PRÁCTICAS DE ESTADÍSTICA CON *R*

PRÁCTICA 2: ESTADÍSTICA DESCRIPTIVA DE UNA VARIABLE

2.1 Lectura de datos

En la mayor parte de los conceptos que vamos a ir viendo a lo largo de esta práctica se va a hacer referencia al ejercicio siguiente:

■Ejemplo 2-1: Se ha medido la tensión de rotura en toneladas por cm^2 de 50 pernos de una nueva aleación de aluminio, obteniéndose los valores que aparecen en el archivo "Tensión de rotura.txt". Se pide:

1º) Tabla de frecuencias absolutas y relativas.

2º) Histograma de frecuencias absolutas.

3º) Diagrama boxplot.

4º) Media y desviación típica.

5º) Mediana, primer cuartil, tercer cuartil y percentil 60.

En primer lugar debemos leer el archivo correspondiente, que en nuestro caso está en un fichero de texto denominado "Tensión de rotura.txt" y cuya ruta de acceso se ha de detallar con precisión:

```
> datos<-
read.table("C:\\Estadística\\Prácticas\\Tensión de
rotura.txt")
> datos
     V1
1  4.05
2  4.58
3  4.42
...
49 3.54
50 4.84
> class(datos) # Con esta sentencia confirmamos el
tipo de objeto que es datos
[1] "data.frame"
```

Como se ve, el programa ha creado un marco de datos denominado **datos** y que consta de 50 valores de una variable a la que automáticamente le ha asignado el nombre V1. Ahora crearemos un vector con los valores numéricos, lo que se consigue eligiendo la variable V1 del *data frame* **datos**:

```
> tenrot<-datos$V1
> tenrot
[1] 4.05 4.58 4.42 4.20 4.41 4.64 4.76 4.58 3.95 4.17
4.56 3.51 3.27 3.80 3.59
```

```
[16] 4.70 3.77 3.80 4.27 3.94 3.96 4.86 4.39 4.04 4.36
3.72 4.00 3.46 4.01 4.08
[31] 3.40 3.89 4.46 4.38 4.41 4.33 4.16 4.58 4.03 3.76
4.05 4.17 4.46 3.60 4.76
[46] 3.99 4.43 4.15 3.54 4.84

> #Ordenamos los valores del vector tenrot
> sort(tenrot)
[1] 3.27 3.40 3.46 3.51 3.54 3.59 3.60 3.72 3.76 3.77
3.80 3.80 3.89 3.94 3.95
[16] 3.96 3.99 4.00 4.01 4.03 4.04 4.05 4.05 4.08 4.15
4.16 4.17 4.17 4.20 4.27
[31] 4.33 4.36 4.38 4.39 4.41 4.41 4.42 4.43 4.46 4.46
4.56 4.58 4.58 4.58 4.64
[46] 4.70 4.76 4.76 4.84 4.86
```

2.2 Tabla de frecuencias

En primer lugar vamos a definir el número de intervalos en que agruparemos los datos para construir una tabla de frecuencias:

```
> sqrt(50)
[1] 7.071068
```

Construiremos 7 intervalos con una amplitud que calculamos a continuación:

```
> min(tenrot);max(tenrot)
[1] 3.27
[1] 4.86
> (max(tenrot)-min(tenrot))/7
[1] 0.2271429
```

Redondeando este valor a 0,3, formaremos 7 intervalos de esa amplitud empezando en 3 y terminando en 5,1. Para ello construimos el vector formado por los extremos de los intervalos en los que agrupamos los datos:

```
> límites<-scan()
1: 3
2: 3.3
3: 3.6
4: 3.9
5: 4.2
6: 4.5
7: 4.8
8: 5.1
9:
Read 8 items

> límites
[1] 3.0 3.3 3.6 3.9 4.2 4.5 4.8 5.1
```

Veamos ahora en qué intervalo queda cada uno de los 50 valores leídos. Por defecto *R* crea intervalos abiertos por la izquierda y cerrados por la derecha. Si queremos hacer intervalos cerrados por la izquierda y abiertos por la derecha indicamos la opción **right=F**:

```
> tenrot.intervalos<-cut(tenrot,límites,right=F)
> tenrot.intervalos
[1] [3.9,4.2) [4.5,4.8) [4.2,4.5) [4.2,4.5) [4.2,4.5)
[4.5,4.8) [4.5,4.8)
[8] [4.5,4.8) [3.9,4.2) [3.9,4.2) [4.5,4.8) [3.3,3.6)
[3,3.3)   [3.6,3.9)
[15] [3.3,3.6) [4.5,4.8) [3.6,3.9) [3.6,3.9) [4.2,4.5)
[3.9,4.2) [3.9,4.2)
[22] [4.8,5.1) [4.2,4.5) [3.9,4.2) [4.2,4.5) [3.6,3.9)
[3.9,4.2) [3.3,3.6)
[29] [3.9,4.2) [3.9,4.2) [3.3,3.6) [3.6,3.9) [4.2,4.5)
[4.2,4.5) [4.2,4.5)
[36] [4.2,4.5) [3.9,4.2) [4.5,4.8) [3.9,4.2) [3.6,3.9)
[3.9,4.2) [3.9,4.2)
[43] [4.2,4.5) [3.6,3.9) [4.5,4.8) [3.9,4.2) [4.2,4.5)
[3.9,4.2) [3.3,3.6)
[50] [4.8,5.1)
7 Levels: [3,3.3) [3.3,3.6) [3.6,3.9) [3.9,4.2)
[4.2,4.5) ... [4.8,5.1)
```

Utilizamos ahora la función **table** para contar el número de veces que aparece cada intervalo, lo que es propiamente una tabla de frecuencias:

```
> table(tenrot.intervalos)
tenrotporintervalos
[3,3.3) [3.3,3.6) [3.6,3.9) [3.9,4.2) [4.2,4.5)
[4.5,4.8) [4.8,5.1)
1         5         7        15        12         8
2
```

Es decir, en el primer intervalo [3,3.3) hay un valor, en el segundo intervalo [3.3,3.6) caen 5 valores, etc.

Si aplicáramos la función anterior al vector **tenrot** se obtendría el resultado siguiente, de poca utilidad en este caso:

```
> table(tenrot)
tenrot
3.27  3.4 3.46 3.51 3.54 3.59  3.6 3.72 3.76 3.77  3.8
3.89 3.94 3.95 3.96 3.99
   1    1    1    1    1    1    1    1    1    1    2
1    1    1    1    1
   4 4.01 4.03 4.04 4.05 4.08 4.15 4.16 4.17  4.2 4.27
4.33 4.36 4.38 4.39 4.41
```

```
   1    1    1    1    2    1    1    1    2    1    1
1    1    1    1    2
4.42 4.43 4.46 4.56 4.58 4.64  4.7 4.76 4.84 4.86
   1    1    2    1    3    1    1    2    1    1
```

2.3 Diagrama de tallos y hojas

Un diagrama de tallos y hojas (*stem and leaf)* es una representación sencilla de los datos similar a un histograma, pero con la ventaja de que se conserva la información numérica de todos y cada uno de los datos.

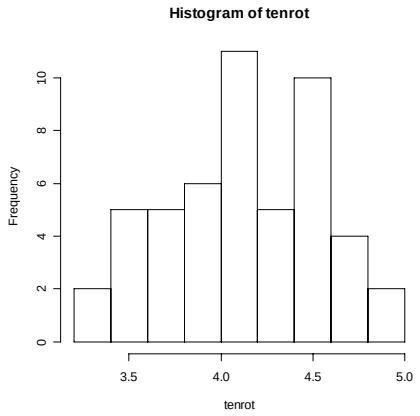
```
> stem(tenrot)
The decimal point is 1 digit(s) to the left of the |

  32 | 7
  34 | 06149
  36 | 0267
  38 | 0094569
  40 | 01345585677
  42 | 073689
  44 | 1123666888
  46 | 4066
  48 | 46
```

2.4 Histograma

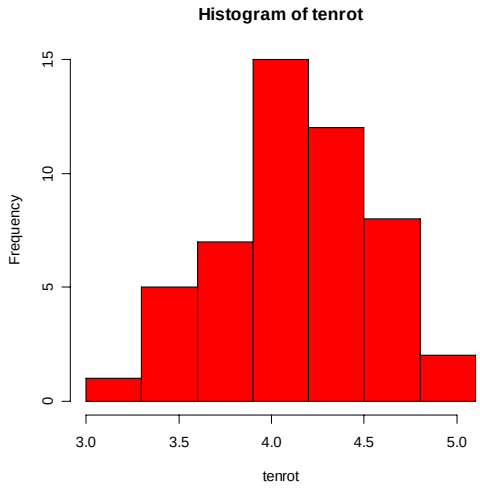
Al igual que el diagrama de tallos y hojas un histograma da una idea de cómo se distribuyen los datos. La instrucción correspondiente es

```
> hist(tenrot)
```



Mediante la instrucción anterior construimos un histograma dividido en 9 intervalos, que el programa genera automáticamente. Con objeto de disponer de un histograma con los siete intervalos definidos por **límites** hacemos lo siguiente:

```
> hist(tenrot,límites,right=F,col="red")  #Dibujamos así el histograma de frecuencias (con barras de color rojo)
```



2.5 Medidas de centralización

Obtenemos a continuación la media y la mediana de los valores que forman el vector **tenrot**:

```
> mean(tenrot);median(tenrot)
[1] 4.1448
[1] 4.155
```

Para obtener la media recortada al 10 %, o sea la media aritmética de todos los valores del vector, exceptuando el 10% de los que están por arriba y el 10% de los que están por abajo hacemos:

```
> mean(tenrot,trim=0.1)
[1] 4.1535
```

2.6 Medidas de dispersión

La instrucción **var** nos devuelve un estimador insesgado de la varianza muestral, que se denomina *cuasivarianza*. Para obtener el valor de la varianza de los datos debemos multiplicar el valor anterior por **(length(tenrot)-1)/length(tenrot))** donde **length(tenrot)** es la longitud del vector, o sea el número de datos (en nuestro ejemplo 50):

```
> varianza.muestra<-((50-1)/50)*var(tenrot)
> varianza.muestra
[1] 0.1583210
> sqrt(varianza.muestra)
[1] 0.3978957
```

El valor anterior es la desviación típica de los datos. Este valor también puede obtenerse a través de la función **sd** que da la *cuasidesviación típica.*

```
> desvtípica.muestra<-sqrt(varianza.muestra)
> desvtípica.muestra
[1] 0.3978957
> sqrt((50-1)/50)*sd(tenrot)
[1] 0.3978957
```

2.7 Percentiles

Los percentiles son los valores que dividen el rango de los datos en cien unidades, de modo que, por ejemplo, el percentil 20 es el valor que deja por debajo de sí el 20% de las observaciones; el percentil 50 deja por debajo la mitad de las observaciones, o sea es la mediana, etc. La función **quantile** sirve para calcular percentiles.

```
>quantile(tenrot,0.3)#Con esta sentencia calculamos el percentil 30
    30%
3.957
>quantile(tenrot,0.5)#Volvemos a calcular la mediana de otro modo
    50%
4.155
```

Mediante la función anterior se calcula, por defecto, un resumen de cinco números: mínimo, primer cuartil (Q1), mediana, tercer cuartil (Q3), y máximo.

```
> quantile(tenrot)
  0%    25%    50%    75%   100%
3.2700 3.9025 4.1550 4.4275 4.8600
```

Si además de los cinco valores anteriores queremos conocer la media podemos hacer:

```
> summary(tenrot)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.270   3.903   4.155   4.145   4.428   4.860
```

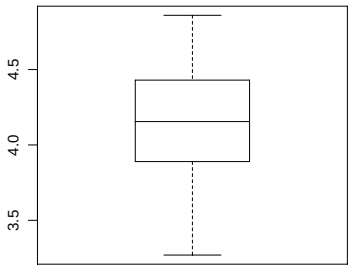
Conocidos los valores anteriores podemos calcular el recorrido intercuartílico RIQ = Q3-Q1.

```
> RIQ<-quantile(tenrot,0.75)-quantile(tenrot,0.25)
> RIQ
   75%
 0.525
```

2.8 Diagrama de cajas (box plot)

Para dibujar un diagrama de cajas utilizamos la función **boxplot**, que en su forma más simple (tiene varias versiones) es:

```
> boxplot(tenrot)
```



Los extremos del diagrama son el valor máximo y el mínimo. Si hay *outliers* (valores mayores que 3/2*RIQ por encima de Q3 o por debajo de Q1) vienen señalados por un pequeño círculo y, entonces, los extremos que aparecen son el máximo y el mínimo de los valores que quedan al eliminar los valores atípicos.

PRÁCTICA 2: EJERCICIOS

■Ejercicio 2-1: Las longitudes en micras de 25 grietas medidas en una pieza de hormigón han sido:

50-68-84-86-64-67-78-87-110-85-52-65-52-93-72-70-

105-85-30-42-74-30-70-65-49

1) Agrupar los datos en las siguientes clases de diferente amplitud:

[30,40),[40,50),[50,60),[60,70),[70,75),[75,85),[85,90), [90,110),[110,∞]

2) Construir una tabla de frecuencias en la que figuren las siguientes columnas:

Clases / Frecuencia absoluta / Frecuencia relativa / Frecuencia absoluta acumulada / Frecuencia relativa acumulada.

■Ejercicio 2-2: Calcular la media, varianza y desviación típica de los siguientes datos agrupados en clases:

Clase	x_i	F_i
[10.5,21.5)	16	6
[21.5,32.5)	27	8
[32.5,43.5)	38	7
[43.5,54.5)	49	17
[54.5,65.5)	60	18
[65.5,76.5)	71	10
[76.5,87.5)	82	3
[87.5,98.5)	93	3

■Ejercicio 2-3: Dado el conjunto de datos: {2,4,4,0,3,2,5,6,26,13/4} calcular la media, mediana, varianza, desviación típica y los cuartiles. Representar este conjunto de datos usando dos representaciones distintas: diagrama de barras y diagrama de sectores.

■Ejercicio 2-4: Generar una lista aleatoria de 100 números naturales comprendidos entre 1 y 10. Para los datos obtenidos calcular la media, mediana, varianza, desviación típica y cuartiles. Representar los datos con un diagrama de barras y dibujar un diagrama boxplot, indicando los datos atípicos, si existen.

■Ejercicio 2-5: Los valores siguientes son los precios de venta en miles de euros de 35 terrenos rurales de Bizkaia en el año 2006:

115,232,181,161,155,137,165,171,139,130,406,69,171,135,135,132,88,410,87,
90,123,157,345,323,411,334,80,87,235,198,450,223,602,415,950

Se pide:

1º) Histograma.

2º) Diagrama boxplot.

3º) Detectar datos atípicos.

4º) Media, mediana, desviación típica y RIQ antes y después de eliminar los datos atípicos, si existen.