



MarkQual[®]

Tecnología aplicada a la educación

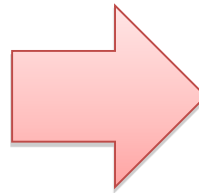
GLOSARIO ESTADÍSTICO

Fuente: Murray R. Spiegel, *Estadística*, McGraw-Hill.

CONCEPTOS Y DEFINICIONES ESPECIALES

Es el estudio científico de los métodos para recoger, organizar, resumir y analizar los datos de una información, así como para obtener conclusiones válidas con rigor y tomar decisiones razonables apoyadas en tal análisis.

Con esta herramienta es posible trabajar en forma objetiva, permite vincular causas y efectos, presentar resultados con claridad y orden. Es un insumo esencial para interpretar información numérica y luego construir diseños curriculares.



La estadística posee tres campos bien definidos:

Descriptiva

Inferencial

Teoría de Probabilidades.

Estadística Descriptiva

Trata de los estudios que se hacen sobre el total de individuos de una población con el fin de establecer las principales características de interés para el investigador.

Estadística Inferencial

Se refiere a los estudios que se hacen sobre una parte de la población (muestra), con el fin de obtener (inferir) conclusiones sobre las características de interés de toda la población. Es un camino de deducción con riesgo, con probabilidad de error.

Teoría de Probabilidades

Es una rama de la matemática de gran importancia en los estudios inferenciales, dado que los valores que se obtienen sobre el análisis de una muestra no son exactamente iguales a los correspondientes parámetros de toda la población. Estudia el comportamiento matemático del azar con un control de los fenómenos aleatorios.

Población

Grupo completo de individuos u objetos que constituyen la base de interés para un estudio estadístico. Es el conjunto de todos los elementos que cumplen una determinada característica que deseamos medir y estudiar.

Muestra

Parte representativa de una población. Es todo subconjunto de una población sobre el que se va a realizar el estudio. El número de elementos de la muestra se denomina **tamaño** de la muestra.

Individuo

En estadística se considera individuo (objeto) a cada uno de los elementos de la población.

Carácter

Cada uno de los aspectos o propiedades que se pueden estudiar en los individuos de una población recibe el nombre de **carácter o estadístico**. Esto permite clasificar a los individuos.

El carácter puede ser **cuantitativo** si se puede medir o bien **cualitativo** si no se puede medir pero se puede comparar.

Dato

Valor o forma que asume una variable para un individuo determinado.

Estimadores

Características cuantificables que posee una muestra, y que se utilizan para calcular los parámetros de la población.

Distribuciones

Formas de organización y representación tabular de los datos.

Experimento Estadístico

Cualquier proceso que genera un conjunto de datos numéricos.

Espacio Muestral

Conjunto de todos los posibles resultados de un experimento estadístico.

Variable

El conjunto de valores que puede tomar un carácter estadístico se llama variable estadística. Son atributos que poseen o se le pueden asignar a los individuos de una población y que difieren de uno a otro.

Clasificación de Variables

Cualitativas: Las que definen cualidades de los individuos; usualmente pueden subdividirse en categorías.

Ejemplo: Variable: Sexo. Categorías: M. F.

Indicadoras: Valores numéricos que se le asignan a las categorías de una Variable Cualitativa.

Cuantitativas: Cuando los atributos que las definen son cuantificables o medibles numéricamente. Las Variables Cuantitativas pueden ser Discretas o Continuas.

Discretas: Cuando las variables sólo pueden tomar determinados valores, (asumen valores de uno en uno); es decir pueden tomar un número finito o bien infinito numerable de valores.

Continuas: Cuando pueden asumir cualquier valor entre dos enteros consecutivos, es decir pueden tomar todos los valores de un intervalo y tan próximos como se quiera.

Intervalos de Clase

Se le denomina **intervalo de clase** a cada uno de los intervalos en que pueden agruparse los datos de una variable estadística. Permiten obtener una idea más clara y concreta de la realidad. Al agrupar los valores de una variable estadística y clasificarla por intervalos, la variable pasa a ser considerada continua.

Existen situaciones como las siguientes que pueden ocurrir:

- Hay demasiados datos, para una sola variable.
- Pueden ser pocos datos, pero sus valores muy dispersos.
- Interesa una clasificación particular de los resultados.

En estos casos el agrupamiento de los datos resulta una buena técnica para el análisis de las variables dentro de un estudio estadístico.

Intervalos de Clase - Método de Trabajo I

1. Aplicar una técnica de recolección de datos, por ejemplo: Tronco y Hoja

2. Determinar el **Rango** de la información: $R = D_M - D_m$.

Donde $D_M = \text{DatoMayor}$ y $D_m = \text{DatoMenor}$

3. Determinar la **Amplitud de los intervalos**: Se divide el rango obtenido en dos por el **Número de intervalos** en los que deseo agrupar $A = \frac{R}{I}$

Donde: A : Amplitud; R : Rango e I : Cantidad de intervalos en los que deseo agrupar

4. Si la amplitud no me da un número entero, se puede realizar el siguiente ajuste:

4.1. Ajusto la amplitud obtenida al entero siguiente.

4.2. Con la amplitud anterior y el número de intervalos (I), hallo un nuevo rango (NR).

4.3. Establezco la diferencia: $NR - VR$; donde VR es el viejo rango. +

4.4. Ajusto los datos originales de acuerdo con la diferencia anterior.

Intervalos de Clase - Método de Trabajo II

5. Establezco los intervalos, partiendo del primer dato (o del primer dato ajustado), y sumando la amplitud hasta cubrir el número de intervalos previamente definidos

6. Se calcula la **Marca de Clase** para cada intervalo $MC_i = \frac{L_{\text{inf}} + L_{\text{sup}}}{2}$

donde MC_i = Marca de Clase del intervalo i ; L_{inf} = Límite inferior del intervalo i ;

L_{sup} = Límite superior del intervalo i .

La **Marca de Clase** es un valor que representa a todos los del intervalo o clase. Es el punto medio entre los extremos de cada intervalo.

7. Se realiza la distribución de frecuencias

Observaciones:

El número de intervalos, puede definirse previamente a criterio de los investigadores, o aplicando algunas técnicas sugeridas para ello de acuerdo con el tipo de estudio; una fórmula empleada con frecuencia es la de **Sturges: $I = 1 + (3.3 \log N)$** , donde **N = Total de datos**. En cualquier caso se recomienda que el número de intervalos no sea inferior a 5, ni superior a 20.

Siempre que se realiza este agrupamiento hay una pérdida de información, se tiene en cuenta la pertenencia o no de cada dato al intervalo pero no su valor exacto.

Asimismo se produce un error en el cálculo posterior de los parámetros estadísticos. Los valores que pertenecen al intervalo se ven representados por su marca de clase, y ellos pueden ser mayores o menores que ésta

Definiciones

Frecuencia Absoluta (f_i)

Es la cantidad de veces que aparece y se repite el valor dato.

Frecuencia Absoluta Acumulada (F_A)

Se denomina frecuencia absoluta acumulada de un valor a la suma de todas las frecuencias absolutas de los valores menores o iguales hasta el valor considerado. $F_A = \sum f_i$

Distribución de frecuencias

Representación tabular de los datos correspondientes a una variable, que incluye:

Frecuencia Relativa (f_r)

Se denomina frecuencia relativa de un valor al cociente entre su frecuencia absoluta y el número total de datos que intervienen en el experimento .

Frecuencia Relativa Acumulada (F_R)

Se denomina frecuencia relativa acumulada de un valor a la suma de todas las frecuencias relativas de los valores menores o iguales al considerado. También puede calcularse como: $F_R = \frac{F_A}{N}$

Presentación de datos: Tablas, Gráficos y Diagramas

Existen muy diversos tipos y estilos de herramientas utilizadas para representar los datos correspondientes a una variable estadística. Es fundamental que sean claros, fácil de entender e interpretar, deben ajustarse estrictamente a la realidad que representan.

Hay dos tipos de gráficas para representar Distribuciones Agrupadas por intervalos:

Histograma: Gráfico de barras verticales pegadas con igual amplitud y centradas en la marca de clase. Asocian a cada intervalo un rectángulo de superficie proporcional a la frecuencia correspondiente a dicho intervalo. Las alturas de esos rectángulos son los cocientes entre las frecuencias absolutas y las longitudes de los intervalos que le corresponden. A intervalos de clase iguales, las alturas son directamente proporcionales a las frecuencias.

Polígono de frecuencias: Gráfico de trazos o líneas, cerrado, que se construye uniendo los puntos extremos medios superiores de un histograma, es decir los puntos correspondientes a las frecuencias de cada valor.



Gráfico de Trazos o Líneas

Son utilizados para una o varias variables cuantitativas discretas.

Gráfico de barras (Verticales y horizontales)

Es un gráfico que asocia a cada valor de la variable una barra (vertical u horizontal), proporcional a la frecuencia que le corresponde. Es apropiado para variables cualitativas solas o en comparación.

Dentro de la representación en barras, se encuentran las **Barras en Porcentaje por componente**, muy útiles cuando se trata de poblaciones con tamaños muy diferentes.

Sectores Circulares o de Pastel

Para una sola variable cualitativa o cuantitativa. Son útiles para representar las distintas partes de un todo, los diversos componentes de un carácter. Cada suceso se representa por un sector circular de una amplitud proporcional a su frecuencia

Pictogramas

Representaciones usualmente de barras, que se apoyan en las facilidades que brindan los gráficos por computador. Junto al gráfico de los datos se puede ver una imagen relativa o referida a los objetos que se miden.

Cartogramas

Son representaciones de los datos sobre un mapa.

Parámetros Estadísticos

Características numéricas cuantificables que posee una población. Se obtienen mediante un proceso de cálculo a partir de datos medibles. Son números que describen el comportamiento y las características generales de un conjunto de datos estadísticos. Se los agrupa en dos categorías: **centralización y dispersión**.

Medidas de Tendencia Central

Valores numéricos que se pueden obtener de la distribución de una variable cuantitativa, y cuyos resultados se ubican por el centro de la misma distribución; ellas son:

La Moda (M_o), la Mediana (M_{ed}), y la Media (\bar{X}).

Moda (M_o)

Es el dato que se presenta con mayor frecuencia absoluta dentro de una distribución. Si en una distribución aparecen dos Modas, se habla de una Distribución Bimodal; si son más de dos modas se dirá una Distribución Multimodal. En el caso de una variable continua este valor carece de sentido.

Mediana (M_{ed})

Se denomina mediana al valor central de los datos cuando éstos se han organizado ordenadamente de menor a mayor. Es un valor que divide a la distribución en dos partes iguales, cada una de las cuales contiene el 50% de los datos por debajo y el otro 50% por encima.

Media (\bar{X})

Se define como la suma de todos los valores (datos) que asume una variable, dividida por el número total de datos.

Fórmulas de cálculo de las Medidas de Tendencia Central

1. Para distribuciones no agrupadas

* **Moda:** Basta con observar la distribución, e identificar el dato que se presenta con la mayor frecuencia absoluta.

* **Mediana:** Dependiendo si el total de datos es impar o par:

* Si el total de datos es impar, la mediana será aquel dato que ocupe el lugar $\frac{n+1}{2}$

* Si el total de datos es par, la mediana será el promedio de los datos que ocupen las posiciones $\left(\frac{n}{2}\right)$ y $\left(\frac{n}{2}+1\right)$, es decir los valores que ocupan las posiciones centrales.

* **Media:** La definición conduce a la siguiente fórmula, que es la más utilizada en el cálculo de la media:

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n} ; \text{ si los datos están agrupados se puede calcular como: } \bar{X} = \frac{\sum_{i=1}^n x_i f_i}{n}$$

donde x_i : Dato i ésimo; f : frecuencia absoluta del dato i ésimo; n : total de datos de la información.

2. Para distribuciones agrupadas (Con intervalos de igual amplitud)

Se emplean las siguientes fórmulas:

MODA $M_o = L_{\text{inf}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) A$

MEDIANA $M_{ed} = L_{\text{inf}} + \left(\frac{\frac{n}{2} - \sum f_{\text{ant}}}{f_{\text{abs}}} \right) A$

MEDIA

*Método largo $\bar{X} = \frac{\sum M_{ci} f_i}{n}$

*Método corto $\bar{X} = M_{ca} + \left(\frac{\sum \mu \cdot f_i}{n} \right) A$

L_{inf} :Límite inferior del intervalo de clase

A :Amplitud de los intervalos

M_{ci} :Marca de clase del intervalo i.

Δ_1 :Diferencia entre la frecuencia absoluta del intervalo modal y la frecuencia absoluta del intervalo anterior

Δ_2 :Diferencia entre la frecuencia absoluta del intervalo modal y la frecuencia absoluta del intervalo siguiente

M_{ca} :Marca de clase arbitraria

μ :Número de veces la amplitud de los intervalos, a partir de la marca de clase arbitraria.

Medidas de dispersión

Son valores numéricos que nos dan información sobre los esparcidos o aglutinados, donde se encuentran los datos correspondientes a una variable cuantitativa dentro de un estudio estadístico. Permiten obtener una imagen del alejamiento de los datos respecto de las medidas de centralización. Pueden existir series de datos con iguales medidas de centralización pero con aspecto muy diverso, según sea la dispersión de sus datos.

Las medidas de dispersión más usadas son:

*Los Rangos:

Común o amplitud: Se define como la diferencia entre los datos extremos de una variable cuantitativa, así: Dato Mayor-Dato Menor

Intercuartil: Q3-Q1; proporciona información sobre el 50% central de la variable.

Percentil: P90-P10; se utiliza cuando se quieren excluir algunos datos extremos de información; recoge información sobre el 90% central de los datos.

*La Varianza: (σ^2)

*Las desviaciones: La Desviación Media y la Desviación Típica o Standard (S.D.)

Coeficiente de variación

Es un valor que da información global sobre el grado de dispersión de la medida estadística utilizada; cuando la medida empleada es la media aritmética se define como:

$$CV = \frac{SD}{\bar{x}}$$

Es muy útil para comparar estadísticamente dos poblaciones o dos muestras diferentes. Nos muestra la variación relativa de cada población. Quien posea el coeficiente de variación mayor será la más heterogénea.

Medidas de posición: Los Cuantiles

Se denominan así a ciertos valores dentro de una información, que permiten dividirla en partes iguales. Hemos visto que la **Mediana** es el valor que divide al conjunto de datos en dos partes iguales, pues bien a las medianas de cada una de estas dos partes iguales que nos han quedado, los denominamos **Cuantiles**.

Los cuantiles más usados son

Los Cuantiles (Q):

Se utilizan para dividir la información en cuatro (4) partes iguales, cada una de las cuales contiene un 25 % del total de los datos. Son 4 cuantiles notados como: Q_1 - Q_2 - Q_3 - y - Q_4 .

Los Deciles (D):

Se usan para dividir una información en diez (10) partes iguales, cada una de las cuales contiene un 10 % de los datos. Son 10 Deciles notados como: D_1 - D_2 - - - D_{10}

Los Percentiles (P):

Se usan para dividir una información en cien partes iguales, cada una de las cuales contiene un 1 % de los datos. Son cien percentiles notados como: P_1 - P_2 - P_3 - - - P_{100}

La Varianza (σ^2)

Brinda información global sobre la forma en cómo varían los datos; juega un papel trascendental en la estadística inferencial cuando de hacer estimaciones se trata; ya que del análisis de la varianza de una información cuantitativa se pueden deducir muchos resultados sobre el comportamiento general de los parámetros de una población. Se define como la media de los cuadrados de las desviaciones medias. Permite descubrir la variación que existe entre dos muestras de la misma o de diferentes poblaciones.

La varianza se define como:

$$\text{I) } Var(x) = \sigma^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{n} ; \text{ para datos sin agrupar.}$$

$$\text{II) } Var(x) = \sigma^2 = \frac{\sum (M_{ci} - \bar{x})^2 f_i}{n} ; \text{ para datos agrupados.}$$

III) Como fórmula alterna de la varianza, la que se emplea comúnmente es la siguiente:

$$Var(x) = \frac{\sum x^2 f_i}{n} - \left(\frac{\sum x_i f_i}{n} \right)^2 = \bar{x^2} - \left(\bar{x} \right)^2$$

Las Desviaciones

Una desviación se define como la diferencia entre el valor del dato y alguna medida estadística; las desviaciones más comunes se toman con respecto a la media aritmética, pero se pueden tomar desviaciones respecto a la moda, a la mediana, a uno de los cuantiles,

La Desviación Media

Se define como el promedio - la media aritmética - de las desviaciones absolutas de una variable, tomadas en valor absoluto, con respecto a la media aritmética de la variable; en fórmulas será:

$$DM = \frac{\sum |x_i - \bar{x}|}{n}$$

$$DM = \frac{\sum |x_i - \bar{x}| f_i}{n}$$

Desviación Típica o Estándar

Se define como la raíz cuadrada de la varianza; en fórmula será:

$$SD = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2 f_i}{n}}, \text{ para datos sin agrupar.}$$

$$SD = \sigma = \sqrt{\frac{\sum (Mc_i - \bar{x})^2 f_i}{n}}, \text{ para datos agrupados.}$$

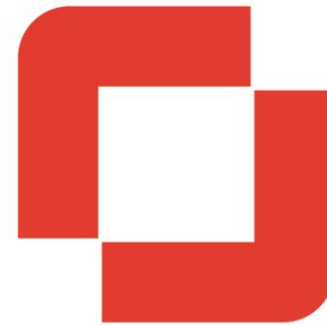
Observación:

La media es el valor del promedio, físicamente puede pensarse como “el centro de gravedad” del conjunto de datos. Se puede imaginar como el valor equitativo que se obtendría al repartir el todo entre sus elementos.

La desviación típica es una medida de lo equitativo que se ha realizado la distribución. Hay menor equilibrio cuando hay mayor desviación típica

GLOSARIO ESTADÍSTICO

Fuente: Murray R. Spiegel, *Estadística*, McGraw-Hill.



MarkQual[®]