Analítica de datos

Estadísticos descriptivos

Pontificia Universidad JAVERIANA Bogotá

Profesor: Nicolás Velásquez

Estadísticos descriptivos

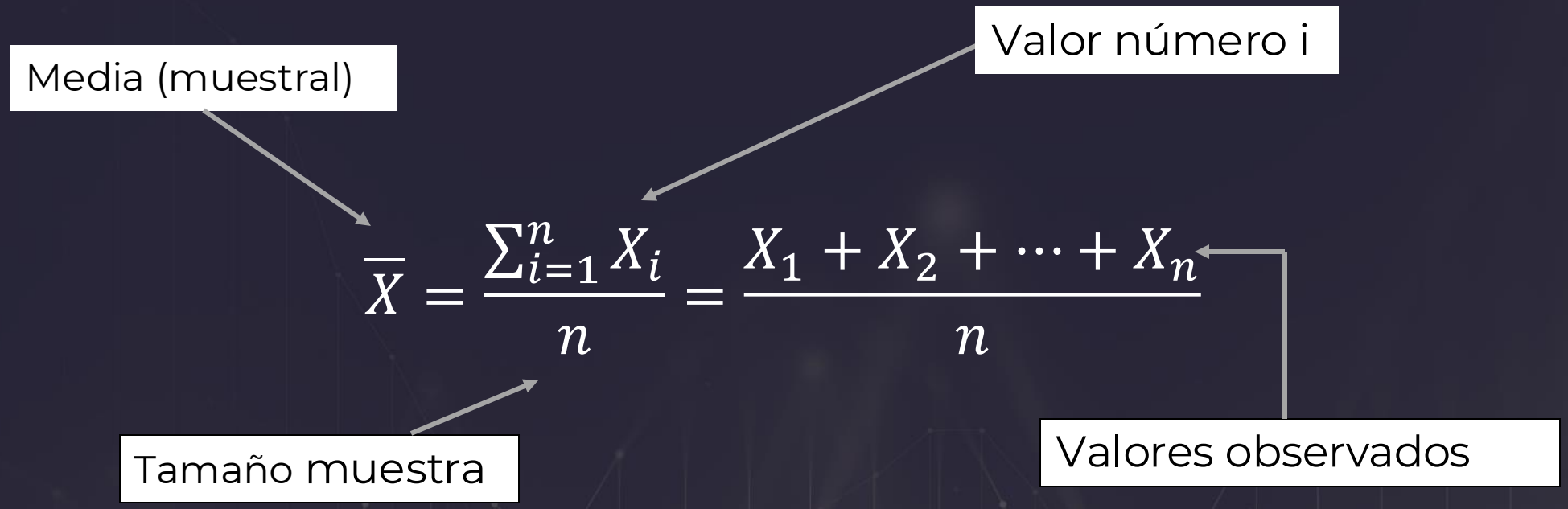
1. Medidas de tendencia central.
2. Medidas de variación.
3. Medidas de forma.
4. El diagrama de caja y brazos.
5. Medidas de relación lineal entre dos variables.

Resumen de Definiciones

- La **tendencia central** nos dice alrededor de qué valor se mueven los valores de una variable numérica.
- La **variación** nos dice qué tan dispersos están los datos o qué tanto se alejan de su tendencia central.
- La **forma** es el patrón de la distribución de los valores.

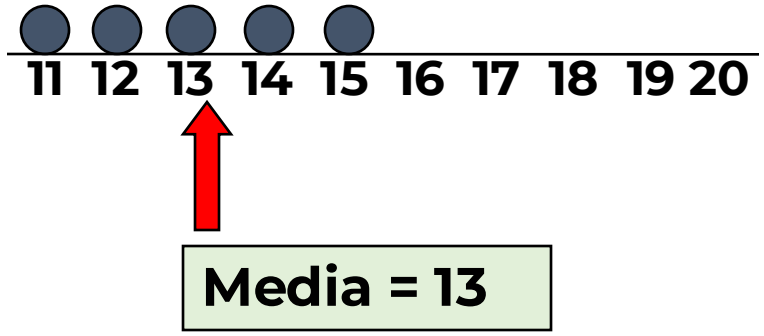
Medidas de Tendencia Central: Media (Average)

- La media es la medida más común de tendencia central:
  - Para una muestra de tamaño n:

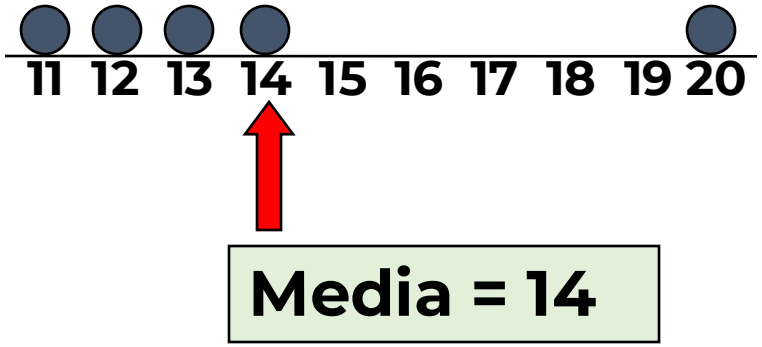


Medidas de Tendencia Central: Media

- Media = suma de los valores dividido el número de valores.
- Afectada por valores extremos (“outliers”)



$$\frac{11+12+13+14+15}{5}=\frac{65}{5}=13$$



$$\frac{11+12+13+14+20}{5}=\frac{70}{5}=14$$

Medidas de Tendencia Central:
Mediana

- La posición de la mediana cuando los valores están ordenados de más bajo a más alto:

$$Posición = \frac{n+1}{2}$$

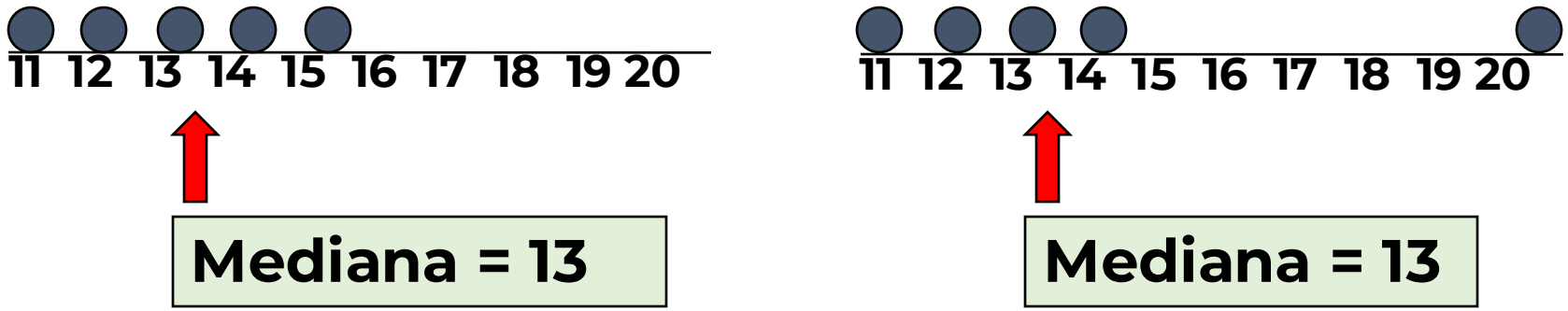
- Si el número de valores es impar, la mediana es el número del medio.

- Si el número de valores es par, la mediana es el promedio de los números del medio.

Note que $\frac{n+1}{2}$ no es el *valor* de la mediana, solamente la *posición* en los datos ordenados!

Medidas de Tendencia Central: Mediana

- Valor para el cuál 50% de las observaciones quedan abajo y 50% arriba.

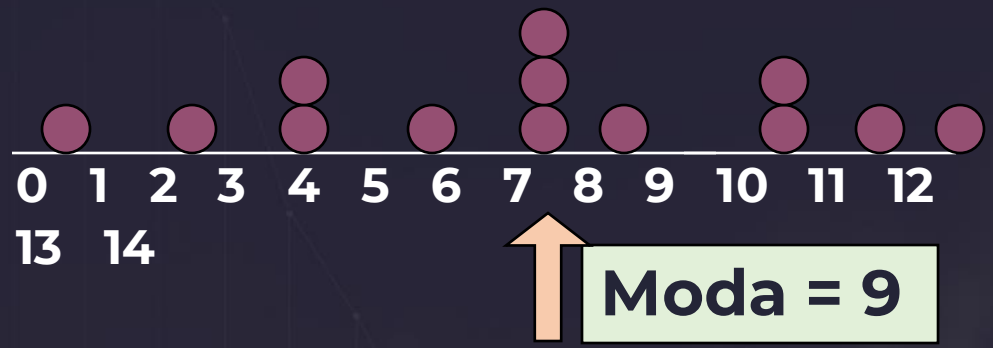


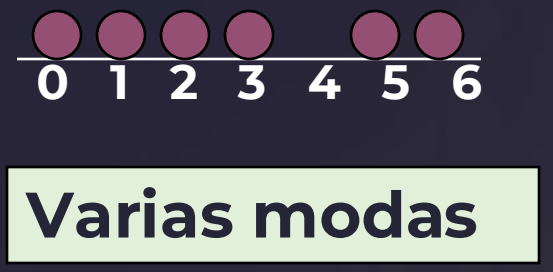
- Menos sensible que la media a valores extremos (outliers).

Medidas de Tendencia Central: Ejemplo

Medidas de Tendencia Central:
Moda

- Valor más frecuente.
- No afectada por valores extremos.
- Podría haber varias modas





Medidas de Tendencia Central: Un ejemplo

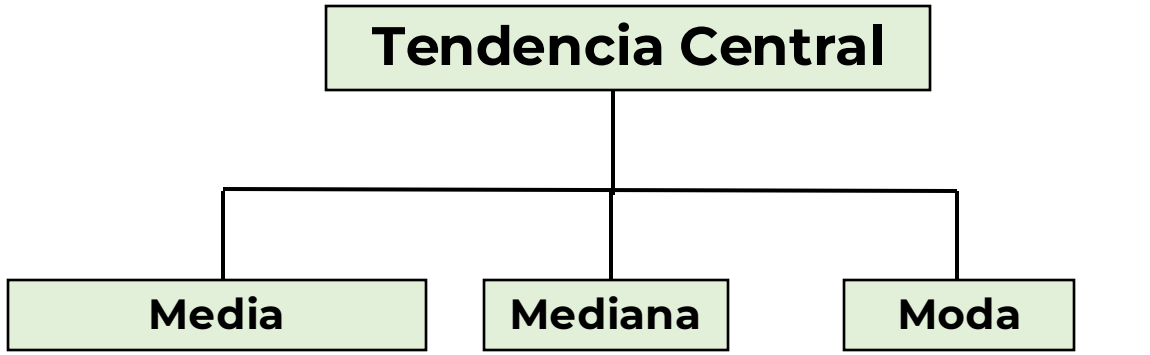
Precios casas:

$2,000,000
$ 500,000
$ 300,000
$ 100,000
$ 100,000

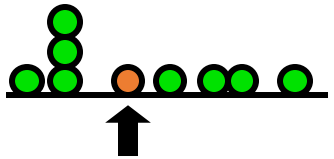
Suma $
3,000,000

- **Media:** ($3,000,000/5) = **$600,000**
- **Mediana** = **$300,000**
- **Moda** = **$100,000**

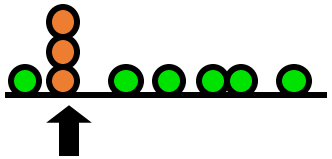
Medidas de Tendencia Central: Resumen



$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$



Valor de en medio en datos ordenados.



Valor más frecuente.

Medidas de Variación

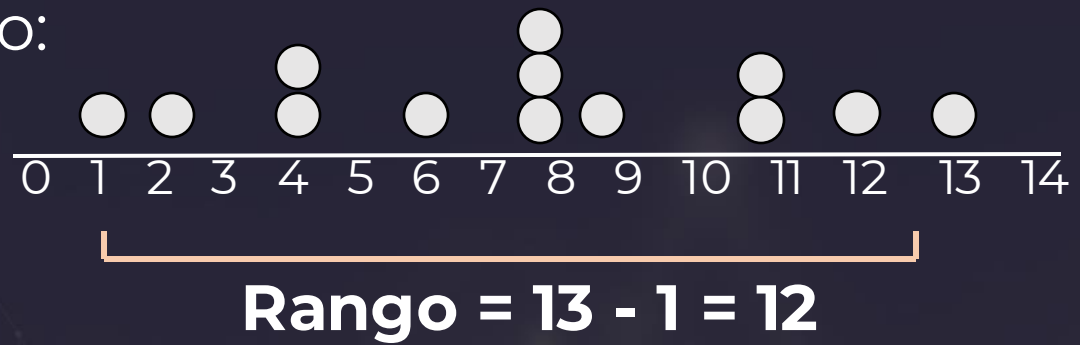
Medidas sobre la dispersión de los datos.

Medidas de Variación: Rango

- Medida más simple de variación.
- Diferencia entre el valor más grande y el más pequeño:

$$\text{Rango} = X_{+\text{grande}} - X_{+\text{pequeño}}$$

Ejemplo:



Rango = 13 - 1 = 12

Medidas de Variación: el rango es normalmente una mala medida

- No toma en cuenta cómo se distribuyen los datos.

Rango = 12 - 7 = 5

Rango = 12 - 7 = 5

- Sensible a outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Rango = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Rango = 120 - 1 = 119

Medidas de Variación: Varianza

- Promedio de las desviaciones de la media, al cuadrado:
  - Varianza (muestral):

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

donde $\overline{X}$ = media

n = tamaño muestra

X_i = valor i-ésimo

Medidas de Variación: Desviación Estándar

- Medida más común.
- Muestra variabilidad alrededor de la media.
- Es la raíz de la varianza.
- Expresada en las mismas unidades que los datos.

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Medidas de Variación: Desviación Estándar

Datos (X_i) :

10	12	14	15	17	18	18	24

n = 8 Media = $\overline{X}$ = 16

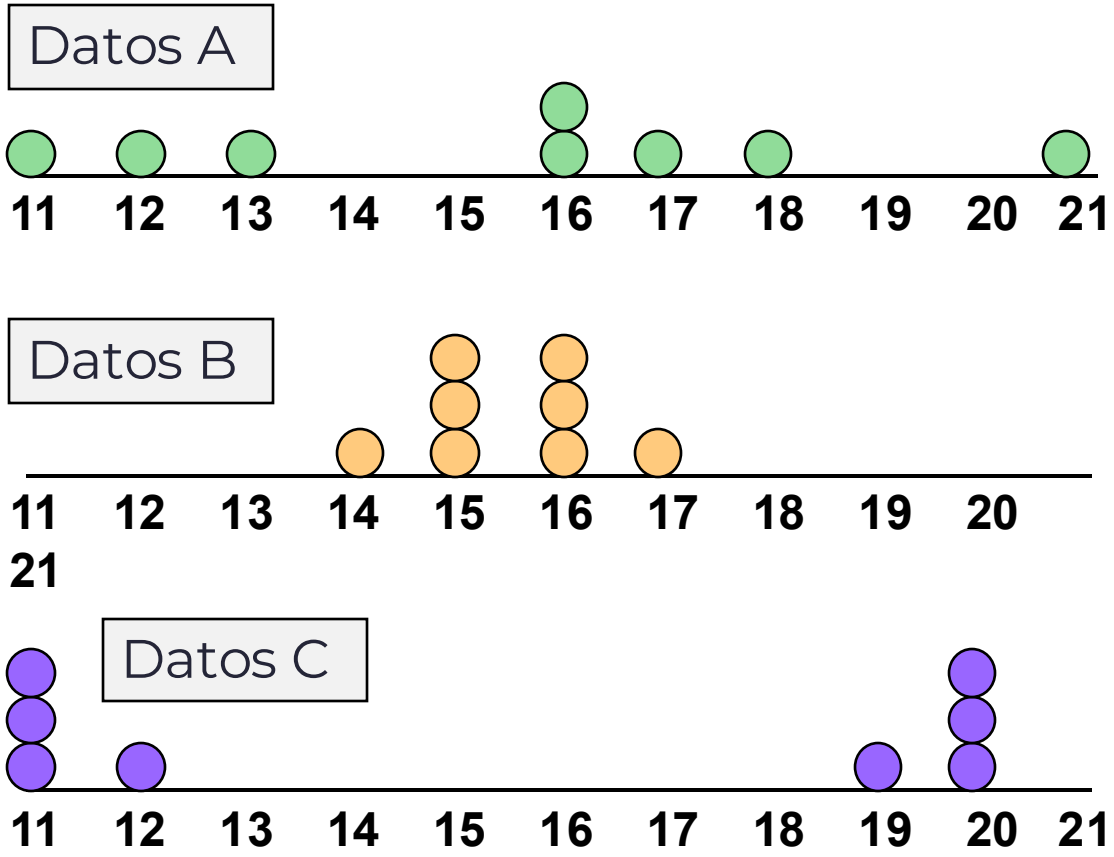
$$S = \sqrt{\frac{(10-\overline{X})^2+(12-\overline{X})^2+(14-\overline{X})^2+\cdots+(24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2+(12-16)^2+(14-16)^2+\cdots+(24-16)^2}{8-1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

Una medida de la desviación media de la media.

Medidas de Variación: Desviación Estándar



Media = 15.5
S = 3.338

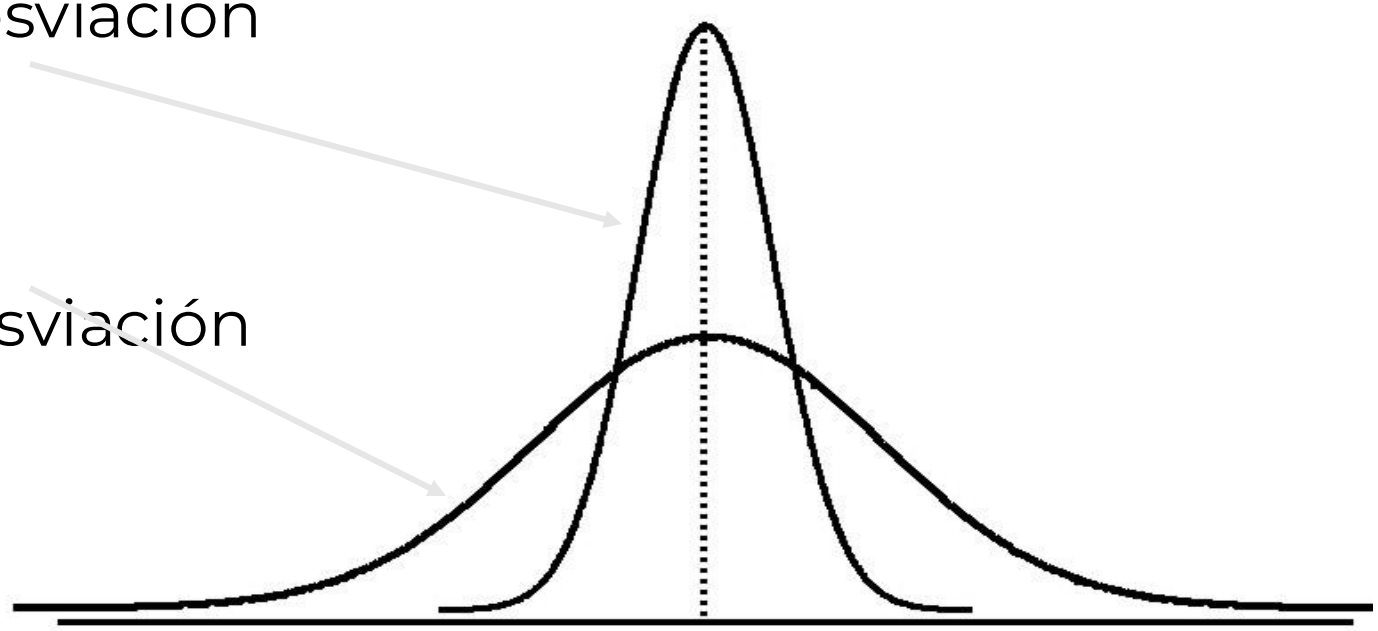
Media = 15.5
S = 0.926

Media = 15.5
S = 4.567

Medidas de Variación: Desviación Estándar

Menor desviación estándar

Mayor desviación estándar



Medidas de Variación: Resumen

- Entre más dispersos los datos, el rango, varianza y desviación estándar serán más grandes.
- Entre más concentrado, serán menores.
- Si todos los valores son los mismos (no hay variación), todas las medidas serían cero.
- Ninguna medida es negativa.
- Mejor utilizar la desviación estándar.

Medidas de Variación:
Coeficiente de Variación

En términos porcentuales (%).

- Muestra variación relativa a la media.
- Utilizado para comparar volatilidad de dos acciones distintas (finanzas)

$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

Medidas de Variación:
Coeficiente de Variación

- Acciones A:
  - Precio medio = $50.
  - Desviación estándar = $5.

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Acciones B:
  - Precio medio = $100.
  - Desviación estándar = $5.

$$CV_B = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Ambas acciones tiene la misma desviación estándar, pero para B, la variabilidad es menor relativo a la media.

Medidas de Variación:
Coeficiente de Variación

- Acciones A:
  - Precio medio = $50.
  - Desviación estándar = $5.

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Acciones C:
  - Precio medio = $8.
  - Desviación estándar = $2.

$$CV_C = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$2}{\$8} \cdot 100\% = 25\%$$

C tiene menor desviación estándar, pero mayor coeficiente de variación.

EJERCICIO

Utilizando los datos de concerts.csv :

Construya un cuadro de resumen estadístico de las variables numéricas (pueden enfocarse en net_sales, impressions ó spend por canal-tactico y variables macro):

1. Las medidas de tendencia central de los datos:

   Media, mediana y moda.

2. Las medidas de dispersión de los datos:

   Varianza, desviación estándar y rango.

3. El coeficiente de variación.

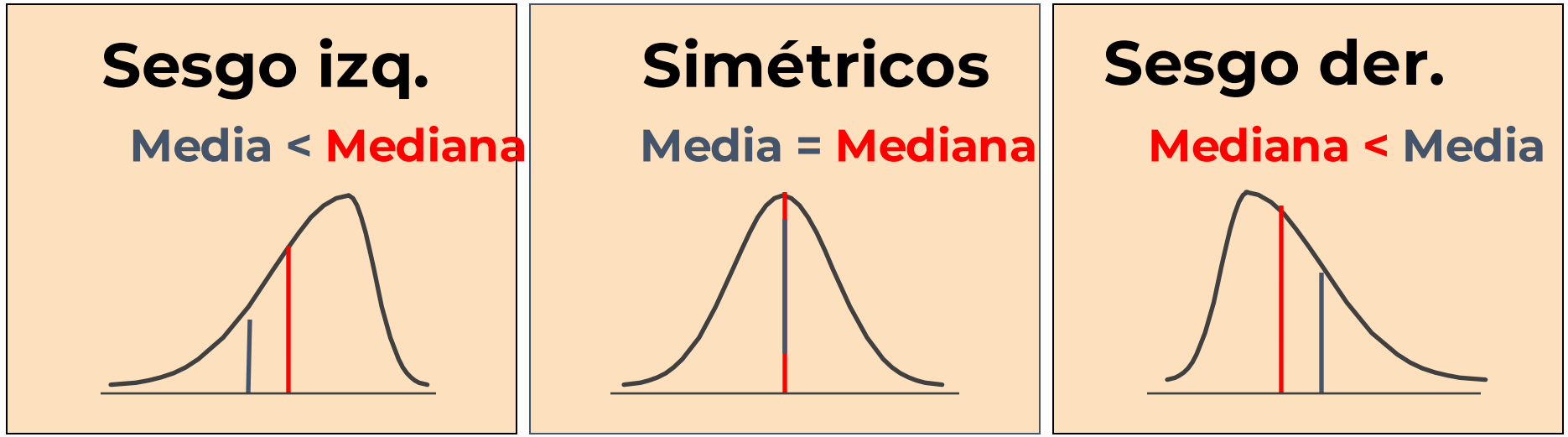
Pista: Puede usar los paquetes {gt} o flextable. Acá pueden ver varios paquetes de tablas: https://towardsdatascience.com/top-7-packages-for-making-beautiful-tables-in-r-7683d054e541

Medidas de Forma de la Distribución

- Describen cómo se distribuyen los datos.
  - Asimetría o (Skewness):
    - Mide qué tan asimétricamente están distribuidos los datos.
  - Curtosis:
    - Mide que tan “picuda” es la distribución.
    - No cubriremos este estadístico.

Forma de la Distribución:
Asimetría

- Mide la asimetría de los datos



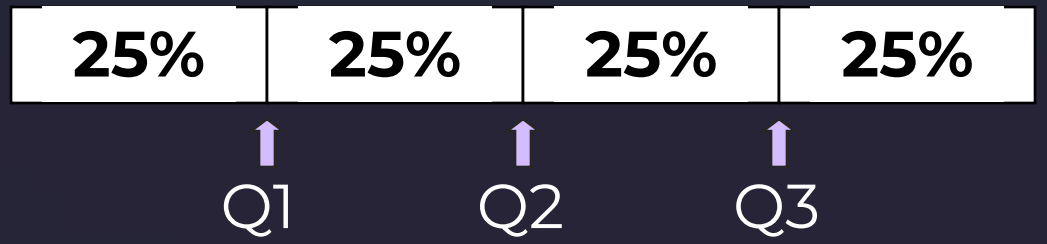
Asimetría	< 0	0	>0

Explorando los datos utilizando Cuartiles

- Se puede visualizar la distribución de los valores:
  - Calculando los cuartiles.
  - Haciendo el resumen de 5 estadísticos y,
  - Construyendo un diagrama de caja y brazos.

Cuartiles

- Se dividen los datos en cuatro segmentos, con el mismo número de observaciones/valores por segmento.



  - El primer cuartil, Q_1, es el valor para el cual el 25% de los valores son más pequeños y el 75% más grandes.
  - Q_2 es la mediana (50% son menores y 50% mayores).
  - Solo el 25% de los valores son más grandes que el tercer cuartil.

Cuartiles

Se encuentra la posición en los datos:

$Q_1 = (n+1)/4$

$Q_2 = (n+1)/2$

$Q_3 = 3(n+1)/4$

donde **n** es el número de valores.

Cuartiles

Datos ordenados: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 está en la (9+1)/4 = 2.5 posición.

Tomamos la el promedio entre la posición 2 y la posición 3**.**

Q_1 = 12.5

Cuartiles

Datos ordenados: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1: (9+1)/4 = 2.5 posición,

Q_1 = (12+13)/2 = 12.5.

Q_2: (9+1)/2 = 5 posición

Q_2 = mediana = 16.

Q_3: 3(9+1)/4 = 7.5 posición,

Q_3 = (18+21)/2 = 19.5.

Cuartiles:
El Rango Intercuartílico (RIC)

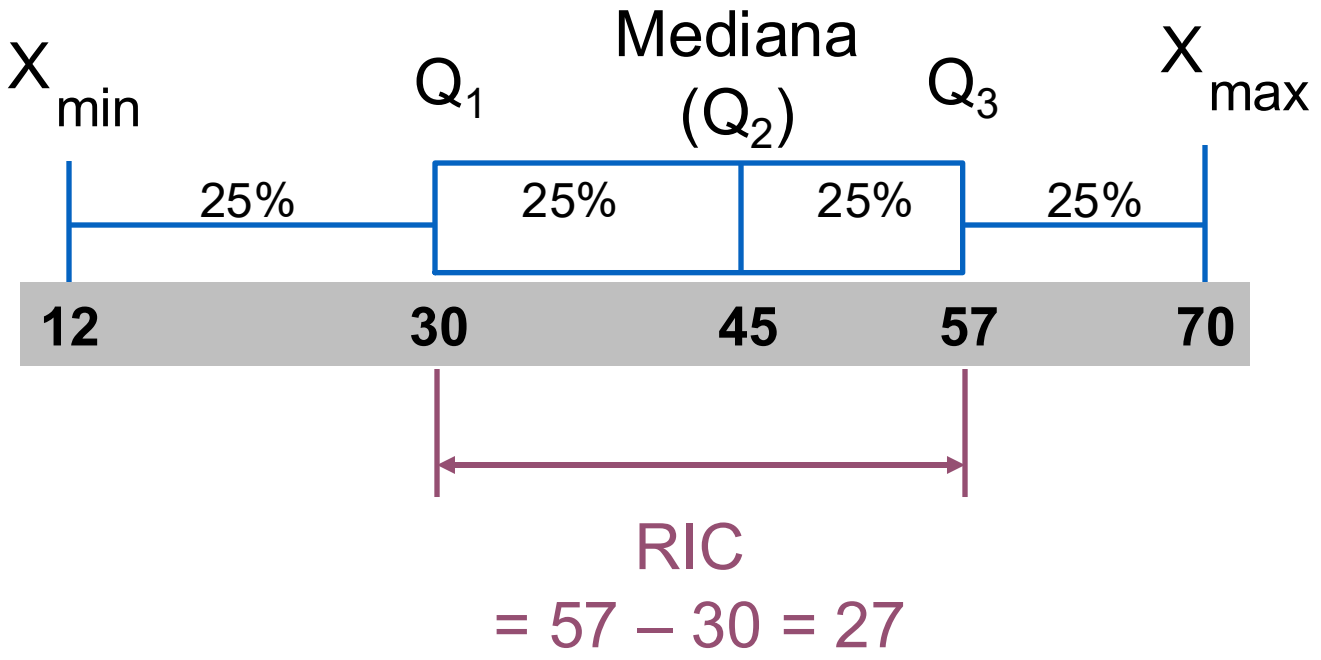
- RIC = $Q_3 - Q_1$ y mide la variabilidad del 50% de los datos que están en medio.

- Una medida de variabilidad no afectada por outliers.

- Las medidas como Q_1, Q_3, y RIC que no son influenciadas por outliers se llaman medidas resistentes.

Rango Intercuartílico

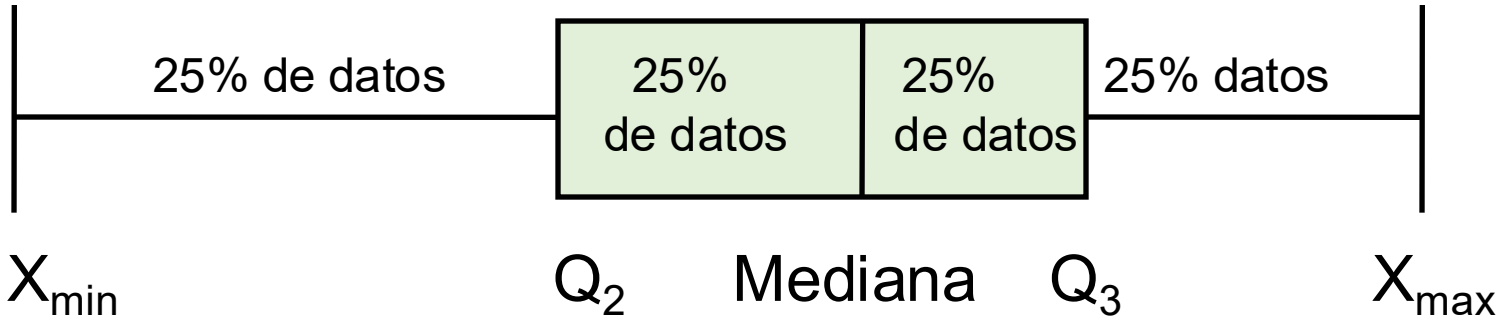
Ejemplo:



Resumen de 5 estadísticos y el Diagrama de Caja y Brazos

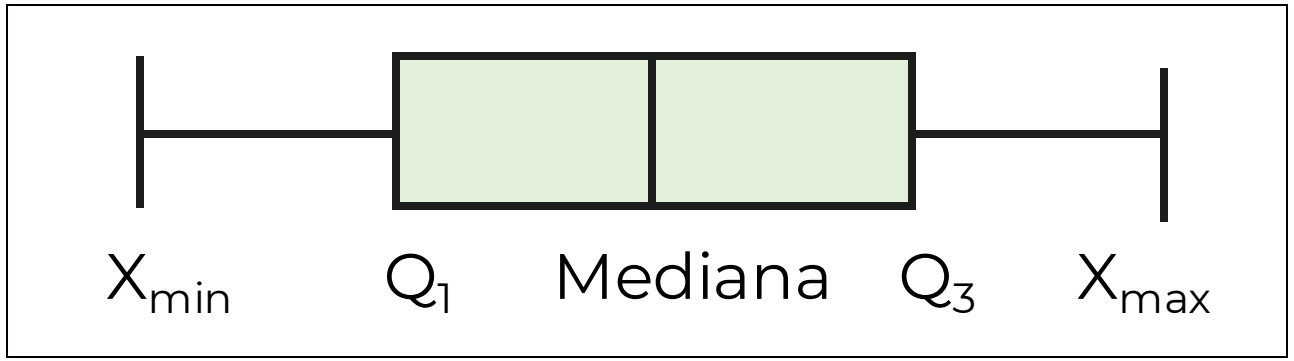
X_{min} -- Q_1 -- Mediana -- Q_3 -- X_{max}

Ejemplo:



Resumen de 5 estadísticos y el Diagrama de Caja y Brazos

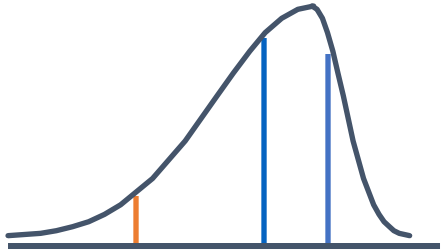
- Si los datos están distribuidos simétricamente alrededor de la mediana, entonces la caja está centrada:



- Se puede mostrar horizontal o verticalmente

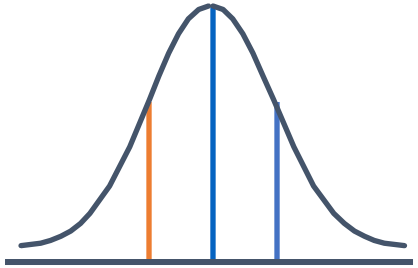
Forma de la Distribución y el Diagrama de Caja y Brazos

Sesgo izq.

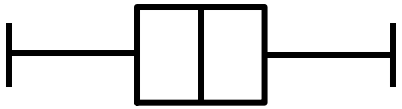


Q_1 Q_2 Q_3

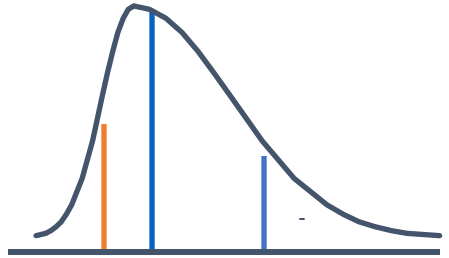
Simétrico



Q_1 Q_2 Q_3



Sesgo der.



Q_1 Q_2 Q_3

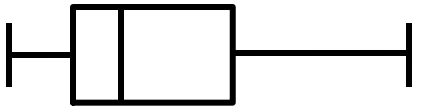
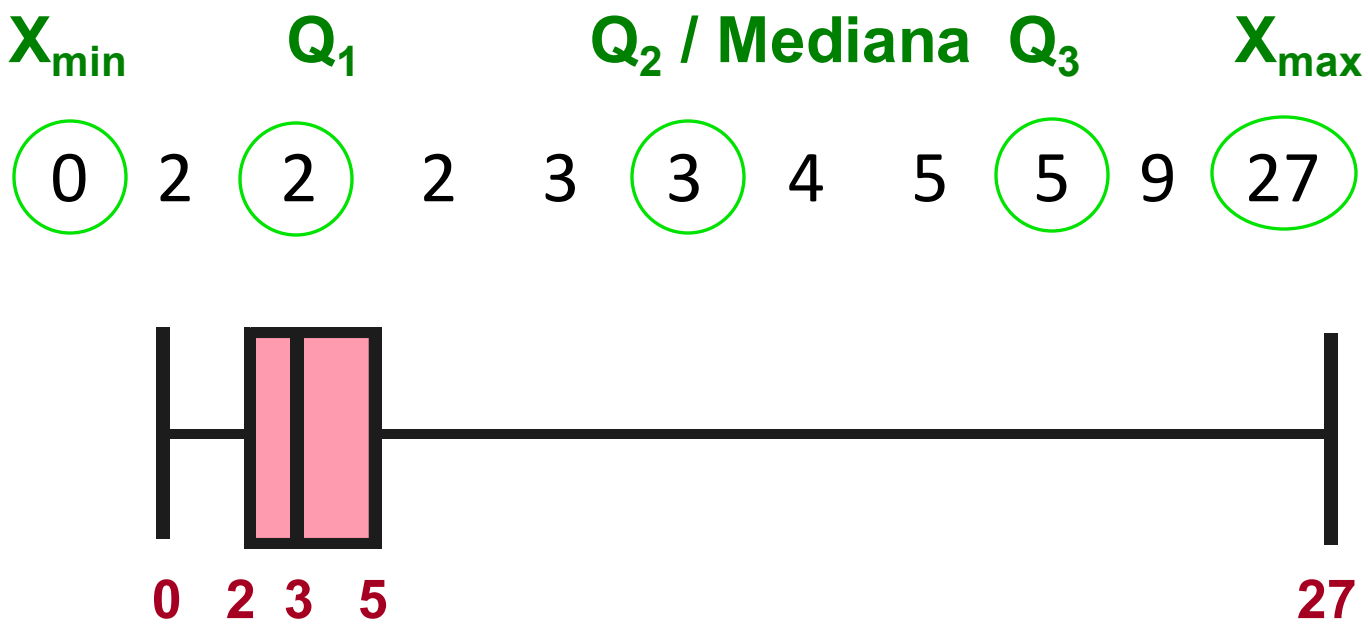


Diagrama de Caja y Brazos: Ejemplo



- Sesgo o Asimetría a la derecha.

Medidas Descriptivas para una Población

- Los estadísticos descriptivos discutidos anteriormente **describen una muestra**, pero **no a la población.**

- Las medidas que describen a una población se llaman **parámetros.** Utilizamos letras griegas para referirnos a ellos.

- Los parámetros más importantes son: la media, mediana, y desviación estándar.

Notación: estadísticos muestrales vs. parámetros poblacionales

Medida	Parámetro poblacional	Estadístico muestral
Media	μ	$\overline{X}$
Varianza	σ^2	S^2
Desviación estándar	σ	S

Medidas de la Relación entre Dos Variables

Como vimos, los gráficos de dispersión nos ayudan a visualizar y examinar la relación entre dos variables numéricas.

Ahora vamos a discutir dos medidas cuantitativas de dichas relación:

- La Covarianza
- El Coeficiente de Correlación.

Covarianza

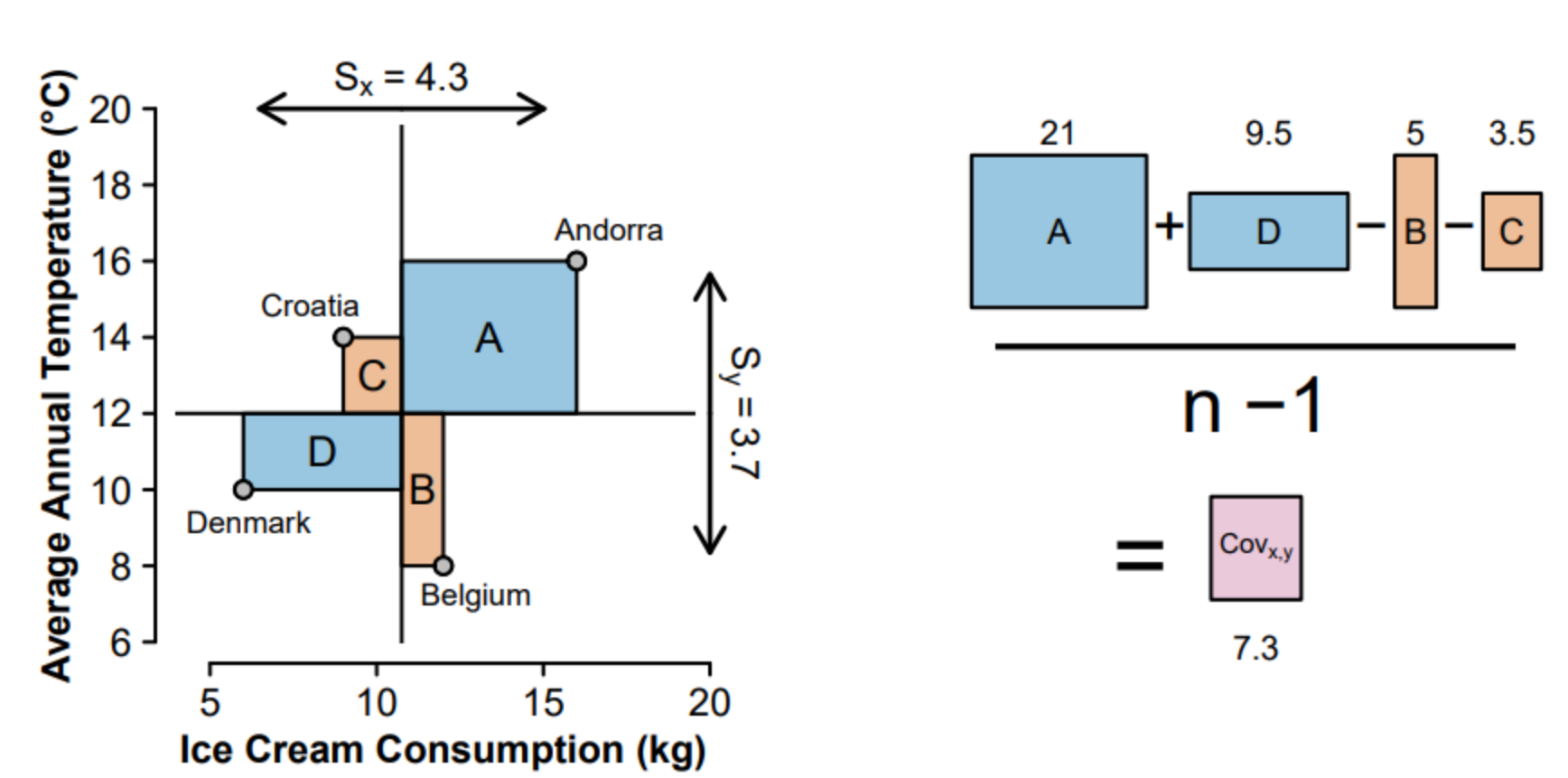
- La covarianza mide qué tan fuerte es la relación (lineal) de **dos variables numéricas** (X & Y).

- La covarianza muestral:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

- Solamente mide la “fuerza” de la relación.
- No habla sobre causalidad.

Covarianza


$S_x = 4.3$
Average Annual Temperature (°C)
20
18
16
14
12
10
8
6
Andorra
Croatia
Denmark
Belgium
A
B
C
D
$S_y = 3.7$
5
10
15
20
Ice Cream Consumption (kg)
21
9.5
5
3.5
A
+
D
−
B
−
C
n −1
=
$Cov_{x,y}$
7.3

Interpretando la Covarianza

- **Covarianza** entre dos variables:

cov(X,Y) > 0 ⟶ X and Y se mueven en la misma dirección.

cov(X,Y) < 0 ⟶ X and Y se mueven en dirección opuesta.

cov(X,Y) = 0 ⟶ X and Y son independientes.

- La covarianza tiene un gran defecto:
  - No se puede determinar la fuerza relativa de la relación solamente mirando la covarianza.

Coeficiente de Correlación

- Mide la fuerza relativa de la relación lineal entre dos variables.
- Coeficiente de correlación muestral:

Donde:

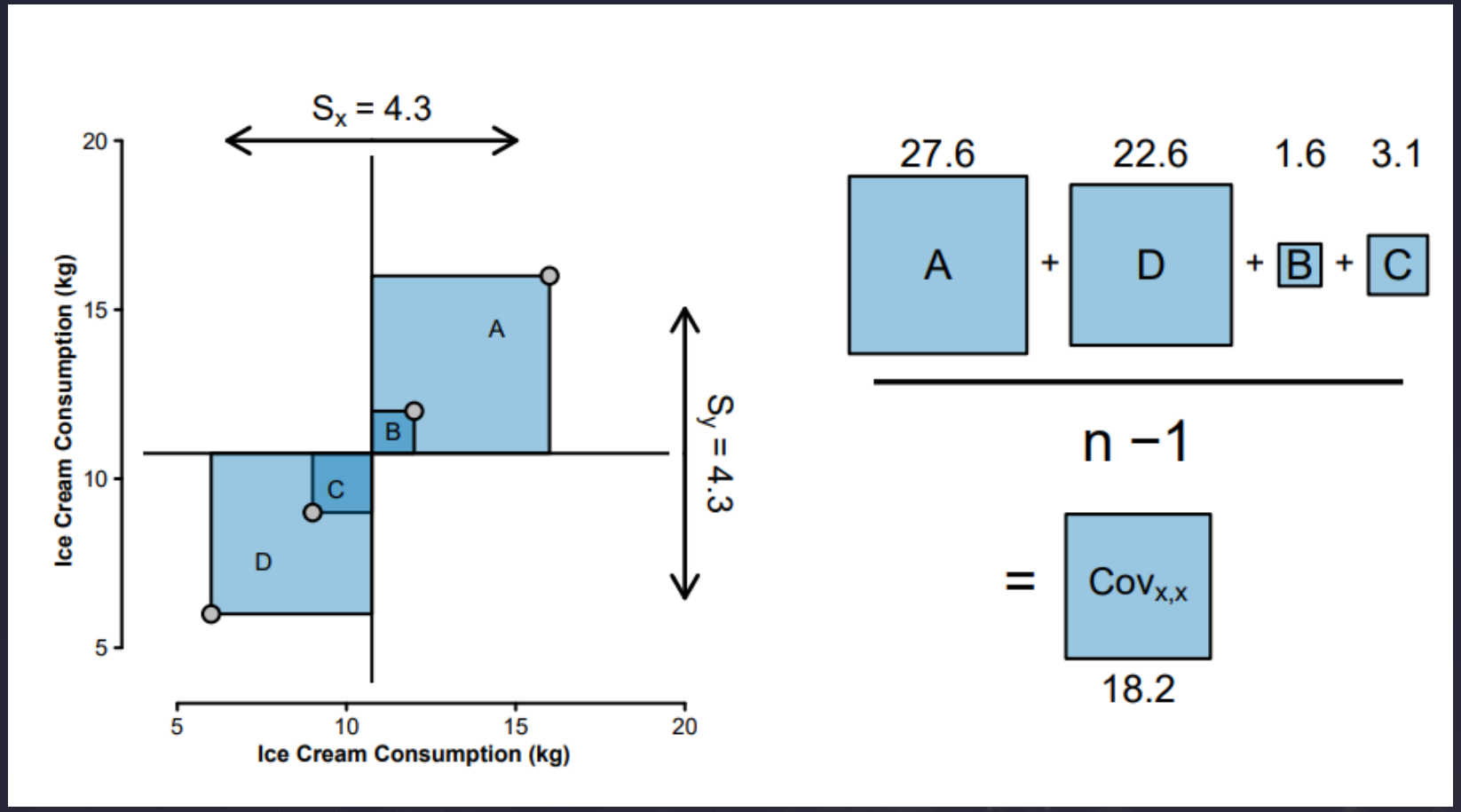
$$r = \frac{\operatorname{cov}(X,Y)}{S_X S_Y}$$

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

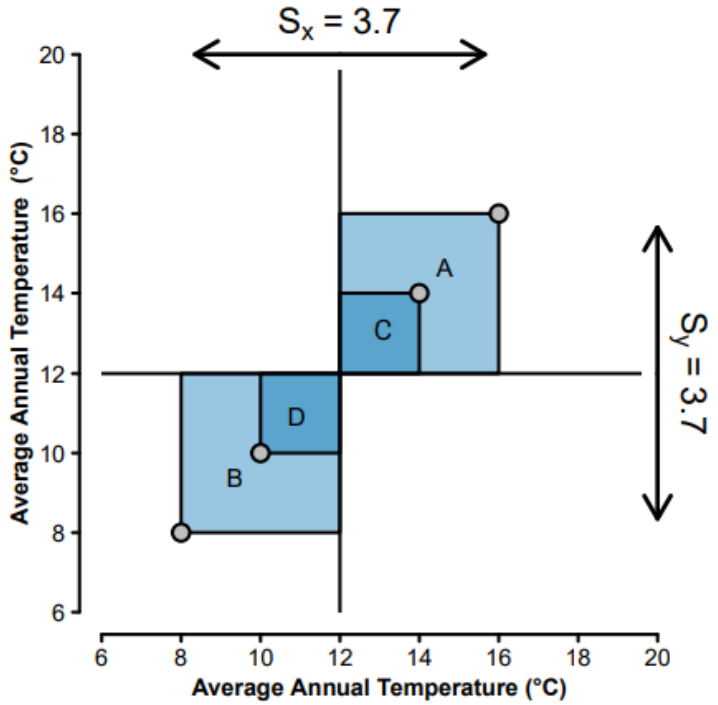
$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

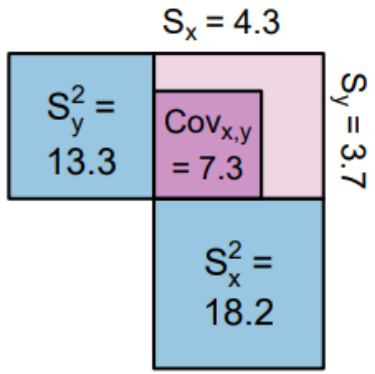
Coeficiente de Correlación



Coeficiente de Correlación



$$\frac{16 + 4 + 16 + 4}{n - 1} = \frac{A + D + B + C}{n - 1} = Cov_{y,y} = 13.3$$

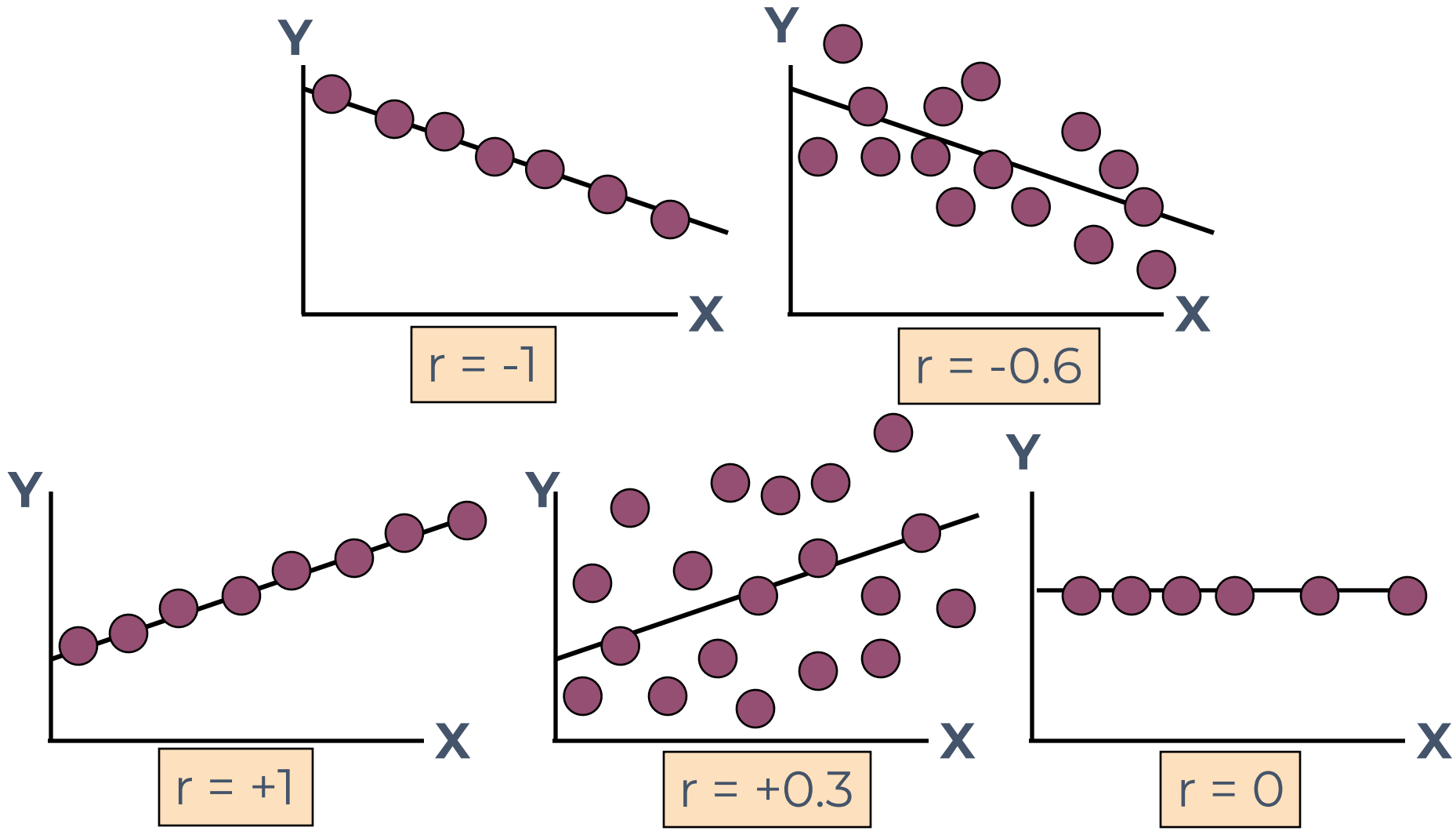


$$r_{x,y} = \frac{Cov_{x,y}}{S_x \times S_y} = \frac{7.3}{4.3 \times 3.7} = 0.47$$

Coeficiente de Correlación

- Al parámetro **poblacional**, el coeficiente de correlación en la población lo llamamos **ρ**.
- Al **muestral** lo llamamos **r**.
- ρ y r tienen las siguientes propiedades:
  - **No están medidos en unidades**.
  - Entre **–1 y 1**.
  - Más cerca a **–1**, más fuerte la **relación lineal negativa.**
  - Más cerca a **1**, más fuerte la **relación lineal positiva**.
  - Más cerca a **0**, más **débil la relación lineal.**

Gráficas de Dispersión con sus respectivos Coeficientes de Correlación



EJERCICIO

Utilizando los datos de concerts.csv:

1. Construya una matriz de correlación entre las variables de spend y net_sales.

2. Grafique la relación entre net_sales y unemployment.

3. ¿Podría unemployment ser un predictor de net_sales? Justifique su respuesta.

A Fine is a Price

Uri Gneezy y Aldo Rustichini (JELS, 2000)

Introducción

- Guarderías en Haifa con moderado problema: algunos padres llegan tarde por sus hijos.
- Una posible solución: introducir una multa por llegar tarde
- Idea: las personas funcionan con incentivos (monetarios) y este es un incentivo para llegar a tiempo.

Introducción

- Estudio elaborado en 10 guarderías en Haifa, por 20 semanas.
- Pequeñas guarderías para máximo 35 niños.
- Guarderías cierran a las 4pm. Un empleado de la guardería se queda cuidando de los niños.
- No hay multas por llegar tarde.

Experimento

- Por las primeras 4 semanas, los investigadores recolectan información sobre el número de padres que llegan tarde cada semana..

- Iniciando la semana 5, una multa en 6 de 10 centros.
    - > 10 minutos.
    - Por niño.
    - 10 NIS ($2.75 of 1998).

Experimento

- A lxs niñerxs se les informa de la multa, pero no del estudio.
- Padres informados en tablero de anuncios, frecuentemente leído.
- Se elimina la multa al final de la semana 17.

Resultados

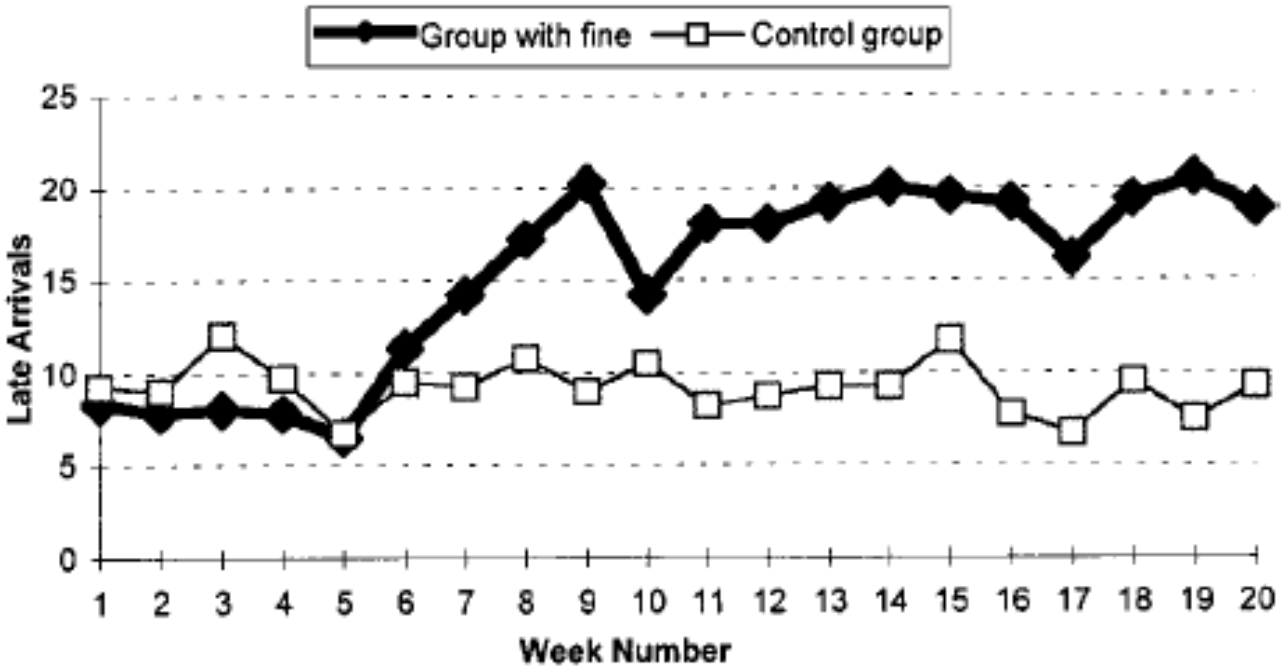


FIGURE 1.—Average number of late-coming parents, per week

Resultados

TABLE 2

AVERAGE NUMBER OF LATE-COMING PARENTS, ACCORDING TO FOUR PERIODS OF THE STUDY

Center	No. of Children	Weeks 1–4	Weeks 5–8	Weeks 5–16	Weeks 17–20
1	37	7.25	9.5	12.5	15.25
2	35	5.25	9	12.2	13.25
3	35	8.5	10.25	16.8	22
4	34	9	15	19.1	20.25
5	33	11.75	20	24.6	29.5
6	28	6.25	10	13.1	12
7	35	8.75	8	7.2	6.75
8	34	13.25	10.5	10.9	9.25
9	34	4.75	5.5	5.5	4.75
10	32	13.25	12.25	13.1	12.25

NOTE.—The four periods of the study are as follows: before the fine (weeks 1–4), the first 4 weeks with the fine (weeks 5–8), the entire period with the fine (weeks 5–16), and the postfine period (weeks 17–2).

Conclusiones

- La pequeña multa aumenta el número de niños que son recogidos tarde. La multa no funciona.
- Una vez se elimina la multa, no hay un regreso a la situación inicial: los padres en las guarderías con multas siguen llegando tarde.
- Explicación: la multa es un incentivo externo que diminuye la motivación interna.