

Ejercicio resuelto 2.1

En una tienda han anotado los precios de los artículos que han vendido en una hora. Los datos son: 9,95, 19,95, 19,95, 14,95, 29,95, 14,95, 9,95, 9,95, 14,95, 19,95, 19,95, 29,9, 9,95, 14,95, 9,95, 29,95, 14,99, 14,95, 19,95, 29,95 y 9,95. Realiza un análisis exploratorio.

Solución:

El **objetivo** es realizar un análisis exploratorio.

Planteamiento: el experimento consiste en seleccionar artículos (individuos) y observar su precio (variable). La población serían todos los artículos y la muestra los 21 vendidos en esa hora. Los posibles valores de la variable son $\{9,95, 19,95, \dots\}$, es decir, los números positivos, luego es una variable cardinal (porque los posibles valores son números) y de razón, (porque un precio de 0 significa que no cuesta nada). En principio, se puede considerar que es una variable continua porque podría haber cualquier precio (9, 9,2, 9,25, 9,254, ...), depende de la precisión, aunque está actuando como discreta, porque hay pocos precios distintos que se repiten con alta frecuencia.

Método y justificación: como la variable es discreta se realizará una tabla de frecuencias y un gráfico de barras con el fin de observar si hay algún valor anómalo y obtener una primera idea de la forma de la distribución.

Cálculos: la Tabla 2.1 muestra los precios distintos y sus frecuencias, mientras que en la Figura 2.1 se han representado los precios en el eje de las equis, asociándolos con barras cuya altura representa la frecuencia absoluta (que es proporcional a la frecuencia relativa).

x_i	n_i
9,95	6
14,95	5
14,99	1
19,95	4
29,95	5
Total	21

Tabla 2.1: Tabla de frecuencias.

Conclusión: en la tabla de frecuencias y en el gráfico de barras se observa que la frecuencia de 14,99 es mucho más baja que la del resto. Además, no sigue la pauta de los otros 20 artículos, cuyo precio es siempre cierta cantidad más 95 céntimos, por lo que parece que puede ser un error en la recogida de datos. Aunque el cambio no afectará prácticamente al

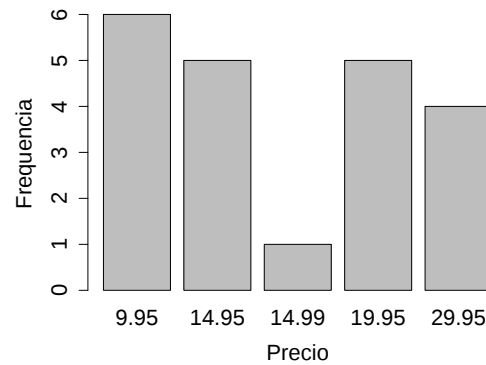


Figura 2.1: Gráfico de barras.

análisis estadístico, se puede corregir el error para que las representaciones tabulares gráficas sean más naturales (ver Tabla 2.2 y Figura 2.2).

x_i	n_i
9,95	6
14,95	5
19,95	4
29,95	5
Total	21

Tabla 2.2: Tabla de frecuencias corregida.

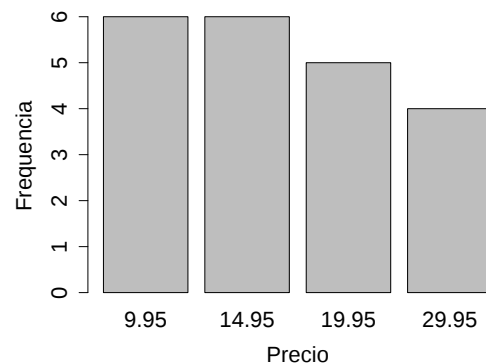


Figura 2.2: Gráfico de barras corregido.

En la Tabla 2.2 y Figura 2.2 no se observa ningún valor anormal. Las barras son al principio más altas y disminuyen luego paulatinamente, aunque la diferencia no es mucha. Eso significa que la distribución se concentra ligeramente más en precios bajos, siendo los precios altos cada vez un poco menos frecuentes.

Ejercicio resuelto 2.2

La siguiente tabla muestra el importe de 60 fincas hipotecadas en cierta región durante el último año (en miles de euros).

167,3	175	125,7	135,4	183,2	178,2	114,1	193,9	235,1	192	145,1	113,2
182,6	128,5	378,3	192,4	127,4	173,2	165,8	126,5	233,2	91	123,1	181,3
182,3	172	189,2	139,4	144,8	119,2	157,3	133,5	168,5	152,3	78,5	137,2
234,2	192,5	178,4	97,5	167,3	118,9	134,2	147,2	136,7	123,5	181,9	124,4
111,63	182,3	127,4	173,5	155,5	193,4	189,3	174,3	168,5	217	198,2	147,3

- ¿Crees que hay algún dato anómalo en la muestra? Describe a grandes rasgos la distribución de esas hipotecas.
- Representa gráficamente y comenta la distribución de las hipotecas comparando los resultados con los obtenidos en el apartado anterior.

Solución:

El **objetivo del Apartado a)** es localizar si hay algún valor anómalo y describir a grandes rasgos la distribución.

Planteamiento: el experimento consiste en seleccionar hipotecas formalizadas en el último año (individuos) y observar su importe en miles de euros (variable). La población la constituyen todas las hipotecas y la muestra las 60 seleccionadas. Los posibles valores de la variable son $\{160, 160,1, 160,4, \dots\}$, es decir, los números positivos, luego es una variable cardinal (porque los posibles valores son números) y de razón (porque una hipoteca de 0 significa que no hay hipoteca). Es una variable continua porque podría haber cualquier valor (depende de la precisión) y está actuando como continua, porque hay muchos valores distintos.

Método y justificación: como la variable es continua, no tiene sentido realizar una tabla de frecuencias y es preferible optar por el análisis exploratorio gráfico. Se realizará un gráfico de tallos y hojas (en inglés, Stem-and-Leaf Plot) y un gráfico de cajas (en inglés, Boxplot) .

A la vez que se describe la construcción de los gráficos y su utilidad se indicarán las **conclusiones**.

En la Figura 2.3, la primera frecuencia es 3, que significa que en la primera fila hay 3 hipotecas. *Stem width* es 100, así que como el primer *Stem* es 0, significa que los 3 datos representados en la primera fila son menos de 100 (decenas). El segundo *Stem* es 1, lo que significa que en esa fila las hipotecas son de “ciento y pico” miles de euros.

```

Importe Hipoteca Stem-and-Leaf Plot

Frequency      Stem &  Leaf

    3,00        0 .   799
   23,00        1 .  1111122222222333334444
   29,00        1 .  555666667777778888888999999
    4,00        2 .   1333
    1,00 Extremes      (>=378)

Stem width:    100,00
Each leaf:      1 case(s)

```

Figura 2.3: Gráfico de tallos y hojas.

En la Figura 2.3, el primer valor de leaf es 7, luego el dato correspondiente sería ‘setenta y pico’. El primer valor de leaf de la segunda fila es 1, luego el dato correspondiente sería ‘ciento diez y pico’.

Se observa que cada hoja representa 1 dato (en inglés each leaf: 1 case), es decir, el primer 7 indica una hipoteca de setenta y pico mil euros, el segundo una de noventa y pico mil euros, etc.

Al llegar a la última frecuencia, aparece un dato que llama *Extreme* con valor mayor o igual a 378. No hay valores extremos bajos, sin embargo hay un valor extremo alto.

Además se observa que la fila entre 50 y 99 es más corta que la de 100 a 149 y casi igual que la de 200 a 249. La fila más larga es la de 150 a 199. Esto significa que hay más hipotecas en esa franja (hay más densidad de datos), hay unas poquitas menos entre 100 y 149 mil euros y menos aún en 200-249 y 50-99. Es decir, la mayoría de las hipotecas están en la zona central/alta, habiendo pocas muy bajas o muy altas.

En el diagrama de cajas de la Figura 2.4 se observa una escala de importes (en miles de euros) de algo menos de 100 a algo más de 350. Aparece un círculo con un valor de algo más de 350 muy alejado del resto del gráfico. Ese valor es elevado en relación con el resto, pero no parece que se trate de un error, ya que es perfectamente posible que un piso se hipoteque por un importe de 350 ó 400 mil euros.

Se observa que el segmento superior está aproximadamente en 240 mil euros y el inferior alrededor de 80 mil euros, es decir, el rango de valores ‘normales’ para esas hipotecas va aproximadamente de 80 a 240 mil euros, coincidiendo el extremo inferior con el mínimo de la muestra.

Además la raya negra está aproximadamente en 160 mil euros, lo que significa que la mitad de las hipotecas de la muestra tienen un importe

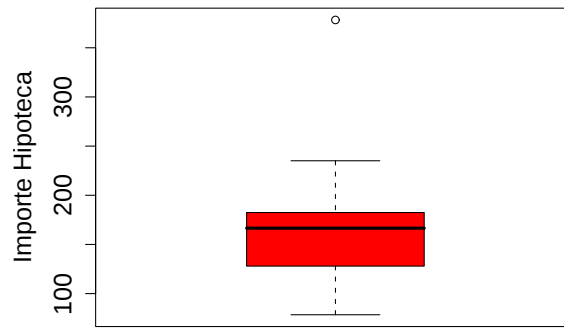


Figura 2.4: Gráfico de cajas.

de aprox. 16000 euros o menos y la otra mitad de aprox. 16000 euros o más.

El borde inferior está alrededor de 120 y el superior alrededor de 189, lo que significa que el 50 % de las hipotecas (sin tener en cuenta el 25 % de las más baratas y el 25 % de las más caras) se sitúan entre los 120 y los 180 mil euros.

Si se elimina el dato extremo (ver Figura 2.5), la forma del gráfico y las conclusiones son similares, aunque la escala es distinta. Esto significa que es un gráfico *robusto*, que no se ve muy influenciado por la presencia de valores anómalos.

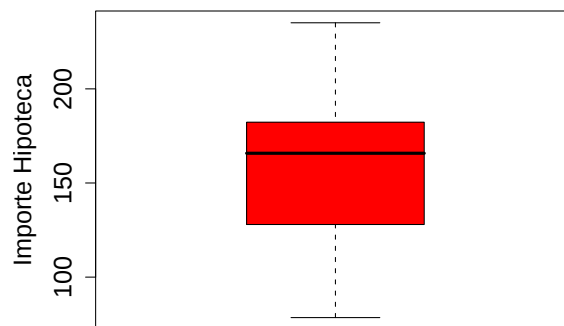


Figura 2.5: Gráfico de cajas sin dato extremo.

El **objetivo del Apartado b)** es representar la distribución del importe de las hipotecas. El **planteamiento** está hecho en el Apartado a).

Método y justificación: se podría utilizar una tabla de frecuencias (agrupando previamente los datos), sin embargo, los gráficos son más sencillos de interpretar mediante una simple referencia visual y contienen la misma información que se puede incluir en una tabla.

Como la variable es continua, el gráfico más adecuado será el histograma porque los valores se representarán agrupados siguiendo su escala natural.

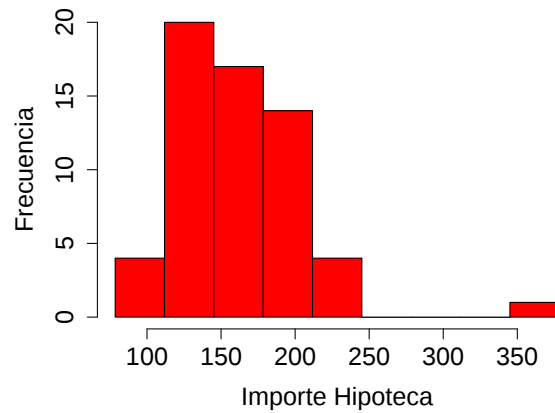


Figura 2.6: Histograma con pocos intervalos.

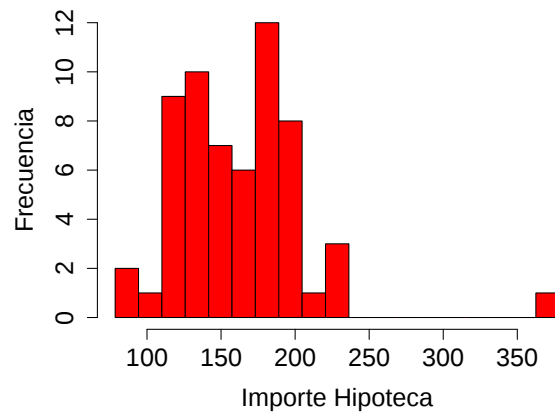


Figura 2.7: Histograma con muchos intervalos.

Cálculos: con ayuda de cualquier programa estadístico se obtienen gráficos como los representados en las Figuras 2.6 y 2.7. La forma de los histogramas depende bastante del número de intervalos elegidos, por lo que es importante hacer pruebas y elegir los que sean más convenientes. En este caso, parece más adecuado alrededor de 6 intervalos al ser la forma más suave y más realista (Figura 2.6).

Conclusión: en la Figura 2.6 se aprecia, en primer lugar, la existencia de un dato extremo. El resto de los datos se concentran entre aprox. 80 y casi 250. El centro de la distribución parece situarse alrededor de 150 ó 170 mil euros (aprox.). Hay pocas hipotecas con importes muy bajos (menos de 100), sin embargo, en valores moderadamente bajos (100-140) se concentra un gran número. Luego la densidad va decreciendo paulatinamente hacia importes más altos.

Las conclusiones acerca de la distribución que se extraen de este histograma son similares a las que se obtuvieron a partir del gráfico de tallos y hojas, aunque aquí es más sencillo observar el detalle de la distribución general (aunque se pierde la información individual).

Ejercicio resuelto 2.3

Los 5 nombres de varón más frecuentes en 2008 en Asturias según datos del SADEI fueron Daniel (134 niños), Diego (126 niños), Hugo (137 niños), Pablo (144 niños) y Pelayo (120 niños). Representa los datos de la forma que consideres más adecuada.

Solución:

El **objetivo** es representar la información acerca de los nombres más frecuentes de la forma más adecuada.

Planteamiento: el experimento consiste en seleccionar niños (individuos) y observar su nombre (variable).

Un error común en este tipo de problemas es plantear el experimento alternativo “seleccionar nombres y observar su frecuencia”. En este caso, la variable sería ‘la frecuencia de cada nombre’ y las estadísticas se referirían a dichas frecuencias numéricas (por ejemplo, se podría calcular la ‘frecuencia media’). Como, en realidad, el interés se centra en los nombres, la variable debe ser el nombre y las estadísticas se referirán a los nombres (por ejemplo, se podría indicar el ‘nombre más común’). Se debe recordar siempre que la variable debe ser la característica de interés sobre la que se pretenden realizar estadísticas.

La población son todas los niños y la muestra los $144 + 137 + 134 + 126 + 120 = 661$ con nombres más frecuentes. Los posibles valores de la variable son {Pablo, Hugo, ...} que son nombres que no admiten un orden natural, luego es una variable nominal.

Método y justificación: en principio se podría utilizar una tabla de frecuencias, sin embargo, los gráficos son más sencillos de interpretar mediante una simple referencia visual y contienen la misma información que se puede incluir en una tabla.

En general las tablas se utilizarán con fines técnicos (como apoyo para hacer estadísticas), pero a la hora de publicar la información es más recomendable utilizar un gráfico.

Como la variable es nominal, el gráfico más adecuado será el de sectores, porque los valores se representarán sobre un círculo sin orden natural.

Cálculos: con ayuda de cualquier programa estadístico se obtiene un gráfico como el representado en la Figura 2.8.

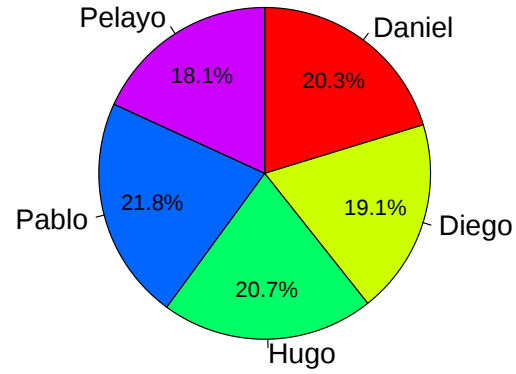


Figura 2.8: Gráfico de sectores sin dato extremo.

Conclusión: se aprecia que casi la frecuencia de los 5 nombres es similar, habiendo poca diferencia entre el más frecuente (Pablo) y el menos (Pelayo).