Medidas de Tendencia Central y de Variabilidad

Unidad de Aprendizaje: Estadística

Unidad 3: Medidas estadísticas.

Centro Universitario Atlacomulco

Licenciatura en Contaduría

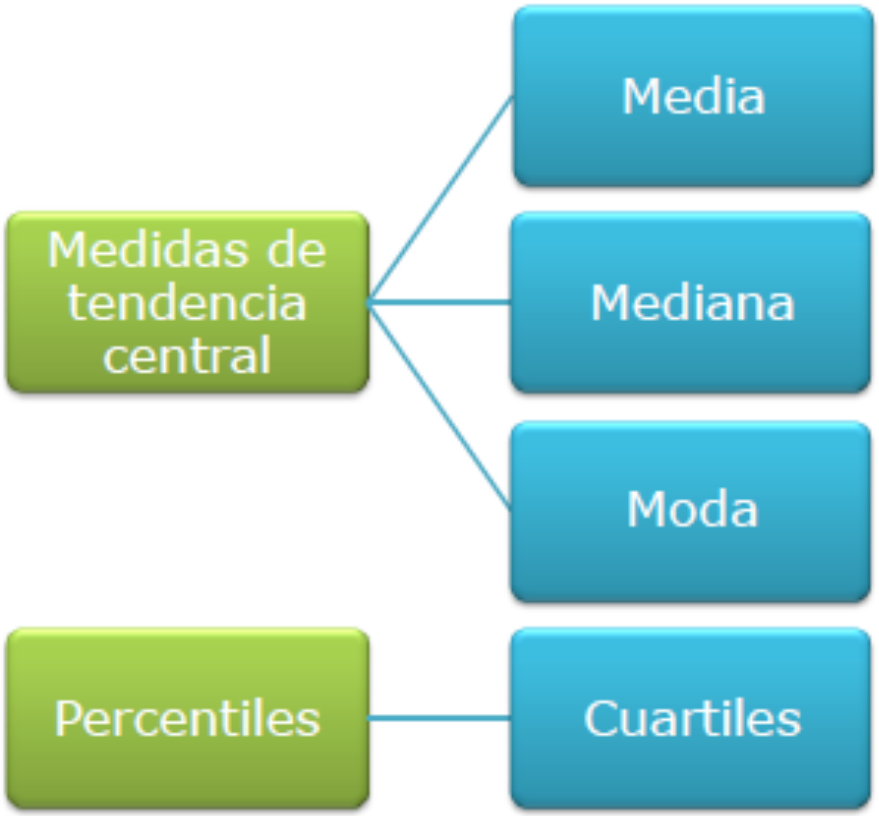
Créditos: 7

Elaboró: M.A. Yenit Martínez Garduño

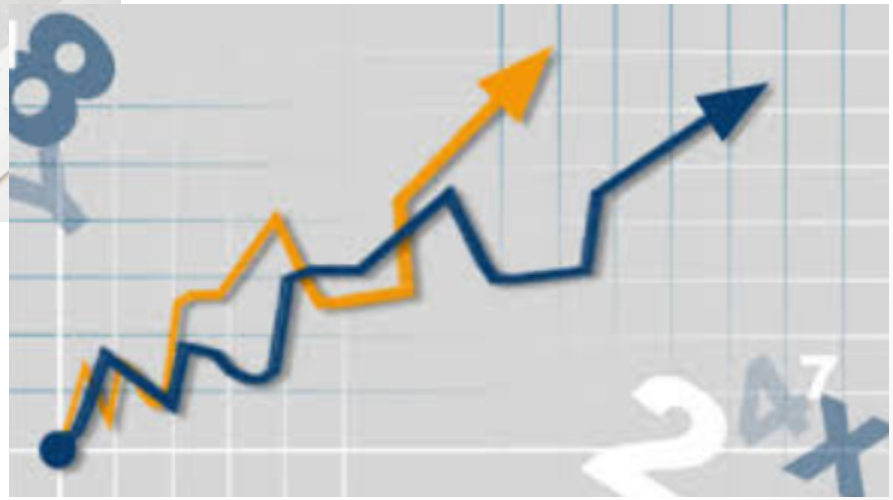
Guión Explicativo:

- El presente material contiene el estudio de los temas medidas de tendencia central y de variabilidad de la **unidad 3 Medidas Estadísticas** de la Unidad de aprendizaje Estadística.

- En éste el alumno encontrará información sobre lo que son las medidas de tendencial central: media, mediana y moda; así como percentiles y cuartiles. Se abordarán también las medidas de variabilidad o de dispersión: rango, varianza y desviación estándar. Para cada una de éstas, el alumno aprenderá a determinarlas e interpretarlas para poder resolver problemas, reconociendo así la aplicación práctica de éste tipo de conceptos.


Medidas de tendencia central
Media
Mediana
Moda
Percentiles
Cuartiles

Las medidas de localización permiten ubicar el comportamiento de un conjunto de datos para respaldar la interpretación y toma de decisiones. A estas medidas numéricas les llamamos genéricamente **estadísticos descriptivos** o simplemente **estadísticos**.



En primer lugar revisaremos las llamadas **medidas de tendencia central**, dichas medidas son la **media**, la **mediana** y la **moda.** Las tres tienen el objetivo de resumir en un solo valor un conjunto de datos numéricos, sin embargo cada una lo hace mediante un proceso diferente, por lo cual la información que proporcionan es diferente en cada caso aunque puede resultar valiosamente complementaria.

La **media** es el estadístico más comúnmente utilizado para resumir conjuntos de datos cuantitativos. Sin embargo en
la práctica común se conoce más con el nombre de **promedio**. Se calcula sumando todos los valores y dividendo el
resultado entre el número de observaciones. Lo cual se puede expresar con la siguiente fórmula:

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

En donde representa la suma de todos los datos y *n* es el número de datos. Si éstos son la totalidad de datos que interesan en un estudio, se les llama **población** y si se trata de sólo una parte que son tomados para obtener información, se les llama **muestra**. Para designar la media de una muestra utilizamos el signo , mientras que para la población usamos la letra griega μ.

Para ejemplificar este concepto consideremos los siguientes datos correspondientes al número de hijos que tienen 20 maestras que trabajan en una universidad.

2 4 2 3 1 2 4 2 3 0
3 3 4 5 2 0 3 2 1 2

El promedio se obtiene sumando todos los valores y dividendo entre el tamaño de la muestra.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{2+4+2+2+3+\cdots+2+1+2}{20} = \frac{48}{20} = 2.4$$

Así pues podemos afirmar que las maestras de esta universidad tienen en promedio 2.4 hijos.

Aunque la media es la medida de tendencia central más común, presenta el inconveniente de que es una medida muy sensible a los valores extremos. Tú sabes bien por experiencia en tus calificaciones cuánto puede afectar a tu promedio el que pierdas o te anulen un examen; o bien cómo puede ayudar obtener la máxima calificación en un examen parcial.

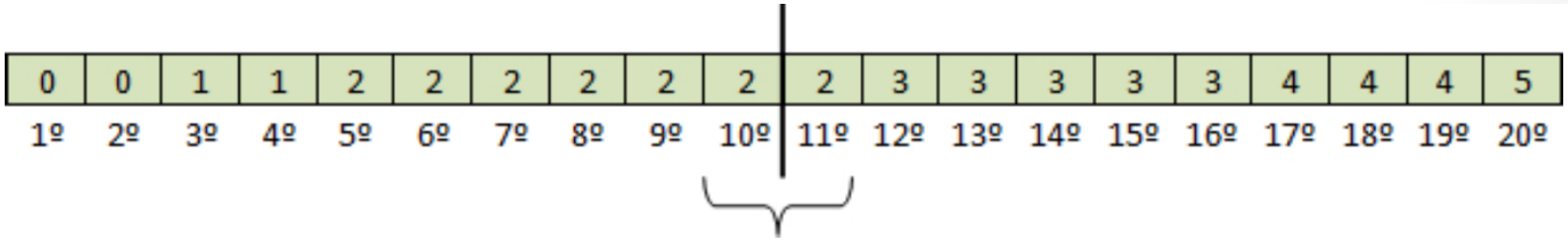
La **mediana** es otra medida de tendencia o localización central de los datos. Aquí lo que se busca es encontrar el valor intermedio de un conjunto de datos cuando éstos se ordenan de menor a mayor.

De esta manera su valor no queda afectado por los valores extremos. No existe un símbolo universalmente aceptado para denotarlo, pero en algunos textos se utiliza $\tilde{x}$ ó ***me***

A diferencia de la media, la mediana no puede calcularse siguiendo una fórmula, pero sí mediante un algoritmo:

1. Ordena los datos en orden ascendente.
2. Si el conjunto tiene un número impar de datos, la mediana es el valor intermedio.
3. Si el conjunto tiene un número par de datos, la mediana es la media de los dos valores intermedios.

Para ejemplificar tomemos el conjunto de 20 datos que usamos para la media y ordenémoslos de menor a mayor:



Tenemos 20 datos, entonces buscaremos el promedio de los datos 10º y 11º que están a la mitad de la secuencia de datos.

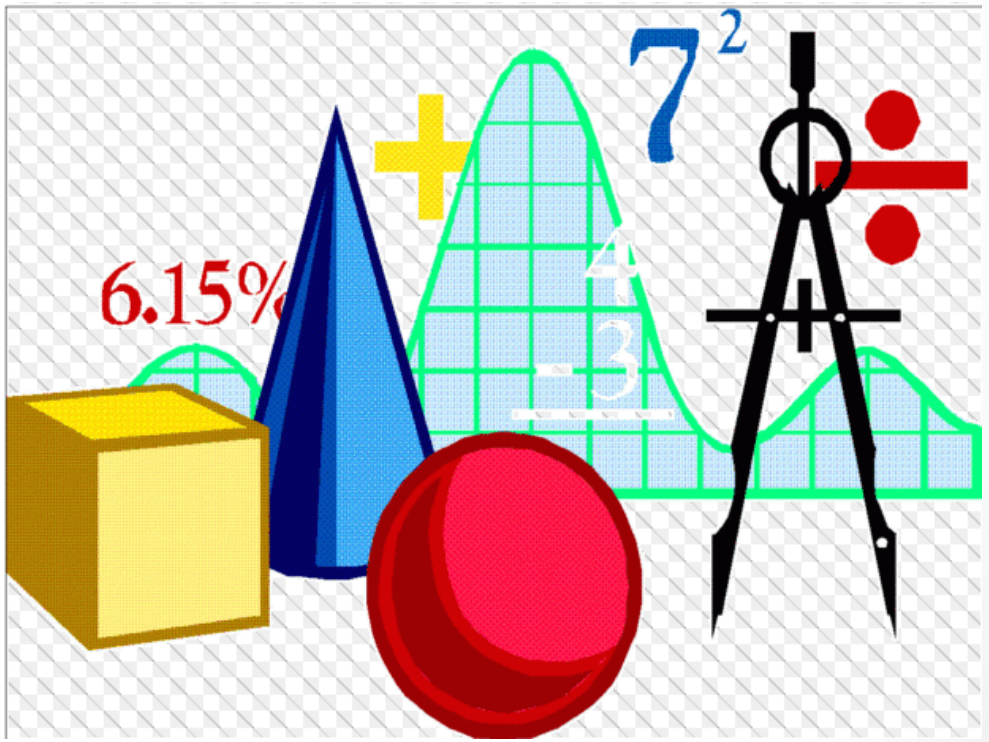
Como en este caso ambos datos son 2, entonces el promedio de ellos también lo es y podemos afirmar que la mediana es igual a 2.

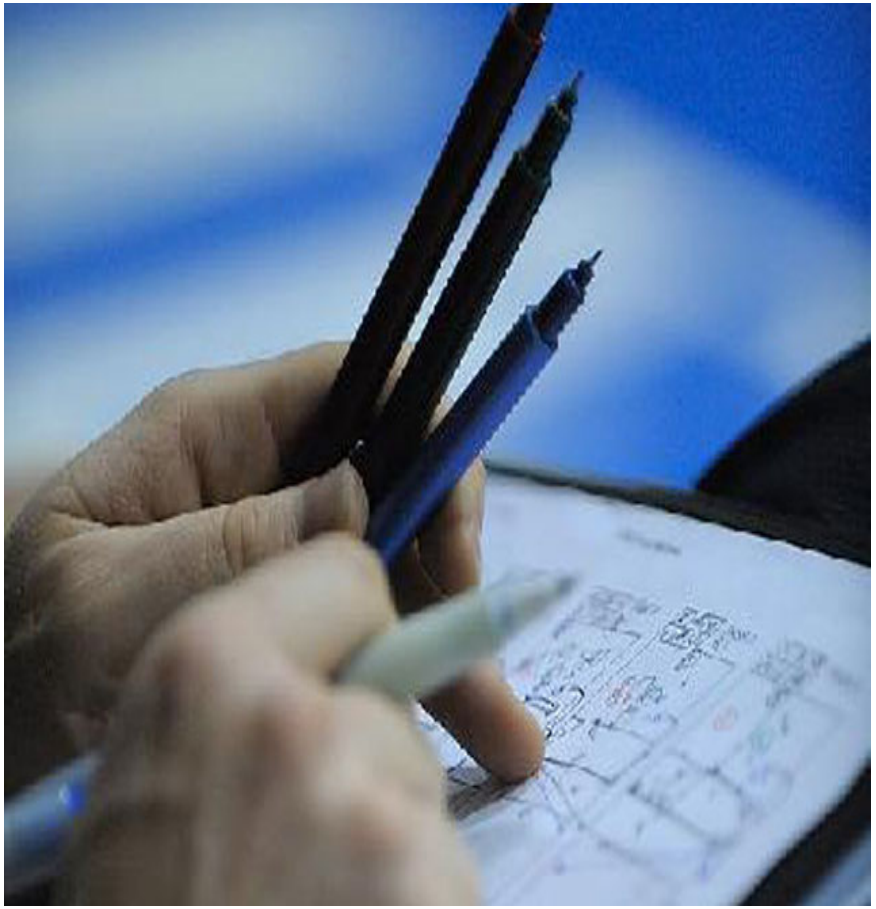
La **moda** es la tercera medida de tendencia central o localización. Se trata del valor que se presenta con mayor frecuencia. Este es un concepto muy sencillo de recordar ya que en nuestra vida diaria nos referimos a él cuando hablamos de la música o la ropa que está de moda y comprendemos que se trata de la más común. No existe un símbolo universalmente adoptado para denotarlo pero en algunos textos se utiliza $\hat{x}$ ó ***mo***

Para nuestro ejemplo, hay 2 madres que no tienen hijos, 2 que tienen un solo hijo, 7 que tienen dos hijos, 5 con 3 hijos, 3 con 4 hijos y finalmente una madre que tiene 5. Es por esto que el valor que se repite con mayor frecuencia es el 2, por lo que éste es la moda.

Cuando conocemos las tres medidas de tendencia central podemos tener una mejor idea de los datos que se pretenden resumir, ya que estos tres estadísticos no necesariamente coinciden.

Resumiendo:
Media = 2.4 hijos
Mediana = 2 hijos
Moda = 2 hijos





A veces nos interesa saber cuál es el lugar que ocupa un determinado valor en un conjunto de datos. Aquí el interés ya no está en centrar la información sino en ubicar un dato con respecto al total. Por ejemplo podríamos preguntarnos cuántos hijos tienen el 10 por ciento de las madres que tienen más hijos. Este es el concepto del estadístico llamado **percentil**.

Un **percentil** determinado (*p-ésimo*) representa el valor tal, que por lo menos ese tanto por ciento de valores son menores o iguales que él.
Para contestar cuál es valor que limita al 10 por ciento mayor, tendríamos que buscar el percentil 90 ya que el 90% de los valores serían menores a él.
Para denotarlo no existe un símbolo de adopción universal pero en algunos textos se utiliza P_{90}

Para calcular un valor percentil, se encuentra por regla de 3 cuántos valores corresponden a ese porcentaje. Si el valor obtenido no es entero, se toma el valor entero inmediato mayor. Si el valor sí es entero, entonces se obtiene el promedio entre ese valor y el siguiente como en el caso de la mediana.

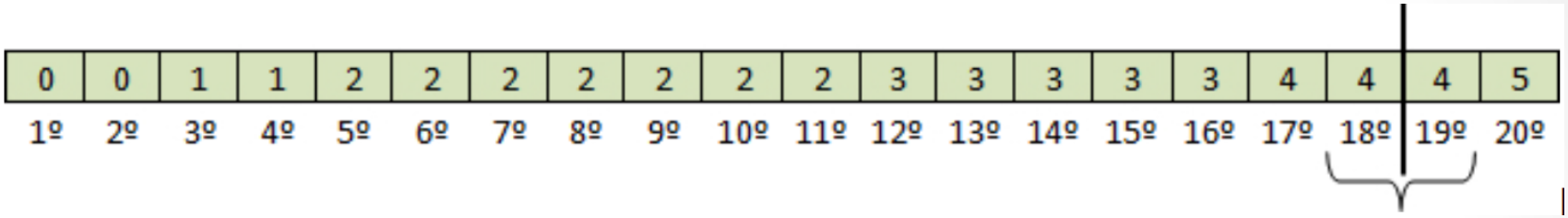
Así pues:

$$\frac{i}{20^o} = \frac{90\%}{100\%}$$

$$i = \frac{(20^o)(90\%)}{100\%}$$

$$i = 18^o$$

Así el percentil 90 equivale al 18º dato. Como se trata de un valor entero, promediaremos el 18º y el 19º valor, por lo que el valor del percentil 90 es 4. Lo que significa que el 90% de las madres tienen 4 o menos hijos.



Por último están los **cuartiles** que corresponden a percentiles específicos:

Primer cuartil	Q*1* equivale al percentil 25
Segundo cuartil	Q*2* equivale al percentil 50 (que es también la mediana)
Tercer cuartil	Q*3* equivale al percentil 75


Medidas de variabilidad
Rango
Varianza
Desviación estándar

Las medidas de localización ofrecen información útil para resumir un conjunto de datos, pero tomándolas por separado no nos permiten inferir la forma en que los datos varían.

Fíjate en las calificaciones parciales de tres alumnos:

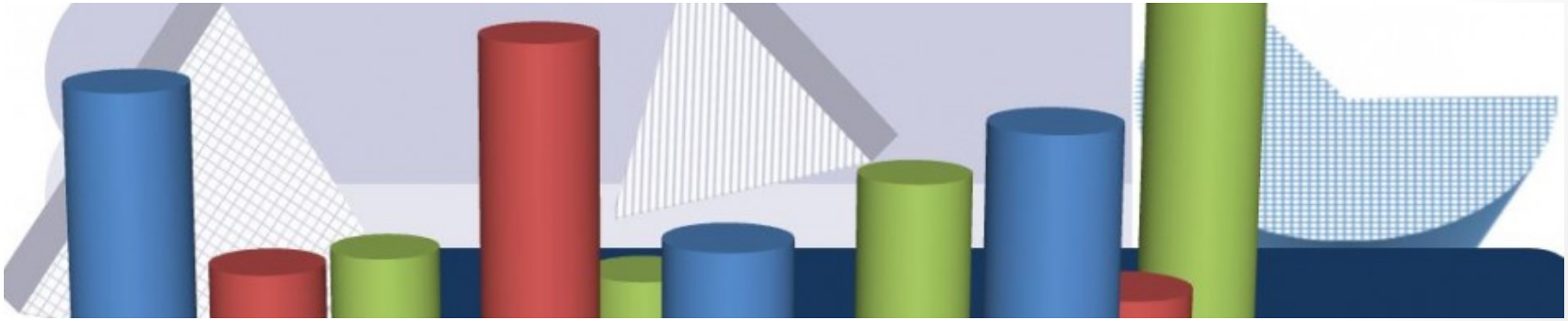
Alberto:	8	8	8	8
Beatriz:	10	10	10	2
Carlos:	6	7	9	10

Los tres tienen el mismo promedio, pero ¿podemos decir que tienen el mismo desempeño? Alberto tiene un aprovechamiento consistente por lo que el promedio es una buena medida de resumen. Beatriz obtiene en las tres primeras calificaciones un desempeño mejor que Alberto, sin embargo, en su última calificación algo sucede que le hace obtener una calificación muy baja. Para Carlos su desempeño ha ido de menos a más, habiendo obtenido al final la calificación más alta. Si la calificación del curso consistiera en el promedio de las cuatro calificaciones parciales, ¿consideras justo que los tres obtengan la misma nota?

Las medidas de variabilidad permiten considerar qué tanta confiabilidad podemos tener en la representatividad de una medida de resumen sobre un conjunto de datos.

El **rango** es la diferencia entre el dato mayor y el menor del conjunto de datos.

R = max - min



Así para el caso de los tres alumnos:

Alberto: $R = 8 - 8 = 0$
Beatriz: $R = 10 – 2 = 8$
Carlos: $R = 10 – 6 = 4$

Sin embargo, el **rango** sólo utiliza dos datos para obtener la dispersión, por lo que éste queda determinado sólo por los valores extremos que pueden no ser representativos.

Otra medida de la dispersión que sí considera a todos los datos es la **varianza**.

Si se trata de la **varianza de la población** se denota con el cuadrado de la letra griega sigma y se calcula mediante la siguiente fórmula:

$$i = 1, N$$

Y si en lugar de la población se trata de la **varianza de una muestra**, se denota con el cuadrado de la letra s, y la fórmula varía en el denominador donde la división se hace entre una unidad menos que el tamaño de la muestra.

$$i = 1, n$$

Ejemplifiquemos el cálculo de la varianza para el caso de las calificaciones de Carlos a las que consideraremos valores de una muestra.

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(6-8)^2 + (7-8)^2 + (9-8)^2 + (10-8)^2}{4-1}$$
$$= \frac{(-2)^2 + (-1)^2 + (1)^2 + (2)^2}{3} = \frac{4+1+1+4}{3} = \frac{10}{3} = 3.33$$

Hace años calcular la varianza de un conjunto de datos podía ser un trabajo extenuante. Afortunadamente hoy podemos hacerlo con el apoyo de las llamadas calculadoras científicas que ya tienen una función programada para hacer este cálculo o también utilizando una hoja de cálculo en cualquier computadora.

La **varianza** tiene para su uso práctico el inconveniente de que, para evitar que las diferencias se cancelaran al promediarse, éstas fueron elevadas al cuadrado para hacerlas todas positivas, con la consecuencia de que la unidad también está elevada al cuadrado lo que hace su interpretación más difícil.

Es decir, ¿qué significa decir que Carlos tiene un promedio de 8 puntos con una varianza de 3.33 puntos cuadrados?

Es por esta razón que para corregir esta situación se introduce el siguiente concepto.

La **desviación estándar** es la raíz cuadrada positiva de la varianza y se convierte en la medida de variabilidad más utilizada.
Desviación estándar poblacional:

$$\sigma = \sqrt{\sigma^2}$$

Desviación estándar de una muestra:

$$s = \sqrt{s^2}$$

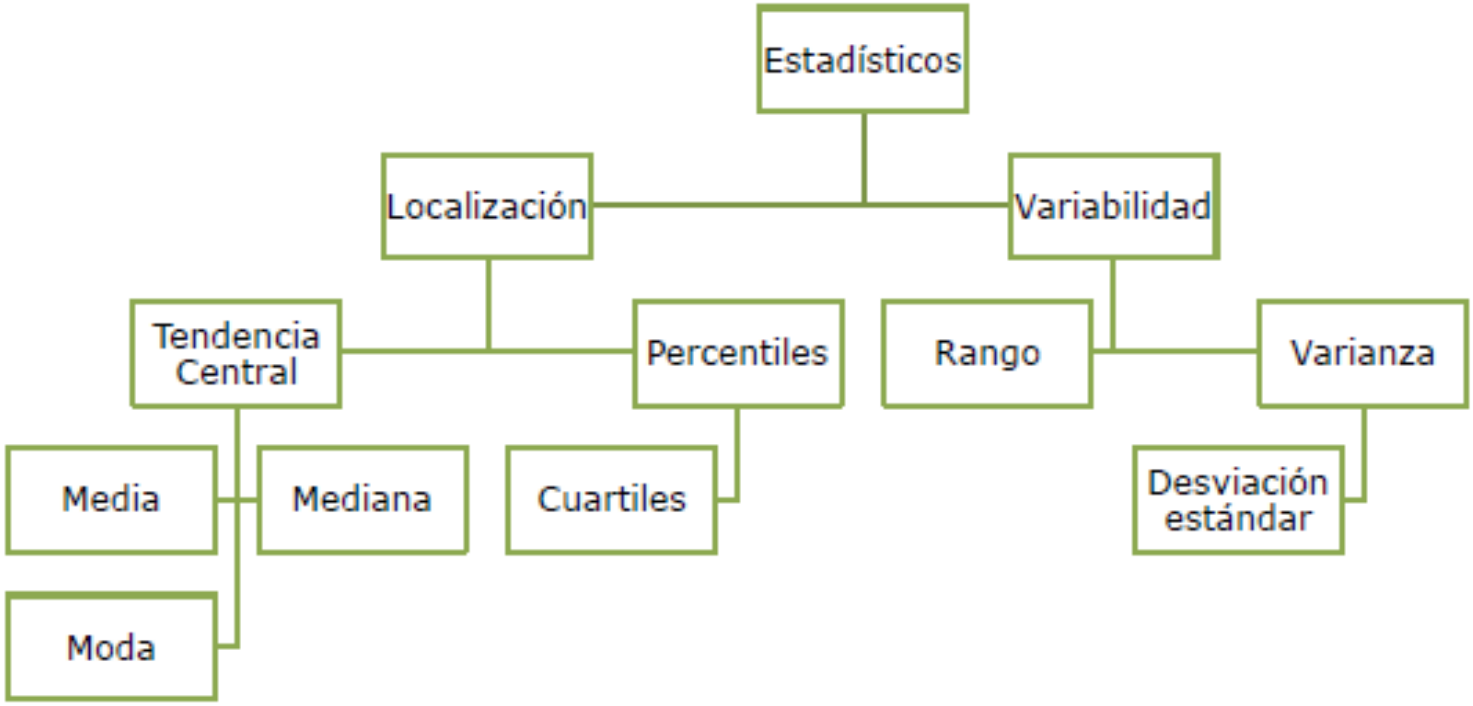
Para nuestro ejemplo, las calificaciones de Carlos tienen una desviación estándar de: 1.82 puntos.

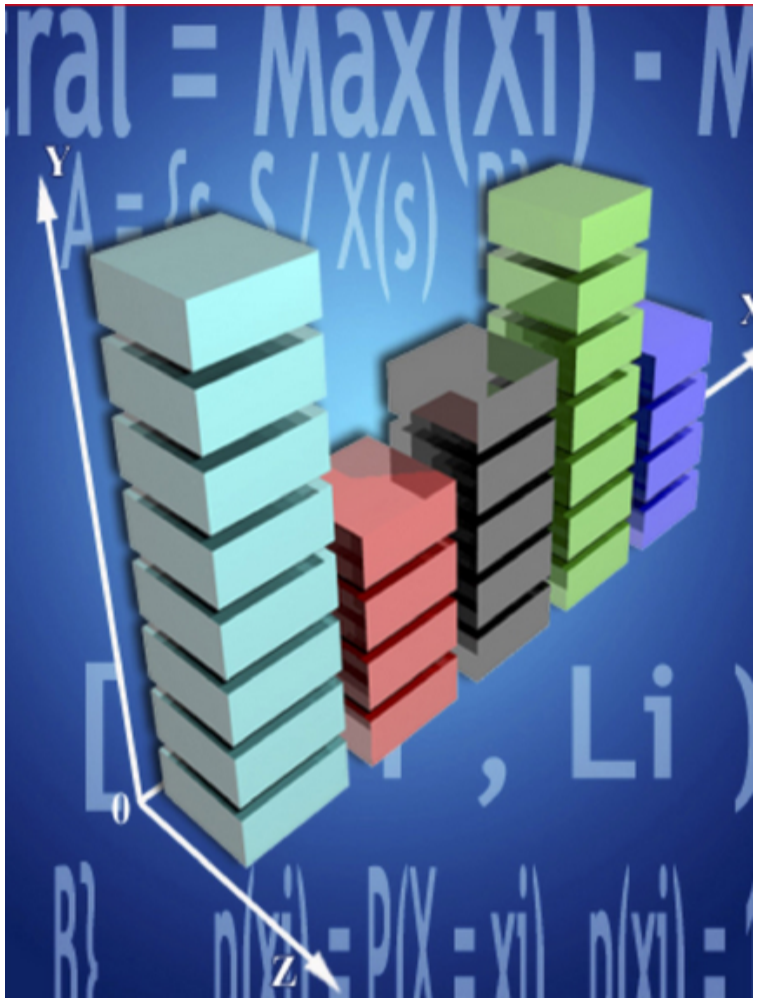
$$s = \sqrt{s^2} = \sqrt{3.33} = 1.82$$

Entre más grande sea la **desviación estándar de un conjunto de datos éstos se encontrarán más dispersos.**

¿Qué podrías decir de las desviaciones estándar correspondientes a las calificaciones de Beatriz y Alberto?

En resumen…


Estadísticos
Localización
Variabilidad
Tendencia Central
Percentiles
Rango
Varianza
Media
Mediana
Cuartiles
Desviación estándar
Moda



En este tema hemos revisado que las principales medidas de localización son las de tendencia central (media, mediana y moda), los percentiles y los cuartiles.

También se estudiaron las principales medidas de variabilidad como el rango, la varianza y la desviación estándar.

Las medidas de localización y las de variabilidad resultan complementarias para describir un conjunto de datos, así es común que se presenten la media y desviación estándar como forma de resumir un conjunto de datos que puede ser numeroso.

Bibliografía:

- Anderson, D., Sweeney, D., y Williams, T. (2004). Estadística para administración y economía (8ª. Ed.). México: Thompson.

- Mendenhall, W. (2002). Introducción a la probabilidad y estadística. México: Thompson.