Centro de Investigación Académica y Desarrollo Tecnológico del Occidente Colombiano

ESTADÍSTICA

Edivar Fernández Hoyos
edivarf@gmail.com

LA ESTADÍSTICA se puede definir como el conjunto de métodos científicos ligados a la toma, organización, recopilación, presentación y análisis de datos, tanto para la deducción de conclusiones como para tomar decisiones razonables de acuerdo con tales análisis.

Se deriva de la palabra italiana **"STATISTA"** que significa **"EXPRESIÓN"** y fue introducida por primera vez en la Inglaterra del siglo XVIII.
Por tradición fueron los Gobiernos los primeros y principales interesados en tomar información y analizarla para una mejor toma de decisiones.
Hoy la estadística se enfoca como una técnica utilizada en la investigación para la toma, ordenación, presentación y análisis de datos.

La estadística se puede diferenciar como **Deductiva e Inductiva.**

Estadística Deductiva o Descriptiva: Tiene como objetivo la reducción de un conjunto o información a unas pocas medidas descriptivas que permitan conocer de manera científica, las características de los grupos u objetos de la información.
Ejemplo: Presentación estadística de las notas promedio de los alumnos de una clase.

Estadística Inductiva, Inferencial o Analítica: Tiene como objetivo encontrar explicaciones al comportamiento de un conjunto de observaciones; intenta descubrir las causa que producen esos resultados, para encontrar nuevas teorías. Rama apta para los investigadores.

NOTA: *Los seres humanos somos únicos, de tal forma que los datos pueden ser cambiantes de uno a otro. Esto hace de la estadística una labor de cuidado y por ello algunos datos pueden ser irrelevantes en comparación con otros más significativos.*

Algunos conceptos básicos a tener en cuenta:

***Hipótesis*:** Enunciado que se basa en conocimientos ya existentes o en hechos nuevos, o en ambos.

***Variable*:** Característica que cambia en una situación experimental o fenómeno. Alo observar el comportamiento de una variable se obtiene un conjunto de observaciones o conjunto de Datos.

***Dato*:** Registro de una información, o agrupaciones de de cualquier número de observaciones relacionadas. Los datos deben organizarse para que sean útiles y se pueda identificar tendencias y así llegar a conclusiones.

***Población*:** Agrupación de todos los elementos a estudiar para tratar de conocer con base en informaciones y así obtener conclusiones.

***Muestra*:** Selección de algunos elementos de una población, no de todos.

***Aleatorio*:** Sinónimo de "al azar".

***Muestreo*:** Selección de una muestra representativa entre toda una población. De tal forma que el análisis de la muestra dé información de toda la población.

***Muestreo aleatorio*:** Mecanismo de selección mediante el cual todo miembro de la población tiene igual posibilidad e ser escogido.

ELABORACIÓN DE TABLA DE FRECUENCIAS Y REPRESENTACIÓN GRÁFICA

TABLA DE FRECUENCIAS

Ejercicio: Realice la práctica lanzando un dado y registrando los valores que obtiene en cada lanzamiento en la tabla de **"datos brutos".** A partir de estos y con ayuda del tutor elabora la tabla de frecuencias, el diagrama circular o de pastel y el diagrama de barras.

La presentación de datos y resultados a partir de un estudio realizado mediante cualquier instrumento de recolección de datos debe ir en orden jerárquico y de complejidad.

1. A los datos obtenidos a partir del instrumento aplicado les llamaremos ***Datos primarios*** y el valor numérico para cada categoría constituirá las ***Frecuencias absolutas*.** Las frecuencias absolutas equivalen al número de veces que se repite determinada categoría
2. Con base en estos valores podremos hallar las ***Frecuencias relativas.*** Estas no son más que los **porcentajes** de cada categoría comparada con el total de personas aportantes de datos. Si al hallar los porcentajes quedan cifras decimales, se deben **aproximar** (redondear) a una, dos o más cifras por **exceso o defecto.**

Para aproximar a dos cifras decimales, por ejemplo; observamos la tercera cifra decimal:

- Si es 0, 1, 2, 3 ó 4, se aproximará por defecto suprimiendo las cifras decimales de tal forma que solo queden dos cifras decimales.
- Si es 5, 6, 7, 8, ó 9, se aproximará por exceso aumentando en uno la segunda cifra decimal.

Observemos:

Por defecto		**Por exceso**	
$11{,}670003 \cong 11{,}67$	$245{,}98444 \cong 245{,}98$	$3{,}45672 \cong 3{,}46$	$143{,}088 \cong 143{,}09$
$5{,}56298 \cong 5{,}56$	$0{,}673422 \cong 0{,}67$	$27{,}04891 \cong 27{,}05$	$0{,}6765 \cong 0{,}68$

Continuamos con el ejercicio anterior:

Valor obtenido (Categorías)	Frecuencia Absoluta		Frecuencia Relativa		Diagrama Circular	Diagrama de Barras
	Simple	Acumulada	Simple (%)	Acumulada (%)	Ángulo (º)	Altura (cm)
1						
2						
3						
4						
5						
6						

REPRESENTACIÓN GRÁFICA

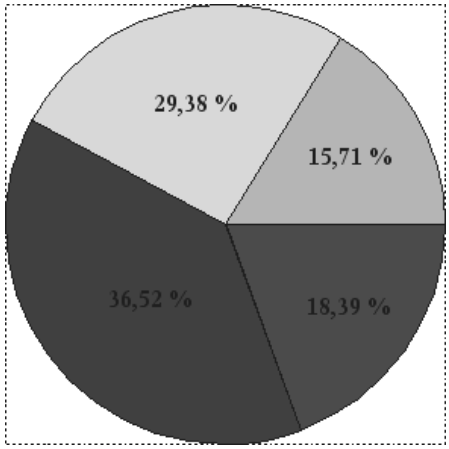
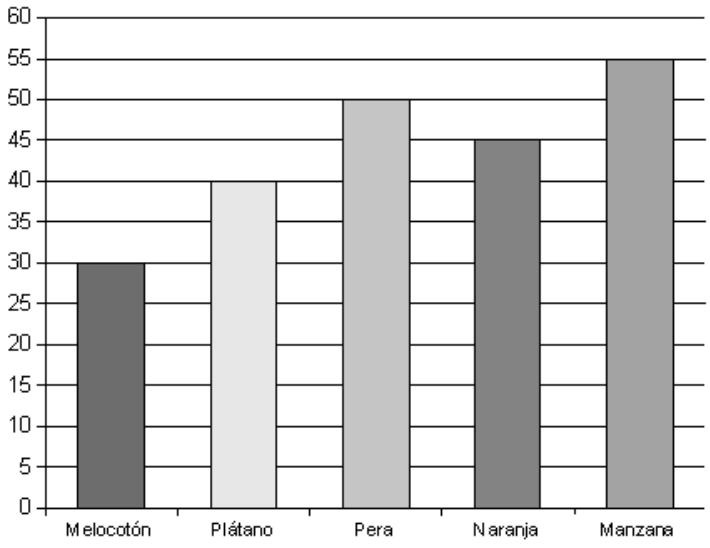


Diagrama circular: Es de los más sencillos de elaborar. Se sugiere que el tamaño no sea muy pequeño ni excesivamente grande y los colores llamativos para facilitar la comparación. Recuerde que el área de cada porción del diagrama debe ser proporcional a la frecuencia relativa que representa.

Diagrama de barras: Es más elaborado y facilita mejor las comparaciones. Tiene unos requerimientos especiales para la altura y el ancho de los rectángulos para facilitar la inspección visual.

- ✓ La base de altura será la mayor frecuencia absoluta y según esta se determina la atura de las otras barras y el ancho de todas ellas. Por ejemplo, si para una frecuencia absoluta máxima de 45 le damos una altura de 12cm, para una frecuencia de 18 le debemos dar una altura de 4.8cm **(Regla de tres).**
- ✓ Se llama **línea base** a la horizontal sobre la que se colocan las barras. Es recomendable que la razón entre la línea base y la altura máxima de las barras sea $\frac{3}{4}$.
- ✓ En el ejemplo anterior si la línea base se toma de 16cm, la barra que representa la mayor frecuencia absoluta tendrá una altura de: $\frac{3}{4}$.16cm=12cm.
- ✓ El ancho de las barras debe ser igual para todas aunque arbitrario, pero se sugiere que dependa del número de categorías a representar.
- ✓ La separación entre las barras no debe ser menor que la mitad del ancho de las barras ni mayos que el ancho de las mismas y siempre se inicia con un espacio (Por comodidad casi siempre los espacios tienen el mismo ancho de las barras). En el ejemplo anterior si hay cinco categorías, dividimos 16cm por el doble de categorías, es decir 10 y por tanto cada barra tendrá un ancho de 1.6cm.



- ✓ Si hay información necesaria para la buena interpretación del diagrama que no queda inmersa en él, se debe colocar al pie un espacio de **"Observaciones".** Por ejemplo el total de personas encuestadas.

Nota 1: En cualquiera de las representaciones debe incluirse un recuadro de convenciones.

TABLA DE FRECUENCIAS AGRUPADAS: Para el ejemplo anterior la tabla de frecuencias es apropiada, pero en casos diferentes no es funcional. **Por ejemplo**, cuando presentamos la información de la estatura de una familia. Si tiene 15 miembros, la tabla de frecuencia tendrá 15 filas con frecuencia absoluta de 1 para cada miembro (Cada miembro de la familia puede tener diferente estatura).
Es mejor presentar la información por rangos, es decir; elaborar una tabla de frecuencias agrupadas. Para ello, se definen **intervalos** en los que se agrupa la información.
El número de intervalos es una decisión que debe tomar el analista: La regla es que **mientras más tramos se utilicen menos información se pierde, pero puede que menos representativa e informativa sea la tabla.**
Para ilustra de mejor forma lo anterior, se presenta una posible tabla de frecuencias donde cada categoría tiene un rango de 10cm:

Estatura (cm)	Frecuencias absolutas		Frecuencias relativas	
	Simple	Acumulada	Simple (%)	Acumulada (%)
1,01 - 1,10	1	1	3,3	3,3
1,11 - 1,20	3	4	10,0	13,3
1,21 - 1,30	3	7	10,0	23,3
1,31 - 1,40	2	9	6,6	30,0
1,41 - 1,50	6	15	20,0	50,0
1,51 - 1,60	4	19	13,3	63,3
1,61 - 1,70	3	22	10,0	73,3
1,71 - 1,80	3	25	10,0	83,3
1,81 - 1,90	2	27	6,6	90,0
1,91 - 2,00	3	30	10,0	100,0

Observe que todos los datos pueden ser incluidos en la categoría respectiva (sin ambigüedades).

MEDIDAS DE POSICIÓN CENTRAL

Las medidas de posición nos facilitan información sobre la serie de datos que estamos analizando. Estas medidas permiten conocer diversas características de esta serie de datos.
Las medidas de posición son de dos tipos:

a) Medidas de posición central: Informan sobre los valores medios de la serie de datos.

b) Medidas de posición no centrales: Informan de como se distribuye el resto de los valores de la serie.

A. MEDIDAS DE POSICIÓN CENTRAL

Las principales medidas de posición central son las siguientes:

1. **MEDIA:** Es el valor medio ponderado de la serie de datos. Se pueden calcular diversos tipos de media, siendo las más utilizadas:

Media Aritmética ($\overline{X}$): Se calcula multiplicando cada valor por el número de veces que se repite. La suma de todos estos productos se divide por el total de datos de la muestra. Es lo que comúnmente se denomina "promedio"

$$\overline{X} = \frac{x_1 . n_1 + x_2 . n_2 + x_3 . n_3 + \ldots + x_n . n_n}{n}$$

En este caso x representa los valores y la muestra tiene n valores.

Media Geométrica ($\hat{X}$) : Se eleva cada valor al número de veces que se ha repetido. Se multiplican todo estos resultados y al producto final se le calcula la raíz "n" (siendo "n" el total de datos de la muestra).

$$\hat{X} = \sqrt[n]{X_1^{n_1} \cdot X_2^{n_2} \cdot X_3^{n_3} \cdot \ldots \cdot X_n^{n_n}}$$

Según el tipo de datos que se analice será más apropiado utilizar la media aritmética o la media geométrica.
La media geométrica se suele utilizar en series de datos como tipos de interés anuales, inflación, etc., donde el valor de cada año tiene un efecto multiplicativo sobre el de los años anteriores. En todo caso, la media aritmética es la medida de posición central más utilizada.
Lo más positivo de la media es que en su cálculo se utilizan todos los valores de la serie, por lo que no se pierde ninguna información.
Sin embargo, presenta el problema de que su valor (tanto en el caso de la media aritmética como geométrica) se puede ver muy influido por valores extremos, que se aparten en exceso del resto de la serie. Estos valores anómalos podrían condicionar en gran medida el valor de la media, perdiendo ésta representatividad.

2. **MEDIANA:** Es el valor de la serie de datos que se sitúa justamente en el centro de la muestra (un 50% de valores son inferiores y otro 50% son superiores).

No presenta el problema de estar influido por los valores extremos, pero en cambio no utiliza en su cálculo toda la información de la serie de datos (no pondera cada valor por el número de veces que se ha repetido).

3. **MODA:** Es el valor que más se repite en la muestra.

A manera de ejemplo vamos a utilizar la tabla de distribución de frecuencias con datos de la estatura de alumnos de la página 2. A partir de esta calculamos los valores de las distintas posiciones centrales:

- **Media aritmética:**

$$\overline{X} = \frac{(1,20*1)+(1,21*4)+(1,22*2)+\ldots+(1,30*3)}{30} = 1,253$$

Luego $\overline{X} = 1,253cm$. La estatura media de este grupo es de 1,253cm.

$\hat{X} = 1,253cm$

- **Media geométrica:**

$$\hat{X} = \sqrt[30]{1,20^1 * 1,21^4 * 1,22^2 * \ldots * 1,30^3} = 1,253$$

Luego $\hat{X} = 1,253cm$.

En este ejemplo la media aritmética y la media geométrica coinciden, pero no siempre da igual.

- **Mediana:** La mediana de esta muestra es 1,26cm, ya que por debajo está el 50% de los valores y por arriba el otro 50%. Esto se puede ver al analizar la columna de frecuencias relativas acumuladas.

En este ejemplo, como el valor 1,26 se repite en 3 ocasiones, la media se situaría exactamente entre el primer y el segundo valor de este grupo, ya que entre estos dos valores se encuentra la división entre el 50% inferior y el 50% superior.

- **Moda:** Hay 3 valores que se repiten en 4 ocasiones: el 1,21, el 1,22 y el 1,28, por lo tanto esta serie cuenta con 3 modas.

MEDIDAS DE POSICIÓN NO CENTRAL

Las medidas de posición no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales. Entre otros indicadores, se suelen utilizar una serie de valores que dividen la muestra en tramos iguales:

- **CUARTILES:** Son 3 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cuatro tramos iguales, en los que cada uno de ellos concentra el 25% de los resultados.
- **DECILES:** Son 9 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en diez tramos iguales, en los que cada uno de ellos concentra el 10% de los resultados.
- **PERCENTILES:** Son 99 valores que distribuyen la serie de datos, ordenada de forma creciente o decreciente, en cien tramos iguales, en los que cada uno de ellos concentra el 1% de los resultados.

Nota: Como ejemplo de aquí en adelante, se tomarán los datos y tabla de frecuencias de la página 4.

Ejemplo: Vamos a calcular los cuartiles de la serie de datos referidos a la estatura del grupo de alumnos que nos ha servido de ejemplo. Los deciles y centiles se calculan de igual manera, aunque haría falta distribuciones con mayor número de datos.
1º cuartil: Es el valor 1,22cm, ya que por debajo suya se sitúa el 25% de la frecuencia (tal como se puede ver en la columna de la frecuencia relativa acumulada).
2º cuartil: Es el valor 1,26cm, ya que entre este valor y el 1º cuartil se sitúa otro 25% de la frecuencia.
3º cuartil: Es el valor 1,28cm, ya que entre este valor y el 2º cuartil se sitúa otro 25% de la frecuencia. Además, por encima suya queda el restante 25% de la frecuencia.

Nota: Cuando un cuartil recae en un valor que se ha repetido más de una vez (como ocurre en el ejemplo en los tres cuartiles) la medida de posición no central sería realmente una de las repeticiones.

MEDIDAS DE DISPERSIÓN

Estudia la distribución de los valores de la serie, analizando si estos se encuentran más o menos concentrados, o más o menos dispersos.
Existen diversas medidas de dispersión, entre las más utilizadas podemos destacar las siguientes:

1. **RANGO:** Mide la amplitud de los valores de la muestra y se calcula por diferencia entre el valor más elevado y el valor más bajo.
2. **VARIANZA:** Mide la distancia existente entre los valores de la serie y la media. Se calcula como la sumatoria de las diferencias al cuadrado entre cada valor y la media, multiplicadas por el número de veces que se ha repetido cada valor. La sumatoria obtenida se divide por el tamaño de la muestra.

$$S_x^2 = \frac{\sum (x_i - x_m)^2 * n_i}{n}$$

Recuerde que n es el número de datos.

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.

3. **DESVIACIÓN TÍPICA O ESTÁNDAR:** Se calcula como la raíz cuadrada de la varianza.
4. **COEFICIENTE DE VARIACIÓN DE PEARSON:** Se calcula como cociente entre la desviación típica y la media.

Ejemplo: Con base en la página 4:
1. **Rango:** Diferencia entre el mayor valor de la muestra (1,30) y el menor valor (1,20). Luego el rango de esta muestra es 10cm.
2. **Varianza:** Recordemos que la media de esta muestra es 1,253. Luego, aplicamos la fórmula: $S_x^2 = \frac{((1,20-1,253)^2*1)+((1,21-1,253)^2*4)+((1,22-1,253)^2*4)+.....+((1,30-1,253)^2*3)}{30} = 0,0010$.
   Por lo cual, la varianza es 0,0010.

3. **Desviación Típica o Estándar:** Es la raíz cuadrada de la varianza. $\sigma = \sqrt{S_x^2} = \sqrt{0.0010} = 0,0316$, luego la desviación típica es 0,0316.
4. **Coeficiente de Variación de Pearson:** se calcula como cociente entre la desviación típica y la media de la muestra; es decir: $Cv = \frac{\sigma}{\bar{X}}$. Luego $Cv = \frac{0,0316}{1,253} = 0,0252$

El interés del coeficiente de variación es el que al ser un porcentaje permite comparar el nivel de dispersión de dos muestras. Esto no ocurre con la desviación típica, ya que viene expresada en las mismas unidas que los datos de la serie.

Por ejemplo, para comparar el nivel de dispersión de una serie de datos de la altura de los alumnos de una clase y otra serie con el peso de dichos alumnos, no se puede utilizar las desviaciones típicas (una viene expresada en cm y la otra en Kg). En cambio, sus coeficientes de variación son ambos porcentajes, por lo que sí se pueden comparar.