Ejercicios Módulo 01

2023-10-07

ESTADÍSTICA DESCRIPTIVA

1. Se ha medido el tiempo en segundos que tarda en arrancar la última versión del programa Macrohard Phrase en los ordenadores de nuestra empresa según el sistema operativo con el que funcionan. Los resultados han sido los siguientes:

- En los ordenadores equipados con Doors95: 27, 25, 50, 33, 25, 86, 28, 31, 34, 36, 37, 44, 20, 59 y 85 segundos.
- En los ordenadores equipados con Doors98: 33, 7, 25, 14, 5, 31, 19, 10, 29 y 18 segundos.

Calculad los cinco números resumen y la media de la distribución correspondiente al tiempo que el programa tarda en arrancar y dibujad algunos gráficos que os parezcan relevantes para comparar el tiempo que el programa tarda en arrancar según el sistema operativo. A partir de estos gráficos, comparad el comportamiento del programa según el sistema operativo y explicad si creéis que hay diferencia entre utilizar Doors95 y Doors98.

1.1. Primero introducimos los vectores y ordenamos los datos:

```
D95 <- c(27, 25, 50, 33, 25, 86, 28, 31, 34, 36, 37, 44, 20, 59, 85)
D98 <- c(33, 7, 25, 14, 5, 31, 19, 10, 29, 18)
#Ordenamos los vectores anteriores
D95O <- sort(D95)
D98O <- sort(D98)
#Combinamos los dos vectores en una sola tabla
sistemas <- cbind(D95O, D98O)
```

```
## Warning in cbind(D95O, D98O): number of rows of result is not a multiple of
## vector length (arg 2)
```

Los datos ordenados son:

D95: 20, 25, 25, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

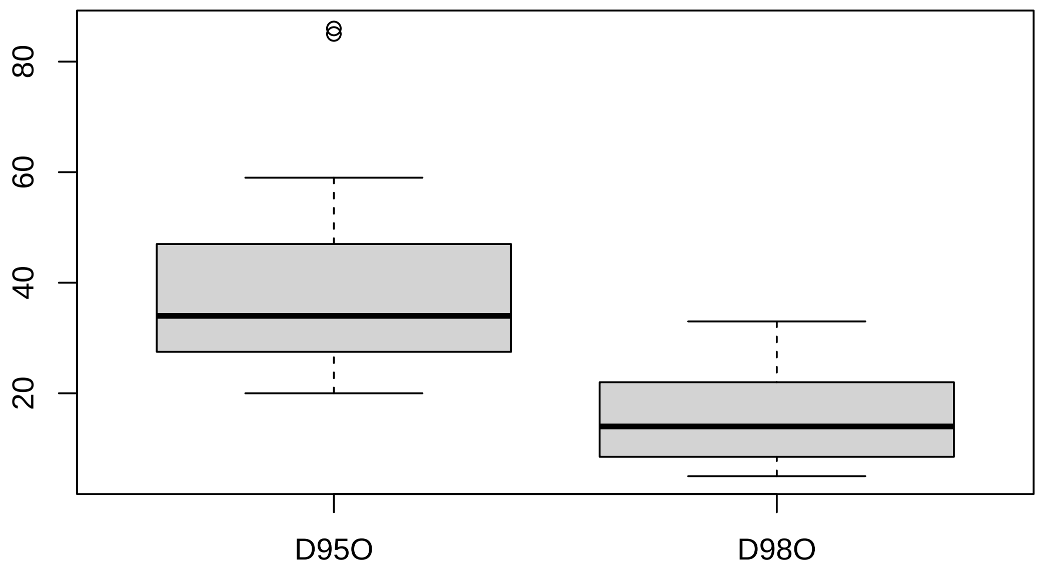
D98: 5, 7, 10, 14, 18, 19, 25, 29, 31, 33

1.2. Dibujamos una tabla con los cinco números resumen y la media:

Sistema	Media	Minimo	Q1	Q2	Q3	Maximo
D95	41.3333333	20	27	34	50	86
D98	19.1	5	9.25	18.5	29.5	33

1.3. Dibujamos los diagramas de caja correspondientes a las dos distribuciones:

```
boxplot(sistemas)
```



A partir de los resúmenes numéricos y de los gráficos, podemos concluir que **D98 tarda mucho menos en arrancar que D95**. Destacamos los aspectos siguientes:

- Tanto la media como la mediana de D98 son aproximadamente la mitad que las de D95.
- La mediana de D98 es menor que el mínimo de D95; esto significa que la mitad de las veces D98 tarda menos en arrancar que la vez que más rápido ha arrancado D95.
- El máximo de D98 es menor que la mediana de D95, es decir, la mitad de las veces que arrancamos D95 tarda más que la vez que más ha tardado D98.
- El diagrama de caja de D98 es mucho más simétrico y no tiene ninguna cola hacia valores alejados o mayores de los normales. Esta simetría sugiere que D98 tiene un comportamiento más regular y más estable que D95.

2. a) Confeccionad una lista de 10 números tales que mínimo = 2, máximo = 20, primer cuartil = 5, mediana = 10 y tercer cuartil = 19.

La mediana ocupará la posición $10 + 1/2 = 5.5$ y por tanto, se calculará como valor medio de la 5^a y 6^a observaciones. En estas circunstancias, el primer cuartil debe ser la 3^a observación y el tercer cuartil, la 8^a. En la tabla siguiente representamos estos datos:

posicion	1	2	3	4	5	6	7	8	9	10
	Minimo		Q1		M1	M2		Q3		Maximo
	2		5					19		20

En la tabla, vemos que el máximo, el mínimo, el primer y el tercer cuartil están fijados. Podemos acabar de rellenar la tabla de diferentes maneras, siempre que las observaciones aparezcan ordenadas y que la mediana tenga el valor 10. Por ejemplo:

posicion	1	2	3	4	5	6	7	8	9	10
	Minimo		Q1		M1	M2		Q3		Maximo
	2	3	5	9	10	10	18	19	19	20
	2	2	5	6	8	12	13	19	19.5	20
	2	3	5	5	5	15	18	19	20	20

b) Igual que antes, pero con la media = 11.

Si la media tiene que ser 11, tenemos que tener cuidado de que la suma de los valores sea igual a $10 \times 11 = 110$. Como las posiciones $1, 3, 8$ y 10 tienen el valor fijado y la suma de los valores 5^o Y 6^o debe ser 20, la suma de los valores que aparecen en las posiciones $2, 4, 7$ y 9 tiene que ser $110 - 20 - 2 - 5 - 19 - 20 = 44$. Un ejemplo podría ser:

posicion	1	2	3	4	5	6	7	8	9	10
	Minimo		Q1		M1	M2		Q3		Maximo
	2	3	5	5	8	12	17	19	19	20

c) En las condiciones del punto a), ¿la media podría ser igual a 21? (Recordad que la suma de las desviaciones con respecto a la media ha de ser 0).

La media no puede superar la observación mayor. Una posible razón es que si lo hiciera, todas las desviaciones con respecto a la media serían estrictamente negativas y por tanto, la suma de las desviaciones no podría ser 0.

3. Los siguientes datos corresponden al número de veces que el programa Minihard Word se "colgó" durante un mes en cada uno de los 50 ordenadores de nuestra empresa. Comprobad que se satisface la regla de Tchebichev para m = 1, m = 2 y m = 3.

0	9	12	14	19
2	10	12	14	20
4	10	12	14	20
5	10	12	15	21
6	11	12	15	22
6	11	12	17	29
6	11	12	17	29
7	11	13	18	32
8	11	13	18	39
9	12	14	19	39

Figure 1: Tabla de observaciones

- Primero calculamos la media y la desviación típica:

```
datos <- c(0,2,4,5,6,6,6,7,8,9,9,10,10,10,11,11,11,11,11,12,12,12,12,12,12,12,12,
13,13,14,14,14,14,15,15,17,17,18,18,19,19,20,20,21,22,29,29,32,39,39)
media <- mean(datos)
desviacion <- sd(datos)
```

La media es 14.28 y la desviación típica es 8.1842058.

- Ahora calculamos los intervalos para m = 1, m = 2 y m = 3 utilizando la regla de Tchebichev $(\bar{x} - m \cdot s_x, \bar{x} + m \cdot s_x)$:
  - Para $m = 1 \Rightarrow (14.28 - 1 \cdot 8.18, 14.28 + 1 \cdot 8.18) = (6.1, 22.46)$
  - Para $m = 2 \Rightarrow (14.28 - 2 \cdot 8.18, 14.28 + 2 \cdot 8.18) = (-2.08, 30.64)$
  - Para $m = 3 \Rightarrow (14.28 - 3 \cdot 8.18, 14.28 + 3 \cdot 8.18) = (-10.26, 38.82)$
- Anotamos las observaciones en cada intervalo:
  - $(6.1, 22.46) \Rightarrow 38$
  - $(-2.08, 30.64) \Rightarrow 47$
  - $(-10.26, 38.82) \Rightarrow 48$
- Calculamos los porcentajes para cada intervalo:
  - $(6.1, 22.46) \Rightarrow \dfrac{38}{50} = 0.76 \Rightarrow 76\%$
  - $(-2.08, 30.64) \Rightarrow \dfrac{47}{50} = 0.94 \Rightarrow 94\%$
  - $(-10.26, 38.82) \Rightarrow \dfrac{48}{50} = 0.96 \Rightarrow 96\%$
- Calculamos el porcentaje mínimo previsto por la regla de Tchebichev $\left(1 - \dfrac{1}{m^2}\right)$:
  - $(6.1, 22.46) \Rightarrow 1 - \dfrac{1}{1^2} = 0$
  - $(-2.08, 30.64) \Rightarrow 1 - \dfrac{1}{2^2} = 0.75 \Rightarrow 75\%$
  - $(-10.26, 38.82) \Rightarrow 1 - \dfrac{1}{3^2} = 0.88 \Rightarrow 88\%$

La regla de Tchebichev indica el porcentaje mínimo que podemos encontrar en el intervalo. Observamos que en todos los casos hay más observaciones en los intervalos que los previstos por la regla.

- Finalmente, dibujamos una tabla donde podemos ver resumidos todos los datos:

m	Intervalo	Observaciones en el intervalo	Porcentaje de observaciones en el intervalo	Porcentaje mínimo previsto por la regla de Tchebichev
1	$(\bar{x} - 1s_x, \bar{x} + 1s_x) = (6.1, 22.46)$	38	$38/50 = 76\%$	0%
2	$(\bar{x} - 2s_x, \bar{x} + 2s_x) = (-2.08, 30.64)$	47	$47/50 = 94\%$	75%
3	$(\bar{x} - 3x_x, \bar{x} + 3s_x) = (-10.26, 38.82)$	48	$48/50 = 96\%$	88%