Práctica 1. Soluciones.

Estadística Grado Biología 2022-23

Actualizado: 2024-10-24

Enunciado

Los datos que aparecen en el fichero `p1-robles.csv` se refieren a un estudio realizado sobre un robledal cercano a una planta industrial, parte de los cuales se visualizan en la tabla. Se han seleccionado robles

- De dos variedades (A y B).
- Ubicados en cuatro zonas distintas.
- La mitad de ellos han sido sometidos a cierto tratamiento (codificados con 1), los no tratados (codificados con 0).

Sobre cada árbol se han medido las concentraciones (mg/kg) de ocho elementos químicos en sus hojas:

- Metales pesados: hierro, manganeso y zinc
- Metales alcalinotérreos: calcio y magnesio
- Metal alcalino: potasio
- No metales: fósforo y nitrógeno.

Instrucciones:

- Abre RStudio
- Pulsa File -> New file -> R script
- Copia en el script y ejecuta este código (ya veremos qué significa en la práctica 2)

con esto has creado la tabla (data.frame, en la jerga de R) `robles` que contiene los datos arriba mencionados

Responde de forma concisa y razonadamente a las siguientes preguntas:

Ejercicio 1 Clasifica las variables del estudio en cualitativas, cuantitativas discretas y cuantitativas continuas.

Cualitativas nominales son `variadad`, `zona` y `tratamiento`. Cuantitativas discretas no hay, y continuas son todas las concentraciones de los 8 elementos químicos.

Ejercicio 2 ¿Hay alguna variable cualitativa ordinal?

No, ninguna.

Ejercicio 3 ¿Cuántos robles hay de cada variedad?

La tabla de frecuencias absolutas nos dice Cuántos robles hay de cada variedad

```
table(robles$Variedad)
```

```
##
##  A  B
## 14 24
```

¿Qué porcentaje de los robles estudiados pertenecen a la variedad B?

El porcentaje se obtiene de la tabla de recuencias relativas

```
n = nrow(robles)
table(robles$Variedad)/n
```

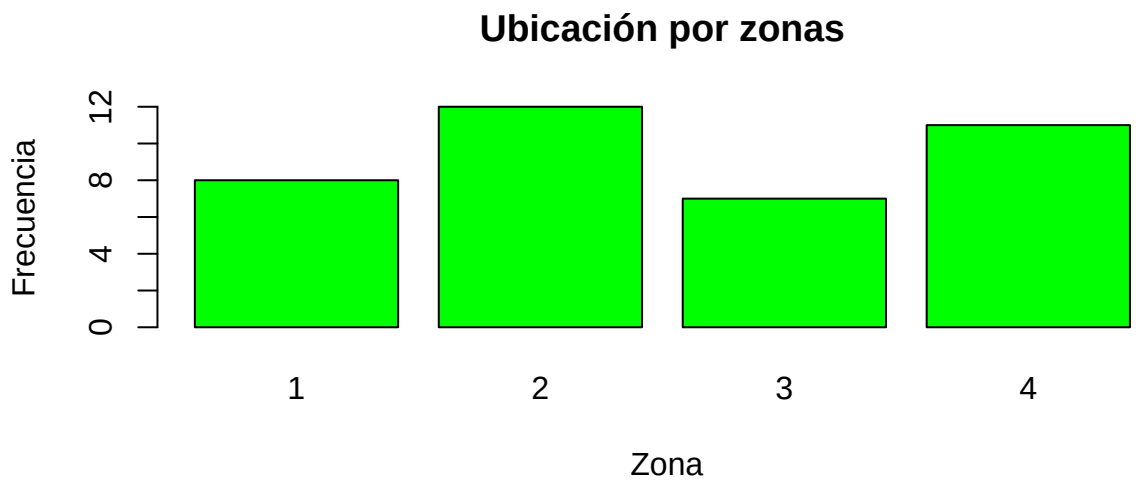
```
##
##         A         B
## 0.3684211 0.6315789
```

que es la tabla en la que cada una de las frecuencias absolutas está dividida entre el total de datos. Esto da la proporción (tanto por uno) de robles de cada tipo. Al multiplicar por 100 se obtiene el tanto por ciento.

Ejercicio 4 Construye el diagrama de barras de la variable `Zona` e interpreta los resultados

Para construir el diagrama de barras se usa

```
barplot(table(robles$Zona),
        col = "green",
        main = "Ubicación por zonas",
        xlab = " Zona",
        ylab = "Frecuencia")
```



se observa que donde más robles hay es en la zona 2, luego en la 4, en la 1 y en la 2. Para obtener valores exactos (número de robles) es preferible calcular la tabla de frecuencias aboslutas

```
table(robles$Zona)
```

```
##
##  1  2  3  4
##  8 12  7 11
```

Ejercicio 5 Calcula la media, la mediana y el histograma de la variable `Nitrogeno`. Comenta los resultados.

La media es

```
mean(robles$Nitrogeno)
```

```
## [1] 3.293737
```

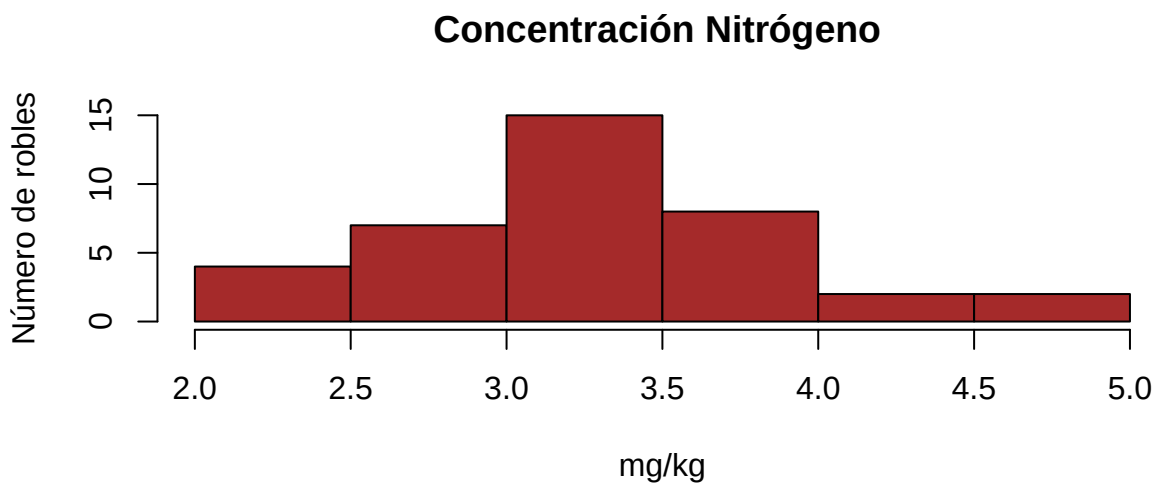
La mediana

```
median(robles$Nitrogeno)
```

```
## [1] 3.2895
```

y el histograma

```
hist(robles$Nitrogeno,
     col = "brown",
     main = "Concentración Nitrógeno ",
     xlab = "mg/kg",
     ylab = "Número de robles")
```



Se observa que la distribuión de los datos es unimodal y muy simétrica, lo que es totalmente coherente con el hecho de que la media y la mediana sean prácticamente iguales.

Ejercicio 6 Calcula la media, la mediana y el histograma de la variable `Hierro`. Comenta los resultados.

En este caso la media es

```
mean(robles$Hierro)
```

```
## [1] 0.02989474
```

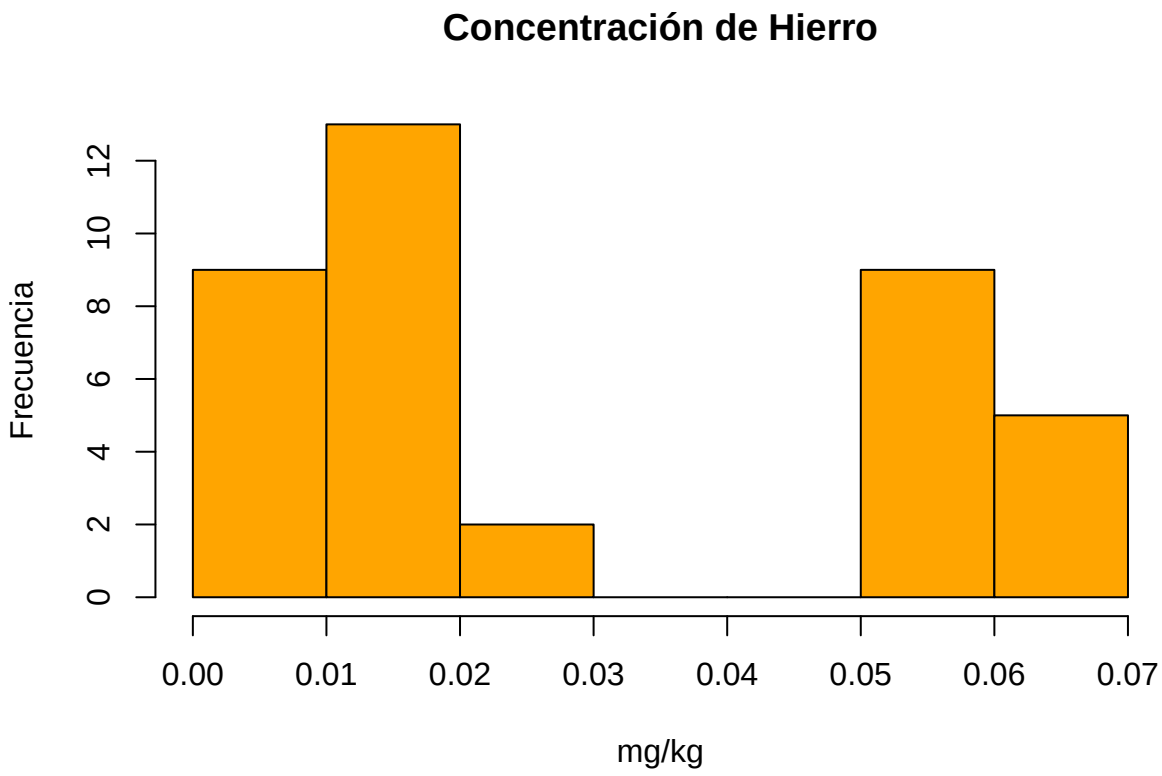
La mediana

```
median(robles$Hierro)
```

```
## [1] 0.0145
```

y el histograma

```
hist(robles$Hierro,
       col = "orange",
       main = "Concentración de Hierro",
       xlab = "mg/kg",
       ylab = "Frecuencia")
```

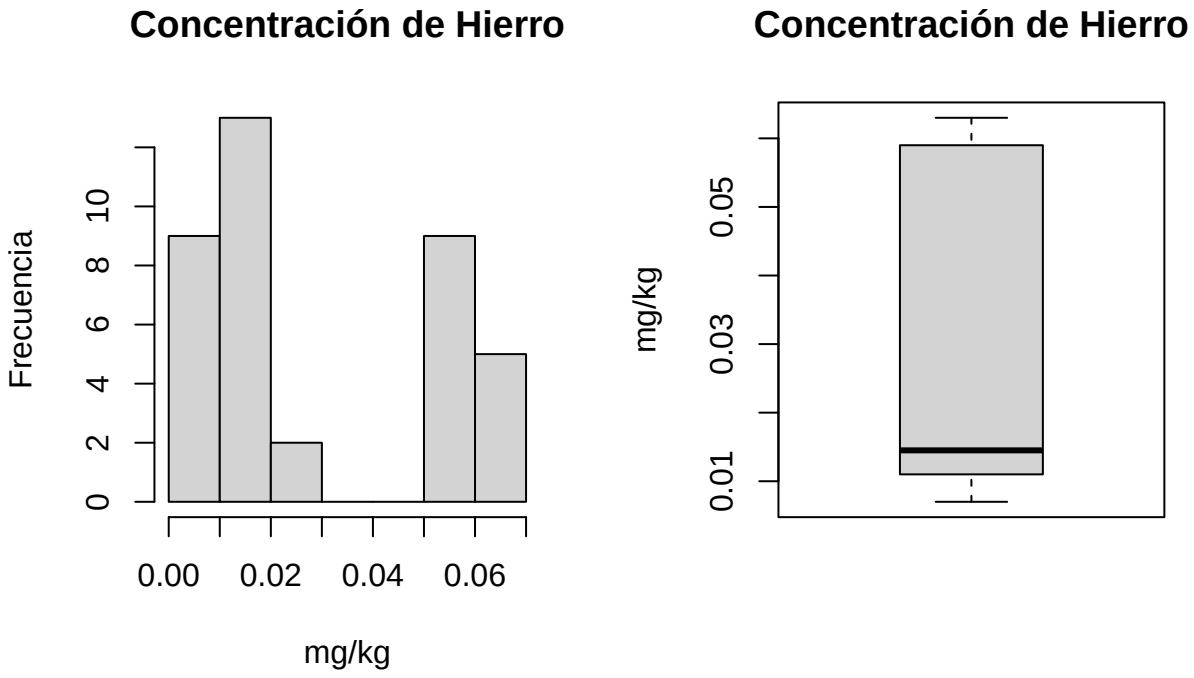


La concentración de hierro varía de 0 a 0.07, aunque ningún individuo tiene concentraciones de hierro entre 0.03 y 0.05. De hecho, los valores del hierro se agrupan en dos zonas: los que tienen concentraciones menores de 0.03 y los que las tienen entre 0.05 y 0.07. Este hecho sugiere la existencia de dos grupos de robles, aunque de momento no tenemos explicación. Es decir, desconocemos si este hecho se relaciona con el comportamiento de alguna otra variable.

La distribuión de los datos sigue siendo unimodal, pero es muy asimétrica; el gráfico incluso sugiere la existencia de dos grupos. La media es mayor (casi el doble) que la mediana, lo que se explica por la asimetréa (a la derecha) de los datos. De hecho, la media (approx 0.03) está en una zona en la que hay muy pocos datos, por lo que no resulta muy descriptiva del conjunto de datos del que se calcula.

Ejercicio 7 Considera la variable `Hierro`, calcula su histograma y su boxplot. Compara lo que te dice cada una de las figuras.

```
par(mfrow = c(1,2))
hist(robles$Hierro,
        main = "Concentración de Hierro",
        xlab = "mg/kg",
        ylab = "Frecuencia")
boxplot(robles$Hierro,
        main = "Concentración de Hierro",
        ylab = "mg/kg")
```



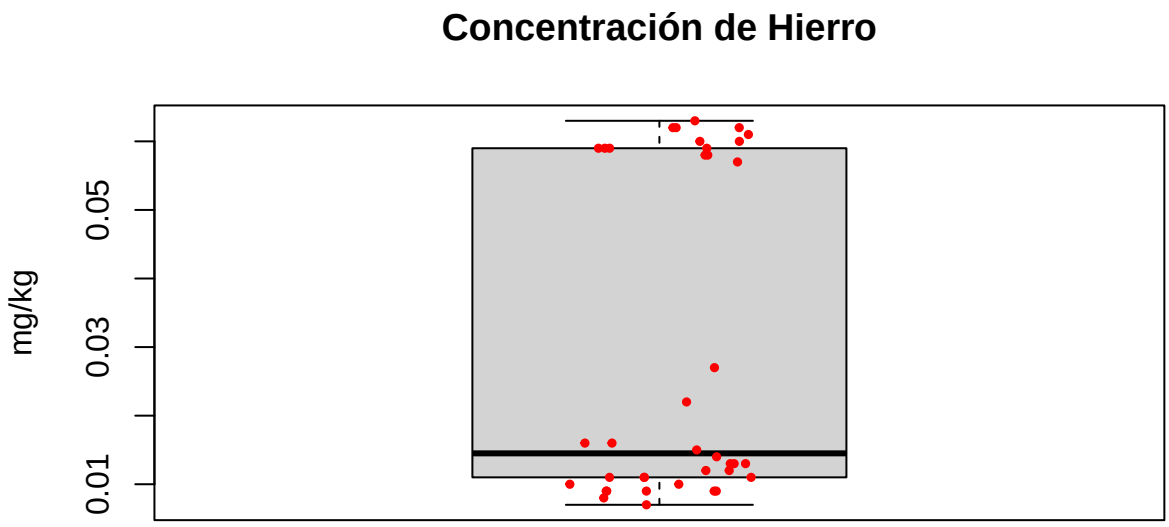

```
par(mfrow = c(1,1))
```

En el histograma se aprecia de forma evidente que los datos están agrupados en dos bloques. No hay ningún roble con una concentración de `Hierro` entre 0.03 y 0.05.

En el boxplot esto no es tan evidente (aunque algo se intuye). Se observa una distribución no uniforme de los datos: están muy concentrados por debajo de la mediana (bigote inferior próximo a la caja y mitad inferior de la caja muy estrecha) y por encima del tercer cuartil (bigote superio próximo a la caja). Están muy dispersos entre la mediana y el tercer cuartil (mitad superior de la caja).

Si sobreimpresionamos los valores en el boxplot

```
boxplot(robles$Hierro,
        main = "Concentración de Hierro",
        ylab = "mg/kg")
stripchart(robles$Hierro,
           method = "jitter", add = TRUE,
           pch=19, col= "red", cex=0.5, vertical = TRUE)
```



se evidencia la no homogeneidad en la distribución de los datos y el hecho de que ningún roble presenta una concentración de `Hierro` entre 0.03 y 0.05.

Cosas de R: fíjate en que en `stripchart` el argumento `vertical` vale `TRUE`. Por defecto, el boxplot se representa verticalmente, pero los puntos horizontalmente (no sólo se usan con el boxplot, claro). Por eso hay que añadir, o bien en `boxplot` el argumento `horizontal = TRUE`, o bien en `stripchart` el argumeto `vertical = TRUE`. Si haces las dos cosas simultaneamente, seguirás sin ver los puntos sobre el boxplot.

Ejercicio 8 Calcula los cuartiles de la variable `Hierro` e interpretalos.

```
quantile(robles$Hierro)
```

```
##      0%     25%     50%     75%    100%
## 0.0070 0.0110 0.0145 0.0590 0.0630
```

Quiere decir que

- El 25% de los robles tienen 0.011mg/kg de hierro o menos
- El 50% de los robles tienen 0.0145mg/kg de hierro o menos (y el otro 50% más de esa cantidad)
- El 25% de los robles tienen 0.063mg/kg de hierro o más

Ejercicio 9 ¿Entre qué valores se mueve el 80% central de la muestra respecto de la variable `Magnesio`? Se trata de los valores que están por encima del percentil 10 y por debajo del percentil 90

```
quantile(robles$Magnesio, probs = c(0.1, 0.9))
```

```
##    10%    90%
## 0.2949 0.4881
```

Ejercicio 10 Construye una tabla de frecuencias para la variable `Hierro` con 8 intervalos de clase entre 0 y 0.08.

Para calcular la tabla de frecuencias absolutas primero hay que definir los extremos de los subintervalos.

```
(cortes = seq(from = 0, to = 0.08, length.out = 9))
```

```
## [1] 0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08
```

A continuación, asigna a cada valor de la variable `Hierro` la clase a la que pertenece

```
(clases = cut(robles$Hierro, breaks = cortes, include.lowest = TRUE))
```

```
##  [1] (0.05,0.06] (0.05,0.06] (0.05,0.06] (0.05,0.06] (0.06,0.07] (0.02,0.03]
##  [7] (0.05,0.06] (0.06,0.07] (0.05,0.06] (0.05,0.06] (0.01,0.02] (0.01,0.02]
## [13] (0.01,0.02] (0.01,0.02] [0,0.01]    (0.01,0.02] (0.01,0.02] (0.01,0.02]
## [19] (0.01,0.02] (0.01,0.02] [0,0.01]    [0,0.01]    (0.02,0.03] (0.06,0.07]
## [25] (0.06,0.07] (0.05,0.06] (0.06,0.07] (0.05,0.06] (0.01,0.02] [0,0.01]
## [31] (0.01,0.02] (0.01,0.02] (0.01,0.02] [0,0.01]    [0,0.01]    [0,0.01]
## [37] [0,0.01]    [0,0.01]
## 8 Levels: [0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] ... (0.07,0.08]
```

Finalmente, recoger todo esto en una tabla

```
table(clases)
```

```
## clases
##    [0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] (0.04,0.05] (0.05,0.06]
##           9          13           2           0           0           9
## (0.06,0.07] (0.07,0.08]
##           5           0
```

¿Cuántos robles tienen una concentración de `Hierro` entre 0.05 y 0.06? Se observa en la tabla que hay 9

¿Qué porcentaje de robles tiene una concentración menor o igual que 0.06? Hay que sumar la el total de robles con una concentración menor o igual que 0.06 y luego dividir entre el total de robles (y multiplicar por 100, claro)

```
((9+13+2+9)/n)*100
```

```
## [1] 86.84211
```

Otra opción ión directa es usar la función `cumsum()`.

```
(tablaAcumulada = cumsum(table(clases)))
```

```
##    [0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] (0.04,0.05] (0.05,0.06]
##           9          22          24          24          24          33
## (0.06,0.07] (0.07,0.08]
##          38          38
```

para obtener la tabla de frecuencias acumuladas y, a continuación calcular la tabla de frecuencias relativas acumuladas

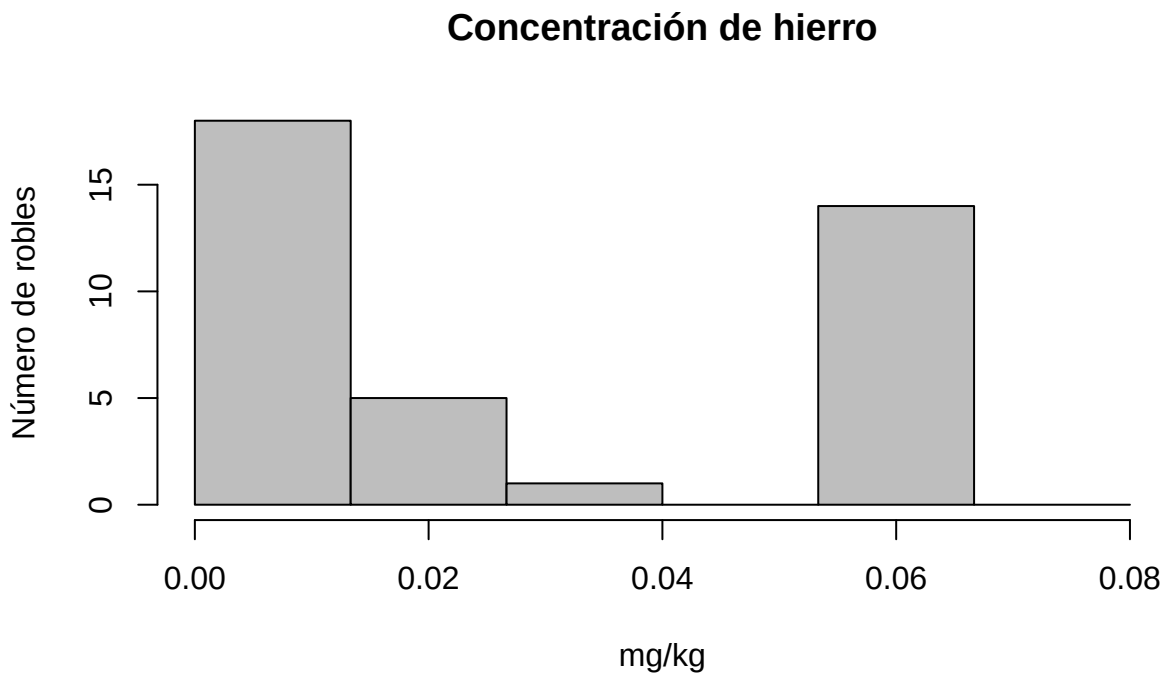
```
tablaAcumulada/nrow(robles)
```

```
##    [0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] (0.04,0.05] (0.05,0.06]
##   0.2368421   0.5789474   0.6315789   0.6315789   0.6315789   0.8684211
## (0.06,0.07] (0.07,0.08]
##   1.0000000   1.0000000
```

Ejercicio 11 Representa un histograma para la variable `Hierro` agrupando los valores en 6 clases de la misma longitud (entre 0 y 0.08). Explica el resultado y compáralo con el otro histograma calculado

Adapta el comando que has usado antes para agrupar la variable `Hierro` en 8 clases al caso de 6 clases

```
cortes = seq(from = 0, to = 0.08, length.out = 7)
hist(robles$Hierro,
        col = "gray",
        breaks = cortes,
        main = "Concentración de hierro", xlab = "mg/kg", ylab = "Número de robles")
```



Aparentemente hay menos barras (4) que clases (hemos definido 6), lo que sucede es con esta nueva forma de agrupar los robles en clases hay 2 clases vacías (ningún roble presenta una concentración de `Hierro` dentro de dichas clases).

Observa que los histogramas son diferentes; resumir datos hace que perdamos información, y no siempre nos daremos cuenta.