ESTADÍSTICA

UNIDAD 4: Resumen de datos en forma numérica

Ciencias Económicas y Empresariales

UCEL

Estadística descriptiva

- Descripción de un conjunto de observaciones de una variable cualitativa
- Descripción de un conjunto de observaciones de una variable cuantitativa
- Medidas descriptivas: de tendencia central y de dispersión.

Medidas descriptivas o resumen

Son medidas resúmenes que permiten expresar las características más notables de un conjunto de datos. Si el conjunto en estudio es la muestra, dichas medidas reciben el nombre de **estadísticos** y si se trata de la población, dichas medidas se conocen como **parámetros**.

1) Posición o ubicación	2) Dispersión
Media aritmética	Rango
Mediana	Rango intercuartílico
Moda	Desvío cuartílico
Cuartiles	Variancia
	Desvío estándar
	Coeficiente de variación

Medidas de posición

Ejemplo: Pensemos que se cuenta con información del número de hermanos de 15 personas encuestadas en la EPH del año 2021. Las observaciones resultan ser:

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

Resulta de interés resumir dicha información, por lo tanto se van a presentar y calcular las medidas descriptivas que se enunciaron anteriormente.

1) Medidas posición o tendencia central: Permiten ubicar el "centro" de un conjunto de datos.

Media aritmética:

La media aritmética de un conjunto de observaciones se obtiene sumando las observaciones y dividiendo por el total de observaciones. Es necesario distinguir entre la media de la población y la media de la muestra:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Media poblacional: Parámetro

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Media muestral: Estadística

Donde $x_1, x_2, \dots, x_n$ representan los valores observados de la variable x en la muestra.

Media aritmética

Cálculo de la media aritmética para el conjunto de datos del ejemplo:

$$\bar{x} = \frac{2 + 1+ 3+ 1+ 1+ 4+ 1+ 1+ 3+ 4+ 1+ 5+ 4+ 1+ 3}{15} = 2{,}33\ hermanos$$

Interpretación: En promedio el número de hermanos en la muestra de 15 familias, es de 2,33 hermanos

Media aritmética

Supongamos que la última observación fue erróneamente registrada y en lugar de 3 se registró 30. ¿Cuál es el promedio ahora?

$$\bar{x} = \frac{2 + 1 + 3 + 1 + 1 + 4 + 1 + 1 + 3 + 4 + 1 + 5 + 4 + 1 + 30}{15} = 4{,}13\ hermanos$$

Cuidado! La media es sensible a observaciones extremas

Características de la media aritmética

- Es un valor único para cada conjunto de datos.
- En su cálculo intervienen todas las observaciones, aprovechando al máximo la información.
- Se ve afectada por valores extremos. No es adecuada para conjuntos de datos con valores inusualmente bajos o altos o con distribuciones asimétricas.

MANOS A LA OBRA N° 1

ACTIVIDAD 1: Encuentre la cantidad promedio de autos vendidos en 9 concesionarias de la ciudad de Rosario en el mes de mayo de 2021.
4 – 5 – 5 – 6 – 6 – 6 – 7 – 7 – 9

ACTIVIDAD 2: Encuentre el ingreso de ventas promedio, en millones de pesos, en 14 concesionarias de la ciudad de Rosario en el mes de junio de 2021.
50,61 – 60,85 – 63,70 – 70,98 – 88,90 – 44,07 – 56,87 – 48,84 – 64,41 – 40,85 – 65,04 – 61,38 – 45,39 – 82,80

ACTIVIDAD 3: La nota promedio de 3 estudiantes es 54 y la nota promedio de otros 4 estudiantes es 76. ¿Cuál es la nota promedio de los 7 estudiantes?

ACTIVIDAD 4: Los resultados de los parciales de Gastón son: 98, 25, 19 y 26. Calcule la nota promedio. Explique por qué la media no hace un buen trabajo para resumir las notas de Gastón.

Mediana (Mna):

Es el valor central de las observaciones una vez que las mismas se han ordenado. Es el valor de la variable que divide al conjunto en 2 partes con igual número de elementos, tal que el 50% de los datos son menores o iguales a él y el 50% restante mayor o igual.

- Si el nº de observaciones es impar, la mediana es el valor central. Ejemplo: 2 – 3 – 4 – 6 – 7. Mna=4.
- Si el nº de observaciones es par, la mediana es el promedio de los dos valores centrales. Ejemplo: 2 – 3 – 4 – 6 – 7 – 9. Mna=(4+6)/2=5.
- Teniendo los datos ordenados, la mediana se encuentra en la posición $Mna^0 = (n+1)/2$.

La mediana es una medida de centro **resistente** o **robusta**

Cálculo de la mediana para el conjunto de datos del ejemplo

En primer lugar se ordenan las observaciones (en este caso de menor a mayor):

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

Observamos que como el número de observaciones es impar la mediana va a ser el valor central. En este caso **Mna=2 hermanos**.

También la podemos obtener calculando la posición de la mediana: $Mna^0 = (n+1)/2.= (15+1)/2=8°$. Cuyo valor indica que la mediana se encuentra en la octava posición. Por lo tanto **Mna=2 hermanos.**

Cálculo de la mediana para el conjunto de datos del ejemplo

En primer lugar se ordenan las observaciones (en este caso de menor a mayor):

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

Mna= 2 hermanos

Interpretación: El 50% de los encuestados en la EPH tiene entre 1 y 2 hermanos, mientras que el 50% restante tiene entre 2 y 5 hermanos.

MANOS A LA OBRA N° 2

☑ **PARA RESOLVER!!!**

4.3.

Encuentre la mediana del número de hijos correspondiente a una muestra de 10 familias.

Nº de observación:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Nº de chicos:	**2**	**3**	**0**	**1**	**4**	**0**	**3**	**0**	**1**	**2**

a) Ordene las observaciones de menor a mayor.
b) Localice el lugar que ocupa la mediana.

c) Mediana = --------------

d) ¿Qué le pasa a la mediana si la observación # 5 en el primer listado fue incorrectamente listada como 40 en lugar de 4?

e) ¿Qué ocurre si la observación # 3 del primer listado fue incorrectamente registrada como -20 en lugar de 0?

Características de la mediana:

- Es un valor único para cada conjunto de datos.
- En su cálculo no intervienen todos los valores de la variable.
- No se ve afectada por valores extremos, por eso resulta ser una medida de posición útil en esos casos o para distribuciones asimétricas.

<u>Modo</u> (Mo):

También conocida como Moda. Es el valor de la variable que ocurre con mayor frecuencia. Puede haber más de una moda.

Esta medida es la única que puede ser utilizada para resumir variables cualitativas. En ese caso, la moda es la categoría de una variable cualitativa que presenta la mayor frecuencia.

Ventajas: A semejanza de la mediana, al modo tampoco le afectan los valores extremos.

Desventajas: A pesar de sus ventajas el modo se usa menos que la media y la mediana, ya que, puede no existir en el caso en que ninguna observación se repita. Otras veces todos los valores son modo ya que ocurren el mismo número de veces.

Cálculo del modo para el conjunto de datos del ejemplo

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

Mo=1 hermano

Interpretación: El número más frecuente de hermanos entre los encuestados de la EPH es de un hermano.

MANOS A LA OBRA N° 3

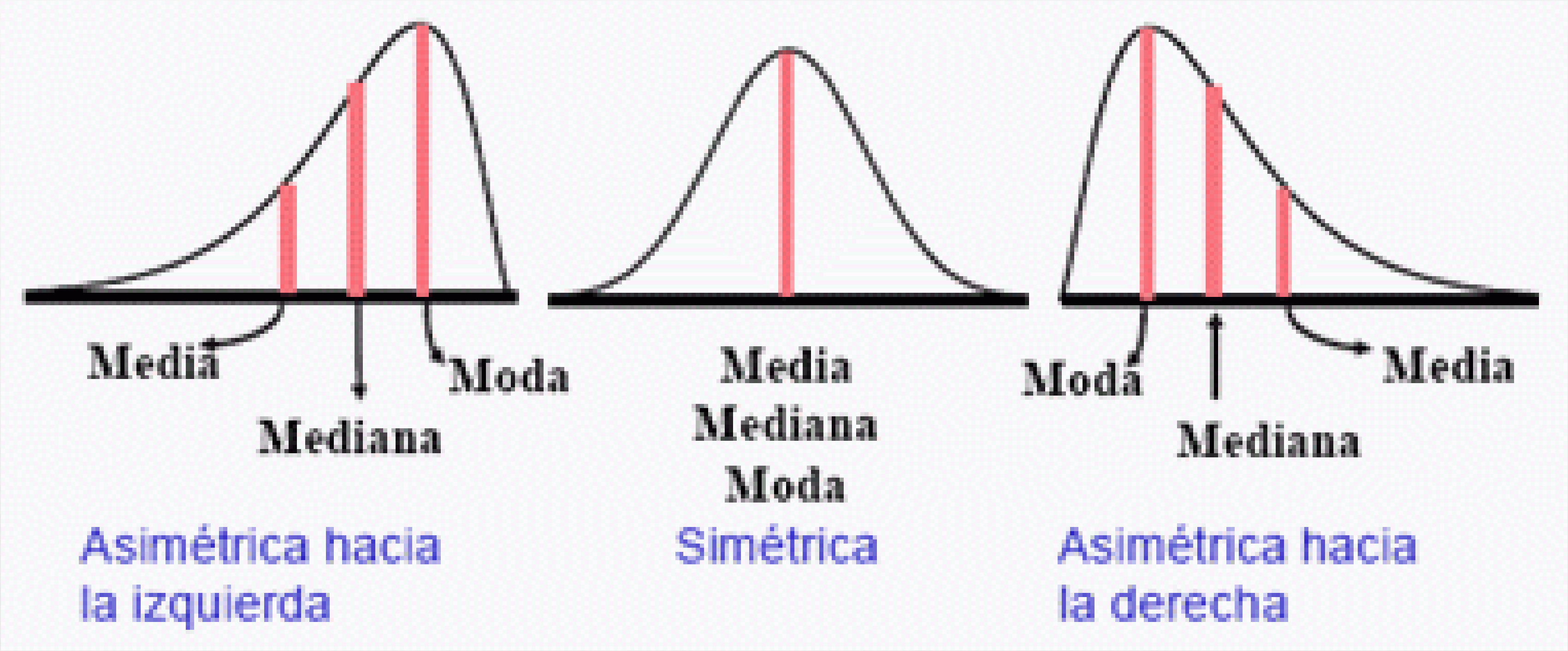
EJEMPLO 4.3 (página 231). Diferentes medidas pueden dar diferentes impresiones

Consideremos los ingresos anuales de 5 familias de una misma zona:

$12.000 - $12.000 - $30.000 - $90.000 - $100.000

¿Cuál es el ingreso típico o característico de esta zona? Calcule la media, la mediana y la moda analice cuál de estas medidas es más apropiada en este caso. Explique por qué.

¿Qué medida de centro usar?



MANOS A LA OBRA N° 4

PARA RESOLVER 4.5 (página 234).

Considere un estudio para comparar dos tipos de antibióticos para el tratamiento de anginas en niños en edad escolar: Amoxicilina y Cefadroxil. En un centro sanitario, 23 chicos fueron asignados aleatoriamente para uno de los dos tratamientos. Se pensó que la edad de los chicos podría influir sobre la efectividad de los antibióticos. Las edades de los chicos para cada tratamiento se presentan a continuación.

T1: Amoxicilina (n=11)
Edad: 14 17 11 10 11 14 9 12 8 10 9
Media:
Mediana:
Moda:

T2: Cefadroxil (n=12)
Edad: 9 14 8 10 13 7 9 11 16 10 12 9
Media:
Mediana:
Moda:

Calcule la media, mediana y moda para cada uno de los grupos.
Compare los grupos con respecto a la edad.

Cuartiles (Q_i)

Son tres medidas de posición no central, se dice que son medidas de orden. Los cuartiles dividen a la distribución en cuatro partes iguales. Su cálculo requiere, al igual que la mediana, que los datos estén ordenados.

- **Cuartil 1 (Q_1):** es el valor de la variable que acumula el 25% de los datos. Por lo tanto, el 25% de los datos son menores o iguales a dicho valor y el 75% son mayores o iguales a el. Se encuentra en la posición $Q_1^0=(n+1)/4$.
- **Cuartil 2 (Q_2):** coincide con la mediana.
- **Cuartil 3 (Q_3):** es el valor de la variable que acumula el 75% de los datos. Por lo tanto, el 75% de los datos son menores o iguales a dicho valor y el 25% son mayores o iguales a el. Se encuentra en la posición $Q_3^0 =3(n+1)/4$.

Cálculo de los cuartiles para el conjunto de datos del ejemplo

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

Una vez ordenado los datos en forma ascendente se calcula la posición del cuartil 1 y del cuartil 3.

$Q_1{}^0 = \frac{n+1}{4} = \frac{15+1}{4} = 4°$ □ Q_1= 1 hermano

Interpretación Q_1: El 25% de los encuestados de la EPH tienen a lo sumo 1 hermano, mientras que el 75% restante tiene como mínimo un hermano.

$Q_3{}^0 = \frac{3(n+1)}{4} = 3\,\frac{15+1}{4} = 12°$ □ Q_3= 4 hermanos

Interpretación Q_3: El 75% de los encuestados de la EPH tienen a lo sumo 4 hermanos, mientras que el 25% restante tiene como mínimo 4 hermanos.

MANOS A LA OBRA N° 5

ACTIVIDAD 1: Encuentre los cuartiles de las edades de cada uno de los dos grupos de niños que participaron del estudio para comparar dos tratamientos para la angina. Interprete los valores en términos del problema.

Amoxicilina: 14 – 17 – 11 – 10 – 11 – 14 – 9 – 12 – 8 – 10 – 9
Cefadroxil: 9 – 14 – 8 – 10 – 13 – 7 – 9 – 11 – 16 – 10 – 12 – 9

ACTIVIDAD 2: Encuentre los cuartiles del ingreso de ventas de autos, en millones de pesos, en 14 concesionarias de la ciudad de Rosario en el mes de junio de 2021.
50,61 – 60,85 – 63,70 – 70,98 – 88,90 – 44,07 – 56,87 – 48,84 – 64,41 – 40,85 – 65,04 – 61,38 – 45,39 – 82,80

ACTIVIDAD 3: Encuentre los cuartiles de la cantidad de autos vendidos en 9 concesionarias de la ciudad de Rosario en el mes de mayo de 2021.
4 – 5 – 5 – 6 – 6 – 6 – 7 – 7 – 9

Medidas de dispersión

Si sólo se toma en cuenta las medidas de posición de un conjunto de datos o si se comparan varios conjuntos de datos utilizando valores centrales, **se llegará a una conclusión incorrecta.**

Además de las medidas de posición, se debe tomar en consideración la **dispersión.**

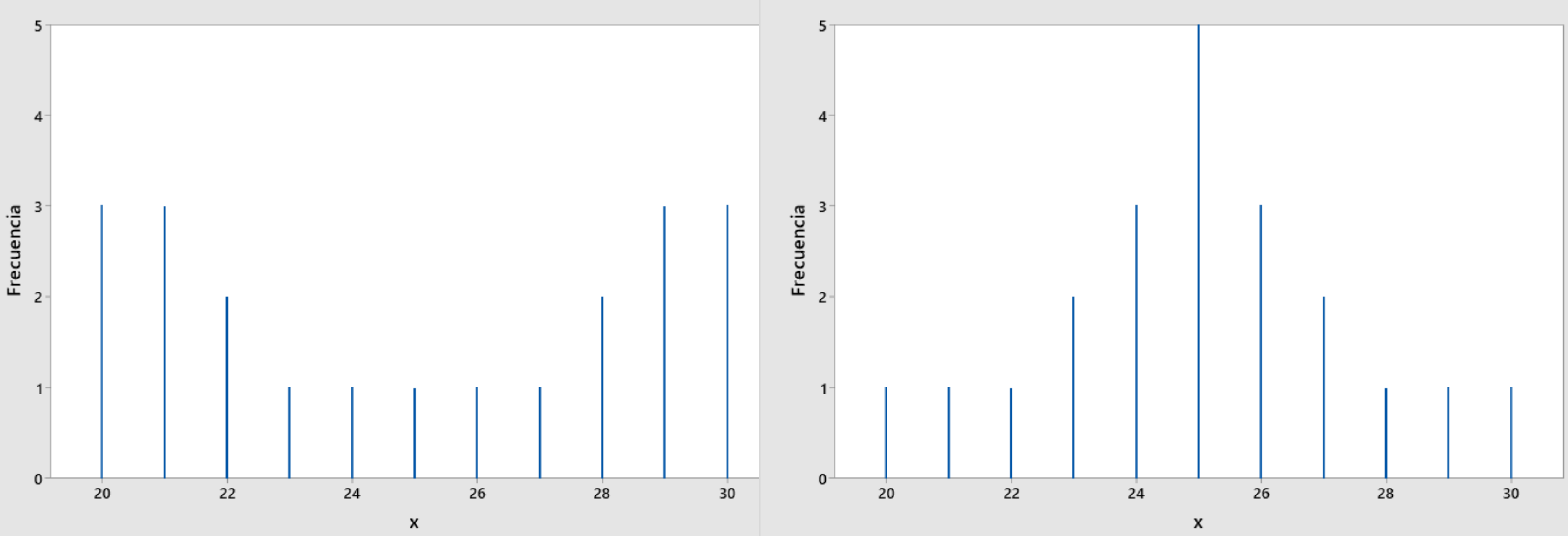
2) Medidas de dispersión o variabilidad: La dispersión se refiere al esparcimiento de los datos. O sea, al grado de variabilidad de las observaciones. Un valor pequeño en una medida de dispersión indica que los datos están estrechamente agrupados alrededor del centro, mientras que un valor grande indica una alta variabilidad.

Rango (R):

Es la diferencia entre el valor máximo y el mínimo que toma la variable.

$$R=X_{máx}-X_{mín}$$

Si bien esta medida de variabilidad es la más fácil de calcular, no es la mejor ya que no tiene en cuenta toda la información de los valores en el centro de la distribución. Veamos el siguiente ejemplo:



Cálculo del rango para el conjunto de datos del ejemplo

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

R=5-1=4 hermanos

Interpretación: La máxima diferencia que existe entre el número de hermanos de los encuestados en la EPH es de 4.

Rango Intercuartílico (RI)

Mide la amplitud del 50% central de la distribución. Esta medida de variabilidad acompaña a la mediana, y como ella, no se ve afectada por valores extremos.

$$RI=Q_3-Q_1$$

Cálculo del rango intercuartílico para el conjunto de datos del ejemplo:

RI= 4 - 1 =3

Interpretación: La dispersión del 50% central del número de hermanos es de 3.

Desvío estándar (S) y Variancia(S^2)

Son las medidas de dispersión más utilizadas, acompañan a la media aritmética.

La variancia es el promedio de la sumas de los cuadrados de los desvíos de los valores de la variable, respecto a la media. Es difícil de interpretar porque está medida en unidades al cuadrado, por lo tanto se interpreta el desvío estándar que tiene la misma unidad de medida que la variable bajo estudio.

Las diferencias se elevan al cuadrado para reflejar la distancia absoluta entre los valores y la media. Si se dejaran las diferencias sin elevar, la suma de esas diferencias sería igual a 0, que es una de las características de la media ya mencionadas.

Desvío estándar (S) y Variancia(S^2)

El desvío estándar es la raíz cuadrada de la variancia e indica cuánto se alejan, en promedio, los valores respecto de la media del grupo. Se ve afectada por valores extremos.

Variancia

Desvío estándar

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

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Cálculo del desvío estándar para el conjunto de datos del ejemplo

$$S = \sqrt{\frac{(2-2{,}33)^2 + (1-2{,}33)^2 + \ldots + (3-2{,}33)^2}{15}} = 1{,}45$$

Interpretación: En promedio, el número de hermanos se aleja en 1,45 hermanos de la media.

El desvío estándar se puede utilizar para comparar la dispersión entre dos o más conjuntos de datos siempre y cuando las medias sean similares!!

Nota: tanto la media aritmética como la desviación estándar y la variancia pueden ser obtenidos al cargar la base de datos en una calculadora científica.

En el siguiente video se ve como se realiza para el modelo más común en el mercado.

https://www.youtube.com/watch?v=qguhqq0xvM0

El enlace siguiente es para otros modelos frecuentes:

https://www.youtube.com/watch?v=4eXrcn6QLgs

MANOS A LA OBRA N° 6

ACTIVIDAD 1: Encuentre las medidas de dispersión de las edades de cada uno de los dos grupos de niños que participaron del estudio para comparar dos tratamientos para la angina. Interprete los valores en términos del problema.

Amoxicilina: 14 – 17 – 11 – 10 – 11 – 14 – 9 – 12 – 8 – 10 – 9
Cefadroxil: 9 – 14 – 8 – 10 – 13 – 7 – 9 – 11 – 16 – 10 – 12 – 9

ACTIVIDAD 2: Encuentre las medidas de dispersión del ingreso por ventas de autos, en millones de pesos, en 14 concesionarias de la ciudad de Rosario en el mes de junio de 2021.
50,61 – 60,85 – 63,70 – 70,98 – 88,90 – 44,07 – 56,87 – 48,84 – 64,41 – 40,85 – 65,04 – 61,38 – 45,39 – 82,80

ACTIVIDAD 3: Encuentre las medidas de dispersión de la cantidad de autos vendidos en 9 concesionarias de la ciudad de Rosario en el mes de mayo de 2021.
4 – 5 – 5 – 6 – 6 – 6 – 7 – 7 – 9

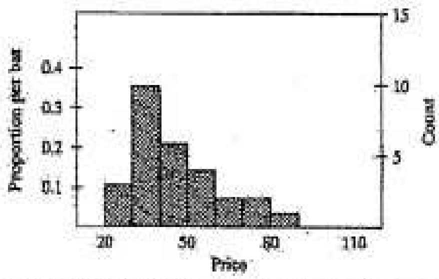
MANOS A LA OBRA N° 7

PARA RESOLVER 4.10 (página 257). Calzado de competición

☑ **PARA RESOLVER !!!**

4.10. Más sobre el calzado de competición

Considere la distribución de los precios de los calzados de competición dada por:



a) ¿Qué valores representarían mejor a la media y a la mediana?

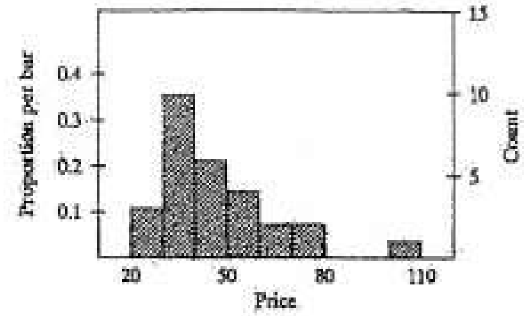
Valores: 40 45

Media:........ ☐ ....

Mediana:...... ☐ ....

Explique:..

b) Suponga que una observación más grande fue incorrectamente ingresada, según lo demuestra la siguiente figura:



Marcar con una cruz:

	Incrementará	Decrecerá	Será la misma
La media			
La mediana			
El desvío estándar			

<u>Coeficiente de variación</u> (CV)

El coeficiente de variación es una medida de variabilidad que elimina las unidades de medida (es un número puro) y cancela el efecto de distorsión que provocan las magnitudes promedios muy diferentes.

Expresa al desvío estándar como un porcentaje de la media. El conjunto de datos que posee un coeficiente de variación menor es más homogéneo, es decir, presenta menor dispersión en los valores, menor variabilidad.

$$CV = \frac{S}{\bar{x}} * 100$$

Cálculo del coeficiente de variación para el conjunto de datos del ejemplo

$$\boldsymbol{CV} = \left(\frac{\mathbf{1,45}}{\mathbf{2,33}}\right) * \mathbf{100} = \mathbf{62,23\%}$$

Interpretación: El desvío estándar representa el 62,23% de la media del número de hermanos.

¡También se pueden utilizar los diagramas de caja para comparar homogeneidad entre dos o más distribuciones!

MANOS A LA OBRA N° 8

ACTIVIDAD: Indique qué grupo es más homogéneo.

Amoxicilina: 14 – 17 – 11 – 10 – 11 – 14 – 9 – 12 – 8 – 10 – 9
Cefadroxil: 9 – 14 – 8 – 10 – 13 – 7 – 9 – 11 – 16 – 10 – 12 – 9

Diagrama de caja (boxplot)

- Es una representación gráfica muy utilizada que permite describir un conjunto de datos. Se basa en 5 medidas de resumen: $X_{mín}$, Q_1, Mna, Q_3 y $X_{máx}$.
- Permite analizar la forma de la distribución, si la Mna divide a la caja en dos partes iguales la distribución es simétrica en su parte central. Si además, las distancias entre la caja (Q_1) y el valor mínimo y la distancia entre la caja (Q_3) y el máximo la distribución es simétrica. En caso contrario presenta una asimetría.
- También permite observar entre qué valores se encuentra el 50% central de los datos (RI), comparar dos o más distribuciones, etc.

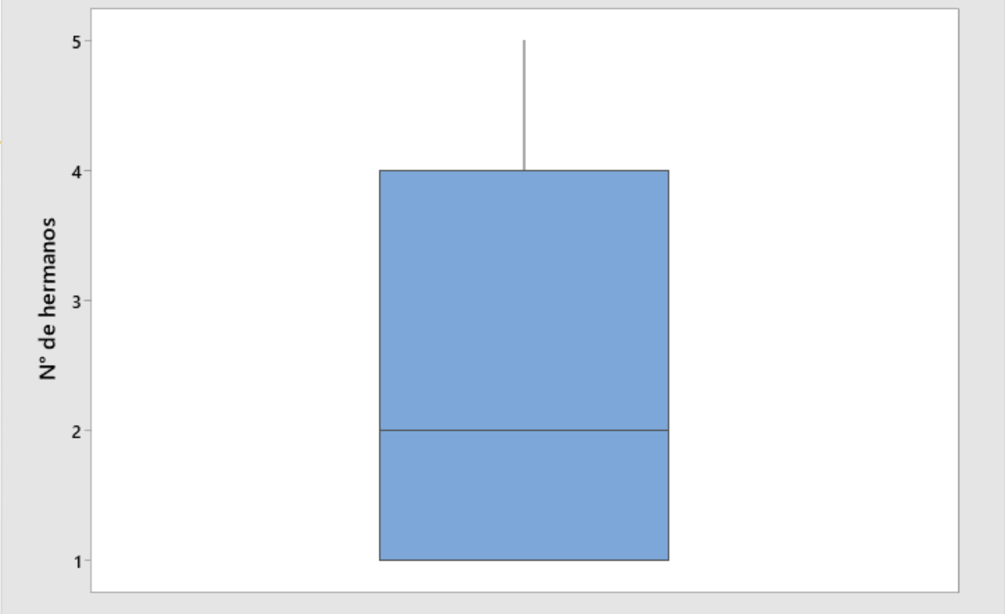
Diagrama de caja (boxplot)

Este gráfico permite detectar potenciales outliers (observaciones que no son típicas del conjunto). Se considerarán potenciales outliers aquellas observaciones que caigan por fuera de:

$$Q_1 - 1.5 \text{ RI} \; ; \; Q_3 + 1.5 \text{ RI}$$

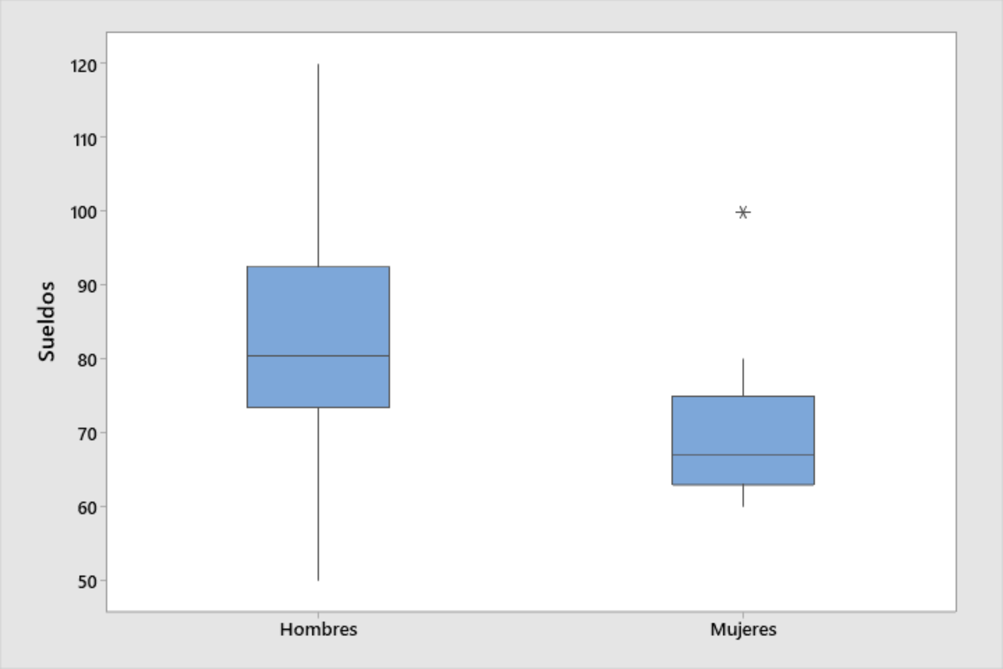
Los outliers se marcan en el gráfico como puntos separados.

Recordando que el mínimo es 1, el Q_1=11, Mna es 22, Q_3=4 y el máximo es 5, el diagrama de caja para el número de hermanos resulta:



Interpretación: La distribución del número de hermanos muestra que la mediana se encuenta más cerca del primer cuartil que del tercero, lo que indica una asimetría hacia la izquierda. El 50% central del número de hermanos se encuentra entre 1 y 4.

Estos diagramas se utilizan mucho para comparar la distribución de la misma variable en distintos grupos. Por ejemplo se evalúa el sueldo, en miles de pesos, de hombres y mujeres de determinada empresa.



Comparando ambas cajas se observa que en el grupo de hombres los datos son más heterogeneous y los sueldos presentan una distribución apróximadamente simétrica. Mientras que los sueldos de las mujeres presentan una distribución asimétrica hacia la derecha.

MANOS A LA OBRA N° 9

<u>PARA RESOLVER! 4.7: COMPARANDO EDADES. ESTUDIO DE LOS ANTIBIÓTICOS (página 246)</u>

Se estudia la edad de los niños que participaron de un estudio sobre dos antibióticos: Amoxicilina y Cefadroxil. En este tipo de estudios es importante comparar si las condiciones generales de los grupos son homogéneas, en este caso si los niños que recibieron los distintos antibióticos tienen edades similares, ya que niños más pequeños o más grandes podrían tener reacciones diferentes a los tratamientos.

Interprete el gráfico obtenido. ¿Cómo es la distribución de la edad en cada uno de los dos grupos de niños? ¿Puede identificarse algún outlier en alguno de los dos grupos?

MANOS A LA OBRA N° 9