Análisis Estadístico de Datos

Fundamentos Descriptivos

Profesor: Nicolás López

2026-02-13

Tabla de Contenidos

1. Introducción

La estadística es una herramienta central en la toma de decisiones. En economía, negocios, políticas públicas e investigación científica:

- Permite transformar datos en información.
- Reduce incertidumbre.
- Apoya decisiones racionales.
- Permite comparar escenarios.

Tomar decisiones sin estadística implica basarse en intuición. Tomar decisiones con estadística implica basarse en evidencia.

La estadística descriptiva es el primer paso en ese proceso.

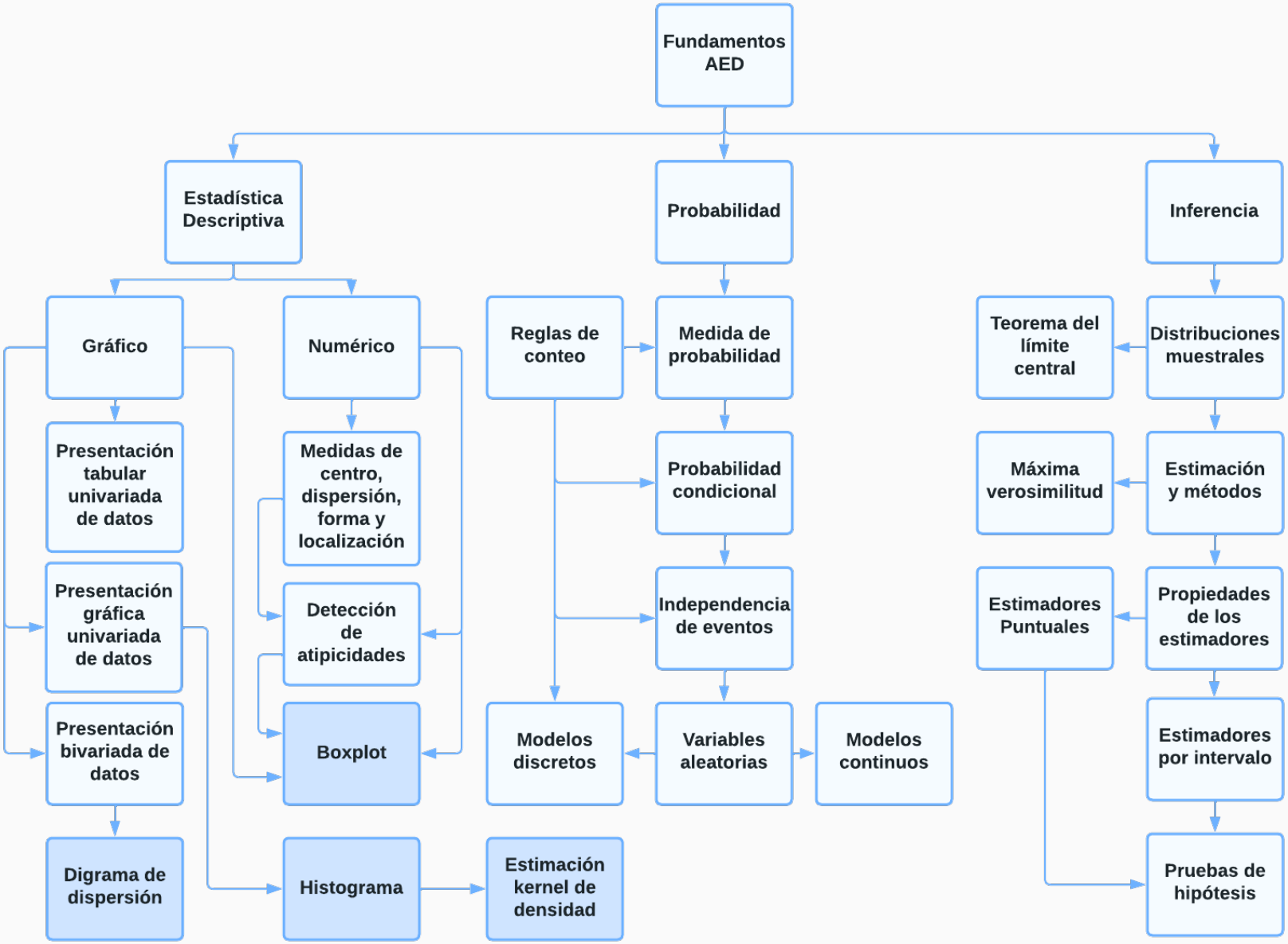


Figure 1: Diagrama detallado de contenidos fundamentales para el desarrollo del curso

2. R como lenguaje de programación

R es un lenguaje de programación diseñado específicamente para análisis estadístico y científico.

No es solo un software para "hacer cuentas".

Es:

- Un lenguaje interpretado.
- Orientado a objetos.
- Optimizado para operaciones vectoriales.
- Altamente extensible mediante paquetes.

R permite:

- Manipulación eficiente de datos.
- Cálculo matricial.
- Simulación.
- Generación de reportes reproducibles.

La filosofía de R se basa en trabajar con objetos completos y no elemento por elemento.

```
x <- c(2, 4, 6, 8)
mean(x)
```

```
## [1] 5
```

```
sd(x)
```

```
## [1] 2.581989
```

3. RStudio y Posit Cloud

RStudio es un entorno de desarrollo integrado que facilita:

- Escritura de código.
- Visualización gráfica.
- Gestión de paquetes.
- Organización de proyectos.

Posit Cloud permite trabajar con R desde el navegador:

- Sin instalación.
- Entorno homogéneo para todos.
- Ideal para enseñanza.

4. tidyverse

El tidyverse es un conjunto coherente de paquetes para:

- Importar datos.
- Manipular datos.
- Visualizar datos.

Principio fundamental:

- Cada variable es una columna.
- Cada observación es una fila.

```
data_charcoal <- read_csv("UNdata_Charcoal.csv",
                          show_col_types = FALSE)

charcoal_prd19 <- data_charcoal %>%
  filter(Year == 2019 &
         Commodity == "Charcoal - Production") %>%
  select(-Commodity)
```

5. Datos y estructura

Definiciones escenciales:

- Unidad estadística: país-área
- Variable de interés: Quantity

Una base de datos es una estructura rectangular donde:

- Filas representan unidades.
- Columnas representan variables.

Entender esta estructura es esencial antes de calcular cualquier medida.

6. Conceptos Fundamentales

- Variable: característica que puede variar entre unidades.
- Dato: realización observada.
- Población: conjunto total.
- Muestra: subconjunto observado.

El análisis descriptivo trabaja con la muestra disponible.

6.1. Clasificación de Variables

Escalas de medición:

- Nominal: categorías sin orden.
- Ordinal: categorías con orden.
- Intervalo: diferencias significativas.
- Razón: posee cero absoluto.

Quantity es de razón.

Esto permite:

- Comparaciones proporcionales.
- Cálculo válido de medias y varianzas.

6.2. Distribución Univariada

Estudia una sola variable.

Responde preguntas como:

- ¿Dónde se concentra?
- ¿Qué tan dispersa está?
- ¿Es simétrica?
- ¿Tiene valores extremos?

Es el paso previo al análisis multivariado.

6.3. Medidas de Tendencia Central

Media

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

Representa el centro de gravedad de la distribución.

Es eficiente cuando la distribución es simétrica.

```
mean(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 429.075
```

Mediana

Valor central de la muestra ordenada.

Es robusta ante valores extremos.

Cuando la distribución es asimétrica, la mediana puede ser más representativa.

```
median(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 21.1975
```

6.4. Medidas de Dispersión

Varianza

$$s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$$

Mide la variabilidad promedio respecto al centro.

El cuadrado amplifica desviaciones grandes.

Desviación estándar

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

Se interpreta en las mismas unidades que la variable.

```
sd(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 962.7235
```

Rango intercuartílico

$$IQR = Q_3 - Q_1$$

Mide la dispersión del 50% central.

```
IQR(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 338.6513
```

6.5. Forma: Asimetría

La asimetría mide el grado de simetría de la distribución.

$$\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

Interpretación:

- Valor cercano a 0: distribución aproximadamente simétrica.
- Positiva: cola larga hacia la derecha.
- Negativa: cola larga hacia la izquierda.

Cuando hay asimetría positiva:

Media $>$ Mediana.

```
skewness(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 3.439393
```

La asimetría es importante porque afecta la interpretación de la media.

6.6. Forma: Curtosis

La curtosis mide la concentración y el peso de las colas.

$$\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4$$

Interpretación:

- Alta curtosis: colas pesadas y mayor probabilidad de valores extremos.
- Baja curtosis: distribución más plana.

No mide solo "pico", sino también frecuencia de eventos extremos.

```
kurtosis(charcoal_prd19$Quantity, na.rm = TRUE)
```

```
## [1] 16.20926
```

Es relevante para evaluar riesgo y estabilidad.

6.7. Atipicidades

Regla basada en IQR:

$$x_i < Q_1 - 1.5IQR \quad \text{o} \quad x_i > Q_3 + 1.5IQR$$

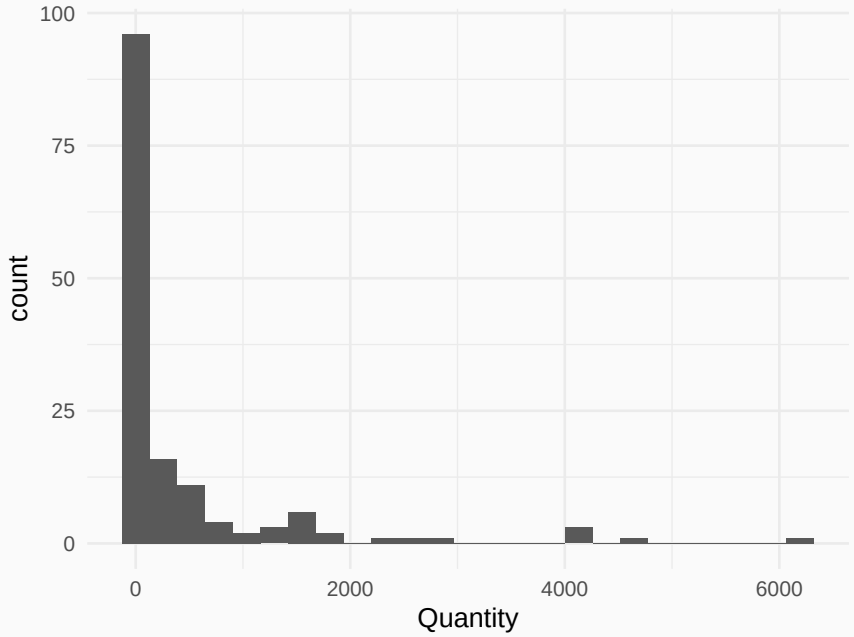
```
Q1 <- quantile(charcoal_prd19$Quantity, 0.25, na.rm = TRUE)
Q3 <- quantile(charcoal_prd19$Quantity, 0.75, na.rm = TRUE)
iqr <- IQR(charcoal_prd19$Quantity, na.rm = TRUE)

sum(charcoal_prd19$Quantity < Q1 - 1.5 * iqr |
    charcoal_prd19$Quantity > Q3 + 1.5 * iqr,
    na.rm = TRUE)
```

```
## [1] 22
```

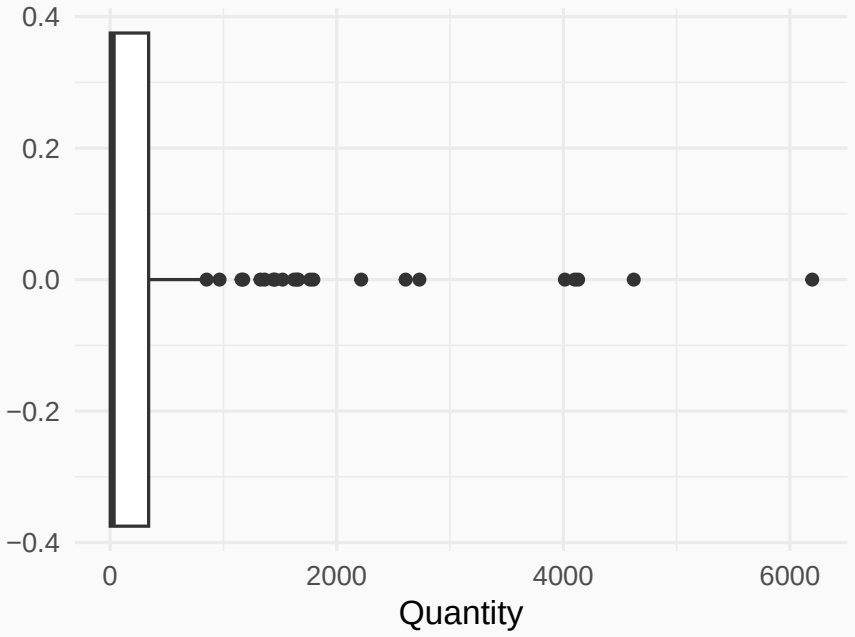
Un valor atípico no es necesariamente un error.

Histograma y boxplot

```
ggplot(charcoal_prd19,
       aes(x = Quantity)) +
  geom_histogram(bins = 25) +
  theme_minimal(base_size = 9)
```



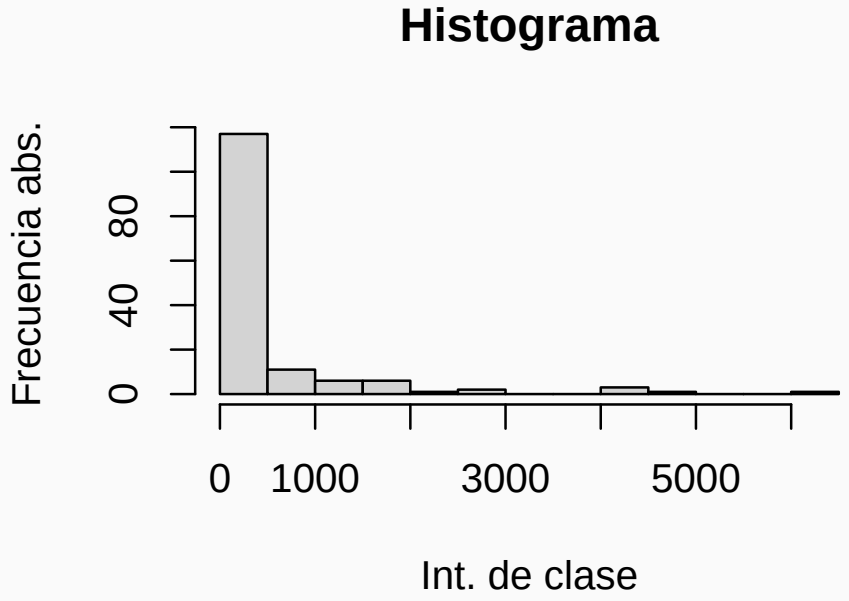

```
ggplot(charcoal_prd19,
  aes(x = Quantity)) +
  geom_boxplot() +
  theme_minimal()
```



Histograma

- Esta herramienta ampliamente utilizada en el análisis descriptivo de datos univariados permite una representación sucinta de la variable de interés.
- Versión continua del conocido diagrama de barras.
- La variable subyacente, al ser cuantitativa (y no cualitativa como en el diagrama de barras) necesita ser **discretizada**

```
hist(charcoal_prd19$Quantity,
     main = 'Histograma',
     xlab = 'Int. de clase',
     ylab = 'Frecuencia abs.')
```



Boxplot

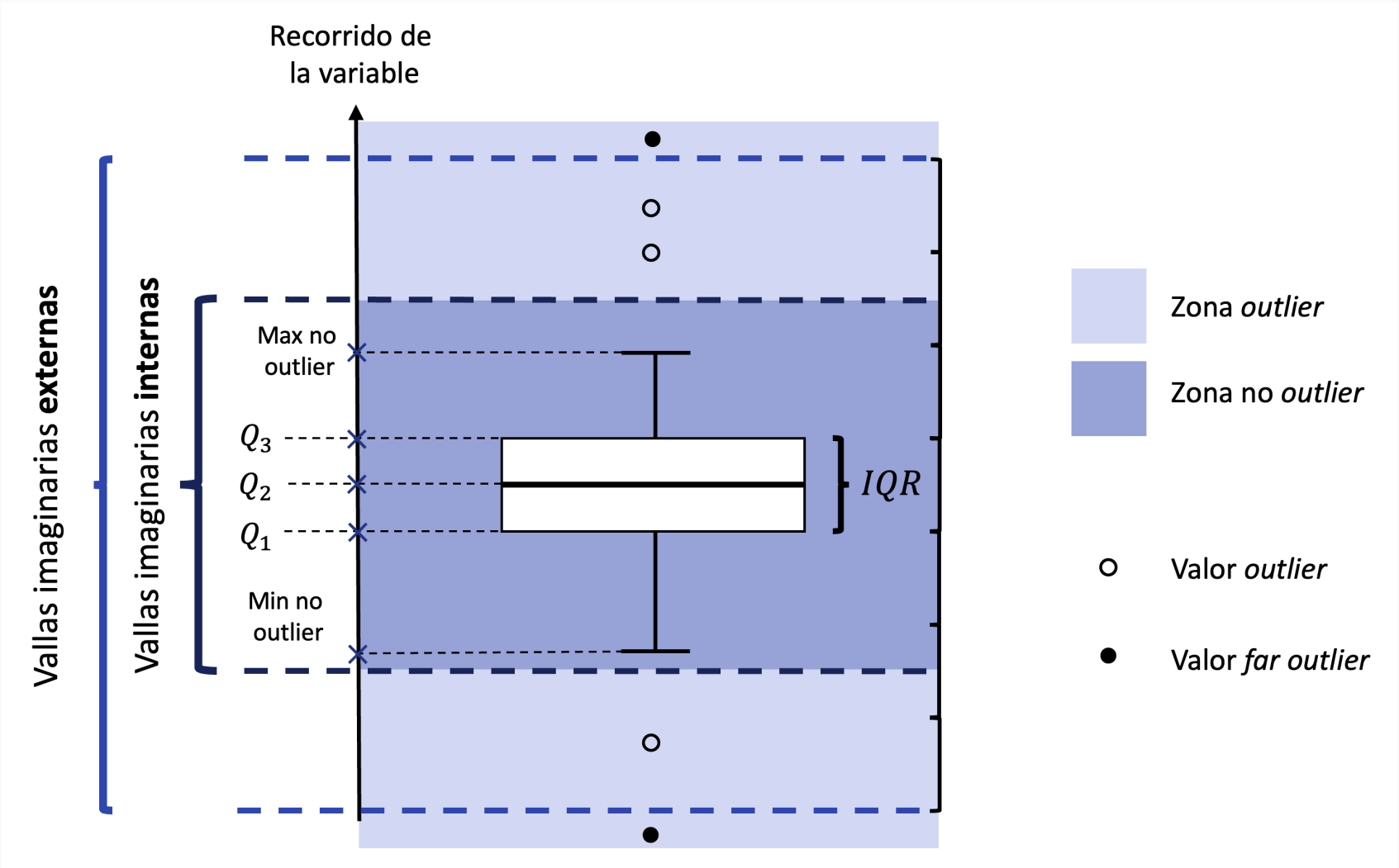


Figure 2: Elaboración de un boxplot o diagrama de caja

8. Covarianza

Hasta ahora analizamos una variable X (Quantity) con realizaciones $(x_1, \ldots, x_n)$. Sea ahora el vector de observaciones:

$$\mathbf{x} = (x_1, \ldots, x_n)$$

Y

$$\mathbf{y} = (y_1, \ldots, y_n)$$

La covarianza se define como:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Interpretación:

- Positiva: ambas variables tienden a moverse en la misma dirección.
- Negativa: tienden a moverse en direcciones opuestas.

9. Ejercicios

1. Compare media y mediana e interprete.
2. Explique por qué la varianza amplifica valores extremos.
3. Interprete una asimetría positiva.
4. ¿Qué implica una curtosis alta?
5. Interprete una covarianza negativa.

10. Síntesis

La estadística descriptiva:

- Resume información.
- Reduce incertidumbre.
- Permite decisiones basadas en evidencia.
- Es el fundamento del análisis multivariado.

La covarianza marca la transición hacia el estudio conjunto de variables.

TAREA Revisar el cuaderno de clase.