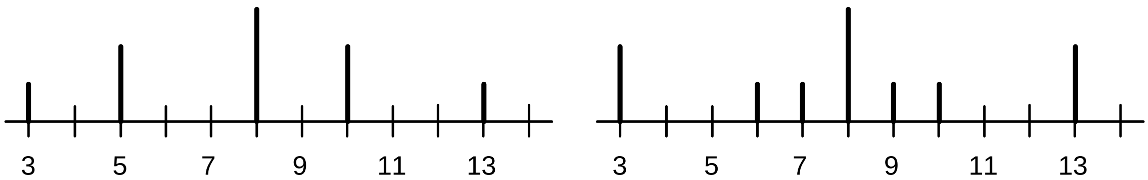
MEDIDAS DE DISPERSION

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.

Si se tiene los siguientes grupos de datos:

Grupo 1	3	5	5	8	8	8	8	8	10	10	13
Grupo 2	3	3	6	7	8	8	8	9	10	13	13

para los cuales $\overline{x}(\text{grupo } 1) = 8$ y $\overline{x}(\text{grupo } 2) = 8$ y se los representa



se advierte que las distribuciones son "diferentes" en cuanto a su dispersión, aún cuando en ambos casos las distancias entre la menor y la mayor observación de cada grupo son iguales (distancia de 10 puntos).

Se ve que 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.

Sería bueno también que todos los datos fueran tenidos en cuenta para describir esa dispersión y por supuesto, que la medida lograda pudiera interpretarse en términos del problema (esto es, con las unidades de medida empleadas al obtener los datos). Debería ser útil esa medida de dispersión para poder decidir si una observación tiene un valor "lógico" o es "raro".

Con todos estos criterios, enumeramos las medidas que indican o manifiestan la dispersión de los datos.

Procederemos de la misma forma en que lo hicimos para las medidas de posición, definiendo, analizando, ejercitándonos.

RECORRIDO, RANGO O AMPLITUD

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

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

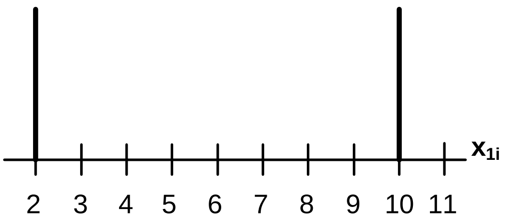
Es la más sencilla y directa medida de dispersión.

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.

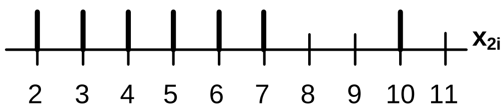
Si se tienen los siguientes grupos de datos:

G_1	2	2	2	2	10	10	10
G_2	2	3	4	5	6	7	10
G_3	2	6	6	6	6	6	10

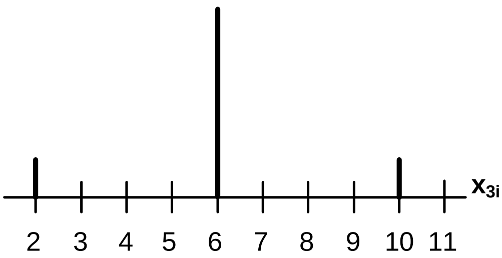
y se los representa, se observa:



G_1 $R_1 = 10 - 2 = \mathbf{8}$



G_2 $R_2 = 10 - 2 = \mathbf{8}$



G_3 $R_3 = 10 - 2 = \mathbf{8}$

lo que confirma lo comentado antes.

•

VARIANCIA O VARIANZA

Se define como media de los cuadrados de los desvíos con respecto a la media aritmética de la variable. Como se analizó previamente, uno de los problemas del D_M es que no se tienen en cuenta los signos de los desvíos ya que el efecto del valor absoluto es que las observaciones se reacomodan todas a la derecha del promedio. Pero si se toman los cuadrados de los desvíos se evita que se compensen los positivos con los negativos sin "obligar" a la información a reacomodarse y entonces es posible hablar con propiedad de la variabilidad alrededor de la media.

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

Tiene ventajas:

- Permite la interpretación precisa de la dispersión de los datos.
- Puede usarse "matemáticamente" para estudios de inferencia.

Cuando se tiene una muestra y se trata sólo de describir la variabilidad de esa muestra se calcula:

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

Sin embargo, cuando el interés radica en estimar a la varianza de la población a partir de un valor obtenido de la muestra, la definición no es apropiada, puesto que S^2_m en realidad tiende a subestimar la varianza de la población, sobre todo si el tamaño de la muestra es pequeño. Esto es, da una estimación sesgada o viciada, que no es lo que se quiere (no es buen estimador).

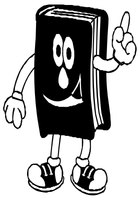
En forma intuitiva se entiende como una estimación insesgada a una que se iguala, en promedio, al valor del parámetro.

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

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

que sí es un buen estimador de la varianza poblacional.

Cuando se desarrolle el tema Esperanza Matemática de una variable aleatoria, se comprobará lo que ahora se indica.
RECUÉRDELO



IMPORTANTE: Aún cuando la diferencia entre trabajar con “n” o con “n-1” parezca mínima o no importante, es necesario recordar que existen fundamentos de la Teoría Estadística que nos permiten asegurar que usar indiscriminadamente la varianza dividiendo por “n”, tamaño de la muestra, **es un error conceptual.**
Analice detenidamente los conceptos, acostúmbrese a criticar, aún a los autores propuestos en la bibliografía y a otros, y también a los que “usan” la Estadística en su tarea diaria, y aseguran que *como los cálculos no se ven demasiado afectados..., entonces puedo...y...*-
NO! No es así. Estamos desarrollando conceptos y fundamentos de la Teoría Estadística, no aprendiendo recetas de trabajo... Esto significa que aún cuando a veces no detallemos todos los fundamentos por falta de tiempo o porque sobrepasaríamos el nivel adecuado para esta cursada, estos existen y deben ser respetados.

A la varianza poblacional la identificamos como σ^2 (sigma cuadrado) y es:

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

No olvide entonces que:

- $\mathrm{Var(x)} = \sigma^2$: Varianza poblacional, parámetro.
- S^2 : estimador no viciado de σ^2; se calcula justamente cuando a través de la muestra se desea hacer alguna inferencia.
- S^2_m: estimador viciado de σ^2; se calcula únicamente cuando el único objetivo del estudio que se realiza es describir la muestra.
- μ : media aritmética poblacional (parámetro)

Es posible pasar de S^2_m a S^2 y viceversa sólo con considerar que

$$n.S_m^2 = (n-1).S^2$$

y entonces

$$\boxed{S^2 = \frac{n}{n-1}.S_m^2}$$

En la expresión del S^2 aparece como divisor "(n-1)" y no "n" mientras que en σ^2 aparece "N", ¿a qué se debe esto?

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

En

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

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.

Por ejemplo, si se tiene n = 10 valores y se hacen los desvíos iguales a

3 - 5 - 2 -(-1) -(-4) - 7 -(-2) - 2 - 3 - ____ ?

este último, el décimo desvío, debe ser igual a -15 para que la suma sea cero,

3 + 5 + 2 - 1 - 4 + 7 - 2 + 2 + 3 - 15 = 0

En cambio, si se tuviera

$$\frac{\sum\left(x_i - \mu\right)^2}{N}$$

ya no es cierto que

$$\sum\left(x - \mu\right) = 0$$

por lo que ya no hay restricciones y los N desvíos son independientes.

Esto explica que σ^2 tiene "N"g.l. y S^2 sólo "(n-1)"g.l.

Esto justifica también que intuitivamente se admita que del total de observaciones (y desvíos) que se tienen, se debe "pagar" un g.l. por cada parámetro que no se conoce y se debe estimar con los datos de la muestra.

CUIDADO... Recuerde lo que se dijo sobre las justificaciones y los fundamentos teóricos...

Se ha visto hasta aquí cómo se calcula la Varianza y cuáles son sus ventajas.

Se deben mencionar ahora las desventajas:

- 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.

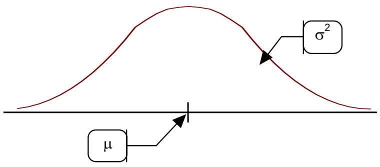
$$\boxed{S_m = \sqrt{S_m^2} = \sqrt{\frac{1}{n}.\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}$$

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.**

En la naturaleza y en la industria es típico encontrar variables con distribuciones bastante simétricas, con una gran cantidad de observaciones cercanas a un valor **lógico o central** (la media aritmética) y pocas veces diferentes a éste valor ya sea en sentido positivo como negativo.

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



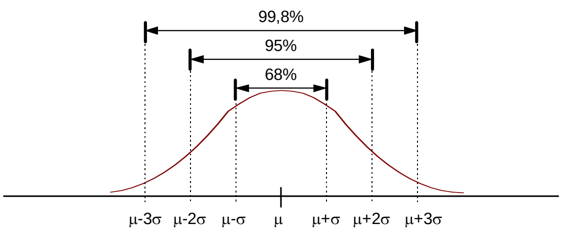
Tienen una forma muy parecida a la de una campana.

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 $(\mu - \sigma)$ y $(\mu + \sigma)$, se encuentra alrededor del 68% de la población; entre $\mu - 2\sigma$ y $\mu + 2\sigma$ se encuentra cerca del 95% de la población; y entre $\mu - 3\sigma$ y $\mu + 3\sigma$ 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 $x - S$ y $x + S$, el 95% entre $x - 2.S$ y $x + 2.S$ y casi todas las observaciones entre $x - 3.S$ y $x + 3.S$.

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.**

(29) Le pedimos ahora que, observando detenidamente las fórmulas presentadas y dados los siguientes datos

7 4 5 3 8 5 6 6 3 9 7 5 6 7 9

calcule:

- la varianza que puede emplear sólo para describir
- la desviación estándar poblacional
- el valor de S

Le aclaramos que este ejercicio tiene como único objetivo que usted calcule "a mano", y empleando la definición, los valores que se le piden. Más adelante se le presentarán fórmulas de trabajo, de uso mucho más sencillo.

Continuamos presentando medidas de dispersión.

COEFICIENTE DE VARIACIÓN

Supóngase que un grupo de leñadores está haciendo un cálculo de la altura de <u>dos</u> árboles.

Obviamente no todas las mediciones que se realizan son iguales. Es posible tener una distribución de mediciones de cada uno de los 2 árboles, basadas en las mediciones del grupo de leñadores.

Esto es: los leñadores hacen mediciones (en metros) sobre un árbol para obtener una distribución y lo mismo hacen sobre el otro árbol.

Se puede calcular el S_m^2 en cada caso, es decir para cada uno de los dos árboles y se tendrían, por ejemplo

S_m(árbol 1) = 6m
S_m(árbol 2) = 0,30 m

En apariencia, la altura del segundo árbol ha sido determinada con mayor precisión.

Pero ¿ésta es una interpretación correcta si el árbol 1 mide en promedio 120m de altura y el árbol 2 sólo 3m de altura?

Es evidente que si se quiere evaluar la precisión de los dos grupos de datos teniendo en cuenta la magnitud de las cantidades originales, será necesario relacionar la variabilidad con algún indicador de esa magnitud, por ejemplo con la media aritmética. Se tendrá así una medida de la **variabilidad relativa** de los grupos de datos.

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

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

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 los que se ha trabajado con distintas unidades de medida.

Si se calcula C.V. en los grupos de mediciones de los dos árboles realizadas por los leñadores se tiene:

$$C.V_{(arbol1)} = \frac{6m}{120m}.100 = 5\%$$

$$C.V_{(arbol2)} = \frac{0{,}30m}{3m}.100 = 10\%$$

lo que indica que fue mayor la precisión relativa con que se midió la altura del árbol 1.

Analicemos otra situación.

Supóngase que se tienen dos grupos de animales vacunos con las mismas características, pero distinto peso promedio.

Se toman 100 animales en cada grupo y se obtiene:

Grupo 1	**G**rupo 2
$n_{(Grupo1)} = n_1 = 100$ animales	$n_{(Grupo2)} = n_2 = 100$ animales
$\overline{x}_{(grupo1)} = \overline{x}_1 = 160 kg.$	$\overline{x}_{(grupo2)} = \overline{x}_2 = 310 kg.$
$S^2_{m(grupo1)} = S^2_{m1} = 100$	$S^2_{m(grupo2)} = S^2_{m2} = 100$

$$C.V_{(grupo\ 1)} = \frac{10}{160}.100 = 6{,}25\%$$

$$C.V_{(grupo\ 2)} = \frac{10}{310}.100 = 3{,}225\%$$

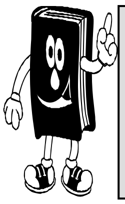
y se advierte que es menos "variable" el grupo dos. Se ve que la misma variabilidad absoluta tiene distinto "peso" en los dos casos.

Teniendo en cuenta lo que hemos desarrollado, esperamos que:

(30) Analice los siguientes ejemplos:

....a) un error de 1/16 de pulgada al unir dos segmentos de un puente podría ser aceptable, pero sería intolerable ese mismo error en uno de los mini-componentes de una cápsula espacial.

....b) no es lo mismo una variabilidad de 1cm. en la longitud de la cola de los ratones que en la longitud de la cola de los elefantes.



NOTA: Se recomienda a los alumnos de Profesorado y Licenciatura en Matemática, a los de Computación y a los de Electrónica prestar atención a las demostraciones de las propiedades y a los procedimientos.
Recurra a los textos citados en la bibliografía para completar los desarrollos.

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

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

Así como en su momento trabajó con las propiedades de la media aritmética, empleando ahora la expresión formal de la varianza y simplemente reemplazando, se espera que...

(31) Demuestre las propiedades de la varianza.

(32) Retome los datos con los que trabajó en la actividad (8)a,b y c, y calcule rango,desvío estándar y coeficiente de variación.

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.

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

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

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

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

$$\boxed{\begin{array}{l} S_m^2 = \dfrac{1}{n}.\left[\sum x_i^2 - \dfrac{\left(\sum x_i\right)^2}{n}\right] \\ S_m^2 = \dfrac{1}{n}\left[\sum x_i^2 - n.\overline{x}^2\right] \end{array}}$$

Si se trabaja con datos agrupados se parte de

$$S_m^2 = \frac{1}{\sum^{k} f_i}\sum^{k}\left(x_i - \overline{x}\right)^2.f_i$$

y se llega a

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

o bien

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

Simplemente para practicar el uso de las fórmulas de trabajo le pedimos que:

RETORNAMOS AL EJERCICIO RESUELTO...

(36) Retomamos la actividad (10) a fin de completar la descripción calculando las medidas de dispersión.

Recuerde que...

La medida mas sencilla y rápida de calcular es el **rango:**
R = 12,82 - 1,20 = 11,62 centésimas de segundo

(Observe que éste es el rango corregido que se obtuvo cuando se formaron los intervalos; no se tienen ahora los valores originales) (Obsérvese la tabla)

Como es sabido, el rango es una medida sencilla pero sólo informa que entre el mayor y el menor de los tiempos de ignición (entre 12,82 y 1,20 seg/100) hay 11,62 centésimos de segundo , pero sólo esto.

Es posible calcular la **varianza** usando la fórmula de trabajo para datos agrupados.

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

$$= \frac{1}{80}\left[2{,}904 - \frac{(431{,}32)^2}{80}\right] =$$

$$= \frac{1}{80}\left[2{,}904 - 2325{,}4617\right] = \frac{578{,}53822}{80} = 7{,}2317277$$

Este resultado, como se sabe, es S_m^2 = 7,23 centésimos de segundo al cuadrado, imposible de interpretar.

Se calcula entonces

$$S_m = \sqrt{S_m^2} = \sqrt{7{,}2317277} = 2{,}6891871.....?$$

que son ... ; esto es, la variabilidad de los datos observados alrededor de su media aritmética es de 2,69 ...

Por lo tanto, aproximadamente el 68% de los datos se encuentra entre $\boxed{\bar{x} \pm 1.S_m}$ es decir entre 2,70 ... y 8,08 ...

Si se quiere conocer la **variabilidad relativa a la media aritmética** se calcula

$$C.V. = \frac{2{,}69 seg/100}{5{,}39 seg/100}.100 = 49{,}90\%$$

que es un valor bastante grande.

PROBLEMAS – EJERCICIOS (a resolver)

Ahora está en condiciones de resolver las siguientes actividades

(37) Analice la situación planteada:

Suponga que se tienen dos grupos de jóvenes en los cuáles se midió el número de pulsaciones por minuto y la capacidad vital respectivamente. Se obtuvo la siguiente información:

<u>**G**rupo 1:</u> Se tomó el número de pulsaciones por minuto de 600 jóvenes de 20 años y se obtuvo $\overline{x}_1 = 74$ pulsaciones por minuto y $S_1 = 11$ pulsaciones por minuto.

<u>**G**rupo 2:</u> Se tomó la capacidad vital en 20 jóvenes de 18 años y se obtuvo $\overline{x}_2 = 4$ litros y $S_2 = 0,6$ litros.

Si el objetivo es comparar la variabilidad de los dos grupos ¿qué medida calcularía y por qué?.

(38) Complete las actividades (19)(25)(28)y(29) de la primera parte y luego las actividades (19) (20) y (28) de esta parte, calculando e interpretando las medidas de dispersión de interés.

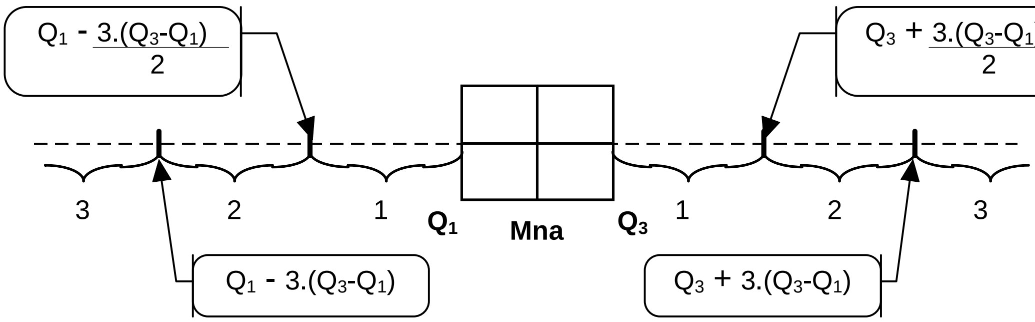
(39) ¿Qué opinión le merece la variación relativa de los dos grupos de datos mencionados en las actividades 3 y 4? Analice.

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.

Consiste en una caja con dos émbolos y las zonas en que caen las observaciones quedan divididas como se indica a continuación:



Las observaciones que caen en la zona:

1. son llamadas observaciones adyacentes.
2. son llamadas observaciones lejanas.
3. son llamadas observaciones muy lejanas.

A partir de la zona de las observaciones lejanas, las observaciones suelen llamarse **outlier** (observaciones: que caen fuera, excepcionales, sospechosas y algunos hasta las llaman groseras).

El encontrar un outlier es de gran información para el experimentador pues indica que se ha cometido algún error al recopilar la información para la cual controla los valores transcriptos y/o calculados. También puede ser que verdaderamente algún valor de la variable tenga un valor extremo (ya sea por exceso como por defecto) como por ejemplo: si se realiza un test de coeficiente intelectual a 15 niños que ingresarán a 1er. grado y uno de ellos tiene coeficiente de "genio" entonces el resultado que éste obtenga en el test estará muy alejado (por exceso) de los resultados obtenidos por los restantes niños.

Con unos pocos cálculos y el diagrama se observan medidas de:

Posición Mna, Q_1, Q_3

Variabilidad $\frac{Q3 - Q1}{2}$ (semi-oscilación intercuartílica)

$Q3 - Q1$ (oscilación intercuartílica)

Simetría o asimetría comparando la medida de la distancia entre Q_1 y Mna con la medida de la distancia entre Mna y Q_3.

- Si dist(Mna, Q_1) = dist(Q_3, Mna) $\Rightarrow$ dist. simétrica
- Si dist(Mna, Q_1) < dist(Q_3, Mna) $\Rightarrow$ dist. asimétrica positiva
- Si dist(Mna, Q_1) > dist(Q_3, Mna) $\Rightarrow$ dist. asimétrica negativa

Ejemplo: suponga que se toman 21 precios de un mismo producto y que a medida que se recoge la información se la fue colocando en un diagrama de tallos y hojas logrando.

	1	2	3	0				
	3	2	4	5	9			
	8	2	5	5	5	5	8	9
13	12	2	6	0	0	5	8	
9		2	7	4	5	9		
6		2	8	9	9	9		
3		2	9	0	9			
1		3	0					
1		3	1					
1		3	2					
1		3	3					
1		3	4	9				

Se calcula

$$Q_1^0 = \frac{n+3}{4} \qquad Q_1^0 = \frac{24}{4} = 6$$

del mismo modo

$$Mna^0 = \frac{n+1}{2} \qquad Mna^0 = \frac{21+1}{2} = 11$$

Una manera de encontrar Q_1, Q_3 y Mna una vez calculadas Q_0^1 y Mna^0 es colocar una columna auxiliar a la izquierda acumulando en cada línea la cantidad de observaciones halladas hasta llegar a superar la mitad, tanto de arriba hacia abajo, como de abajo hacia arriba.

Salta a la vista que:

- la 6ta. observación en forma descendente es 255 $\Rightarrow$ Q_1 = 255
- la 6ta. observación en forma ascendente es 289 $\Rightarrow$ Q_3 = 289
- la observación central (debido a que la cantidad de observaciones es impar) es 265 $\Rightarrow$ Mna = 265

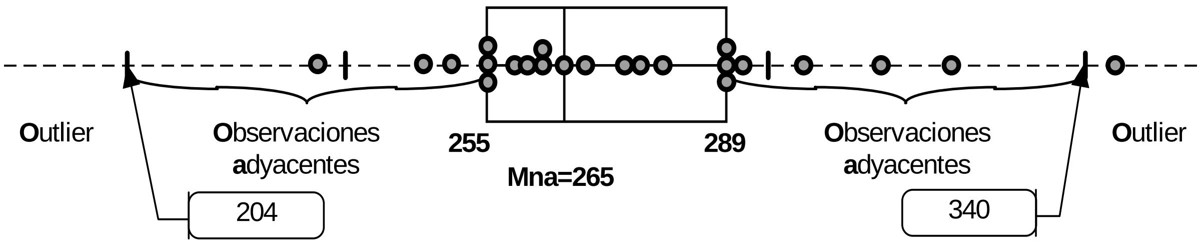
Con lo que se está en condiciones de calcular
Q_3 - Q_1 = 289 - 255 Q_3- Q_1 = 34 y los puntos

$$Q_1 - \frac{3(Q_3 - Q_1)}{2} = 255 - \frac{3.(34)}{2} = 204$$

$$Q_3 + \frac{3.(Q_3 - Q_1)}{2} = 289 + 51 = 340$$

$$Q_1 - 3(Q_3 - Q_1) = 255 - 3.(34) = 153$$

$$Q_3 + 3(Q_3 - Q_1) = 289 + 102 = 395$$



Si se colocan las observaciones (●) en el diagrama se observa que:

- sólo existe 1 (un) outlier.
- las observaciones se encuentran "bastante" concentradas, es decir hay "poca" dispersión.
- la distribución es asimétrica positiva (pues dist(Mna,Q_1) es menor dist(Q_3, Mna) lo que se observa gráficamente).
- las observaciones se encuentran centradas alrededor de 265.

IMPORTANTE

Notó que hay errores en el desarrollo?
Corríjalos.
Analice, critique, reconstruya correctamente.

Ya está en condiciones de trabajar con este método interesante propuesto por Tuckey.

(40) Dadas las siguientes observaciones provenientes de mediciones en milímetros

21	35	2	13	25	31	28
17	14	57	24	22	30	13

Hacer un diagrama "Box-Plot" y obtener alguna conclusión basándose en la información que muestra.

PROBLEMAS – PREGUNTAS – EJERCICIOS
(a resolver)

Se propone ahora una actividad integradora de contenidos. Responda individualmente para auto evaluarse y luego discuta con sus compañeros.

1. ¿Qué medidas descriptivas puede utilizar en una escala nominal?

2. Defina media aritmética, mediana y modo. De semejanzas y diferencias.

3. Lea e interprete el artículo "El rey estadístico", cuya copia se adjunta en el “Bloque Anexo”.

4. Una media aritmética puede no ser una buena medida de posición. Explique cuándo y ejemplifique.

5. Relacione los tres promedios principales cuando la distribución es simétrica y cuando es asimétrica.

6. Proporcione ejemplos de su disciplina en los cuáles usted espera una distribución asimétrica y analice.

7. Describa un problema de su disciplina en el cual la media aritmética no le proporcione suficiente información. Analice.

8. En cada uno de los siguientes casos, explique si el promedio al cual se hace referencia es la media aritmética, la mediana o el modo.

....**a)** Puede calcularse a partir de una distribución de frecuencias con intervalos desiguales.
....**b)** Puede calcularse a partir de una distribución de frecuencias con intervalos extremos abiertos.
....**c)** Se toma en consideración todos los valores para su cálculo.
d) No da información sobre la distribución en general sino que sólo está "localizada".
e) Los valores de los elementos extremos no influyen.
f) Puede no ser único en una distribución.
g) Indefectiblemente los datos deben estar ordenados de acuerdo con su magnitud.

9. ¿Tiene claro lo que significa mediana de orden ¿Cómo la ubica gráficamente? ¿Cómo la interpreta?

10. Distinga como procede para calcular la mediana cuando tiene pocos datos y un número par, y como cuando el número es impar. ¿Puede esto tener importancia en la interpretación de la mediana?

11. ¿Cuántos son los cuartilos? ¿Qué significa Q_1^0?

12. ¿Es cierto que entre los dos centilos extremos está el 98% de las observaciones?

13. Haga el mismo razonamiento que en (11) para los decilos y percentilos

14. Hay un cuartilo, un decilo y un centilo que coinciden con la mediana? ¿Cuáles son?

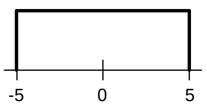
15. Repase el concepto de grados de libertad.

16. ¿Es siempre una varianza muestral un buen estimador de la varianza poblacional? Razone.

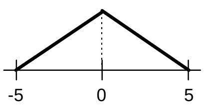
17. Explique las diferencias y semejanzas entre el promedio de desviaciones (o desviación promedio) y la desviación típica.

18. ¿Cuáles de las siguientes distribuciones tienen mayor varianza?

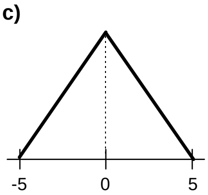
a)



b)



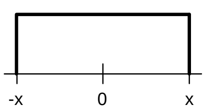
c)



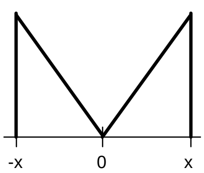
19. ¿Cuál de las siguientes distribuciones tiene idéntica la media y la mediana?

....a) Ninguna.
....b) Unicamente 1 y 3.
....c) Unicamente 3.
....d) Todas ellas.

1)



2)



3)