EXPECTATIVAS DE LOGRO:

- Establecer las medidas de posición más utilizadas.
- Diferenciar las medidas de tendencia central para datos agrupados y para datos no agrupados.
- Analizar el uso de la medida de posición más adecuada, dependiendo del problema planteado.
- Comprender la importancia de la utilización de las medidas de dispersión para el análisis exploratorio de datos y su aplicación a problemas concretos
- Aplicar los conceptos y formas de calculo aprendidos, utilizando software especifico de estadística.

CONTENIDOS

2.1-Medidas de Posición: Modo, Mediana, Media aritmética., Cuartiles, Percentiles. Su aplicación al Diseño Industrial. Características. Propiedades. Usos.

2.2 Medidas de dispersión: Rango o amplitud de variación, valores cuarticos y percentiles. Desviación Media, varianza, desviación estándar, Coeficiente de variación. Características, propiedades y uso de las medidas de dispersión.

INTRODUCCIÓN

Las medidas descriptivas son medidas que se calculan a partir de los datos disponibles en una Distribución de Frecuencias, aportando información sobre distintos aspectos de la misma. Básicamente nuestro interés está centrado en dos grandes tipos de medidas: de Posición o tendencia Central y de Dispersión o Variación.

2.1 MEDIDAS DE POSICIÓN

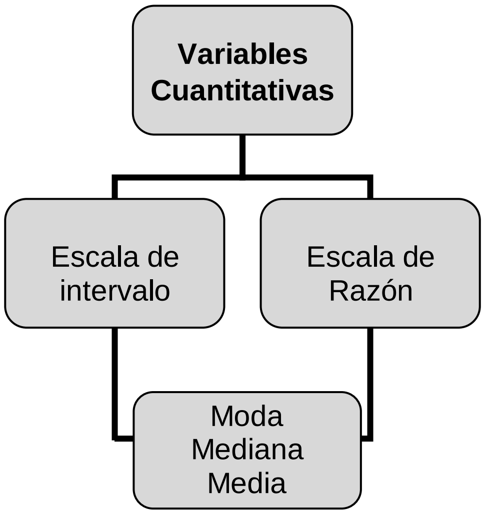
INTRODUCCIÓN

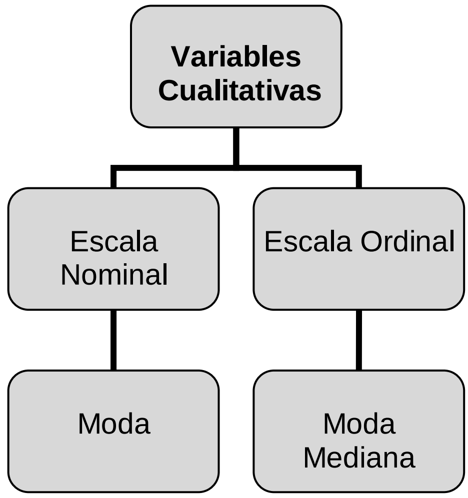
En la Unidad N° 2 se ha visto que después de recolectar un conjunto de datos, ellos pueden ser organizados y resumidos a los fines de un análisis posterior. En esta unidad avanzaremos un poco mas en el análisis exploratorio de datos, presentando algunas medidas resúmenes referidas concretamente al posicionamiento de una distribución.

En primer lugar, aprenderemos a calcularlas y luego analizaremos la posibilidad y conveniencia del uso de cada una de ellas de acuerdo al tipo de variables y/o forma de la distribución de frecuencias.

Se conocen varias medidas que cumplen con el requisito de posicionar a un conjunto de datos y veremos a continuación las tres más comúnmente utilizadas: moda, mediana y media aritmética.

El cálculo de estas medidas difiere de acuerdo al tipo de variables con que se trabaja y presentan pequeñas modificaciones según se disponga de datos agrupados o no.





En esta unidad solo se obtendrán las medidas de posición y dispersión mas usadas en el tratamiento estadístico de los datos.

MODA: (Mo)

Corresponde al valor de la variable que más se repite. Ésta sirve para describir una distribución, si sólo se desea tener una idea aproximada y rápida de donde está la mayor concentración de observaciones. También se la utiliza para describir la forma de algunas distribuciones.

Se debe hacer notar que la moda es el valor de la variable y la frecuencia de este valor sugiere su importancia estadística.

La moda se puede obtener tanto a las variables categóricas como a numéricas.

EJEMPLO N° 1

Máximo nivel educacional alcanzado	Frecuencia absoluta
Primario	**38**
Secundario	21
Universitario	8

En este ejemplo, la categoría de la variable que se presenta con mayor frecuencia es la categoría **nivel primario**, ya que para esta variable se presenta la mayor frecuencia

Si la variable es discreta la moda se calcula de igual forma. En la siguiente tabla se presenta la información referida al número de materias rendidas por los alumnos al finalizar el primer año.

EJEMPLO N° 2

N° de materias rendidas	Frecuencia Absoluta
0	8
1	12
2	**14**
3	6
4	3

La **moda** correspondiente a la variable N° de materias rendidas es **2,** ya que allí se tiene la mayor frecuencia

Cuando se mide una variable continua tal como altura, peso, etc. todas las mediciones pueden ser diferentes, en este caso no se observa ningún valor modal, pues cada valor observado tiene frecuencia 1. Sin embargo ya hemos visto que los datos al ser agrupados en intervalos de clase y luego la moda se determina en función de la frecuencia de dichos intervalos. Cuando se tienen datos cuantitativos (numéricos) agrupados, existe un **intervalo modal** que es el intervalo de mayor frecuencia. Algunos autores toman la moda como el punto medio del intervalo modal, este valor se llama **marca de clase.**

Para la obtención del modo o moda a partir de un grafico éste se obtiene en el caso de variable discreta en el bastón de mayor frecuencia (el más grande) y cuando los datos están agrupados en los intervalos de mayor frecuencia.

EJEMPLO N° 3

En la tabla adjunta se muestra cómo se han repartido 1.200 calificaciones entre 0 y 10, en 10 intervalos iguales

(a) INTERVALO	(b) MARCA CLASES	(c) FRECUENCIA
[0-1)	0.5	20
[1-2)	1.5	15
[2-3)	2.5	18
[3-4)	3.5	25
[4-5)	4.5	44
[5-6)	5.5	88
[6-7)	6.5	222
[7-8)	**7.5**	**335**
[8-9)	8.5	218
[9-10]	9.5	215

Tabla de frecuencias con datos agrupados en intervalos

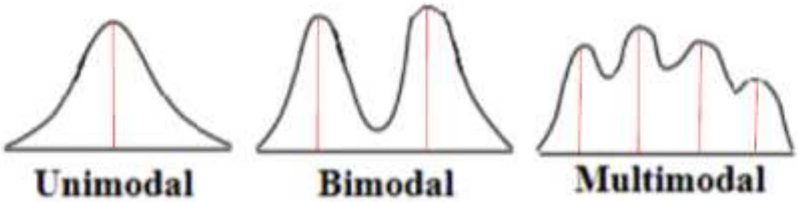
En este caso la moda se encuentra en el octavo intervalo ya que la frecuencia para ese intervalo es la mayor.
Se puede afirmar que **el intervalo modal** para la variable calificaciones es el [7-8) o que el **valor modal** es la calificación 7,5 que corresponde a la marca de clase de dicho intervalo.

Puede suceder que en una distribución existan dos modas, es decir dos variables con la misma frecuencia, en ese caso se denomina **distribución bimodal.**

La moda es un buen indicador del centro de los datos solo si hay una frecuencia dominante.

Puede ocurrir que en un conjunto de datos **no** haya moda, como en: 3; 4; 7; 9; 10; 11; 13.

O también que haya **varios** valores con la mayor frecuencia, en estos casos la moda se dice multimodal**.**



MEDIANA (Me)

La mediana de un conjunto de observaciones es un **valor de la variable** que divide a ese conjunto **ordenado de menor a mayor** en dos subconjuntos que contiene la misma cantidad de datos. **La mediana es aquel valor de la variable que ocupa el lugar central, de modo que la mitad de los casos queda por debajo de ese valor y la otra mitad por encima**

SIN AGRUPAMIENTO DE DATOS (serie simple)

- Si el número de valores es **impar**, consideramos el valor central: 2; 3; 5; 7; **11**; 13; 16; 18; 25. La mediana es Me = 11.
- Si el conjunto de valores es un número **par**, entonces se calcula la media aritmética a los dos valores del centro. 2; 3; 5; 7; **11; 13**; 16; 18; 25, 27. La mediana es 12.

La expresión analítica para encontrarla es:

Se ordenan los datos de menor a mayor.

Si n (cantidad de observaciones) es impar entonces la mediana es el lugar de la variable que ocupa el lugar central $M_e = x_{\frac{n+1}{2}}$

- Si n es par entonces se calcula como un promedio de los dos valores centrales

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

CON AGRUPAMIENTO DE DATOS

- ***SERIE DE FRECUENCIAS* (Variable Discreta).**

En esta se considera la frecuencia acumulada: f_{ia} La mediana corresponde a la observación cuya frecuencia acumulada contiene a n=22, ya que quedan 21 datos antes y 21 después, la frecuencia acumulada que contiene a n=22 es 28 .Luego **el valor de la variable es 2** por ello concluimos que $\mathbf{M_e = 2}$ o si ingresamos en la columna de las frecuencias acumuladas relativas, la mediana corresponde a la observación donde está incluido el 50%

Tabla N° 1

Variable X_i	**Frec. Abs.** f_i	**Frec. Relativa** f_{ir}	**Frec. Acumulada** f_{ia}	**Frec. Rel. Acum.** f_{ira}
0	8	18,6%	8	18,6%
1	11	25,6%	19	44,2%
2	9	20,9%	**28**	**65,1%**
3	6	13,9%	34	79,0%
4	3	7,0%	37	86,0%
5	3	7,0%	40	93,0%
6	2	4,7%	42	97,7%
7	0	0,0%	42	97,7%
8	1	2,3%	43	100%

Para los casos en que no haya un valor de X que acumule exactamente la mitad de los datos, no importa si n es par o impar, y la Me será el primer valor de X cuya f_{ia} supere n/2

Determinación analítica
Para encontrar el valor de la mediana el procedimiento consiste en buscar sobre las frecuencias absolutas acumuladas (F_{ia}), el valor de la variable que acumula hasta n/2 observaciones.
Determinación Gráfica
Utilizando el grafico de frecuencias absolutas acumuladas ingresamos sobre el eje de ordenadas (vertical) con n/2 y trazamos una línea horizontal por este valor hasta tocar el grafico. Luego bajamos hasta el eje de abscisas (horizontal) y el punto que encontramos es el valor de la mediana.

- ***DISTRIBUCIÓN DE FRECUENCIA EN INTERVALOS DE CLASE***[1] **(***Generalmente para Var. Continuas o para un rango muy grande de Var. Discretas)*

Consideremos los intervalos de clase y las frecuencias acumuladas correspondientes a dichos intervalos.

EJEMPLO N° 4

x	f_i	f_{ai}
[150,155)	3	3
[155,160)	10	13
[160,165)	13	**26**
[165,170)	11	37
[170,175]	3	40

Se calcula en primer término la ubicación del dato n/2; para este caso n/2 =40/2=20
La **mediana** está comprendida en el intervalo [160,165) que denominaremos **Intervalo Mediano**, que corresponde a la frecuencia acumulada 26, ya que la mediana esta entre el dato 20 y 21, se encuentra ambos en ese intervalo.
$160 < M_e < 165$ $\quad M_e \cong 162$
El valor de la mediana en este caso es aproximado

[1] Existen expresiones más exactas para determinar la mediana cuando trabajamos con una distribución de frecuencias en intervalos de clases. Pero como el objetivo de este curso no es el calculo en si, sino su interpretación es que usaremos esta expresión simplificada..

Cuando la distribución de frecuencia viene dada para valores de la variable agrupados en intervalos de clase, no puede obtenerse exactamente el valor de la mediana porque se desconocen las observaciones individuales.

Determinación analítica

Para encontrar el valor de la mediana con una mayor aproximación se utiliza una formula, pero en este curso la obtendremos en forma simplificada y se procede en forma similar al caso de series de potencias

Determinación Gráfica

Utilizando el grafico de frecuencias absolutas acumuladas ingresamos sobre el eje de ordenadas (vertical) con n/2 y trazamos una línea horizontal por este valor hasta tocar el grafico. Luego bajamos hasta el eje de abscisas (horizontal) y el punto que encontramos es el valor de la mediana.

La expresión matemática usada para su cálculo es:

$$M_e = L_i + \frac{\frac{n}{2} - f_{a(i-1)}}{f_{ai}} h_i$$

L_i: límite inferior del intervalo correspondiente a la frecuencia absoluta acumulada que contiene a la cantidad n/2
$f_{a(i-1)}$:la frecuencia absoluta acumulada hasta el intervalo anterior al que contiene la mediana.
f_{ai}=la frecuencia absoluta del intervalo en que ubicamos la mediana.
h_i=la amplitud del intervalo en el que se encuentra la mediana

CUARTILES

Los cuartiles de una distribución, como su nombre lo indica, son los valores de la variable que dividen al conjunto de datos (ordenados de menor a mayor) en cuatro subconjunto que contienen la misma cantidad de datos. La mediana coincide con el segundo cuartil.

EJEMPLO N° 5

Supongamos que se han registrado los pesos de 8 paquetes de algodón, estos se ordenaron de menor a mayor obteniendo:

146 _ 148 _ 149 _ 150 _ 152 – 154 _ 157 _ 158

$\downarrow$ $\boxed{Q_1=148,5}$ $\downarrow$ $\boxed{Me=151=Q_2}$ $\downarrow$ $\boxed{Q_3=155,5}$

PRIMER CUARTIL: Q1 El primer cuartil es un valor de la variable tal que el 25% de los datos son menores y el 75% son mayores.

TERCER CUARTIL: Q3 El tercer cuartiles el valor de la variable tal que e 75% de los datos son menores que dicho valor y el 25% son mayores.

Las reglas para obtener los cuartiles son las mismas que para la mediana.

Si el número de datos (n) es múltiplo de 4 entonces el cálculo es el siguiente

$$Q_1 = Q_{INFERIOR} = \frac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2} \quad y \quad Q_3 = Q_{Superior} = \frac{x_{\frac{3n}{4}} + x_{\frac{3n}{4}+1}}{2}$$

En el caso que el número de datos (n) no sea múltiplo de 4 entonces los cuartiles coinciden con algún valor de la serie ordenada de observaciones.

En el caso de trabajar con datos agrupados y si la variable es discreta, se procede de la misma manera que con la mediana. Para considerar el orden para Q_1 se debe considerar n/4 y en el caso de Q_3 se debe utilizar ¾ n, así se obtiene la posición en que se encuentran los cuartiles, y para obtener el cuartil correspondiente se determina el **valor de la variable** que se encuentra en la posición antes mencionada

EJEMPLO N° 6.

En el siguiente arreglo ordenado de dan las horas extras trabajadas por un grupo de trabajadores de una sección.

10, 20,28, 29 ,29,29,30,30,**30**,31,32,32,32 , 33,35,37,38

Q1 Q3

Q1= 29.Se obtiene del promedio del 4to y 5to valor y que es para ambos de 29; Q3= 32,5 obtenido como promedio entre los valores de la variable 32 y 33

Para el caso de datos agrupados continuos se suelen aplicar diferentes formulas de interpolación, en nuestro caso lo simplificaremos y los obtendremos como para variable discreta. Una forma de encontrarlas es trabajando con la grafica de las **frecuencias relativas acumuladas.**

PERCENTILES

Son valores de la variable, que dividen al conjunto de datos (ordenados de menor a mayor) en cien partes iguales. Es una medida estadística muy utilizada, que nos permite saber cómo está posicionado **un valor respecto al total de una muestra**.

Cuando se habla del percentil 10, se quiere expresar que es el valor de la variable que déjale 10% de los datos por debajo de ese valor y el 90% de los mismos por encima. Para calcularlo se divide el conjunto de observaciones en 100 partes iguales o subconjuntos, tal que cada uno de ellos contiene la misma cantidad de datos.

Los datos antropométricos se expresan en **percentiles**[2]; la población se fracciona en categorías de porcentajes, ordenados de menor a mayor, de acuerdo con alguna medida del cuerpo. El primer percentil en estatura, por ejemplo, indica que el 99% de la población estudiada

[2] Algunos autores a los percentiles le suelen llamar **centiles**, Nosotros adoptaremos el termino percentil ya que es el usado por los autores sajones y también en Ergonomía.

superaría esta dimensión. De igual manera el percentil 98, diría que solo el 2% de la población sobrepasaría esta estatura, mientras que el 98% restantes tendría alturas inferiores.

Utilizaremos al polígono de frecuencias relativas acumuladas para calcular algunos percentiles en forma gráfica.

En el gráfico N° 6 de la unidad anterior (pag. 11) en el polígono de frecuencias relativas acumuladas encontraremos los siguientes percentiles: P_{15} , $P_{50}= Q_2=$ Me P_{80}

MEDIA ARITMÉTICA O PROMEDIO

Corresponde a la suma de todos los datos dividido por el número total de ellos. Es lo que se conoce como "promedio". La media aritmética es uno de los estadígrafos más usados, por el hecho de ser de muy fácil cálculo

Modo de calcular el promedio

Sin agrupamiento de datos $\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + ... + x_n}{n}$

EJEMPLO:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+3+5+8+10+12+14+16+18+20}{10} = \frac{108}{10} = 10.8$$

Con agrupamiento de datos

- **Promedio de una serie de frecuencias**

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

- **Promedio de una distribución de frecuencias en intervalos de clase**

$$\bar{x} = \frac{\sum f_i . x_{im}}{n}$$

Siendo f_i la frecuencia absoluta del intervalo y x_{im} la marca de clase o valor medio del intervalo respectivo.

En estadística, la media es una medida de centralización. Se llama media de una distribución estadística a la media aritmética de los valores de los distintos individuos que la componen.

Encontrar la media aritmética con los datos de las tablas de los ejemplos N° 2 y 3.

Existen otras medias como son: **Media geométrica, Media armónica y Media ponderada**. Estas son de aplicación frecuente en economía.

RANGO MEDIO

El rango medio es el promedio de la observación más pequeña y la observación mas grande en un conjunto de datos. Se obtiene con la suma del valor más pequeño y el valor mas grande dividido entre 2.

$$\text{Rango medio} = \frac{(X\ mas\ pequeño + X\ mas\ grande)}{2}$$

El rango medio se utiliza como una medida de resumen tanto para análisis financieros como para reportes metereológicos, porque puede proporcionar una medida adecuada, rápida y sencilla que caracteriza a todo el conjunto de datos.: Sin embargo a pesar de estas ventajas, debe utilizarse con cuidado.

La media o promedio actúa como un centro de gravedad que equilibra los desvíos positivos y negativos. Esta es la medida de posición por excelencia, porque en su cálculo intervienen todos los valores observados de X ., además es indispensable para el calculo de las medidas de dispersión, como veremos a continuación.

La mediana es el valor central de la distribución; separa a la serie en dos partes: todos los valores anteriores a la mediana son menores y todos los valores posteriores son mayores.

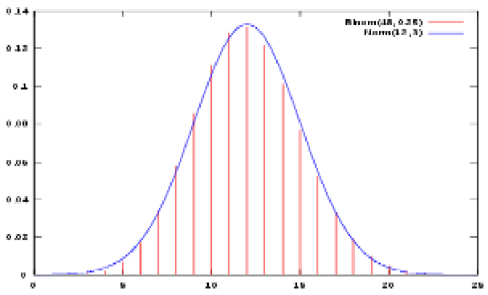
La moda indica el valor de la frecuencia máxima.

Los tres valores centrales dan una idea bastante aproximada de la distribución de la serie.

FORMAS DE LAS CURVAS

Para las distribuciones unimodales se pueden clasificar en distribuciones simétricas o asimétricas.

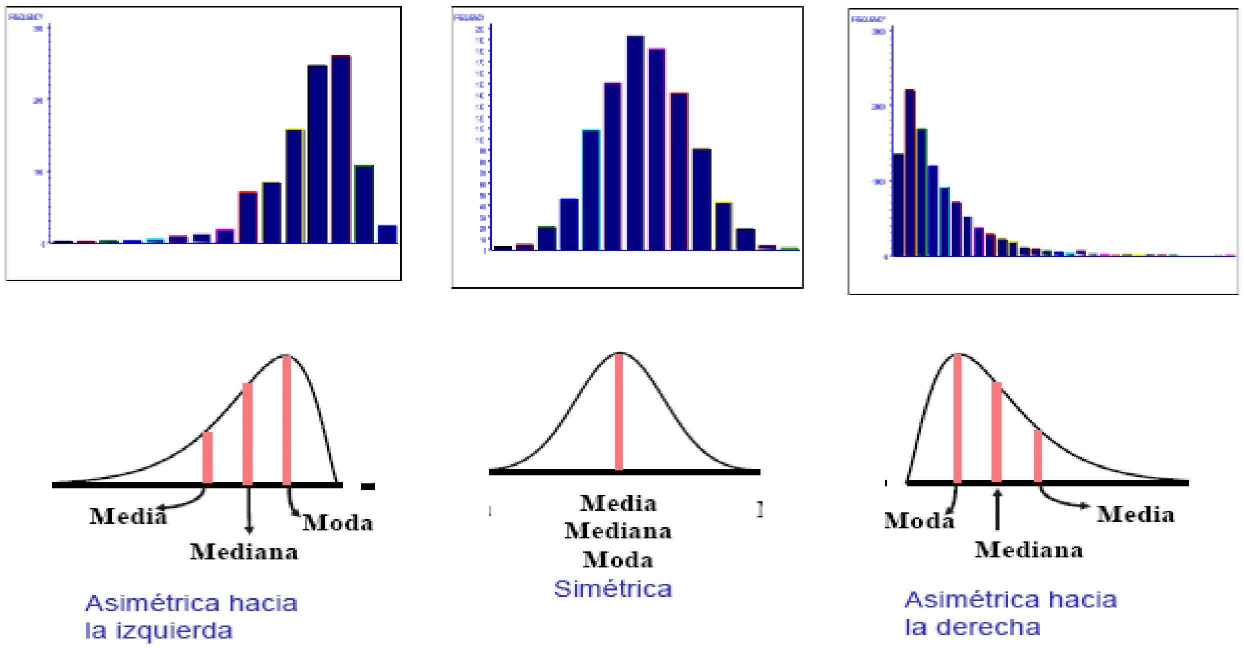
Distribución normal. Las curvas que representan una distribución en intervalos de clase adoptan distintas formas.



Cuando coinciden los tres parámetros de posición: Promedio, mediana y modo, se dice que la distribución es normal. La curva es simétrica y se denomina curva de Gauss

Cuando no coinciden los parámetros de posición las curvas son asimétricas a derecha o izquierda.

Distribuciones asimétricas. Cuando hay asimetría a la izquierda la ubicación de menor a mayor será: $\bar{x}$, Me, Mo. Si es asimétrica a la derecha, las posiciones se invertirán quedando de menor a mayor Mo, Me, $\bar{x}$.



Diremos que una distribución es asimétrica a la derecha si las frecuencias (relativas o absolutas) descienden más lentamente por la derecha que por la izquierda;

Si la distribución desciende más lentamente hacia la izquierda que a la derecha se dice que la distribución es asimétrica a la izquierda.

ELECCIÓN DE UNA MEDIDA DE POSICIÓN ADECUADA

La media aritmética al intervenir todos los valores observados de X hace que sea una medida muy sensible a la aparición de valores extremos en la Distribución de frecuencias, y puede llegar a perder representatividad cuando estos valores extremos rompen marcadamente la simetría de la misma porque tiende a alejarse de la zona donde están la mayoría de los datos. Esto sucede en las distribuciones asimétricas donde hay valores extremos concentrados en una dirección de la distribución. En estos casos la media aritmética no es un valor representativo de la distribución y es más conveniente usar la moda o la mediana como medidas de posición.

Cuando la distribución es bimodal, como en los dos últimos gráficos de la sección anterior ninguna medida de posición provee información útil. En esté caso se encuentran mezcladas dos poblaciones bastante diferentes. En este caso conviene calcular las medidas de posición de cada una de las poblaciones independientemente.

La Mediana, no se ve tan afectada en distribuciones asimétricas, porque no depende de la magnitud de los datos, y es por lo tanto mas representativa en estos casos.

3.2 MEDIDAS DE DISPERSIÓN

INTRODUCCIÓN

Aunque para algunos propósitos un promedio puede ser una descripción suficiente de una población o de una muestra, generalmente se necesita mayor información acerca del conjunto de observaciones en estudio. Nuestras descripción de un conjunto de datos es insuficiente con las medidas de posición, son indispensables otras medidas como son las de dispersión que nos permitirán analizar el comportamiento de los datos y también para comparar dos o mas distribuciones.

Se estudiaran algunos parámetros estadísticos que miden cómo de diseminados o "separados" se encuentran los datos de una distribución. Los más utilizados **se refieren al grado de lejanía de los datos respecto a la media** y son la desviación media, la varianza, la desviación típica y el coeficiente de variación.

CÁLCULO DE LOS PARÁMETROS DE DISPERSIÓN DE UNA SERIE SIMPLE

RANGO O RECORRIDO (R): Se denomina rango de un conjunto de observaciones a la diferencia entre el mayor y el menor valor de la variable.

Si es pequeño, significa que los datos están concentrados en un corto intervalo dentro del campo de variación de X, y por ende habrá poca variación. Si es grande significa en principio lo contrario, pero tiene el problema que no nos dice nada sobre lo que ocurre con los datos intermedios.

Tomado este valor como medida de dispersión es completamente dependiente de los dos valores extremos que toma la variable; es por ello que aporta poca información de la variabilidad de los datos.

El *recorrido intercuartílico* es la diferencia, Q3 - Q1, entre el cuartil superior, Q3, y el cuartil inferior, Q1. El par de parámetros formado por la mediana, Me, y el recorrido intercuartílico, Q3 - Q1, proporciona una buena información sobre la forma de la distribución.

DESVIACIÓN*:* Se llama desviación o desvío de un valor de la variable a la diferencia entre el valor de la variable y el promedio d= $x_i - \bar{x}$. Para el siguiente ejemplo, donde la variable es el número de materias rendidas, se observa que la suma de los desvíos es cero.

X	$\bar{x}$	d= $x-\bar{x}$
4	8	-4
6	8	-2
6	8	-2
7	8	-1
9	8	1
11	8	3
13	8	5
		Suma=0

DESVIACIÓN MEDIA, (D.m)

Es un promedio de los valores absolutos de las desviaciones, de cada elemento, *xi*, de la distribución respecto a su media, $|x_i - \bar{x}|$

$$D.m. = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n} = \frac{\Sigma |x_i - \bar{x}|}{n}$$

Por ejemplo, en la distribución 4, 6, 6, 7, 9, 11, 13, cuya **media es 8**, la desviación media es:

$$D.m. = \frac{|4-8| + |6-8| + |6-8| + |7-8| + |9-8| + |11-8| + |13-8|}{7} =$$

$$= \frac{4+2+2+1+1+3+5}{7} = \frac{18}{7} = 1{,}14$$

La desviación media mide el promedio de las diferencias entre los valores observados respecto de la media del grupo, sin tener en cuenta el signo de la desviación. A diferencia del rango, la desviación media toma en cuenta todas las observaciones en una distribución.

VARIANZA (V)

Es la suma de las desviaciones con respecto a la media aritmética elevadas al cuadrado dividida por el número de observaciones

$$V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

La fórmula anterior es equivalente a esta otra:

$$V = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \bar{x}^2 = \frac{\Sigma x_i^2}{n} - \bar{x}^2$$

que resulta más cómoda de aplicar, sobre todo cuando la media, no es un número entero.

En la distribución 4, 6, 6, 7, 9, 11, 13, de media 8, la varianza es:

$$V = \frac{(4-8)^2 + (6-8)^2 + (6-8)^2 + (7-8)^2 + (9-8)^2 + (11-8)^2 + (13-8)^2}{7} =$$

$$= \frac{16+4+4+1+1+9+25}{7} = \frac{60}{7} = 8{,}57$$

Aplicando la segunda fórmula se obtiene, obviamente, el mismo resultado:

$$V = \frac{4^2+6^2+6^2+7^2+9^2+11^2+13^2}{7} - 8^2 =$$

$$= \frac{16+36+36+49+81+121+169}{7} - 64 = \frac{508}{7} - 64 = 8{,}57$$

Como se desprende de las expresiones de la varianza, al elevar las desviaciones al cuadrado, se altera la unidad de medida, quien también quedará al cuadrado. Para evitar este inconveniente se emplea como medida de dispersión la desviación estándar.

DESVIACIÓN TÍPICA O DESVIACIÓN ESTÁNDAR (σ)

Es la raíz cuadrada positiva de la varianza.

$$\sigma = \sqrt{V} = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma x_i^2}{n} - \bar{x}^2}$$

La razón de ser de este parámetro es conseguir que la medida de dispersión se exprese en las mismas unidades que los datos a los que se refiere. Por ejemplo, en una distribución de estaturas en la que los datos están dados en centímetros (cm), la media viene dada en centímetros, pero la varianza en centímetros cuadrados (cm2). Para evitar este inconveniente se calcula su raíz cuadrada, obteniéndose así la desviación típica en centímetros.

El par de parámetros formado por la media y la desviación típica $(\bar{x}, \sigma)$ aporta una información suficientemente buena sobre la forma de la distribución.

Tanto la Varianza como la Desviación típica toman valores que están relacionadas con la dimensión de los datos estudiados, de modo que una S=200, puede ser muy grande para datos que se mueven entre 150 y 350, y muy pequeña para datos que están entre 1000 y 2000. Esto hace que estas medidas no sean adecuadas para comparar las variaciones entre dos o mas Distribuciones de Frecuencias con valores de X que se mueven en distintos campos de variación. Tampoco será posible la comparación si las variables están dadas en unidades de medida diferentes. Para ello se empleará otra medida de dispersión.

COEFICIENTE DE VARIACIÓN (C.V)

Es el cociente entre la desviación típica y la media de la distribución:

$$C.V. = \frac{\sigma}{\bar{x}}$$

Este parámetro sirve para **relativizar** el valor de la desviación típica y así poder comparar la dispersión de dos poblaciones estadísticas con gamas de valores muy distintas.

El coeficiente de variación no tiene unidad, a veces se suele dar en porcentajes.

Por ejemplo, si en una compañía mexicana los salarios de los empleados tienen una media $\bar{x}_1$ = 7.000 pesos y una desviación típica σ_1 = 500 pesos y en otra empresa española la media de los salarios es $\bar{x}_2$ = 2.000 euros y la desviación típica σ_2 = 400 euros, para comparar la dispersión de salarios se recurre al coeficiente de variación:

C.V.$_1$ = 500/7.000 = 0,07 C.V$_1$= 7%

C.V.$_2$ = 400/2.000 = 0,2 C.V$_2$ =20%

Se aprecia así que en la primera compañía los salarios tienen menor dispersión que en la segunda.

CÁLCULO DE LOS PARÁMETROS DE DISPERSIÓN DE UNA SERIE DE FRECUENCIAS

Para el desvío medio la expresión usada es: $D.m = \frac{\sum |x_i - \bar{x}| . f_i}{n}$

Para la varianza $V = \frac{\sum (x_i - \bar{x})^2 . f_i}{n}$ Para la **desviación estándar**: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 . f_i}{n}}$

A partir del siguiente ejemplo completaremos la tabla y calcularemos las medidas de dispersión

Tabla N° 2

N° de materias rendidas	Frecuencia Absoluta	$\bar{x}$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert^2$	$\lvert x_i - \bar{x} \rvert^2$ fi
0	7	1,65	1,65	2,72	19,06
1	11	1,65	0,65	0,42	4,65
2	**14**	1,65	0,35	0,12	1,72
3	5	1,65	1,35	1,82	9,11
4	3	1,65	2,35	5,52	16,57
	n=40				51,10

Varianza: $\sigma^2 = \frac{51,10}{40}$ =1,28 Desviación Estándar: $\sigma = \sqrt{1,28} = 1,13$ CV= $\frac{1,13}{1,65} = 0,68$

CÁLCULO DE LOS PARÁMETROS DE DISPERSIÓN EN UNA DISTRIBUCIÓN EN INTERVALOS DE CLASE.

Desvío medio: la expresión usada es: $D.m = \frac{\sum |x_m - \bar{x}| . f_i}{n}$

Varianza: $\sigma^2 = \frac{\sum (x_{mi} - \bar{x})^2 . f_i}{n}$ **Desviación estándar**: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 . f_i}{n}}$

Completando la Tabla II

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
X	f_i	X_m	$\overline{x}$	$\lvert x_m - \overline{x}\rvert$	$\lvert x_m - \overline{x}\rvert \cdot f_i$	$\lvert x_m - \overline{x}\rvert^2$	$\lvert x_m - \overline{x}\rvert^2 \cdot f_i$
[150,154)	3	152	162	10	30	100	300
[155,159)	10	157	162	5	50	25	250
[160,164)	13	162	162	0	0	0	0
[165,169)	11	167	162	5	55	25	275
[170,174)	3	172	162	10	30	100	300
	40				165		1125

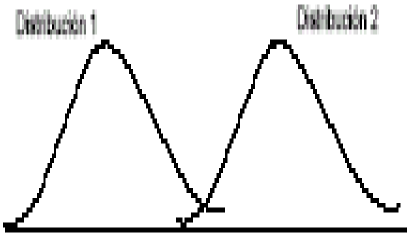
Para el ejemplo de la tabla anterior es:

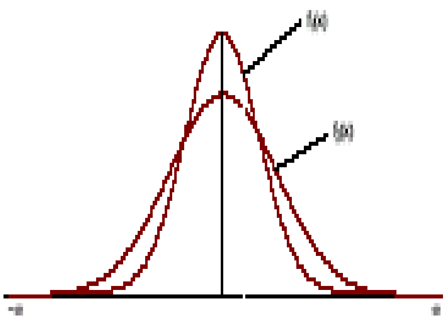
$$Dm = \frac{165}{40} = 4{,}125 \qquad \sigma^2 = \frac{1125}{40} = 28{,}1 \qquad \sigma = \sqrt{\frac{1125}{40}} = \sqrt{28{,}1} = 5{,}3$$

ANÁLISIS DE LAS CURVAS DE DISTRIBUCIÓN DE FRECUENCIAS.

En el caso en que los datos recolectados correspondan a una población muy grande, aun tomando intervalos muy pequeños habrá un gran número de observaciones en cada intervalo. A medida que se achican los intervalos, se achica la base de los rectángulos y el *polígono de frecuencias* se aproxima a una *curva de frecuencias*. A continuación analizaremos algunas particularidades que se pueden presentar en las curvas de frecuencias.

Dos conjuntos de datos pueden diferir tanto en la tendencia central como en la variación de los datos; En el primer gráfico aparecen la distribución 1 y 2 que difieren solo en la tendencia central (diferentes medias). En la segunda figura ambos polígonos tienen las mismas medidas de tendencia central pero difieren en las medidas de variación: el polígono B es menos variable que el polígono A. El polígono A tiene una desviación mayor, e indica que los valores están más dispersos; mientras que en el polígono B la desviación estándar es mas pequeña, esto indica que los valores se concentran entorno al valor medio.





DISTRIBUCIÓN NORMAL CURVA DE GAUSS

Para el estudio de la distribución se utiliza fundamentalmente la desviación estándar, esta es muy útil en la distribución normal de frecuencias.

El coeficiente de variación es el indicador de la forma en que los valores de la serie se agrupan en torno a la media aritmética. Por eso constituye una de las características sobresalientes de la serie.

La distribución normal esta representada por la curva de Gauss. Esta queda determinada por la media aritmética y la desviación estándar.

Conocidos estos dos parámetros se puede calcular la frecuencia de cualquier valor de la variable o de un intervalo limitado por dos valores de la variable.

PROPIEDADES DE LA DISTRIBUCIÓN NORMAL

En una distribución normal se cumple la siguiente relación:

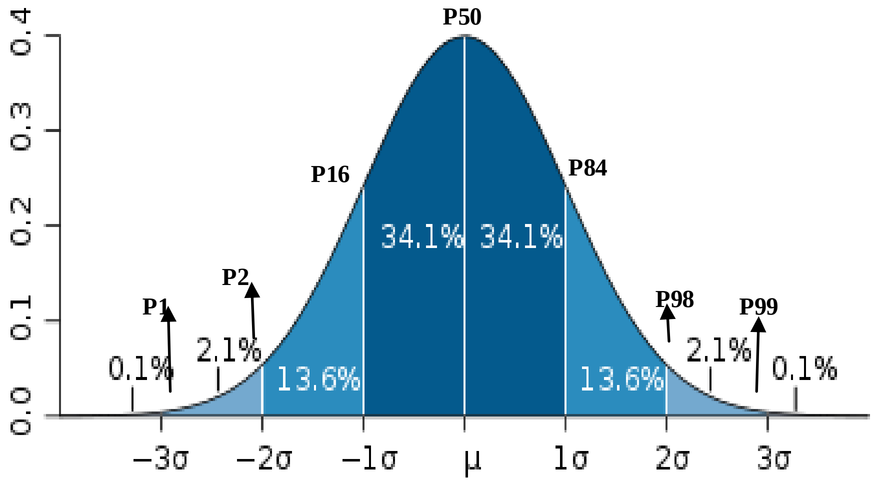
Si a partir del punto que corresponde al valor medio se llevan hacia ambos lados los valores σ , 2σ y 3σ quedan determinados los siguientes intervalos:

$\left[\overline{x}-\sigma;\overline{x}+\sigma\right]$ Contiene el 68% [3] de las observaciones.

$\left[\overline{x}-2\sigma;\overline{x}+2\sigma\right]$ Contiene el 95% de las observaciones.

$\left[\overline{x}-3\sigma;\overline{x}+3\sigma\right]$ Contiene el 99% de las observaciones.

Gráficamente tendremos:



[3] Los valores exacto de los porcentajes es: 68,2% , 95,4% y 99,8%.

Ejemplo:

Consideremos la variable altura de una determinada edad escolar de una ciudad. Una vez medidas las alturas de una muestra de niños y sistematizada la información se obtienen los siguientes valores:

$\bar{x}$ = 1,30m
Mo = 1,29m
Me = 1,29m
σ = 0,12

La distribución de frecuencias de la variable altura es aproximadamente simétrica y se pueden aplicar las propiedades anteriores, obteniendo las siguientes conclusiones:

- El 68% de los mismos medirá entre:

$$[\bar{x} - \sigma]; [\bar{x} + \sigma]$$
$$[1{,}30 - 0{,}12]; [1{,}30 + 0{,}12]$$
$$[1{,}18; 1{,}42]$$

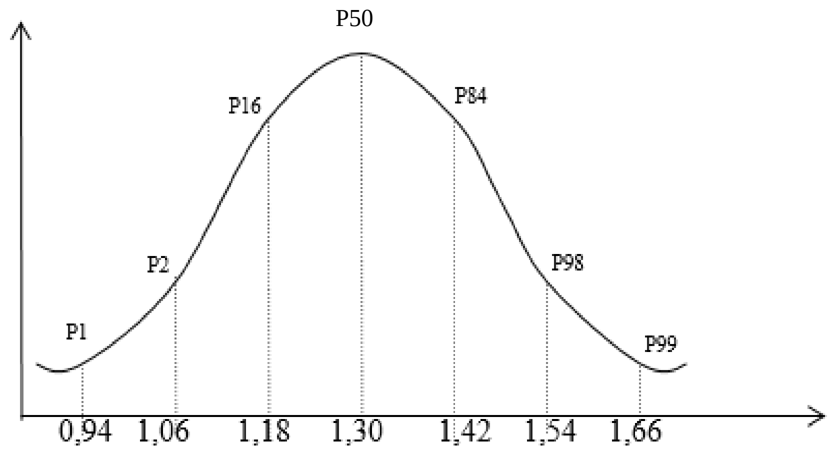
- El 95% de los niños tendrá una altura comprendida en el intervalo :

$$[\bar{x} - 2\sigma]; [\bar{x} + 2\sigma]$$
$$[1{,}06; 1{,}54]$$

- El 99% de los niños tendrá una altura comprendida en el intervalo:

$$[\bar{x} - 3\sigma]; [\bar{x} + 3\sigma]$$
$$[0{,}94; 1{,}66]$$

Gráfico Indicando los valores obtenidos. (Teniendo en cuenta el gráfico q tienen en el apunte les quedará de la siguiente manera)



SIN AGRUPAMIENTOS DE DATOS	CON AGRUPAMIENTO DE DATOS
MEDIANA (Me)	**MEDIANA (Me)**
$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$	• **Serie de frecuencias (Variable Discreta).** $M_e = L_i + \frac{\frac{n}{2} - f_{a(i-1)}}{f_{ai}} h_i$
PROMEDIO	**PROMEDIO**
$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + ... + x_n}{n}$	• **Promedio de una serie de frecuencias** $\bar{x} = \frac{\sum f_i . x_i}{n}$ • **Promedio de una distribución de frecuencias en intervalos de clase** $\bar{x} = \frac{\sum f_i . x_{im}}{n}$
DESVIACIÓN MEDIA	**DESVIACIÓN MEDIA**
$D.m = \frac{\sum \lvert x_i - \bar{x} \rvert . f_i}{n}$	$D.m = \frac{\sum \lvert x_{mi} - \bar{x} \rvert . f_i}{n}$
DESVIACIÓN ESTANDAR	**DESVIACIÓN ESTÁNDAR**
$\sigma = \sqrt{\frac{\sum (xi - \bar{x})^2 . fi}{n}}$	$\sigma = \sqrt{\frac{\sum (xm - \bar{x})^2 . fi}{n}}$
VARIANZA	**VARIANZA**
$V = \sigma^2 = \frac{\sum (x_i - \bar{x})^2 . f_i}{n}$	$V = \sigma^2 = \frac{\sum (x_{mi} - \bar{x})^2 . f_i}{n}$
COEFICIENTE DE VARIACIÓN	
CV= $\frac{\sigma}{\bar{X}}$	