INDICADORES ESTADÍSTICOS

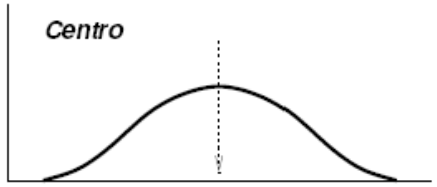
Explorando los Datos

"Generalmente las decisiones se toman fundamentadas en indicadores resumen"

Medidas de Resumen

*"En ocasiones **la evaluación grafica** puede ser una fuente de error por **percepción**, los **indicadores contribuyen a corroborar lo observado**"*

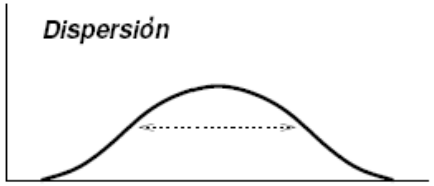
TENDENCIA CENTRAL



- MEDIA (PROMEDIO) → $\boldsymbol{\mu}$ ó $\boldsymbol{\bar{X}}$
- MEDIANA → $\boldsymbol{Me}$
- MODA → $\boldsymbol{Mo}$

Representan en un solo valor central, a un conjunto de datos.

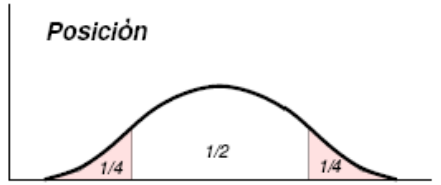
DISPERCIÓN



- RANGO → $\boldsymbol{R}$
- DESVAICIÓN ESTANDAR → $\boldsymbol{\sigma}$ ó $\boldsymbol{s}$
- VARIANZA → $\boldsymbol{\sigma^2}$ ó $\boldsymbol{s^2}$
- COEFICIENTE DE VARIACIÓN → $\boldsymbol{CV}$

Indican sobre la variabilidad de los datos, con respecto a un valor central.

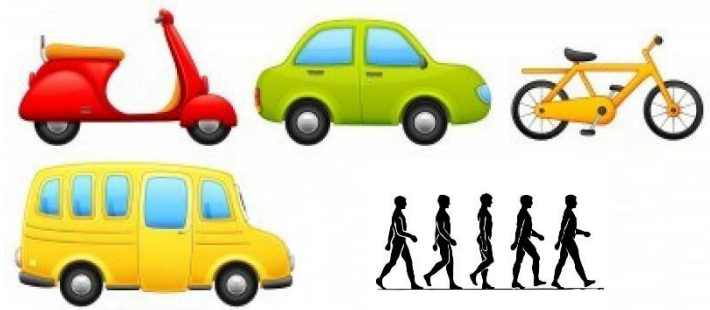
POSICIÓN



- Cuartiles → $\boldsymbol{Q_1}$, $\boldsymbol{Q_2}$, $\boldsymbol{Q_3}$
- Percentiles → $\boldsymbol{P_{25}}$, $\boldsymbol{P_{50}}$, $\boldsymbol{P_{75}}$

Medidas para resumir la distribución de los datos.

¿Cuánto tiempo tarda un estudiante en desplazarse desde su casa hasta la universidad?



Indicadores de tendencia central

- Es un valor que esta en el centro o punto medio de un conjunto de datos.
- Tiene como objetivo resumir los datos en un valor típico o representativo del conjunto de valores.

- **MEDIA** $\rightarrow \mu$ ó $\overline{X}$

La **media** es el valor representativo de todos los datos. A la media se le suele llamar *promedio*.

Si las observaciones son $x_1, x_2, \ldots, x_n$, su media es:

Media Aritmética

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + \ldots + x_n}{n}$$

Datos puntuales

=PROMEDIO(DATOS)

Media Ponderada:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$$

Datos agrupados

Ejemplo:

Los siguientes datos se han obtenido al observar el número de chocolatinas defectuosas en una muestra de 10 cajas de un lote de producción:

2 3 2 2 2 3 1 3 0 4

¿Cuál es el número promedio de unidades defectuosas por caja?

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2+3+2+2+2+3+1+3+0+4}{10} = \frac{22}{10} = 2.2$$

El proceso presenta en promedio 2.2 unidades defectuosas por caja.

- **MEDIA** $\rightarrow \mu$ ó $\overline{X}$

Ejemplo

Calcula el numero promedio de unidades defectuosas por caja :

Unidades defectuosas	ni
0	3
1	11
2	13
3	8
4	4
5	1
Total	40

$$\overline{x} = \frac{\sum_{i=1}^{40}(x_i n_i)}{40} = \frac{(0*3)+(1*11)+\cdots+(5*1)}{40} = 2.05 \approx 2$$

Las cajas en promedio tiene 2 unidades defectuosas.

OBSERVACIONES:

- La media puede ser la medida de centralidad más utilizada y más informativa. No obstante la media tiene una desventaja, la cual radica en que es sensible a los valores extremos.

Ejemplo:

Sean los valores: 2, 4, 3, 6, 5

$$\overline{x} = \sum_{i=1}^{5} \frac{x_i}{5} = \frac{2+4+3+6+5}{5} = \frac{20}{5} = 4$$

Sean los valores: 2, 4, 3, 6, 50

$$\overline{x} = \sum_{i=1}^{5} \frac{x_i}{5} = \frac{2+4+3+6+50}{5} = \frac{65}{5} = 13$$

- **MEDIA** $\rightarrow \boldsymbol{\mu}$ ó $\boldsymbol{\bar{X}}$

Ejemplo:

Carlos quiere decidir entre dos ciudades donde puede trabajar. Sin embargo, considera que la temperatura de cada ciudad es un factor importante en su elección.

Las mediciones de la temperatura máxima (°C), en cada una de las ciudades durante los últimos ocho meses son las siguientes:

Ciudad 1: 25 24 26 27 28 27 26 27

Ciudad 2: 16 22 24 28 32 31 30 27

¿Qué le podemos recomendar con base a la temperatura máxima promedio?

$$\bar{x}_1 = 26.25 \quad y \quad \bar{x}_2 = 26.25$$

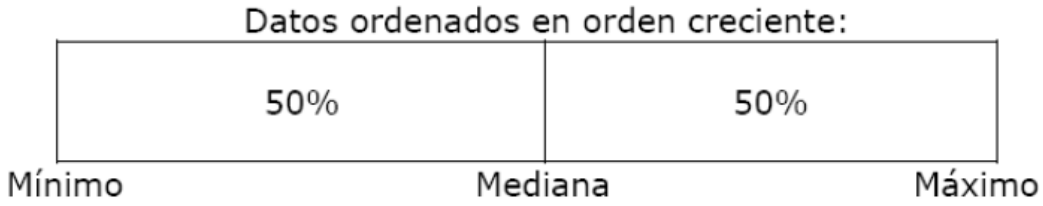
Las dos ciudades presentan la misma temperatura máxima promedio en los últimos ocho meses, que es de 26.25 °C.

Indicadores de tendencia central

- Es un valor que esta en el centro o punto medio de un conjunto de datos.
- Tiene como objetivo resumir los datos en un valor típico o representativo del conjunto de valores.

- **MEDINA → Me**

La **mediana** (*Me*) representa el punto medio de las observaciones ordenadas; la mitad (50%) de las observaciones están por debajo de la mediana y la otra mitad (50%) están por encima.



Si $x_1, x_2, \ldots, x_3$ corresponde a un conjunto de datos ordenados de manera ascendente de una variable **X**, entonces la mediana puede ser calculada como:

$$Me = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & si\ \boldsymbol{n}\ es\ impar \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & si\ \boldsymbol{n}\ es\ par \end{cases}$$

=MEDIANA(DATOS)

Ejemplo:

Número de chocolatinas defectuosas en una muestra de 10 cajas de un lote de producción:

0 1 2 2 2 2 3 3 3 4

¿Cuál es la mediana del numero de unidades defectuosas por caja?

$$Me = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{2+2}{2} = 2$$

El 50% de las cajas presentan 2.0 o menos unidades defectuosas.

OBSERVACIONES:

- La mediana no es influenciable por los valores extremos.

Ejemplo:

Sean los valores: 2, 3, 4, 5, 6

$$\overline{x} = \sum_{i=1}^{5} \frac{x_i}{5} = \frac{2+4+3+6+5}{5} = \frac{20}{5} = 4$$

$$Me = X_{\left(\frac{5+1}{2}\right)} = X_{(3)} = 4$$

Sean los valores: 2, 3, 4, 6, 50

$$\overline{x} = \sum_{i=1}^{5} \frac{x_i}{5} = \frac{2+4+3+6+50}{5} = \frac{65}{5} = 13$$

$$Me = X_{\left(\frac{5+1}{2}\right)} = X_{(3)} = 4$$

Indicadores de tendencia central

- Es un valor que esta en el centro o punto medio de un conjunto de datos.
- Tiene como objetivo resumir los datos en un valor típico o representativo del conjunto de valores.

- **MODA** → **Mo**

***Mo*: Es el valor que más se repite**

=MODA(DATOS)

- Cuando la variable de interés, es de naturaleza discreta, la moda (Mo) corresponde al dato de la muestra que tiene mayor frecuencia.
- Cuando se trata de una variable de naturaleza continua, la moda corresponde al(os) valor(es) alrededor del(os) cual(es) se produce una mayor concentración de datos, es decir a los puntos de mayor densidad de frecuencia.
- A diferencia de los otros indicadores este es el único que puede ser calculado cuando observamos variables cualitativas.

Ejemplo:

Número de chocolatinas defectuosas en una muestra de 10 cajas de un lote de producción:

0 1 2 2 2 2 3 3 3 4

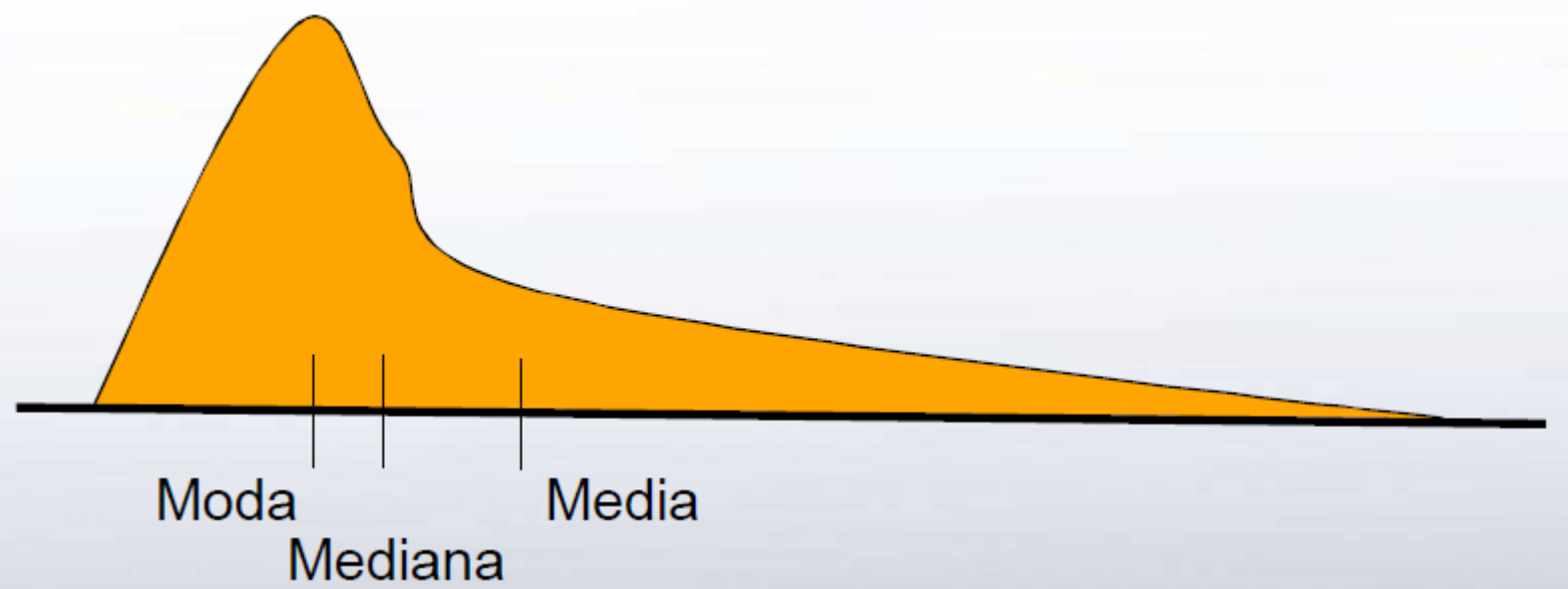
¿Cuál es la moda del del numero de unidades defectuosas por caja?

$$Mo = 2$$

La mayoría de las cajas presentan 2.0 unidades defectuosas.

Medidas de Tendencia Central

¿Cuál elegir?



Medidas de Tendencia Central

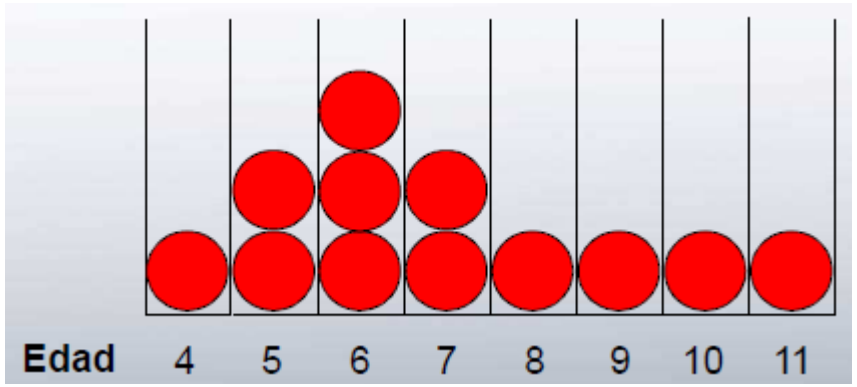
Consideremos la siguiente tabla de datos:

Individuo	Nombre	Edad
1	Juan	4
2	Alberto	5
3	Inés	6
4	Aurora	5
5	Rodrigo	6
6	Martha	7
7	Juana	8
8	Roberto	6
9	Silvia	7
10	Ana	11
11	Andres	10
12	Carlos	9

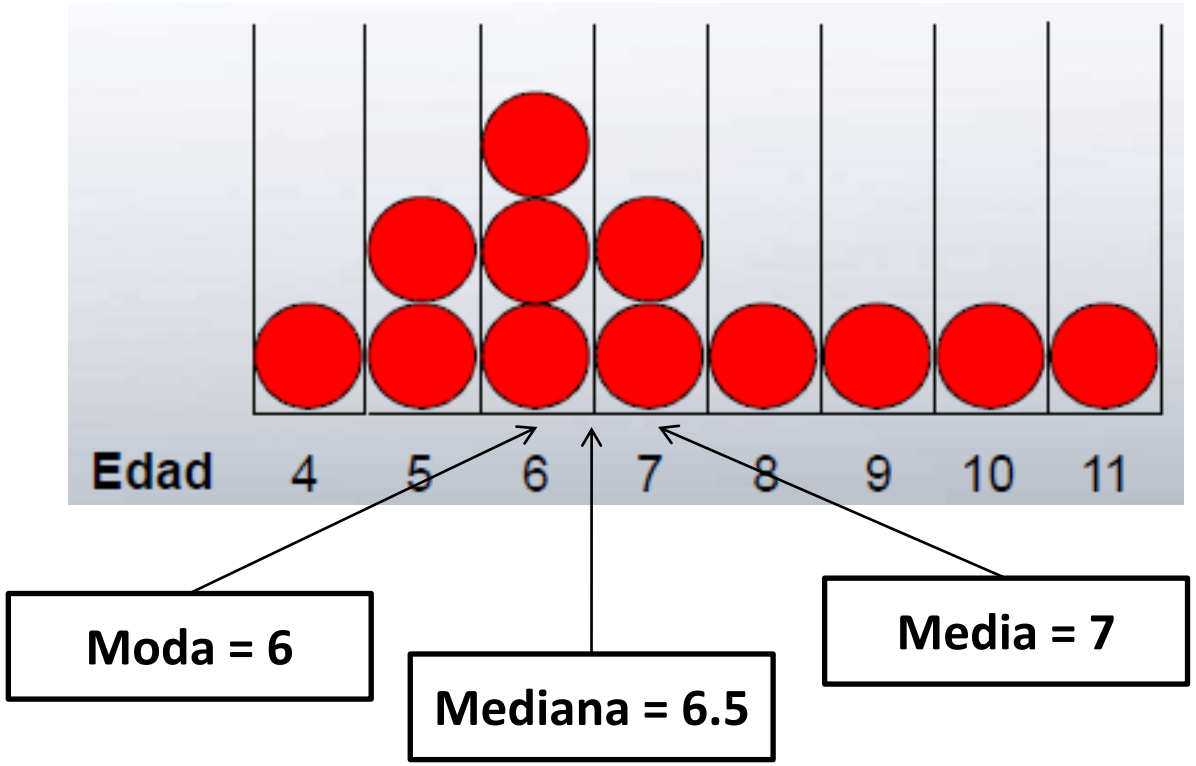
Medidas de Tendencia Central

Consideremos la siguiente tabla de datos:

Edad	4	5	6	7	8	9	10	11	Total
Frecuencia	1	2	3	2	1	1	1	1	12

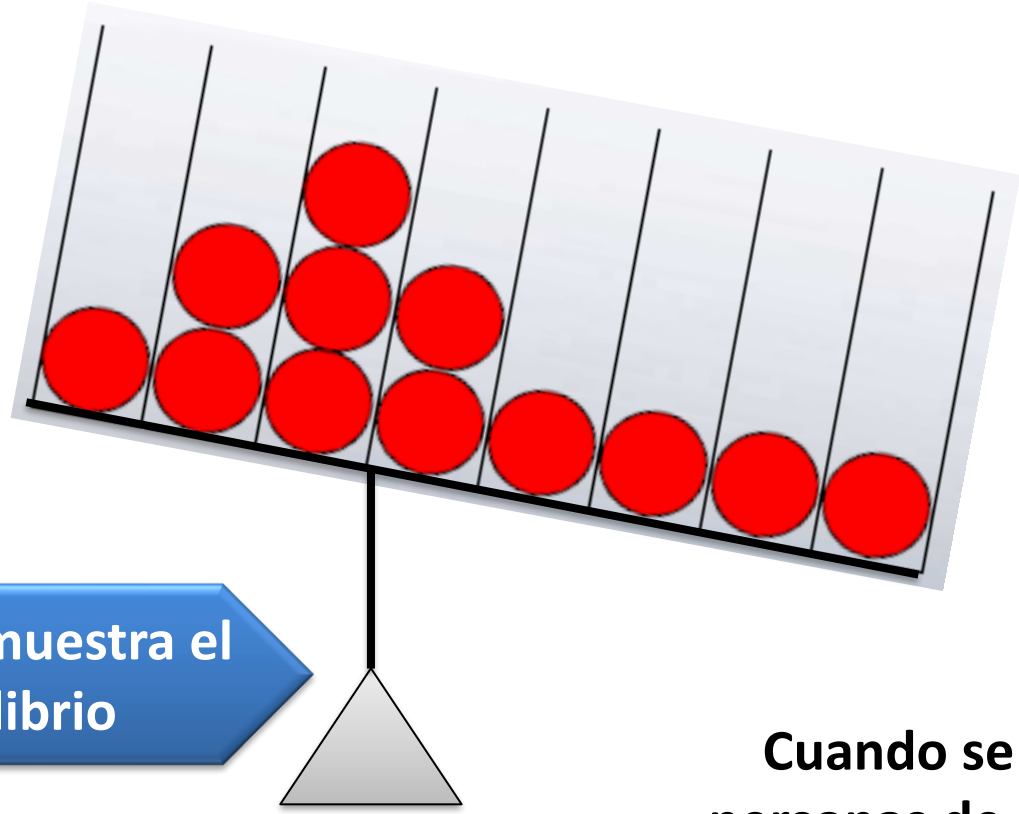


Medidas de Tendencia Central



Medidas de Tendencia Central

Medidas de Tendencia Central



La mediana muestra el desequilibrio

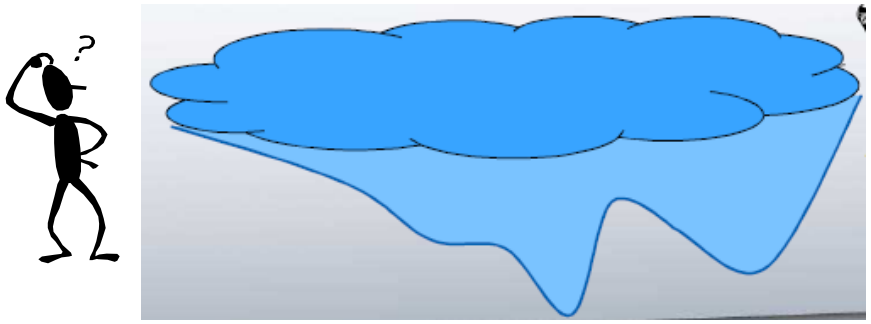
Cuando se pone 6 personas de cada lado

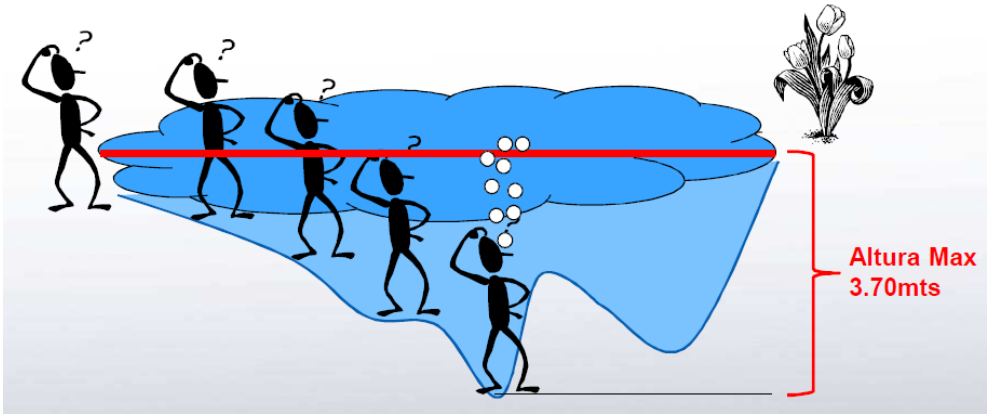
INDICADORES DE DISPERSION

¿Por que se estudia la dispersión?

Pregunta: ¿Un Guía turístico le dice a usted que cierto río tiene una profundidad los siguientes indicadores:

Media:	100 cms
Mediana:	50 cms
Moda:	30 cms





¿Con está información cruzaría usted el río?

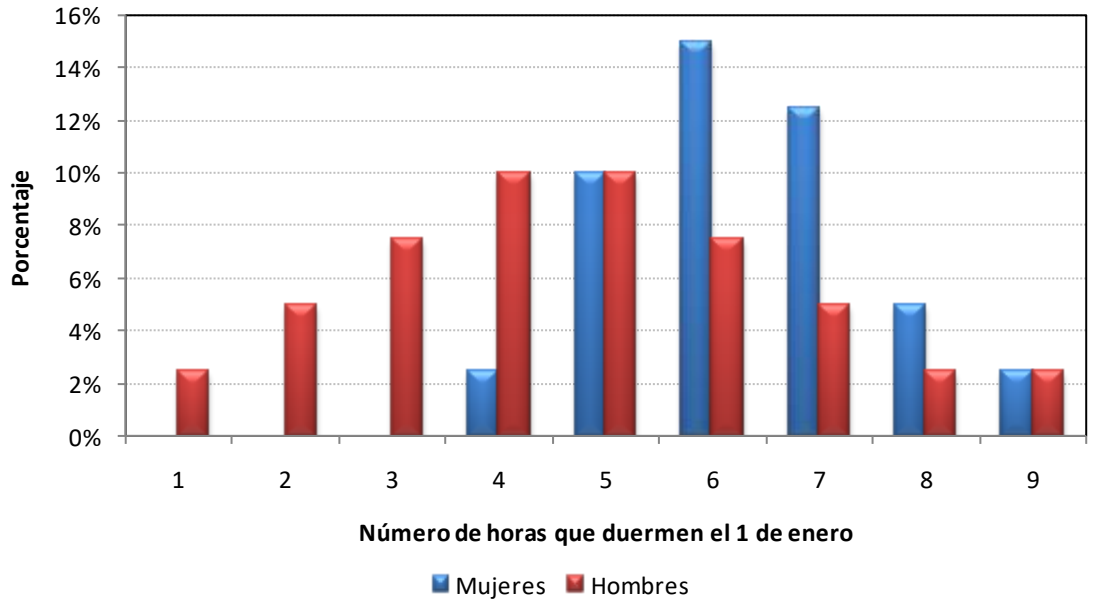
- Los Indicadores de Tendencias central no trabajan solos, deben apoyarse con alguna medida de dispersión
- Un Índice de dispersión pequeño, indica baja variabilidad, por ende el valor de tendencia central será mas confiable.
- Un índice de dispersión grande, indica gran variabilidad, esto implica que el Indicador de tendencia central sea poco confiable

INDICADORES DE DISPERSION

Todo conjunto de datos tiene al menos dos características principales:

CENTRO Y DISPERSIÓN

Los gráficos de barra, histogramas, de puntos, entre otros, nos dan cierta idea sobre ellos.



Indicadores de Dispersión

- Un índice de dispersión pequeño, indica baja variabilidad, por ende el valor de tendencia central será mas confiable.
- Un índice de dispersión grande, indica gran variabilidad, esto implica que el Indicador de tendencia central sea poco confiable.

• Varianza → s^2

La **varianza** representa el promedio de las desviaciones al cuadrado entre cada observación y la media.

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

- Es siempre positiva.
- Si todos los valores son iguales, entonces la varianza es cero.
- Le afectan los valores extremos.
- La varianza no se interpreta.

=VAR(DATOS)

• Desviación Estándar → s

La **desviación estándar**, indica en promedio cuanto se desvían las observaciones con respecto a la media.

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

- Mientras más alejadas están las observaciones del promedio, mayor será la desviación estándar.

=DESVEST(DATOS)

Ejemplo:

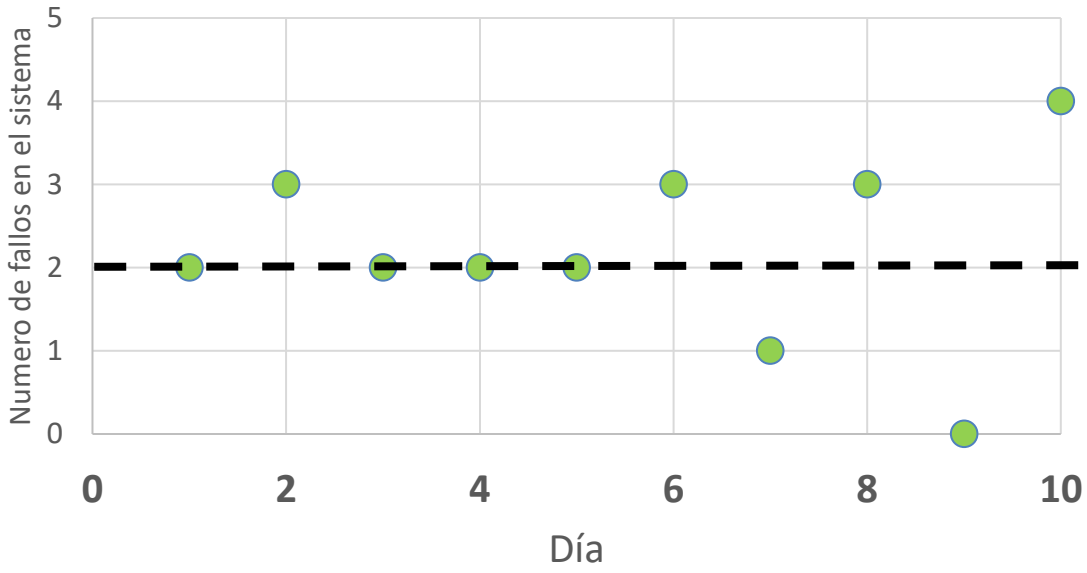
Los siguientes datos se han obtenido al observar el número de chocolatinas defectuosas en una muestra de 10 cajas de un lote de producción:

2 3 2 2 2 3 1 3 0 4

¿Qué tanto varía la cantidad de unidades defectuosas respecto al promedio?

$$s = 1.14$$

La cantidad de unidades defectuosas del proceso varia respecto a su promedio en 1.14 unidades.



Indicadores de Dispersión

Por la estructura de la varianza se sabe que cuando aumenta la dispersión el valor de la varianza aumenta, al igual que la desviación estándar.

pero, qué se respondería a las preguntas:

- ¿una desviación estándar de 200 metros indica que hay poca o mucha dispersión?
- ¿una desviación estándar de 100 kilogramos podría ser grande?

Depende de la magnitud de los datos.

Coeficiente de Variación

- **Coeficiente de Variación → *CV***

Indica el grado de variabilidad porcentual de los datos con respecto a la media. Se suele interpretar en términos porcentuales.

$$CV = \frac{S}{\bar{X}}$$

- Útil para comparar la variabilidad relativa de una característica, en poblaciones que tiene diferente media.

Si ***CV*** < 30% → la variabilidad es baja.
Si 30% ≤ ***CV*** ≤ 80% → la variabilidad es moderada.
Si *CV* > 80% → la variabilidad es alta.

=DESVEST(DATOS)/PROMEDIO(DATOS)

$$CV = \frac{s}{\bar{x}} = \frac{1.14}{2.2} = 0.52 \simeq 52\%$$

Ejercicio

El entrenador de un equipo de baloncesto duda entre seleccionar a Eva o a Tatiana. Los puntos conseguidos por cada una, en una semana de entrenamiento, fueron los siguientes:

Eva	18	23	22	24	19	25	16
Tatiana	18	26	18	28	22	17	18

Calcule e interprete:

a) ¿Cuál de las dos tiene mejor promedio de puntos?

Ambas tienen el mismo promedio de 21 puntos

b) ¿Cuál de las dos es más regular en sus puntos?

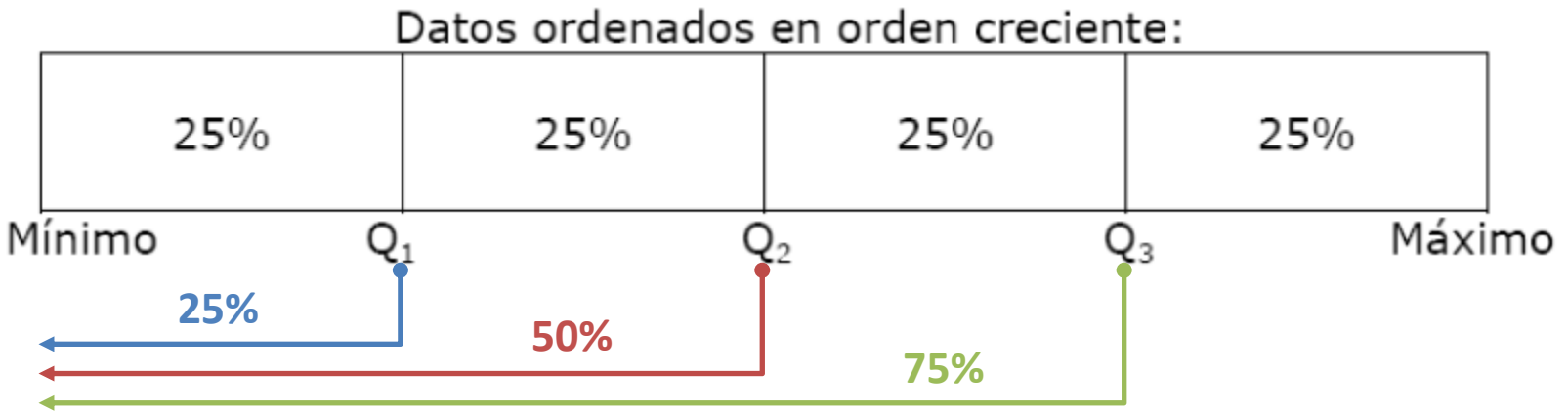
Es más regular Eva, porque la dispersión de los puntos es menor.

Desv. Est. (Eva) = 3.367 Desv. Est. (Tatiana) = 4.435

Indicadores de Posición

Cuartiles:
Q1, Q2, Q3

=CUARTIL(DATOS;#)



Los cuartiles, son tres valores Q_1, Q_2 y Q_3, que dividen a las observaciones de forma ordenada, en cuatro partes que contienen aproximadamente el mismo número de datos. Estos tres indicadores junto al **mínimo (*mín*)** y el **máximo (*máx*)** conforman los 5 números resumen.

«La representación grafica de los Cuartiles es el diagrama de cajas y alambres»

Ejemplo

La resistencia a la tensión del caucho de silicio se considera una función de la temperatura de vulcanizado. Se llevó a cabo un estudio en el que se prepararon muestras de 12 especímenes del caucho utilizando temperaturas de vulcanizado de 20°C y 45°C. Los siguientes datos presentan los valores de resistencia a la tensión en megapascales:

20°C

2.07	2.14	2.22	2.03	2.21	2.03
2.05	2.18	2.09	2.14	2.11	2.02

45°C

2.52	2.15	2.49	2.03	2.37	2.05
1.99	2.42	2.08	2.42	2.29	2.01

¿Determinar las cinco medidas resumen de la resistencia a la tensión para cada temperatura?

Indicador	20°C	45°C
Mínimo	2.02	1.99
Cuartil 1	2.05	2.05
Cuartil 2	2.10	2.22
Cuartil 3	2.15	2.42
Máximo	2.22	2.52

«La representación grafica de los Cuartiles es el diagrama de cajas y alambres»

Boxplot (Diagrama de cajas y alambres)

Permiten hacerse una idea acerca de la forma de la distribución de una variable y su dispersión.

Q1= Valor que es superior al **25%** de las observaciones
Q2= Valor que es superior al **50%** de las observaciones
Q3 = Valor que es superior al **75%** de las observaciones

- **Rango Inter Cuartilico (RIC):**
  Se define como la diferencia entre el 3er y 1er cuartil

RIC = Q3 – Q1

Contiene el 50% de las observaciones

- La caja se construye entre los cuartiles Q1 y Q3 con un ancho arbitrario. Dentro de la caja se marca Q2, con un trazo.

- Los alambres que sales de Q1 y Q3, van hasta el dato más próximo al cerco inferior y superior (sin cruzarlos).

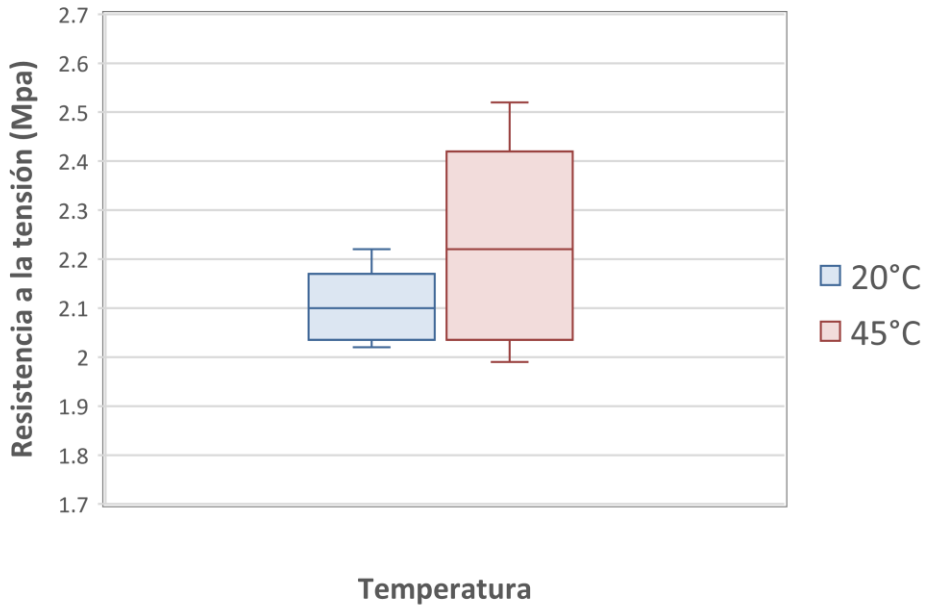
CI= Q1 – (1.5*RIC) **(Cerco Inferior)**
CS = Q3 + (1.5*RIC) **(Cerco Superior)**

- Los valores que salen de los cercos son marcados sobre el gráfico como puntos **(puntos atípicos)**.

Ejemplo

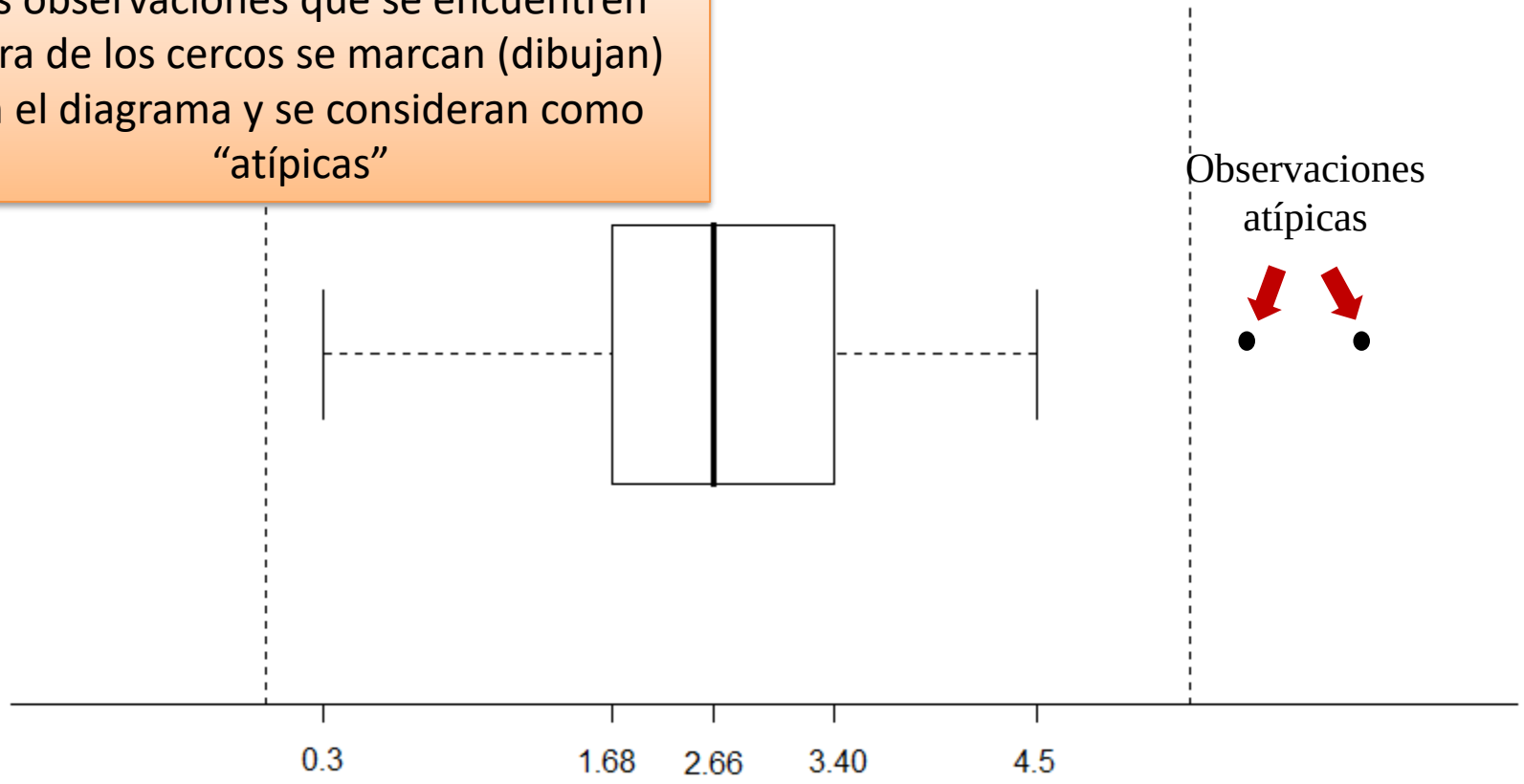
La resistencia a la tensión del caucho de silicio se considera una función de la temperatura de vulcanizado. Se llevó a cabo un estudio en el que se prepararon muestras de 12 especímenes del caucho utilizando temperaturas de vulcanizado de 20°C y 45°C. Los siguientes datos presentan los valores de resistencia a la tensión en megapascales.

Indicador	20°C	45°C
Mínimo	2.02	1.99
Cuartil 1	2.05	2.05
Cuartil 2	2.10	2.22
Cuartil 3	2.15	2.42
Máximo	2.22	2.52
RIC	0.11	0.38
CI	1.86	1.43
CS	2.38	3.08



Boxplot (Diagrama de cajas y alambres)

Las observaciones que se encuentren fuera de los cercos se marcan (dibujan) en el diagrama y se consideran como “atípicas”

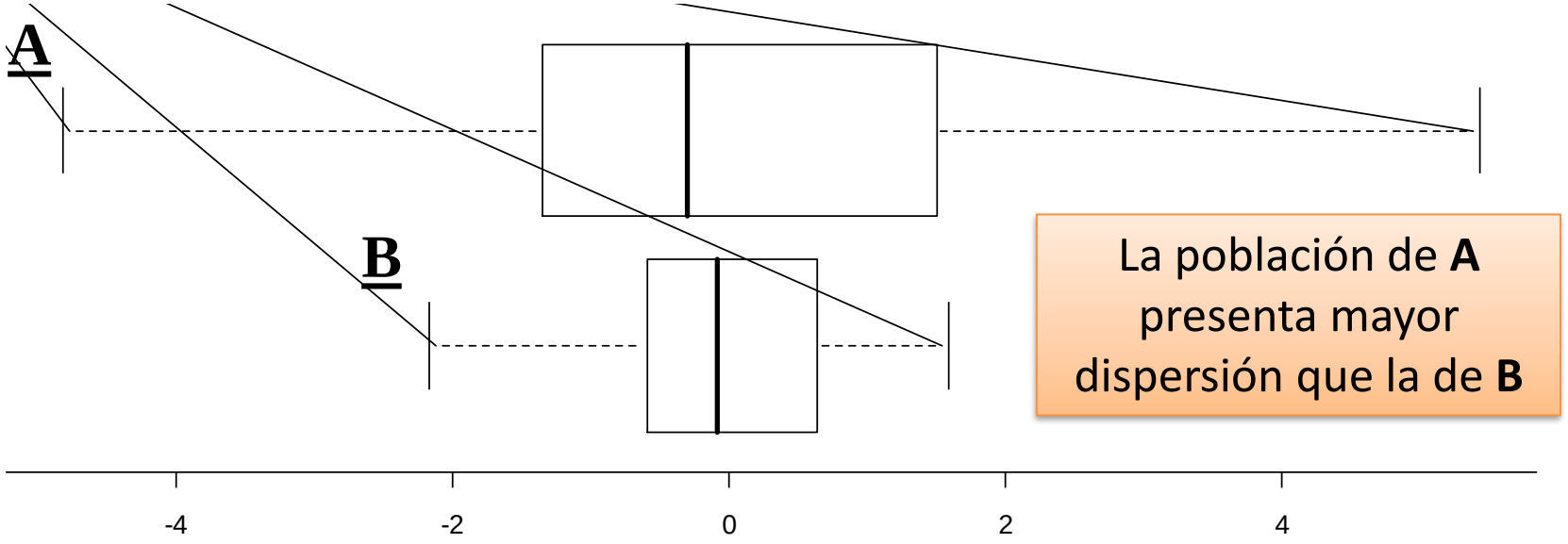


Las observaciones atípicas, son datos que tienen magnitudes “raras” con respecto al conjunto de datos.

Boxplot (Diagrama de cajas y alambres)

Los diagramas de cajas y alambres son útiles, entre otros para los siguientes propósitos:

1. Para identificar la localización de los datos alrededor de la mediana.
2. Para hacerse una buena idea de la dispersión de los datos, basándose en la longitud de la caja. Además se aprecia el rango de los datos

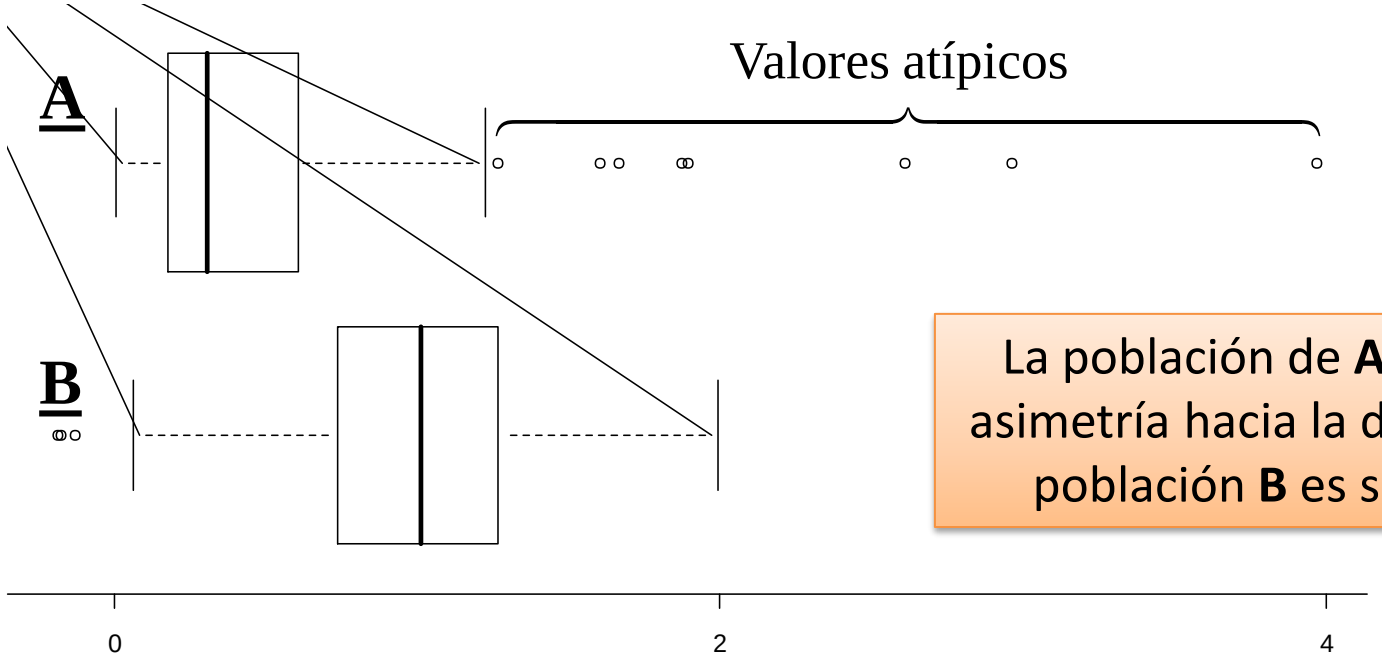


La población de **A** presenta mayor dispersión que la de **B**

Boxplot (Diagrama de cajas y alambres)

Los diagramas de cajas y alambres son útiles, entre otros para los siguientes propósitos:

3. Permite observar el grado de asimetría de una distribución, comparando las proporciones de la caja que queda a los lados de la mediana.
4. Útil para identificar posibles valores atípicos (fuera de los cercos)

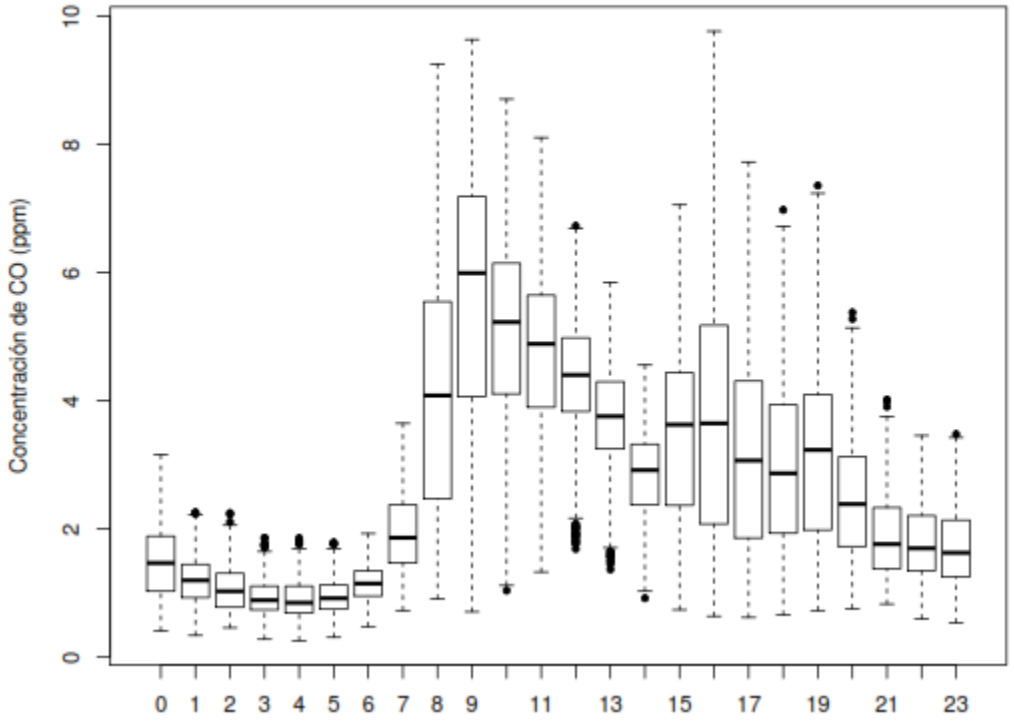


La población de **A** presenta asimetría hacia la derecha y la población **B** es simétrica

Boxplot (Diagrama de cajas y alambres)

En el siguiente gráfico se observa el comportamiento de los niveles de monoxido de carbono (CO) durante un día ordinario (lunes a viernes).

Figura: Diagrama de cajas y alambres de la concentración de CO por hora



Boxplot (Diagrama de cajas y alambres)

En un experimento se observó la longitud de los dientes de conejillos de indias para dos tipos de administración (zumo de naranja o acído ascórbico) y tres niveles de dosis de vitamina C (0.5, 1 y 2).

Figura: Longitud del diente según dosis y tipo de administración