

Unidad 5. Análisis exploratorio de datos

1. Introducción

El proceso científico

El campo de la estadística se enfoca en el proceso científico de

- recolectar
- organizar
- analizar y
- extraer conclusiones

a partir de los “datos”.

Recordamos que:

“todo dato es una construcción”.

Las otras ciencias

La estadística se vincula con disciplinas o líneas disciplinares como:

- bioinformática
- genética (bioestadística genética)
- agronomía
- astronomía
- economía
- educación
- electrónica
- geología
- ciencias de la salud
- teoría de comunicación
- teoría de la información
- inteligencia artificial
- ...

¿Cómo la estadística colabora en la investigación?

La estadística puede asistir a otros campos científicos a través de

- El diseño de experimentos y encuestas
- La organización, descripción y resumen de los datos, abriendo paso a la estadística descriptiva.
- La producción de inferencias y toma de decisiones

¿Cómo la estadística colabora en la investigación?

La recolección de datos incluye los siguientes pasos:

1. Definir los objetivos del problema y proceder al desarrollo del experimento o encuesta.
2. Definir las variables y parámetros de interés.
3. Definir los procedimientos de recolección de datos y técnicas de medición: muestreo, tamaño muestral, dispositivos de medición de datos (cuestionarios, encuestas telefónicas, perfil de las/los encuestadoras/res, etc).

2. Conceptos básicos

Población y muestra

Definición

La **población** es el conjunto o colección de todos los objetos o sujetos o mediciones que son objeto de interés del estudio.

Definición

Un **parámetro** es una medida resumen sobre la población.

Definición

Cada elemento de la población se denomina **unidad de análisis** (unidad experimental en el contexto de un experimento).

Definición

La **muestra** de unidades es un subconjunto seleccionado de la población y el **tamaño muestral** es su cardinal.

Producción industrial

Ejemplo

Se quiere estimar el porcentaje o proporción de partes averiadas en una fábrica durante una semana dada (cinco días) producidas por día. Se seleccionan al azar por día 20 piezas fabricadas durante los cinco días.

Así:

- La población consiste en “todas las partes producidas en la fábrica durante la semana”.
- El parámetro es la proporción
- La muestra de unidades es el conjunto de 100 piezas elegidas al azar durante la semana.

Variables y datos

Definición

Una **variable** es una característica que varía en cada unidad de análisis.

Definición

Un **dato** es el valor observado de la variable medido en una unidad de análisis.

Definición

La **muestra estadística** son todos los datos correspondientes a la muestra de unidades.

Definición

La **población estadística** son todos los datos correspondientes a la población de unidades.

Observación: La noción de variable sigue siendo la misma de antes, cambia el sentido por el nuevo contexto.

Tipos de datos

Definición

*Un dato **cuantitativo** es una observación medida en escala numérica. Se clasifica en continuo o discreto.*

Definición

*Un dato **categorico** define la pertenencia de un objeto a una categoría o clase. Se clasifica en nominal u ordinal.*

Observación: Las variables reciben una denominación análoga.

Aplicaciones

Ejemplo

1. *La respuesta a un tipo de terapia puede ser clasificada en mejora, mejora parcial o sin mejora. Estos datos son cualitativos (categóricos nominales)*
2. *Los números de almaceneros minoristas según barrio en Río Cuarto son datos cuantitativos.*
3. *El grupo sanguíneo de una persona de Moldes es un dato categórico nominal.*
4. *El nivel educativo de un sujeto de la ciudad de Moldes puede ser pensado como un dato categórico ordinal.*

3. Resumen estadístico y visualización

Medidas de resumen y visualización

En estadística descriptiva se exploran los “datos” mediante:

- gráficos y
- medidas de resumen o **estadísticos**.

Seguimos una estrategia con dos momentos:

- Análisis descriptivo univariado: se resume y representa la información por cada variable separadamente.
- Análisis descriptivo multivariado: se representan y resumen las variaciones conjuntas de las variables.

Para ambos análisis seguiremos los siguientes materiales

- Bianco, A. (2019)¹
- Maronna, R. (2021).

¹http://cms.dm.uba.ar/academico/materias/2docuat2019/estadistica_M/

Encuesta a estudiantes

Los datos provienen de una encuesta a estudiantes de un curso introductorio de estadística². Hay 362 observaciones y 17 variables. Las variables son:

1. Year: Año en la institución: FirstYear, Sophomore, Junior, o Senior
2. Gender: Género: F o M
3. Smoke: Fuma? No ó Yes
4. Award: Premio preferido: Academy, Nobel, o Olympic
5. HigherSAT: Qué puntaje SAT es más alto? Math o Verbal
6. Exercise: Horas semanales de ejercicio
7. TV: Horas semanales de TV
8. Height: altura (en pulgadas)
9. Weight: peso (en libras)
10. Siblings: número de hermanos
11. BirthOrder: orden de nacimiento, 1 = mayor, 2 = segundo mayor, etc.
12. VerbalSAT: puntaje SAT en lengua
13. MathSAT: puntaje SAT en matemáticas
14. SAT: puntaje SAT combinado de lengua y matemáticas
15. GPA: puntaje promedio obtenido en el colegio
16. Pulse: pulso cardíaco (latidos por minuto)
17. Piercings: número de piercings

²<https://rdrr.io/rforge/Lock5Data/man/StudentSurvey.html>

Base de datos

En R, levantamos la base de datos

```
base_datos=read.csv("StudentSurvey.csv",header=TRUE)
```

Utilizando la función `colnames()` le cambiamos los nombres de las variables del inglés al castellano.

```
> colnames(base_datos)
[1] "Año"          "Genero"        "Fuma"          "Premio"
[5] "SATmax"       "Ejercicio"     "TV"            "Altura"
[9] "Peso"        "Hermanos"      "Ordennacimiento" "SATLengua"
[13] "SATMatematica" "SAT"          "GPA"           "Pulso"
[17] "Piercings"
```

Y, utilizando la función `gsub()` modificamos para las variables categóricas los nombres de las categorías, del inglés al castellano.

Representación de datos categóricos

Distribución empírica de una variable categórica.

- Tabla de frecuencias absolutas o relativas: indica el número de observaciones que caen en cada una de las clases de la variable o la frecuencia relativa de una clase (el cociente entre la frecuencia absoluta y el número total de observaciones).
- Gráfico de Barras: a cada categoría o clase de la variable se le asocia una barra cuya altura representa la frecuencia o la frecuencia relativa de esa clase. Las barras difieren sólo en altura, no en ancho y se representan separadas por un espacio.
- Gráfico de Tortas: Se representa la frecuencia relativa de cada categoría como una porción de un círculo, en la que el ángulo se corresponde con la frecuencia relativa correspondiente.

Tablas y gráficos de barras

La siguiente tabla nos da la distribución (empírica) de la variable Fuma:

Fuma	
No	Si
319	43

Tabla: Tabla de frecuencias absolutas para la variable Fuma

Gráficos de barras

Diagrama de barras para Fuma

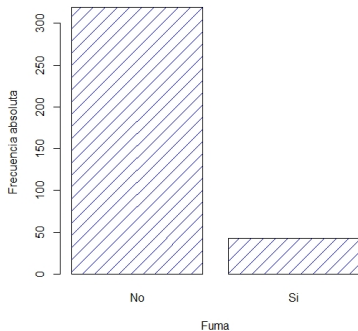


Gráfico de barras para Fuma

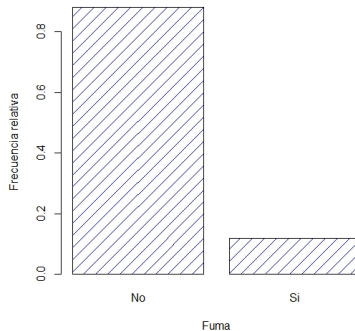
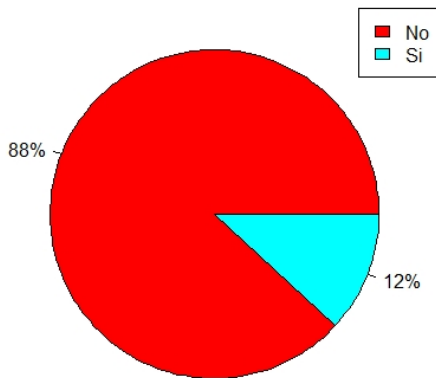


Figura: Gráfico de barras de la variable Fuma.

Gráfico de torta

Porcentajes para Fuma



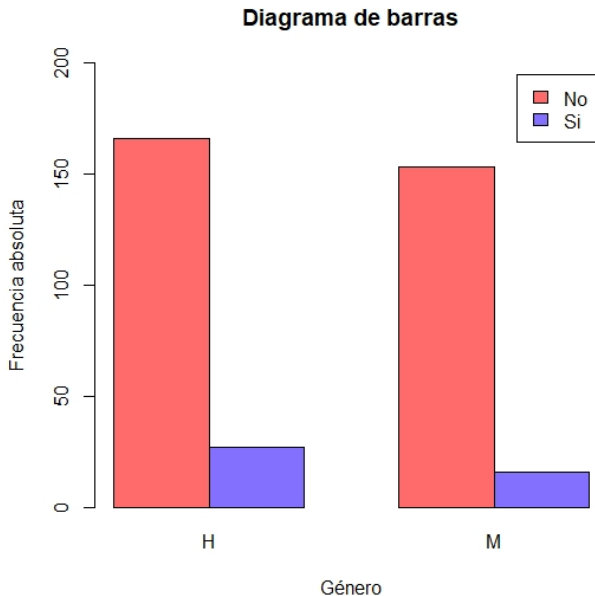
Tablas de contingencia

Tabla de contingencia para la distribución empírica de dos variables categóricas

	Género	
	H	M
Fuma		
No	166	153
Si	27	16
Total	193	169

Tabla: Tabla de contingencia para las variables Fuma y Género

Gráficos de barras para datos categóricos bivariados



Histograma para datos cuantitativos

Dada una muestra x_1, \dots, x_n de datos cuantitativos y un intervalo (a_0, a_m) que los contenga³, un histograma se construye del siguiente modo:

- Se elige una partición en m sub-intervalos $a_0 < a_1 < \dots < a_m$ y sea $L_j = a_j - a_{j-1}$, $j = 1, \dots, m$.
- Sea $f_j = \# \{i : x_i \in [a_{j-1}, a_j)\}$ (o $f_{j,r} = f_j/n$).
- Se grafica la función igual a f_j/L_j (ó $f_{j,r}/L_j$) en el intervalo $[a_{j-1}, a_j)$ y 0 fuera de los intervalos.

³Maronna (2021)

Histograma para datos cuantitativos

Notar que:

- Se obtiene un conjunto de rectángulos con área f_j (ó $f_{j,r} = f_j/n$)
- Un histograma es una versión discreta de la densidad, en la que áreas miden frecuencias.
- En los lenguajes, como R o Python, hay diferentes definiciones de histogramas (ver en R la documentación para la función `hist`).

Histograma de Altura

```
hist(base_datos$Altura, ylab='Frecuencia absoluta',main='Histograma para la variable altura',  
xlab='Altura (en pulgadas)', breaks = 5)
```

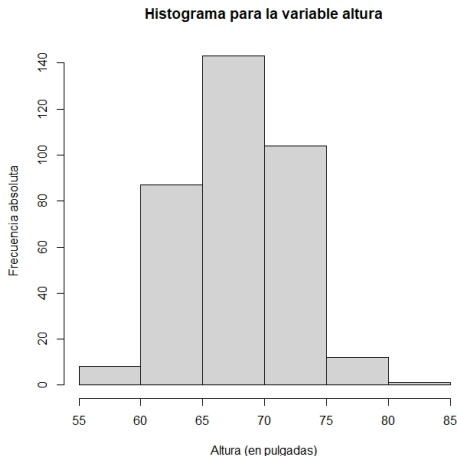


Figura: Histograma para Altura

¿Qué preguntas puede responder un histograma?

A partir de un histograma, como en Bianco (2021) nos preguntamos:

- ¿Cuál es el rango de variación de los datos (mínimo y máximo)?
- ¿Cuáles son los intervalos de mayor frecuencia?
- ¿La distribución es unimodal o multimodal (más de un pico)?
- ¿La distribución es simétrica?
- Si es asimétrica, ¿la asimetría es a derecha o a izquierda?
- ¿En torno a qué valor están aproximadamente centrados los datos?
- ¿Cuán dispersos en torno a este centro están los datos?
- ¿Hay datos atípicos en relación a la mayoría de los datos?

Algunos tipos de histogramas

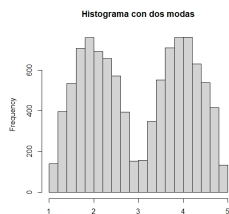
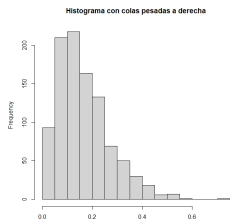
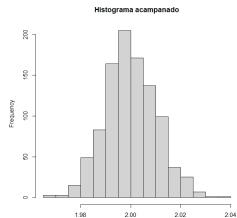
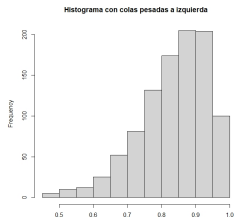
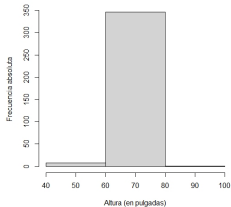


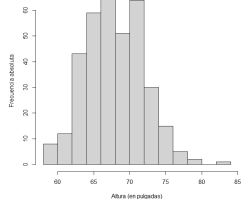
Figura: Diferentes tipos de histogramas

Selección del número de subintervalos

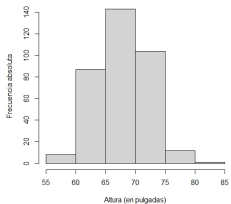
Histograma para la variable altura con breaks = 1



Histograma para la variable altura con breaks=default



Histograma para la variable altura con breaks = 5



Histograma para la variable altura con breaks=22

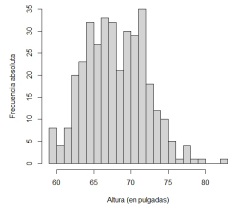


Figura: Histogramas con diferentes número de breaks (extremos de subintervalos) para la variable Altura

Selección del número de subintervalos

Hay varios métodos para seleccionar el número de subintervalos:

- Método de Sturges.
- Regla de Freedman - Diaconis.
- Regla de Scott.

En R se utiliza un método de dos pasos.

Diagramas de dispersión para datos bivariados cuantitativos

Consideremos dos variables cuantitativas X e Y , dadas

- una muestra estadística x_1, \dots, x_n de la variable X
- una muestra estadística y_1, \dots, y_n de la variable Y

en un diagrama de dispersión se grafican (x_i, y_i) , $i = 1, \dots, n$ en un sistema de ejes coordenados cartesianos.

El objetivo de tal diagrama o gráfico es visualizar alguna estructura entre los datos bivariados.

Correlación entre variables de la encuesta

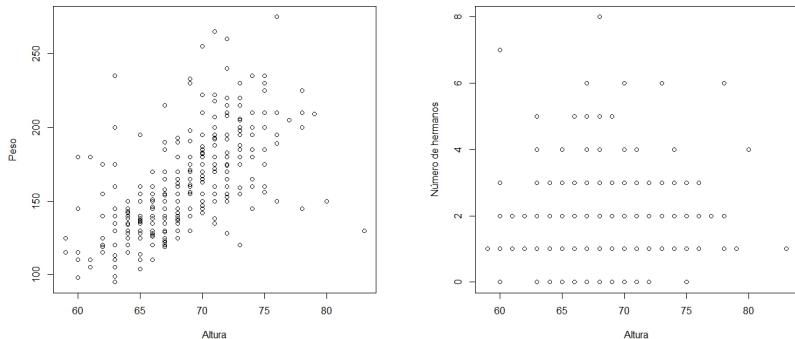


Figura: Diagramas de dispersión

Claramente entre Altura y Número de hermanos no hay ningún patrón, en cambio el otro diagrama sugiere que a mayor Altura mayor Peso.

Medidas de resumen

Consideremos una variable X con codominio Δ tal que

$\Delta \subseteq \mathbb{R}$ o Δ es un conjunto de categorías.

Un estadístico es una función

$$T : \Delta^n \rightarrow \mathbb{R}$$

que permite resumir el comportamiento de los datos.

Medidas de resumen para datos categóricos

- Las frecuencias de cada categoría o nivel de la variable.
- La moda (o modas): el valor (o valores) más frecuente (s).

Ejemplo

Para los datos de la variable Fuma

<i>Fuma</i>	
No	Si
319	43

la moda es No.

Medidas de resumen para datos cuantitativos

La función de distribución empírica correspondiente a una muestra x_1, \dots, x_n de una variable cuantitativa se define como:

$$\forall x \in \mathbb{R} : \quad F_n^*(x) = \frac{1}{n} \# \{i : x_i \leq x\} .$$

Las medidas de resumen (para datos cuantitativos) se pueden dividir en:

- medidas de posición
- medidas de dispersión

Medida de posición central

Representa el “centro” de la muestra ordenada (Bianco, 2019).

- Media o promedio muestral:

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

- Dados los datos ordenados

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(k)} \leq x_{(k+1)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}$$

se definen:

Mediana muestral

$$\tilde{x}_n = \begin{cases} x_{(k+1)} & \text{si } n = 2k + 1 \\ \frac{x_{(k)} + x_{(k+1)}}{2} & \text{si } n = 2k \end{cases}$$

También se denota como $\text{med}_{1 \leq i \leq n}(x_i)$.

Media α -podada muestral. Si $0 \leq \alpha < 1/2$ y $m = [n\alpha]$:

$$\bar{x}_{\alpha,n} = \frac{1}{n - 2m} \sum_{i=m+1}^{n-m} x_{(i)}$$

Ejemplo “toy o juguete”

Consideremos los datos:

4.4, 5.2, 4.2, 6.6, 5.3, 4.2, 5.5, 5.7, 5.6, 4.7

- Media Muestral:

$$\bar{x}_{10} = \frac{51.4}{10} = 5.14$$

- Ordenamos los datos:

4.2, 4.2, 4.4, 4.7, 5.2, 5.3, 5.5, 5.6, 5.7, 6.6

Mediana muestral: $10 = 2 \times 5 \Rightarrow k = 5$

$$\tilde{x}_{10} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{5.2 + 5.3}{2} = 5.25$$

- Media 10% podada muestral ($\alpha = 0.1$):

$$\bar{x}_{0.1} = \frac{x_{(2)} + \dots + x_{(9)}}{8} = 16$$

Resumen de Altura

Consideremos los datos de la variable Altura (hay 7 NA o valores perdidos):

```
> mean(u$Altura); median(u$Altura); mean(u$Altura, trim=0.3)
[1] 68.42254
[1] 68
[1] 68.34266
```

¿Cómo interpretamos estos estadísticos en términos del contexto de la encuesta?:

- las alturas de los 355 estudiantes - que informaron el dato - están alrededor de 68.4 pulgadas
- la mitad de los estudiantes tienen una altura igual o inferior a 68 pulgadas.

¿Qué significa “centro”?

Los siguientes histogramas corresponden a dos muestras de datos diferentes.

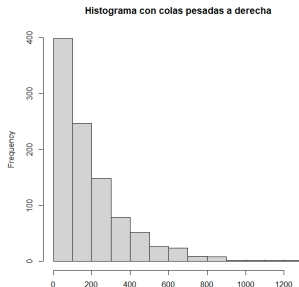
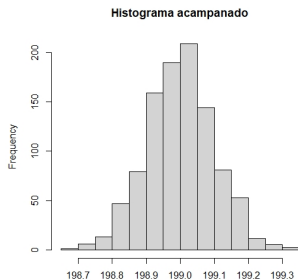


Figura: De izquierda a derecha: $\bar{x} = 199 \approx \tilde{x} = 198.99$; $\bar{x} = 187.4 > \tilde{x} = 125.8$

Medidas de resumen: el cuantil

Sea F_n^* la función de distribución empírica asociada a x_1, \dots, x_n y sean $x_{(1)} \leq \dots \leq x_{(n)}$ los valores de la muestra ordenados (estadísticos de orden).

A $x_{(i)}$ se le denomina el i -ésimo estadístico de orden, $i = 1, \dots, n$.

Dado $\alpha \in (0, 1)$, el cuantil muestral α se define como⁴

$$x_\alpha^* = (1 - h)x_{(k)} + hx_{(k+1)} \quad \text{para } \alpha \in [1/2n, 1 - 1/2n],$$

donde k y h son respectivamente la parte entera y la parte fraccionaria de $u = n\alpha + 0.5$; o sea, $k = [u]$ y $h = u - [u]$.

El cuantil x_α^* deja una proporción α de las observaciones x_1, \dots, x_n por debajo de él.

⁴

ver Maronna, 2021 y la documentación de R
<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile>

Medidas de resumen: cuartiles

Hay tres percentiles muy utilizados:

- El cuantil $\alpha = 0.25$ llamado primer cuartil y denotado por Q_1 .
- El cuantil $\alpha = 0.5$ que es la mediana denotado por Q_2 ⁵.
- El cuantil $\alpha = 0.75$ llamado tercer cuartil y denotado por Q_3 .

Un resumen de los 5 números consiste de los tres cuartiles, el mínimo y el máximo.

⁵también denotado por \tilde{x} , como antes.

Cuartiles y cuantiles para el puntaje SAT Lengua

Basados en

```
> quantile(base_datos$SATLengua, prob=.10)
10%
500
> summary( base_datos$SATLengua)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
390.0  550.0   600.0   594.2  640.0   800.0
```

podemos establecer que

- El 10 por ciento de los estudiantes tiene un puntaje SAT en Lengua igual o inferior a 500.
- El 25% del grupo de estudiantes tiene un puntaje en dicha asignatura menor o igual a 550, la mitad igual o menos que 600, el 75% posee un puntaje igual o inferior a 640. El puntaje mínimo es 390 y el máximo es 800.

Medidas de dispersión

Capturan la variabilidad presente en los datos.

- Rango:

$$\text{rango} = x_{(n)} - x_{(1)}$$

- Varianza:

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

- Desvío muestral:

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

- Coeficiente de variación:

$$cv_n = \frac{\sqrt{s_n}}{\bar{x}_n} \times 100\%$$

Medidas de dispersión

Medidas de dispersión

- Mediana de Desvíos Absolutos:

$$\text{mad} = 1.4826 \text{ med}_{1 \leq i \leq n} (|x_i - \tilde{x}|)$$

- Distancia Intercuartil:

$$d_I = Q_3 - Q_1$$

Nota: asumiendo normalidad, para que s , la mad y d_I “identifiquen” a σ se introduce el factor de corrección $d_I/1.349$.

Medidas de sesgo o asimetría

- Medida de sesgo

$$\bar{\eta}_3 = \frac{\hat{\mu}_3}{\sigma^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

- Medida de sesgo de Fisher

$$\hat{\eta}_3 = \frac{\frac{n}{(n-1)(n-2)} \sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

donde s^2 es la varianza muestral.

Si los coeficientes son cercanos a cero indican simetría, caso contrario la presencia de sesgo. Ver skewness para su cómputo en R.

Medidas de dispersión para Altura

A partir de

```
Altura<- base_datos[!is.na(base_datos$Altura), ]$Altura # Removemos los NA
> rango
[1] 24
> var(Altura)
[1] 16.63452
> sd(Altura)
[1] 4.078544
> cv=(sd(Altura)/mean(Altura))*100; cv
[1] 5.960819
> mad(Altura)
[1] 4.4478
IQR<- quantile(Altura,.75)-quantile(Altura,.25);IQR # ó IQR(Altura)
75%
6
```

concluimos que:

- La diferencias entre las alturas máxima y mínima es de 24 pulgadas.
- En promedio las alturas de los estudiantes distan del promedio en 4.08 pulgadas.
- El porcentaje de variabilidad de la muestra de alturas relativa a su media es del 6%.

Diagrama de caja o boxplot.

Como en Bianco (2019), para construir un diagrama de caja:

- Representamos una escala vertical u horizontal.
- Dibujamos una caja cuyos extremos son los cuartiles Q_3 y Q_1 y dentro de ella un segmento que corresponde a la mediana.
- A partir de cada extremo dibujamos un segmento, llamado bigote, hasta el dato más alejado que está a lo sumo 1.5 veces el rango intercuartílico desde cada uno de los extremos de la caja.
- Marcamos con \circ a aquellos datos que están a más de $1.5d_I$ de cada extremo de la caja.

Diagrama de caja: ejemplo juguete.

En R computamos los estadísticos de dispersión para una muestra juguete:

```
toy
-2 10  0  0 -4  3  3  1  4 15
  Q3<- quantile(toy,.75);  Q2<- quantile(toy,.5); Q1<- quantile(toy,.25)
Q1;Q2;Q3
25%
0
50%
2
75%
3.75
max(toy); min(toy)
15
4
1.5*IQR(toy)+Q3
9.375
Q1- 1.5*IQR(toy)
-5.625
```

Diagrama de caja: ejemplo juguete.

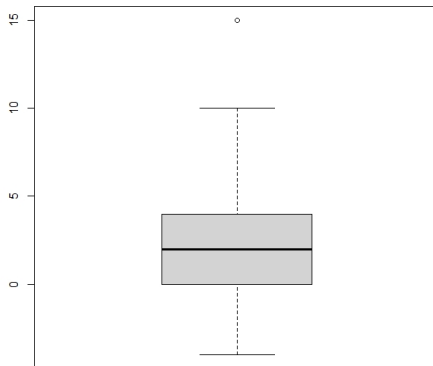


Figura: Diagrama de caja para Altura

4. Aprendizaje estadístico

Definiciones “elementales”

Para Hastie (2015) ⁶:

- El **aprendizaje de máquinas** (machine learning) construye algoritmos que pueden aprender de los datos.
- El **aprendizaje estadístico** (statistical learning) es una rama de la estadística aplicada que emergió en respuesta al aprendizaje de máquinas, enfatizando en los modelos estadísticos y en la evaluación de la incertidumbre.
- La **ciencia de datos** es la extracción de conocimiento de los datos, utilizando ideas de la matemática, la estadística, el aprendizaje de máquinas, la ciencia de la computación, ingeniería...

⁶ https://web.stanford.edu/~hastie/TALKS/SLBD_new.pdf

Algunos problemas más complejos

En estadística además de describir y resumir interesa **inferir** y **predecir**.

Los problemas de aprendizaje se clasifican en:

- Supervisados: el objetivo es predecir el valor de una variable de salida (respuesta) en función de un número de variables de ingreso (inputs)
- No supervisados: no hay variable de salida y el objetivo es describir las asociaciones y patrones de un conjunto de variables.

Algunos problemas más complejos

Algunos problemas del aprendizaje estadístico: ⁷

- Predecir las ventas de una empresa sobre la base de tres presupuestos en medios
- Estimar el riesgo que un paciente hospitalizado por un infarto tenga un segundo infarto. La predicción se basa en datos demográficos, dieta y mediciones clínicas para el paciente.
- Identificar los números de un código postal escrito a mano, a partir de un imagen.
- Estimar la cantidad de glucosa en sangre de una persona diabética, a partir de la absorción del espectro de la sangre de la persona.
- Identificar los factores de riesgo del cáncer de próstata basado en variables clínicas.

⁷Hastie, T., Tibshirani, R., Friedman, J. 2016. The element of Statistical Learning. Springer.

Estudio de mercado

En este estudio contamos con las siguientes variables registradas:⁸

- *Variables de entrada (inputs), características (features) o predictores:* “presupuesto” en TV, Radio y Newspaper.
- *Variable output, de salida, respuesta o variable dependiente:* “valor de venta” o sales.

Estamos frente a un problema de aprendizaje supervisado con variable respuesta sales.

⁸James, G.; Witten, D.; Hastie T.; Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. Springer.

Estudio de mercado

Datos "Advertising"

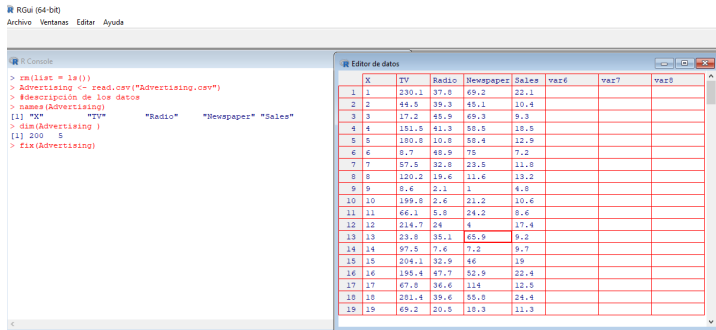


Figura: $n = 200$ observaciones correspondientes, cada una, a un producto vendido en un mercado. Para cada producto conocemos el valor de venta y la inversión publicitaria en tres diferentes medios: tv, radio y diario impreso (James et al., 2021)

Estudio de mercado

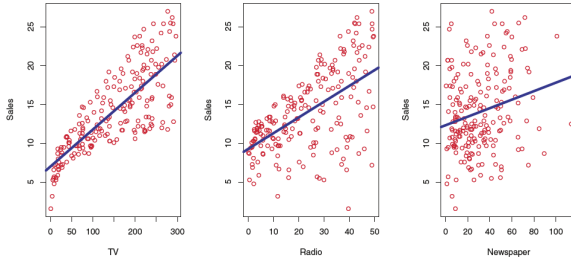


Figura: Datos “Advertising”. Representación de ventas (**sales**) versus inversión según medio publicitario (**TV**, **Radio** y **Newspaper**)

Estudio de mercado

- No se pueden incrementar el valor de venta del producto pero sí se puede controlar el gasto en publicidad.
- Si determinamos que existe una asociación entre publicidad y ventas entonces podríamos recomendarle al vendedor cómo ajustar los montos publicitarios.

Objetivo:

desarrollar un modelo que pueda ser utilizado para predecir ventas sobre la base de tres presupuestos en medios.

Mails spam

Contamos con 4601 mensajes de email en un estudio para predecir, cuando un email es basura (spam o junk) o no.

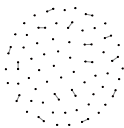
Objetivo:

diseñar un detector (clasificador) de spam que pueda filtrar este tipo de mensajes antes que saturen las cuentas de los usuarios.

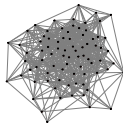
Este es un problema de aprendizaje supervisado, con una variable outcome “email/spam”; es un problema de clasificación.

Genómica

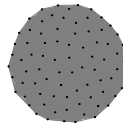
- Base de datos con expresiones génicas de **cáncer de mama**, con $p = 80$ variables (genes) y $n = 133$ pacientes.
- Se asume que (X_1, \dots, X_p) tiene una distribución multivariada con matriz de covarianza Σ .
- Los elementos fuera de la diagonal de las entradas de Σ^{-1} caracterizan la asociación condicional entre dos variables dadas las restantes y se representa con un grafo del tipo:



(a)



(b)



(c)

Figura: Grafos posibles para las 80 variables. Cada lado representa asociación condicional entre dos genes.

Objetivos:

- A partir de los datos “descubrir” (estimar) el grafo real.
- Con el grafo hallado a partir de los datos entrenar un clasificador que permita clasificar si un paciente bajo quimioterapia preoperativa se encuentra en estado pCR (erradicación completa del cancer) o bien en RD (enfermedad residual, el cáncer persiste)

Este es un problema de aprendizaje supervisado con una variable respuesta (estado del paciente) y X_1, \dots, X_p características.

Segmentación de mercado

En un estudio de segmentación de mercado registramos código postal, ingreso familiar, hábitos de compra, etc.

Nos planteamos la hipótesis de que los clientes caen en diferentes grupos, como por ejemplo de altos, medios y bajos gastos.

Pero, si no se cuenta con la información de gastos entonces estamos frente a un problema de clasificación no supervisada (no hay una variable respuesta).

Objetivo:

a partir de las variables registradas, descubrir la existencia de grupos.

Segmentación de mercado

Para un total de 150 observaciones elegimos dos variables y las representamos como en la figura siguiente:⁹

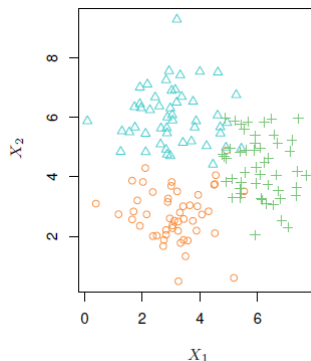


Figura: 150 observaciones de las variables X_1 y X_2 .

Es un típico problema de cluster análisis.

⁹Tomada de James et al., 2021.

El Reemplazante: cluster análisis.

¿Quién podría reemplazar a Nacho Fernández?¹⁰

Un problema en el fútbol es hallar reemplazo para un jugador con el mismo perfil que el que se retira.

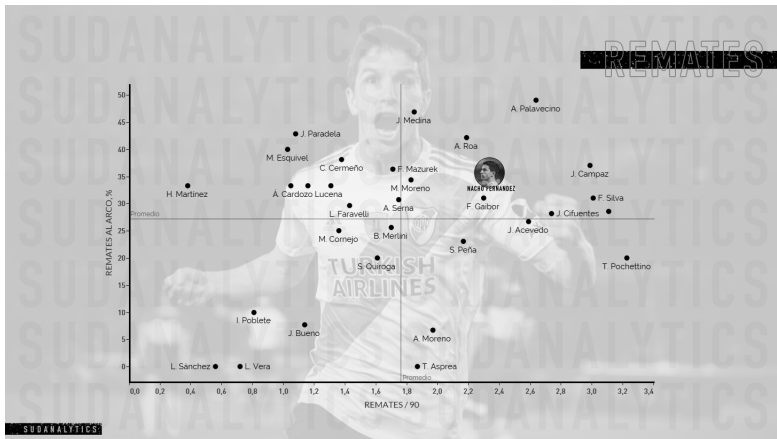


Figura: Distribución de los jugadores de acuerdo a las variables Remate y Asistencias esperadas. Fuente Sudanalytics. Problema propuesto por Matías Bricco, estudiante de Computación de UNRC