MEDIDAS ESTADÍSTICAS

Generalidades

Las medidas estadísticas son herramientas fundamentales para analizar y comprender datos.

Cuando se busca resumir más la información de una tabla o de una gráfica, y de encontrar algunos valores lo más simples posible que nos permitan dar información sobre la muestra o comparar dos muestras entre sí, se recurre a las medidas estadísticas .

Para hacer ese resumen o información de los datos hay tres enfoques fundamentales:

- En primer lugar, dar un valor lo más representativo posible de todos los valores de la muestra, que no sea, por tanto, ni de los más bajos ni de los más altos. Así se crean las medidas o parámetros de centralización, tendencia central o posición central.
- En segundo lugar, y como complemento a lo anterior, dar una valoración de hasta qué punto los datos se parecen entre sí o bien están muy diferenciados (dispersos); además, cuanto más se parezcan entre sí los valores que nos salen, más se parecerán al parámetro de centralización que elijamos, y mejor sería éste. Por todo esto conviene medir las diferencias internas de los datos mediante las medidas o parámetros de dispersión.
- En tercer lugar, se puede también tratar de medir qué valor supera a una cierta porción o proporción de valores, o lo que es lo mismo, tratar de informar sobre la distribución de la variable diciendo a cuántos de sus valores supera uno dado. Para ello se usan los cuantiles como medidas o parámetros de posición.

Los procedimientos para obtener las medidas estadísticas difieren levemente dependiendo de la forma en que se encuentren los datos.

1. Si los datos se encuentran ordenados en una tabla estadística diremos que se encuentran "agrupados".

2. Si los datos no están en una tabla hablaremos de datos "no agrupados".

Aplicaciones

Toma de decisiones. Las medidas estadísticas son herramientas fundamentales para analizar y comprender datos. En este conjunto de secciones, exploraremos las principales medidas de tendencia central: media, mediana y moda, aprendiendo sus definiciones, fórmulas y aplicaciones prácticas.

Resumen de datos. Permiten resumir grandes conjuntos de datos en valores clave que facilitan su interpretación y comprensión.

Identificación de tendencias. Revelan patrones y tendencias en los datos, lo que permite detectar oportunidades y desafíos.

Tipos de medidas estadísticas

Se distinguirán tres tipos de medidas: medidas de posición, medidas de dispersión y medidas de forma. A continuación, se describen cada una de estas.

1. Medidas de posición de tendencia central

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: media, mediana y moda. Estas medidas nos proporcionan valores alrededor de los cuales se distribuyen los datos observados en la muestra, tal como se muestra en la siguiente figura.

1.1. Media aritmética ($\overline{X}$)

Es el centro de gravedad de la distribución entre todos los datos. Podría definirse como el promedio aritmético de los valores de una variable cuantitativa, y se calcula sumando los valores de la variable y dividendo su resultado por la cantidad de datos presentes, es decir:

Propiedades de la media aritmética:

1. Es la medida de localización más usada.
2. El cálculo de la media no exige la ordenación de la variable cuantitativa.
3. Su principal desventaja es que se ve afectada por los valores extremos o atípicos de la variable (puede decirse que la media no es resistente o robusta).
4. Un conjunto de datos sólo tiene una media. La media es única.
5. La media es una medida útil para comparar dos o más poblaciones (distribuidas normalmente).
6. La media aritmética es la única medida de posición en la que las sumas de las desviaciones de los valores de la media es siempre cero: $\sum(X - \bar{X}) = 0$

Existen otras medidas de centralización de uso menos frecuente, como la media ponderada (que es una media aritmética con distintos pesos de importancia para los distintos datos), la media geométrica (raíz enésima del producto de los datos) o la media armónica (la inversa de la media aritmética de los inversos de los datos). Este tipo de medias no serán tratadas en el curso, sin embargo, los invito a consultar sobre ellas.

El cálculo de la media puede hacerse, como en cualquier otra medida estadística, de manera manual o computarizada. En el caso de realizarse los cálculos manualmente se utilizan las siguientes fórmulas:

Para datos no agrupados	Para datos agrupados o en intervalos
$\overline{\mathrm{X}} = \frac{\sum \mathrm{X_i}}{\mathrm{n}}$	$\overline{\mathrm{X}} = \frac{\sum \mathrm{f_i}.\,\mathrm{Xi}}{\mathrm{n}}$ $\mathrm{fi} = \text{frecuencia absoluta}$ $\mathrm{Xi} = \text{punto medio}$

Cuando utilizar la media aritmética como promedio:

1. Cuando la distribución sea simétrica o aproximadamente simétrica.
2. Cuando se quiera hacer un análisis inferencial o se requieren otros estadísticos complementarios como la desviación estándar o el coeficiente de correlación.

3. Cuando las escalas de los datos sean de intervalo o de razón y no sea recomendable otro promedio.
4. Cuando la distribución de los datos sea uniforme.

Otras medias

Media ponderada

$$\bar{X}_p = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

En una empresa en la que hay 80 empleados, 60 ganan $10 por hora y 20 ganan $13 por hora. Determine el sueldo medio ponderado de los empleados de la empresa.

$$\bar{X}_p = \frac{60 * 10 + 20 * 13}{60 + 20} = 10{,}75\$$$

El sueldo medio ponderado de los empleados es de aproximadamente $10,75 por hora.

Media geométrica

Datos no agrupados

$$G = \sqrt[n]{x_1.x_2.x_3 \ldots x_n} = \sqrt[n]{\pi x_i}$$

$$\log G = \frac{\sum log x_i}{n}$$

Datos agrupados

$$G = \sqrt[n]{\pi x_i^{fi}}$$

$$\log G = \frac{\sum f_i log x_i}{n}$$

Puede hacerse uso del Excel mediante la función

= MEDIA.GEOM(número1; [número2]; …)

Si el crecimiento de las ventas en un negocio fue en los tres últimos años de 26%,32% y 28%, hallar la media anual del crecimiento.

Método 1:

$$\mathrm{G} = \sqrt[n]{\mathrm{x_1.x_2.x_3 \ldots x_n}} = \sqrt[3]{0{,}26 \text{ x } 0{,}32 \text{ x } 0{,}28} = 0{,}286$$

$$= 28{,}6\% \text{ es la media anual de crecimiento}$$

Método 2:

$$\log \mathrm{G} = \frac{\sum \log \mathrm{x_i}}{\mathrm{n}} = \frac{\log 26 + \log 32 + \log 28}{3} = 1{,}456$$

$$\mathrm{G} = \text{antilog} 1{,}456 = 28{,}6\%$$

En la siguiente tabla se muestra la cantidad de muestra de la cantidad de accidentes laborales registrados durante el primer semestre del año 2021 en la empresa J/A.

$\mathbf{x_i}$	$\mathbf{f_i}$	$\mathbf{x_i^{fi}}$	$\mathbf{log x_i}$	$\mathbf{f_i * log x_i}$
1	4	1	0	0
2	5	32	0,3010	1,505
3	7	2.187	0,4771	3,3397
4	8	65.536	0,6021	4,8168
6	6	46.656	0,7782	4,6692
Σ	**30**			**14,3307**

Método 1

$$\mathrm{G} = \sqrt[n]{\pi \mathrm{x_i^{fi}}} = \sqrt[30]{1 * 32 * 2187 * 65536 * 45656} = 3$$

Método 2

$$\log \mathrm{G} = \frac{\sum \mathrm{f_i log x_i}}{\mathrm{n}} = \frac{14{,}3307}{30} \qquad \log \mathrm{G} = 0{,}4777 = 3$$

En promedio se han registrado durante el primer semestre del año 2021 una cantidad de 3 accidentes laborales.

Media armónica

Datos no agrupados

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_1}}$$

Datos agrupados

$$H = \frac{n}{\sum_{i=1}^{n} {f_i}/{x_i}}$$

En Excel mediante la función:

= MEDIA.ARMO(número1; [número2]; …)

Encuentre la media armónica de la cantidad de artículos defectuosos de una máquina que arrojó los siguientes datos: 2, 3, 3, 7, 6

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_1}} = \frac{5}{\frac{1}{2}+\frac{1}{3}+\frac{1}{3}+\frac{1}{7}+\frac{1}{6}} = 3{,}387\ artículos$$

Se concluye que la maquina está arrojando una media armónica 3,387 artículos defectuosos.

La siguiente tabla de distribuciones de frecuencia registra las longitudes en centímetros que en una semana tienen 100 plantas de frijol; con esta información obtener la media armónica.

$Li - Ls$	x_i	f_i	f_i/x_i
5,4 - 5,7	5,6	7	1,261
5,9 - 6,1	6,0	16	2,689
6,2 - 6,5	6,4	21	3,307
6,6 - 6,9	6,8	29	4,296
7,0 - 7,3	7,2	18	2,517
7,4 - 7,7	7,6	9	1,192
Σ		**100**	**15,263**

$$H = \frac{n}{\sum_{i=1}^{n} {f_i}/{x_i}} = \frac{100}{15{,}263} = 6{,}55\ cm$$

La media armónica de la longitud de las 100 plantas de frijol es de 6,55 cm.

Media recortada

Consiste en eliminar un porcentaje de los datos originales. Se eliminan los valores extremos.

Se hicieron las mediciones de resistencia a la tensión en MPa de 12 cauchos a 20°C: 2,07 – 2,14 – 2,22 – 2,03 – 2,21 – 2,03 – 2,05 – 2,18 – 2,09 – 2,14 – 2,11 – 2,02. Calcule la media recortada al 10%.

Primero se ordenan los datos en forma ascendente:

2,02 – 2,03 – 2,03 – 2,05 – 2,07 – 2,09 – 2,11 – 2,14 – 2,14 – 2,18 – 2,21 – 2,22

Luego se elimina 10% de los datos menores y el 10% de los datos mayores, es decir, el 10% de 12 es igual 1,2; significa que descartamos el 2,03 y 2,22, los cuales son los datos extremos.

Con el restante de observaciones calculamos la media aritmética:

$$\bar{x}_{rec} = \frac{2{,}03 + 2{,}03 + 2{,}05 + 2{,}07 + 2{,}09 + 2{,}11 + 2{,}14 + 2{,}14 + 2{,}18 + 2{,}21}{10}$$

$$2{,}105\ MPa$$

1.2. Mediana (Me)

Es el valor que está en el centro de la distribución, es decir, el valor que supera a la mitad de los de la muestra y se ve superado por la otra mitad, es decir, la mitad de las observaciones de la variable cuantitativa está por debajo del 50% y la otra mitad por encima del 50%.

Cuando se realiza el cálculo de la mediana en forma manual, lo más recomendable es ordenar el conjunto de datos en forma creciente.

Cuando *n* es impar, la mediana (Me) es el valor central, y cuando *n* es par la mediana es el promedio de los dos valores centrales, es decir:

$$\text{Me} = X_{\left(\frac{n+1}{2}\right)} \quad \text{n es impar} \qquad \text{Me} = \frac{X_{\frac{n}{2}} + X_{\left(\frac{n}{2}\right)+1}}{2} \quad \text{n es par}$$

En Excel podemos hacer uso de la fórmula: **=MEDIANA(___:____)**

Propiedades de la mediana:

1. Es única, sólo existe una mediana para un conjunto de datos cuantitativos.
2. Al calcular la mediana no usamos todos los valores observados en la variable, lo que la limita como medida de tendencia central.
3. La mediana no es sensible a valores extremos., es robusta.
4. Es la medida más representativa en el caso de variables que solo admitan la escala ordinal.
5. La mediana no cambia por mucho cuando se incluyen solo unos pocos valores extremos, por lo que la mediana es una medida de tendencia central resistente.
6. No puede ser aplicada a distribuciones de variables cualitativas.
7. Al representar el 50% de los datos es equivalente a decir que es el cuartil 2.
8. Puede obtenerse para datos de nivel de razón, de intervalo y ordinal (excepto para el nominal).

En caso de que los datos estén agrupados o en intervalos se utilizan las siguientes fórmulas:

$$\text{Me} = \text{Li} + \frac{\frac{n}{2} - F_{a-1}}{fi} * \text{IC} \quad \text{cuando n es par}$$

$$\text{Me} = \text{Li} + \frac{\frac{n+1}{2} - F_{a-1}}{fi} * \text{IC} \quad \text{cuando n es impar}$$

Li = Límite inferior de la clase medianal

F_{a-1} = Frecuencia acumulada anterior a la clase medianal

f_i = Frecuencia simple de la clase de la medianal

IC = intervalo de la clase medianal

1.3. Moda (Mo)

Es el valor de la variable que tiene mayor frecuencia en la muestra, es decir, el que se repite más (moda se asocia con lo más frecuente) dentro de una serie de datos, y es la única medida estadística que puede ser empleada para variables cualitativas y cuantitativas.

Por ejemplo, supongamos que tenemos las calificaciones de un test de 11 personas: 56, 49, 73, 55, 55, 64, 66, 77, 50, 70, 55. Para calcular la moda ordenamos la serie de datos: 49, 50, 55, 55, 55, 56, 64, 66, 70, 73, 77. Notamos que el la calificación 55 es la que más se repite, por lo tanto, es la moda.

En Excel podemos hacer uso de la fórmula: **=MODA(___:___)**

Propiedades de la moda

1. La moda puede determinarse para cualquier conjunto de datos.
2. Al igual que la mediana no se ve afectada por la presencia de valores extremos.
3. Puede ser determinada para categorías con intervalos abiertos.
4. En su determinación no intervienen todos los valores de la distribución.
5. En una serie de datos pueden existir una o varias modas: si existe una moda se llama unimodal, o simplemente modal, si hay dos se denomina bimodal, y si hay más dos se llamara multimodal. Incluso, no tener ningún valor que se repita.

Para el cálculo de la moda puede darse dos casos:

1. Para la variable cualitativa o numérica discreta: Su cálculo es sumamente sencillo, pues basta hallar en la tabla de frecuencias el valor de la variable que presenta la frecuencia máxima.
2. Cuando la variable está agrupada en intervalos de clases (intervalos), la moda se encontrará en la clase de mayor frecuencia, pudiendo calcular su valor por medio de la expresión:

$$\mathrm{Mo} = \mathrm{Li} + \frac{\mathrm{d}_1}{\mathrm{d}_1 + \mathrm{d}_2} * \mathrm{IC}$$

Li = límite inferir de la clase modal

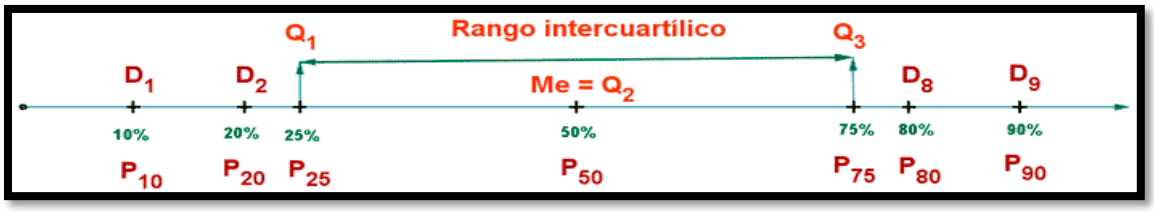
d_1 = diferencia entre la frecuencia de la clase modal y la frecuencia que antecede a la clase modal , $d_1 = f_m - f_{m-1}$

d_2 = diferencia entre la frecuencia de la clase modal y la frecuencia que precede a la clase modal , $d_1 = f_m - f_{m+1}$

IC = Intervalo de clase de la clase modal

2. Medidas de posición de tendencia no central (Cuantiles)

Las medidas de posición no central (o medidas de tendencia no central) permiten conocer puntos característicos de una serie de valores, que no necesariamente tienen que ser centrales. La intención de estas medidas es dividir el conjunto de observaciones en grupos con el mismo número de valores. Como medidas de posición de tendencia no central, están: cuartiles, deciles y percentiles, según se muestra en la siguiente figura.



2.1. Cuartiles (Q_i)

Se entiende por cuartil (Q_i) al estadístico que divide una distribución en cuarto partes iguales en términos de porcentajes, siendo estos los siguientes:

- El primer cuartil (Q_1) es aquel valor de la variable que no supera el 25% de las observaciones, o que supera el 75% de las mismas.
- El segundo cuartil (Q_2) es aquel valor de la variable que no supera el 50% de las observaciones, o que supera el 50% de las mismas. Representa la mediana del conjunto de datos.
- El tercer cuartil (Q_3) es aquel valor de la variable que no supera el 75% de las observaciones, o que supera el 25% de las mismas.

Para el cálculo de los cuartiles recurrimos a las siguientes fórmulas:

Datos no agrupados	Datos agrupados
$\mathrm{PQ_i} = \frac{\mathrm{i} * \mathrm{n}}{4}$ cuando n es par , $\mathrm{i} = 1,2,3$ $\mathrm{PQ_i} = \frac{\mathrm{i} * (\mathrm{n}+1)}{4}$ cuando n es impar	$\mathrm{PQ_i} = \mathrm{Li} + \frac{\mathrm{PQi} - \mathrm{F_{a-1}}}{\mathrm{f_i}} \mathrm{IC}$ PQi depende si n es par o impar

$\mathrm{Q_i}$ = cuartil i $\quad$ $\mathrm{PQ_i}$ = posición del cuartil i

Li = límite inferior de la clase del cuartil $\quad$ n = número de datos

$\mathrm{F_{a-1}}$ = frecuencia acumulada de la clase anterior a la clase cuartilar

$\mathrm{f_i}$ = frecuencia simple de la clase cuartilar $\quad$ IC = intervalo de clase

En Excel podemos hacer uso de la fórmula: **=CUARTIL(___:___;k),** donde k =1,2,3

2.2. Deciles (D_i)

Se entiende por deciles (D_i) al estadístico que divide una distribución ordenada de datos en diez partes iguales. Los deciles van del 1 al 9, y se lee de la siguiente manera:

D_1 indica que el 10% de las observaciones están por debajo del D_1 o el 90% están por encima

D_5 indica que el 50% de las observaciones están por debajo del D_5 (es la mediana)

D_9 indica que el 90% de las observaciones están por debajo del D_9 o el 10% están por encima

Para el cálculo de los deciles recurrimos a las siguientes fórmulas:

Datos no agrupados	Datos agrupados
$\mathrm{PD_i} = \frac{\mathrm{i} * \mathrm{n}}{10}$ cuando n es par , $\mathrm{i} = 1,2,3, \ldots 9$ $\mathrm{PD_i} = \frac{\mathrm{i} * (\mathrm{n}+1)}{10}$ cuando n es impar	$\mathrm{PD_i} = \mathrm{Li} + \frac{\mathrm{PDi} - \mathrm{F_{a-1}}}{\mathrm{f_i}} \mathrm{IC}$ PDi depende si n es par o impar

$\mathrm{D_i} =$ decil i $\quad\quad$ $\mathrm{PD_i} =$ posición del decil i

$\mathrm{Li} =$ límite inferior de la clase del decil $\quad\quad$ $\mathrm{n} =$ número de datos

$\mathrm{F_{a-1}} =$ frecuencia acumulada de la clase anterior a la clase decilar

$\mathrm{f_i} =$ frecuencia simple de la clase decilar $\quad\quad$ $\mathrm{IC} =$ intervalo de clase

2.3. Percentiles (P_i)

Los percentiles (P_i) es un estadístico que divide una distribución de datos ordenados en cien partes iguales en términos de porcentajes. Los percentiles van del 1 al 99, y se lee de la siguiente manera:

P_1 indica que el 1% de las observaciones están por debajo del P_1 o el 99% por encima

P_{50} indica que el 50% de las observaciones están por debajo del P_{50}

P_{99} indica que el 99% de las observaciones están por debajo del P_{99} o el 1% por encima

Para el cálculo de los cuartiles recurrimos a las siguientes fórmulas:

Datos no agrupados	Datos agrupados
$\mathrm{PP_i} = \frac{\mathrm{i} * \mathrm{n}}{100}$ cuando n es par , $\mathrm{i} = 1,2,3, \ldots 99$ $\mathrm{PP_i} = \frac{\mathrm{i} * (\mathrm{n}+1)}{100}$ cuando n es impar	$\mathrm{PP_i} = \mathrm{Li} + \frac{\mathrm{PPi} - \mathrm{F_{a-1}}}{\mathrm{f_i}} \mathrm{IC}$ PPi depende si n es par o impar

$P_i =$ percentil i $\quad\quad$ $\mathrm{PP_i} =$ posición del percentil i

$\mathrm{Li} =$ límite inferior de la clase del percentil $\quad\quad$ $\mathrm{n} =$ número de datos

$\mathrm{F_{a-1}} =$ frecuencia acumulada de la clase anterior a la clase del percenti

$\mathrm{f_i} =$ frecuencia simple de la clase decilar $\quad\quad$ $\mathrm{IC} =$ intervalo de clase

En Excel podemos hacer uso de la fórmula: **=PERCENTIL(___:___;k)** donde k oscila entre 0,1 a 1.

3. Medidas de dispersión o variabilidad

Las medidas de dispersión entregan información sobre la variación de la variable, es decir, permite conocer como se comporta la distribución de los datos. A través de estas medidas se resume en un solo valor la dispersión que tiene un conjunto de datos.

Cabe mencionar que, las medidas de posición indican en torno a qué valores se sitúan los datos, pero para obtener una descripción más precisa de los mismos, es necesario conocer cuál es la dispersión que presentan. Para ello, se recurre a las medidas de dispersión, las cuales se basan en la idea de medir las diferencias entre unos datos y otros, midiendo las diferencias de cada dato con la media, esto es, usando las desviaciones; sin embargo, como éstas siempre suman cero, también es preciso considerar su valor absoluto o su cuadrado para que ello no ocurra.

Así que, las medidas de dispersión, tienen por finalidad medir el grado dispersión de las observaciones en torno a una unidad de medida (como por ejemplo la media aritmética), lo cual puede llevar a tres situaciones:

1. Si la dispersión es poca, nos indica que los datos están muy juntos alrededor de la unidad de medida, es decir, hay una gran uniformidad entre los datos (ver Figura 1a).
2. Si la dispersión es cero quiere decir que todos los datos son iguales (ver Figura 1b)
3. Si la dispersión es grande, nos indica que los datos están muy alejados o dispersos en torno de la unidad de medida, es decir, hay poca uniformidad entre los datos (ver Figura 1c).

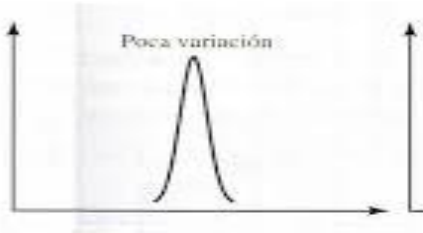


Figura 1a

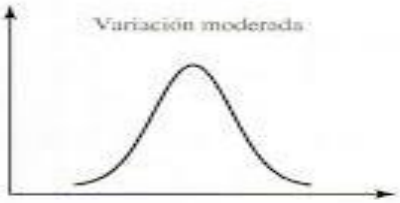


Figura 1b

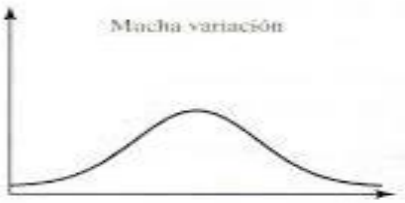
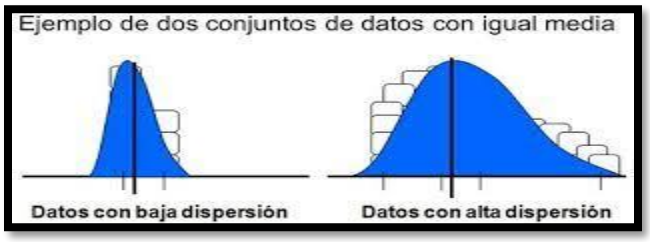


Figura 1c

Por ejemplo, tenemos una situación en relación a la variable tiempo de espera en minutos de los clientes en un supermercado para la cancelación de los artículos. Sabemos que en el supermercado existen varias colas para el pago de los artículos, supongamos seis cajeros. Un grupo de clientes deciden hacer cola en un solo cajero, los demás clientes deciden hacer cola en los restantes cinco cajeros. Es de suponer que el tiempo de espera tiene más dispersión en las colas donde hay más diversidad de clientes.

En la siguiente figura se muestra un ejemplo de dos conjuntos de datos con igual media, pero con diferente dispersión.



Existen dos tipos de medidas de dispersión o variabilidad: absolutas y relativas, las cuales se describen a continuación.

a. Las medidas de dispersión absolutas dependen de las unidades en las que se miden las observaciones, siendo las más conocidas: rango o amplitud de variación, varianza, desviación típica, rango intercuartílico y rango percentílico. Entre ellas tenemos: rango, varianza, desviación estándar, rango intercuartílico y rango percentílico.

b. Las medidas de dispersión relativa, a diferencia de las medidas de dispersión absolutas que dependen de las unidades de los datos, por lo que no son adecuadas para comparar variables, las medidas de dispersión relativas no dependen de las unidades de los datos. Entre ellas se tiene: coeficiente de variación y puntuación Z o estandarización de una variable.

3.1. Rango, amplitud o recorrido (R)

Es una medida que describe la longitud de la variable de estudio, el cual se calcula mediante la siguiente fórmula:

$$\boldsymbol{R = Dato\ mayor - Dato\ Menor}$$

Cabe destacar lo siguiente sobre el rango:

1. No brinda información respecto a la dispersión existente en el conjunto de observaciones, solo permite determinar el alcance de las variaciones extremas, por ejemplo, cuando se indica la variación de temperatura o las operaciones bursátiles.
2. No toma en cuenta todos los valores, por lo tanto, no refleja realmente la variación entre todos los valores de los datos.
3. El rango utiliza solo los valores mínimo y máximo de los datos, por lo que es muy sensible a los valores extremos que se puedan presentar en una sucesión de datos. El rango no es resistente.
4. Entre sea menor el rango, son más confiables las observaciones obtenidas. Esto es útil cuando se compara dos grupos de datos con la misma media aritmética.

En Excel no puede calcularse directamente el rango, para ello utilizamos las funciones, MAX y MIN, por ejemplo: $\mathbf{Rango} = (= \mathbf{MAX}(\mathbf{A1}:\mathbf{A15})) - (= \mathbf{MIN}(\mathbf{A1}:\mathbf{A15}))$

3.2. Varianza (S^2)

La varianza se define como el promedio de las desviaciones con respecto a la media, elevadas al cuadrado. Cuando se utiliza la varianza como medida de variabilidad, resulta que el promedio obtenido de las desviaciones elevadas al cuadrado siempre serán unidades cuadradas. Así, si el conjunto de datos está medido en kilogramos, la varianza de estos se medirá en kg^2, si es en años, la varianza se medirá en$años^2$ y así sucesivamente.

Se denota de la siguiente manera:

- Varianza poblacional: $\boldsymbol{\sigma^2}$(sigma) se lee cuadrado de la desviación estándar de la población (parámetro).
- Varianza muestral: $\mathbf{s^2}$, se lee cuadrado de la desviación estándar de la muestra (estimador).

Vale recordar que en estadística inferencial se utiliza el estimador o estadígrafo para estimar el valor del parámetro.

Para el cálculo manual de la varianza se utilizan las siguientes fórmulas:

Datos no agrupados	Datos agrupados
$S^2 = \frac{\sum(x_i - \overline{X})^2}{n-1}$ o $\frac{\sum x_i^2 - n\overline{X}}{n-1}$	$S^2 = \frac{\sum f_i(x_i - \overline{X})^2}{n-1}$

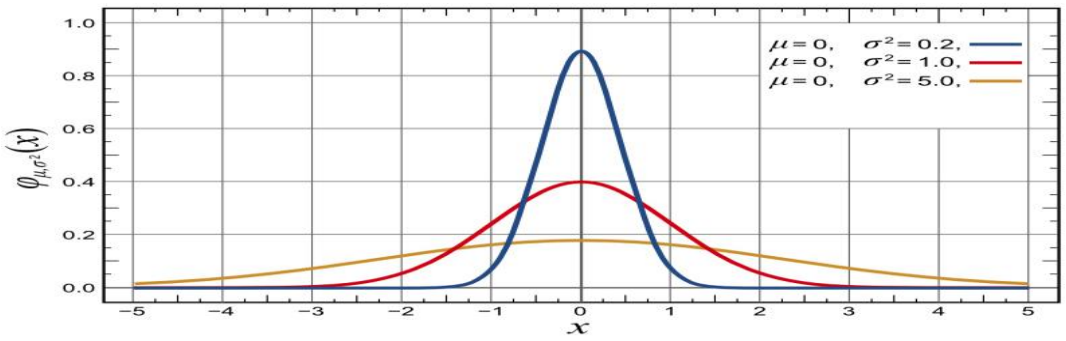
En Excel podemos hacer uso de la fórmula: **=VAR.S(___:___)**

Propiedades de la varianza

1. Todos los valores son utilizados en el cálculo de la varianza.
2. El valor de la varianza puede aumentar dramáticamente con la inclusión de valores atípicos. (La varianza no es resistente).
3. Las unidades de la varianza son los cuadrados de las unidades de los valores de datos originales.
4. La varianza debe ser siempre un valor positivo.
5. La varianza de una constante es igual a cero.
6. La varianza muestral es un estimador no sesgado de la varianza poblacional.
7. La varianza de una constante más una variable es igual a la varianza de la variable. $V(k) + V = V(v)$
8. La varianza de una constante por una variable, es igual al producto de la constante al cuadrado por la varianza de la variable. $V(k) * V = k^2S^2$

Para tener en cuenta sobre la varianza

1. Cuando el valor de la varianza de un conjunto de datos es grande, se dice que tiene mayor variabilidad.
2. Si la varianza de un conjunto de datos es pequeña, entonces la variabilidad es pequeña. Este conocimiento es útil cuando se comparan dos o más conjuntos de datos.
3. También es claro que a mayor variabilidad (mayor extensión de los datos) mayor será: el recorrido, el recorrido intercuartílico, la varianza y, como consecuencia, la desviación estándar.
4. Si se quiere describir la variabilidad de un solo conjunto de datos, la varianza no es de gran ayuda porque esta no se expresa en las unidades originales, sino en unidades al cuadrado.



La línea amarilla indica una mayor dispersión de los datos. Mientras la línea azul indica una menor dispersión. Estas diferencias de variación se deben a los valores de las varianzas.

3.3. Desviación estándar (S)

La desviación estándar de un conjunto de valores muestrales, expresada por ***s***, es una medida de cuánto se desvían los valores de datos de la media. En el caso de la desviación estándar poblacional se simboliza como $\boldsymbol{\sigma}$ (se lee como sigma) y corresponde al parámetro.

La desviación estándar es igual a la raíz cuadrada positiva de la varianza; es decir, en la desviación estándar, las unidades con que se mide este estadístico de dispersión serán las mismas que tienen las observaciones y la media aritmética de estas.

Para el cálculo manual de la desviación estándar se utilizan las siguientes fórmulas:

Datos no agrupados	**Datos agrupados**
$S = \sqrt{\frac{\sum(x_i - \overline{X})^2}{n-1}}$ o $\sqrt{\frac{\sum x_i^2 - n\overline{X}}{n-1}}$ $S = \sqrt{S^2}$	$S = \sqrt{\frac{\sum f_i(x_i - \overline{X})^2}{n-1}}$

En Excel podemos hacer uso de la fórmula: **= DESVEST.M(___:____)**

Propiedades importantes de la desviación estándar

1. La desviación estándar es una medida de cuanto se desvían los valores de datos de la media.
2. El valor de la desviación estándar nunca es negativo. Es cero solo cuando todos los valores de datos son exactamente iguales.
3. Los mayores valores de *s* indican mayores cantidades de variación.
4. La desviación estándar puede aumentar dramáticamente con uno o más valores atípicos.
5. Las unidades de la desviación estándar (como minutos, pies, libras, etc.) son las mismas que las unidades de los valores de datos originales.
6. La desviación estándar muestral *s* es un **estimador sesgado** de la desviación estándar σ de la población, lo que significa que los valores de la desviación estándar muestral *s* no se centran en torno al valor de s.

Relación entre el Rango y la Desviación Estándar (Regla práctica del rango para estimar un valor de la desviación estándar s)

Esta relación es válida en distribuciones aproximadamente simétricas. La misma consiste en que la desviación estándar es aproximadamente igual a la cuarta parte del rango:

$$\boldsymbol{S \approx \frac{Rango}{4}}$$

Debido a que esta estimación se basa solo en los valores mínimo y máximo, es generalmente una estimación aproximada que podría estar alejada a una distancia considerable.

3.4. Rango intercuartílico (RIQ)

Es la distancia promedio desde los cuartiles Q_1 y Q_3 hasta la mediana. Se calcula a través de la siguiente expresión:

$$\boldsymbol{RIQ = Q_3 - Q_1}$$

Es decir, el 50% de las observaciones de la variable están entre Q_1 y Q_3

Existe una relación entre el rango intercuartílico y la desviación estándar, que es válida cuando los datos siguen una distribución aproximadamente normal, cuya expresión es la siguiente:

$$\boldsymbol{RIQ = \frac{2}{3}S}$$

3.5. Rango percentil (RP)

Es el recorrido de la variable desde el P_{90} hasta el P_{10}, y se calcula de la siguiente manera:

$$\boldsymbol{RP = P_{90} - P_{10}}$$

Es decir, el 80% de las observaciones de la variable están entre P_{10} y $P_{90.}$

3.6. Coeficiente de variación (CV)

El coeficiente de variación permite comparar variables, aunque estas estén registradas en distintas unidades de medida. También es de utilidad para comparar variables que, aunque de la misma magnitud, están en escalas distintas.

El coeficiente de variación se emplea para:

1. Comparar la variabilidad entre dos grupos de datos referidos a distintos sistemas de unidades de medidas, por ejemplo: metros y centímetros, kilogramos y gramos, m/s y cm/s.
2. Comparar la variabilidad entre dos grupos de datos obtenidos por dos o más personas.
3. Comparar dos grupos de datos que tienen distinta media.
4. Determinar si cierta media es consistente con cierta varianza.

Se emplea la siguiente formula:

$$\mathbf{CV = \frac{\sigma}{\mu} * 100\ (población)} \quad \mathbf{o} \quad \mathbf{CV = \frac{S}{\overline{X}} * 100\ (muestra)}$$

Puede utilizarse el siguiente baremo para su interpretación:

- Menos del 10% Baja variabilidad
- De 10% a 20% Moderada variabilidad
- De 21% a 30% Regular variabilidad
- Mayor de 31% Alta variabilidad

A mayor valor de CV, mayor variabilidad.

Nota. Para calcular el coeficiente de variación con ayuda de Excel, debemos calcular primero la media aritmética y la desviación estándar.

3.7. Puntuación Z o estandarizada

Los puntajes Z son transformaciones que se pueden hacer a los valores o puntuaciones individuales ($\boldsymbol{X_i}$) de una distribución normal, con el propósito de analizar su distancia respecto a la media, expresándolas en unidades de desviación estándar.

La fórmula para transformar un valor de una distribución normal en una unidad de desviación estándar es:

$$\mathbf{Z} = \frac{\mathbf{X} - \bar{\mathbf{X}}}{\mathbf{S}}$$

Siendo **Z** la puntuación Z, **X** es la puntuación o valor a transformar, $\bar{\mathbf{X}}$ es la media de la distribución original y **S** es la desviación estándar.

Propiedades de las puntuaciones típicas

1. La media aritmética de todas las puntuaciones típicas siempre es igual a 0.
2. La desviación típica de las puntuaciones típicas es equivalente a 1.
3. Las puntuaciones típicas son adimensionales, ya que las unidades del numerador se cancelan con las unidades del denominador.
4. Si una puntuación típica es positiva, significa que la puntuación directa está por encima de la media. Por otro lado, si la puntuación típica es negativa quiere decir que la puntuación directa está por debajo de la media.
5. Las puntuaciones típicas son muy útiles para comparar diferentes distribuciones.
6. La puntuación tipificada se utiliza para estandarizar datos y hacerlos comparables, independientemente de la escala en la que se miden. Además, permite identificar valores extremos o atípicos en un conjunto de datos.

4. Medidas de forma o distribución

Las medidas de distribución de nos proporciona una idea de la forma cómo se distribuyen los datos. Es decir, permite identificar ciertas características especiales como simetría, asimetría, nivel de concentración de datos y nivel de apuntamiento que la clasifiquen en un tipo particular de distribución.

Estas medidas describen la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se hallen dentro de la información. Su utilidad radica en la posibilidad de identificar las características de la distribución sin necesidad de generar el gráfico. Estas medidas de forma que proporcionan información numérica sobre dos características de la distribución, su simetría y su apuntamiento o curtosis.

En general, las medidas de distribución permiten:

1. Identificar si una distribución de frecuencia presenta uniformidad.
2. Identificar la forma en que se separan o aglomeran los valores de acuerdo a su representación gráfica.
3. Describir la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se hallen dentro de la información.
4. Determinar el comportamiento de los datos y así, poder adaptar herramientas para el análisis probabilístico

Sus principales medidas son: Coeficiente de asimetría y Coeficiente de curtosis.

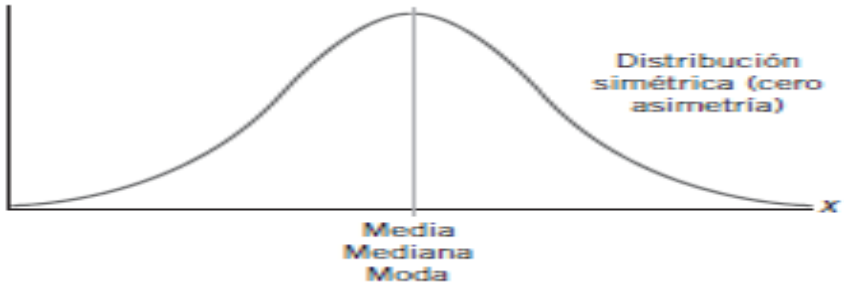
4.1. Coeficiente de sesgo o asimetría (CS)

El coeficiente de sesgo o asimetría es un número que mediante su signo determina si los datos representados en una curva (campana de Gauss) tienen distribución simétrica o sesgada.

En una distribución simétrica las observaciones de la variable tienden a situarse en igual proporción a ambos lados del valor medio. Cualquier medida que recoja alteraciones de esta situación proporcionará una cuantificación de la asimetría de la distribución.

La asimetría presenta tres estados diferentes, cada uno de los cuales define de forma concisa como están distribuidos los datos respecto al eje de asimetría. Cabe decir, que la ubicación de la media en relación con las otras dos medidas de tendencia central indica si los datos presentan una distribución simétrica o asimétrica. Estos tres estados son los siguientes:

1. Cuando la distribución es simétrica (no tiene sesgo), la media, la mediana y la moda se ubicarán en el centro de la distribución. En este caso, estas tienen el mismo valor $\bar{X} = Me = Mo$.



2. Cuando una distribución es asimétrica (tiene sesgo) se presenta dos situaciones:

- Asimetría a la derecha o asimetría positiva o distribución con sesgo positivo. En estas condiciones $\bar{X} > Me > Mo$ (figura 1a).
- Asimetría a la izquierda o asimetría negativa o distribución con sesgo negativo. En estas condiciones: $\bar{X} < Me < Mo$ (figura 1b)

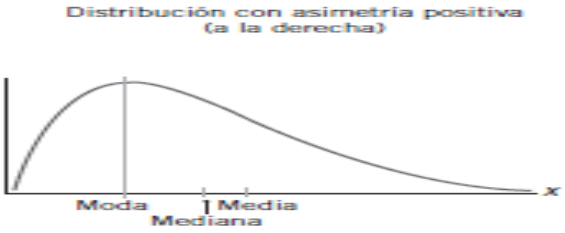


Figura 1a

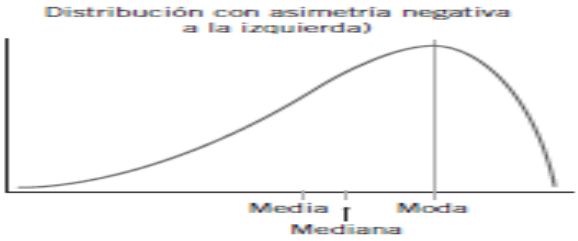


Figura 1b

Para el cálculo del Coeficiente de Simetría puede utilizarse cualquiera de las siguientes formulas, dependiendo si los datos están agrupados o no.

	Datos no agrupados	**Datos agrupados**
Coeficiente de simetría de Fisher	$CS = \frac{\sum(x_i - \bar{X})^3}{nS^3}$	$CS = \frac{\sum f_i(X_i - \bar{X})^3}{nS^3}$
Coeficiente de simetría de Pearson	$CS = \frac{\bar{X} - Mo}{S}$	
Coeficiente de Arthur Bowley	$CS = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$	

El Coeficiente de Simetría se interpreta del siguiente modo:

1. Si *CS* = 0, entonces los datos (de la curva) se distribuyen de manera simétrica.
2. Si *CS* > 0, entonces los datos (de la curva) son sesgados a la derecha.
3. Si *CS* < 0, entonces los datos (de la curva) son sesgados a la izquierda

Otra forma de determinar el CS es mediante el coeficiente de asimetría, a través de la diferencia de la media y la mediana en relación con la desviación estándar, tal como se muestra en las siguientes expresiones:

Asimetría de la población	**Asimetría de la muestra**
$C.S. = \frac{3(\mu - Me)}{\sigma}$	$C.S. = \frac{3(\bar{X} - Me)}{S}$

En Excel existe la función **=COEFICIENTE.ASIMETRIA(___:___).**

4.2. Coeficiente de curtosis (K)

El coeficiente de curtosis es un número cuya magnitud nos indica si los datos se distribuyen simétricamente de forma normal (curva mesocúrtica), más empinada que la curva normal (curva leptocúrtica) o más aplanado de la curva normal (curva platicúrtica).

1. Si el coeficiente es positivo, la distribución se llama leptocúrtica, más puntiaguda que la anterior. Hay una mayor concentración de los datos en torno a la media. **K > 0** (figura 1a)
2. Si este coeficiente es nulo, la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de mesocúrtica. **K = 0** (figura 1b)
3. Si el coeficiente es negativo, la distribución se llama platicúrtica y hay una menor concentración de datos en torno a la media. Sería más achatada que la primera. **K < 0** (figura 1c)

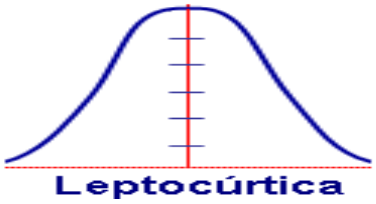


Figura 1a

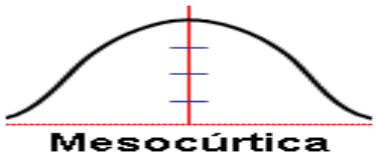


Figura 1b

Figura 1c

En los histogramas correspondientes suele dibujarse en ocasiones la forma teórica que correspondería a una distribución *normal*, con respecto a la cual se está haciendo la comparación.

Para el cálculo del coeficiente de curtosis recurrimos a las siguientes fórmulas:

Datos no agrupados	Datos agrupados
$\mathrm{K} = \frac{\sum(X_i - \overline{X})^4}{nS^4} - 3$	$\mathrm{K} = \frac{f_i(X_i - \overline{X})^4}{nS^4} - 3$

Otra forma de calcular el coeficiente de curtosis es a través de una fórmula en función de los cuantiles Q_1, Q_2, P_{10} y P_{90}, la cual se ilustra a continuación:

$$\mathbf{K} = \frac{\mathbf{Q_3 - Q_1}}{\mathbf{2(P_{90} - P_{10})}}$$

Algunos autores consideran el siguiente criterio de interpretación del coeficiente de curtosis:

1. Si $K = 3$, entonces los datos (de la curva) se distribuyen de manera simétrica en forma de una curva de distribución normal estandarizada.
2. Si $K > 3$, entonces los datos (de la curva) presentan un pico mayor que los de la curva de distribución normal estandarizada (curva leptocúrtica).
3. Si $K < 3$, entonces los datos (de la curva) se presentan más aplanados que los de la curva distribución normal (curva platicúrtica).

Actividad de autoevaluación

1. En la siguiente lista de variables, seleccione a aquellas que admiten el cálculo de las medidas de tendencia central y de dispersión

a) Color de los coches

b) Nivel de estudios

c) Número de hijos de una familia

d) Profesiones de las personas

e) Sueldo mensual de los trabajadores de las empresas del sector cerámico

f) Puntajes obtenidos de un test de inteligencia emocional

g) Clasificación de las personas según la raza

2. Un investigador está realizando investigaciones sobre las diferencias individuales de los estudiantes en cuanto a su susceptibilidad para ser hipnotizados. Como parte del experimento, el investigador decide administrar una parte de la Escala de Susceptibilidad Hipnótica de Stanford a 24 estudiantes que se ofrecieron como voluntarios para el experimento. Los resultados obtenidos fueron calificados de O a 12. correspondiendo a 12 el grado más alto de susceptibilidad hipnótica y O al más bajo. Las calificaciones se presentan a continuación:

Determine e interprete (Datos no agrupados):

a) Media aritmética y moda

b) Rango y varianza

c) ¿La distribución es simétrica o sesgada?

d) El rango intercuartílico y rango percentílico.

3. En una medición de angustia, la media es 79 y el desvío estándar es 12. ¿Cuáles son las puntuaciones Z correspondientes a cada una de las siguientes puntuaciones brutas? a) 81, b) 68, c) 103.

4. Un psicólogo está interesado en los hábitos de los estudiantes de licenciatura en materia de citas románticas. Con ese propósito, elige una muestra de 10 estudiantes y determina el número de citas que tuvieron durante el mes pasado. A partir de bs datos 1, 8, 12, 3, 5, 10, 4, 5, 10, 2. Calcule lo siguiente:

a. Media b. Mediana c. Rango d. Desviación estándar

e. Coeficiente de variación f. Coeficiente de asimetría g. Coeficiente de curtosis

5. En una investigación sobre el rendimiento académico en estudiantes de secundaria, se obtuvieron las calificaciones medias en la materia Ciencias Naturales en tres grupos de estudiantes de diferentes unidades educativas, los resultados se resumen en la siguiente tabla:

Unidad educativa	Media	Desviación estándar
A	12,5	3,5
B	17,0	2,4
C	14,8	2,9

¿Cuál de las tres unidades educativas presentan una mayor variabilidad respecto al rendimiento académico de los estudiantes en la materia Ciencias Naturales?

6. La puntuación de una persona en una prueba de aptitud verbal es de 81, y de 6,4 en una prueba de aptitud numérica. En el caso de la prueba de aptitud verbal, la media para las personas en general es 50 y el desvío estándar es 20. En el caso de la prueba de aptitud numérica, la media para las personas en general es 0 y el desvío estándar es 5. ¿Cuál es la mayor aptitud de esta persona, la verbal o la numérica?

7. La siguiente tabla de frecuencias agrupadas corresponde a la medición del estrés.

Li	**Ls**	**Xi**	**fi**	**Fa**
0	1,9		3	3
2	3,9		15	18
4	5,9		34	52
6	7,9		43	95
8	9,9		41	136
10	11,9		14	150
			150	

Determine e interprete:

a. Media b. Cuartil 50 c. Decil 4 d. Percentil 45 e. Coeficiente de asimetría