SESIÓN 3 PROGRAMACIÓN EN R

ESTADISTICA DESCRIPTIVA

Hugo Andrés Dorado

Contenido

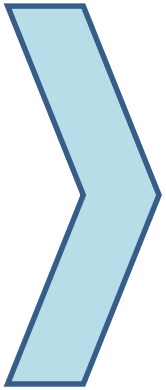
- **Estadística**
- Variables
- Medidas descriptivas
- Gráficos

¿Qué es la estadística?

“Es la ciencia que busca comprender al mundo a través de los datos”

Proporciona métodos para:

- Recolectar
- Describir
- Evaluar
- Interpretar



DATOS

Transfórmalos en información útil para tomar decisiones.

Usuarios de la estadística

- Diarios y revistas.
- Políticos.
- Marketing.
- Control de calidad.
- Investigadores científicos.
- Médicos.
- Seguros de vida
- Bancos
- Etc.

Estadística descriptiva

Recolectar, ordenar y clasificar datos obtenidos por observaciones.

- Los datos del censo poblacional 2001
- La cantidad de homicidios en Cali la ultima semana
- La cantidad de goles anotados por Colombia en el último partido con Uruguay.

Estadística inferencial

Proporciona métodos para **estimar** las características de un grupo total **(población)**, basándose en datos de un conjunto pequeño **(muestra)**.

- Índices de la encuesta nacional de hogares.
- Preferencia por los candidatos a la presidencia

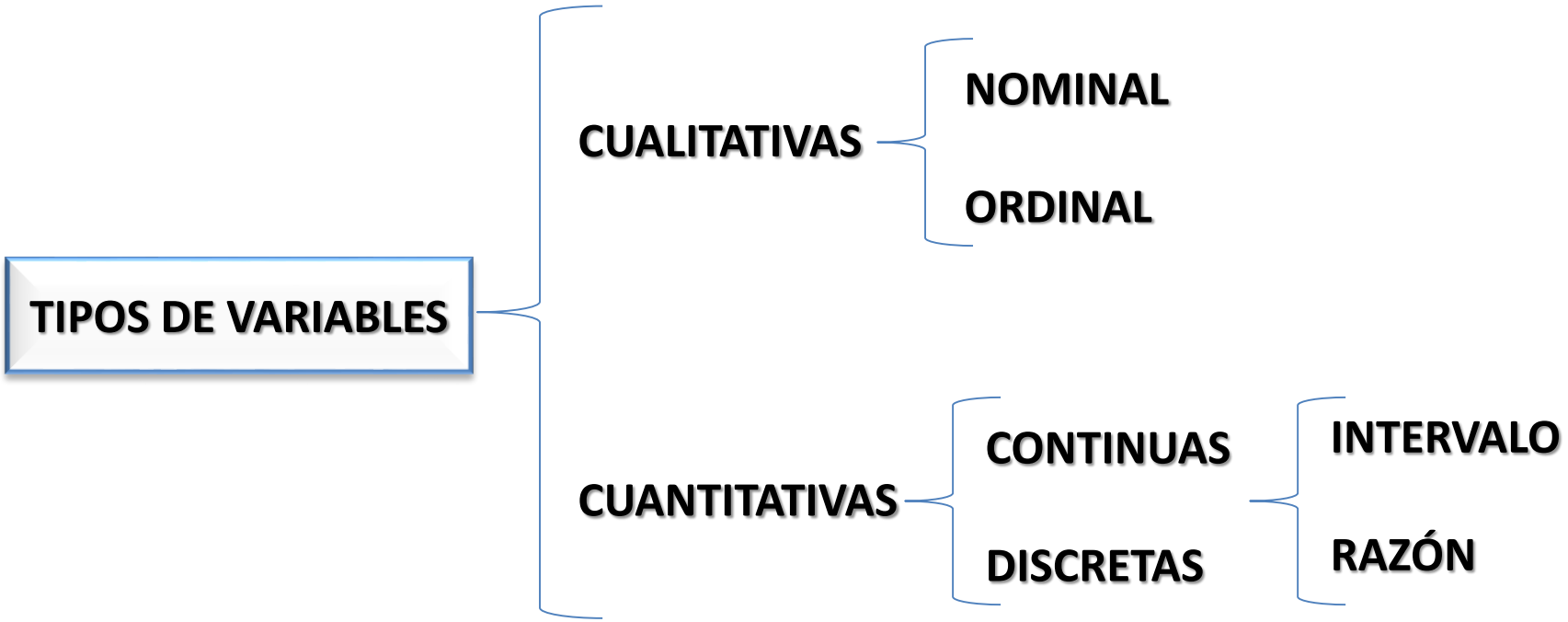
Población y muestra

- **Población**: Es la colección, o conjunto de individuos, objetos o eventos cuyas propiedades serán analizadas
- **Muestra**: Un subconjunto de la población de interés.

Contenido

- Estadística
- **Variables**
- Medidas descriptivas
- Gráficos

VARIABLES


TIPOS DE VARIABLES
CUALITATIVAS
NOMINAL
ORDINAL
CUANTITATIVAS
CONTINUAS
INTERVALO
DISCRETAS
RAZÓN

VARIABLES

CUALITATIVAS

- Si sus valores (modalidades) no se pueden asociar naturalmente a un número.
- No se pueden hacer operaciones algebraicas con ellos.

Nominales:
Si sus valores no se pueden ordenar.
Sexo, Religión, Nacionalidad, Fumar (Si/No)

Ordinales:
Si sus valores se pueden ordenar.
Grado de satisfacción, Intensidad de dolor, Mejoría a un tratamiento.

CUANTITATIVAS

- Si sus valores son numéricos.
- Tiene sentido hacer operaciones algebraicas con ellos.

Discretas:
Si toma valores enteros.
Número de hijos, Número de carros.

Continuas:
Si entre dos valores, son posibles infinitos valores intermedios.
Altura, Temperatura, Duración de una batería, Peso(kg).

ESCALAS DE MEDICIÓN

CUALITATIVAS

1. Escala Nominal:
No puede establecer un orden jerárquico entre las opciones de respuesta.

Color de Ojos (Verde, Azul, Gris, Negro, Café).

2. Escala Ordinal:
Existe un ordenamiento natural de las opciones de respuesta.

Calificación de un servicio (Excelente, Bueno, Regular, Malo).

CUANTITATIVAS

3. Escala de Intervalo:
El valor 0 es un valor arbitrario, no implica la no presencia de una característica.

Temperatura = 0ºC ¿No hay temperatura?

4. Escala de Razón:
El valor 0 refleja ausencia de la característica.

Altura = 0 mts

Contenido

- Estadística
- Variables
- **Medidas descriptivas**
- Gráficos

Medidas descriptivas

Son valores numéricos calculados a partir de la muestra o población y nos resumen la información contenida en ella.

(Posición, centralización, dispersión)

Posición: Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos.

Medidas descriptivas

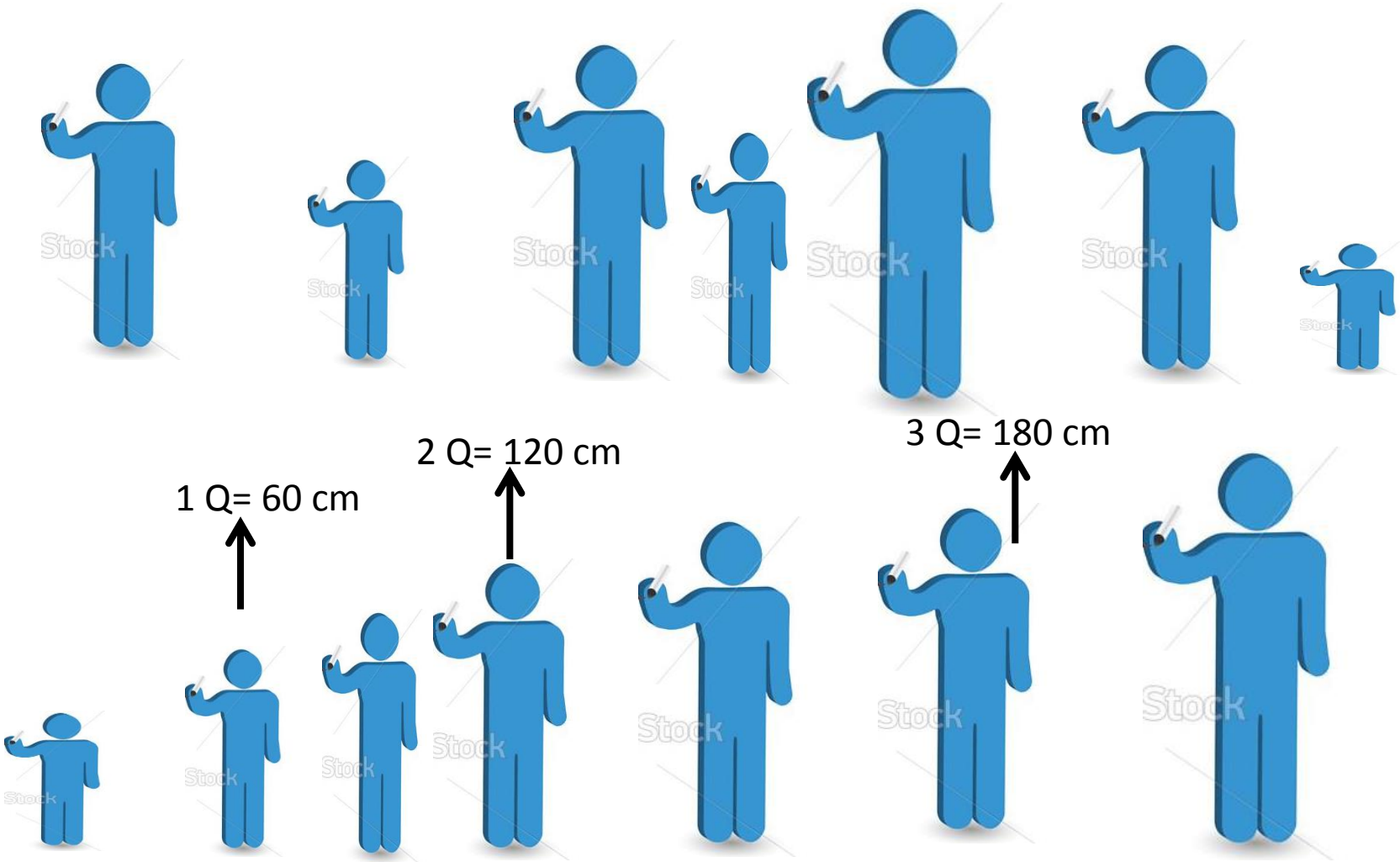
CUANTILES: Valores de la distribución que la dividen en partes iguales los mas usados son:

- **Percentiles:** son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados.

- **Cuartiles:** son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales.

- **Deciles:** son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales.

Medidas descriptivas



Medidas descriptivas

Centralización: Indican valores con respecto a los que los datos parecen agruparse.

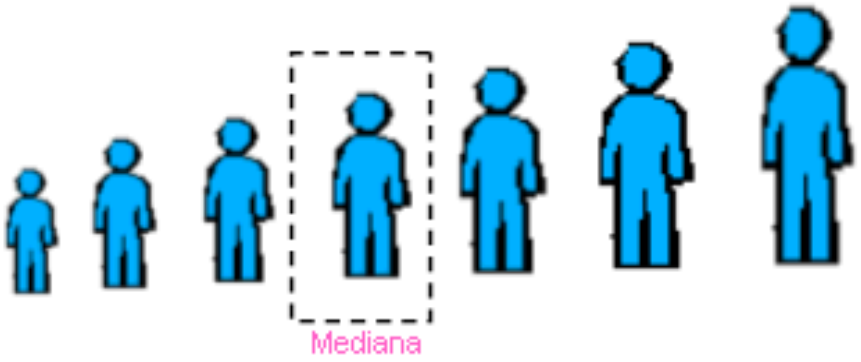
- **Media:** promedio aritmético de las observaciones

$$\bar{x} = \sum_{i=0}^{n} \frac{x_i}{n}$$

$$\frac{49 + 51}{2} = \frac{1 + 99}{2}$$

Medidas de Centralización

- **Mediana:** es el valor que separa por la mitad las observaciones ordenadas de menor a mayor.



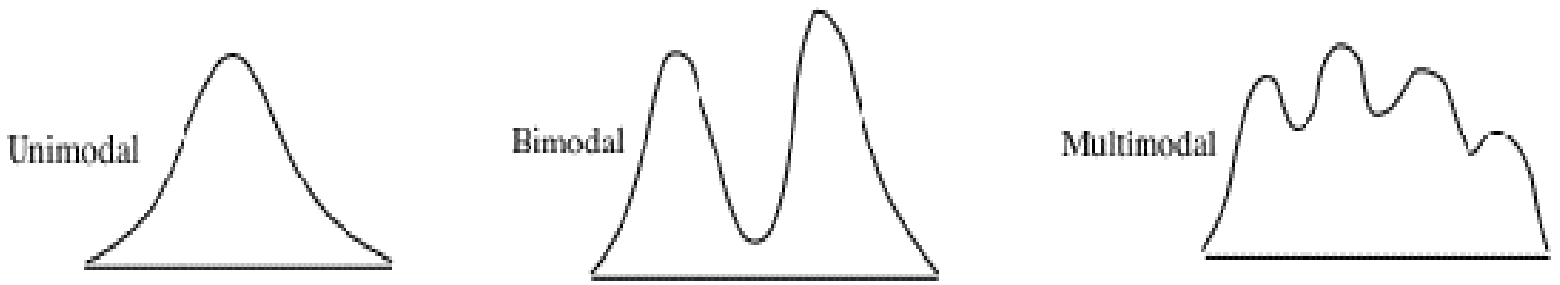
(18, 21, 24, 26, 53)

Me = 24

Mu= 30.8

Medidas Centralización

- **Moda:** es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor.



Medidas de dispersión

Cuantifican la separación, la dispersión y la variabilidad de los valores de la distribución respecto al valor central.

Medidas de dispersión

- **Varianza:** es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto de observaciones.

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

Medidas de dispersión

- **Desviación típica:** La varianza viene dada por las mismas unidades que la variable pero al cuadrado.

$$s = \sqrt{s^2}$$

Medidas de dispersión

- **Recorrido o rango muestral :** Es la diferencia entre el valor de las observaciones mayor y el menor.

$$R_e = x_{max} - x_{min}$$

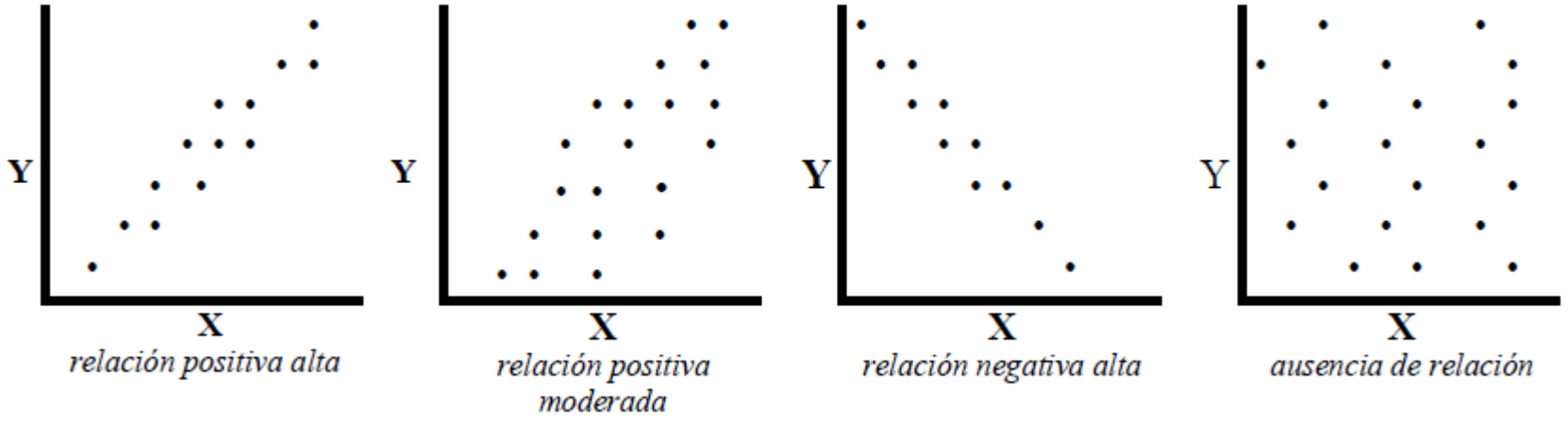
Medidas de dispersión

- **Coeficiente de variación de pearson:** Representa el número de veces que la desviación típica contiene a la media aritmética.

$$CV = \frac{S}{|\bar{x}|}$$

Medidas de covariación y correlación

Relación existente entre dos o mas variables cuantitativas.



Medidas de covariación y correlación

Coeficiente de correlación de Pearson

El coeficiente de correlación comprueba y cuantifica solamente *relaciones lineares*

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

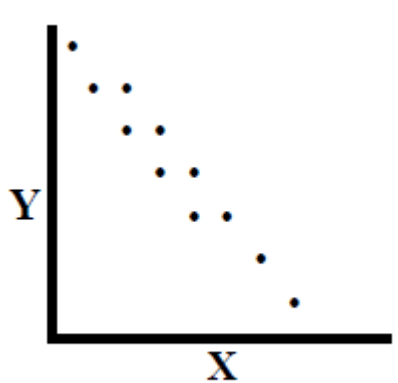
$$r = \frac{S_{XY}}{S_X\, S_Y}$$

$$< -1 \quad 0 \quad 1 >$$

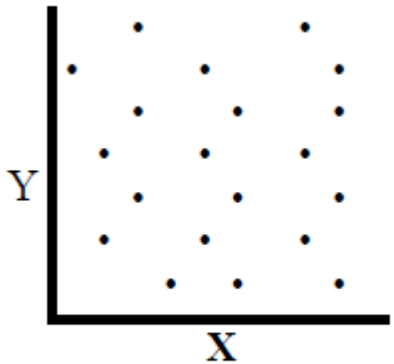
Medidas de covariación y correlación

r

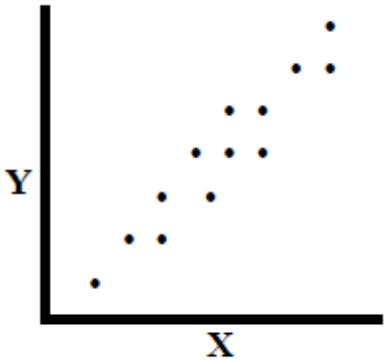
-1 0 1



relación negativa alta



ausencia de relación



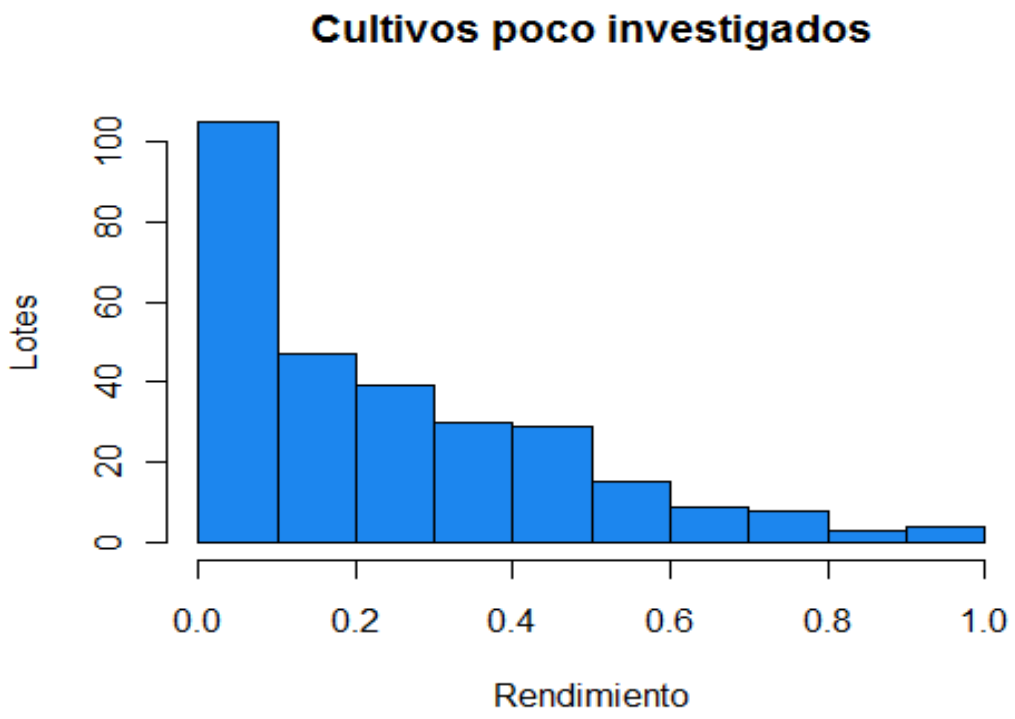
relación positiva alta

Contenido

- Estadística
- Variables
- Medidas descriptivas
- **Gráficos descriptivos**

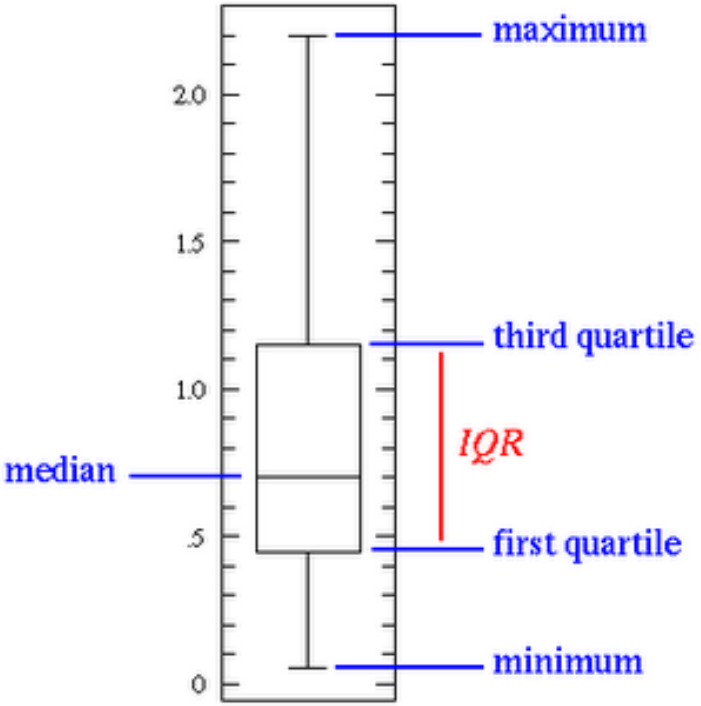
Gráficos descriptivos

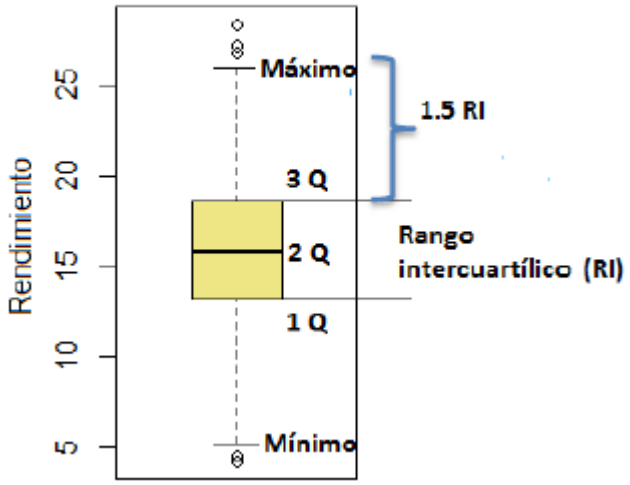
- **Variables continuas:** Histograma



Gráficos descriptivos

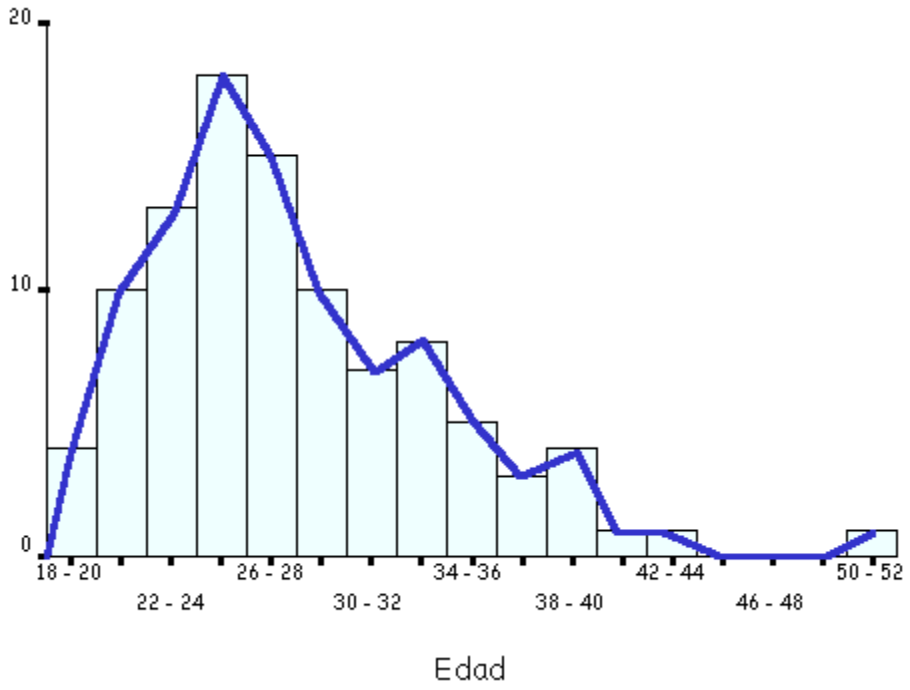
- **Variables continuas:** Boxplot





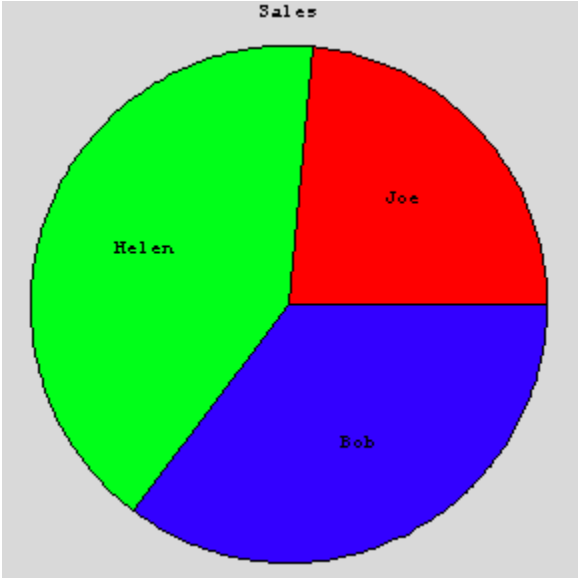
Gráficos descriptivos

- **Variables continuas:** Polígonos de frecuencia



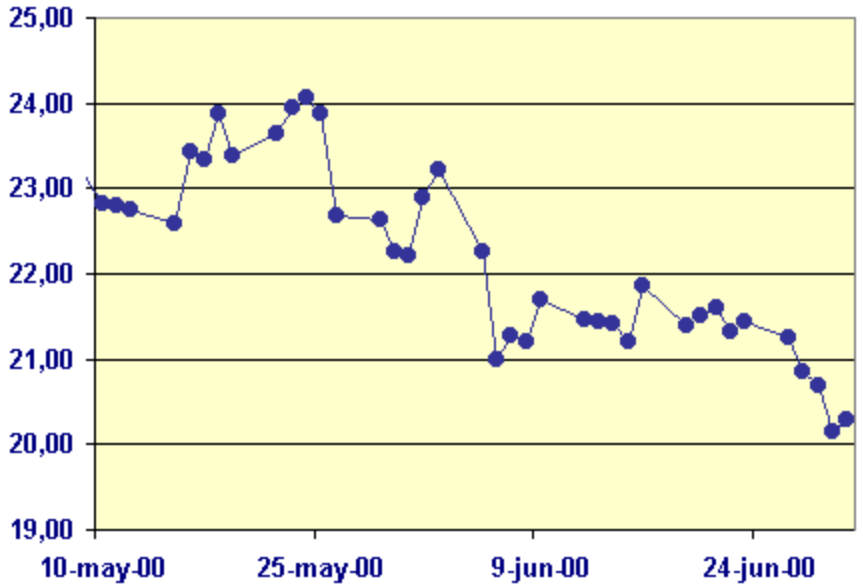
Gráficos descriptivos Variables cualitativas

- Grafico circular.



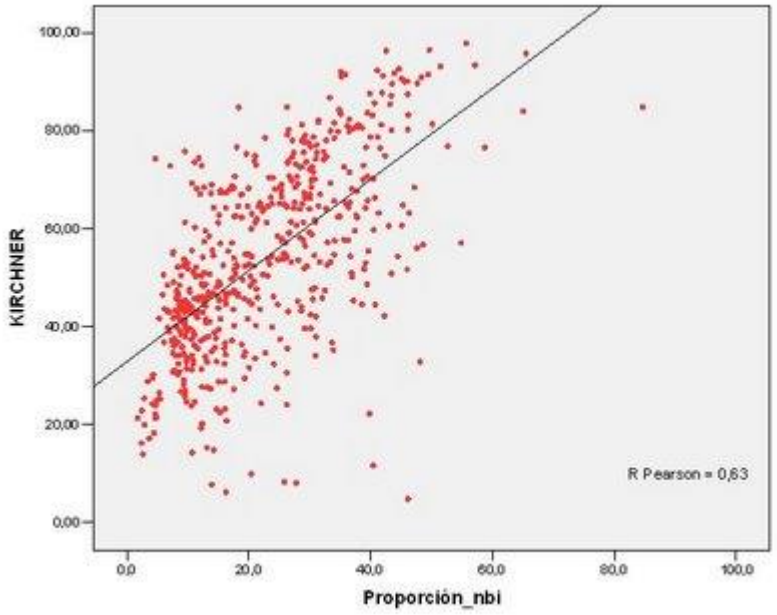
Gráficos descriptivos bi variados

- **Gráfico de líneas.**



Gráficos descriptivos bi variados

- **Gráficos de dispersión.**



Gráficos descriptivos bivariados.

- **Histogramas de frecuencias**

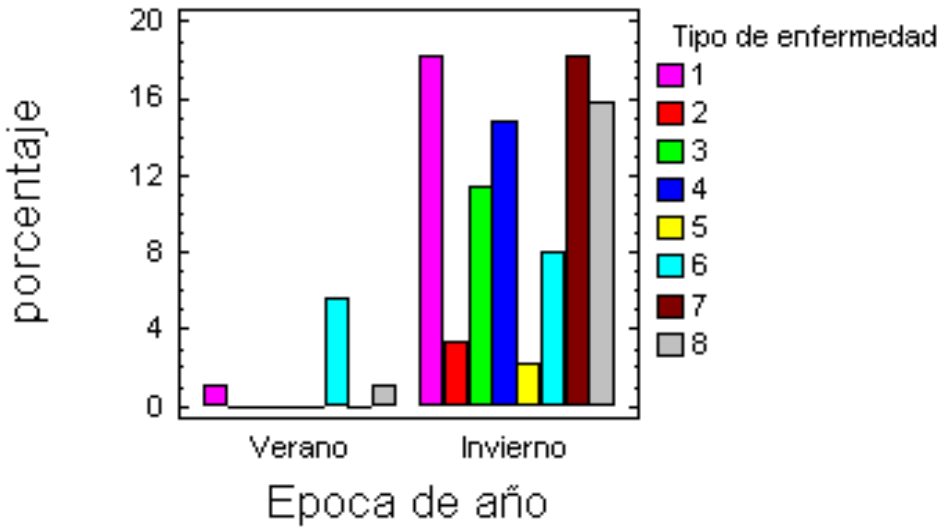


FIGURA 2. Gráfico del tipo de enfermedad y la época del año.

Ejercicios propuestos

Bases de datos

1. Leer el archivo eventos_de_platano.csv, especificar row.names.
2. Ajustar el formato de la fecha.
3. Calcular el rendimiento (PN_ANIO/AREA_UM) y agregarlo directamente usando $.
4. Resumir la base de datos para explorar sus variables.
5. Realizar análisis exploratorio de datos para cada una de las variables de la base de datos. (Medidas de tendencia central y dispersión; complementar con gráficos)
6. Hacer un gráfico de puntos Edad vs Rendimiento; indique la media dentro del gráfico.
7. Realizar un gráfico boxplot en ggplot con rendimiento, comparando las variedades y agrupado por dibujo de siembra.
8. Cruzar distancia de fecha vs Rendimiento (geom_smooth) pero agrupar por dibujo de siembra y variedad.

Bibliografía

- http://www.tuveras.com/estadistica/estadistica02.htm
- http://es.scribd.com/doc/70141495/Curso-breve-de-Estadistica