y se representa como sigue:

X \ Y	y ₁	y ₂	y _j	y _k	n _{i .}
x ₁	n ₁₁	n ₁₂	n _{1j}	n _{1k}	n _{1 .}
x ₂	n ₂₁	n ₂₂	n _{2j}	n _{2k}	n _{2 .}
.
.
x _i	n _{i1}	n _{i2}	n _{ij}	n _{ik}	n _{i .}
.
.
.
x _h	n _{h1}	n _{h2}	n _{hj}	n _{hk}	n _{h .}

En este caso, n_{11} nos indica el número de veces que se repite x_1 conjuntamente con y_1 , n_{12} , nos indica la frecuencia conjunta de x_1 con y_2 , etc.

DISTRIBUCIÓN CONDICIONAL.- De una tabla de frecuencias bidimensionales se puede formar varias distribuciones unidimensionales en las que previamente hace falta definir una condición. Las distribuciones surgen al fijar un valor de una de las variables (condicionante) y considerar la distribución de los valores de la otra variables (condicionada).

- Al condicionar reducimos el número de elementos de la distribución defina por un valor específico de la otra variables.
- El número total de distribuciones condicionadas es $h+k$

h número de filas
 k número de columnas

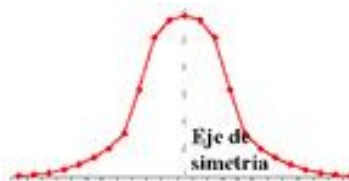
DISTRIBUCIÓN MARGINAL.- Es la distribución de frecuencias de una variable independientemente de cómo se comporta la otra variable de una distribución bidimensional. De cada distribución bidimensional se pueden deducir dos distribuciones marginales: una correspondiente a la variable «x» y otra correspondiente a la variable «y».

X	n_{i.}
X ₁	n _{1.}
X ₂	n _{2.}
.....	...
X _{n-1}	n _{n-1.}
X _n	n _{n.}

Y	n_{.j}
y ₁	n _{.1}
y ₂	n _{.2}
.....	...
y _{m-1}	n _{.m-1}
y _m	n _{.m}

DISTRIBUCIÓN LEPTOCÚRTICA.- Es aquella que presenta un elevado grado de concentración alrededor de los valores centrales de la variable.

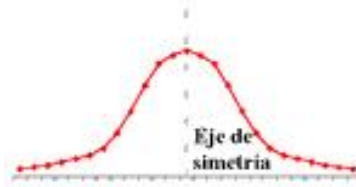
CURVA LEPTOCÚRTICA



DISTRIBUCIÓN MESOCÚRTICA.- Es conocida también como curva normal o campana de Gauss. Es aquella que presenta un grado de

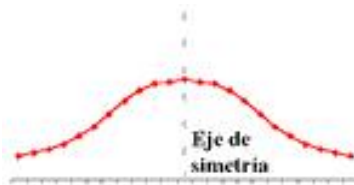
concentración alrededor de los valores centrales de la variable. (Ver curva normal).

CURVA MESOCÚRTICA



DISTRIBUCIÓN PLATICÚRTICA.- Es aquella que presenta un reducido grado de concentración alrededor de los valores centrales de la variable.

CURVA PLATICÚRTICA



DISTRIBUCIÓN UNIDIMENSIONAL.- Es una tabla resumen en la que se estudia una sola variable. Los datos se disponen según agrupamientos o categorías convenientemente establecidas. Puede construirse con variable cualitativa o cuantitativa.

Categorías o clases	Frecuencias Absolutas n_i	Frecuencias Relativas h_i	Frecuencias Absolutas acumuladas N_i	Frecuencias Relativa acumuladas H_i
1	n_1	h_1	N_1	H_1
2	n_2	h_2	N_2	H_2
.
i	n_i	h_i	N_i	H_i
.
.
m	n_m	h_m	N_m	H_m
Total	$\sum_{i=1}^m n_i = N$	$\sum_{i=1}^m h_i = 1$		

DISTRIBUCIÓN NORMAL O CURVA NORMAL.- Llamada también como distribución de Gauss, es la distribución de probabilidad más utilizada en estadística y teoría de probabilidad. Esto se debe a dos razones:

- Su función de densidad es simétrica y con forma de campana lo que favorece su aplicación como modelo a gran número de variables.
- Es además límite de otras distribuciones y aparece relacionada con resultados ligados a la teoría de las probabilidades gracias a sus propiedades matemáticas. La función de densidad está dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad - \quad \infty < x < \infty$$

Donde:

- μ Media
- σ^2 Varianza
- σ Desviación estándar
- π Constante = 3,1415.....
- e Constante = 2,7182.....

E

ENCUESTA.- Es un método de recolección de datos. Es llevada a cabo generalmente a través de algún cuestionario que puede o no ser diligenciado por el encuestado y/o encuestador.

ENTREVISTA.- Es un método de recolección de datos. Consiste en una serie de preguntas realizadas por el entrevistador, personalmente, a cada uno de los entrevistados.

ERROR DE MUESTREO.- Conocido también como error muestral, es la diferencia que existe entre el valor real (parámetro) obtenido con los valores de la población y el valor estimado en base a los valores de una muestra (estimación).

ERROR TIPO I.- En la teoría de decisiones, es el error que se comete al rechazar la hipótesis nula H_0 , cuando es verdadera.

ERROR TIPO II.- En la teoría de decisiones, es el error que se comete al aceptar la hipótesis nula H_0 cuando es falsa.

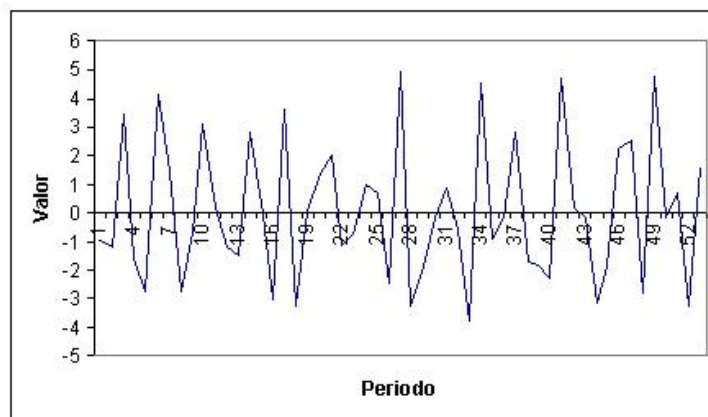
DECISIONES POSIBLES	HIPÓTESIS NULA H_0 VERDADERA	HIPÓTESIS NULA H_0 FALSA
Se acepta la H_0	Correctamente aceptada	Error de tipo II
Se rechaza H_0	Error de tipo I	Correctamente rechazada

ESPACIO MUESTRAL.- Es el conjunto de todos los resultados posibles de un experimento aleatorio. Cada experimento aleatorio tiene definido su espacio muestral (es decir, un conjunto con todas las soluciones posibles).

Ejemplo: si tiramos una moneda al aire una sola vez, el espacio muestral será cara o sello.

Si el experimento consiste en lanzar una moneda al aire dos veces, entonces el espacio muestral estaría formado por (cara-cara), (cara-sello), (sello-cara) y (sello-sello).

ESTACIONARIA.- Es la serie de datos cuyas propiedades estadísticas básicas como la media y la varianza permanecen constantes en el tiempo, es decir cuando la serie no presenta crecimiento o declinación es estacionaria.



ESTADÍSTICA.- Es la ciencia que comprende una serie de métodos y procedimientos destinados a la recopilación, tabulación, procesamiento, análisis e interpretación de datos cuantitativos y cualitativos. Un objetivo de la estadística es describir "la población del estudio" en base a información obtenida de elementos individuales. Se divide en dos ramas: Estadística descriptiva y Estadística inferencial

ESTADÍSTICA DESCRIPTIVA.- Rama de la ciencia estadística que se encarga desde la recopilación, procesamiento y análisis de la información siendo sus conclusiones válidas sólo para el grupo analizado.

ESTADÍSTICA INFERENCIAL.- Rama de la ciencia estadística que proporciona métodos y procedimientos que permiten obtener conclusiones para una población a partir del estudio de una o más muestras representativas.

ESTADÍSTICO.- Conocido también como estadígrafo, es el valor calculado en base a los datos que se obtienen sobre una muestra y por lo tanto es una estimación de los parámetros. Entre los más usados se tiene la media muestral y la desviación estándar muestral.

ESTIMADOR.- Es un estadístico empleado para estimar un parámetro.

ESTIMADOR INSESGADO.- Es un tipo de estimador que posee la propiedad de que el promedio de las estimaciones efectuadas a partir de todas las muestras posibles de un determinado tamaño es igual al valor verdadero o valor poblacional.

ESTRATIFICACIÓN.- Es un procedimiento por medio del cual una población se divide en grupos llamados estratos, con el propósito de seleccionar una muestra separada en cada grupo. Cada uno de estos grupos o estratos debe ser internamente lo más homogéneo posible.

ESTRATO.- Es una subpoblación o parte de una población que reúne características comunes que le hacen ser homogénea. Los estratos son mutuamente excluyentes. Ello significa que los elementos que pertenecen a un estrato no pueden pertenecer a otro.

EXACTITUD.- Es la cercanía de una medición al verdadero valor que se pretende medir.

EXPERIMENTO.- Es un método de investigación mediante el cual se determina la incidencia de variables independientes sobre la variable dependiente.

EXPERIMENTO ALEATORIO.- Es cualquier acto que implique la observación de los valores de una variable aleatoria. Es aquel que puede dar lugar a varios resultados, sin que pueda ser previsible enunciar con certeza cuál de éstos va a ser observado en la realización del experimento.

F

FACTOR DE EXPANSIÓN.- Es un número constante (factor o multiplicador) por medio del cual el valor de la variable muestral se expande o eleva a nivel de la población total. El factor de expansión es el recíproco o inverso de la fracción de muestreo.

FRACTIL O CUANTIL.- Es el valor que se obtiene al fraccionar el conjunto de datos en partes o fracciones iguales. Los más conocidos son: mediana, cuartiles, deciles y percentiles.

FRECUENCIA ABSOLUTA.- Es el número de veces que la variable asume un valor dado o pertenece a una clase dada. Se representa simbólicamente por n_i .

FRECUENCIA ABSOLUTA ACUMULADA.- Es el número de observaciones hasta (inclusive) un valor dado de una variable numérica. Se representa por N_i .

$$N_i = n_1 + n_2 + \dots + n_i = \sum_{j=1}^i n_j$$

FRECUENCIA CONDICIONAL.- En una distribución conjunta, son las frecuencias de una de las variables estando fijo un valor de la (s) otra (s) variable(s).

FRECUENCIA CONJUNTA.- Es un número n_{ij} que representa la ocurrencia de dos variables (x, y) en los elementos de población o de la muestra. (Ver distribución bidimensional)

FRECUENCIA MARGINAL.- En una distribución conjunta, son las frecuencias de cada una de las variables sin tener en cuenta el valor de la (s) otra (s).

FRECUENCIA RELATIVA.- Es un valor que se obtiene como el cociente de la frecuencia absoluta (n_i) sobre el tamaño de la muestra (N). Simbólicamente se representa por h_i .

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

FRECUENCIA RELATIVA ACUMULADA.- Es una cantidad que se obtiene como el cociente de la frecuencia absoluta acumulada (N_i) sobre el tamaño de la muestra (N). Simbólicamente se representa por H_i .

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

FUENTES DE DATOS.- Medios de donde procede la información. Los datos pueden reunirse de diferentes fuentes de información ya existentes o pueden obtenerse mediante censos, encuestas y estudios experimentales para conseguir nuevos datos.

FUENTE PRIMARIA.- Es aquella en la que los datos estadísticos se obtienen a partir de un relevamiento propio, como por ejemplo a partir de una encuesta.

FUENTE SECUNDARIA.- Es aquella persona o institución que proporciona datos estadísticos, es decir los datos se obtienen a partir de un relevamiento de otros recopiladores.

G

GRADO DE URBANIZACIÓN.- Es el porcentaje de población que reside en las zonas urbanas (ciudades) de un país, región o lugar. Se define como el cociente de la población urbana entre el total de la población, multiplicado por 100. Se expresa como porcentaje:

$$PNU_i^z = \frac{NU_i^z}{N_i^z} \times 100$$

donde:

PNU_i^z representa el porcentaje de población urbana del lugar "i" en el año "z".

NU_i^z representa la población urbana que reside en el lugar "i" en el año "z".

N_i^z representa la población total del lugar "i" en el año "z".

GRADOS DE LIBERTAD.- En estadística grados de libertad de un estadístico calculado en base a «n» datos, se refiere al número de cantidades independientes que se necesitan en su cálculo, menos el número de restricciones que ligan a las observaciones y el estadístico. Simbólicamente se representa por gl.

Ejemplo: Sea $X_i : 2, 5, 7, 9, 12$ su media es $\bar{X} = 7$ y se ha calculado a partir de $n=5$ observaciones independientes, que están ligadas por la media aritmética. Luego el número de grados de libertad de la media es $n-1=4$

GRÁFICO CIRCULAR.- Conocido también como gráfico de sectores circulares. Está formado por un círculo dividido en sectores, de modo que cada uno de ellos representa una categoría distinta de la variable observada, manteniendo su proporción relativa respecto del total de la muestra. (Ver diagrama de sectores).

GRÁFICO DE ÁREAS.- Gráfico que busca mostrar la tendencia de la información generalmente en un período de tiempo. Pueden ser para representar una, dos o más series en dos, o tres dimensiones.

GRÁFICO DE BARRAS.- Ver diagrama de barras.

GRÁFICO DE CAJAS.- (Ver diagrama de cajas).

GRÁFICO DE LÍNEAS.- Diagrama donde se representa con líneas los valores de los datos en dos ejes cartesianos ortogonales entre sí. Se puede usar para representar una, dos o más series.

GRÁFICO SEMILOGARÍTMICO.- Es un diagrama donde uno de los ejes está en escala logarítmica. Se utiliza cuando hay grandes incrementos entre sí.

H

HIPÓTESIS ESTADÍSTICA.- Es una afirmación respecto a alguna característica de la población en estudio que se formula para ser sometida a la denominada prueba de hipótesis, para ser aceptada o rechazada.

HISTOGRAMA.- Gráfico utilizado para representar la distribución de frecuencias de una variable continua. Describe el comportamiento de un conjunto de datos en cuanto a su tendencia central, forma y dispersión. Está formado por un conjunto de rectángulos unidos, cuya base es igual a la amplitud del intervalo, y la longitud proporcional a la frecuencia.

I

INDEPENDENCIA ESTADÍSTICA.- Se dice que dos variables X e Y son independientes, estadísticamente, cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales en todos los casos, es decir:

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} ; \forall i, j > 0$$

Si esta condición no se cumple para todos los valores, se dice que hay dependencia estadística.

ÍNDICE.- Es la relación expresada en porcentaje entre el precio, cantidad o valor de un bien y servicio o conjunto de bienes y servicios, en un período

de estudio y el precio, cantidad o valor del mismo bien y servicio o conjunto de bienes y servicios en el periodo de referencia o período base

El número índice es :
$$I_t^0 = \frac{X_{it}}{X_{io}} * 100$$

X_{io} Representa el valor de la magnitud en el periodo base

X_{it} Representa el valor de la magnitud en el periodo que se quiere estudiar

El Índice mide la variación (expresado en porcentaje) que ha sufrido la magnitud X entre los dos periodos considerados. Puede referirse a precios, cantidad y valor. (Ver número índice).

ÍNDICE DE CANTIDAD.- Es un número que refleja la variación en las cantidades de un producto o un conjunto de productos en dos momentos en el tiempo. Ejemplos: índice de exportación del algodón, el índice de producción industrial, variación de la carga transportada, entre otros.

$$q_t^0 = \frac{q_{it}}{q_{io}} * 100$$

ÍNDICE DE PRECIOS.- Es un número que refleja el cambio en el precio de un bien, servicio o conjunto de bienes y servicios en un período de tiempo, en relación con el precio en un período de referencia (período base).

$$P_t^0 = \frac{P_{it}}{P_{io}} * 100$$

ÍNDICE DE VALOR.- Es un número que expresa la variación en el valor de un conjunto de productos en dos momentos en el tiempo o el espacio. Ejemplo: índice de ventas comerciales, valor de las exportaciones, deuda externa, entre otros.

$$V_t^0 = \frac{P_{it}q_{it}}{P_{io}q_{io}} * 100$$

ÍNDICE AGREGATIVO.- Es aquel que expresa la variación de un conjunto de artículos agregados. Entre ellos tenemos el índice de Laspeyres, Paasche y Fisher.

ÍNDICE DE PRECIOS DE LASPEYRES.- Describe la variación de precios de una canasta de bienes y servicios elegidos en un año base, que permanece inalterable durante los períodos sucesivos.

$$IP_L = \frac{\sum Q_0 P_t}{\sum Q_0 P_0} \times 100$$

Donde:

- P_0 Precio del año base
 Q_0 Cantidad del año base
 P_t Precio del año dado

ÍNDICE DE PRECIOS DE PAASCHE.- Es un número que describe la relación existente entre el precio actual de un grupo de bienes y servicios y el precio de dichos bienes y servicios en el año base. A diferencia del índice de precios de Laspeyres donde se mantenían fijas las cantidades de la canasta de bienes y servicios del período base, para el índice de precio de Paasche estas cantidades van variando y corresponden a las del período corriente (período actual).

El índice de precios de Paasche está definido por:

$$IP_p = \frac{\sum Q_t P_t}{\sum Q_t P_0} \times 100$$

ÍNDICE IDEAL DE FISHER.- Es un índice de precios que se obtiene como la media geométrica de los números índices de Laspeyres y de Paasche. El índice ideal de Fisher satisface los criterios de inversión temporal y de inversión de factores, lo que le confiere una cierta ventaja teórica sobre otros números índice.

Se obtiene de la combinación de los índices de Laspeyres y Paasche:

$$I_{t/o} = \sqrt{\left(\frac{\sum P_t Q_0}{\sum P_0 Q_0}\right) \left(\frac{\sum P_t Q_t}{\sum P_0 Q_t}\right)} = \sqrt{I_L \times I_P}$$

ÍNDICE DE CARLI.- Es un índice agregado simple. Si los precios de un conjunto de bienes en el período base están dados por $P_{o1}, P_{o2}, P_{o3}, P_{o4}$, etc., y los precios de estos mismos bienes para el período dado t son $P_{t1}, P_{t2}, P_{t3}, P_{t4}$, etc., entonces el índice de Carli se define como la media aritmética de la evolución de los precios relativos:

$$I_{t/o} = \frac{1}{n} \sum \left(\frac{P_t}{P_0} \right) \times 100$$

Donde n es el número de bienes y la suma de (P_t / P_0) se extiende a todos los bienes.

ÍNDICE DE CONCENTRACIÓN DE GINI.- Es el coeficiente expresado en porcentaje. Aunque el coeficiente de Gini se utiliza, sobre todo, para medir la desigualdad en los ingresos también puede utilizarse para medir la desigualdad en la riqueza.

El coeficiente se calcula como el doble del área encerrada por la Curva de Lorenz y la diagonal.

Este índice se calcula aplicando la siguiente fórmula:

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

En donde p_i mide el porcentaje de individuos de la muestra que presentan un valor igual o inferior al de x_i .

$$p_i = \frac{n_1 + n_2 + \dots + n_i}{n} * 100$$

Mientras que q_i se calcula aplicando la siguiente fórmula:

$$q_i = \frac{(x_1 * n_1) + (x_2 * n_2) + \dots + (x_i * n_i)}{(x_1 * n_1) + (x_2 * n_2) + \dots + (x_n * n_n)} * 100$$

El Índice Gini (IG) puede tomar valores entre 0 y 1:

- IG = 0 Concentración mínima. Indica que la muestra está uniformemente repartida a lo largo de todo su rango. Distribución perfecta equitativa.
- IG = 1 Concentración máxima. Indica que un solo individuo acumula el 100% de los resultados. Distribución perfecta desigual.

ÍNDICE DE MARSHALL-EDGEWORTH.- Índice que se calcula por el método de agregación ponderada. Utiliza como ponderación la media aritmética de las cantidades consumidas en el año base y en el año de estudio. (período en que se calcula el índice).

$$\text{Índice de Marshall-Edgeworth} = \frac{\sum p_n (q_0 + q_n)}{\sum p_0 (q_0 + q_n)}$$

Donde:

- q_0 Representa cantidades del año base.
- q_n Representa cantidades del año dado.

ÍNDICE DE MASCULINIDAD.- Es un número que relaciona el número de hombres por cada 100 mujeres, expresado como porcentaje.

Es un indicador básico para el análisis de la distribución por sexo de la población. Se expresa como la relación por cociente entre el número de varones y el número de mujeres en una población dada o en parte de ella. Se define como:

$$IM(X) = \frac{NH(X)}{NF(X)} * 100$$

donde:

$IM(X)$ es el índice de análisis correspondiente a la edad X .

$NH(X)$ es el total de varones a la edad X .

$NF(X)$ es el número total de mujeres a la edad X

ÍNDICE DE PRECIOS AL CONSUMIDOR (IPC).- Es un indicador económico que muestra la variación en los precios de un conjunto de bienes y servicios (canasta familiar) que consume habitualmente un grupo representativo de familias de diversos estratos socio-económicos de un país.

Esto nos indica qué tanto más cara o más barata está la canasta (los bienes y servicios seleccionados) en el periodo actual, en comparación con el periodo base, expresándolo como un porcentaje.

La ponderación de los bienes y servicios (artículos) que componen la canasta familiar son los pesos relativos medidos en términos de valores de gasto, con relación al gasto total de los hogares. Las ponderaciones permanecen fijas hasta un nuevo cambio de base del índice.

ÍNDICE DE ENVEJECIMIENTO.- Es un valor que se obtiene dividiendo el número de personas de 60 y más años entre el número de los menores de 15 años, multiplicado por 100.

El descenso de los niveles de mortalidad y fecundidad a través del tiempo produce el envejecimiento de la población; esto es, disminuye la proporción de la población menor de 15 años y a la vez aumenta la proporción de adultos mayores, fenómeno que se conoce como envejecimiento de la población. Se expresa como

$$IV = \frac{N(60 y +)}{N(0 - 14)} \times 100$$

donde:

IV representa el índice de envejecimiento o vejez.

$N(60 y +)$ representa la población de 60 y más años de edad.

$N(0 - 14)$ representa la población de menores de 15 años de edad.

INFERENCIA ESTADÍSTICA.- Es una parte de la estadística cuya finalidad es obtener conclusiones respecto a la población a partir de datos observados en muestras. Es el proceso por medio del cual se hacen aseveraciones o estimaciones de un todo, a partir de sus partes o elementos.

INTERVALO DE CLASE.- Es el conjunto de datos cuantitativos comprendido entre dos valores. Generalmente se ubican en la primera columna en una tabla de distribución de frecuencias.

Se conoce intervalos abiertos, semiabiertos, cerrados y semicerrados, en función a la inclusión de los valores extremos.

INTERVALO DE CONFIANZA.- Conocido también como límites de confianza. Es un rango de valores en el cual se encontraría el valor del parámetro, con una probabilidad determinada.

Generalmente se construye intervalos de confianza con 95% de probabilidad (Ver parámetro).

L

LÍMITE INFERIOR.- Es el menor valor de un intervalo de clase.

LÍMITE SUPERIOR.- Es el mayor valor de un intervalo de clase.

M

MARCA DE CLASE.- Es la denominación que se le da al punto medio de un intervalo en una tabla de frecuencias de datos agrupados. Hay tantas marcas de clase como intervalos tenga la variable. Simbólicamente se representa por x_i .

MARCO MUESTRAL.- Es la totalidad de unidades de muestreo de la que se selecciona una muestra. El marco puede ser una lista de personas, o unidades de vivienda, hogares, un archivo de registros, un mapa subdividido, una foto aérea con detalles, entre muchos otros.

MEDIA ARITMÉTICA PARA DATOS SIMPLES.- Es una medida de tendencia central que denota el promedio de un conjunto de datos. Se calcula dividiendo la suma del conjunto de datos entre el total de ellos. Simbólicamente se representa por: \bar{X}

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{N}$$

MEDIA ARITMÉTICA PARA DATOS AGRUPADOS.- Es una medida de tendencia central. Se calcula multiplicando cada valor de los elementos por el número de veces que se repite. La suma de todos estos elementos se divide entre el total de datos:

$$\bar{X} = \frac{(X_1 * n_1) + (X_2 * n_2) + (X_3 * n_3) + \dots + (X_{m-1} * n_{m-1}) + (X_m * n_m)}{N}$$

La media aritmética de una variable se define como la suma ponderada de los valores de la variable por sus frecuencias relativas. Se denota por \bar{X} y se calcula mediante la expresión:

$$\bar{X} = \frac{\sum_{i=1}^m x_i * n_i}{N}$$

x_i representa el valor de la marca de clase o punto medio del intervalo.
 n_i representa la frecuencia absoluta
 N representa el total de datos.

MEDIA ARMÓNICA.- Es un valor que se obtiene como la inversa de la media de las inversas de las observaciones. Se le denota por H.

$$H = \frac{1}{\sum_{i=1}^n \frac{1}{x_i} \cdot n_i}$$

Donde:

x_i representa el valor de la variable o en su caso la marca de clase.

n_i representa la frecuencia absoluta

MEDIA GEOMÉTRICA.- Es una medida de tendencia central. Dado dos números y_1 e y_2 , llamaremos media geométrica (G) de estos números a la raíz cuadrada del producto de los mismos. Cuando se tiene N observaciones (más de dos datos): x_1, x_2, \dots, x_p y cada uno de ellos se repite n_1, n_2, \dots, n_p veces entonces, generalizando la primera expresión se tiene:

$$G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_p^{n_p}}$$

Solo se puede calcular si no hay observaciones negativas o valores cero. Es menos sensible que la media aritmética a los valores extremos. Su valor es siempre menor o igual que la media aritmética. Su uso más frecuente es el de promediar porcentajes, tasas, números índices, entre otros, es decir en los casos que se supone que la variable presenta variaciones acumulativas.

MEDIANA.- Es una medida de tendencia central. Es el valor que divide al conjunto de datos ordenados, en aproximadamente dos partes: 50% de valores son inferiores y otro 50% son superiores. Por ejemplo, si decimos que la mediana de los sueldos de los obreros de una empresa es de 800 soles mensuales, estamos indicando que el 50% gana menos que 800 y el otro 50% gana más.

Simbólicamente se representa por \tilde{x} ó Mn

Cálculo de la mediana para datos no agrupados en intervalos: Tendremos en cuenta el número de datos N :

Si tenemos N datos y N es impar, hay un término central entonces este valor central es la mediana.

Si N es par, hay dos términos centrales, la mediana será la semisuma de esos dos valores.

Cálculo de la mediana en datos agrupados en intervalos:

Si la variable se encuentra representada por intervalos, se calcula mediante la siguiente fórmula:

$$\tilde{X} = LI + \frac{(N/2) - N_{i-1}}{n_i} * c_i$$

Donde:

- LI Es el límite inferior de la clase mediana.
- N_{i-1} Es la frecuencia absoluta acumulada anterior o igual a la frecuencia de la clase mediana.
- n_i Frecuencia de la clase mediana
- N Total de datos.
- c_i Es la amplitud del intervalo de la clase mediana.

MEDIDA DE ASOCIACIÓN.- Es un valor o medida que indica cuánto varían conjuntamente dos o más variables. (Ver coeficiente de correlación).

MEDIDAS DE ASIMETRÍA.- Son aquellas orientadas a elaborar un indicador para establecer el grado de simetría (o asimetría) que presenta la distribución, sin necesidad de una representación gráfica. Se mide con el coeficiente de Fisher y el de Pearson. (Ver coeficiente de asimetría).

MEDIDAS DE DISPERSIÓN.- Son aquellas medidas de resumen que, de acuerdo a algún criterio, reflejan la heterogeneidad de las observaciones. Dan una idea sobre la representatividad de las medidas de tendencia central, a mayor dispersión menor representatividad. Entre ellas: desviación media, varianza, desviación típica, coeficiente de variación, entre otros.

MEDIDAS DE FORMA.- Permiten conocer que forma tiene la curva que representa la serie de datos. Entre estas medidas tenemos las de concentración, asimetría y curtosis. (Ver índice de concentración de Gini, coeficiente de asimetría y coeficiente de curtosis).

MEDIDAS DE POSICIÓN.- Resumen características generales de la ubicación de la distribución de los datos dentro de un conjunto de valores posibles. Estas pueden ser de tendencia central y no central.

MEDIDAS DE POSICIÓN DE TENDENCIA CENTRAL.- Son medidas de resumen que, de acuerdo a algún criterio, indican un valor alrededor del cual se distribuyen las observaciones. Se tiene a: la media, mediana y moda, media geométrica y media armónica.

MEDIDAS DE POSICIÓN DE TENDENCIA NO CENTRALES.- Conocido también como medidas de localización. Son aquellos valores que permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Son valores de la distribución que la dividen en partes iguales, es decir en intervalos que comprenden el mismo número de datos como los cuartiles, deciles y percentiles.

MÉTODO DE MÍNIMOS CUADRADOS.- Es un método utilizado para determinar los parámetros de una ecuación de regresión que mejor se ajuste al conjunto de puntos. El método consiste en minimizar la suma de las diferencias de los valores observados y estimados al cuadrado. Cuando se utiliza este método en regresión, la función ecuación se llama ecuación de regresión mínimo cuadrática.

MODA.- Es una medida de tendencia central es el valor de la variable que tiene mayor frecuencia absoluta, la que más se repite es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo. Por su propia definición, la moda no es única, pues puede haber dos o más valores de la variable que tengan la misma frecuencia siendo esta máxima. Entonces tendremos una distribución bimodal o polimodal según el caso.

Considerando distribuciones unimodales, el cálculo de la moda (M_o) para datos agrupados en intervalos se obtiene mediante la fórmula:

$$M_o = LI + \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})} * c_i$$

Donde:

LI Es el límite inferior de la clase modal.

$n_i - n_{i-1}$ Es la diferencia de la frecuencia absoluta de la clase modal menos la frecuencia del intervalo anterior.

$n_i - n_{i+1}$	Es la diferencia de la frecuencia absoluta de la clase modal menos la frecuencia del intervalo posterior
c_i	Es la amplitud del intervalo.

Clase modal es el intervalo que tiene mayor frecuencia o frecuencia relativa.

MUESTRA.- Es un subconjunto representativo de la población a partir del cual se pretende realizar inferencias respecto a la población de donde procede. Los elementos seleccionados con cierta técnica reúne ciertas características que la hacen ser representativa, significativa y confiable y que en base a ella se pueden hacer inferencias respecto a la población. La muestra puede ser probabilística y no probabilística.

MUESTRA NO PROBABILÍSTICA.- Es aquella que se obtiene mediante juicio de la persona que selecciona los elementos de la muestra que usualmente es un experto en la materia. Este método está basado en los puntos de vista subjetivos de una persona y la teoría de la probabilidad no puede ser empleada para medir el error de muestreo. Las principales ventajas de una muestra de juicio son la facilidad de obtenerla y que el costo usualmente es bajo.

MUESTRA PROBABILÍSTICA.- Es aquella muestra obtenida por un mecanismo de probabilidades, en el cual cada elemento de la población total o universo tiene una probabilidad conocida de selección. (Ver muestreo aleatorio).

MUESTREO.- Es un conjunto de métodos y procedimientos estadísticos destinados a la selección de una o mas muestras es la técnica seguida para elegir muestras. El objetivo principal de un diseño de muestreo es proporcionar procedimientos para la selección de muestras que sean representativas de la población en estudio.

MUESTREO ALETORIO SIMPLE.- También llamado irrestrictamente aleatorio. Es un método de muestreo donde una muestra aleatoria simple es seleccionada de tal manera que cada muestra posible del mismo tamaño tiene igual probabilidad de ser seleccionada de la población. Una muestra

aleatoria es también llamada una **muestra probabilística** es aquella cuyos elementos se seleccionan individualmente de la población en forma aleatoria, y es preferida por los estadísticos porque la selección de las muestras es objetiva y el error muestral puede ser medido en términos de probabilidad bajo la curva normal. Por conveniencia, este método puede ser reemplazado por una tabla de números aleatorios cuando una población es infinita. Se aplica cuando los datos son casi homogéneos. Una variante del muestreo aleatorio simple es el muestreo aleatorio sistemático. Otros tipos más comunes de muestreo aleatorio son: muestreo aleatorio estratificado y muestreo por conglomerados.

MUESTREO SISTEMÁTICO.- Es una variante del método aleatorio simple de selección de cada elemento de la muestra. Se aplica cuando la población está listada en algún orden. Consiste en seleccionar un número aleatorio menor que N/n y luego los $(n-1)$ elementos de la muestra se eligen agregando al primer aleatorio: el entero K obtenido por $K=N/n$ y así sucesivamente. El primer elemento de la muestra es seleccionado al azar. Por lo tanto, una muestra sistemática puede dar la misma precisión de estimación acerca de la población que una muestra aleatoria simple cuando los elementos en la población están ordenados al azar.

MUESTREO ESTRATIFICADO ALEATORIO.- Es un método de muestreo que se aplica cuando se divide la población en grupos, llamados estratos, donde los datos son más homogéneos pero un estrato frente al otro muy distintos. Para extraer la muestra aleatoria se aplica el muestreo aleatorio simple a cada estrato y el tamaño es la suma de los tamaños de todos los estratos. Para determinar los tamaños de los estratos se puede utilizar la asignación proporcional, óptima y óptima económica. Si no se conoce la variabilidad de los datos se aplica la asignación proporcional.

MUESTREO POR CONGLOMERADOS.- Es un método de muestreo en el cual la población está en grupos debido a la organización administrativa u otro (conglomerados). Ejemplo: Colegios, Universidades, manzanas de casas, entre otros. Al interior de los conglomerados no se puede garantizar homogeneidad. Cada conglomerado es una unidad donde la muestra se selecciona como en el muestreo aleatorio simple y se

aplica la encuesta a todos los elementos del conglomerado. Una muestra de conglomerados, usualmente produce un mayor error muestral (por lo tanto, se obtiene menor precisión de las estimaciones acerca de la población) que una muestra aleatoria simple del mismo tamaño. Los elementos individuales dentro de cada "conglomerado" tienden frecuentemente a ser iguales.

MUESTREO CON REPOSICIÓN.- Es el método para obtener una muestra con reposición. Esta muestra consiste en que al seleccionar un segundo elemento, el primero debe haber sido devuelto a la población. De este modo un elemento puede repetirse en la muestra. Es decir, con este método una unidad en particular puede quedar incluida más de una vez en la muestra, pudiendo ser hasta "n veces". El universo o población se mantiene permanentemente con un tamaño N.

MUESTREO SIN REPOSICIÓN.- Es el procedimiento para seleccionar cada elemento de la población éste no se repone o considera de nuevo en la población, por lo que no puede ser seleccionado nuevamente. En este caso el tamaño de la población o universo se irá reduciendo en cada selección N-1, N-2, N-3,...,etc. unidades de muestreo hasta N - n elementos.

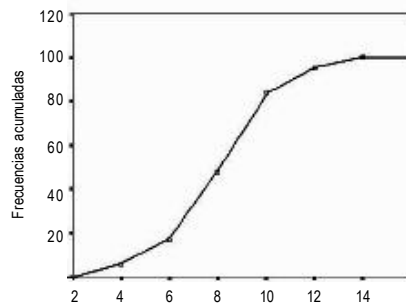
N

NIVEL DE SIGNIFICACIÓN.- Se define como la probabilidad de rechazar la hipótesis nula cuando ésta es verdadera. Se le conoce también con el nombre de «error de tipo 1», simbólicamente se denota por α .

NÚMERO ÍNDICE.- Es aquella medida estadística que permite estudiar los cambios que se producen en una magnitud simple o compleja con respecto al tiempo o al espacio; es decir, se va a comparar dos situaciones, una de las cuales se considera de referencia, llamado también período base. Los números índices pueden ser simples y complejos. Estos índices pueden ser de precios, cantidades y valor (Ver índice).

O

OJIVA.- Es un gráfico acumulativo de frecuencias o frecuencias relativas. Existen las ojivas mayor que y menor que.



Ojiva menor que

P

PARÁMETRO.- Es cualquier valor característico de la población. Ejemplo: la media de la población, la desviación típica de la población. Sin embargo estos valores son desconocidos porque no siempre podemos tener todos los datos de la población para calcularlos.

PERMUTACIONES.- Son las distintas disposiciones de los elementos en que se pueden ordenar los objetos. El número de permutaciones de n objetos se obtiene como el factorial de n !

$$\text{Permutaciones de } n \text{ objetos} = n!$$

Pero generalmente interesa conocer el número de subgrupos de r elementos que se puede tomar del total de n objetos, se obtiene con la siguiente fórmula:

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

Donde:

- n! Representa el factorial de n.
- r El número de elementos de cada subgrupo.

PERCENTIL.- Es el valor que resulta de dividir el conjunto de datos en 100 partes iguales. Cada parte representa al 1% del total, se pueden calcular los 99 percentiles mediante la fórmula:

$$Pr = Li \frac{(rN/100) - N_{i-1}}{n_i} * c$$

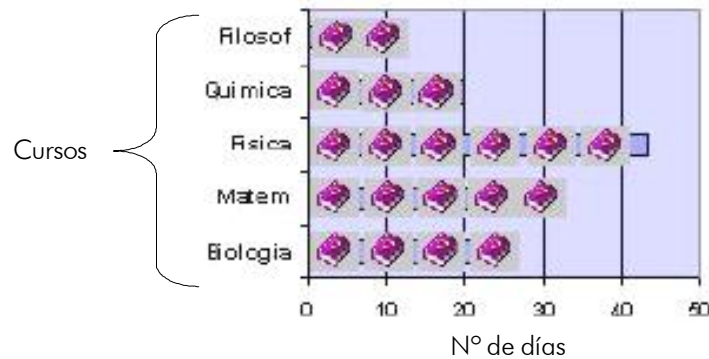
r = 1, 2, 3,99

Donde:

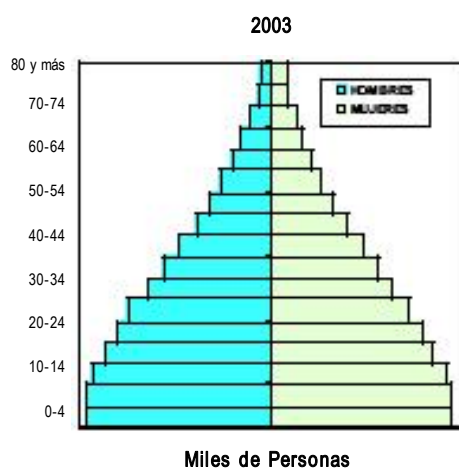
- r Es el número del percentil que se desea calcular
- Li Límite inferior de la clase percentilica
- N Total de datos
- n_i Frecuencia absoluta de la clase percentilica
- N_{i-1} Frecuencia absoluta acumulada anterior o igual a la clase percentilica
- c Amplitud o tamaño del intervalo

PERIODO DE REFERENCIA DE UNA ENCUESTA.- Es el lapso o espacio de tiempo durante el cual se levanta la información de la encuesta y la referencia cronológica respecto a la cual es válida la información inherente a ella.

PICTOGRAMAS.- Son gráficos vistosos, similares a los gráficos de barras, pero empleando un dibujo alusivo al tema que representa, en una determinada escala para expresar la unidad de medida de los datos.



PIRÁMIDE DE POBLACIÓN.- Es la representación gráfica de la estructura por sexo y edad de una población en un instante temporal determinado. La pirámide de población consta de dos histogramas horizontales: usualmente el del lado izquierdo representa la distribución por edad de los hombres y el derecho el de las mujeres. Cada barra horizontal representa la proporción de población de un determinado sexo y rango de edad. En el eje de abscisas se representa los efectivos de población, normalmente en porcentajes, y en el eje de ordenadas las edades.



POBLACIÓN FINITA.- Es aquella en la que es posible enumerar (contar) físicamente los elementos que pertenecen a la población.

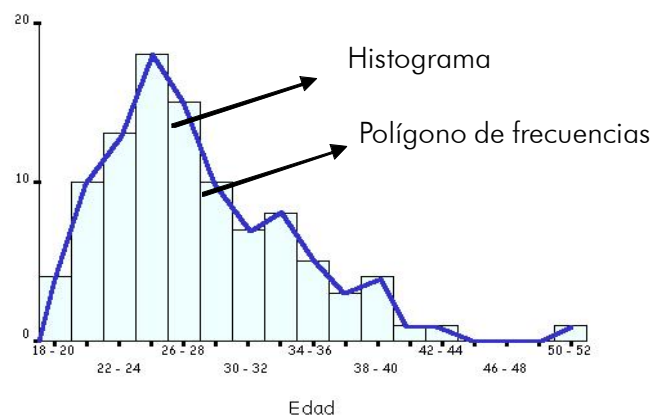
POBLACIÓN INFINITA.- Es aquella en la que no es posible enumerar (contar) físicamente los elementos que pertenecen a la población. Dicho de otra manera, cuando los elementos de la población son ilimitados.

POBLACIÓN O UNIVERSO.- Es cualquier conjunto de unidades o elementos claramente definido, en el espacio y el tiempo, donde los elementos pueden ser personas, granjas, hogares, manzanas, condados, escuelas, hospitales, empresas, y cualquier otro. Las poblaciones pueden ser finitas e infinitas.

POBLACIÓN OBJETIVO.- Es el conjunto de elementos de los que nos interesa obtener información o tomar decisiones.

POLÍGONO DE FRECUENCIAS ACUMULADAS.- Es un gráfico utilizado para representar una **distribución de frecuencias** acumuladas de una variable numérica. Se conoce también como ojiva. (Ver ojiva).

POLÍGONO DE FRECUENCIAS.- Es un gráfico utilizado para representar una distribución de frecuencias de una variable numérica, teniendo en cuenta su marca de clase.



PRECISIÓN.- La precisión de una variable es la inversa de su variabilidad, es decir: $\text{Precisión} = 1/\text{variabilidad}$.

PROBABILIDAD.- Es un número que se le asigna a un suceso como una medida de su incertidumbre. Este número puede tomar valores entre cero y uno inclusive. Cuando los sucesos son equiprobables, es decir todos tienen la misma probabilidad para calcularla, se utiliza la **Regla de Laplace**. Sea un suceso A, entonces:

$$P(A) = \text{Casos favorables} / \text{casos posibles}$$

El experimento aleatorio debe cumplir **dos requisitos**:

- El número de resultados posibles (sucesos) debe ser finito.
- Todos los sucesos deben tener la misma probabilidad.

A la regla de Laplace también se le denomina "**probabilidad a priori**",

ya que para aplicarla hay que conocer antes de realizar el experimento cuáles son los posibles resultados y saber que todos tienen las mismas probabilidades.

PROBABILIDAD DE SELECCIÓN.- Es la oportunidad que tiene cada elemento de la población o universo de ser incluida en una muestra.

PROBABILIDAD PROPORCIONAL AL TAMAÑO.- Es un método de selección de muestreo en el que las unidades se eligen con probabilidad de selección desigual, siendo la probabilidad para cada unidad proporcional a una medida de tamaño. La medida de tamaño para cada unidad es un número asignado antes de la selección de esa unidad, que se supone altamente correlacionada con el estadígrafo a estimar. Usualmente la probabilidad proporcional al tamaño se abrevia como PPT.

PROMEDIO.- Es cualquier medida de posición de tendencia central. Cuando se obtiene sumando los datos y dividiendo entre el número de ellos, se conoce como promedio simple.

PROMEDIO PONDERADO.- Es un número conocido también como media aritmética ponderada. Es el promedio de datos a los que se les asigna distinta importancia llamada ponderación.

PRUEBA DE HIPÓTESIS.- Es una técnica que permite rechazar o aceptar la hipótesis en base de la información proporcionada por la muestra. (Ver contraste de hipótesis).

PRUEBA JI-CUADRADO.- Es una prueba que permite contrastar si la hipótesis H_0 es coherente con los datos obtenidos en la muestra. Se le denota χ^2 . Puede utilizarse para:

1. Bondad de un ajuste.
2. Criterio de independencia.
3. Criterio de homogeneidad.

Una forma de comparar las O_x con las e_x es calculando el valor de χ^2

$$\chi^2 = \sum_x \frac{(o_x - e_x)^2}{e_x}$$

Donde:

O_x Es el valor observado

e_x Es el valor esperado

PUNTO MUESTRAL.- El conjunto de todos los resultados posibles de un experimento aleatorio se le denomina espacio muestral. Un punto de este conjunto es un punto muestral.

Q

QUINTIL.- Es un fractil se obtienen dividiendo al conjunto de datos en cinco partes iguales cada parte representa el 20% del total. Se pueden calcular 4 quintiles.

R

RANGO.- Conocido también como recorrido, es un número que mide la amplitud de los valores de un conjunto de datos y se calcula por diferencia entre el valor mayor y el valor menor. Lo notaremos como **R**. No constituye una medida muy significativa en la mayoría de los casos, pero es muy fácil de calcular.

$$R = X_{mayor} - X_{menor}$$

RAZÓN.- Es la relación entre dos categorías o partes. Señala el tamaño de una parte con respecto a otra que se toma como unidad.

RECORRIDO INTERCUARTÍLICO.- Es una medida de dispersión. Su valor se obtiene como la diferencia del tercer cuartil (Q_3) menos el primer cuartil (Q_1), definido por la expresión: $R1 = Q_3 - Q_1$

REDONDEO.- Es el procedimiento para expresar un número de acuerdo a una precisión establecida.

REGIÓN DE ACEPTACIÓN.- Es la región formada por el conjunto de valores con los cuales decidimos aceptar la hipótesis nula.

REGIÓN DE RECHAZO.- Conocida también como región crítica, está formada por el conjunto de valores con los cuales se rechaza la hipótesis nula.

REGRESIÓN.- Es una técnica de análisis para poner de manifiesto la estructura de dependencia que mejor explique el comportamiento de la variable dependiente o explicada (**y**) a través de un conjunto de variables independientes o explicativas (x_1, x_2, \dots, x_p), con las que se supone está relacionada.

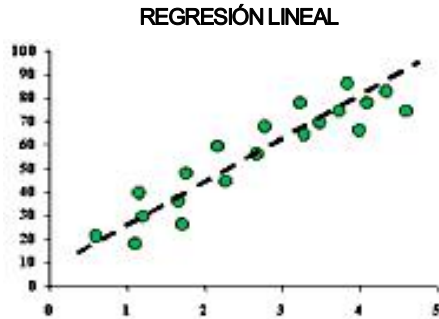
El método más utilizado es el de los mínimos cuadrados. La ecuación a ajustar puede ser lineal o no lineal. En ambos casos el objetivo es el mismo: encontrar las mejores estimaciones de los parámetros y cuantificar la precisión de los mismos.

REGRESIÓN LINEAL.- La regresión será lineal cuando la curva obtenida o seleccionada sea una recta. Es la recta que mejor se ajusta a los datos. Se obtiene mediante el método de mínimos cuadrados. Para ello se debe calcular primero el coeficiente de correlación lineal que permite determinar, si efectivamente, existe relación entre las dos variables. Una vez encontrada la relación, la regresión permite definir la recta que mejor se ajusta a la nube de puntos (gráfico de pares ordenados).

Una recta viene definida por la siguiente fórmula:

$$Y = a + bX$$

Donde "Y" sería la variable dependiente, es decir, aquella que viene definida a partir de la otra variable "X" (variable independiente). Para definir la recta hay que determinar los valores de los parámetros "a" y "b":



El **parámetro "b"** determina la pendiente de la recta, es decir su grado de inclinación.

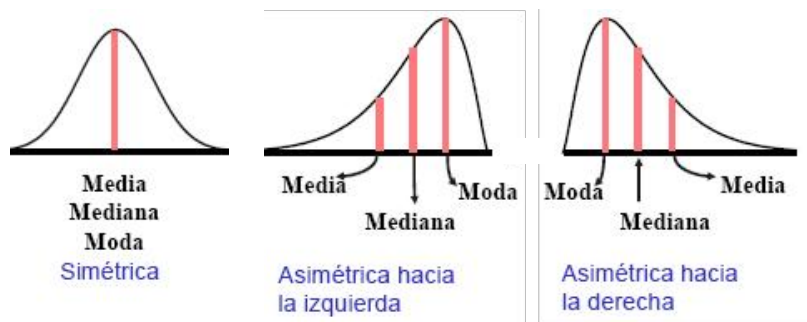
$$\hat{b} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}$$

El **parámetro "a"** es el valor que toma la variable dependiente "Y", cuando la variable independiente "X" vale 0, y es el punto donde la recta cruza el eje vertical.

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

RELACIÓN ENTRE LA MEDIA (\bar{x}), MEDIANA (\tilde{x}) Y MODA (\hat{x}). Al comparar los valores de la media mediana y moda puede ocurrir:

- 1.- $\bar{X} = \tilde{X} = \hat{X} \implies$ La curva es simétrica
- 2.- $\bar{X} \neq \tilde{X} \neq \hat{X} \implies$ La curva es asimétrica
 - a) $\bar{X} < \tilde{X} < \hat{X} \implies$ asimetría hacia la izquierda o negativa
 - b) $\hat{X} < \tilde{X} < \bar{X} \implies$ asimetría hacia la derecha o positiva



S

SERIES TEMPORALES.- Conocida también como serie cronológica. Es la sucesión de observaciones cuantitativas ordenadas en el tiempo de un fenómeno. Los datos tienen un orden que no es posible variar. La información puede ser mensual, trimestral, anual o de cualquier otro intervalo temporal.

SESGO.- Se denomina así a la asimetría que presenta una distribución de frecuencias. Puede ser sesgo negativo o a la izquierda y sesgo positivo o a la derecha.

SUCESO CIERTO.- Se conoce también como suceso seguro. Es aquel suceso que siempre se realiza. Estará formado por todos los resultados posibles del experimento; es decir, coincide con el espacio muestral.

SUCESO IMPOSIBLE.- Es aquel suceso que no se realiza nunca. Se designa por un \emptyset .

SUCESOS COMPUESTOS.- Se llama sucesos compuestos, a los sucesos formados por dos o más puntos muestrales; es decir, por más de un resultado del experimento.

SUCESOS CONTRARIOS.- Dado un suceso cualquiera A del espacio de sucesos S, se llama suceso contrario del suceso A a un suceso que se realiza cuando no se realiza A, y recíprocamente.

Por tanto los sucesos A y \bar{A} son contrarios, ya que si se realiza A no se realiza \bar{A} y si se realiza \bar{A} no se realiza A .

SUCESOS ELEMENTALES.- Conocido también como sucesos aleatorios, son los sucesos formados por un solo punto muestral; es decir, por un solo resultado del experimento aleatorio.

T

TABLA DE CONTINGENCIA.- Es una tabla de doble entrada. Se representa genéricamente como $(x_i; y_j; n_{ij})$. (Ver distribución bidimensional).

TABLA DE DATOS AGRUPADOS.- Es un arreglo matricial que contiene el número de veces (frecuencia) que aparece un dato, de acuerdo a las clases de interés especificadas (puede ser intervalos). (Ver distribución unidimensional)

TASA.- Es la relación del número de casos, frecuencias o eventos de una categoría entre el número total de observaciones, multiplicada por un múltiplo de 10, generalmente 100 ó 1000.

La fórmula es:

$$\text{Tasa} = \frac{\text{Número de eventos durante un período } t}{\text{Número total observaciones en el período } t} \times 1000$$

TASA BRUTA DE MORTALIDAD.- Es un número que expresa la frecuencia de muertes en una población en un período de tiempo, por lo general un año calendario.

La tasa bruta de mortalidad se calcula dividiendo el número de defunciones ocurridas en un período de tiempo determinado entre la población donde ocurren estas defunciones, generalmente expresadas por 1000 habitantes. Se define como:

$$m^z = \frac{D^z}{N^{30-VI-Z}} \times 1,000$$

donde:

- m^z representa la tasa bruta de mortalidad para el año "z".
 D^z representa las defunciones ocurridas durante el año "z".
 $N^{30-VI-Z}$ representa la población al 30 de junio del año "z", o población media.

TASA DE ANALFABETISMO.- Es un valor que se obtiene como la relación del número de analfabetos (personas que no saben leer ni escribir) de 15 y más años de edad en el año «z» dividido entre la población total de 15 y más años de edad en el año "z". Se calcula como sigue:

$$TA^z = \frac{NA_{(15,y+)}^z}{N_{(15,y+)}^z} \times 100$$

Donde:

- TA^z representa la tasa de analfabetismo en el año "z".
 $NA_{(15,y+)}^z$ representa la población de analfabetos de 15 y más años de edad en el año "z".
 $N_{(15,y+)}^z$ representa la población total de 15 y más años de edad en el año "z".

TASA DE ESCOLARIZACIÓN POR EDAD.- Es un número que se obtiene como la relación del número de matriculados de la edad "x" en el año "z" entre la población total de la edad "x" en el año "z". Se calcula del modo siguiente:

$$TE_x^z = \frac{M_x^z}{N_x^z} \times 100$$

Donde:

- TE_x^z representa la tasa de escolarización por edad "x" en el año "z".
 M_x^z representa el número de matriculados de edad "x" en el año "z".
 N_x^z representa la población total de edad "x" en el año "z".

TASA DE INFLACIÓN.- Es un indicador del aumento en los precios de los bienes y servicios, referidos a un periodo de tiempo. Más utilizado para medir la inflación es el índice de precios al consumidor

$$T \text{ de } I = \frac{\text{IPC año actual} - \text{IPC año base}}{\text{IPC año base}} \times 100$$

TASA DE LETALIDAD.- Es un número que se define como la proporción de personas que mueren por causa de una enfermedad determinada entre el total de quienes contrajeron la enfermedad.

$${}^c t^z = \frac{{}^c D^z}{{}^c E^z} \times 1000$$

Donde:

${}^c t^z$ representa la tasa de letalidad del período "z" debido a la causa "c".

${}^c D^z$ representa las defunciones del período "z" debido a la causa "c".

${}^c E^z$ representa las personas que contrajeron la enfermedad "c" en el período "z".

TASA DE MASCULINIDAD.- Es un número que expresa la proporción de varones en la población total o en una parte de ella. Se usa para estudiar la distribución por sexo en la población. Se expresa:

$$TM(X) = \frac{NH(X)}{NH(X) + NF(X)} \times k$$

Donde:

TM(X) representa la tasa de masculinidad de la población de edad X.

NH(X) representa el número total de varones de edad X.

NF(X) representa el número total de mujeres de edad X.

k representa una constante, generalmente 100.

TASA DE MORTALIDAD INFANTIL.- Es un número que expresa la mortalidad de niños menores de un año y se obtiene dividiendo las defunciones infantiles (menores de un año) ocurridas en un año calendario entre el número de nacidos vivos ocurridos en el transcurso del mismo año, multiplicado por mil.

$$TMI^z = \frac{D^z}{B^z} \times 100$$

Donde:

TMI^z representa la tasa de mortalidad infantil en el año "z"

D^z representa las defunciones de menores de un año ocurridas en el año "z"

B^z representa el número de nacidos vivos del año "z".

TASA DE MORTALIDAD MATERNA.- Es un valor que representa las defunciones de las mujeres durante el embarazo o dentro de los 42 días de su término (embarazo, parto, puerperio).

La tasa de mortalidad materna se obtiene dividiendo el número de muertes maternas ocurridas en un año, entre el número promedio de mujeres en edad fértil para ese año, multiplicado por 100,000.

$$TMM^z = \frac{MM^z}{MEF^{30-VI-Z}} \times 100,000$$

Donde:

TMM^z representa la tasa de mortalidad materna del año "z".

MM^z representa las muertes por causa materna ocurridas en el año "z".

$MEF^{30-VI-Z}$ representa el número promedio de mujeres en edad fértil en el año "z".

TASA DE MORTALIDAD POR CAUSAS.- Es un número que representa la mortalidad por causas y se calcula dividiendo el número de defunciones debidas a cierta causa o grupo de causas entre la población total, multiplicado por 100,000.

$${}^c m^z = \frac{{}^c D^z}{N^{30-VI-Z}} \times 100,000$$

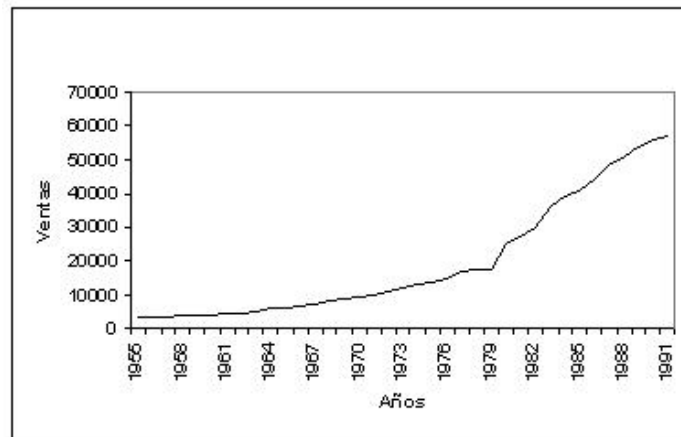
Donde:

${}^c m^z$ representa la tasa de mortalidad por la causa "c" en el año "z".

${}^c D^z$ representa el número de muertes debidas a la causa "c" en el año "z".

$N^{30-VI-Z}$ representa la población al 30 de junio del año "z", o población media.

TENDENCIA.- Es un componente del análisis clásico de series temporales. Refleja el movimiento de la serie en el largo plazo (crecimiento, decrecimiento o estancamiento). Es necesario un número suficientemente grande de observaciones para determinar una tendencia.



U

UNIDAD DE MUESTREO.- Es la unidad estadística que se selecciona para constituir la muestra. La elección de la unidad de muestreo más eficiente es una consideración importante en el diseño de una muestra.

UNIDAD ESTADÍSTICA.- Conocido también como unidad elemental. Es el elemento o unidad base de la población o de la muestra que permite obtener información o datos referidos a ciertas características o variables, que nos interesan para explicar un determinado fenómeno.

V

VARIABLE.- Es una característica de la población o de la muestra cuya medida puede cambiar de valor. Se representa simbólicamente mediante las letras del alfabeto. Según su naturaleza puede ser cualitativa y cuantitativa.

VARIABLE ALEATORIA.- Conocida también como variable estocástica o probabilística. Es la característica considerada en un experimento aleatorio cuyo valor de ocurrencia sólo puede saberse con exactitud una vez observado.

VARIABLE BIDIMENSIONAL.- Es aquella que proporciona información sobre dos características de la población (por ejemplo: edad y altura de los alumnos de una clase). (Ver distribución bidimensional).

VARIABLE CONTINUA.- Es una variable cuantitativa. Es la característica de la población, cuyos valores están representados mediante el conjunto de los números reales. Puede tomar cualquier valor real dentro de un intervalo. Por ejemplo, la velocidad de un vehículo puede ser 80,3 km/h, 94,57 km/h.

VARIABLE CUALITATIVA.- Es aquella que representa cualidades, atributos o características no numéricas y estas pueden ser nominales y ordinales. (Ver dato cualitativo).

VARIABLE CUANTITATIVA.- Es aquella característica de la población o de la muestra que es posible representar numéricamente. Éstas pueden ser continua y discreta. (Ver dato cuantitativo).

VARIABLE DETERMINÍSTICA.- Es aquella cuyo valor puede ser predicho con exactitud.

VARIABLE DISCRETA.- Es una variable cuantitativa. Es la característica de la población, cuyos valores están representados mediante el conjunto de los números naturales. Por ejemplo, el número de alumnos de un aula.

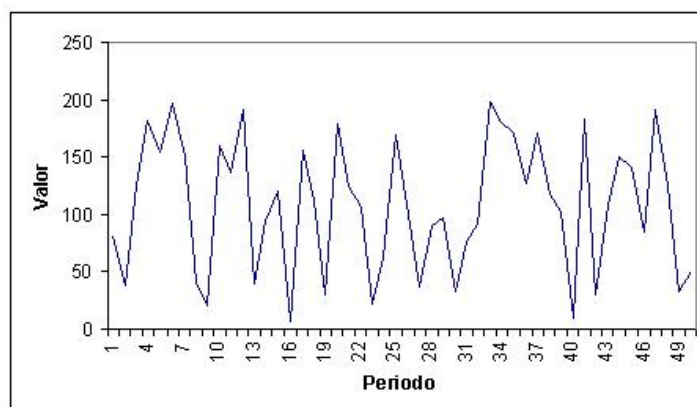
VARIABLE NOMINAL.- Es una variable cualitativa la cual sólo permite asignar nombres a los datos y no implica ningún orden. Ej. el idioma de los habitantes de la tierra.

VARIABLE ORDINAL.- Es una variable cualitativa cuyos valores solamente pueden ser ordenados con algún criterio.

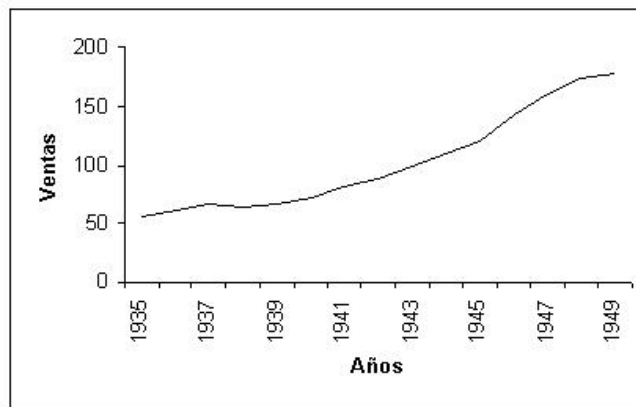
VARIABLE PLURIDIMENSIONAL.- Es aquella que proporciona información sobre tres o más características (por ejemplo: edad, altura y peso de los alumnos de una clase).

VARIABLE UNIDIMENSIONAL.- Es aquella que proporciona información sobre una sola característica (por ejemplo: edad de los alumnos de una clase). (Ver distribución unidimensional).

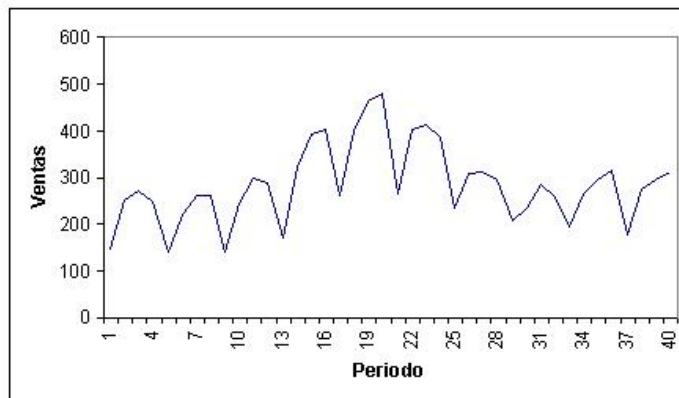
VARIACIONES IRREGULARES.- Conocido como variación de aleatoriedad. Es un comportamiento irregular que está compuesto por fluctuaciones causadas por sucesos impredecibles o no periódicos, movimientos de muy corto plazo, sin un carácter periódico reconocible, ocasionados por fenómenos singulares o fortuitos que producen efectos casuales y no permanentes como el clima poco usual, huelgas, guerras, rumores, elecciones, cambio de leyes, entre otros.



VARIACIONES O FLUCTUACIONES CÍCLICAS.- Se conoce también como ciclos o ciclicidad. Es un conjunto de fluctuaciones en forma de onda o ciclos, de más de un año de duración, producidas por cambios en las condiciones económicas. Refleja movimientos de la serie a medio plazo producidos con un período superior al año, debido a alternancias de prosperidad y de depresión en la actividad económica. Se suelen superponer distintos ciclos, siendo muy difíciles de aislar.



VARIACIONES O FLUCTUACIONES ESTACIONALES.- Son oscilaciones a corto plazo producidas en un período inferior al año (mes, trimestre) y que se repiten de forma reconocible dentro de cada periodo de 12 meses, año tras año. Se deben a factores climatológicos, biológicos, institucionales, culturales, de tradición y otros.



VARIANZA.- Conocida también como variancia, es una medida de dispersión de la información. Se obtiene como el promedio de los cuadrados de las desviaciones de los valores de la variable respecto de su media aritmética.

Fórmula para datos simples.

$$S^2 = \frac{\sum (x_i - \bar{X})^2}{N}$$

Fórmula para datos agrupados

$$S^2 = \frac{\sum (x_i - \bar{X})^2 * n_i}{N}$$

Mide la distancia existente entre los valores de la serie y la media. La varianza siempre será mayor que cero. Mientras más se aproxima a cero, más concentrados están los valores de la serie alrededor de la media. Por el contrario, mientras mayor sea la varianza, más dispersos están los datos. Este estadístico tiene el inconveniente de ser poco significativo, pues se mide en el cuadrado de la unidad de la variable. Por ejemplo, si la variable viene dada en cm, la varianza vendrá en cm².

BIBLIOGRAFIA

BERENSON, Mark. ESTADÍSTICA BASICA EN ADMINISTRACIÓN. (1992). New York: Prentice Hall.

MAZA, Domingo. TRATADO MODERNO DE ECONOMÍA. (1992). Caracas: Panapo.

MURRAY, Spiguel. PROBABILIDADES Y ESTADÍSTICA. (1997). Madrid: Mc Graw Hill.

RÍOS, Sixto. ANÁLISIS ESTADÍSTICO APLICADO. (1972). Madrid: Paraninfo.

SALINAS O., José. ANÁLISIS ESTADÍSTICO PARA LA TOMA DE DECISIONES EN ADMINISTRACIÓN Y ECONOMÍA. 1998. Universidad del Pacífico. Lima-Perú. Números Índices . Pág. 361-376.

SERRANO RODRÍGUEZ, Javier. INTRODUCCIÓN A LA ESTADÍSTICA. Ed universitaria de América LIDA, Bogotá, Colombia. Pág. 30-49

SIERRA BRAVO. R. DICCIONARIO PRÁCTICO DE ESTADÍSTICA, Ed Paraninfo S.A. Madrid. España, Pág. 56-57, 177-187, 427-432.

YA-LUN, Chou. ANÁLISIS ESTADÍSTICO. (1980). Tokio: Mc Graw Hill.

PAGINAS WEB

<http://www.bioestadistica.uma.es/libro/> Universidad de Málaga

<http://www.estadistico.com/dic.html>

<http://www.fvet.edu.uy/estadis/diagrth>

<http://www.fvet.edu.uy/estadis/glosario.htm>

http://www.uhu.es/89009/ficheros_datos/ Universidad de Huelva de Andalucía

Doctor
ALEJANDRO TOLEDO MANRIQUE
Presidente Constitucional de la República

**PRESIDENCIA DEL
CONSEJO DE MINISTROS**

Doctor
PEDRO PABLO KUCZYNSKI
Presidente

**INSTITUTO NACIONAL DE
ESTADISTICA E INFORMATICA**

Señor
FARID MATUK
Jefe

Señor
FRANCISCO COSTA APONTE
Sub-Jefe de Estadística

Señora
LUPE BERROCAL DE MONTESTRUQUE
Directora Técnica del Centro de Investigación
y Desarrollo

LEY DE ORGANIZACION Y FUNCIONES DEL INSTITUTO NACIONAL DE ESTADISTICA E INFORMATICA

DECRETO LEGISLATIVO N° 604

- Artículo 1° Los Sistemas Nacionales de Estadística e Informática tienen por finalidad asegurar, en los respectivos campos, que sus actividades se desarrollen en forma integrada, coordinada y racionalizada y bajo una normatividad técnica común, contando para ello con autonomía técnica y gestión.
- Artículo 2° Son objetivos de los Sistemas Nacionales de Estadística e Informática:
- a. Normar las actividades de estadística e informática oficial.
 - b. Coordinar, integrar y racionalizar las actividades de Estadísticas e Informática; y
 - c. Promover la capacitación, investigación y desarrollo de las actividades de Estadística e Informática.
- Artículo 3° Los ámbitos de competencia de los Sistemas Nacionales de Estadística e Informática son:
- a. Del Sistema Nacional de Estadística
Los levantamientos censales, estadísticas continuas, las encuestas por muestreo, las estadísticas de población, los indicadores e índices en general, las cuentas nacionales y regionales, los esquemas macroestadísticos, análisis e investigación. Corresponde a éste las tareas técnicas y científicas que se desarrollan con fines de cuantificar y proyectar los hechos económicos y sociales para producir las estadísticas oficiales del país.