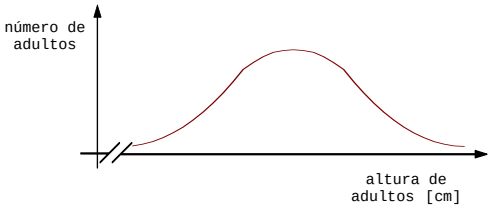
MEDIDAS DE POSICION Y DISPERSION

Existen distribuciones de distintas formas.

- muchas se caracterizan por ser "simétricas" y con un sólo "pico" y muestran una forma de campana con su altura máxima para los valores centrales de la variable disminuyendo gradualmente hacia los extremos, donde se observa un número reducido de valores.

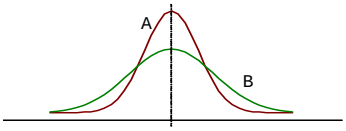
Ejemplo:

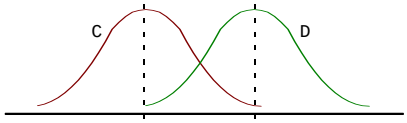
distribución del número de adultos de acuerdo con sus alturas.



- otras distribuciones son "asimétricas" ya sea hacia la derecha o hacia la izquierda, otras presentan dos o más "picos" (son bimodales o plurimodales)

Supongamos tener cuatro grupos de individuos y las distribuciones de sus alturas.





Observando los diagramas podemos decir que:

- Los grupos A y B presentan la misma altura "media" pero las alturas del grupo B se encuentran mas dispersas.(O lo que es lo mismo, menos concentradas)
- Los grupos C y D tienen igual dispersión pero se "ubican" de distinta manera, es decir, las concentraciones de datos se corresponden con distintos valores de la variable.

A partir de lo visto, encontramos y definimos **"medidas de POSICIÓN y de DISPERSIÓN"**

La medida de posición se emplea para ubicar el **centro** de un conjunto de observaciones. Pero esto no es suficiente, pues los datos pueden estar diseminados, dispersos de cierta forma a ambos lados de ese **centro**. Esta diseminación es generalmente llamada **dispersión** o **variación**.

Enumeraremos

- **medidas de posición**
  - media aritmética
  - mediana
  - modo
  - cuartilos, decilos, centilos
  - semirango o rango medio
- **medidas de dispersión**
  - rango, recorrido o amplitud
  - varianza
  - desvío estándar
  - coeficiente de variación
  - desviación intercuartílica

MEDIA ARITMÉTICA

Es el valor que tomaría la variable si estuviese uniformemente repartida entre todos los individuos que forman la muestra; se corresponde con el concepto de centro de gravedad en mecánica.

Se calcula simplemente sumando todos los valores de las observaciones y dividiendo por el número de datos considerados.

Si se tiene una muestra de tamaño n es **estimador**.

M(x) (se lee media de x (operador medio))

$$M(x) = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{X}$$

(se lee **x barra** o **x raya** o **x media**)

Si se tiene una población de tamaño N es **parámetro**

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

(se lee "**mu**")

Si se tienen los datos agrupados en intervalos o por valores distintos de la variable, sea ésta discreta o continua, se calcula la media aritmética ponderada.

$$\overline{X} = \frac{1}{\sum_{i=1}^{k} f_i} \cdot \sum_{i=1}^{k} x_i' \cdot f_i$$

(muestral)

k : número de clases o categorías

$$\sum_{i=1}^{k} f_i = n$$

$$\mu = \frac{1}{\sum_{i=1}^{k} f_i} \cdot \sum_{i=1}^{k} x_i' \cdot f_i$$

(poblacional)

$$\sum_{i=1}^{k} f_i = N$$

MEDIANA

Es el valor central de la distribución, o sea, el valor del carácter a ambos lados del cual se reparten por mitades las observaciones.

MODO

Es el valor del carácter que se presenta con mayor frecuencia en la muestra o población, o sea, aquel al que le corresponde el mayor número de observaciones.

CUARTILOS, DECILOS, CENTILOS

Con el mismo sentido que la mediana, son los valores que dividen alas observaciones por cuartos, décimos y centésimos, respectivamente.(son 3, 9 y 99, respectivamente)

SEMI-RANGO

Es el valor "central" de los valores distintos de la variable, entre el valor mínimo y el valor máximo.

$$\bar{R} = \frac{X_M + x_m}{2}$$

A continuación se enuncian las

propiedades de la media aritmética.

PROP.1.- La suma de los desvíos entre cualquier valor observado y el promedio es cero, cualquiera sea la distribución.
Es decir

PROP.2.- La suma de los cuadrados de los desvíos entre cada valor observado y el promedio es menor o igual que la suma de los cuadrados de los desvíos con respecto a cualquier otro valor.

PROP.3.- El promedio de un grupo de medias aritméticas, cada una de ellas ponderada por la cantidad de observaciones que le dio origen, coincide en el promedio de las observaciones individuales

PROP.4.- La media aritmética de una constante es la misma constante.

PROP.5.- La media aritmética de una variable mas (o menos) una constante es igual a la media aritmética de la variable mas (o menos) la constante.

PROP.6.- La media aritmética de una variable por una constante es igual a la media aritmética de la variable multiplicada por la constante.

PROP.7.- La media aritmética de una variable que es suma (o resta) de otras variables originales es igual a la suma (o resta) de las medias aritméticas de las variables originales.

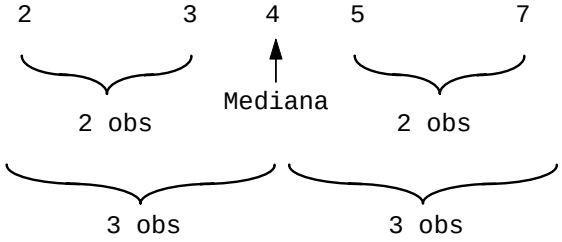
EJEMPLOS

(8)....a) Supongamos que se tiene una variable que toma los siguientes valores: 3, 5, 7, 4, 2 Se tiene entonces que el **tamaño de la población** es N = 5 . La **MEDIA ARITMÉTICA** es:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{3+5+7+4+2}{5} = 4{,}2$$

Es decir, 4,2 es la media aritmética de los cinco valores de la variable.

Para calcular la MEDIANA se debe, en primer lugar, ordenar los valores a fin de encontrar aquella observación que deja el mismo número de observaciones a su izquierda que a su derecha. En nuestro caso se tiene:



La mediana es cuatro (4) y es el valor de la variable que deja dos observaciones tanto a su derecha como a su izquierda.

Es el valor de la variable que ocupa la posición central, en este caso la tercera posición. Notemos que esencialmente la mediana "divide" los datos en dos subconjuntos.

...¿Cómo determinar la posición de la mediana o el número de orden que ella ocupa?

Calculando lo que llamamos "**mediana de orden**":

M°=(N+1)/2,

donde 1 es la posición que ocupa el dato con el valor mínimo, y N es la posición que ocupa el dato con el valor máximo. Evidentemente la posición central está bien calculada.

En este ejemplo la mediana ocupa el tercer lugar, esto es **"el orden de la mediana"** es 3. O dicho de otra manera, "la mediana de orden" es 3 y se denota $Mna^0 = 3$

¿Cómo lo calculamos?

Haciendo $\frac{N+1}{2}$ en este caso $M^0 = 3$

y $\Rightarrow$ es $Mna = x_{(3)} = 4$

Al ser N un número impar será exactamente la mediana el valor central del conjunto. Para encontrarlo se puede comenzar a contar a partir del primero o del último de los datos ordenados.(Conviene contar a partir de los dos extremos, para verificar)

Recuerde que:

Cuando el subíndice se encuentra entre paréntesis se trata de valores ordenados de la variable.

Observando las frecuencias con que se dan los valores de la variable se ve que no existe **MODO.**

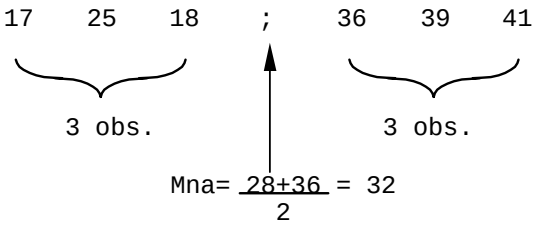
....b) Supóngase que se tiene una **muestra de tamaño "n = 6"** de una población y que los valores observados de la variable aleatoria son: 41 , 36 , 17 , 28 , 25 , 39

Se tiene la **media aritmética** que es:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{41+36+17+28+25+39}{6} = 31$$

La **mediana** no será ahora un valor observado de la variable. Es evidente que la posición de la mediana no es un entero si n es par.

Si ordenamos tendremos:



Mna = 32 es un valor no observado de la variable que separa las observaciones ordenadas dejando tres valores a su izquierda y tres a su derecha.

La mediana de orden en este caso es $\frac{n+1}{2} = \frac{7}{2} = 3,5$ lo que indica que la mediana se encuentra entre los datos ordenados tercero y cuarto, justamente en la mitad. Esto hace que se considere el

valor de la mediana como la semi-suma entre la tercera y la cuarta observación:

$Mna^0 = 3,5 \quad \Rightarrow \quad Mna = \frac{x_{(3)} + x_{(4)}}{2} = 32$

El **modo,** al igual que en el ejemplo anterior, no existe.

- **Modo**:

Al tener la variable agrupada en intervalos se halla primeramente el intervalo modal, es decir, el intervalo de mayor frecuencia absoluta.

Si se quiere hallar un valor dentro de tal intervalo como valor modal se utiliza la fórmula:

$$M_0 = L_i + \left[\frac{f_{M_0} - f_{M_0-1}}{\left(f_{M_0} - f_{M_0-1}\right) + \left(f_{M_0} - f_{M_0+1}\right)} \right].c$$

Donde: L_i límite inferior del intervalo modal
f_{M_0} frecuencia absoluta del intervalo modal
f_{M_0-1} frecuencia absoluta del intervalo anterior al modal
f_{M_0+1} frecuencia absoluta del intervalo posterior al modal
c amplitud de los intervalos

Si se representa la distribución de frecuencias absolutas en escala es posible gráficamente (y en forma aproximada) hallar este valor.

Para hallar el modo gráficamente se ubica el intervalo modal y luego...

1. Se traza un segmento desde la altura del extremo derecho del intervalo anterior al modal a la altura del extremo derecho del modal (segmento ab).
2. Se traza un segmento desde la altura del extremo izquierdo del intervalo posterior al modal a la altura del extremo izquierdo del intervalo modal (segmento cd).
3. Se traza una perpendicular por la intersección de dichos segmentos al eje de abscisas.
4. El punto de intersección de la perpendicular con el eje de la variable es el valor del modo.

- **Mediana:**

En forma similar a la usada para hallar el modo se debe hallar primeramente el intervalo mediana hasta el cual se acumula, por lo menos, la mitad de las observaciones.

En este caso

$$Mna^0 = \frac{n+1}{2} \qquad Mna^0 =$$

Para hallar el valor de la mediana dentro de dicho intervalo se aplica la fórmula:

$$Mna^{0}$$

$$Mna = L_i + \left[\frac{\left[\left(\frac{n+1}{2} \right) - F_{i-1} \right]}{\left(F_i - F_{i-1} \right)} \right].c$$

Donde: L_i límite inferior del intervalo mediana
F_i frecuencia acumulada hasta el intervalo mediana
F_{i-1} frecuencia acumulada hasta el intervalo anterior a la mediana
c amplitud del intervalo

Dado que al trabajar con el concepto de mediana se utilizan las frecuencias acumuladas, si se quiere representar la mediana gráficamente, se debe utilizar igualmente el polígono de frecuencias acumuladas u ojivas.

Para hallar la mediana por el modo gráfico:

1. Se traza una paralela al eje horizontal por la mediana de orden hasta intersectarse con la ojiva.
2. Desde ese punto de intersección se traza una perpendicular y donde ésta corta al eje horizontal se encuentra la mediana.

MEDIDAS DE DISPERSION

Se ha visto que las medidas de posición informan parcialmente sobre la distribución de la información. Falta aún encontrar e interpretar medidas que informen sobre la variación o variabilidad de los datos.

No es fácil interpretar y/o explicar el concepto de medida de dispersión, pero intuitivamente puede reconocerse que sería bueno que se lograra una medida que indicara que la dispersión es menor cuando los datos están más concentrados alrededor de la media aritmética, por ejemplo.

RECORRIDO, RANGO O AMPLITUD

Es la diferencia entre la mayor y la menor de las observaciones.

$$\boxed{R = x_{maximo} - x_{mínimo} = x_M - x_m}$$

Es la más sencilla y directa medida de dispersión.

No proporciona una medida de la variabilidad con respecto a alguna medida de posición, ni informa sobre la ubicación o dispersión de los datos, ya que sólo se calcula en base a los dos valores extremos.

VARIANCIA O VARIANZA

Se define como media de los cuadrados de los desvíos con respecto a la media aritmética de la variable

Si la varianza de un grupo de datos es grande, éste grupo tiene mayor variabilidad absoluta que otro grupo con varianza más pequeña.

Cuando el interés radica en estimar a la varianza de la población a partir de un valor obtenido de la muestra, se calcula la varianza con "n-1" como denominador.

Está demostrado que si en lugar de " n " se consideran "n-1" (grados de libertad), la estimación es insesgada.

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

que sí es un buen estimador de la varianza poblacional.
A la varianza poblacional la identificamos como

σ^2 (sigma cuadrado) y es:

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

La cantidad (n - 1) es llamada **grados de libertad** (g.l.) y daremos sólo una interpretación intuitiva de su significado.
En

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

el numerador representa una sumatoria de n cuadrados de desvíos con respecto a la media aritmética, pero es sabido (por propiedad de la media) que la sumatoria de los desvíos con respecto a la media aritmética es cero, esto es $\sum(x - \bar{x}) = 0$ y si esto es así, podremos dar valor libremente a "n-1" de los desvíos puesto que el n-ésimo queda determinado por los anteriores.
Esto es, se tienen sólo "n-1" desvíos independientes.
Se deben mencionar las desventajas de la varianza:

- Para un sólo conjunto de datos no proporciona ayuda inmediata puesto que no puede interpretarse en términos del problema (si S^2= 36 ¿qué significa esto? ¿cuáles serán sus unidades?).
- Es cierto que varianzas pequeñas indican variabilidades pequeñas, pero si se tiene un sólo grupo, ¿indica pequeña variabilidad? ¿qué es pequeña? ¿cuánto es?
- Será necesario encontrar otra medida de dispersión que permite solucionar, aunque sea en parte, estos problemas.

DESVÍO ESTANDAR O DESVIACIÓN TÍPICA

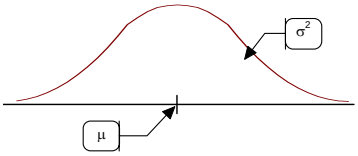
Es la raíz cuadrada de la varianza. Se toma en consideración sólo la raíz positiva, porque se la emplea como medida.

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

Con este desvío estándar se soluciona el inconveniente de interpretar en términos del problema, puesto que S_n sí está expresado en las mismas unidades de medida que la variable original.

Para tener un significado cuantitativo de la magnitud de esta medida de dispersión será necesario referirse a un tipo de distribución particular que se denomina **población o distribución normal.**

Son más altas en la parte central y bajan con bastante rapidez en sus extremos.



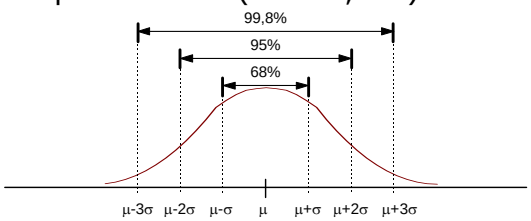
Tienen una forma muy parecida a la de una campana.

Una distribución teórica con estas características es la curva Normal ó de Gauss ó de Laplace, que está centrada en μ, con una varianza σ^2.

La altura de los hombres de raza blanca, por ejemplo, sigue una distribución normal (la mayoría tienen alturas lógicas; hay pocos enanos y pocos gigantes).

Como bajo una curva de una población Normal se encuentran todas las observaciones de esa población, todas las frecuencias, el área total bajo la curva abarca el 100% de esas observaciones.

Puede demostrarse que entre el promedio μ y una vez la desviación estándar en más y en menos, es decir entre (μ - σ) y (μ + σ), se encuentra alrededor del 68% de la población; entre μ - 2σ y μ + 2σ se encuentra cerca del 95% de la población; y entre μ - 3σ y μ + 3σ casi la totalidad de la población (el 99,8%).



Cuando se tienen muestras, con n suficientemente grande, que hayan sido seleccionadas en forma aleatoria de una población normal o casi normal, el polígono de frecuencias que se obtiene tiene características similares a las de la distribución normal (más similares en tanto n sea mas grande y los intervalos considerados mas chicos a fin de que la línea del polígono se "suavice")

Es entonces aceptable decir, cuando la muestra es grande, que alrededor del 68% de las observaciones están comprendidas entre $\bar{x} - S$ y $\bar{x} + S$, el 95% entre $\bar{x} - 2.S$ y $\bar{x} + 2.S$ y casi todas las observaciones entre $\bar{x} - 3.S$ y $\bar{x} + 3.S$.

Esta posibilidad de interpretar la desviación estándar y expresar resultados concretos si bien aproximados de ubicación de determinados porcientos de las observaciones alrededor del promedio es válido.

Por esto es la **medida de dispersión más usada.**

COEFICIENTE DE VARIACIÓN

Se calcula el **coeficiente de variación** que mide el porcentaje de la variabilidad relativa al promedio, haciendo:

$$C.V. = \frac{S_m}{\bar{x}}.100$$

Esta medida tiene la ventaja de ser independiente de las unidades de medida y por lo tanto resulta útil para comparar variabilidades de poblaciones o muestras diferentes, en las que se ha trabajado con distintas unidades de medida.

Las Propiedades De La Varianza (Var(X)) Son:

Prop.1.- La varianza es una cantidad no negativa Var(x) = 0, puesto que se trata de una sumatoria de números positivos porque son cuadrados.

Prop.2.- La varianza de una constante es cero.

Var(c) = 0 (c es constante)

En este caso se tendrían todos los valores de la variable iguales a ese valor c, el cual sería también el valor del promedio; todos los desvíos serían ceros y en consecuencia la Varianza será cero.

Prop.3.- Si a una variable se le suma una constante la varianza no cambia, no se altera.

$$\mathbf{Var(x + c) = Var(x)}$$

El efecto de sumar (o restar) a cada valor de la variable una constante es que toda la distribución se traslada hacia la derecha (o izquierda si la constante se resta) una distancia igual a la constante pero sin cambiar la dispersión.

Puede también pensarse en el sentido de que como la varianza de una suma o diferencia de variables es suma de variabilidades (las varianzas **siempre** se suman, se agregan) será

$$\mathbf{Var(x + c) = Var(x) + Var\ c) = Var\ (x) + 0 = Var(x)}$$

Prop.4.- Si los valores de la variable se multiplican (o dividen) por una constante, cambia la dispersión puesto que al multiplicar el producto depende del valor de la variable a considerar. Los desvíos se verán afectados y como estos se elevan al cuadrado, aparecerá la constante al cuadrado

$$\mathbf{Var(x\ .\ c) = c^2 Var(x)}$$

Fórmula de trabajo de la varianza

Si se tiene un número grande de observaciones muestrales será difícil (si es que no se tiene computadora o calculadora para hacerlo) calcular S_m^2 construyendo los desvíos y usando la expresión de la fórmula teórica dada.

Se trabaja a partir de la fórmula y se llega a una expresión que recibe el nombre de fórmula de trabajo, pues resulta sumamente fácil y rápido hacer los cálculos y resolver con unas pocas operaciones.

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

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

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

$$= \frac{1}{n-1}\left(\sum x_i^2 - \frac{2.(\sum x_i)^2}{n} + \frac{n.(\sum x_i)^2}{n^2}\right) =$$

$$S^2 = \frac{1}{n-1}.\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right]$$

$$S^2 = \frac{1}{n-1}\left[\sum x_i^2 - n.\bar{x}^2\right]$$

Si se trabaja con datos agrupados se parte de

$$S^2 = \frac{1}{\sum_k f_i} - - - \sum^k (x_i - \bar{x})^2 . f_i$$

y se llega a

$$S^2 = \frac{1}{\sum_k f_i} - - - \left[\sum x_i^2 . f_i - \frac{\left(\sum x_i . f_i\right)^2}{n}\right]$$

o bien

$$S^2 = \frac{1}{\sum f_i} - - - \left[\sum x_i^2 . f_i - n\bar{x}^2\right]$$

DIAGRAMA DE CAJA

El profesor J.W.Tukey ha estudiado métodos estadísticos para que los "dueños" de los datos aprendan a conocer la información que de esos datos se desprende sin necesidad de tener gran base matemática. Pues, cuanto más entiende el "dueño" de sus datos, mejor es el trabajo del estadístico. Propone el llamado **Box-Plot** o **diagrama de caja**, para el cual preciso calcular Q_1, Mna y Q_3.

El diagrama de caja (box plot) consiste en una caja cuyos bordes inferior y superior son los cuartiles 1º y 3º y la línea central representa la mediana. Los bigotes desde la caja indican el rango de los datos. La longitud de la caja es la distancia entre el primer y el tercer cuartil, de forma que la caja contiene los datos centrales de la distribución. La línea dentro de la caja señala la posición de la mediana. Si esta cae cerca del final de la caja, indica la presencia de asimetría. Las líneas que se extienden desde cada caja (llamadas bigotes) representan la distancia entre la mayor y la menor de las observaciones.