Estadística descriptiva

Ángel Berihuete, Carmen Ramos, Juan Antonio García

2018-11-14

Índice general

Prólogo

Este libro es una recopilación de apuntes y presentaciones impartidos en las clases de Estadística descriptiva. No es un libro completo. Al contrario, se está insertando y actualizando material de manera continua. ¡Cualquier sugerencia a través del repo en GitHub será bienvenida!

El libro ha sido escrito en R-Markdown utilizando el paquete `bookdown` y está disponible en el repositorio Github: AngelBerihuete/EstNavBook.

Capítulo 1

Introducción al análisis de datos

En las siguientes secciones iremos desarrollando los elementos básicos de un análisis exploratorio, calculando medidas que nos ayuden a interpretar de forma adecuada el conjunto de datos analizados. Para ello utilizaremos varios conjuntos de datos que podrán ser descargados del repositorio en GitHub.

Planteamos el siguiente problema sencillo: los precios (en euros) de un conjunto de botellas de vino en cierta sección de un centro comercial son los siguientes

42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53 53 53 62 62 62 62 62 62 62 62 62 63 63 63 63 63

Podríamos hacernos las siguientes preguntas:

1. ¿Cómo preparo los datos para su análisis?
2. ¿Qué representación gráfica hago de los datos?
3. ¿Puedo elegir uno o varios valores que representen con garantías al conjunto de los datos?
4. ¿Cómo puedo sistematizar el análisis de un conjunto de datos?

Un **análisis exploratorio de datos** implica el uso de los elementos numéricos y gráficos que describan lo más exhaustivamente posible un conjunto de datos. Además, deberá poner de manifiesto los aspectos más relevantes de la distribución de los datos: regularidades, datos atípicos, especificidades, relaciones, etc.

En este capítulo nos marcamos los siguientes objetivos:

- Ser capaces de identificar el tipo de variable estadística en estudio.
- Decidir qué tipo de medidas y gráficos serán los adecuados para resumir los diferentes tipos de variables estadísticas.
- Entender que **la media no siempre es la adecuada para resumir los datos**.
- Ser capaces de comparar diferentes medidas entre diferentes conjuntos de datos.
- Diferenciar entre **análisis exploratorio** de datos e **inferencia estadística**.

1.1 Variables estadísticas

Las variables a analizar pueden ser de diferentes tipos:

- **Cualitativas** o **factores**: son variables no expresables numéricamente.

- **Cuantitativas**: pueden ser expresadas numéricamente. Las variables cuantitativas se subdividen en:

1. **Cuantitativas Discretas**, si el conjunto de sus posibles valores tiene cardinal finito o infinito numerable. (Ejemplo: "Número de trabajadores en una bodega")
2. **Cuantitativas Continuas**, si pueden tomar los infinitos valores de un intervalo.

A veces, por cuestiones prácticas, conviene discretizar las variables cuantitativas continuas. Por ejemplo la variable "Antigüedad del vino en una bota, medida en años".

1.2 Distribuciones de frecuencias

A partir de un conjunto de datos queremos clasificarlos de modo que la información contenida en ellos quede presentada de forma clara, concisa y ordenada.

Si representamos por N al número total de datos, entre los que consideraremos que hay k valores distintos x_1, x_2,..., x_k (que en el caso de las variables cuantitativas se presentarán ordenados de menor a mayor), se conoce como **frecuencia**:

- **Absoluta** del valor x_i, al número de veces que se presenta dicho valor en el conjunto de datos. Se representa por n_i.
- **Absoluta acumulada** del valor x_i, al número de datos que hay iguales o inferiores a x_i. Se representa por N_i.
- **Relativa** del valor x_i, al cociente $\frac{n_i}{N}$. Se representa por f_i.
- **Relativa acumulada** del valor x_i, al cociente $\frac{N_i}{N}$. Se representa por F_i.

Ejemplo 1.1 (El problema del fluoruro)**.** Para realizar un estudio sobre el contenido en ion fluoruro en vinos embotellados de mayor comercialización en Canarias se procedió al análisis de 79 vinos embotellados de los cuales 50 son de la comunidad de Canarias y 29 de la Península. La siguiente tabla muestra los 10 primeros registros:

denominacion

tipo

conc_fluor

procedencia

precio

El Hierro

Tinto

0.13

Canarias

42

Ycoden Daute Isora

Tinto

0.07

Canarias

42

Ycoden Daute Isora

Tinto

0.10

Canarias

42

Ycoden Daute Isora

Tinto

0.11

Canarias

42

La Palma

Tinto

0.26

Canarias

42

Lanzarote

Tinto

0.12

Canarias

42

La tabla de frecuencias del factor tipo que tiene niveles Tinto, Blanco, Rosado

x_i

n_i

N_i

f_i

F_i

Blanco

27

27

0.3418

0.3418

Rosado

18

45

0.2278

0.5696

Tinto

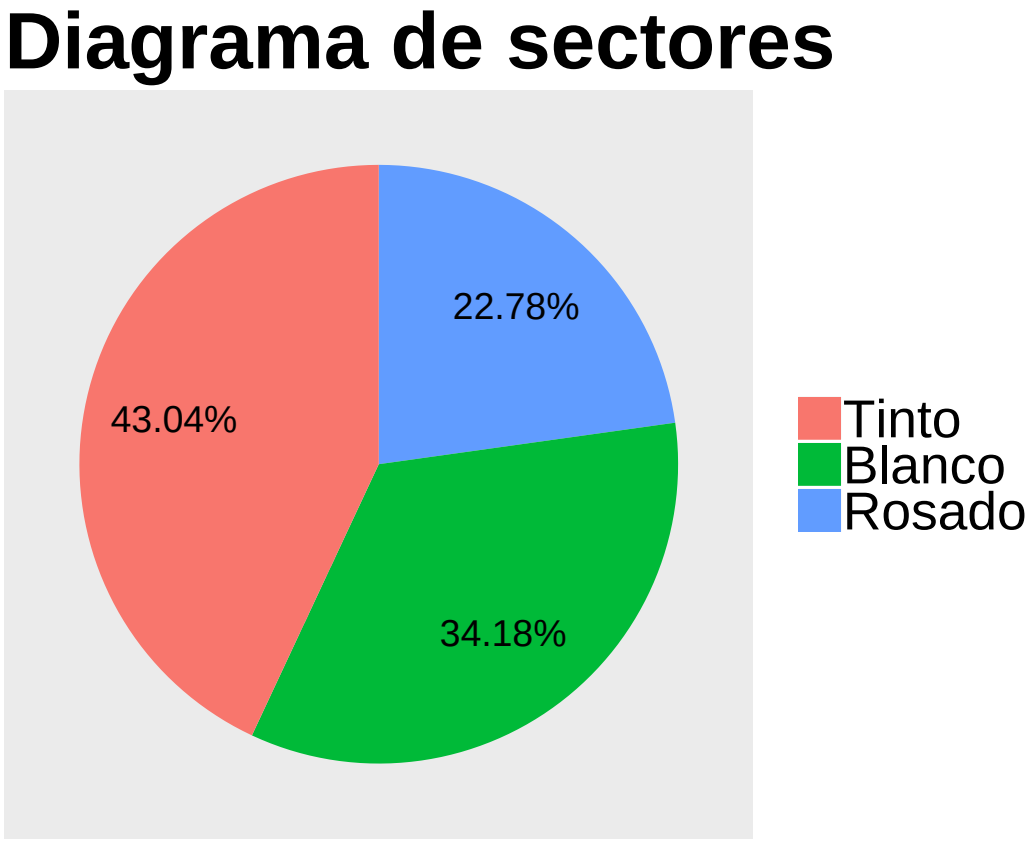
34

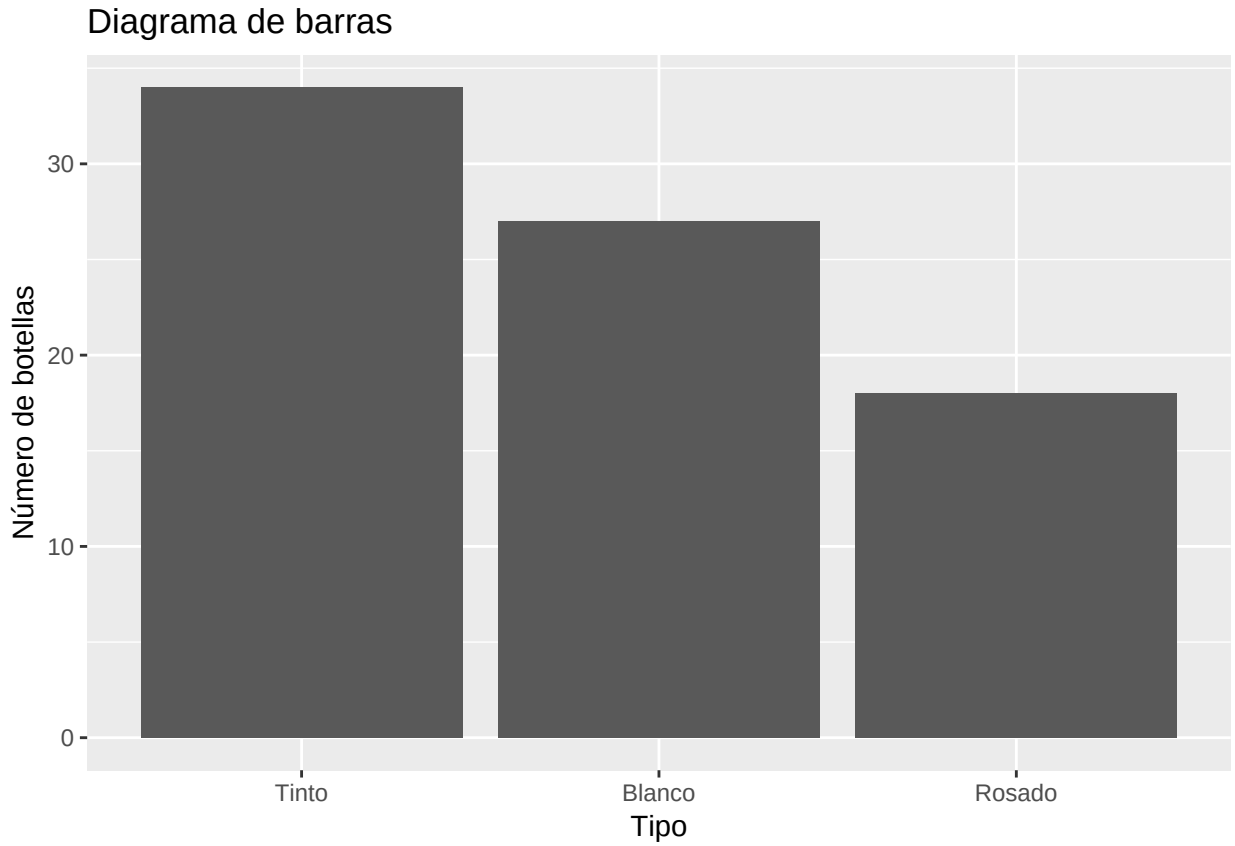
79

0.4304

1.0000

siendo el gráfico más adecuado para esta variable el **diagrama de sectores** o el **diagrama de barras**.





1.2.1 Variables cuantitativas discretas

Tabla de frecuencias de la variable precio

42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 52 53 53 53 53 53 53 53 53 53 53 62 62 62 62 62 62 62 62 62 63 63 63 63 63

x_i	n_i	N_i	f_i	F_i
42	20			
20				

0.2532

0.2532

43

15

35

0.1899

0.4430

52

20

55

0.2532

0.6962

53

10

65

0.1266

0.8228

62

9

74

0.1139

0.9367

63

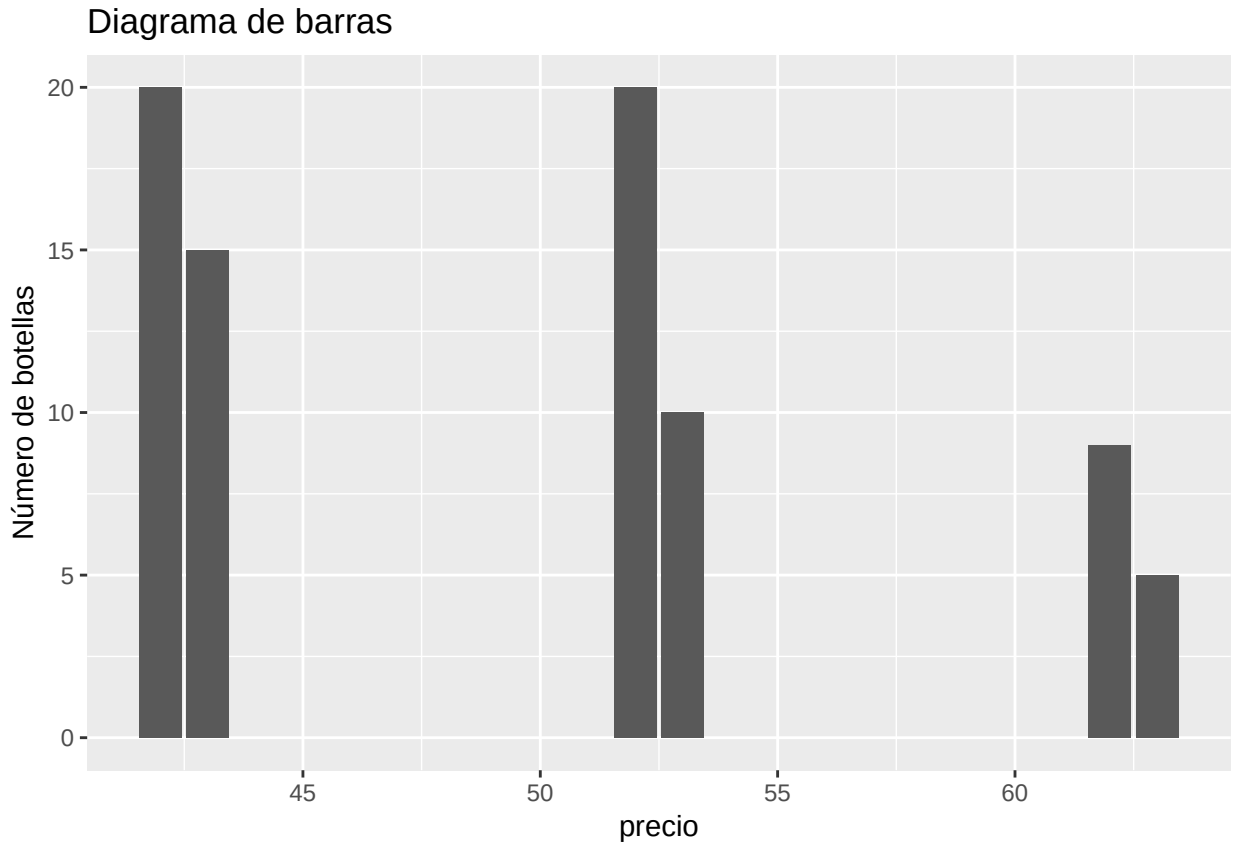
5

79

0.0633

1.0000

y el gráfico adecuado para este tipo de variable será el **diagrama de barras**:



x_i	n_i	N_i	f_i	F_i
42	20	20	0.2532	0.2532
43	15	35		

0.1899
0.4430
52
20
55
0.2532
0.6962
53
10
65
0.1266
0.8228
62
9
74
0.1139
0.9367
63
5
79
0.0633
1.0000

1.2.2 Variables cuantitativas continuas

La tabla de frecuencias de la variable cuantitativa continua conc_fluor

$(l_{i-1}, l_i]$

n_i

N_i

f_i

F_i

(0,0.2]

45

45

0.5696

0.5696

(0.2,0.3]

22

67

0.2785

0.8481

(0.3,0.4]

7

74

0.0886

0.9367

(0.4,0.6]

5

79

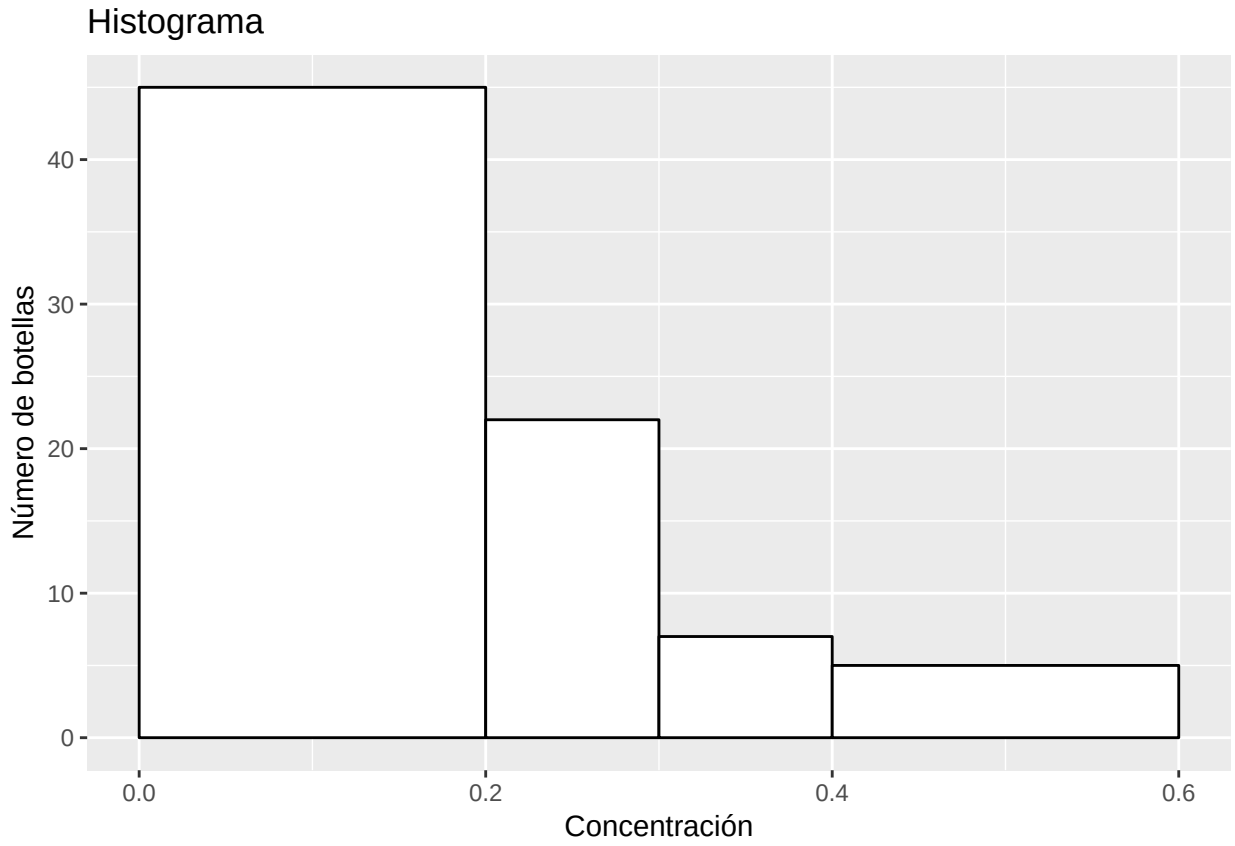
0.0633

1.0000

Obsérvese que hemos agrupado en intervalos los valores

0.13 0.07 0.10 0.11 0.26 0.12 0.20 0.17 0.08 0.14 0.17 0.13 0.16 0.16 0.15 0.28 0.18 0.20 0.09 0.13 0.13 0.06 0.13 0.22 0.12 0.14 0.12 0.17 0.12 0.50 0.15 0.22 0.13 0.20 0.18 0.13 0.10 0.10 0.18 0.09 0.15 0.15 0.08 0.11 0.16 0.17 0.10 0.09 0.15 0.18 0.25 0.26 0.27 0.28 0.28 0.29 0.27 0.27 0.28 0.44 0.37 0.46 0.24 0.28 0.33 0.37 0.46 0.22 0.28 0.38 0.29 0.27 0.29 0.29 0.36 0.23 0.32 0.36 0.45

El gráfico adecuado para este tipo de variable es el **histograma**:



¡Cuidado! Longitud de intervalos distinta. Tendremos que utilizar la **densidad** (h_i) para representar el histograma. Ahora la tabla de frecuencias de la variable cuantitativa continua conc_fluor

$(l_{i-1}, l_i]$

n_i

N_i

f_i

F_i

h_i

(0,0.2]

45

45

0.5696

0.5696

2.8481

(0.2,0.3]

22

67

0.2785

0.8481

2.7848

(0.3,0.4]

7

74

0.0886

0.9367

0.8861

(0.4,0.6]

5

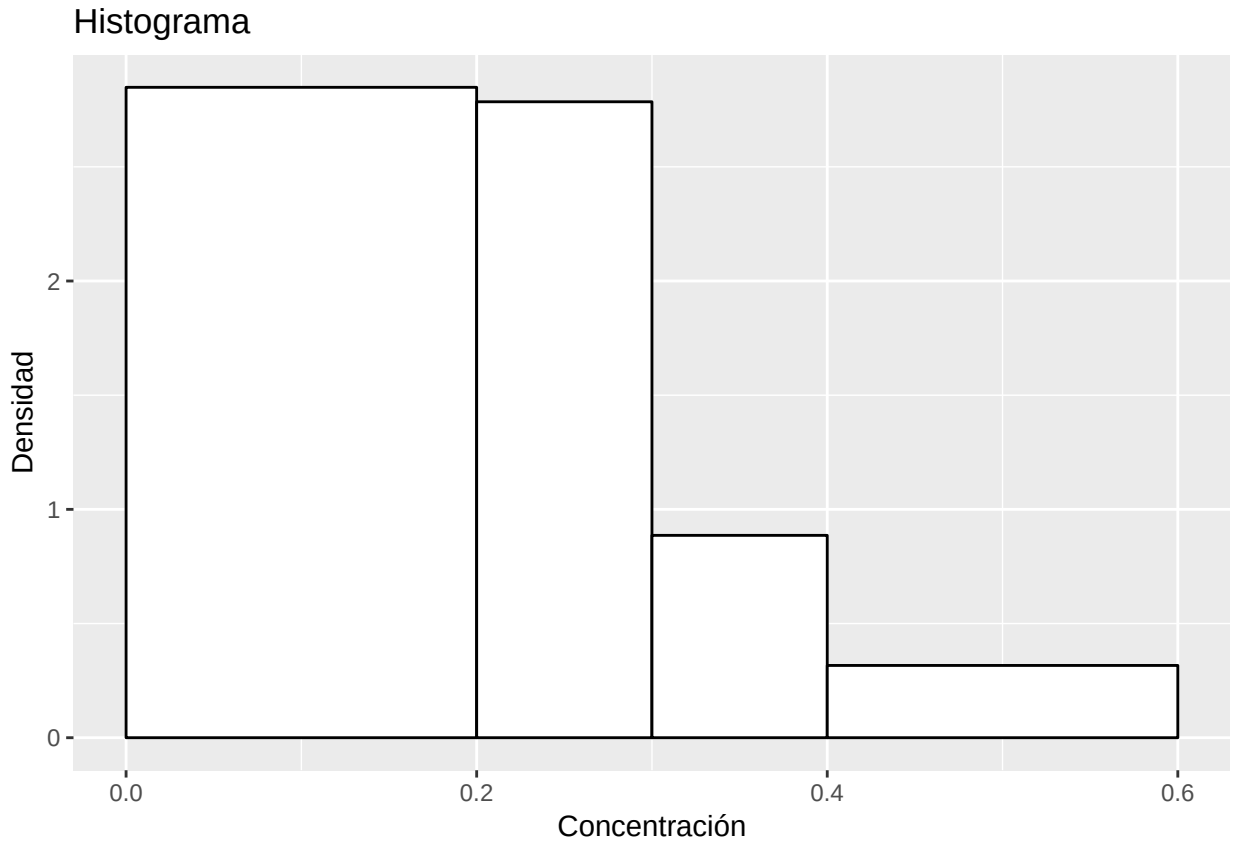
79

0.0633

1.0000

0.3165

Siendo $h_i = \frac{f_i}{a_i}$



1.2.3 ¿Qué número de intervalos y qué amplitud es la adecuada?

No hay una fórmula con la que obtener un valor óptimo del número de intervalos y su amplitud. Existen muchas dependiendo de las hipótesis que hagamos en nuestro problema. Algunas de las más utilizadas si tenemos N observaciones:

- $k = \sqrt{N}$ lo usa de forma predefinida el programa Excel.
- $k = \lceil \log_2 N \rceil + 1$ lo usa de forma predefinida el programa R. Da resultados pobres si $n < 30$ (Nota: $\lceil 2.4 \rceil = 3$).

Una vez tenemos el número de intervalos, basta calcular su amplitud mediante

$$a = \frac{\max x - \min x}{k}$$

1.2.4 ¿Pueden manipular nuestra opinión con los gráficos?

En general hay atentos a que todos los elementos de un gráfico estén representados adecuadamente. Algunos ejemplos de gráficos erróneos en los médios de comunicación podrían ser los siguientes:

1. Los ejes no tiene la escala adecuada o están cortados
2. No se mantiene la escala en todo el gráfico

Se pueden encontrar todos estos gráficos y su expliacción en el diario El País así como en la cuenta de Twitter de Kiko Llaneras

Figura 1.1: Fuente: Verne (El País)

Figura 1.2: Fuente: Verne (El País)

Figura 1.3: Fuente: Verne (El País)

1.3 Medidas de posición central

1.3.1 Medias

- Una **media** es una medida de representación central que necesariamente debe cumplir tres requisitos:

1. Para su obtención deben utilizarse todas las observaciones.
2. Debe ser un valor comprendido entre el menor y el mayor de los valores de la distribución.
3. Debe venir expresada en la misma unidad que los datos.

La media aritmética se calcula mediante la fórmula

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \sum_{i=1}^{r} x_i f_i$$

Por ejemplo, la media aritmética de la variable cuantitativa discreta

$$x_i$$

$$n_i$$

$$N_i$$

$$f_i$$

Figura 1.4: Fuente: Verne (El País)

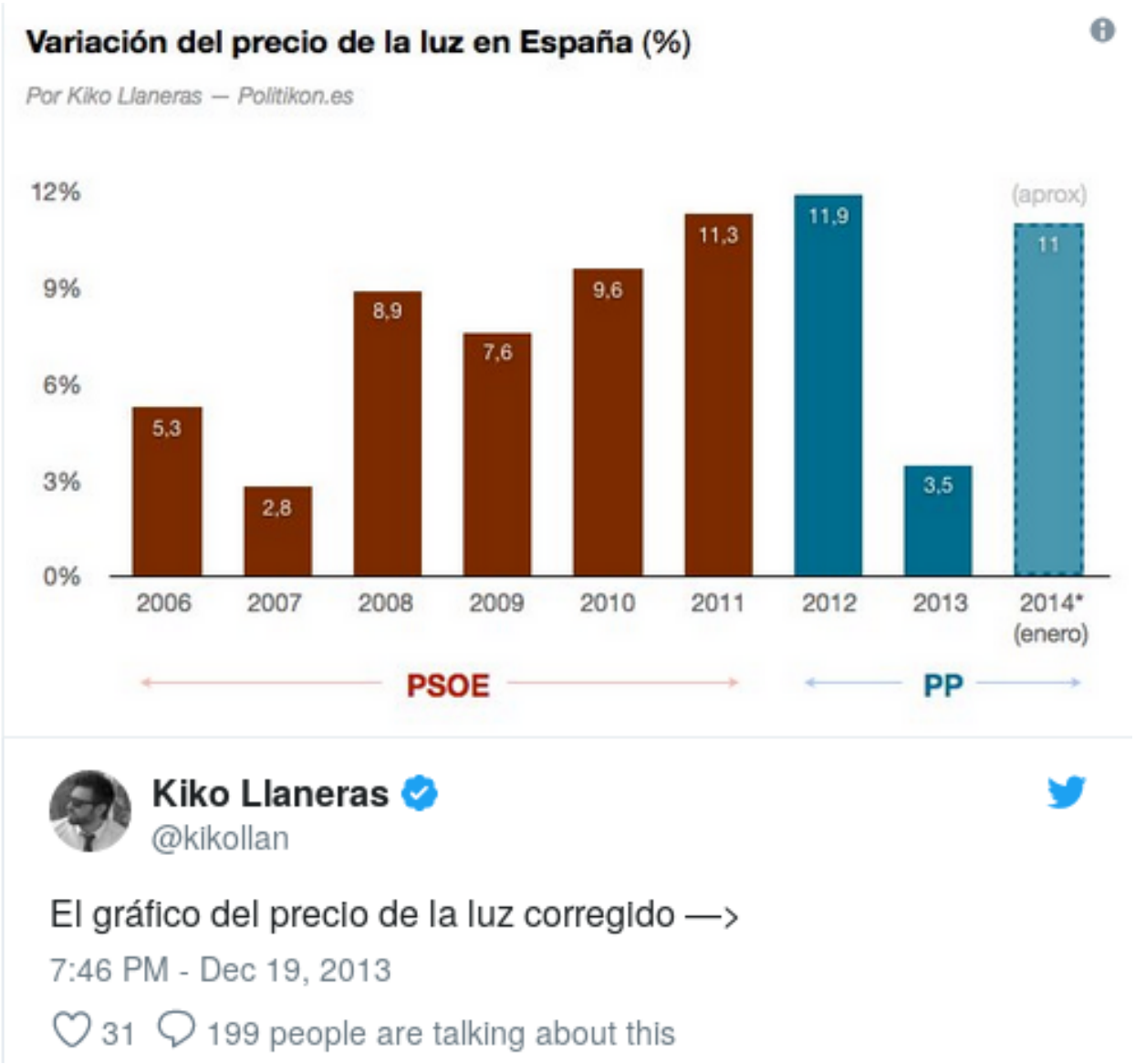


Figura 1.5: Fuente: Verne (El País)

F_i

42

20

20

0.2532

0.2532

43

15

35

0.1899

0.4430

52

20

55

0.2532

0.6962

53

10

65

0.1266

0.8228

62

9

74

0.1139

0.9367

63

5

79

0.0633

1.0000

$$\bar{x} = \frac{\sum_{i=1}^{r} x_i n_i}{N} = \frac{42 \cdot 20 + 43 \cdot 15 + \cdots + 63 \cdot 5}{20 + 15 \cdots + 5} = \frac{3928}{79} = 49.7215$$

En el caso de **variables continuas** agrupadas en intervalos tendremos que obtener primero la **marca de clase**

$$x_i = \frac{l_{-1} + l_i}{2}$$

$(l_{i-1}, l_i]$

n_i

N_i

f_i

F_i

h_i

x_i

(0,0.2]
45
45
0.5696
0.5696
2.8481
0.10
(0.2,0.3]
22
67
0.2785
0.8481
2.7848
0.25
(0.3,0.4]
7
74
0.0886

0.9367

0.8861

0.35

(0.4,0.6]

5

79

0.0633

1.0000

0.3165

0.50

$$\bar{x} = \frac{0.1 \cdot 45 + 0.25 \cdot 22 + \cdots + 0.5 \cdot 5}{45 + 22 \cdots + 5} = \frac{14.95}{79} = 0.1892$$

Otras medias que suelen utilizarse son:

- **media geométrica**:
$$\bar{x}_g = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_r^{n_k}}$$
- **media armónica**:
$$\bar{x}_a = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$
- **media ponderada**: Se asigna a cada valor x_i un peso w_i que depende de la importancia relativa de cada uno de estos valores bajo algún criterio:
$$\bar{x}_p = \frac{\sum_{i=1}^{r} n_i w_i x_i}{\sum_{i=1}^{r} n_i w_i}$$

1.3.2 Propiedades de la media

1. Se cumple que $\sum_{i=1}^{r}(x_i - \bar{x})n_i = 0$
2. Dada una transformación lineal $Y = aX + b$, se cumple que $\bar{y} = a\bar{x} + b$.

Por ejemplo, si sabemos que la media de temperaturas en grados Celsius de cierta región es $\bar{x} = 15°\text{C}$, ¿qué media en grados Farenheit tendrá dicha región?

$$\bar{y} = 1.8 \cdot 15 + 32 = 59°\text{F},$$

porque el paso de una temperatura a otra se hace mediante la transformación $Y = 1.8X + 32$

3. La media es el valor ϕ que hace mínima la expresión:
$$\sum_{i=1}^{r}(x_i - \phi)^2 n_i$$

1.3.3 La mediana

La **mediana**, M_e, es un valor que, **una vez ordenados los datos**, deja la mitad de las observaciones a la izquierda de él y la otra mitad a la derecha.

El cálculo de la mediana en el caso de las **variables discretas** se realiza:

1. Si N es impar, entonces
$$M_e = x_{(\frac{n+1}{2})}$$
2. Si N es par, entonces
$$M_e = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Por ejemplo, en la siguiente tabla de frecuencias

x_i	n_i	N_i	f_i	F_i
42				
20				
20				
0.2532				
0.2532				
43				
15				
35				
0.1899				
0.4430				
52				
20				
55				
0.2532				
0.6962				
53				
10				

65

0.1266

0.8228

62

9

74

0.1139

0.9367

63

5

79

0.0633

1.0000

como N es impar, entonces la mediana es el valor que ocupa la posición

$$\frac{79+1}{2} = \frac{80}{2} = 40$$

Mirando en la tabla $M_e = 52$

Sin embargo, en la siguiente tabla

x_i

n_i

N_i

f_i

F_i

42

20

20

0.2564

0.2564

43

15

35

0.1923

0.4487

52

20

55

0.2564

0.7051

53

10

65

0.1282

0.8333

62

9

74

0.1154

0.9487

63

4

78

0.0513

1.0000

como N es par, entonces la mediana es la media de los valores que ocupan la posión $\frac{78}{2} = 39$ y $\frac{78}{2} + 1 = 40$. Mirando en la tabla

$$M_e = \frac{52 + 52}{2} = 52$$

La mediana en variables continuas agrupadas en intervalos necesita de una fórmula. Por ejemplo, en el caso de la tabla de frecuencias:

$$(l_{i-1}, l_i]$$

$$n_i$$

$$N_i$$

$$f_i$$

F_i

(0,0.2]

45

45

0.5696

0.5696

(0.2,0.3]

22

67

0.2785

0.8481

(0.3,0.4]

7

74

0.0886

0.9367

(0.4,0.6]

5

79

0.0633

1.0000

El primer N_i que es mayor o igual a $\frac{50 \cdot 79}{100} = 39.5$ es $N_1 = 45$, entonces M_e está dentro del intervalo $(0, 0.2]$, por tanto

$$M_e = l_{i-1} + \frac{\frac{50 \cdot N}{100} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

$$M_e = 0 + \frac{\frac{50 \cdot 79}{100} - 0}{45 - 0} \cdot 0.2 = 0.1755$$

1.3.4 La moda

- La moda absoluta de una distribución es el valor que más veces se repite.
- Además de la moda absoluta, aquellos valores que tengan frecuencia mayor a la de los valores adyacentes serán relativas.

En la distribución 2, 3, 3, 4, 6, 7, 7, 7, 10

- $M_o = 7$ es la moda
- $M_{o_r} = 3$ es una moda relativa

La moda para variables discretas es sencillo. Miramos la frecuencia absoluta:

x_i	n_i	N_i	f_i	F_i
42				
20				
20				
0.2564				
0.2564				
43				
15				
35				
0.1923				
0.4487				
52				
20				
55				
0.2564				
0.7051				
53				
10				
65				
0.1282				
0.8333				
62				
9				
74				
0.1154				
0.9487				

63

4

78

0.0513

1.0000

En este caso, el valor que más se repite es $M_o = 42$

En el caso de variables continuas tendremos que utilizar una fórmula:

$(l_{i-1}, l_i]$

n_i

N_i

f_i

F_i

h_i

x_i

(0,0.2]

45

45

0.5696

0.5696

2.8481

0.10

(0.2,0.3]

22

67

0.2785

0.8481

2.7848

0.25

(0.3,0.4]

7

74

0.0886

0.9367

0.8861

0.35

(0.4,0.6]

5

79

0.0633

1.0000

0.3165

0.50

En lugar de fijarnos en n_i nos fijamos en h_i. Como el mayor h_i es 2.8481 la moda es

$$M_o = l_{i-1} + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} \cdot a_i$$

$$M_o = 0 + \frac{2.7848}{0 + 2.7848} \cdot 0.2 = 0.2$$

1.4 Medidas de dispersión

1.4.1 Varianza y Desviación Típica

- La **varianza**, S^2, y su raíz cuadrada positiva, la **desviación típica**, S, son las medidas de dispersión más importantes.

$S^2 = \frac{\sum_{i=1}^{r}(x_i - \bar{x})^2 n_i}{N} = \frac{\sum_{i=1}^{r} x_i^2 n_i}{N} - \bar{x}^2 \qquad S = +\sqrt{S^2}$

1. La desviación típica viene dada en las mismas unidades que la variable estadística.
2. Tanto S como S^2 son siempre no negativas y valen cero sólo en el caso de que todos los valores coincidan con la media.
3. Los valores atípicos influyen en las dos debido al cálculo de la media.

1.4.2 Propiedades de la varianza

- Si $Y = aX + b$ entonces se verifica $S_Y^2 = a^2 S_X^2$. Por ejemplo

Dada X con $\bar{x} = 12$ y $S_X = 9$, entonces para $Y = 3X - 4$:

$$\bar{y} = 3\bar{x} - 4 = 3 \cdot 12 - 4 = 32$$
$$S_Y = \sqrt{a^2} \cdot S_X = \sqrt{9} \cdot 9 = 27$$

1.4.3 Coeficiente de Variación

$$C_v = \frac{S}{\mid \bar{x} \mid}$$

1. Permite comparar la dispersión de varias distribuciones.
2. Da información sobre la representatividad de la media. De forma orientativa, si el coeficiente de variación es menor o igual a 0.5 diremos que la media es representativa.

1.4.4 Otras medidas de dispersión

- El **rango** es la diferencia entre el mayor y el menor de los valores de la variable.
- La **desviación media**

$$D_m = \frac{\sum_{i=1}^{r} |x_i - \bar{x}| n_i}{n}$$

1.4.5 Desigualdad de Tchebychev (para ampliar)

La desigualdad de Tchebychev proporciona una cota inferior para el porcentaje de observaciones en un determinado intervalo con centro la media de la distribución:

Dada una distribución con $\bar{x} = 25$, y $S = 4$, $[\bar{x} - 3S, \bar{x} + 3S] = [13, 37]$ garantiza la presencia en su interior de, al menos, el 88.88% de la distribución.

1.5 Medidas de posición no central

Se llaman medidas de posición o de orden k a aquellas que dividen a la distribución en k partes, de tal forma que en cada una de esas partes haya el mismo número de elementos. Las más importantes son:

- Los **cuartiles**, $Q_i, i = 1, 2, 3$ que dividen a la distribución en cuatro partes iguales.
- Los **deciles** , $D_i, i = 1, 2, \ldots, 9$ que dividen a la distribución en diez partes iguales.
- Los **percentiles**, $P_i, i = 1, 2, \ldots, 99$ que dividen a la distribución en cien partes iguales.

El cálculo de los percentiles será suficiente para calcular tanto los deciles como los cuartiles. En el caso de variables discretas:

- Si el $k\%$ de N, donde N es el número de datos, es un número entero, r, entonces

$$P_k = \frac{x_r + x_{r+1}}{2}$$

- Si el $k\%$ de N no es un número entero, lo redondeamos al siguiente, r, u entonces

$$P_k = x_r$$

Para el caso de datos agrupados en intervalos utilizaremos la fórmula:

$$P_k = l_{i-1} + \frac{\frac{k \cdot N}{100} - N_{i-1}}{N_i - N_{i-1}} \cdot a_i$$

donde N_i es la primera frecuencia absoluta acumulada que cumple $N_i \geq \frac{k \cdot N}{100}$

1.5.1 Diagrama de caja

Es una representación gráfica de los cuartiles:

- La caja representa el 50% central de la distribución, la línea situada en el interior de la caja es la mediana (algunos programas estadísticos representan también un símbolo que se corresponde con la media).
- La longitud de la caja es el **rango intercuartílico** $R_I = Q_3 - Q_1$
- Los extremos inferiores y superiores de los segmentos encierran los valores "normales".
- Los candidatos a valores anómalos se etiquetan como **atípicos** y se encuentran fuera del intervalo $(Q_1 - 1.5 \cdot R_I, Q_3 + 1.5 \cdot R_I)$:

$(l_{i-1}, l_i]$

n_i

N_i

(0,0.2]

45

45

(0.2,0.3]

22

67

(0.3,0.4]

7

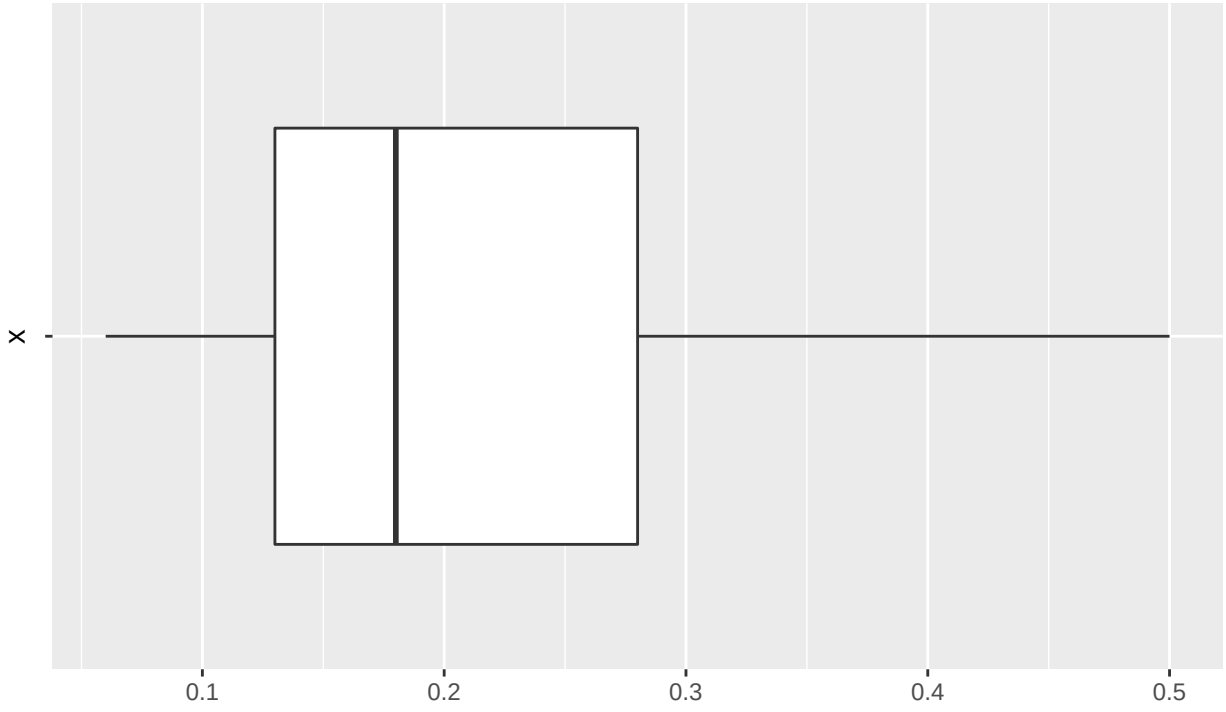
74

(0.4,0.6]

5

79

Los percentiles 25, 50 y 75 de la concentración de fluoruro son 0.13, 0.18, 0.28 respectivamente. Además $1.5 \cdot R_I$ es 0.225



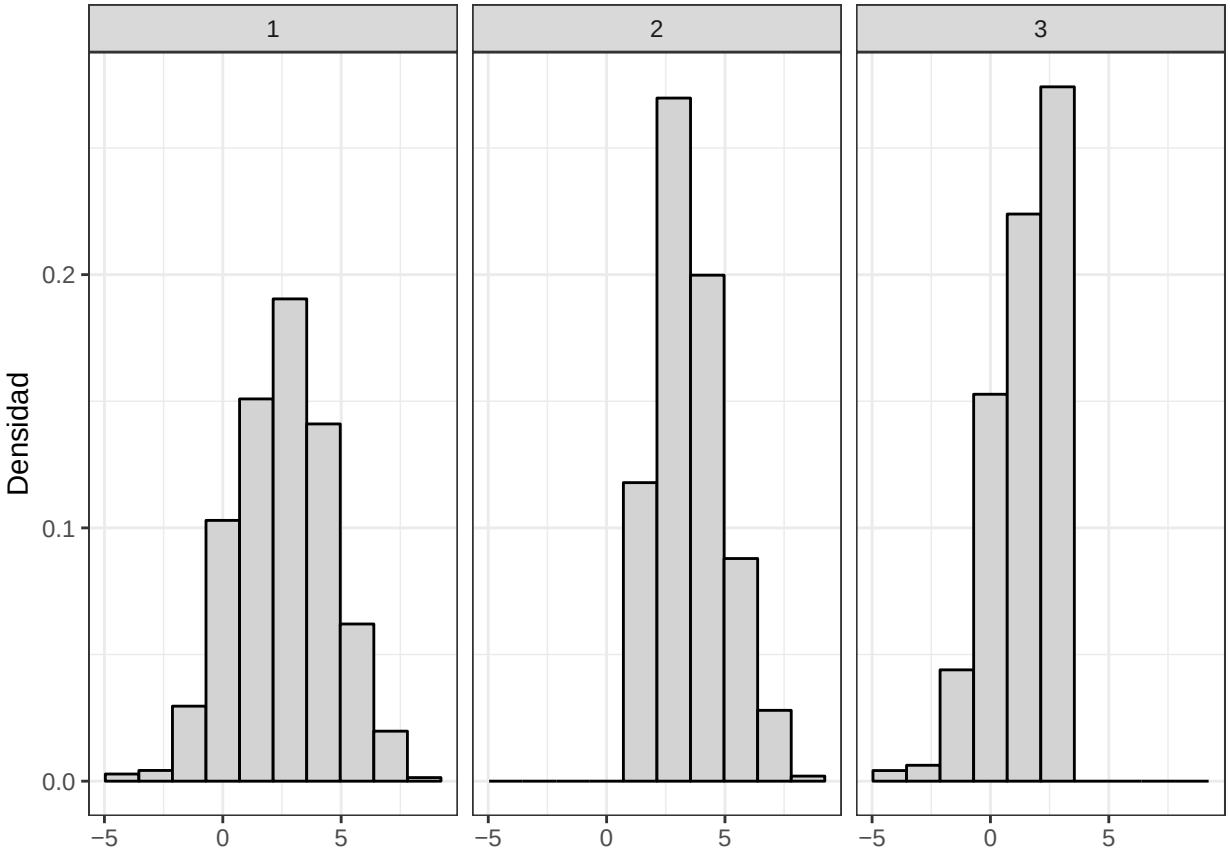
1.6 Medidas de forma

1.6.1 Asimetría

- El **coeficiente de simetría** viene dado por:

$$\gamma_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N \cdot S^3}$$

Observe que cuando la distribución es simétrica coinciden la media y la mediana. Si la distribución tiene además forma de campana, la media y la mediana se aproximan a la moda.



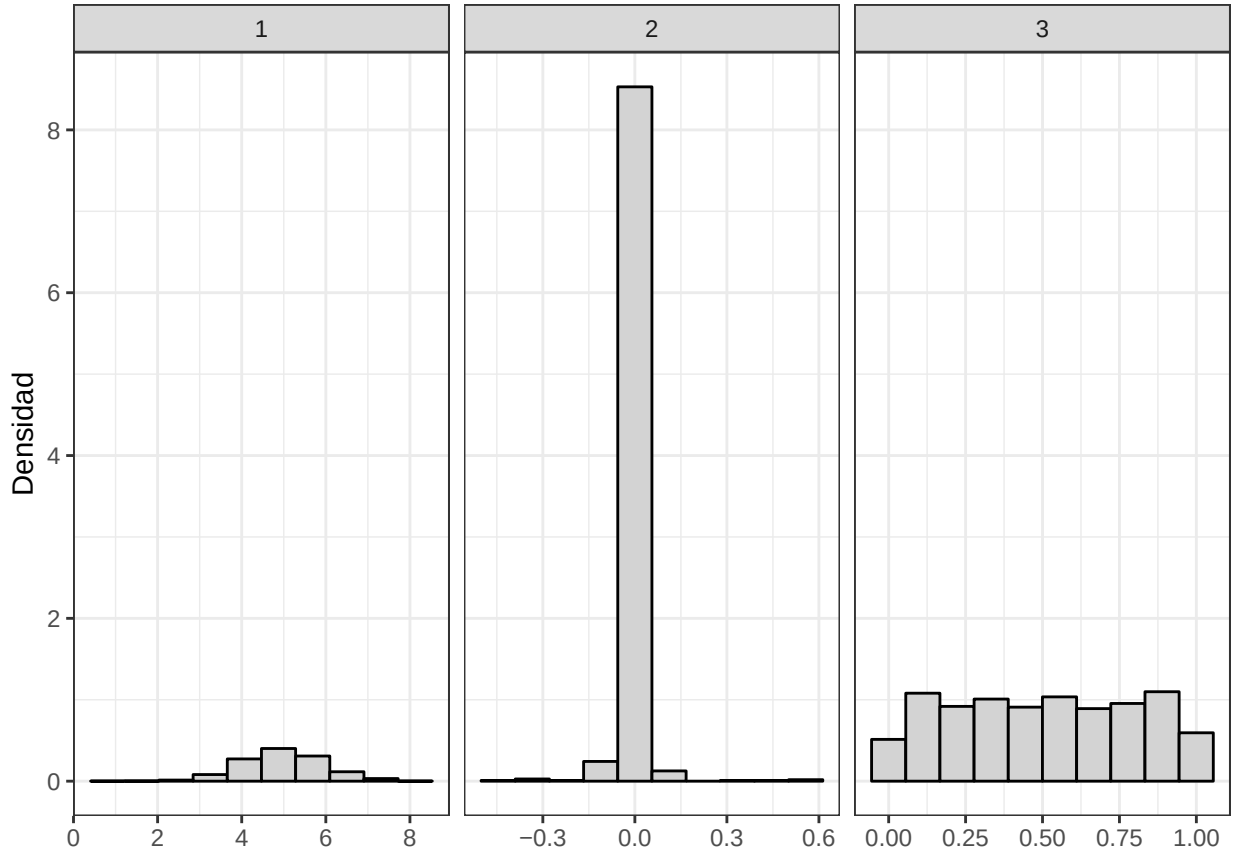
En estos casos, la asimetría es, de izquierda a derecha, -0.0109214, 0.7214576, -0.8380423, respectivamente.

1.6.2 Curtosis

Para el cálculo de la curtosis utilizaremos

$$\gamma_2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N \cdot S^4} - 3$$

- $\gamma_2 = 0$, distribución **mesocúrtica**
- $\gamma_2 < 0$, distribución **platicúrtica**
- $\gamma_2 > 0$, distribución **leptocúrtica**



En las distribuciones anteriores, y de izquierda a derecha, las curtosis tienen un valor de

0.2082964, 79.4672652, -0.0070202, respectivamente.

Puede verse una aplicación de cómo el histograma y el diagrama de caja tienen un comportamiento similar ante diferentes formas de la distribución en la siguiente aplicación:

1.7 Tipificación

A veces la distribución presenta muchas irregularidades, como asimetrías acentuadas, valores extremos, etc. En estos casos es recomendable efectuar una transformación que la haga más regular. Un caso particular de transformación es la **tipificación**.

Dada X con media $\bar{x}$ y desviación típica S

$$Z = \frac{X - \bar{x}}{S}$$

A Z se le llama variable o y tiene media 0 y desviación típica 1.

La tipificación tiene la propiedad de hacer comparables individuos que pertenezcan a distintas distribuciones, aún en el caso de que éstas vinieran expresadas en diferentes unidades.

Dos trabajadores del mismo sector ganan 620 y 672 euros, respectivamente. El primero pertenece a la empresa A, cuya retribución media y desviación típica vienen dados por: $\bar{x}_A = 580$ euros y $S_{x_A} = 25$ euros, mientras que para la empresa del segundo trabajador se tiene: $\bar{x}_B$ =640 euros y $S_{x_B} = 33$ euros.

¿cuál de los dos ocupa mejor posición relativa dentro de su empresa?

$$z_A = \frac{620 - 580}{25} = 1.6$$

$$z_B = \frac{672 - 640}{33} = 0.97$$

Capítulo 2

Análisis bivariante. Ajuste y regresión bidimensional.

Capítulo 3

Objetivos del tema:

Cuando analizamos dos variables estadísticas a la vez, hablamos de **estadística descriptiva bivariante**. Si hay más de dos, hablamos de **estadística descriptiva multivariante**. Por ejemplo:

X	Y
14	3
12	4
15	1
13	5
16	0

Una observación de esta variable bidimensional es $(14, 3)$.

Los objetivos que nos planteamos para este capítulo son:

- Organizar los datos en tablas de frecuencias, buscando la mejor forma de representarlos gráficamente.
- Medir la relación entre dos variables estadísticas cualitativas.
- Medir la relación entre dos variables estadísticas cuantitativas.
- Realizar predicciones y medir la fiabilidad de dicha predicción.
- Entender que **la correlación no implica causalidad**.