Estadística descriptiva. Selección de actividades resueltas

2023-10-08

Actividad 1: cómputo del tiempo de CPU.

Medidas de tendencia central. Medidas de dispersión. Histogramas. Diagramas de caja.

El fichero TCPU.csv contiene los resultados de un test que consiste en ejecutar aleatoriamente diferentes programas en un ordenador y medir el tiempo de CPU consumido (en milisegundos) para cada programa (variable TCPU). También conocemos la longitud del código de cada uno de los programas ejecutados (variable LCODI). En este problema estudiaremos la variable TCPU.

a) Importad el fichero TCPU.csv.

Importamos los datos con la instrucción siguiente (indicando que las tres primeras filas contienen información sobre las variables, no datos; se ve abriendo el fichero con un procesador de textos o una hoja de cálculo cualquiera).

```
setwd("/home/xto/Documentos/01_UOC/Estadistica/Reto01/datosreto1/")
test1 <- read.table("ActR01TCPU.csv", skip = 3, sep =";", header = TRUE)
```

Siempre es conveniente después de la importación ver cómo han quedado las variables:

```
head(test1)
```

```
##   TCPU LCOD
## 1  127  146
## 2   83   80
## 3   85   60
## 4   93   90
## 5  103   58
## 6   80   88
```

b) Indicad el tipo de variable considerada.

La variable TCPU es una variable cuantitativa continua. Al tratarse de tiempo (milisegundos), podemos tomar un tiempo entre dos datos cualesquiera.

c) Calculad la media, la mediana, la desviación típica y los cuartiles, el máximo y el mínimo.

```
tmean <- mean(test1$TCPU)
tmean
```

```
## [1] 99.87
```

```
tmedian <- median(test1$TCPU)
tmedian
```

```
## [1] 101
```

```
tsd <- sd(test1$TCPU)
tsd
```

```
## [1] 21.55831
```

```r
tquantile1 <- quantile(test1$TCPU, 0.25)
tquantile1
```

```
## 25%
##  87
```

```r
tquantile3 <- quantile(test1$TCPU, 0.75)
tquantile3
```

```
##    75%
## 115.25
```

```r
tmax <- max(test1$TCPU)
tmax
```

```
## [1] 149
```

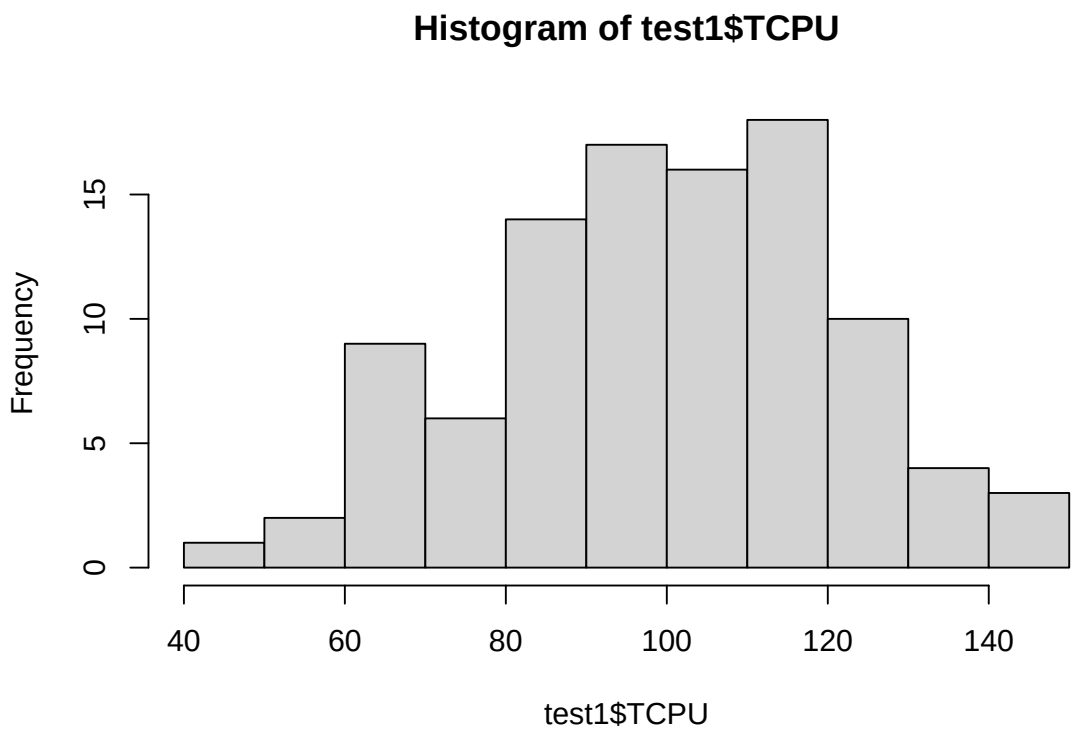
```r
tmin <- min(test1$TCPU)
tmin
```

```
## [1] 48
```

Por tanto, la media de la variable *TCPU* vale 99.87, la mediana 101, la desviación típica 21.55831. Los cuantiles primero y tercero valen 87 y 115.25 respectivamente, y el máximo y el mínimo, 149 y 38 respectivamente.

d) Dibujad un histograma de la variable y comentad su forma.

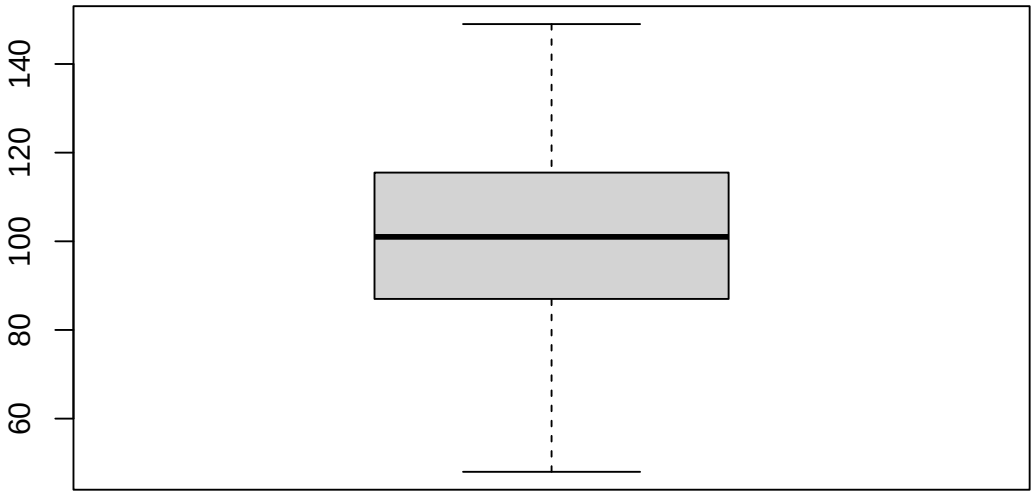
```r
hist(test1$TCPU)
```



Tiene una forma bastante simétrica sin ningún valor atípico.

e) Construid un diagrama de caja de la variable y comentad su forma. Indicad si hay datos anómalos o atípicos.

```
boxplot(test1$TCPU)
```



Comprobamos de nuevo la simetría de la variable, vemos cómo la caja es simétrica y no existen datos atípicos.

f) Comentad el estudio realizado.

Como conculusión, podemos afirmar que la variable *TCPU* es una variable continua con una distribución bastante simétrica y no hay datos atípicos.

Actividad 2: cómputo del tiempo de *CPU* agrupado.

Agrupamiento de datos estadísticos.

Con los datos de la actividad anterior, queremos tabular los datos para estudiar mejor la variable. Para hacerlo distribuiremos los tiempos de ejecución en tres categorías: "TC" (tiempo en el intervalo [48,81]), "T" (tiempo en el intervalo (81,114]), "TL" (tiempo en el intervalo (114,149]) creando la variable CLT. Para estudiar la variable CLT, se piden los resúmenes numéricos que ayuden a entender la distribución de la variable y un gráfico explicativo de la variable.

Como indica el enunciado de la actividad, agrupamos la variable *TCPU* usando la función *cut* de *R* e indicando los intervalos de agrupamiento de la forma siguiente:

```
CLT <- cut(test1$TCPU, breaks=c(48,81,114,149), labels=c('TC','T','TL'), include.lowest=TRUE)
```

Hemos creado una variable *CLT* que representa la variable *TCPU* agrupada y computamos los primeros valores:

```
head(CLT)
```

```
## [1] TL T  T  T  T  TC
## Levels: TC T TL
```

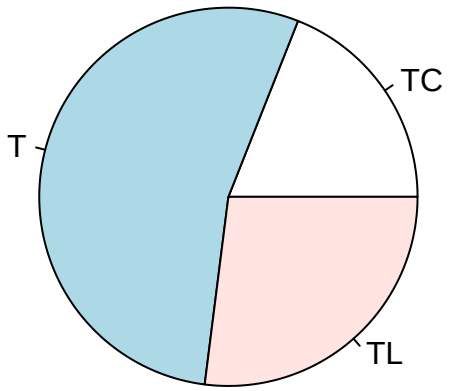
Los resúmenes numéricos para la variable *CLT* serán una tabla de frecuencias. Para esto, usamos la función *table* de *R*:

```
table(CLT)
```

```
## CLT
## TC  T TL
## 19 54 27
```

Vemos que los programas de media duración son los más abundantes y los de duración corta, los menos abundantes. Un gráfico explicativo de la variable podría ser un gráfico de sectores. Para ello, usamos la función *pie* de *R*:

```
pie(table(CLT))
```



Al observar el gŕafico obtenemos las mismas conclusiones que antes.

Actividad 3: inmersión de las tecnologías de la información y comunicación en los municipios.

Medidas de tendencia central. Medidas de dispersión. Histogramas. Diagramas de caja.

En el fichero *TICM.csv* se recogen los resultados de unas encuestas en diferentes municipios sobre el uso de las *TIC* el año 2007. De cada municipio tenemos cuatro valores: *PUORD* (proporción de hogares que tienen

ordenador), *PBA* (proporción de hogares que tienen banda ancha), *PUSUA* (proporción de habitantes que han utilizado el último mes el ordenador) y *PEMAIL* (proporción de habitantes que han usado el correo electrónico el último mes). En este problema estudiaremos y compararemos las variables *PUORD* y *PUSUA*.