2. <u>Mediana</u>

La mediana de una colección de datos, que previamente han sido *ordenados*, es aquél valor más central o que está más en medio en el conjunto de datos.

En otras palabras, la mediana es *mayor* que aproximadamente la mitad de los datos y *menor* que (aproximadamente) la otra mitad.

Así se tiene que aproximadamente 50% de las observaciones se encuentran por arriba y 50% (aproximadamente) por debajo de ella.

La mediana se denota *Md* (también algunos autores la denotan como $\tilde{X}$.

Ejemplo:

Los tiempos de los miembros de un equipo de atletismo en una carrera de 1,6 Km están dados en la siguiente tabla, calcule la mediana.

Miembro	1	2	3	4	5	6	7
Tiempo (en minutos)	4.2	9.0	4.7	5.0	4.3	5.1	4.8

En primer lugar se deben ordenar los datos: 4.2 4.3 4.7 4.8 5.0 5.1 9.0.

Mediana

Md **=** 4.8 minutos, es el valor que está en el centro de los datos.

Ejemplo:

Calcule la mediana para el número de pacientes tratados en la sala de emergencias de un hospital durante ocho días consecutivos:

Día	1	2	3	4	5	6	7	8
No. de pacientes	49	52	86	30	35	31	43	11

Los datos ordenados son: 86 52 49 43 35 31 30 11

Centro de los datos

$$Md = \frac{43+35}{2} = 39$$

Nota:

Si se tienen n observaciones ordenadas, la mediana es la observación que ocupa la posición $\frac{n+1}{2}$ cuando n es impar y la media de las observaciones que ocupan las posiciones $\frac{n}{2}$ y $\frac{n+2}{2}$ cuando n es par.

Medidas Descriptivas Numéricas

Ejemplo:

En ejemplo de los tiempos del equipo de atletismo, se obtuvo que Md = 4.8 minutos.

Miembro	1	2	3	4	5	6	7
Tiempo (en minutos)	4.2	9.0	4.7	5.0	4.3	5.1	4.8

Si se calcula la media aritmética se obtiene que $\overline{x}$ = 5.3 minutos.

¿Por qué la mediana y la media aritmética son tan distintas?

Propiedades de la mediana

i. La mediana es una medida de tendencia central de fácil comprensión pero que solamente toma en cuenta la posición que ocupan las observaciones y no el valor en sí de las mismas.

Esto hace que la mediana no sea susceptible de operaciones algebraicas y en consecuencia limita su utilidad, por ejemplo para fines de inferencia estadística.

ii. No se ve afectada ante la presencia de unos pocos *valores atípicos* y es por ello que se recomienda su uso en el caso de distribuciones marcadamente asimétricas.

3. Moda

La moda es el valor que más se repite, es decir el que aparece con mayor frecuencia.

La moda es el valor más común de los datos, se denota por *Mo* y viene expresada en las mismas unidades que los datos.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4.

En este caso el valor que más se repite es el 5, por tanto $Mo = 5$.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 6, 5, 4, 5, 2, 4, 4.

En este conjunto de datos existen dos valores que se repiten con la misma frecuencia: 4 y 5. Así, se tienen dos modas: $Mo_1 = 4$ y $Mo_2 = 5$.

Ejemplo:

Calcule la moda de los siguientes datos: 5, 3, 3, 5, 6, 2, 6, 4, 2, 4.

En este caso no existe la moda dado que no hay datos que se repitan más que otros.

Nota:

Cuando hay una sola moda la distribución de datos se llama *unimodal*;

Con dos modas *bimodal;*

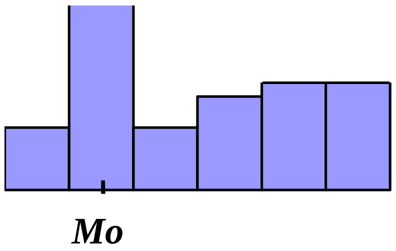
Con tres modas *trimodal;*

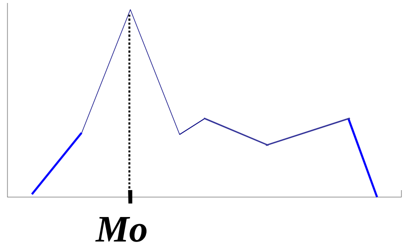
Con 4 o más modas se llama *polimodal* o *multimodal*.

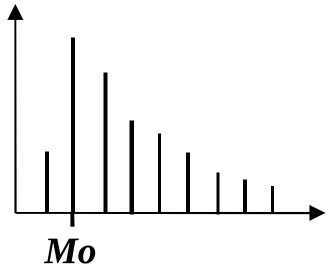
Si todos los valores se presentan la misma cantidad de veces, la distribución se llama *amodal*.

Representación gráfica de la moda:

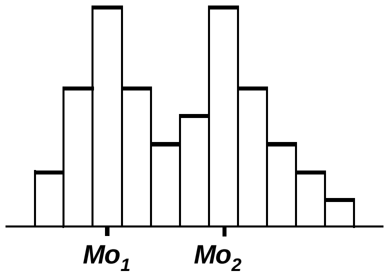
Distribuciones unimodales:



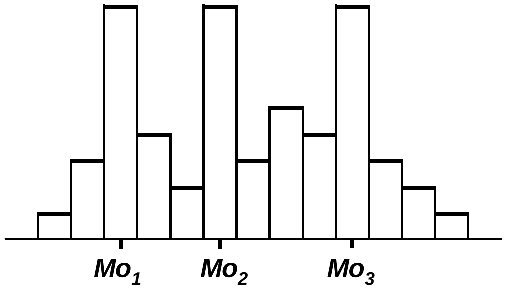




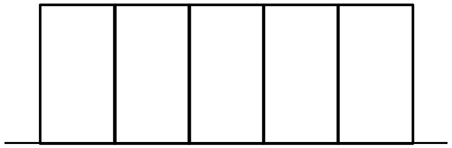
Distribución bimodal:



Distribución trimodal:



Distribución amodal:



***Propiedades de la Moda*:**

1. La moda en realidad no es una medida de tendencia central, sino más bien indica punto(s) de concentración de datos.
2. No es susceptible de operaciones algebraicas y de allí que su uso es limitado.
3. Es la única de las medidas descriptivas que puede utilizarse para datos cualitativos de cualquier tipo.
4. La moda es de utilidad en aquellos casos donde la naturaleza de los datos así lo indique.

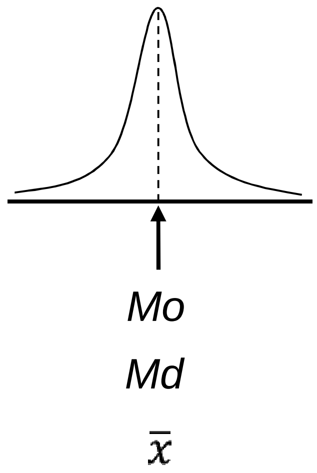
   <u>Ejemplo</u>:

   Para una fábrica de zapatos, el interés está en conocer la o las tallas más frecuentes en la población.

Relación entre la Media Aritmética, la Mediana y la Moda

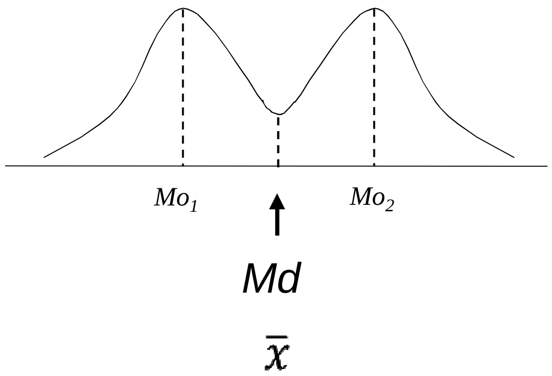
En función de la simetría de una distribución se presentan las siguientes relaciones entre esas tres medidas:

1. En distribuciones simétricas unimodales:



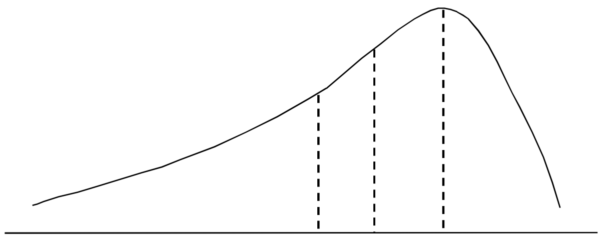
La media, la mediana y la moda coinciden

2. En distribuciones simétricas bimodales:



La media y la mediana son iguales pero no coinciden con las modas.

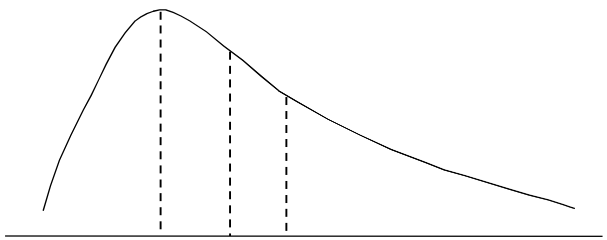
3. En distribuciones asimétricas negativas o por la izquierda:



$\bar{x}$ *Md* *Mo*

Se cumple que $\bar{x} < Md < Mo$

4. En distribuciones asimétricas positivas o por la derecha:



Mo *Md* $\bar{x}$

Se cumple que $\bar{x} > Md > Mo$

Selección de la medida de posición adecuada

Los siguientes factores deben tomarse en cuenta en el momento de la selección de la medida numérica apropiada para describir la posición o tendencia central de los datos:

1. De acuerdo al tipo de dato se puede utilizar una u otra medida de tendencia central. Las medidas que pueden aplicarse con cada tipo de dato son las siguientes:
   a. Datos Nominales:
      Moda
   b. Datos Ordinales:
      Moda y Mediana
   c. Datos Discretos:
      Todas
   d. Datos Continuos:
      Todas

2. Teniendo en cuenta lo anterior se recomienda tener presente los siguientes aspectos:

   a. ***La naturaleza de la distribución de los datos***. Gráficamente se puede observar la forma general en que se distribuyen los datos. Esto es determinante en la selección del promedio adecuado.

      - Si se trata de una *distribución simétrica* o aproximadamente simétrica, se sabe que la media, la mediana y la moda coinciden y en consecuencia se puede utilizar cualquiera de ellas.
      - Si la *distribución es asimétrica*, la media aritmética no va a ser adecuada y es preferible inclinarse por la moda o la mediana.

b. ***El concepto de tendencia central o de posición que interese reflejar en una situación dada.***

- Si interesa conocer el valor más común de una serie de datos como por ejemplo la estatura típica de las personas que ingresan al ejército, es necesario usar *la moda*.
- Si se desea ubicar a una persona en cuanto a su salario anual diciendo que gana por encima o por debajo de lo que gana la mitad de los trabajadores del país, entonces habrá que usar *la mediana*.
- Cuando interesa el total de datos o reflejar el punto de equilibrio de los mismos se utiliza *la media aritmética*.

c. ***Riesgos que se corren ante la presencia de valores atípicos.***

Si existen valores atípicos, hay que verificar si se incurrió en algún error en la recolección de la información o puede ser el alerta de alguna situación no esperada por el investigador.

Hay que tener presente que la media aritmética se ve seriamente afectada ante la presencia de valores atípicos y será necesario recurrir a alguna de las otras medidas conocidas.

c. ***Posibilidad de realizar inferencia estadística***

Cuando el análisis estadístico se realiza sobre una muestra de la población con la intención de generalizar a la totalidad, lo que se conoce como inferencia estadística, prácticamente la única medida de tendencia central utilizada hasta ahora satisfactoriamente es la media aritmética y esto se debe a que existe un fundamento teórico que la respalda.

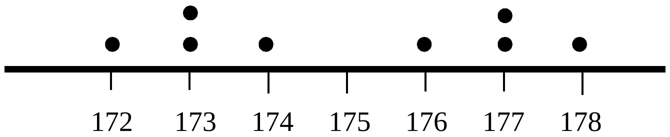
Medidas de Dispersión

Además de obtener la información que reúnen las medidas de tendencia central es muy conveniente tener conocimiento sobre el grado de dispersión o variabilidad que presentan los datos.

Las *medidas de dispersión* indican si los valores están relativamente cercanos uno del otro o si se encuentran dispersos.

Medidas Descriptivas Numéricas

En el ejemplo de los pesos de los cerdos se tenían los siguientes datos: 172, 177, 178, 173, 177, 174, 176, 173. El diagrama de puntos para esos valores es:



Si los cerdos de otro granjero tienen los siguientes pesos: 165, 182, 185, 168, 170, 173, 180, 177. Entonces el diagrama de puntos está dado por:

Obsérvese que ambos grupos de datos poseen la misma media aritmética y la misma mediana, $Md = \overline{x}$ = 175 Kg.

<u>Una medida de tendencia central, casi nunca es suficiente por sí sola, para resumir adecuadamente las características de un conjunto de datos</u>. Por lo general, es necesario, adicionalmente, una medida de la *dispersión* de los datos.

Clasificación de las medidas de dispersión

En general, se pueden clasificar las medidas de dispersión en *absolutas* y *relativas*.

Las *medidas de dispersión absolutas* son aquellas que vienen expresadas en las mismas unidades que los datos.

Las *medidas de dispersión relativas* no vienen expresadas en las unidades de los datos sino en porcentaje.

Entre las más usadas se encuentran:

Medidas de dispersión absolutas

- **Rango o recorrido**
- **Varianza**
- **Desviación Estándar**
- **Basadas en Percentiles**

Medida de dispersión relativa

- **Coeficiente de Variación**

Las medidas anteriores, con excepto el rango, toman la media como punto de referencia.

En cada caso un valor cero indica que no hay dispersión, mientras que la dispersión aumenta a medida que se incrementa el valor del indicador (varianza, coeficiente de variación, etc.)

1. **Rango o recorrido**

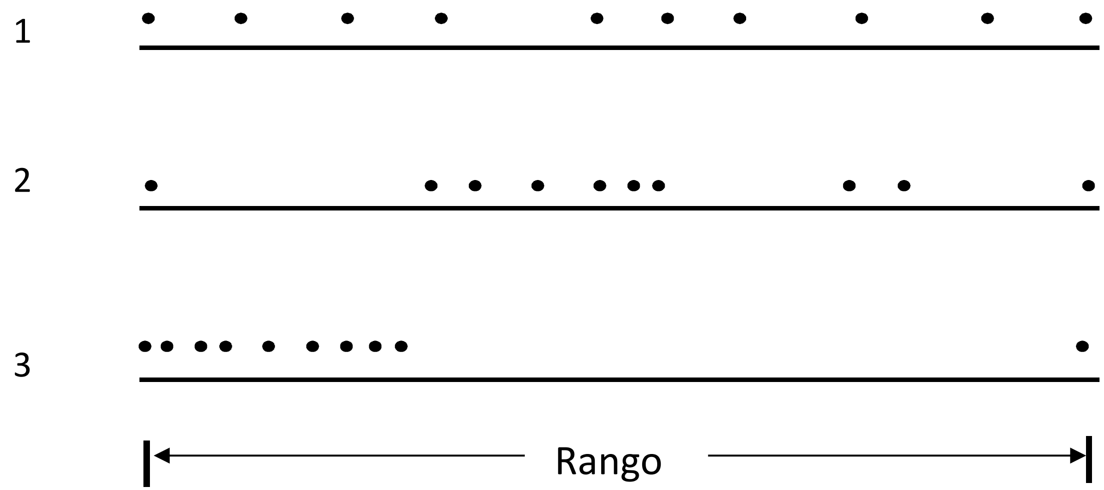
   Se concentra en el valor máximo y mínimo de la colección de datos y viene dada por:

$$R = \text{Valor máximo} - \text{Valor mínimo}$$

La principal limitación del rango es que considera solamente los valores extremos de los datos, y no proporciona información respecto a los demás valores.

Medidas Descriptivas Numéricas

En el siguiente ejemplo se presentan tres conjuntos de datos bastante diferentes, que poseen el mismo rango:



Debido a los problemas que se observan en el ejemplo, el rango tiene una limitada utilidad ya que no resulta una medida de dispersión confiable.

2. **Varianza**

Supóngase que $x_1, x_2, \ldots, x_n$ son las observaciones de una muestra aleatoria, cuya media es $\overline{x}$.

Dado que se está interesado en analizar la dispersión de estos valores, será natural fijarse en sus distancias con respecto a la media, esto es, en las diferencias:

$$x_1 - \overline{x}, x_2 - \overline{x}, \cdots, x_n - \overline{x}$$

Algunas de estas diferencias serán positivas y otras negativas.

Sabemos que la suma de estas diferencias es igual a: _____

Para analizar la dispersión de los datos, no interesa el signo de las diferencias. Así se tratará una diferencia negativa exactamente igual que una diferencia positiva de la misma cantidad.

Una forma de conseguir este objetivo consiste en fijarse, no en las diferencias, sino en sus cuadrados:

$$(x_1 - \overline{x})^2, (x_2 - \overline{x})^2, \cdots, (x_n - \overline{x})^2$$

El promedio de los cuadrados de las diferencias proporciona una medida de la dispersión que se conoce con el nombre de *varianza.*

Definición de varianza

Dada una colección de datos $x_1, x_2, \ldots, x_n$, cuya media aritmética es $\overline{x}$, se define la *varianza* de esos datos como el promedio de las diferencias elevadas al cuadrado de cada uno de esos valores con respecto a su media. Es decir:

$$S_*^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n}$$

Nota

De la fórmula anterior se deduce que:

i. La varianza es un indicador que mide la dispersión de los datos con respecto a su media aritmética

ii. Mientras más alejados estén los valores de su media mayor será el valor de la varianza y mientras más concentrados se encuentren alrededor de su media, menor será el valor de la varianza.

iii. La varianza nunca es negativa, ya que se están sumando cantidades elevadas al cuadrado.

iv. El valor mínimo que puede tomar es cero, el cual se logra cuando todos los valores son iguales entre sí, es decir, que no existe variabilidad entre ellos.

Ejemplo:

Los datos de los pesos de los cerdos son: 172, 177, 178, 173, 177, 174, 176, 173 y la media es 175 Kg. El cálculo de la varianza es como sigue:

$$S_*^2 = \frac{(172-175)^2+(177-175)^2+(178-175)^2+(173-175)^2+(177-175)^2+(174-175)^2+(176-175)^2+(173-175)^2}{8}$$

$$S_*^2 = 4,5 \text{ Kg}^2$$

Si se desarrolla la fórmula anterior, se obtiene otra expresión equivalente de la varianza, más cómoda de calcular y además reduce los errores de redondeo:

$$S_*^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\overline{x}\right)^2$$

Calculando nuevamente la varianza de los pesos de los cerdos, con esta última fórmula:

$$S_*^2 = \frac{172^2+177^2+178^2+173^2+177^2+174^2+176^2+173^2}{8} - 175^2$$

$$S_*^2 = 4,5 \text{ Kg}^2$$

Nótese que la varianza viene expresada en las unidades de los datos pero elevadas al cuadrado. Por esta razón, la varianza resulta difícil de interpretar. Para solucionar esta situación se define la *desviación estándar*.

3. **Desviación Estándar (Desviación Típica)**:

La *desviación estándar* o *desviación típica* de una colección de datos, denotada por S_*, se define como:

$$S_* = +\sqrt{S_*^2}$$

La cual viene dada en las mismas unidades de los datos.

Ejemplo:

Tomando el ejemplo anterior se tiene que:

$$\begin{aligned} S_* &= \sqrt{4,5 \text{ Kg}^2} \\ &= 2,12 \text{ Kg} \end{aligned}$$

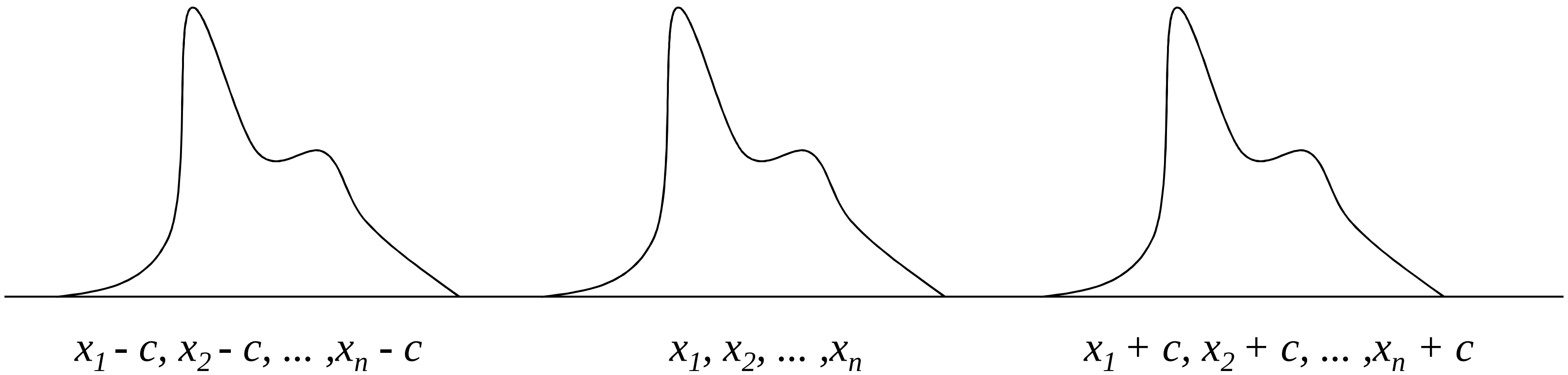
Ejercicio:

Calcular la varianza y la desviación estándar para las distribuciones de frecuencias de las variables peso, número de hermanos, estatura, ingreso mensual del hogar, número de visitas a la discoteca y número de visitas al cine.

Propiedades de la varianza y la desviación estándar

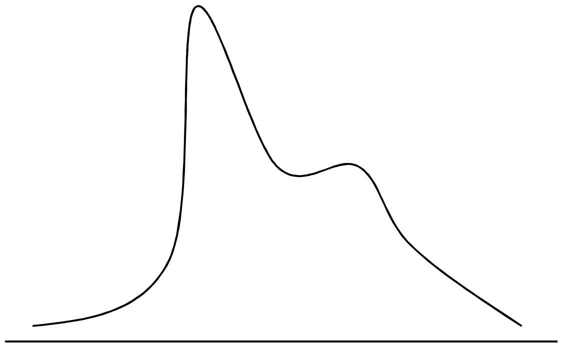
Sea $x_1, x_2, \ldots ,x_n$ una colección de datos, cuya media, varianza y desviación estándar son $\overline{x}$, S_*^2, y S_* respectivamente.

1. S_*^2 y S_* son no negativas, es decir $S_*^2 \geq 0$ y $S_* \geq 0$ para cualquier conjunto de datos.
2. Si cada uno de los datos $x_1, x_2, \ldots ,x_n$ es igual a un mismo valor fijo o constante c, entonces la varianza y la desviación estándar son iguales a cero.
3. Si a cada uno de los datos originales se le suma un mismo número real c, positivo o negativo, entonces la nueva colección de datos que se origina $x_1+c, x_2+c, \ldots ,x_n+c$ tiene la misma S_*^2 y S_* que los datos originales.

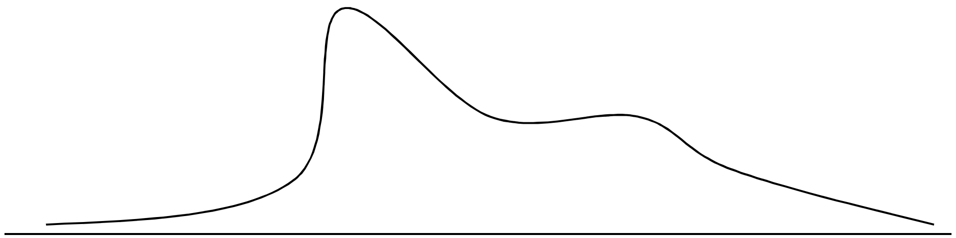


Medidas Descriptivas Numéricas

4. Si cada uno de los datos se multiplica por un mismo número real cualquiera *c*, la varianza y la desviación estándar de los "nuevos datos" x_1*c, x_2*c, ... ,x_n*c vienen dadas por $c^2 S_*^2$ y $|c| S_*$ respectivamente.



$x_1, x_2, \ldots ,x_n$



$x_1 * c, x_2 * c, \ldots ,x_n * c$

Ejercicio:

En un estudio realizado en un hospital se determinó que se gastaba en medicinas un promedio de Bs. 80.000 semanalmente por paciente con una desviación estándar de Bs. 15.000.

a. Si se produce un aumento del 100% en el precio de las medicinas, cuanto será el gasto promedio por paciente y cuanto será la varianza.
b. Cuanto será el gasto promedio por paciente y cuanto será la desviación estándar si el aumento es del 20%.

Varianza muestral y varianza poblacional

Al definir la varianza se ha estado suponiendo que la colección de datos $x_1, x_2, \ldots, x_n$ constituye una muestra de tamaño *n* de una población y que $\overline{X}$ es la media de esa muestra. La *varianza poblacional*, denotada por σ^2, de una población de *N* elementos cuya media poblacional es μ, se define por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} = \frac{\sum_{i=1}^{N}x_i^2}{N} - \mu^2$$

y la *desviación estándar poblacional* es:

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

Nota:

La varianza muestral también puede definirse como:

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

Se utiliza con la finalidad de, además de tener fines descriptivos, realizar inferencias sobre una población usando S^2 y no S_*^2 por cuanto se demuestra que S^2 es un mejor estimador de la varianza poblacional σ^2 que S_*^2 como se verá en el tema de *estimación*.

Coeficiente de variación

La *medida de dispersión relativa* más conocida es el *coeficiente de variación*.

En algunas ocasiones es de interés comparar la dispersión de dos colecciones de datos.

Si los datos están medidos en las mismas unidades y las respectivas medias aritméticas son iguales o muy parecidas, es posible utilizar la desviación estándar.

Si esto no se cumple, no se puede utilizar la desviación estándar para comparar las dispersiones de los dos grupos de datos.

Una medida de dispersión que permite la comparación de la dispersión en cualquier situación, que no viene expresada en ninguna unidad es el *coeficiente de variación*.

Definición de Coeficiente de Variación

El *coeficiente de variación* se define como:

$$CV = \frac{S_*}{\bar{x}} * 100\%$$

El *coeficiente de variación* es la proporción o porcentaje de la media que representa la desviación estándar.

Obsérvese como la fórmula anterior proviene de una regla de tres simple:

$$\bar{x} \rightarrow 100\%$$

$$S_* \rightarrow ?$$

Si por ejemplo el *CV*=20%, significa que la desviación estándar representa el 20% del valor de la media aritmética.

Ejercicio:

Supóngase que se desea comparar las dispersiones de los sueldos de los empleados de las empresas "Cervecería El Cóndor" y "Aguardiente Tropical". Los sueldos promedio para estas empresas son Bs. 70000 y Bs. 25000 respectivamente; las desviaciones estándar correspondientes son Bs. 30000 y Bs. 3000.

Ejercicio:

En una encuesta sobre bienes raíces en la Urbanización Santa Cecilia de una ciudad, se obtiene entre otras cosas, información sobre el valor actual de la casa y el tamaño del lote de terreno. Se está interesado en determinar si el valor de avalúo tiene mayor variabilidad que el tamaño del lote. De la mencionada encuesta se consigue lo siguiente:

Valor de la casa	**Tamaño del terreno**
$\overline{X}$ = 1.550.000,000 Bs.	$\overline{X}$ = 650 mts^2
S_* = 500.000.000 Bs.	S_* = 350 mts^2

Ejercicio:

Compare la dispersión de la `variable peso de los varones con la dispersión de la variable peso de las hembras.