

Diplomado en Ciencia de Datos

Primer Semestre de 2019

Docente: Dora Suárez

Estadística

Descriptiva

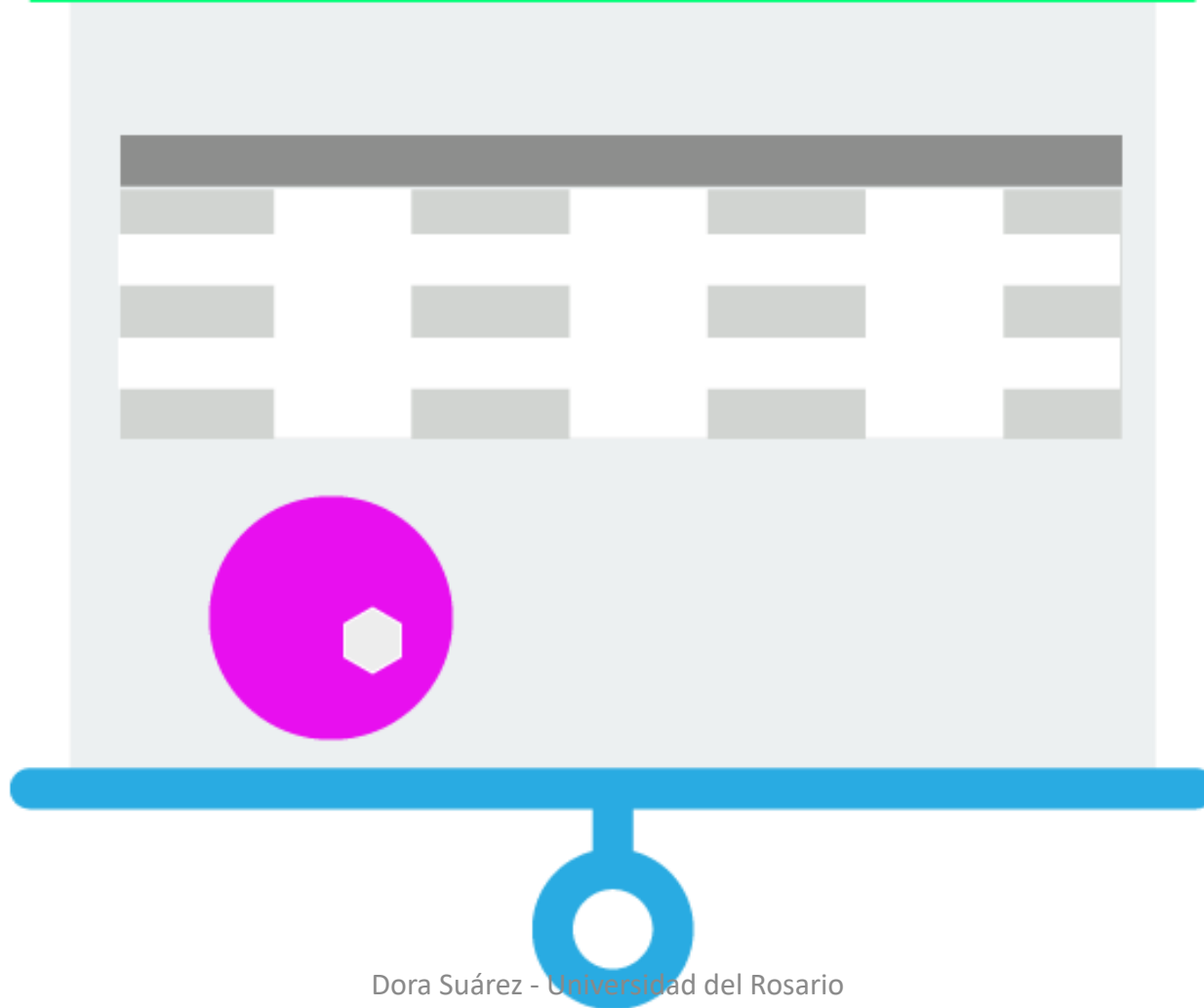


Contenido

- Conceptos Básicos y manipulación de datos
- Tipos de datos
- Resúmenes de Variables cualitativas
- Resúmenes de variables cuantitativas

Conceptos

Básicos



Librerías

dplyr: Gramática de manipulación de datos, que proporciona un conjunto consistente de verbos que lo ayudan a resolver los desafíos mas comunes de manipulación de datos

tidyr: Creacion de datos ordenados.

- Cada variable es una columna
- Cada observación es una fila
- Cada valor es una celda

Operaciones de una table - dplyr

- **filter**: deja las filas que cumplan con cierto criterio
- **select**: selecciona columnas por nombre
- **arrange**: reordena las filas
- **mutate**: agrega nuevas
- **summarize**: reduce las variables a valores

Estructura:

Primer parámetro es data frame

Siguientes parámetros indican qué hacer con los datos

Siempre retorna data frame

Nunca modifica el data frame de entrada

Operaciones de una tabla

```
df <- data.frame(  
  color = c("blue", "black", "blue", "blue", "black"),  
  value = 1:5)
```

df

color	value
blue	1
black	2
blue	3
blue	4
black	5

→

color	value
blue	1
blue	3
blue	4

```
filter(df, color == "blue")
```

Operaciones de una tabla

```
df <- data.frame(  
  color = c("blue", "black", "blue", "blue", "black"),  
  value = 1:5)
```

df

color	value
blue	1
black	2
blue	3
blue	4
black	5







→

color	value
blue	1
blue	4

```
filter(df, value %in% c(1, 4))
```


Operaciones de una tabla

```
df <- data.frame(  
  color = c("blue", "black", "blue", "blue", "black"),  
  value = 1:5)
```

	a
	b
	a b
	a & b
	a & !b
	xor(a, b)

x > 1

x >= 1

x < 1

x <= 1

x != 1

x == 1

x %in% ("a", "b")

Operaciones de una tabla

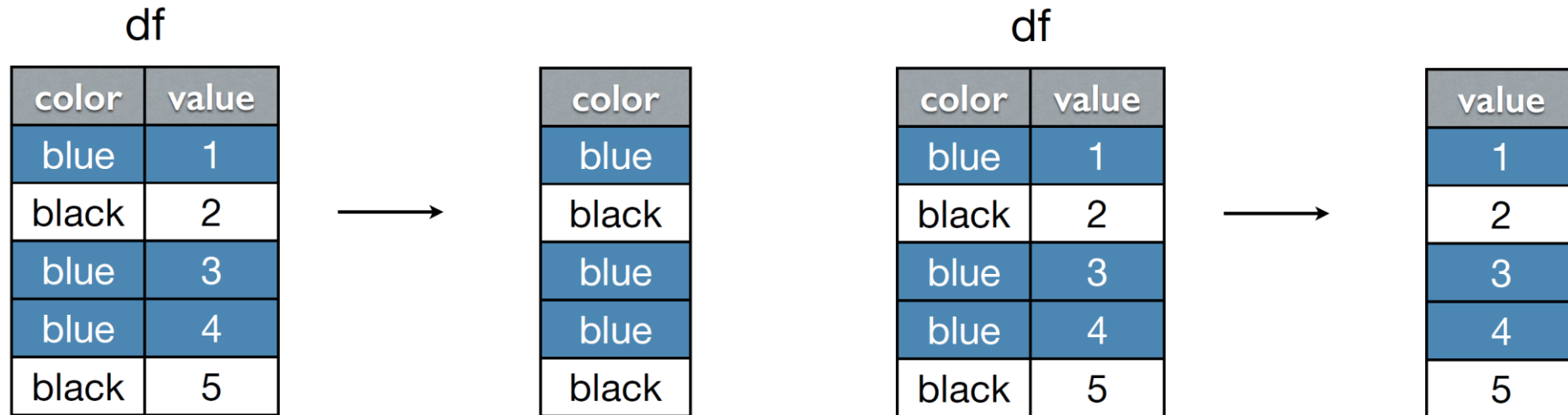
```
df <- data.frame(  
  color = c("blue", "black", "blue", "blue", "black"),  
  value = 1:5)  
  
# Just prints out results  
filter(flights, dest %in% c("IAH", "HOU"))  
# The original is unchanged:  
flights  
  
# To create a new variable use <-  
houston <- filter(flights, dest %in% c("IAH", "HOU"))  
houston  
  
# BE CAREFUL!  
flights <- filter(flights, dest %in% c("IAH", "HOU"))
```

Operaciones de una tabla

Encuentre todos los hurtos que ocurrieron

- En el departamento de CUNDINAMARCA
- En el municipio de DUITAMA
- Donde la víctima era menor de edad
- Donde la víctima tenía entre 35 y 50 años
- Después de las 5 de la tarde

Operaciones de una tabla



`select(df, color)`

`select(df, -color)`

Operaciones de una tabla

- Revise la ayuda de `select()`
- Escriba tres formas diferentes de seleccionar las variables `MOVIL AGRESOR` y `MOVIL VICTIMA`

Operaciones de una tabla

- Calcule el número de hurtos por arma empleada

OPERACIONES DE RESUMEN

- `min(x)`, `median(x)`, `max(x)`,
`quantile(x, p)`
- `n()`, `n_distinct(x)`, `sum(x)`, `mean(x)`
- `sum(x > 10)`, `mean(x > 10)`
- `sd(x)`, `var(x)`, `IQR(x)`, `mad(x)`

Operaciones con dos tablas

- Joins

A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA

left_join(x, y, by = NULL,
copy=FALSE, suffix=c(".x",".y"),...)
Join matching values from y to x.

A	B	C	D
a	t	1	3
b	u	2	2
d	w	NA	1

right_join(x, y, by = NULL, copy =
FALSE, suffix=c(".x",".y"),...)
Join matching values from x to y.

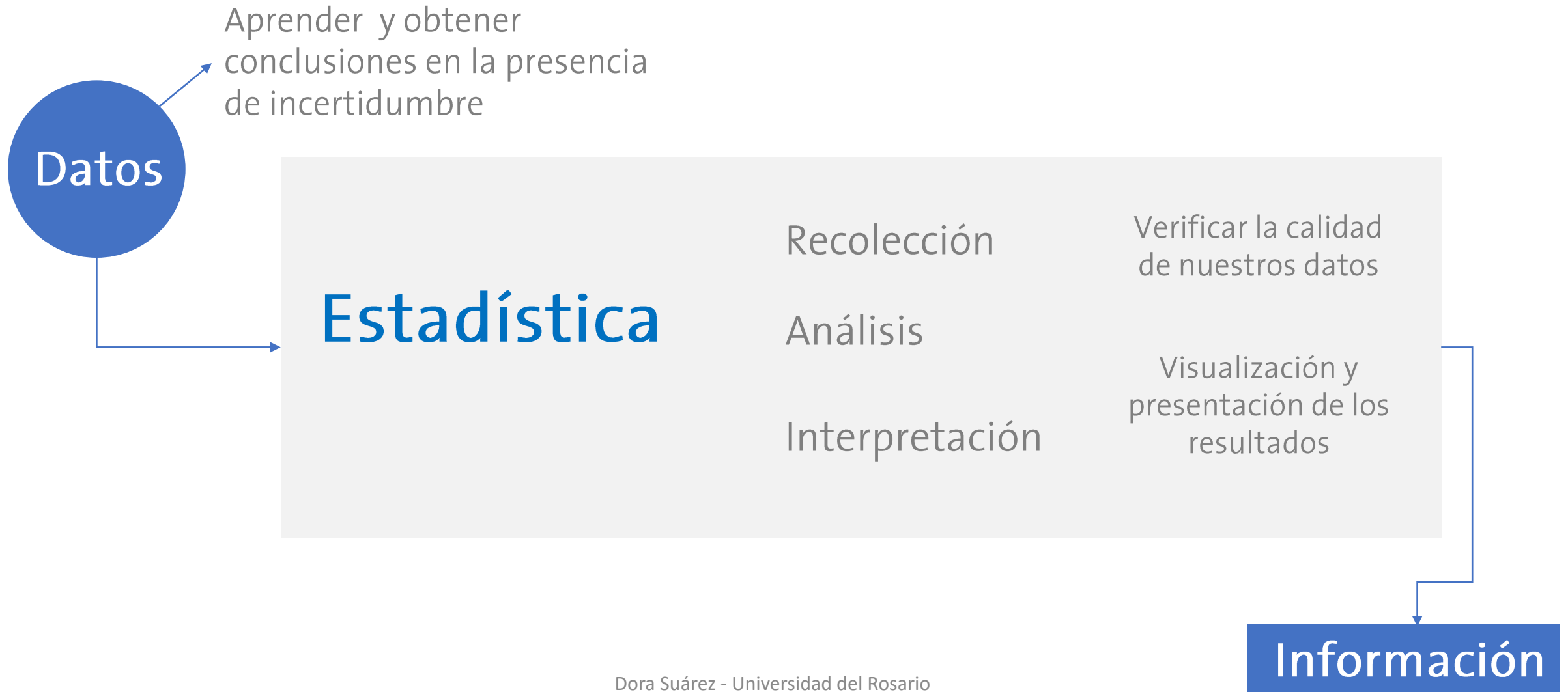
A	B	C	D
a	t	1	3
b	u	2	2

inner_join(x, y, by = NULL, copy =
FALSE, suffix=c(".x",".y"),...)
Join data. Retain only rows with
matches.

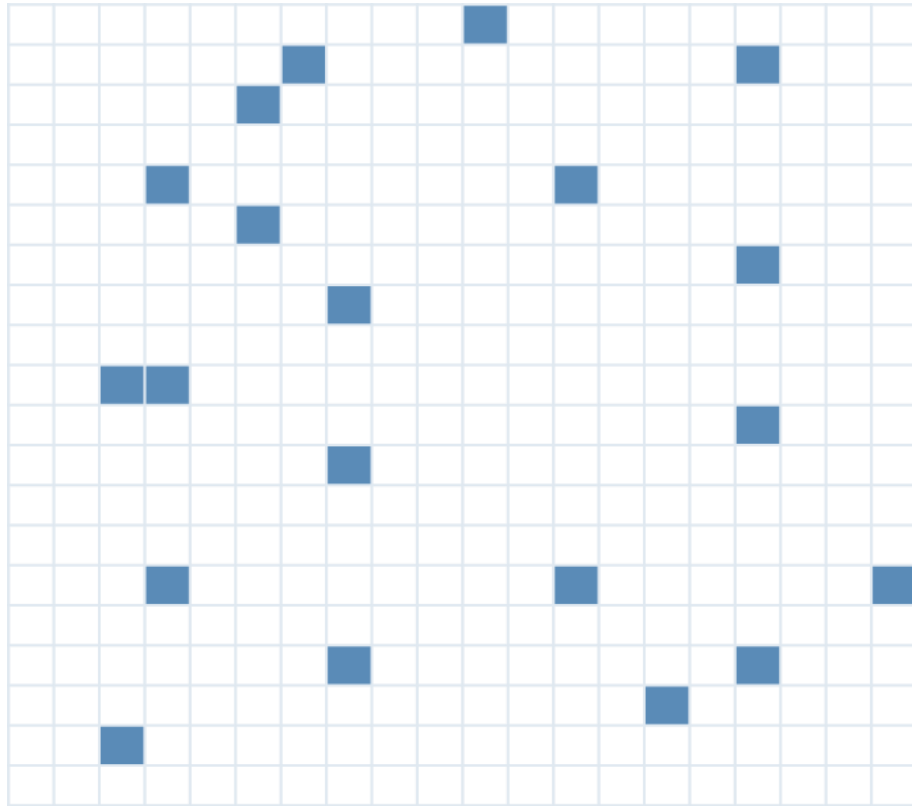
A	B	C	D
a	t	1	3
b	u	2	2
c	v	3	NA
d	w	NA	1

full_join(x, y, by = NULL,
copy=FALSE, suffix=c(".x",".y"),...)
Join data. Retain all values, all rows.

De los datos a la información

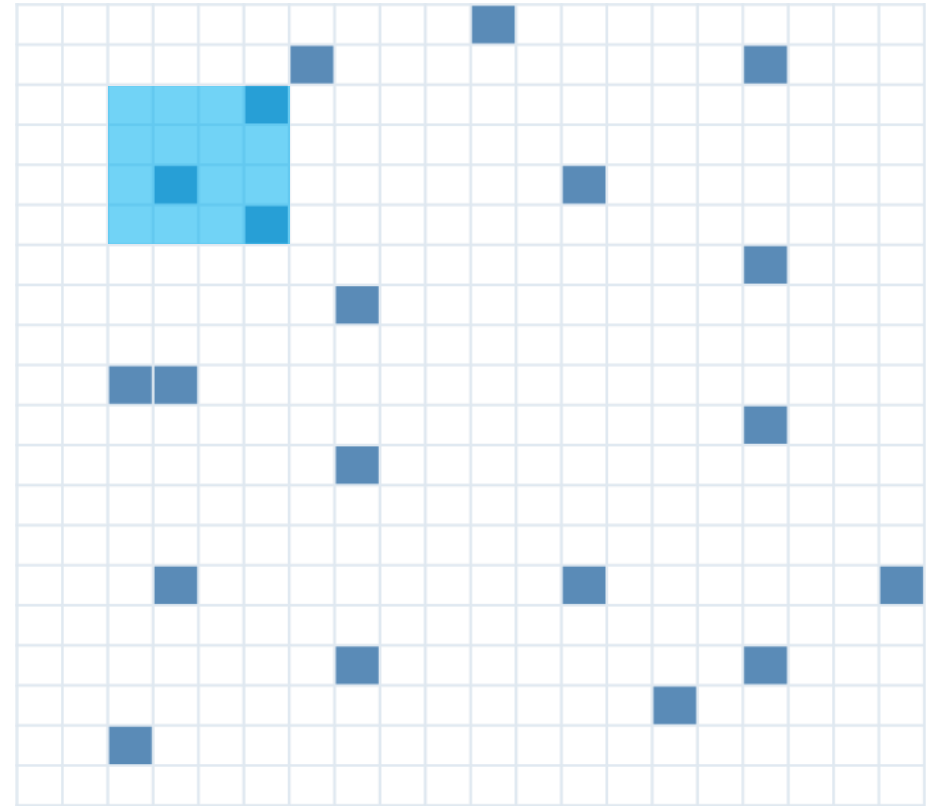


POBLACIÓN $N = 400$ elementos
20 defectuosos



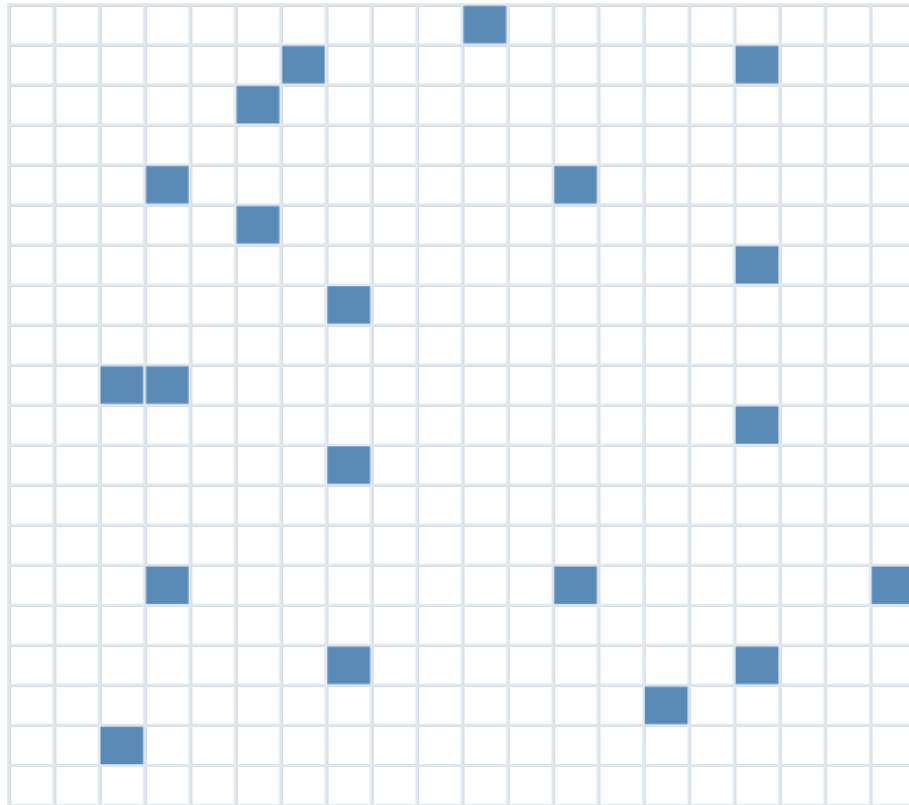
$$\text{Porcentaje de defectos} = \frac{20}{400} (100) = 5\%$$

MUESTRA $n = 16$ elementos



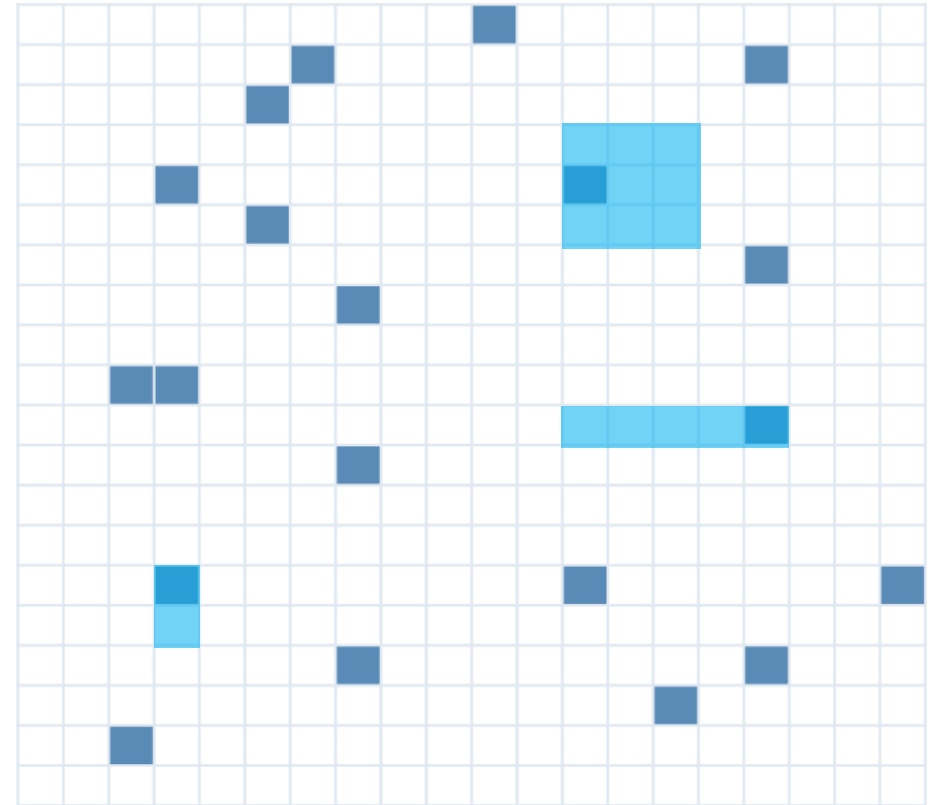
Porcentaje de defectos = ?

POBLACIÓN $N = 400$ elementos
20 defectuosos



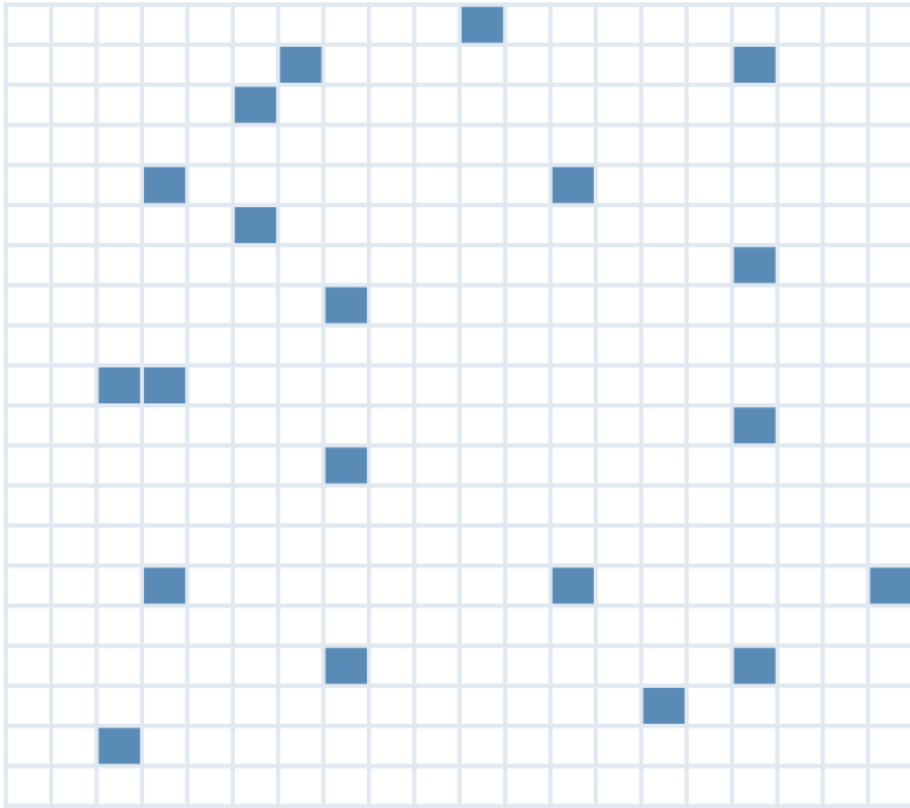
$$\text{Porcentaje de defectos} = \frac{20}{400} (100) = 5\%$$

MUESTRA $n = 16$ elementos



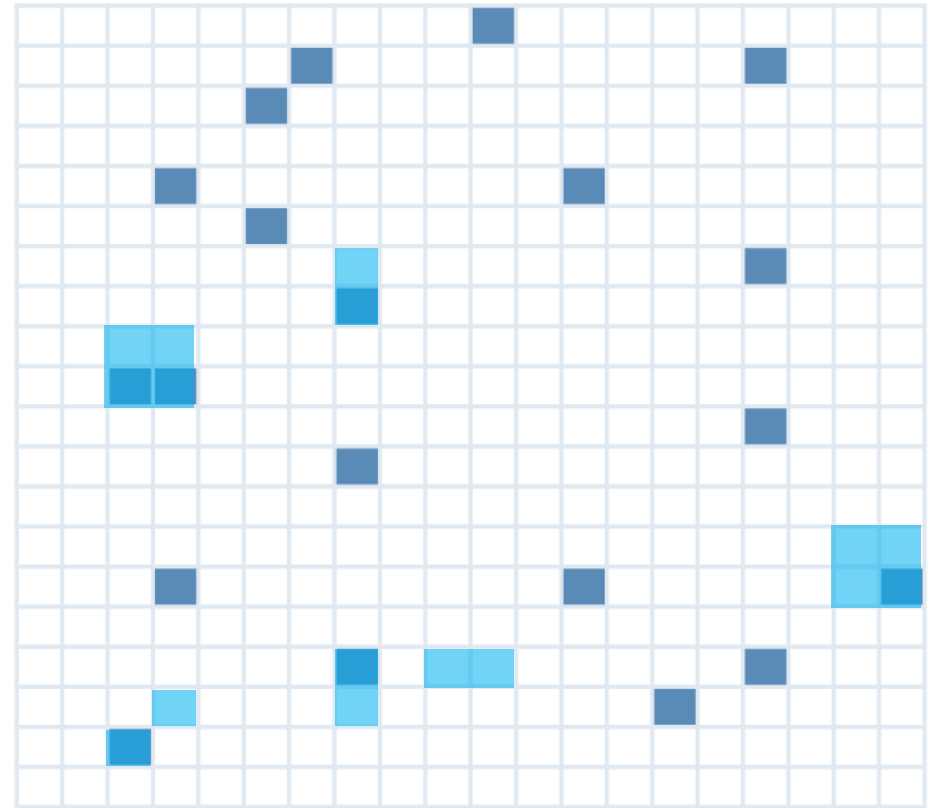
Porcentaje de defectos = ?

POBLACIÓN N = 400 elementos
20 defectuosos

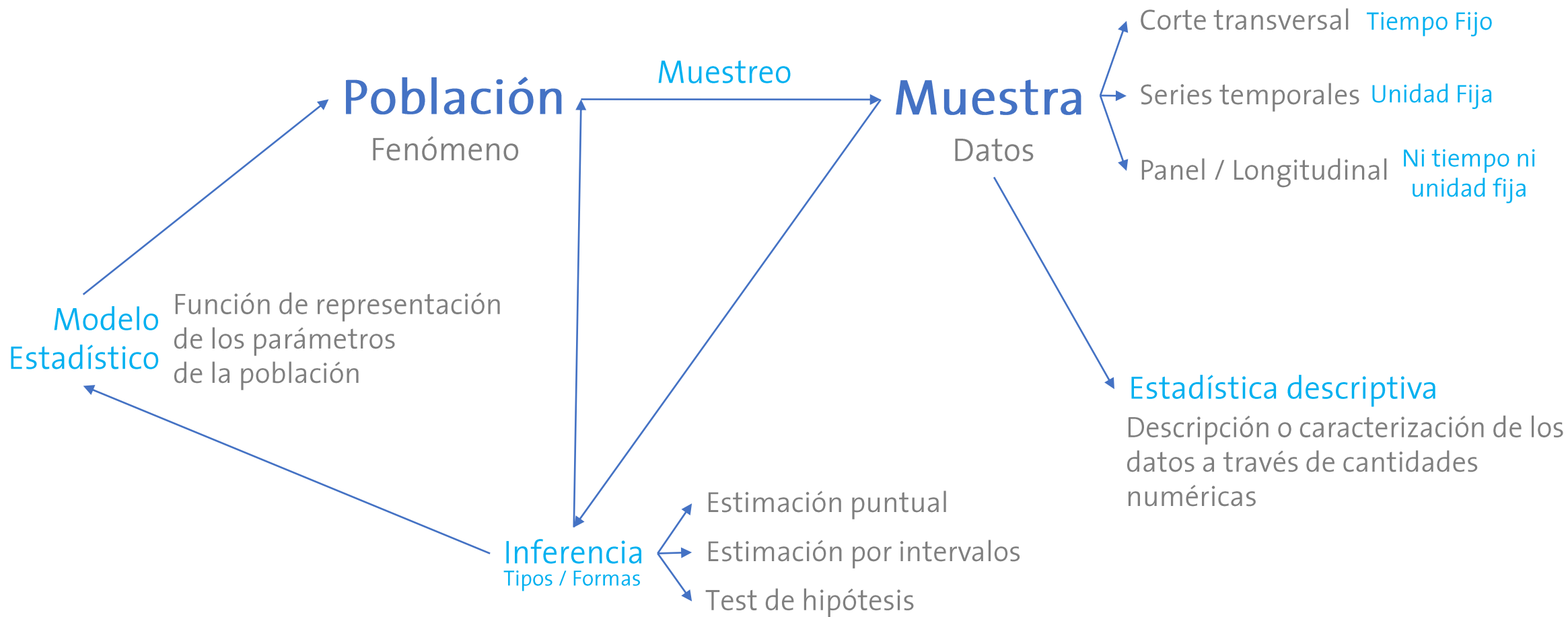


$$\text{Porcentaje de defectos} = \frac{20}{400} (100) = 5\%$$

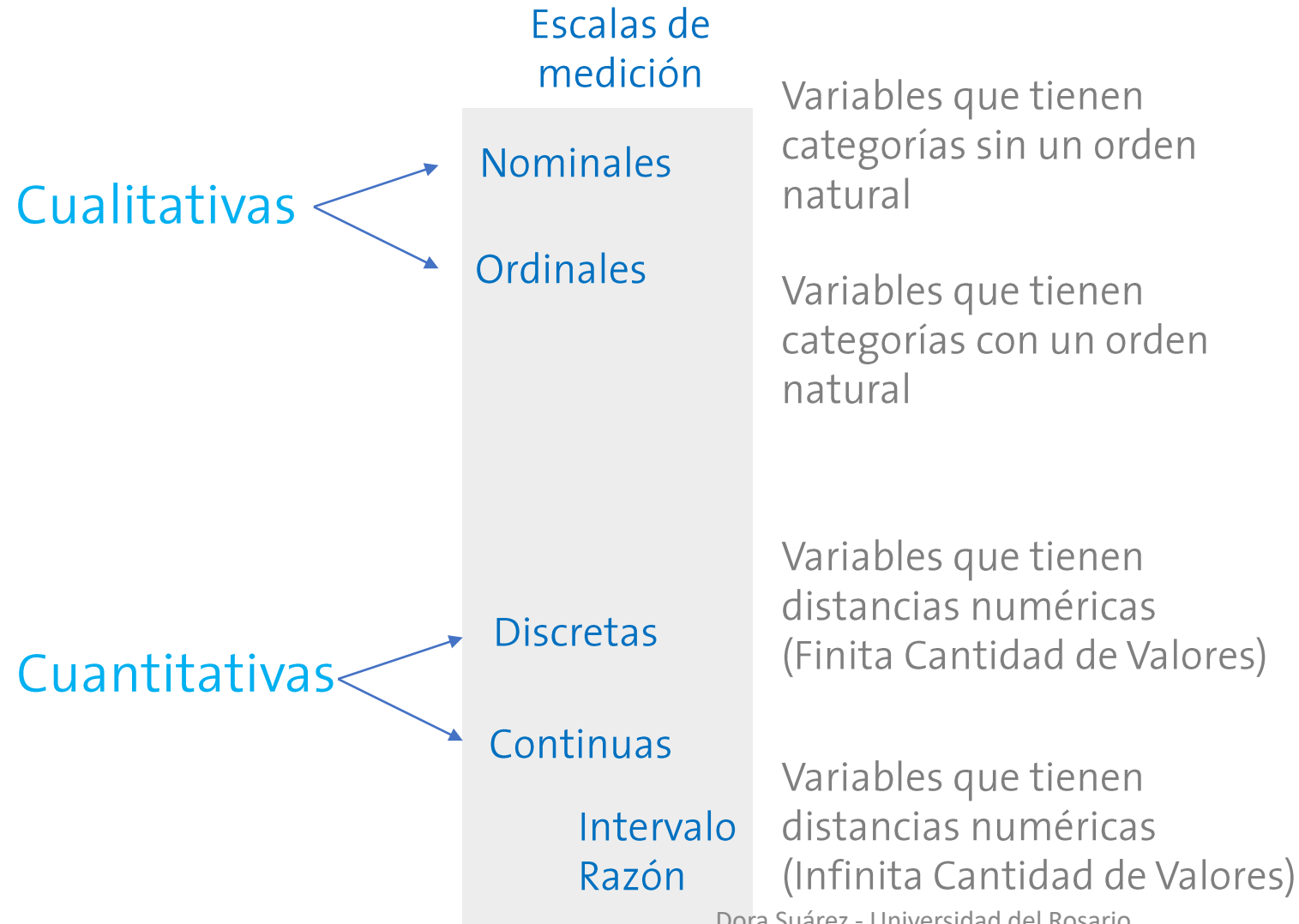
MUESTRA n = 16 elementos



Porcentaje de defectos = ?



Tipos de variables



En R	
character	as.character
logical	as.logical
factor	as.factor
integer	as.integer
numeric	as.numeric

Falso o verdadero

La probabilidad de seleccionar una muestra aleatoria de una población que no tenga características similares a la de la población es alta

El estrato de las personas es una variable aleatoria cuantitativa discreta

Para poder ir de la muestra a la población se utiliza la estadística descriptiva

La única forma de poder llegar de la población a la muestra es a través de la inferencia estadística

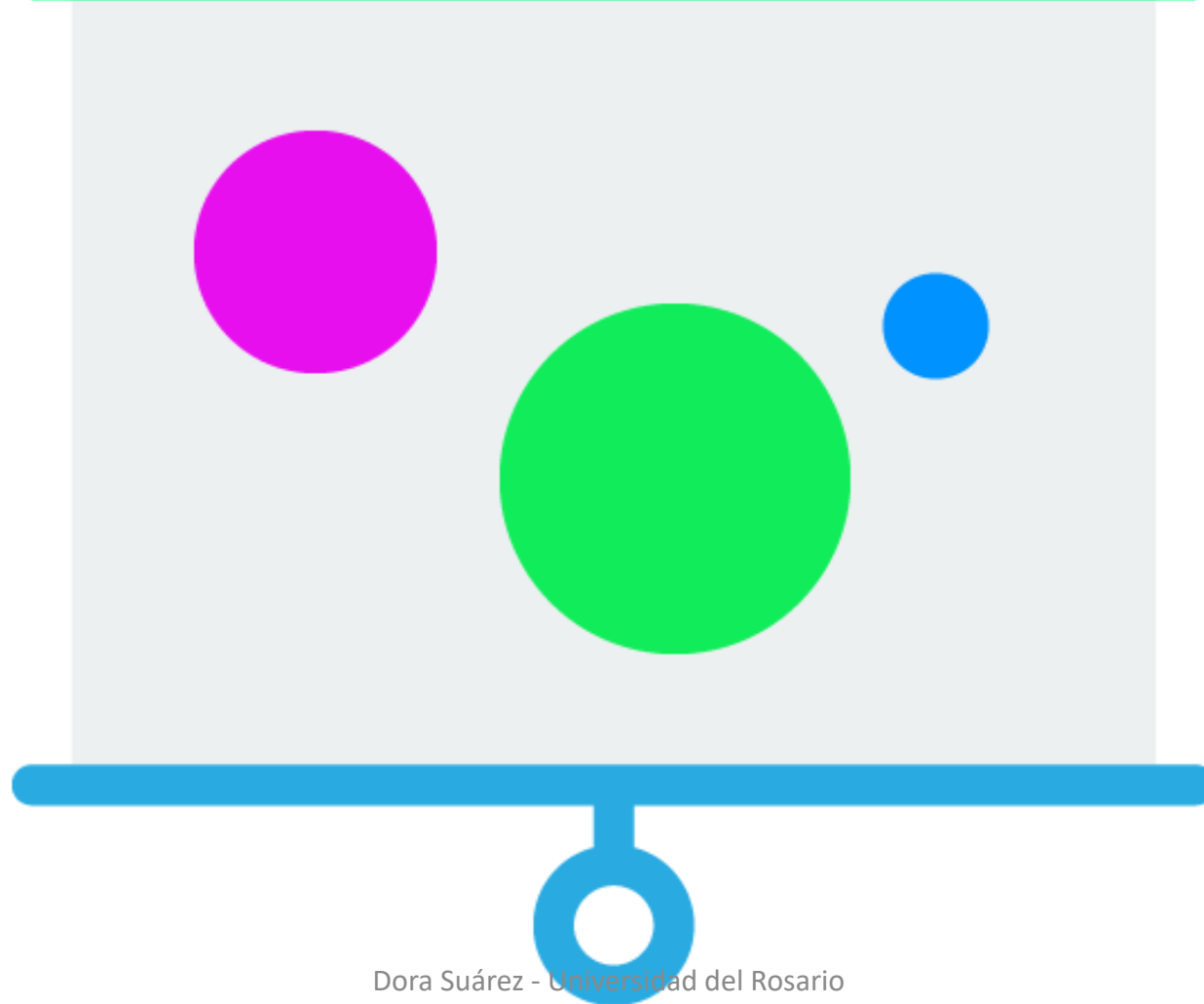
Ejercicio

Clasificar cada una de las variables presentes en la siguiente base de datos:



www.datos.gov.co/Seguridad-y-Defensa/Delito-Hurto-Personas/v6p7-acxt

Resumen de Variables Cualitativas



Mapa de los navegadores en dispositivos móviles

Navegador de Internet más utilizado en dispositivos móviles por país (octubre de 2016)

Chrome

Opera

Safari

Firefox

UC Browser

Android



El 51,3% del tráfico mundial de Internet proviene de dispositivos móviles



@Statista_ES

Fuente: StatCounter

Dora Suárez - Universidad del Rosario

statista

Resumen de variables cualitativas

Tablas de frecuencias

Número de ocurrencias en cada una de las categorías

Porcentaje de ocurrencias en cada una de las categorías

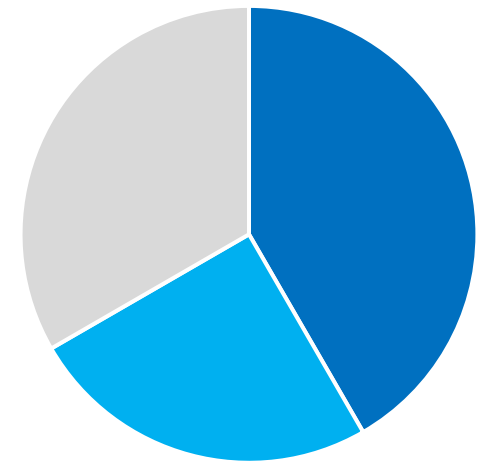
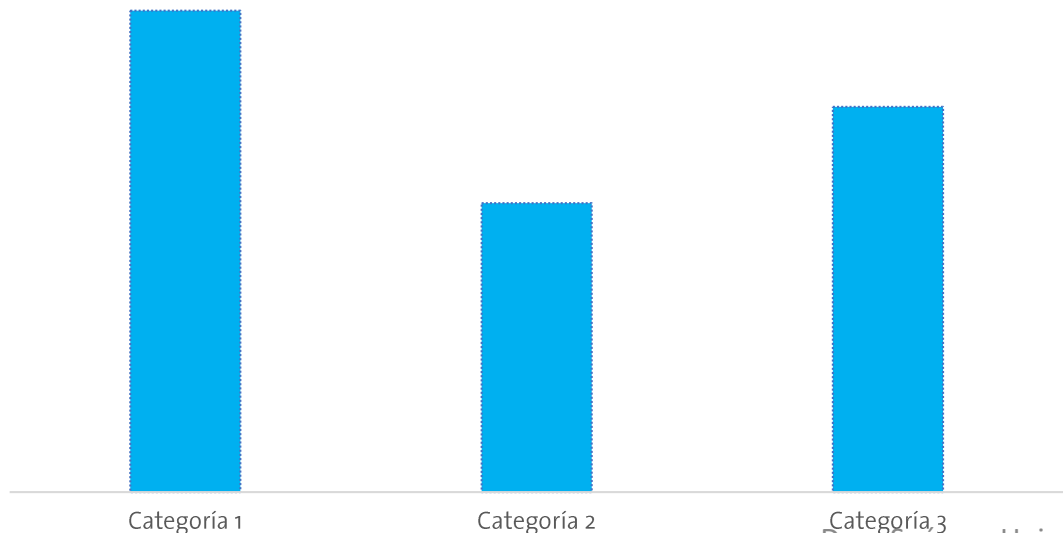
	Número de ocurrencias	Porcentaje de ocurrencias
Categoría 1	A1	$A1/N * 100$
Categoría 2	A2	$A1/N * 100$
...
Categoría k	Ak	$A1/N * 100$
Total	N	100

Resumen de variables cualitativas

Diagramas de barras y pastel

Barras: Frecuencias absolutas

Pastel: Frecuencias relativas



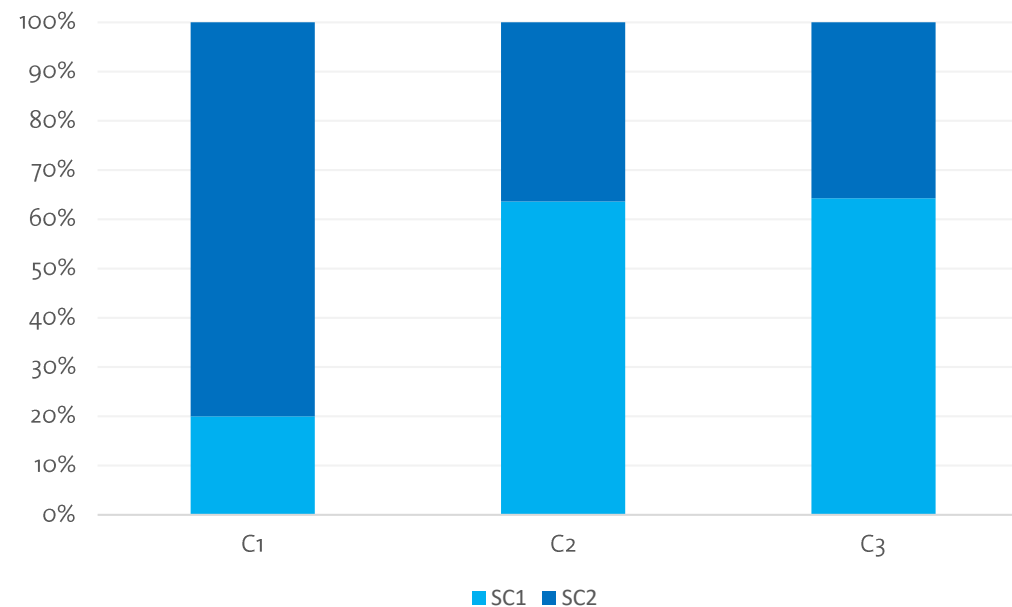
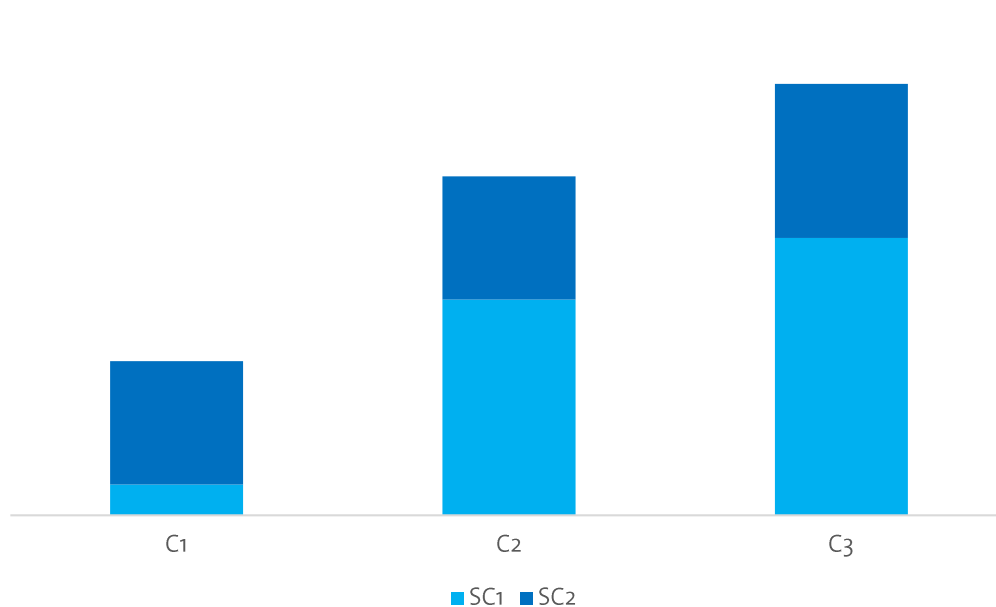
■ Categoría 1 ■ Categoría 2 ■ Categoría 3

Resumen de variables cualitativas

Diagramas de barras y pastel

Barras: Frecuencias absolutas

Pastel: Frecuencias relativas



Resumen de Variables Cuantitativas



Notación

Muestra aleatoria:

$$x_1, x_2, \dots, x_n$$

Resumen de variables cuantitativas

Rango Diferencia máxima observada $\text{Max}(x) - \text{Min}(x)$

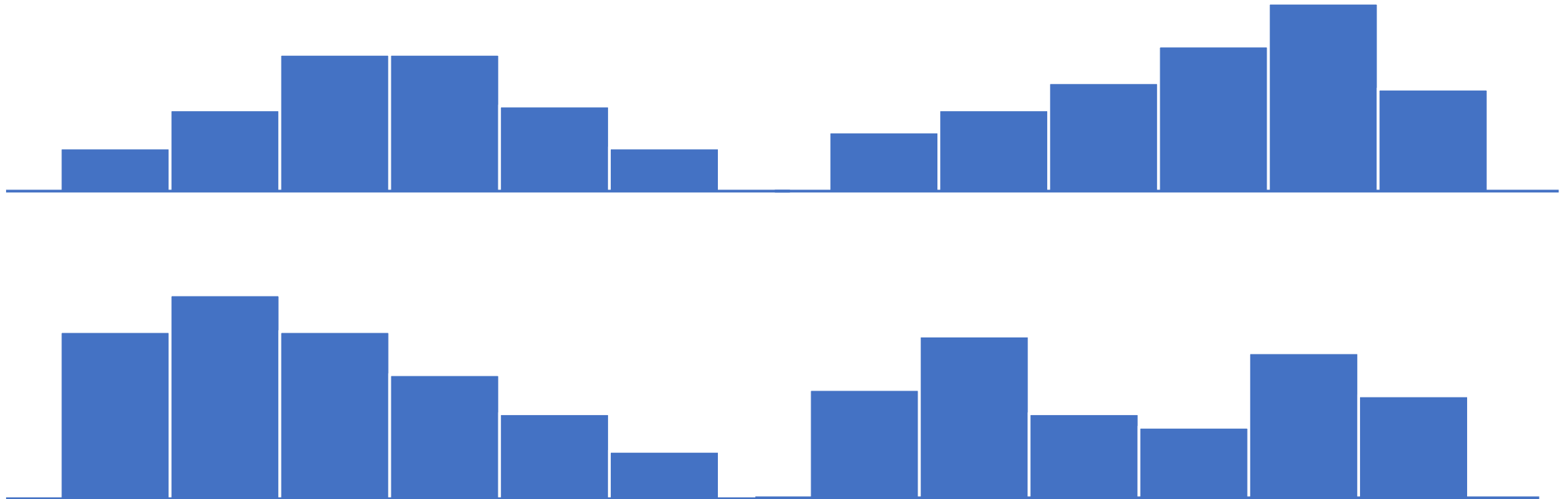
Tablas de frecuencias Número de ocurrencias por intervalos
Los intervalos pueden ser definidos de varias formas

Intervalo	Frecuencia absoluta	Frecuencia Relativa	Frecuencia Absoluta acumulada	Frecuencia Relativa Acumulada
$[a_1, a_2)$	f_1	$h_1 = f_1/N$	$F_1 = f_1$	$H_1 = h_1$
$[a_2, a_3)$	f_2	$h_2 = f_2/N$	$F_2 = F_1 + f_2$	$H_2 = H_1 + h_2$
...
$[a_{k-1}, a_k)$	f_k	$h_k = f_k/N$	$F_k = F_{k-1} + f_k = N$	$H_k = H_{k-1} + h_k = 1$
Total	N	1		

Amplitud = $\text{Rango} / \# \text{Categorías}$

El proceso de elaboración del histograma es igual que el de un diagrama de barras, se grafican los rangos vs las frecuencias absolutas

Resumen de variables cuantitativas



Tipos de medidas de Resumen

Medidas de tendencia central: Indican alrededor de que valores se espera que se pueda encontrar determinada característica

Medidas de dispersión: Indican que tan diferentes son las observaciones respecto a una medida de tendencia central

Medidas de posición: Ayudan a identificar cómo es la forma de la distribución, dónde se agrupan los datos y posibles datos atípicos

Medidas de tendencia central – Promedio

El **promedio** presenta una ponderación de los valores que asume una variable de acuerdo a su ocurrencia.

Puede ser calculada como:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

Medidas de tendencia central – Mediana

La **mediana** representa el valor de la variable de la posición central en un conjunto de datos ordenados.

Se calcula ordenando los datos de menor a mayor, si la cantidad de datos es impar, la mediana corresponderá al dato de la posición $\frac{n+1}{2}$, si la cantidad de datos es par, la mediana corresponderá al promedio de los dos datos centrales.

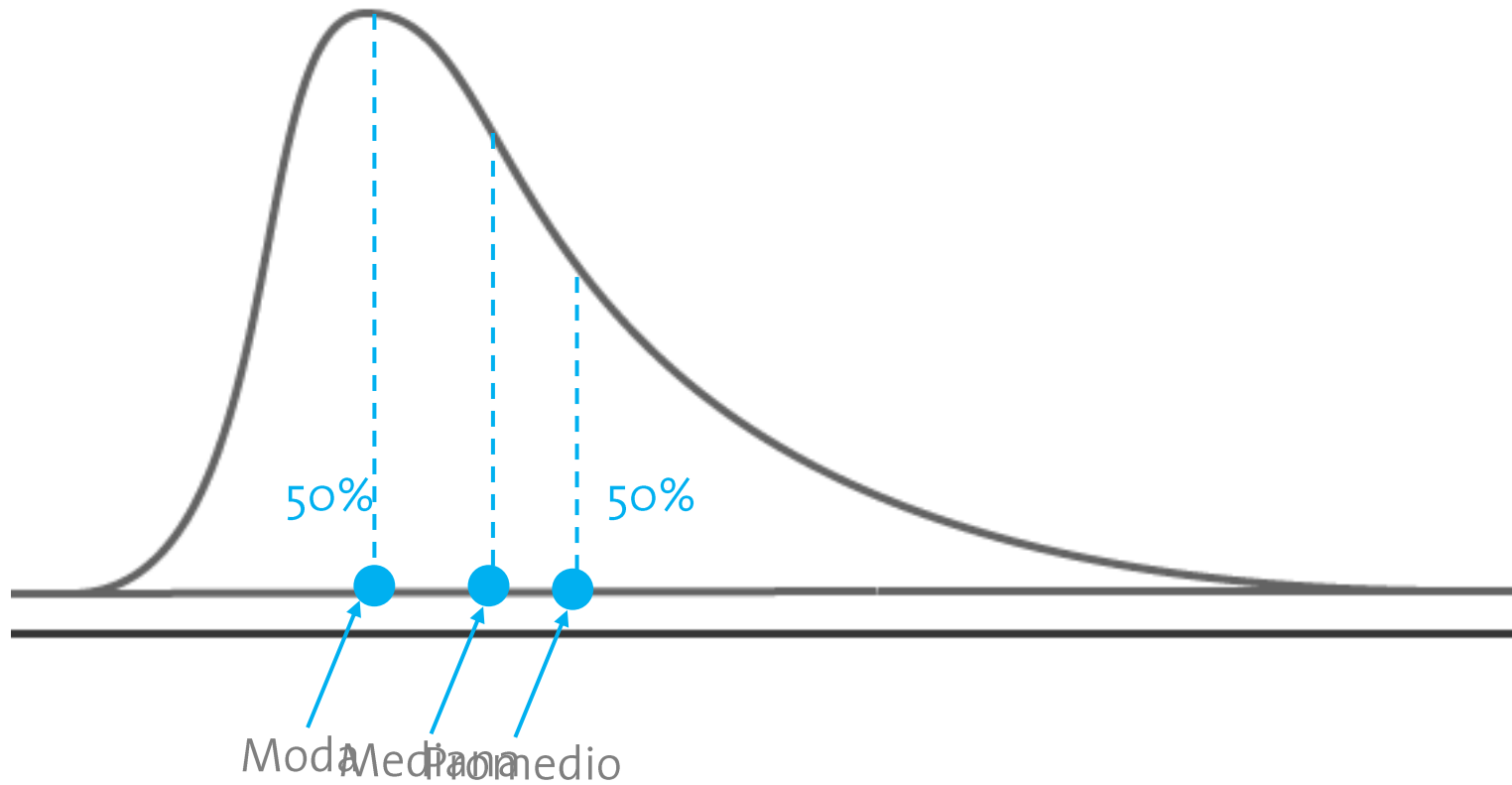
Medidas de tendencia central – Moda

La **moda** es el dato que tiene mayor frecuencia en una distribución de datos.

Una distribución puede tener una o varias modas.

La moda puede ser una categoría, un valor o un intervalo modal.

Medidas de tendencia central



Ejercicio 2 – ¿Falso o verdadero?

La moda media y mediana pueden ser iguales

El promedio de las observaciones menos el promedio es siempre cero

Si la mediana de dos muestras es la misma, entonces la distribución de los datos es la misma

El promedio es una medida robusta (no afectada por datos atípicos)

Medidas de dispersión – Coeficiente de variación

Variación respecto a la media en términos porcentuales. Sirve para hacer comparaciones entre datos que no tengan la misma unidad de medida.

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

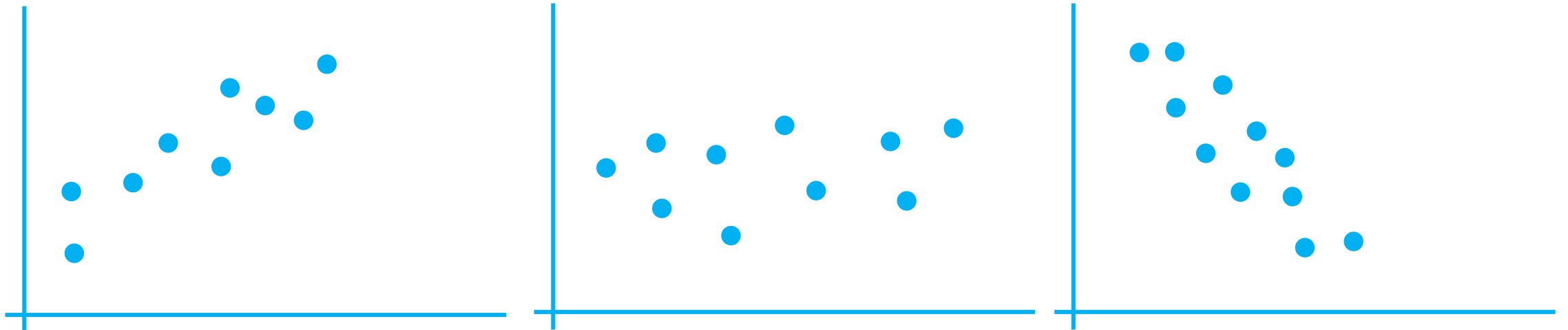
Ejercicio 3 – Calcular el coeficiente de variación y comparar

	Media	Desviación estándar
Estatura	176	7,7
Peso	78	12,3

Medidas de dispersión – Coeficiente de correlación

Si tenemos dos variables numéricas, el valor del coeficiente de correlación está dado por:

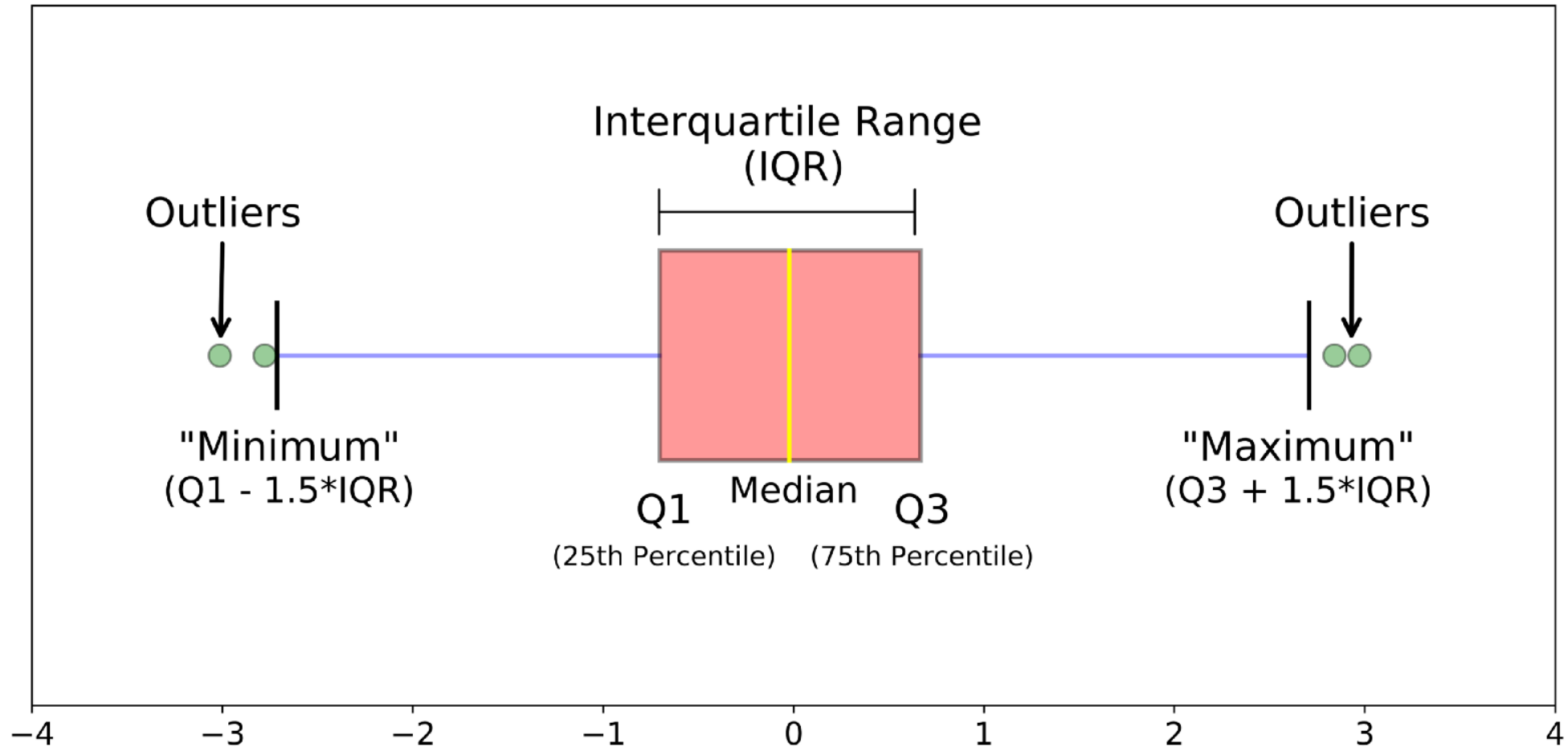
$$r = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{(n - 1)s_x s_y}$$



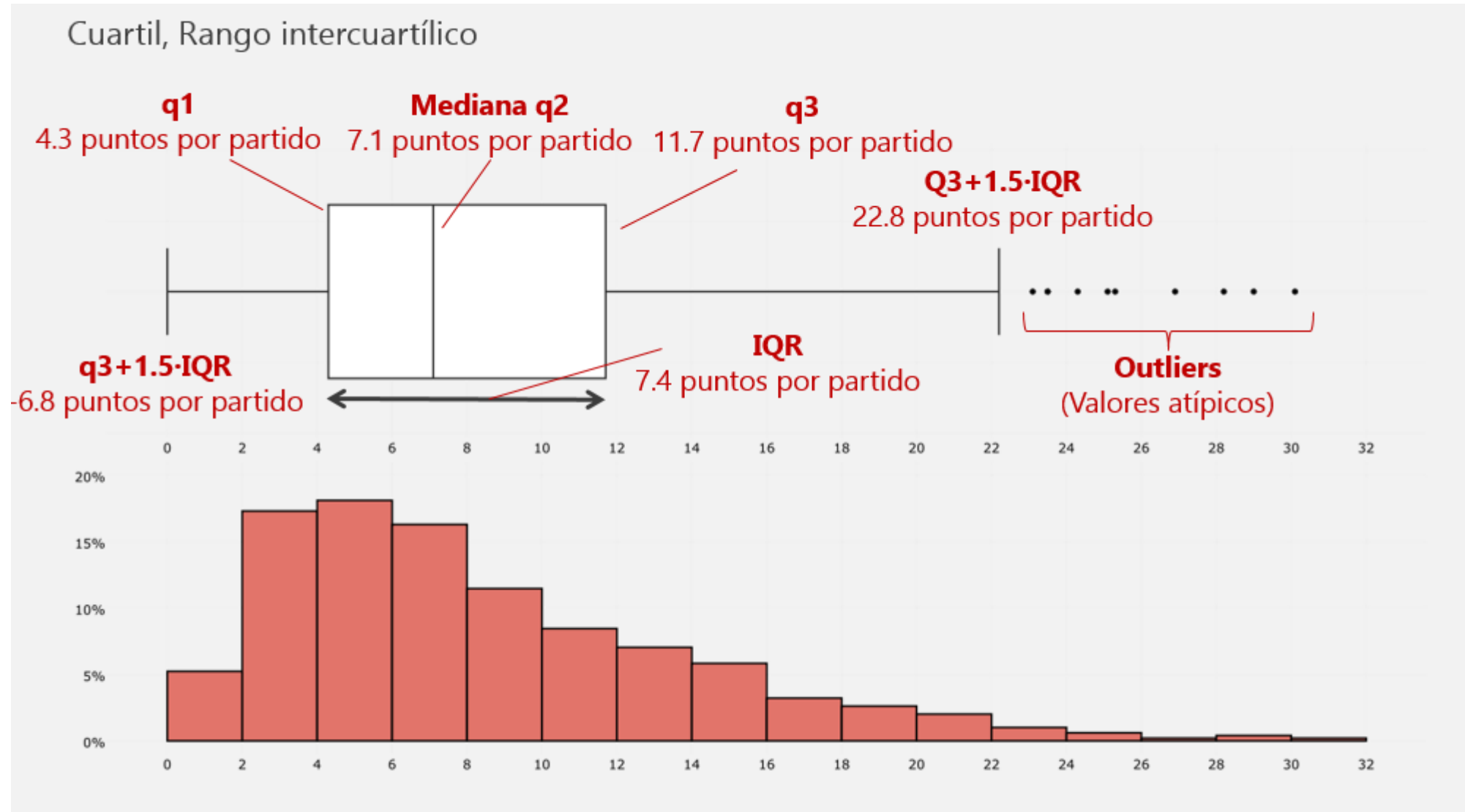
Medidas de localización – Percentiles

Formalmente el **percentil** corresponde al valor tal que por lo menos el p por ciento de las observaciones son menores o iguales que este valor y por lo menos el $(100 - p)$ por ciento de las observaciones son mayores o iguales que este valor.

Medidas de localización – Boxplot



Medidas de localización – Boxplot



Medidas de dispersión – Rango

Distancia máxima observada en un conjunto de datos



$$\text{Rango} = \max(x_i) - \min(x_i)$$

Medidas de dispersión – Varianza

Indica la dispersión de un conjunto de observaciones respecto al promedio. Se define como la media de las diferencias con la media elevadas al cuadrado.

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

Ejercicio 1

Mostrar que

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

Medidas de dispersión – Desviación estándar

Es la raíz cuadrada de varianza y se encuentra en las mismas unidades de medida de la variable original.

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

