Unidad 4

Análisis de datos estadísticos

- Medidas de centralización, moda, media y mediana: ventajas y desventajas de cada una de ellas, cálculo, interpretación y uso.
- Medidas de posición, cuartiles y percentiles, interpretación.
- Medidas de variación, varianza, desvío, y coeficiente de variación: interpretación.

Roberto Fiadone
13/10/2024

Introducción

Los gráficos y las tablas de frecuencias nos dan una idea global del comportamiento de una variable, pero suele ser necesario resumir los datos calculando algunas medidas que caracterizan en forma única a una población o a una muestra. Estas medidas son valores que se interpretan fácilmente y nos sirven para un análisis más profundo que el obtenido por medio de los resúmenes -tablas y gráficos- con los que trabajamos en la unidad anterior.

Cuando las medidas se refieren a la población se las llama ***parámetros*.** Si en cambio se refieren a una muestra se las llama ***estadísticos*.** Y los estadísticos se calculan a partir de la muestra para poder así estimar, de manera aproximada, cual es el valor del parámetro de la población. Por ejemplo, supongamos querer tener una idea aproximada de cuál es el promedio de edad de los alumnos que asisten a una escuela. Entonces, los alumnos de la escuela son la *población* y ese promedio sería el *parámetro*. Pero como los alumnos son muchos, y queremos tener solo una idea aproximada de ese parámetro "*edad promedio de los alumnos de la escuela*", entonces tomaremos al azar a algunos pocos alumnos, les preguntaremos su edad, y calcularemos el promedio de ellos: ese conjunto de alumnos elegidos son la muestra, y el promedio que estoy calculando es mi *estadístico*. El valor que obtenga al calcular este estadístico de la muestra no tiene por qué coincidir con el de mi parámetro de la población, pero si elegí a varios alumnos al azar, es de esperar que obtenga un valor estimado muy aproximado del parámetro de la población.

Trabajaremos en esta unidad con algunas de las medidas o estadísticos más utilizados y analizaremos en qué casos se emplean cada uno de ellos.

Empezaremos por las llamadas ***medidas de centralización* o *de tendencia central*:**

Medidas de centralización o de tendencia central

Una medida de centralización es un estadístico cuyo valor intenta describir un conjunto de datos proporcionando un valor central o típico alrededor del cual los datos tienden a agruparse. Estas medidas son esenciales porque resumen un conjunto de datos con un solo valor representativo, facilitando la comprensión y comparación de los datos.

Se conocen numerosas medidas de centralización o de tendencia central. Las más comunes y que estudiaremos aquí son la moda, la mediana; y la media aritmética o promedio.

Moda

En estadística, la moda es una medida de tendencia central que se define como el valor que aparece con mayor frecuencia en un conjunto de datos.

Esta medida puede utilizarse tanto para variables **cualitativas como cuantitativas**.

Ejemplo 1

Supongamos que una empresa desea clasificar a sus empleados según el máximo nivel educativo alcanzado.

Con la información de la que dispone se elabora la tabla que se muestra.

Nivel Educativo	Frecuencia (f)
Primario	12
Secundario	25
Universitario	8
Total	45

La variable en estudio es la variable cualitativa $x = nivel\ educativo.$ Los valores que toma la variable son: *primario*, *secundario* y *universitario*.

En este ejemplo, el valor de la variable que se observa con mayor frecuencia es *secundario.* Por lo tanto, la moda es "nivel secundario" y lo escribimos así:

$$\boldsymbol{Moda = Mo = Nivel\ Secundario}$$

Por lo que podemos decir que el máximo nivel educativo alcanzado por la mayoría de los empleados es el nivel secundario.

Ejemplo 2

Supongamos ahora que se desea conocer la edad de los alumnos inscriptos en un curso de informática.

La variable en estudio es cuantitativa discreta "*edad de los estudiantes*". Recogidos los datos los resultados que se obtuvieron se resumen en la tabla

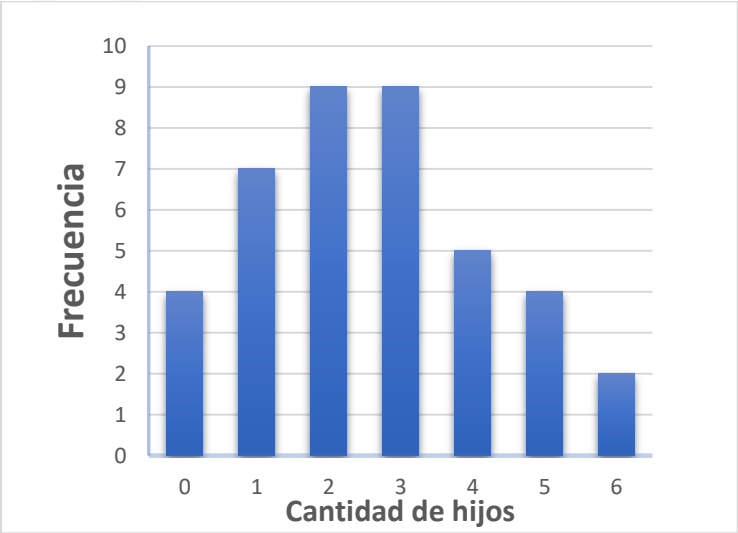
Edad	**Frecuencia (f)**
17	5
18	8
19	6
20	7
21	4
Total	30

Luego de observar la información podemos afirmar que la edad de los estudiantes que presenta mayor frecuencia es la de 18 años entonces:

$$\boldsymbol{Moda = Mo = 18\ años}$$

ATENCIÓN: Un error común es mirar la frecuencia más grande y decir que esa es la moda. Por ejemplo, en este caso, decir que Moda= 8: NO, la moda es siempre el valor de la variable, y **no** el valor de la frecuencia. ¡Ojo!

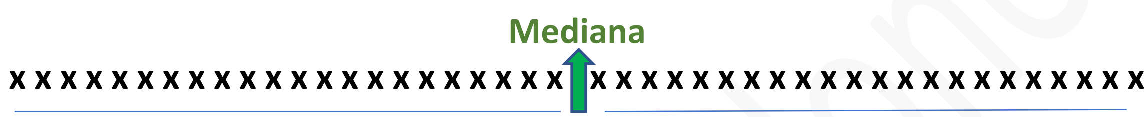
Observación: la moda es una medida muy fácil de calcular, pero tiene el inconveniente de no ser necesariamente única. Dentro de una misma distribución de frecuencias pueden aparecer dos o más categorías o valores de la variable a la que corresponde la máxima frecuencia. Por ejemplo, supongamos que en una encuesta de hogares en un pueblo se le preguntó a cada familia cuantos hijos tiene. El siguiente gráfico resume lo encuestado:



Vemos que hay dos valores de la variable *"cantidad de hijos"* que tienen la máxima frecuencia: para dos y tres hijos la frecuencia es 9. Es decir, hay dos valores para la Moda, el 2 y el 3. En este caso se dice que la distribución es **bimodal.**

<u>Mediana</u>

La **mediana** de un conjunto de observaciones **numéricas** es un valor del recorrido de la variable que deja por encima y por debajo al 50% de las observaciones que hayan sido **previamente <u>ordenadas</u> de menor a mayor.**



50 % de las observaciones <u>menores</u> | 50 % de las observaciones <u>mayores</u>

Veamos algunos ejemplos.

Ejemplo 3

Supongamos que la variable en estudio es la cantidad de materias aprobadas por alumno que concurren a una escuela. Se entrevistó a 11 estudiantes y las respuestas obtenidas fueron:

4 4 6 5 8 3 2 2 7 6 3

Si **ordenamos los datos del menor al mayor** obtenemos:

2 2 3 3 4 4 5 6 6 7 8

En este caso tenemos un número impar de datos, N= 11. La mediana será entonces el valor que quedó en el medio. O sea, el que quede en la posición sexta.

1°	2°	3°	4°	5°	**6°**	7°	8°	9°	10°	11°
2	2	3	3	4	**4**	5	6	6	7	8

Por debajo del **4** destacado quedaron cinco observaciones y por encima de ese **4** otras cinco.

Decimos entonces que la mediana nos dio 4 y lo escribimos así:

$$Me = \ 4 \ materias \ aprobadas$$

En este ejemplo fue fácil darse cuenta de que la mediana estaba en la posición sexta con solo visualizar la tabla. Pero si son muchos números, lo que conviene

es, luego de ordenarlos de menor a mayor, saber que la mediana se encuentra siempre en la posición que se obtiene de redondear hacia arriba $\frac{N}{2}$. Por ejemplo, en nuestro caso, como $\frac{11}{2} = 5{,}5$, el redondeo hacia arriba de 5,5 nos da 6, o sea que la mediana es, como efectivamente vimos, el valor que se sitúa en la posición sexta, el valor **4.** Este valor es nuestra medida central, nos da una idea inmediata de lo que pasó en el curso:

el 50% obtuvo nota menor o igual a 4, el otro 50% una nota mayor o igual a 4.

Ejemplo 4

Se registraron la cantidad de frutos que tenían cada uno de los 12 árboles frutales de una plantación y se obtuvieron los siguientes resultados para cada árbol:

4 4 9 5 8 3 2 2 7 9 3 9

Como antes, primero ordenamos las observaciones de menor a mayor:

2 2 3 3 4 4 5 7 8 9 9 9

Si queremos separar el 50% de los registros, la distribución nos queda dividida así:

1°	2°	3°	4°	5°	6°		7°	8°	9°	10°	11°	12°
2	2	3	3	4	4		5	7	8	9	9	9

A la izquierda de la línea quedan las seis observaciones más chicas y a la derecha las seis observaciones más grandes. Al ser N **par**, no podemos encontrar un único valor que quede justo en el medio. En este caso, para hallar la mediana, se conviene en tomar el promedio de los dos valores centrales, el 6° y el 7°, que resultan en este ejemplo ser el **4** y el **5**.

$$\boldsymbol{Me = \frac{4+5}{2} = 4{,}5 \text{ frutos}}$$

Si bien obviamente no hay un árbol que contenga 4,5 frutos, **la mediana está indicando que *la mitad de los árboles tiene menos de 4,5 frutos y la otra mitad más de 4,5.***

Al igual que para N impar, hay una manera rápida de darse cuenta en que posición se ubican los valores centrales: son los números que se encuentren en las posiciones $\frac{N}{2}$ y $(\frac{N}{2} + 1)$. Por ejemplo, en este caso, si calculamos $\frac{12}{2}$ y $\frac{12}{2} + 1$, obtenemos 6 y 7. En efecto, los valores centrales son el 6° y el 7°.

<u>Conclusión</u>:
Una vez **ordenadas de menor a mayor** las N observaciones de una variable cuantitativa:

- **Si N es impar**, la posición central es el valor que se toma como la mediana. Ese valor está ubicado en la posición que se obtiene de redondear hacia arriba $\frac{N}{2}$**.**
- **Si N es par**, se obtienen **dos posiciones centrales,** $\frac{N}{2}$ **y** $\frac{N}{2}+1$**,** y se toma como mediana el **promedio de los dos valores** que se encuentran en esas dos posiciones.

<u>Cálculo de la mediana cuando los datos están agrupados</u>.

Ejemplo 5:
Consideremos esta tabla que con la información sobre los ingresos de 110 trabajadores de una empresa. Queremos calcular la mediana de este conjunto de datos.

Ingresos	**Frecuencias**
1800	10
1900	23
2000	25
2100	31
2200	21
Total	**110**

Tenemos entonces **N=110** observaciones (N **par**). Si para calcular la mediana pretendiéramos proceder como en el ejemplo anterior, deberíamos anotarlos de menor a mayor uno al lado del otro…:

1800 1800 1800… así en total 10 veces…**1900 1900 1900**…23 veces…**2000 2000 2000**…25 veces…**2100 2100 2100**…31 veces…**2200 2200 2200**…21 veces.

Para luego buscar cuánto valen los dos valores centrales que estén situados en las posiciones $\frac{N}{2}=\frac{110}{2}=55$ y $\frac{N}{2}+1=56$ y promediarlos. Pero sería una tarea muy ardua y poco conveniente escribir los 110 datos. Veamos

cómo podemos proceder de una manera mucho más simple aprovechando, justamente, que los datos están agrupados.
Agreguemos a la tabla las frecuencias acumuladas F (recordemos que la frecuencia acumulada va sumando todas las frecuencias absolutas hasta el valor en consideración, es un subtotal formado por la suma de todas las frecuencias absolutas menores o iguales al valor considerado).

Ingresos	f_i	F_i
1800	10	10
1900	23	33
2000	25	58
2100	31	89
2200	21	110
Total	**110**	

Retornando a lo que queremos, buscamos ahora que observaciones se ubican en las posiciones 55° y 56°. Pues bien, notemos que desde la posición 34° hasta la 58° (ver la fila del $F = 58$) solo encontraremos ingresos iguales a 2000.

1°	2°	…	10°	11°	12°	…	33°	**34°**	**35°**	…	**58°**
1800	1800	…	1800	1900	1900	…	1900	**2000**	**2000**	…	**2000**

Notemos entonces que, en particular, tanto en la posición 55° como en la 56°, las observaciones valen 2000. Evidentemente el promedio de esos dos valores centrales nos va a dar 2000:

$$\boldsymbol{Me = \frac{2000 + 2000}{2} = 2000\ pesos}$$

Ejemplo 6

Supongamos el mismo enunciado del ejercicio anterior, información sobre los ingresos de 110 trabajadores de una empresa, pero que la distribución **hubiera tenido en cambio los siguientes valores de frecuencias f y F.**

Ingresos	f_i	F_i
1800	15	15
1900	13	28
2000	27	55
2100	34	89
2200	21	110
Total	**110**	

En principio procederíamos igual que antes, buscaríamos las posiciones:

$\frac{N}{2} = \frac{110}{2} = 55$ y la $\frac{N}{2} + 1 = 56$

Pero al mirar el cuadro, vemos que en la posición 55 se encontrará un dato con el valor 2000, mientras que en el 56 hallamos el 2100.

De manera que ahora el cálculo nos da:

$$\boldsymbol{Me} = \frac{2000 + 2100}{2} = \boldsymbol{2050\ pesos}$$

Media aritmética o promedio

A lo que en la vida diaria denominamos usualmente promedio, en estadística se lo llama "media", y es también una medida de la posición central, o sea, un estadístico que pretende darnos una idea de donde se localiza el centro de las observaciones, o sea, el dato más representativo posible, lejos del valor más bajo y a su vez lejos del más alto.

Entonces, si estamos considerando una variable x, la media (que se simboliza con el nombre de la variable y una raya arriba $\boldsymbol{\overline{x}}$) se calcula como la suma de las N observaciones x_i que se estén considerando de esa variable dividida por N. En símbolos:

$$\boldsymbol{\overline{x}} = \frac{\sum_{i=1}^{N} x_i}{\boldsymbol{N}}$$

La letra griega "sigma" mayúscula (Σ) se utiliza para significar que todo lo que sigue a ella debe ser sumado desde donde indican el subíndice i hasta donde indica el supra índice hasta N. También se estila, si no hay confusión, omitir eso índices, y simplemente escribir:

$$\bar{x} = \frac{\sum x_i}{\boldsymbol{N}}$$

Ejemplo 7

En el anterior ejemplo de la cantidad de frutos recogidos en cada árbol en una plantación con 12 árboles, teníamos los registros (ya ordenados):

4 4 9 5 8 3 2 2 7 9 3 9

Para calcular el número de frutos promedio o media aritmética de los frutos recogidos simplemente sumamos los datos y los dividimos por el total de observaciones:

$$media = \bar{x} = \frac{\sum x_i}{12} = \frac{4+4+9+5+8+3+2+2+79+3+9}{12} = \frac{65}{12} \approx \mathbf{5,42} \text{ frutos}$$

Observemos que no tiene sentido hablar de 5,42 frutos porque la cantidad de frutos siempre es un número natural, pero la idea es que recogimos una cifra que está más cercana a 5 que a 6.

Veamos cómo podemos calcular la media cuando los datos están agrupados.

Ejemplo 8

Supongamos que en un poblado de 15 familias se le preguntó a cada jefe de hogar cuantos hijos vivían en su casa y se obtuvo la siguiente tabla de frecuencias:

Hijos (x_i)	Familias (f_i)
1	3
2	4
3	5
5	2
7	1
N=	**15**

La variable x es entonces "*cantidad de hijos*". Vemos que hay tres familias con un hijo, cuatro con dos hijos, etc. No hay familias con cuatro o seis hijos, ni con más de siete hijos.

Si pretendiésemos calcular la media como en el ejemplo anterior, deberíamos tener en cuenta cuantas veces se repite cada cantidad:

1	1	1	2	2	2	2	3	3	3	3	3	5	5	7

Y recién luego hacer la cuenta:

$$\bar{x} = \frac{\sum x_i}{15} = \frac{1+1+1+2+2+2+2+3+3+3+3+3+5+5+7}{15}$$

$$\bar{x} = \frac{43}{15} = \mathbf{2,87 \text{ hijos}}$$

Pero esto es tedioso y largo. Aprovechemos que los datos están agrupados en la tabla para ver un procedimiento alternativo que suele ser mejor, sobre todo cuando el valor de N es muy grande.

Notemos en el numerador de la fórmula anterior que, tal como lo indica la segunda columna de la tabla, el 1 se repite tres veces, el 2 cuatro veces, etc.

Entonces, en vez de sumar de "a uno", hubiésemos podido escribir:

$$\bar{x} = \frac{1 \cdot 3 + 2 \cdot 4 + 3 \cdot 5 + 5 \cdot 2 + 7 \cdot 1}{15} = \frac{43}{15} \approx \mathbf{2,87\ hijos}$$

Observemos lo que hemos hecho: en el numerador se han sumado los valores de la variable multiplicado por su frecuencia. Por ejemplo, como con $x = 2$ hijos hay $f_i = 4$ hijos, entonces el segundo término es el "$2 \cdot 4$".

La forma de representar el procedimiento aplicado sería:

$$\boldsymbol{\overline{x} = \frac{\sum x_i \cdot f_i}{N}}$$

Donde $x_i$representa los distintos valores que toma la variable x y f_i su frecuencia de aparición:

Ejemplo 9

Aprovechemos el enunciado del ejercicio 2 en el que se deseaba conocer la edad de los alumnos inscriptos en un curso de informática y se había obtenido el cuadro:

Edad	**Frecuencia (f)**
17	5
18	8
19	6
20	7
21	4
Total	30

Calculemos el promedio en base a la fórmula $\bar{x} = \frac{\sum x_i \cdot f_i}{N}$. Para ello agregamos a la tabla una columna auxiliar donde calculamos los productos $x_i \cdot f_i$:

Edad x_i	f_i	$x_i \cdot f_i$
17	5	(17·5=) 85
18	8	144
19	6	114
20	7	140
21	4	84
Total	30	$\sum x_i \cdot f_i = 567$

O sea que

$$\overline{x} = \frac{\sum x_i \cdot f_i}{30} = \frac{567}{30} = \mathbf{18,9\ a\tilde{n}os}$$

---o---

Hasta aquí explicamos tres maneras distintas (de muchas otras que existen) para encontrar la medida central, es decir, como hemos dicho, un valor que nos parezca representativo de donde está nuestro centro de datos, pero ¿cuál es la más conveniente, que ventajas y desventajas tiene el uso de cada uno de estos estadísticos? Vamos a intentar dar respuesta a estos interrogantes.

Ejemplo 10

Consideremos las siguientes distribuciones de sueldos correspondientes a seis empleados de una oficina y, en cada caso, calculemos los estadísticos:

Distribución 1

1400; 1500; 1500; 1600; 1700; 1900

Obtenemos:

Mo= 1500; Me= 1550; $\overline{x}$=1600

Distribución 2

1400; 1500; 1500; 1600; 1700; **3100**

(*la misma que antes, salvo el último dato que es más grande*)

Obtenemos:

Mo= 1500; Me= 1550; $\overline{x}$=1800

Distribución 3

600; 1500; 1500; 1600; 1700; 1900

(*la misma que la distribución 1, salvo el primer dato que es más chico*)

Obtenemos:

Mo= 1500; Me= 1550; $\overline{x}$=1467

Observamos que la moda y la mediana son las mismas en las tres distribuciones, pero la media aritmética se modificó sustancialmente en las distribuciones 2 y 3, pese a haberse modificado **un solo dato**.

En la segunda distribución, el hecho de que la media de los sueldos sea relativamente alta, se debe a la presencia de un sueldo de $ 3100 que hizo que su cálculo fuese afectado por este valor que es mucho más grande que el resto. Algo similar sucede en la tercera distribución: al haber colocado un dato mucho más chico, la media se redujo significativamente.

Notemos que en los dos últimos casos la mediana es más representativa de los sueldos de la oficina, por ejemplo, en la distribución 2, está claro que lo más común es que un oficinista gane entre 1400 y 1700 pesos, sin embargo, el promedio nos dio un valor igual a 1800, pese a que **solo una** persona gana más que eso. La mediana dio un valor más razonable: decir que lo que ganan los empleados está centrado en unos 1550 pesos es bastante representativo, pues solo uno de los seis oficinistas gana *mucho más* que esa cifra.

Similar observación se puede hacer en la distribución 3, solo que ahora el promedio "se fue para el otro lado": nos da una cifra excesivamente baja como para considerarla el centro" de nuestras observaciones.

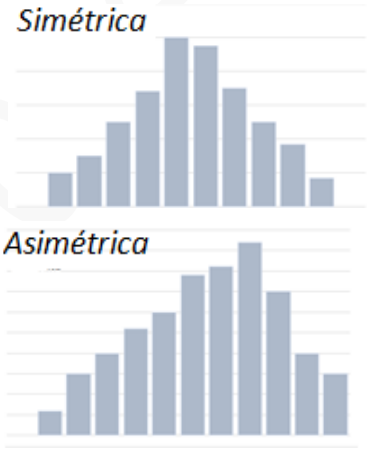
La **moda y la mediana no se ven fuertemente afectados por los valores extremos** que pudiera haber en un conjunto de datos numéricos.
En cambio, la **media aritmética es "atraída" por los valores extremos** que estén muy alejados del resto. Se acercará a ellos "olvidándose" de la mayoría restante.

Con este ejemplo que acabamos de ver, queremos dar a entender que la media aritmética o promedio, aun cuando sea la media de centralización más utilizada en la vida cotidiana, no siempre caracteriza bien a una distribución.
Hay muchas maneras en estadística de estimar el centro de los datos, acá hemos considerado solo la moda, la media y la mediana. En general, la elección de una medida adecuada depende del tipo de variable que se estudia y de la forma que adopte su distribución.
Vamos a ahondar en lo que se refiere a las ventajas y desventajas de cada uno de esos estadísticos.

Elección de una medida de centralización adecuada

Según el tipo de variable (cualitativa o cuantitativa):
Si la variable es de tipo cualitativa, evidentemente no queda otra que usar la moda, pues la media y la mediana están definidas solo para datos numéricos.
Si los datos son numéricos (cuantitativos) entonces, sin pretender profundizar en el tema, puede decirse que la moda no es lo más efectivo para representar aquello que entendemos por el centro de nuestros datos.

Según la forma de la distribución:
Si al realizar un histograma o representación con barras de frecuencia, notamos que la distribución de los datos es *simétrica*, entonces la moda, la mediana y la media aritmética coinciden o son bastante parecidas.
Si en cambio es *asimétrica*, y hay unos pocos valores muy grandes o pocos valores muy chico, o sea, los valores extremos están concentrados en una dirección de la distribución, entonces la media aritmética se ve afectada por estos valores y deja de ser representativa y es más conveniente utilizar la mediana.



	Media	Mediana	Moda
Ventajas	*Es la medida de tendencia central más empleada cotidianamente, de fácil cálculo. *Existen métodos que permiten medir que tan confiable es la media a partir de un conjunto de valores dados.	*No es sensible a los valores extremos. *Recomendable para distribuciones **no** simétricas.	*Es recomendable para el tratamiento de variables cualitativas. *Es muy fácil de visualizar en un gráfico de frecuencias. *Fácil de interpretar.
Desventajas	*Es muy sensible a los valores extremos ***No** recomendable en distribuciones asimétricas.	*Requiere ordenar los datos de menor a mayor. *No hay métodos sencillos para medir que tan fiable es el resultado obtenido como para representar a la medida central.	*Puede existir más de una moda. *No siempre se sitúa en el centro de la distribución. *En distribuciones muy asimétricas suele ser un dato muy poco representativo.

Medidas de posición o cuantiles

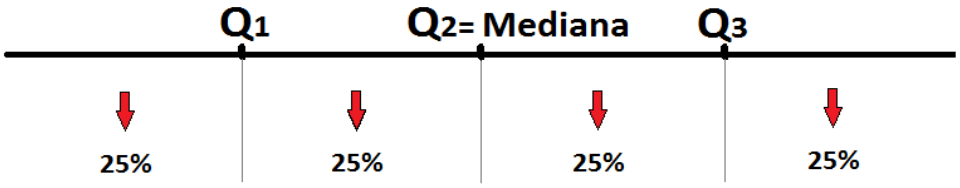
Las medidas de **posición** o **cuantiles** son valores que dividen la distribución en partes iguales; entendidas estas como intervalos que comprenden la misma proporción de observaciones. Los más usados son:

- Los **cuartiles**, que dividen a la distribución en *cuatro* partes.
- Los **quintiles**, que dividen a la distribución en *cinco* partes.
- Los **deciles**, que dividen a la distribución en *diez* partes.
- Los **percentiles**, que dividen a la distribución en *cien* partes.

Cuartiles

Los cuartiles de un conjunto de observaciones (previamente ordenadas de menor a mayor) de una variable cuantitativa son valores que dividen a dicho conjunto en cuatro subconjuntos que contienen la misma cantidad de datos.

Habrá, por tanto, **tres** cuartiles que dividirán al conjunto de N datos en intervalos que contendrán cada uno $\frac{N}{4}$ datos.



Ejemplo 11

Consideremos los siguientes datos **ya ordenados de menor a mayor** que corresponden al salario de 12 trabajadores:

1000 1400 1500 1550 1550 1550 1600 1700 1750 1800 1850 1900

Para hallar los cuartiles, dividimos la cantidad de datos en 4 partes iguales (los datos ya están ordenados de menor a mayor, caso contrario hay que ordenarlos previamente).

Como tenemos N=12 datos, cada parte debe tener entonces el 25% de los datos, o sea

$$\frac{1}{4} \cdot N = 0{,}25 \cdot N = 0{,}25 \cdot \ 12 = 3 \ datos.$$

1°	2°	3°		4°	5°	6°		7°	8°	9°		10°	11°	12°
1400	1400	1500		1550	1550	1550		1600	1700	1750		1800	1850	1900

El primer cuartil Q_1 lo obtenemos como promedio de la tercera y cuarta observación:

$$\boldsymbol{Q_1} = \frac{1500 + 1550}{2} = \mathbf{1525}$$

El segundo cuartil Q_2 (que va a dejar el 50% de valores a la izquierda y el 50% a la derecha, por lo tanto, notemos que Q_2 no es más que la mediana, pero que acá la llamamos cuartil segundo) sería:

$$\boldsymbol{Q_2} = \frac{1550 + 1600}{2} = \mathbf{1575}$$

Por último:

$$\boldsymbol{Q_3} = \frac{1750 + 1800}{2} = \mathbf{1775}$$

		Q_1=1525				Q_2=1575				Q_3=1775				
1400	1400	1500	|	1550	1550	1550	|	1600	1700	1750	|	1800	1850	1900

Advirtamos que la distancia entre los cuartiles **no tiene por qué** ser la misma, tal cual lo podemos observar en nuestro ejemplo. Así, entre el valor mínimo de 1400 y el Q_1 hay una distancia de:

$$\boldsymbol{Q1} \ - \ \mathbf{1400} = 1525 - 1400 = \ \mathbf{125}$$

Entre Q_1 y Q_2 hay

$$\boldsymbol{Q2} \ - \ \boldsymbol{Q1} = \ 1575 \ - \ 1525 = \ \mathbf{50},$$

Y entre Q_2 y Q_3 hay **200.**

Esto ocurrió porque la distancia o diferencia entre cada par de datos **no** es la misma. No están uniformemente separados. Lo importante es que **siempre queden entre cada par de cortes el 25% de las observaciones totales.** Para este caso eso implica que deben de quedar entre cada par de cuartiles:

$$0{,}25 \cdot N = 0{,}25 \cdot 12 = 3\ observaciones$$

No pretendemos en este curso explicar cómo se calculan los cuartiles en cada caso. Solo acabamos de ver un ejemplo para saber de qué se trata. Nos contentamos con que se entienda claramente su definición, propiedades y significado.

Importante: ¿Qué hubiese pasado si nos informaran qué al mes siguiente de tomados estos datos, a los 12 empleados de les abonó una suma extra de $200? ¿Cuánto darían ahora los valores de los cuartiles?

Pues los mismos valores del mes pasado **más** $200.

		Q_1=1725				Q_2=1775				Q_3=1975				
1600	1600	1700	|	1750	1750	1750	|	1800	1900	1950	|	2000	2050	2100

Es decir, si incrementamos (o decrementamos) un total "a" de unidades a TODOS los datos, los cuartiles se correrán la misma cantidad "a" de unidades (hacia la derecha, si incrementamos, hacia la izquierda, si decrementamos).

Otros cuantiles

Quintiles: El concepto de **quintil** es muy similar al de cuartil: ahora se trata de dividir el conjunto numérico (*previamente ordenado de menos a mayor*) en cinco partes, en vez de en cuatro. Habrá por tanto cuatro quintiles que dividirán al conjunto de N observaciones en 5 intervalos que contendrán cada uno $\frac{N}{5}$ datos, o sea, el 20% del total de observaciones.

Los quintiles de una distribución de frecuencias, de un conjunto de observaciones (previamente ordenadas de menor a mayor) de una variable cuantitativa son valores que dividen a dicho conjunto en 5 subconjuntos que contienen la misma cantidad de datos.

Genera algo de confusión, pero a los quintiles también se los simboliza con Q_i. Habrá entonces 4 quintiles, Q_1; Q_2; Q_3 y Q_4

Así, por ejemplo, el valor del segundo quintil divide al conjunto en un 40% de observaciones menores que su valor, y un 60% mayores a él.

Deciles: son 9 que dividen a la distribución en *diez* partes. Se los simboliza desde el D_1 al D_9

Percentiles: En este caso se trata de dividir el conjunto en 100 partes. Habrá por tanto 99 percentiles que dividirán al conjunto de N observaciones en intervalos que contendrán cada uno $\frac{N}{100}$ datos, o sea, cada uno conteniendo el 1% del total de observaciones.

A los percentiles lo denotamos P_k , donde el subíndice k indica el porcentaje de valores que son menores que él. Por ejemplo, P_{30} es el percentil que deja un 30% de observaciones menores que su valor (y por lo tanto, un 70% mayores a él).

Equivalencias

Y claro, notemos que entonces, por ejemplo:

$$cuartil\ Q_1 = P_{20}$$

$$Me = \ cuartil\ Q_2 = D_5 = P_{50}$$

$$quintil\ Q_2 = D_4 = P_{40}$$

Medidas de dispersión y variabilidad

Introducción

Las llamadas medidas de dispersión son valores estadísticos que nos permiten tener una idea de que tan alejados o separados están los valores observados de una variable cuantitativa respecto de su media. Esto permite tener una visión más acorde con la realidad en el momento de tomar las decisiones con respecto al valor del promedio.

Como ejemplo de lo pobre que puede ser contar como estadístico que describa a los datos de una muestra o población con tan solamente un valor promedio, supongamos que alguien que no sabe nadar quiere cruzar una laguna caminando por su fondo y sabe que el valor promedio de profundidad de la laguna es de 1,50 metros. Si él mide 1,70 metros de altura y no lo piensa

demasiado, creerá que puede cruzarla dado que asomarán 20 centímetros de su cabeza por sobre el nivel del agua. Pero ese promedio de 1,70 metros no me aclara si hay puntos en donde la profundidad es de apenas 50 centímetros contra otros en que es de más de 2 metros, o si en realidad la profundidad es muy uniforme, y no hay lugares mucho más profundos o menos profundos que ese valor de 1,50 metros. Es decir que el dato de la media aisladamente no aporta mucho. Este no es más que un ejemplo sencillo de las muy diversas situaciones en que es indispensable saber que tan desviados, que tan dispersos, pueden encontrarse los valores de la variable en cuestión respecto de su media.

Se hace entonces indispensable poder contar con una herramienta estadística para medir esa **lejanía de los valores respecto a la media**, y para ello se cuenta con distintos métodos, donde los más universales; y la mayoría de las veces más apropiados, son los estadísticos que se describen a continuación.

Varianza y desvío estándar (o *desviación típica*).

Para comprender que es lo que mide la ***varianza*** y el ***desvío estándar***, supongamos tener la siguiente situación. Consideremos tres grupos (A, B y C) formado por N=6 amigos, de distintas edades, como sigue:

Grupo	Edades					
A	13	14	15	15	16	17
B	12	13	13	16	17	19
C	11	12	13	15	18	21

Si nos tomamos el trabajo de calcular la media de cada uno de los grupos A, B y C, vamos a notar que en los tres grupos se obtiene el mismo valor de 15 años:

$$Media\ grupo\ A = \bar{X}_A = \frac{13+14+15+15+16+17}{6} = \frac{90}{6} = 15\ años$$

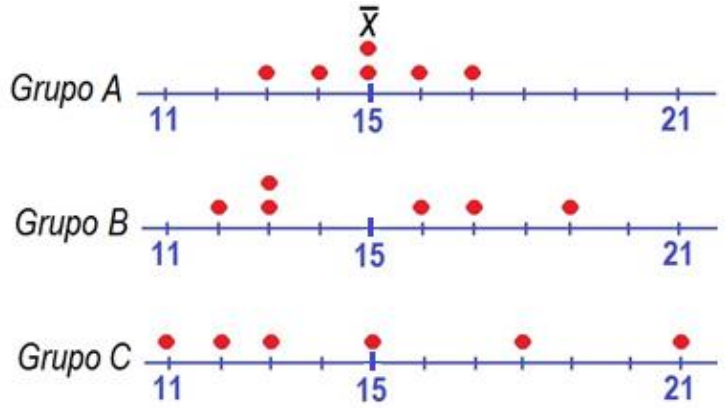
$$Media\ grupo\ B = \bar{X}_B = \frac{12+13+13+16+17+19}{6} = \frac{90}{6} = 15\ años$$

$$Media\ grupo\ C = \bar{X}_C = \frac{11+12+13+15+18+21}{6} = \frac{90}{6} = 15\ años$$

Sin embargo, es evidente que la distribución de las edades es muy diferente, es decir, si observamos el primer grupo A, vemos que los chicos de ese grupo

tienen edades parecidas, más cercanas, más homogéneas. En el grupo B, las edades son algo más diversas, más dispersas, y en el C lo son aún más, es un grupo muy heterogéneo con edades de 11 a 21 años.

Consideremos un gráfico para cada grupo en el que representemos la edad de cada individuo mediante un punto sobre una recta:



Con la finalidad de cuantificar y tener una idea más exacta de que tan cercanamente concentrados están los datos alrededor de la media (sobre todo en casos en que estuviésemos trabajando con muchos más datos o grupos que en este ejemplo) vamos a definir un par de estadísticos, llamados varianza y desviación típica.

Supongamos tener "N" valores correspondientes a un grupo o conjunto:

$$x_1, \quad x_2, \quad x_3, \quad \dots \quad x_N$$

Por ejemplo, para nuestro primer grupo A de amigos, tenemos N= 6. Además:

$$x_1 = 13, \quad x_2 = 14, \quad x_3 = 15, \quad x_4 = 15, \quad x_5 = 16, \quad x_6 = 17$$

Entonces, se define como "varianza" a un número, que simbolizaremos con el símbolo de la letra griega minúscula llamada "sigma" elevada al cuadrado (σ^2) que se obtiene a partir de la siguiente fórmula:

$$\boldsymbol{Varianza} = \boldsymbol{\sigma^2} = \frac{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}{\boldsymbol{N}}$$

Y definimos como "desviación típica" o "desvío estándar" al valor que se calcula como la raíz cuadrada de la varianza y que simbolizaremos con la letra sigma:

$$\textbf{\textit{Desvío estándar}} = \boldsymbol{\sigma} = \sqrt{\sigma^2} = \sqrt{\frac{\sum(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^2}{\boldsymbol{N}}}$$

Más adelante explicaremos cómo se calcula esta fórmula, por el momento concentrémonos en que es lo que pretende medir. Para ello volvamos al ejemplo de los tres grupos de amigos. Si se calcula el valor tanto de la varianza σ^2 como del desvío estándar σ se obtienen los siguientes valores:

	Grupo		
	A	**B**	**C**
$\boldsymbol{\sigma^2}$	1,67	6,33	12,32
$\boldsymbol{\sigma}$	1,29	2,52	3,51

Como vemos, el valor tanto del desvío como de la varianza es menor en el grupo A (en que las edades eran más parecidas a la edad media), algo más grande en el B (donde las edades están más alejadas de la media) y mayor aun en el grupo C (donde la dispersión o distancia de los valores observados están más alejados de la edad media del grupo).

Esta es la idea de estos estadísticos, que su valor sea una **referencia de cuan alejados están las observaciones respecto de la media**. Cuanto más chico es el valor del desvío o de la varianza, más "**homogéneo**" se dice que es el conjunto de valores observados, porque justamente, cuanto más chico el desvío, más parecidos son los valores observados entre sí.

Se hace evidente que esto es así al analizar la fórmula: notemos por ejemplo en que consiste calcular la varianza del *grupo A*: para hallar el valor, debemos calcular primero **la diferencia entre cada valor x_i y la media $\overline{x}$**. A este valor $(x_i - \bar{x})$ se lo denomina "desvío". Y luego elevemos al cuadrado a cada uno de esos desvíos, o sea calcularemos los $(x_i - \bar{x})^2$:

$(x_1 - \bar{x})^2 = (13 - 15)^2 = (-2)^2 = 4$

$(x_2 - \bar{x})^2 = (14 - 15)^2 = (-1)^2 = 1$

$(x_3 - \bar{x})^2 = (15 - 15)^2 = 0^2 = 0$

$(x_4 - \bar{x})^2 = (15 - 15)^2 = 0^2 = 0$

$(x_5 - \bar{x})^2 = (16 - 15)^2 = 1^2 = 1$

$(x_6 - \bar{x})^2 = (17 - 15)^2 = \quad 2^2 \quad = \quad 4$

Notemos, por ejemplo, que el valor obtenido al calcular $(x_1 - \bar{x})^2$ para el chico de 13 años resultó igual al de 17 años $(x_6 - \bar{x})^2$, o sea, ambos igual a 4. Esto es porque el chico de 13 años está dos unidades alejado de la media de 15 y lo mismo le sucede al de 17 años.

En cambio, el valor obtenido fue de solo 1 para los chicos de 14 y 16 años, y para los que tenían una edad de 15, que coincide con el valor de la media $\bar{x}$, el valor obtenido fue de 0. Es decir, cuanto más alejado están los valores del promedio, sea por superar ese valor o por estar por debajo de él, mayor va a ser el valor del desvío elevado al cuadrado $(x_i - \bar{x})^2$.

A la inversa, cuanto más cercanos a la media estén, más cercanos a cero nos van a dar. Por lo tanto $(x_i - \bar{x})^2$ nos dice que tan desviados están los valores con respecto a la media. Hecho esto, para calcular la denominada varianza, solo falta **sumar todos los desvíos $(x_i - \bar{x})^2$ y dividirlos por N=6:**

$$\boldsymbol{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{\sum(x_i - 15)^2}{6} =$$

$$\frac{4 + 1 + 0 + 0 + 1 + 4}{6} = \frac{10}{6} \cong \mathbf{1,67\ a\tilde{n}os^2}$$

(Aunque suene extraño, a las unidades de la varianza se las escribe como las unidades de la variable, pero como si estuvieran elevadas al cuadrado, por eso dice **años**[2]).

Por lo tanto, $\boldsymbol{\sigma} = \sqrt{\sigma^2} = \sqrt{1{,}67} \cong \mathbf{1,29}\ \boldsymbol{a\tilde{n}os}$

Lo mismo podemos hacer con el *grupo **B***:

$(x_1 - \bar{x})^2 = (12 - 15)^2 = (-3)^2 = 9$

$(x_2 - \bar{x})^2 = (13 - 15)^2 = (-2)^2 = 4$

$(x_3 - \bar{x})^2 = (13 - 15)^2 = (-2)^2 = 4$

$(x_4 - \bar{x})^2 = (16 - 15)^2 = \quad 1^2 \quad = 1$

$(x_5 - \bar{x})^2 = (17 - 15)^2 = \quad 2^2 \quad = 4$

$(x_6 - \bar{x})^2 = (19 - 15)^2 = \quad 4^2 \quad = 16$

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N} = \frac{\sum(x_i - 15)^2}{6} =$$

$$\frac{9+4+4+1+4+16}{6} = \frac{38}{6} \cong \mathbf{6,33\ años^2}$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{6{,}33} \cong \mathbf{2,52\ años}$$

Para el *grupo* ***C*** la varianza σ^2 da **12,33** y el desvío σ **3,51** (comprobarlo).

En el grupo A, como hay **menos separación**, el desvío resultó menor.

Imaginen que se hubiese tratado de 6 amigos de la misma edad, por ejemplo, **todos** de 15 años. ¿Cuánto vale la media? Obviamente va a dar $\bar{x} = 15$. ¿Y que tan separados están los valores de las edades respecto de la media? Obviamente, dado que la media es 15, y todos tienen 15 años, no hay desvío. En efecto, todos los $(x_i - \bar{x})^2$ van a dar $(15-15)^2 = 0$

¡La varianza y el desvío dan cero! Como era de esperar, pues no se apartan para nada de la media.

ATENCIÓN: diferencias entre $\boldsymbol{\sigma^2}$ **y** $\boldsymbol{S^2}$

Hasta aquí utilizamos el símbolo σ^2 para referirnos a la varianza y su fórmula:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

En estadística, por razones rigurosamente fundamentadas pero que escapan al nivel de este curso, las fórmulas y símbolos a utilizar en el cálculo de la varianza y del desvío no siempre son los que acabamos de ver. En realidad, hay dos fórmulas muy parecidas pero distintas cuyo uso depende de si los datos a analizar se han extraído de una *muestra* de la población, o si los datos que estamos analizando son *TODOS* los datos de esa población.

Si los N datos que se analizan son de la totalidad de la población, o sea, no quedó ninguno afuera (como cuando se hace un censo) entonces la fórmula que se utiliza es la que vimos y se simboliza con σ^2.

Varianza para datos de una población: $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$

Y el **desvío estándar** es: $\sigma = \sqrt{\sigma^2}$

En cambio, si no tenemos todas las observaciones, sino que los N datos son solo una muestra, un subconjunto de la población, entonces la varianza se representa con una S mayúscula al cuadrado, y la fórmula es igual a la anterior, pero dividiendo por N-1

Varianza para datos de una muestra: $S^2 = \frac{\sum(x_i - \bar{x})^2}{N-1}$

Y el **desvío estándar** es: $S = \sqrt{S^2}$

Esto parece caprichoso ¿por qué usar una fórmula distinta, y porqué dividir por N-1 si son N datos? Como se dijo, hay motivos bien fundados, pero **a los fines prácticos de esta materia, y dado que para un N "grande" ambas fórmulas dan prácticamente el mismo resultado,**

En este curso nos manejaremos siempre con $\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$ sin importar de si se trata de una población o de una muestra.

No obstante, importa saber esto, pues si buscan ejercicios o bibliografía afín a este tema, ya saben ahora por qué a veces se utiliza un σ^2 y otras veces un S^2 al referirse a la varianza y porque la fórmula difiere según el caso.

Intervalos de dispersión

El desvío sirve para algo aún más interesante que nos permitirá tener una idea más concreta aun de que tan dispersos están los valores. Vamos a definir unos intervalos numéricos que, como veremos luego, son muy útiles para poder saber mucho más en cuanto a cómo están distribuidos los datos con respecto a la media.

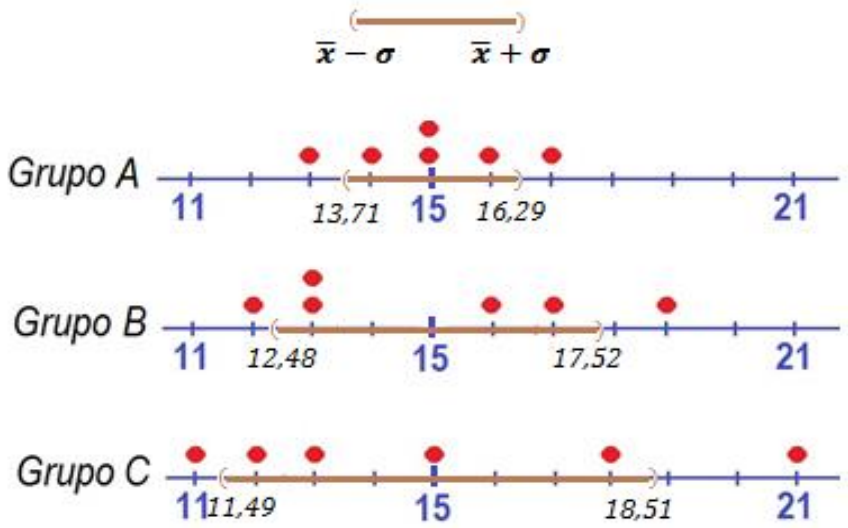
Primero hagamos lo siguiente: en el ejemplo dado, para cada uno de los grupos calculamos estos dos valores:

$$\bar{x} - \sigma \quad y \quad \bar{x} + \sigma$$

y construimos luego unos intervalos que los tengan por extremo, de esta manera:

Grupo	$\bar{x}$	σ	$\bar{x} - \sigma$	$\bar{x} + \sigma$	$(\bar{x} - \sigma\ ;\ \bar{x} + \sigma)$
A	15	1,29	(15-1,29=) 13,71	(15+1,29=) 16,29	**(13,71; 16,29)**
B	15	2,52	12,48	17,52	**(12,48; 17,52)**
C	15	3,51	11,49	18,51	**(11,49; 18,51)**

Para interpretar los que significan los intervalos de la última columna, marquemos esos valores en la línea en que representamos los datos de cada grupo:



Lo que vemos es lo que suele ocurrir cuando se calcula ese intervalo numérico **$(\bar{x} - \sigma \; ; \; \bar{x} + \sigma)$:** en cada uno de los grupos quedó encerrado dentro del intervalo las dos terceras partes de los datos observados (en este caso, 4 datos de un total de 6). En general, los intervalos construidos como los del ejemplo, tienen la propiedad (que no vamos a demostrar aquí) por la cual, bajo ciertas condiciones, **se puede asegurar que el 68% de los de las observaciones numéricas van a encontrarse en ellos**.

Dicho de otro modo: de los N datos numéricos que uno analice, y siempre y cuando se cumplan ciertos requisitos que mencionaremos luego, puede asegurarse que aproximadamente el 68% de los valores de la variable "x" cumplirán que:

$$\bar{x} - \sigma < x < \bar{x} + \sigma.$$

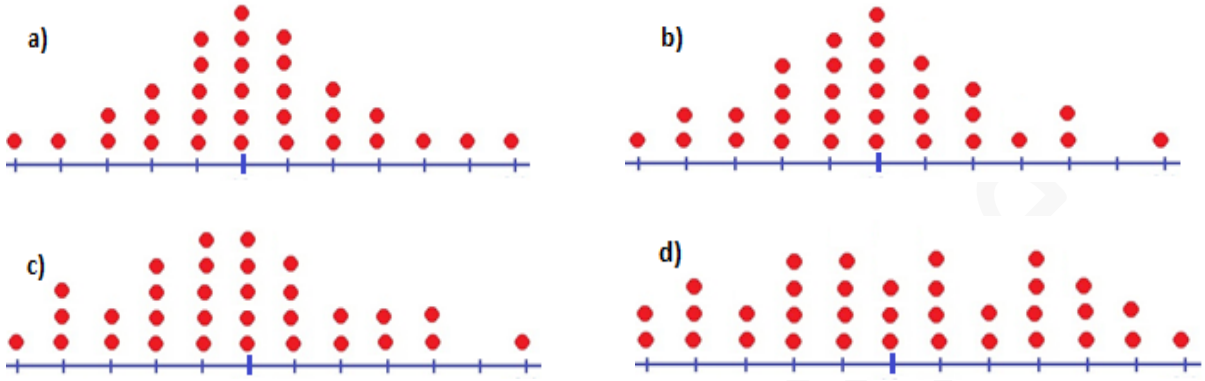
Es más, puede demostrarse que si en cambio hubiésemos calculado intervalos cuyos extremos fueran $\bar{x} - 2 \cdot \sigma$ y $\bar{x} + 2 \cdot \sigma$, entonces aproximadamente el 95% de los datos se encontrarían en el intervalo

$$(\bar{x} - 2 \cdot \sigma \; ; \; \bar{x} + 2 \cdot \sigma)$$

Y también que el 99% se hallaría en $(\bar{x} - 3 \cdot \sigma \; ; \; \bar{x} + 3 \cdot \sigma)$.

Pero habíamos dicho que había que pedir algo para que efectivamente esos intervalos cumplan con lo mencionado.

Si bien no es propósito de este curso explicar con detalle cuando se cumplen esos requisitos, digamos que lo principal es pedir que los datos se distribuyan de manera simétrica con respecto a la media, y que se concentren en su mayoría cerca de ella. Por ejemplo, analicemos estos gráficos de frecuencias:



a) los datos se distribuyen bien simétricamente y con más observaciones hacia el centro que en los extremos. Media y mediana darán casi lo mismo.
b) Los datos se distribuyen algo más asimétricamente.
c) No están muy concentrados en el centro y hay algo de asimetría.
d) No está muy claro donde se concentran, y hay poca simetría.

En los casos a) y b) seguramente los intervalos ($\overline{x} - \sigma$; $\overline{x} + \sigma$) contendrán casi el 68% de los datos.

En cambio, muy probablemente esto no ocurra para los casos c) y d).

Ejemplo 12

En dos cursos de alumnos, A y B, se ha tomado el mismo examen, calificándoles por su nota de 1 a 100. Los datos estadísticos de las notas fueron los siguientes:

x = nota del alumno	$\overline{x}$	σ
A	62	12
B	82	3

Un docente encuentra un examen que se había perdido, pero no saben de qué curso es. Si la nota que se sacó el alumno es de 73 ¿A qué curso más probablemente pertenecerá?

La nota 73 está a 11 puntos de la media del grupo A (62) mientras que está a 9 puntos del promedio de 82. Si nos quedamos solo con esta observación, estaríamos tentados a pensar que la nota de ese examen corresponde al curso B. Pero miremos el desvío: el del curso B es mucho menor que el del curso A. Esto me está indicando que las notas del curso B son mucho menos dispersas y, por lo tanto, todos sacaron más o menos la misma nota, cercana a 82. En cambio, aunque el promedio del curso A fue de 62, sus notas fueron mucho más diferentes.

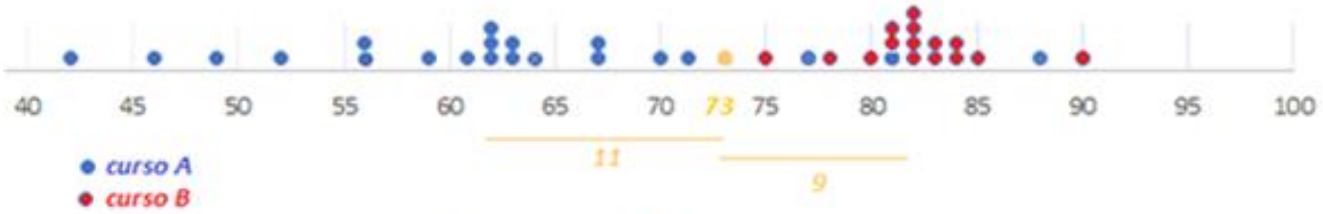
Si calculásemos los intervalos para cada curso, obtenemos:

	$\boldsymbol{Curso\ A}$	$\boldsymbol{Curso\ B}$
$\boldsymbol{(\overline{x} - \sigma\ ;\ \overline{x} + \sigma)}$	(50; 74)	(79; 85)
$\boldsymbol{\overline{x}}$	62	82

A partir de lo estudiado, vemos que:

Más de las dos terceras partes (68%) de las notas del curso A fueron de entre 50 y 74 puntos.
Más de las dos terceras partes (68%) de las notas del curso B fueron de entre 79 y 85 puntos.
La nota 73 que sacó el alumno está dentro del intervalo del curso A, pero no del B. Lo que indica que es más probable que se trate de un alumno del curso A. En otras palabras: dado que las notas del curso B no se alejan mucho de la nota promedio 82, sería raro que el alumno sea de ese curso. En cambio, las notas del grupo A son tan variadas que, pese a que el promedio dio 62, hay alumnos que sacaron notas bastante altas.

Así que concluimos que lo más probable es que el alumno pertenezca al curso **A**.



Las notas del curso B se "amontonan" mientras que las del curso A están muy dispersas. Por eso, pese a que la media del grupo B es más cercana a la nota 73, es más lógico pensar que esta nota es del curso A.