Inferencia Estadística

Clase 4 - Estadística Descriptiva

Paula Spano

Universidad Torcuato Di Tella

Estadística Descriptiva

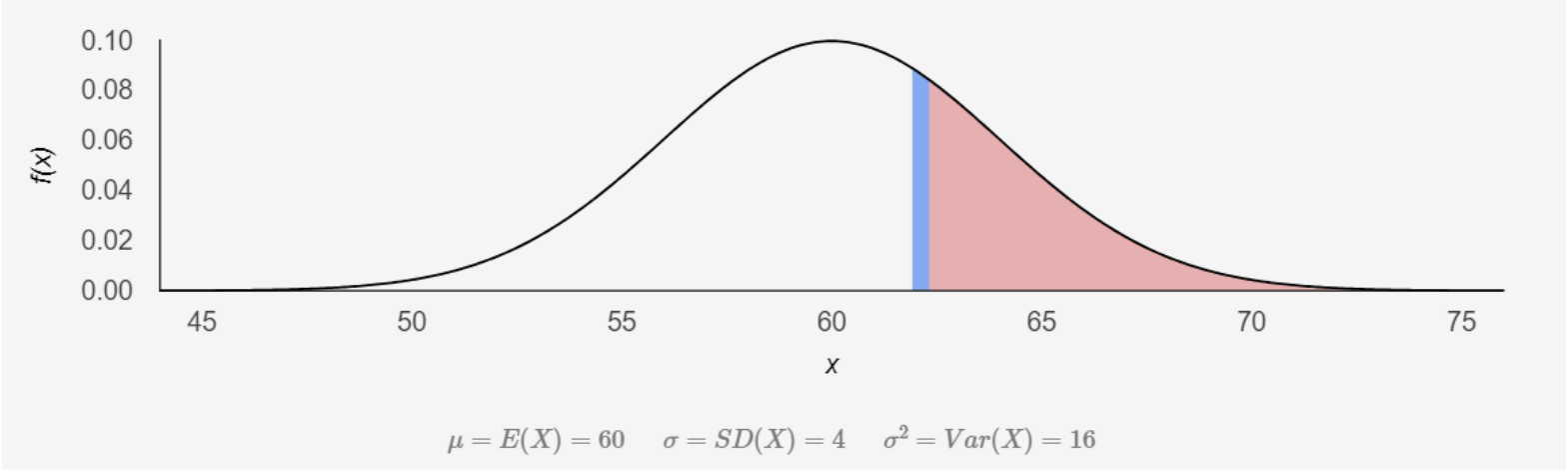
(Bibliografía: Lock - cap 2: Describing Data)

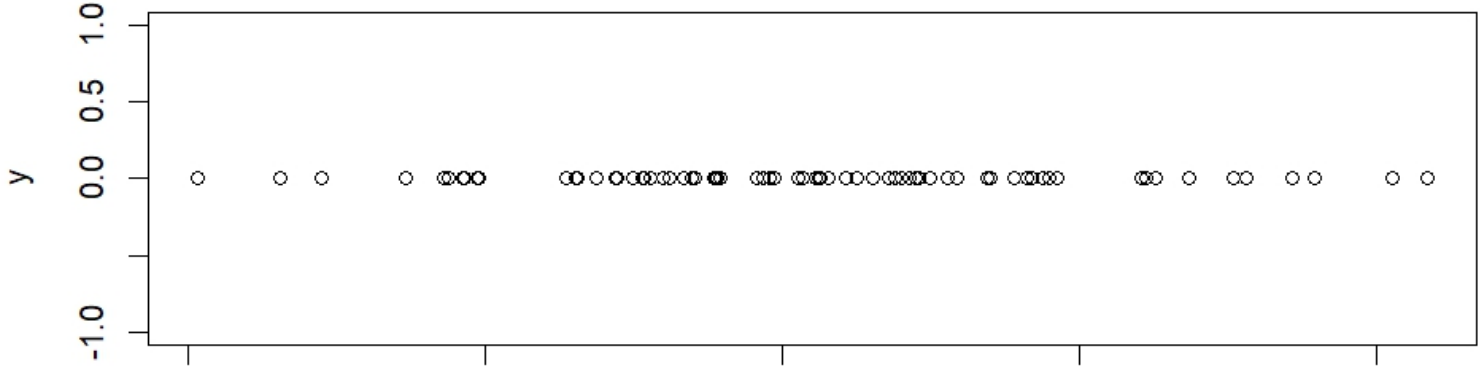
Medidas resumen

Percentiles:

¿Cuál es el tiempo que es superado por el 30% de los usuarios?

`qnorm(0.7,60,4) --> 62.0976`



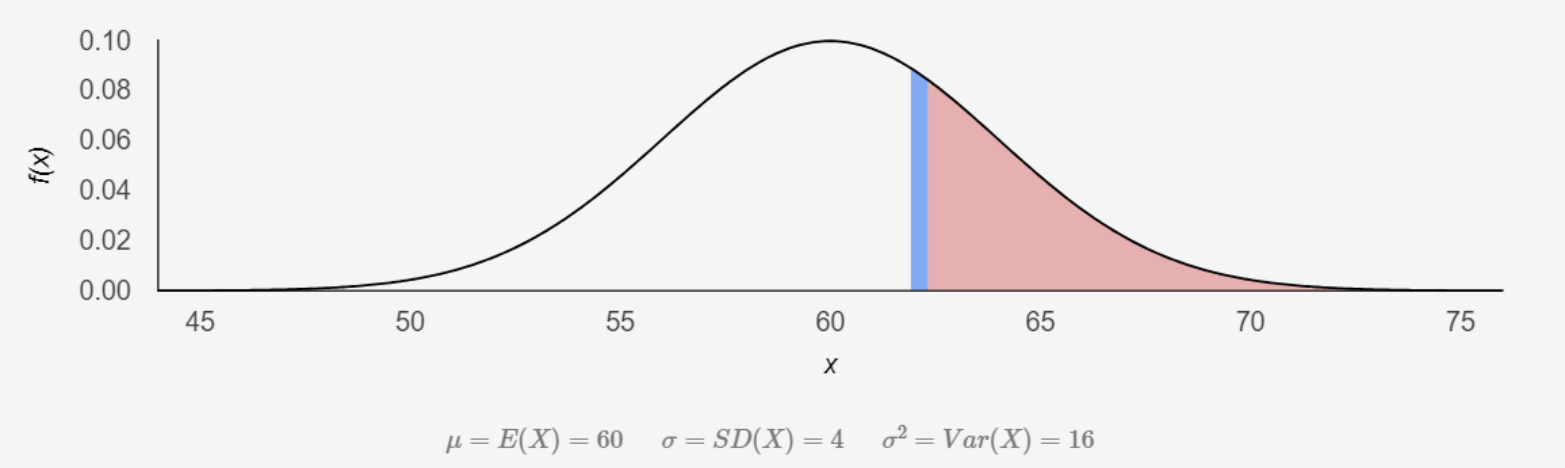


Medidas resumen

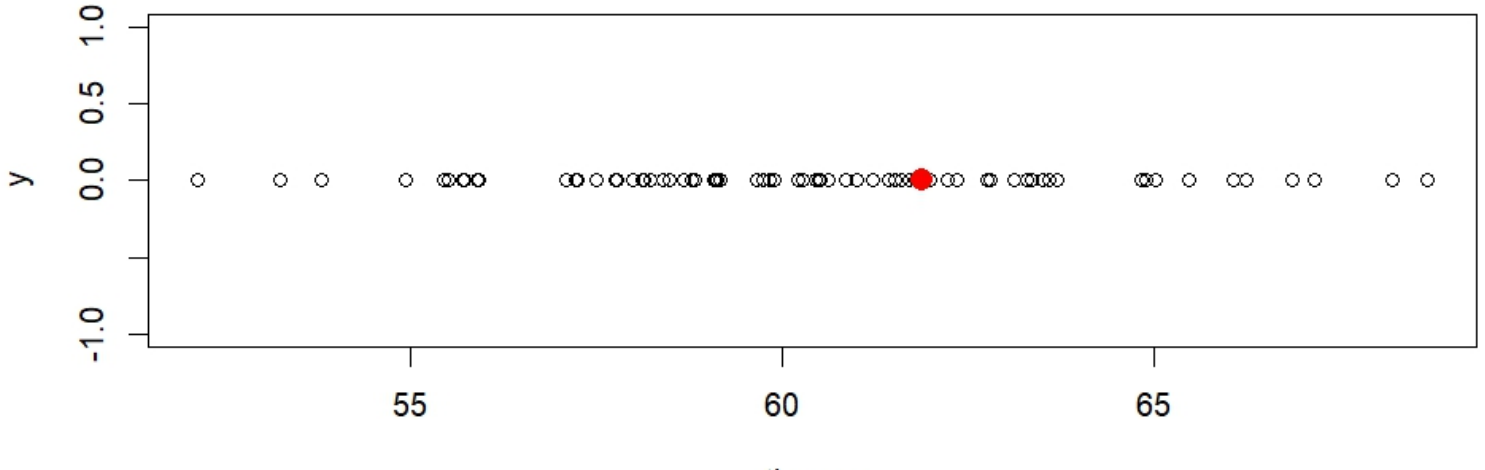
```
quantile(tiempo, probs = 0.7)
     70%
61.88799

sort(tiempo)

 [1] 52.13353 53.25323 53.80499 54.93841 54.93976 55.44745
 [7] 55.50757 55.71284 55.72871 55.89598 55.92570 57.08444
[13] 57.22117 57.25259 57.49984 57.75810 57.77664 57.99071
[19] 58.10883 58.13338 58.21735 58.38846 58.47812 58.66717
[25] 58.77615 58.81971 59.07929 59.09692 59.12810 59.16833
[31] 59.66652 59.75235 59.82852 59.88581 60.21202 60.28203
[37] 60.44273 60.49542 60.51715 60.61349 60.86377 61.01327
[43] 61.21411 61.43926 61.51856 61.60309 61.70586 61.79284
[49] 61.84366 61.99140 62.21567 62.33845 62.75456 62.80542
[55] 63.11986 63.28632 63.35115 63.51253 63.58050 63.68907
[61] 64.83185 64.89633 65.01526 65.47441 66.06588 66.23483
[67] 66.86026 67.14765 68.20034 68.67582

n*0.7
[1] 49

0.7*61.84366+0.3*61.99140
[1] 61.88798
```

Medidas resumen

`qnorm(0.7,60,4) --> 62.0976`



`quantile(tiempo, probs = 0.7) --> 61.88799`



Medidas resumen

Definición

p_α: Es el percentil muestral $\alpha\%$ de un conjunto de datos y será el valor que deja aproximadamente el $\alpha\%$ de los datos a izquierda en la muestra ordenada.

Definición

Definimos como mediana al percentil muestral 50%, es decir aquel valor que deja aproximadamente el 50% de los datos a su izquierda. Si llamamos m a la mediana de un conjunto de n datos, entonces si n es impar, m es el valor central en el conjunto de datos ordenados y, si n es par, m es el promedio de los dos valores centrales en el conjunto de datos ordenados.

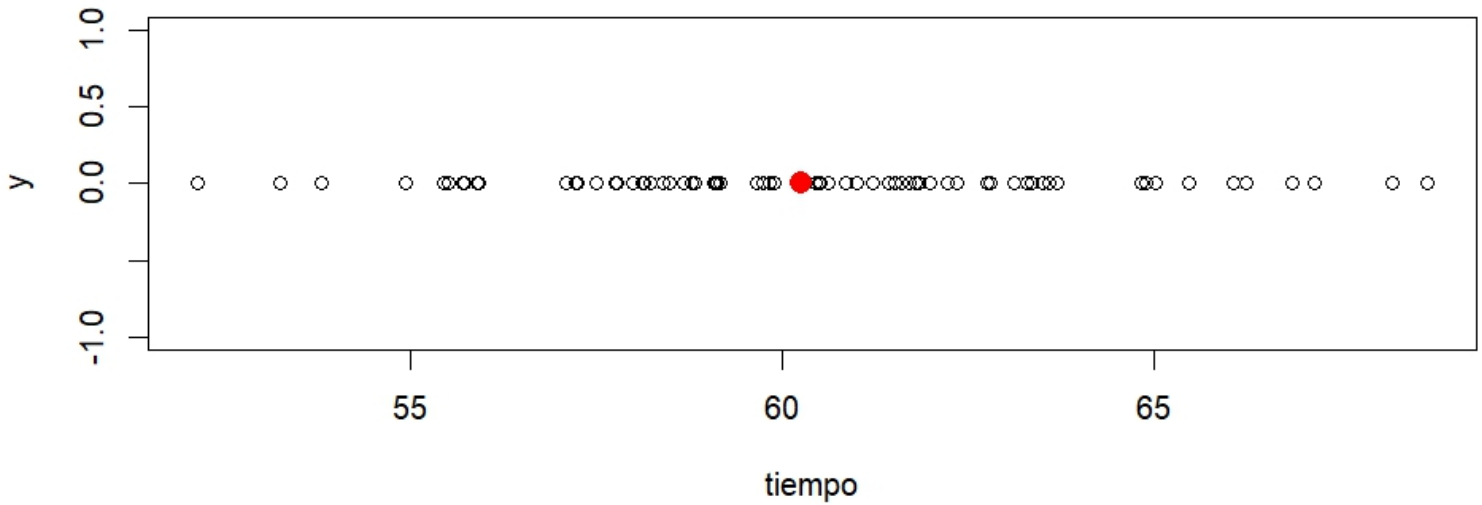
Medidas resumen

Mediana:

```
median(tiempo)
[1] 60.24703

sort(tiempo)
 [1] 52.13353 53.25323 53.80499 54.93841 54.93976 55.44745
 [7] 55.50757 55.71284 55.72871 55.89598 55.92570 57.08444
[13] 57.22117 57.25259 57.49984 57.75810 57.77664 57.99071
[19] 58.10883 58.13338 58.21735 58.38846 58.47812 58.66717
[25] 58.77615 58.81971 59.07929 59.09692 59.12810 59.16833
[31] 59.66652 59.75235 59.82852 59.88581 60.21202 60.28203
[37] 60.44273 60.49542 60.51715 60.61349 60.86377 61.01327
[43] 61.21411 61.43926 61.51856 61.60309 61.70586 61.79284
[49] 61.84366 61.99140 62.21567 62.33845 62.75456 62.80542
[55] 63.11986 63.28632 63.35115 63.51253 63.58050 63.68907
[61] 64.83185 64.89633 65.01526 65.47441 66.06588 66.23483
[67] 66.86026 67.14765 68.20034 68.67582

n*0.5
[1] 35

(60.21202 + 60.28203)/2
[1] 60.24703
```

Medidas resumen

Mediana:

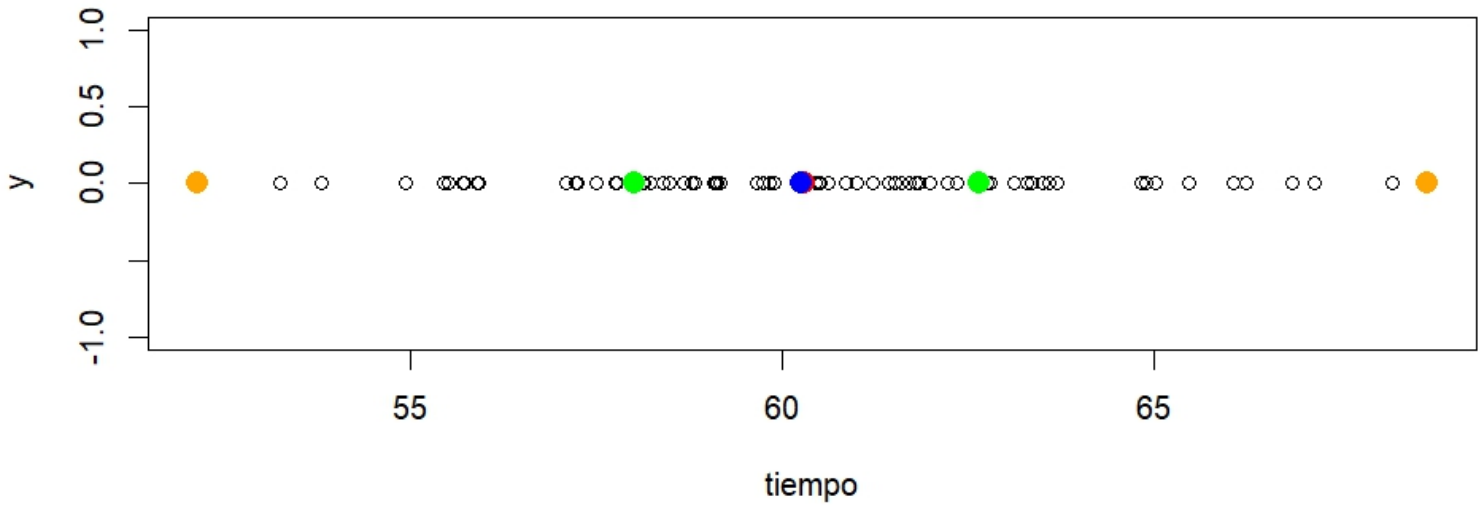


La mediana es otra medida de la centralidad de los datos

Medidas resumen

- Mínimo valor muestral
- Primer cuartil: Es el percentil muestral 25%, y se denota como $Q_1 = p_{25}$
- Tercer cuartil: Es el percentil muestral 75%, y se denota como $Q_3 = p_{75}$
- Máximo valor muestral

Medidas resumen



Ejemplo - Tiempo de Lectura (Adaptación del Lock)

```
lectura <- read.table("lectura.txt", header = TRUE)
lectura

#       persona  tiempo
# 1        1          4
# 2        2          1
# 3        3          3
# 4        4         12
# 5        5          7
# 6        6          8
# 7        7         11
# 8        8         15
# 9        9         13
# 10      10         14
```

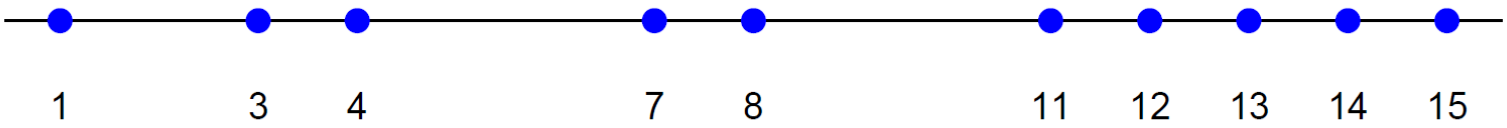
Los ordenamos:

```
x <- lectura$tiempo
sort(x)

# 1  3  4  7  8 11 12 13 14 15
```

y los graficamos:

Media muestral

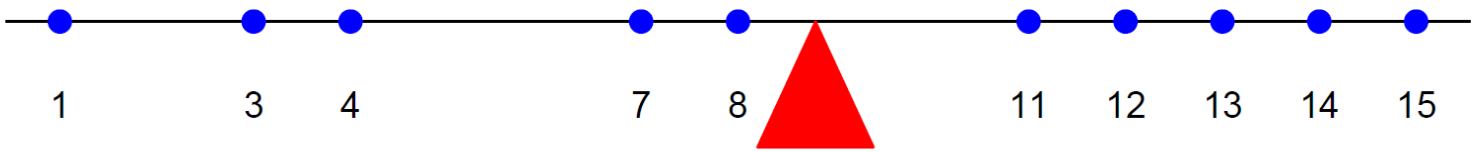
Un posible valor que resume todos los datos



Media muestral: $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

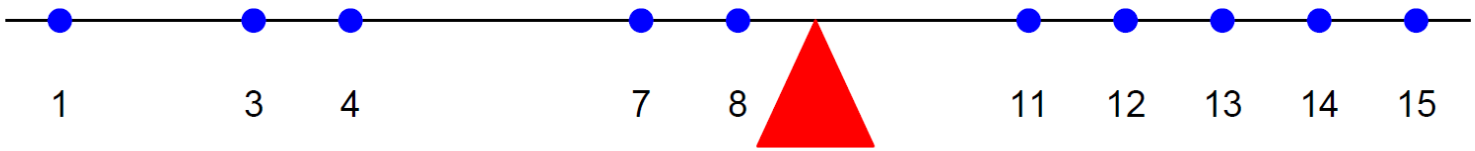
`mean(x) = 8.8`

Interpretación gráfica: centro de masa



Media muestral

Media muestral: $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$



- Medida de **centralidad** de un conjunto de datos
- Centro de masa o punto equilibro
- Estimación de la esperanza.
- En `R`: `mean()`

Ejemplo - Tiempo de Lectura

Otra medida que agregua información

Varianza muestral:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

`var(x) = 24.4`

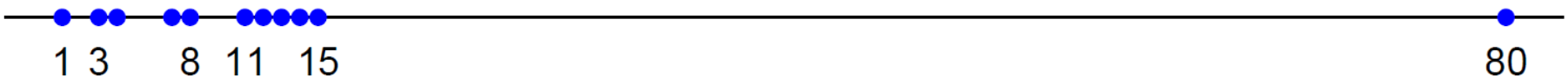
Desvío estándar muestral:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}$$

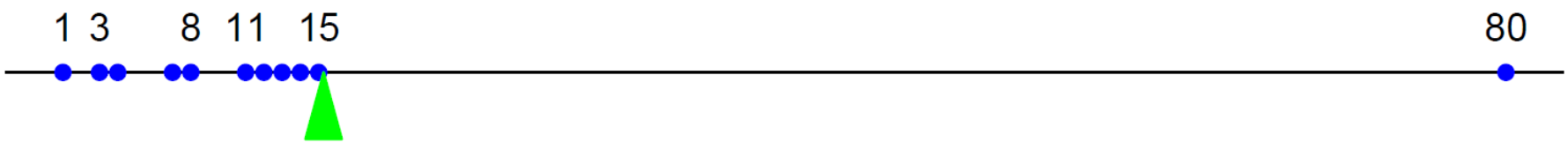
`sd(x) = 4.9`

Medidas resumen

Agreguemos a otra persona: `x_extra <- c(x, 80)`

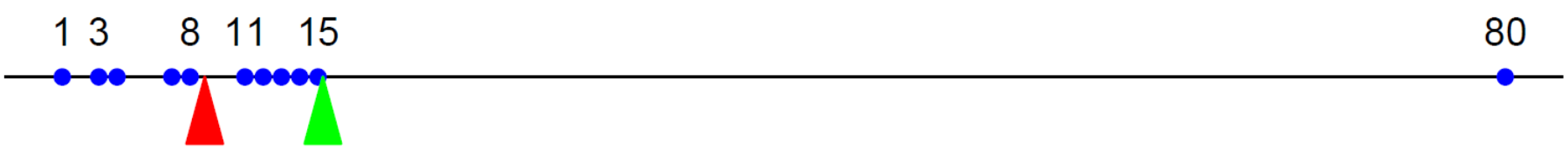


Calculemos la media: `mean(x_extra) = 15.3`

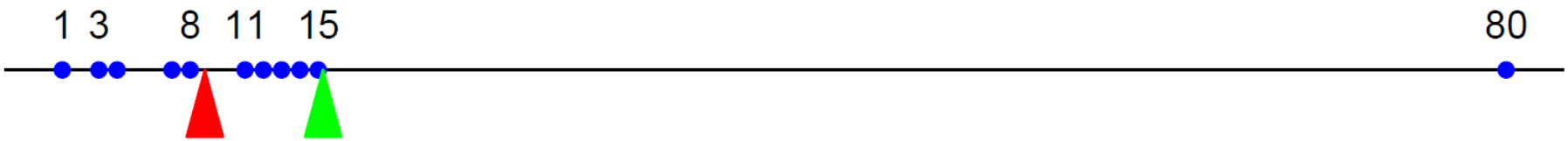


¿La media es representativa del centro de los datos?

¿Cómo es respecto de la anterior (`mean(x) = 8.8`)?



Medidas resumen



- El persona nueva está haciendo un "efecto palanca".
- La media muestral es muy sensible a "outliers" (datos atípicos).
- Queremos una medida de centralidad más "robusta".

Otra medida de centralidad: mediana

La **Mediana** es un valor que deja aproximadamente la mitad de los datos a izquierda.

Calculemos la mediana en cada conjunto de datos.

¿Cómo cambió la mediana al agregar a la persona extra?

¿Qué pasaría si la cantidad de horas que le dedica a la lectura la persona extra fuera 200 en lugar de 80?

La mediana es *robusta* a datos atípicos.

Mediana

Definición

Si llamamos m a la mediana de un conjunto de n datos, entonces si n es impar, m es el valor central en el conjunto de datos ordenados y, si n es par, m es el promedio de los dos valor centrales en el conjunto de datos ordenados.

En R: `median()`

Medidas resumen

¿Qué medida de centralidad (posición) elegimos?

¿Qué medida de dispersión elegimos? Depende de qué medida de posición elijamos:

- Si informamos la media ($\overline{x}$), la medida de dispersión natural es el desvío estándar s.
- Si informamos la mediana (m), vamos a querer una medida de dispersión robusta.

¿El desvío estándar s es robusto?

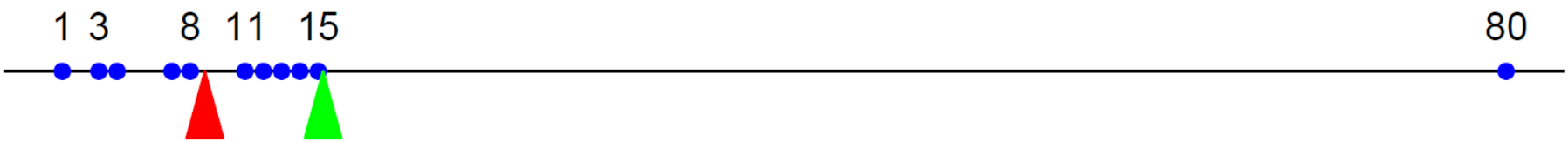
```
sd(x) = 4.9
sd(x_extra) = 21.97
```

¿Por qué el desvío es sensible a datos atípicos?

Medidas resumen

¿Por qué el desvío es sensible a datos atípicos?

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$$



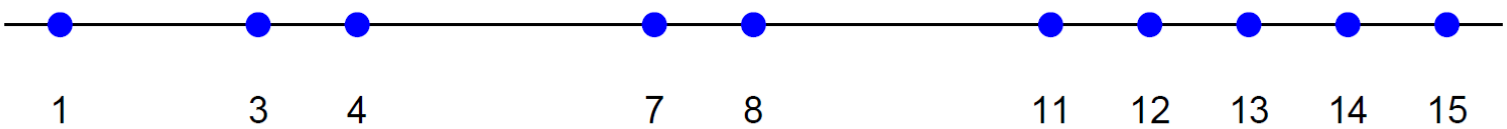
Queremos una medida de dispersión robusta, para ello vamos a utilizar los percentiles o cuantiles.

Otras medidas: percentiles (cuantiles)

Definición

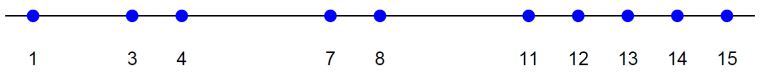
p_α: Es el percentil $\alpha\%$ de un conjunto de datos y será el valor que deja aproximadamente el $\alpha\%$ de los datos a izquierda en el conjunto ordenado.

Calculemos los percentiles 10% y 90% en el ejemplo original.




```
quantile(x, probs = 0.1) = 2.8
quantile(x, probs = 0.9) = 14.1
quantile(x, probs = 0.5) = 9.5
```

5 medidas resumen



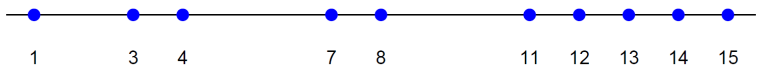
- Mínimo, máximo y mediana:

  `min(x) = 1,  median(x) = 9.5, max(x) = 15`

  - dividen al conjunto de datos en dos
  - permiten ver asimetría
- Cuartiles

Definición

$Q_1 = p_{25} \rightarrow$ *Primer cuartil* $\quad Q_3 = p_{75} \rightarrow$ *Tercer cuartil*



- dividen cada mitad en dos:
- `quantile(x, probs = 0.25) = 4.75`
- `quantile(x, probs = 0.75) = 12.75`
- ¿Quién será Q_2?
- Q_1 y Q_3 encierran el 50 % central de los datos
- Junto con la mediana, permiten ver asimetría en el 50 % central de los datos.

5 medidas resumen en R

```
summary(x)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.00    4.75    9.50    8.80   12.75   15.00
```

```
summary(x_extra)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.00    5.50   11.00   15.27   13.50   80.00
```

Nos faltaba definir una medida de dispersión robusta.

Distancia intercuartil

Definición

La distancia intercuartil de un conjunto de datos es

$$IQR = Q_3 - Q_1$$

¿Por qué es una medida de dispersión?

¿Es robusta? ¿Por qué?

En R:

```
quantile(x, probs = 0.75) - quantile(x, probs = 0.25)
= 12.75 - 4.75 = 8
```

```
IQR(x) = 8
```

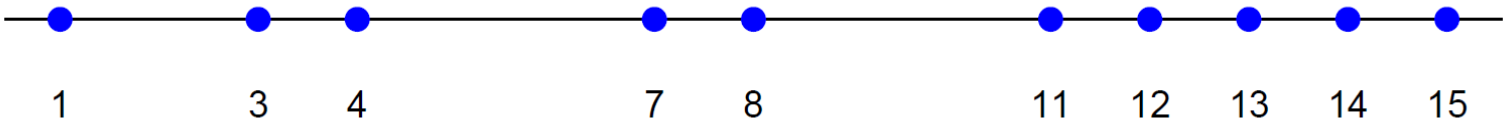
```
IQR(x_extra) = 8
```

Medidas resumen

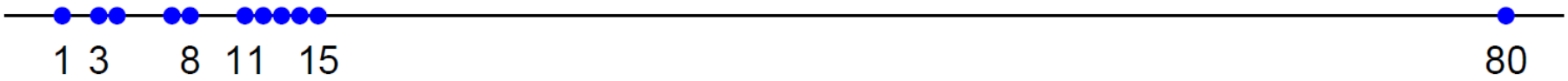
Medida	Fórmula	Función de R
media	$\frac{1}{n}\sum_{i=1}^{n} x_i$	`mean()`
desvío estándar	$\mathrm{s} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$	`sd()`
varianza	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2$	`var()`
percentil α	p_α	`quantile(, probs =` $\frac{\alpha}{100}$`)`
mediana	$m = p_{50}$	`median()`
1er cuartil	$Q_1 = p_{25}$	`quantile(,probs =` 0.25`)`
3er cuartil	$Q_3 = p_{75}$	`quantile(,probs =` 0.75`)`
dist. intercuartil	$IQR = Q_3 - Q_1$	`IQR()`
mínimo	$\min_{1\leq i\leq n}\{x_i\}$	`min()`
máximo	$\max_{1\leq i\leq n}\{x_i\}$	`max()`

Ejemplo: Tiempo de Lectura

Datos:



Agreguemos a otra persona: `x_extra <- c(x, 80)`



Datos atípicos

Dijimos que la persona extra es un dato atípico (outlier), pero, ¿cómo definimos formalmente cuándo un dato es un outlier?

Outlier: definición formal

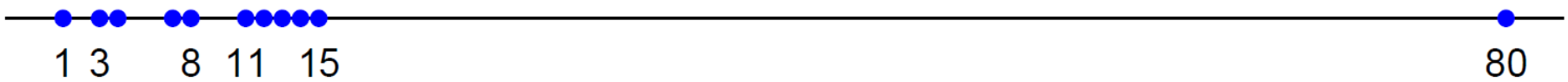
Definición

Un dato x_i se considera un outlier en un conjunto de datos si

$$x_i < Q_1 - 1.5\,IQR \quad \text{ó} \quad x_i > Q_3 + 1.5\,IQR$$

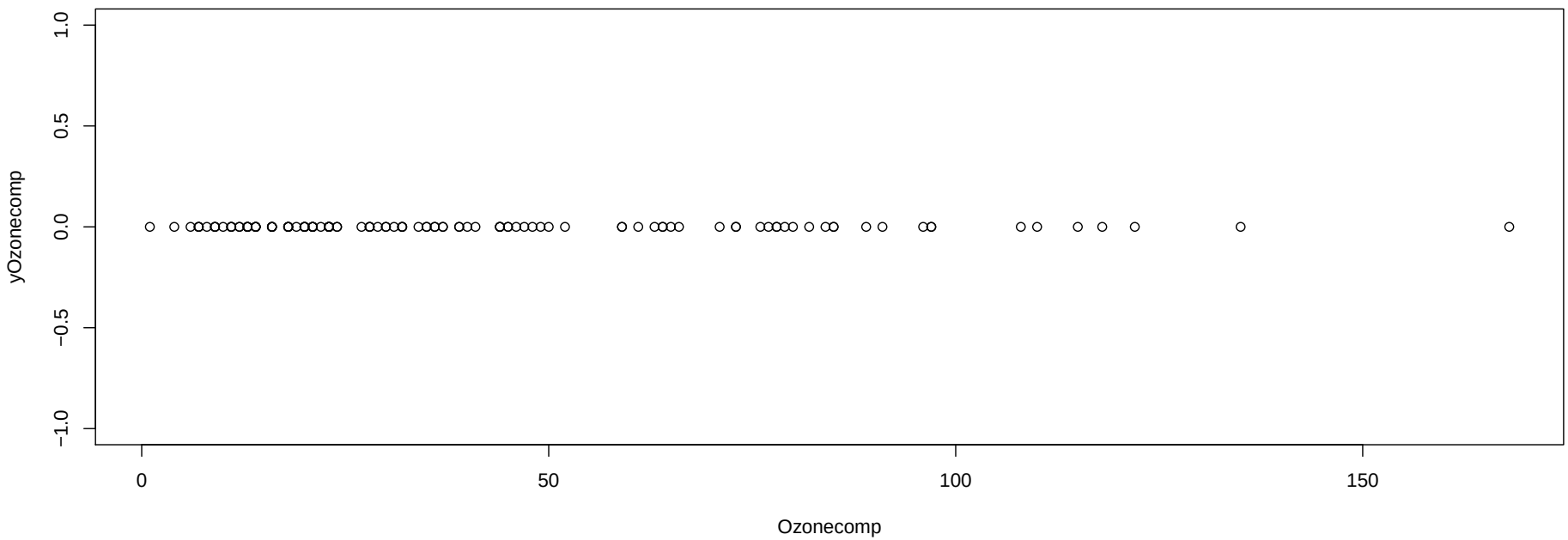
La persona extra, ¿es un outlier?

Q_1 =`quantile(x_hum, probs = 0.25) = 5.5`
Q_3 =`quantile(x_hum, probs = 0.75) = 13.5`



Airquality

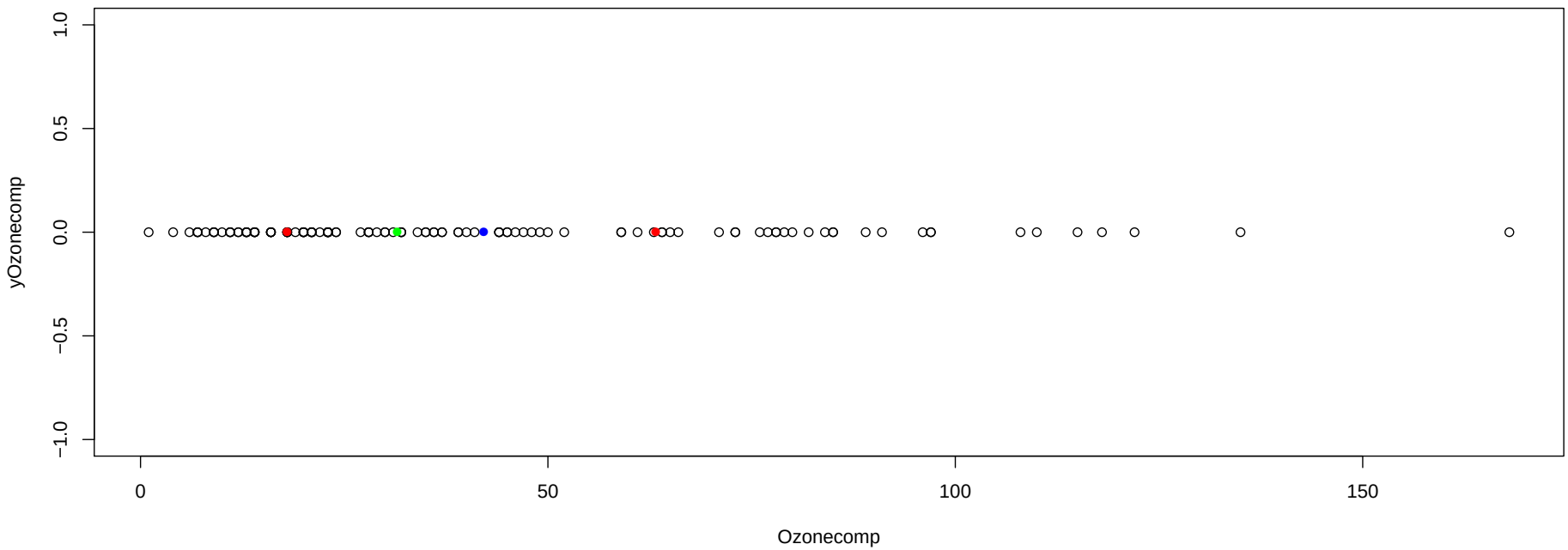
```
OzoneComp <- na.exclude(airquality$Ozone)
n <- length(OzoneComp)
n
116
```



Airquality

```
OzoneMed <- summary(OzoneComp)
OzoneMed
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   18.00   31.50   42.13   63.25  168.00
```



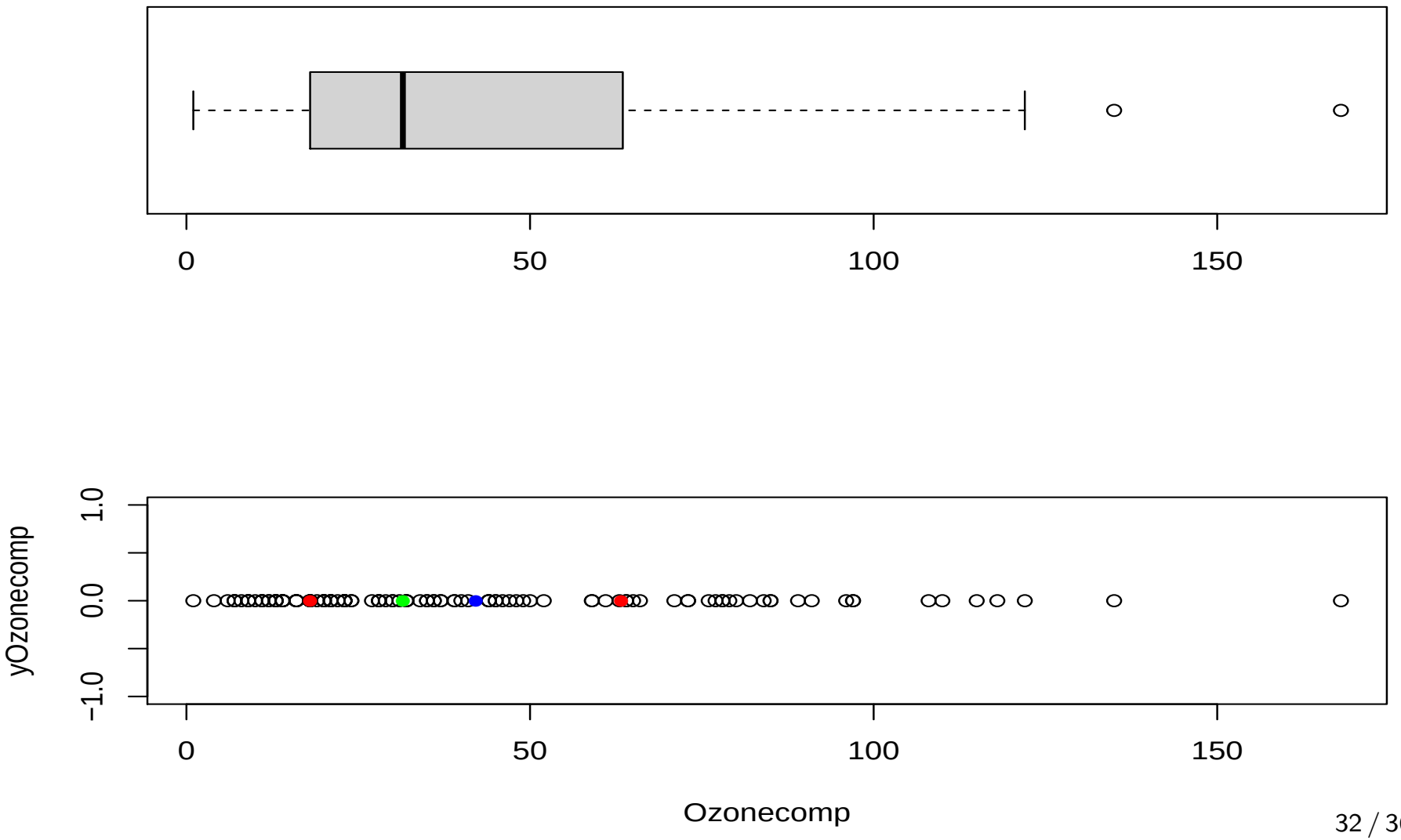
Boxplot

Boxplot: Es un gráfico que se utiliza para visualizar la forma general del conjunto de datos, identificando el 50% de los centrales, los cuartiles y posibles datos atípicos.

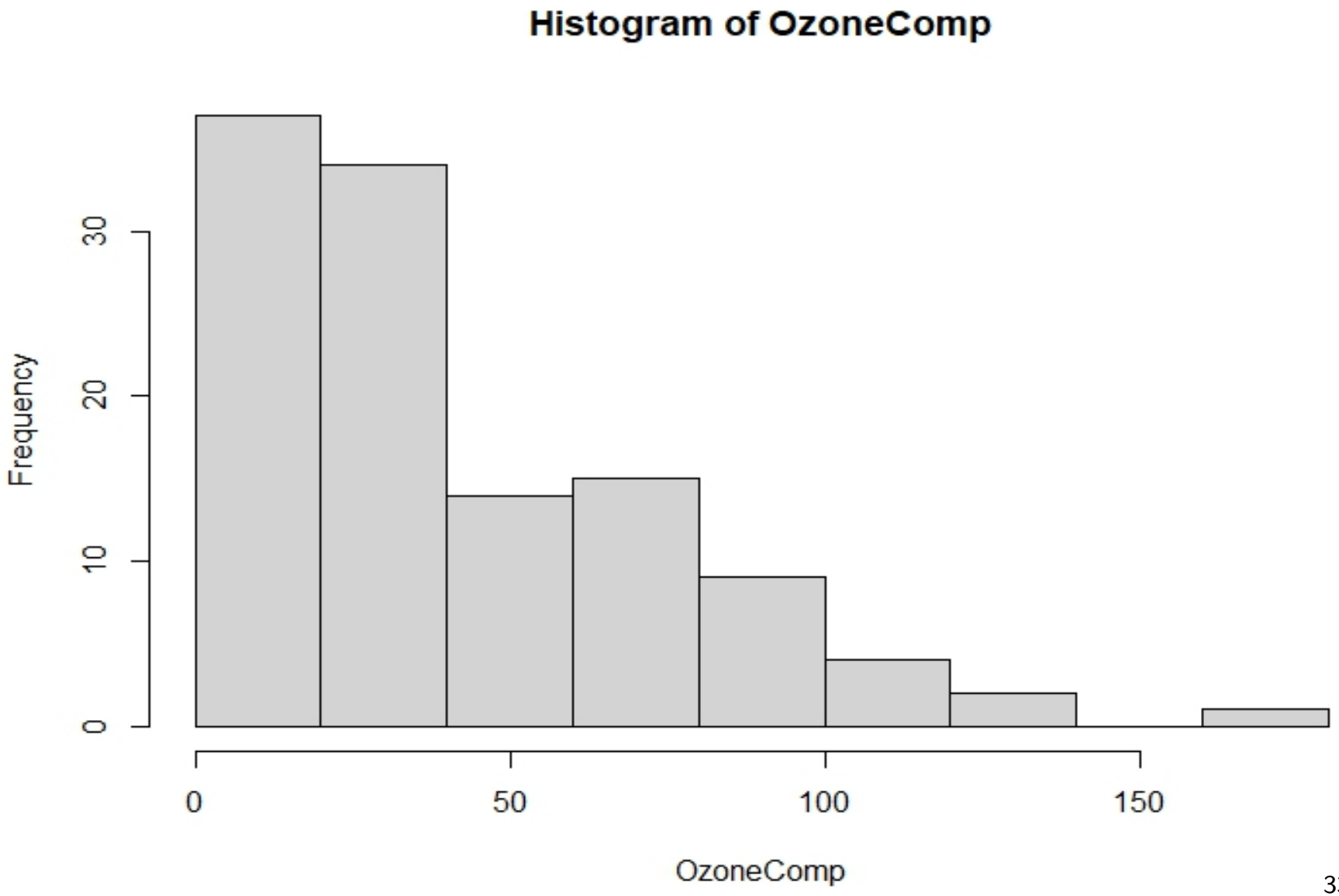
Construcción:

- Dibujar una caja entre Q_1 y Q_3
- Dividir la caja con una línea a la altura de la Mediana
- Para representar los datos que NO son outliers: Dibujar una línea desde cada cuartil (Q_1 y Q_3) hasta la observación más extrema que no sea un dato atípico. (A estas líneas se las denomina bigotes)
- Marcar un punto por cada valor clasificado como outlier.

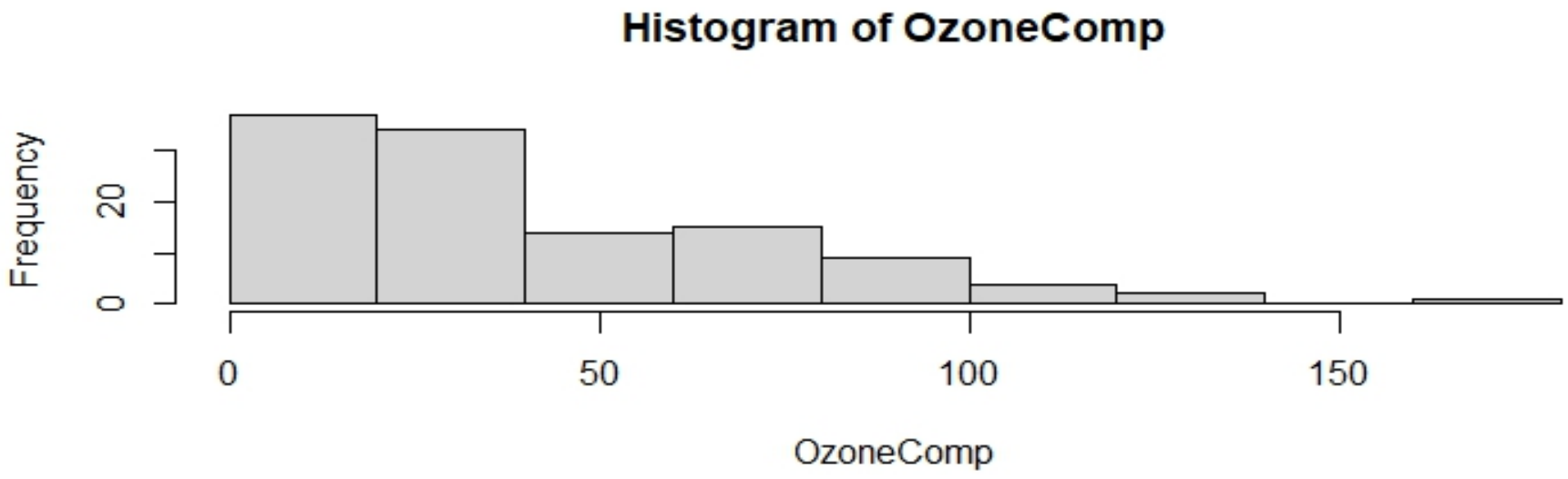
Airquality

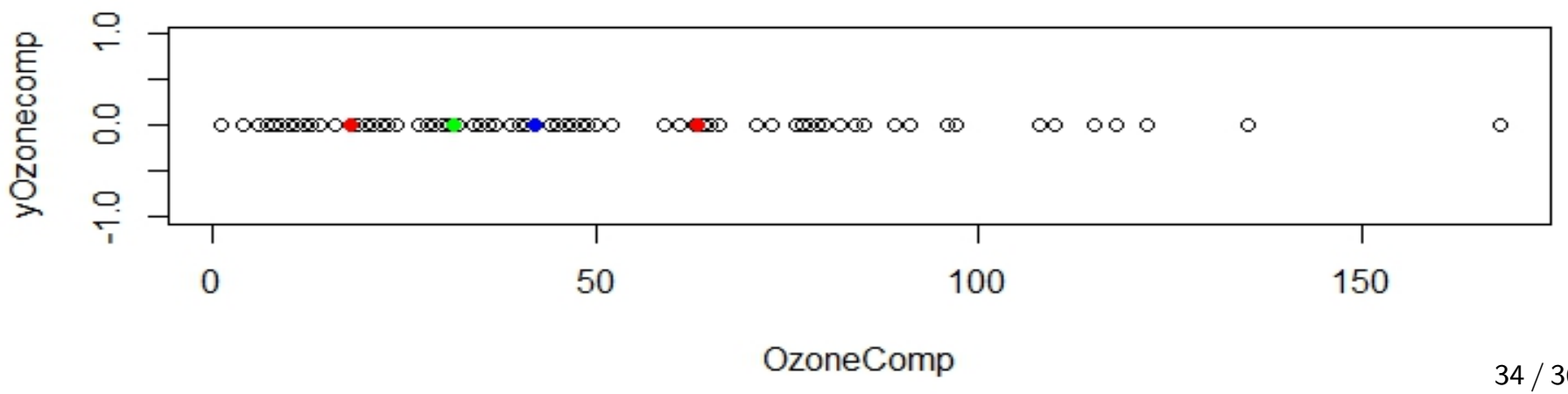


Airquality - Histograma

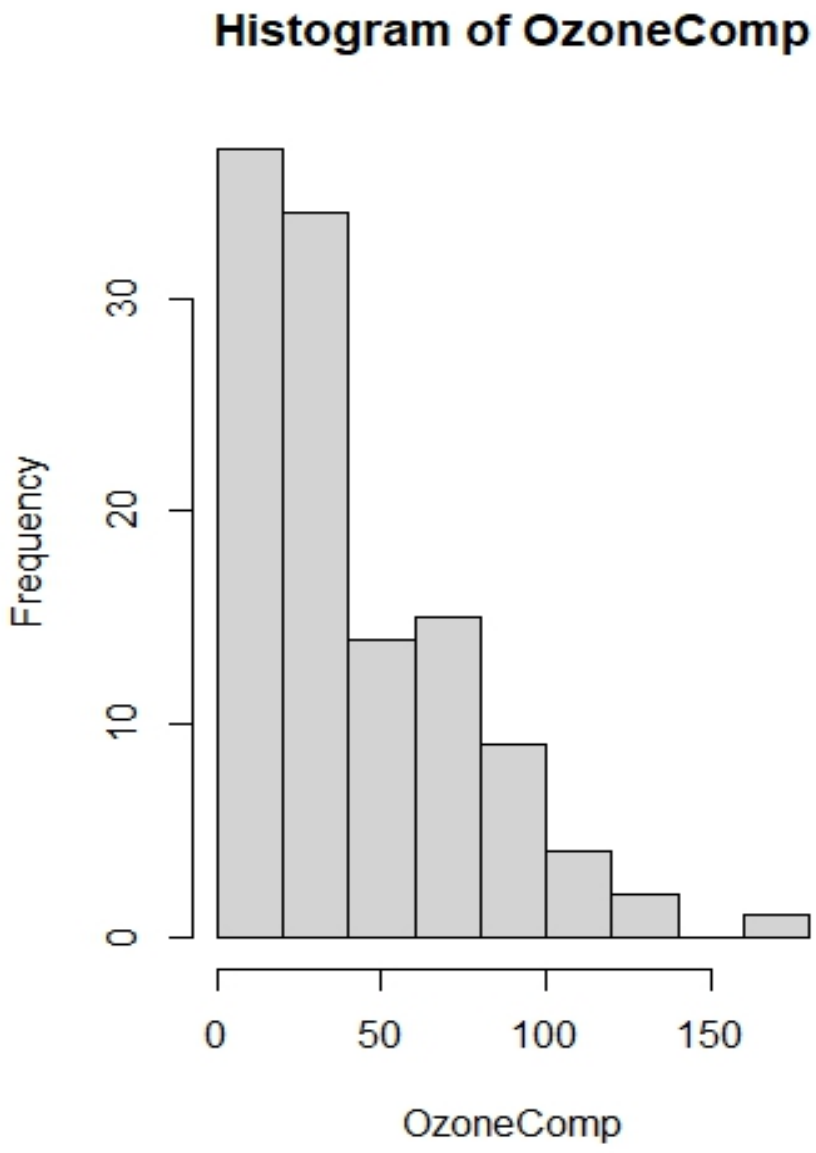


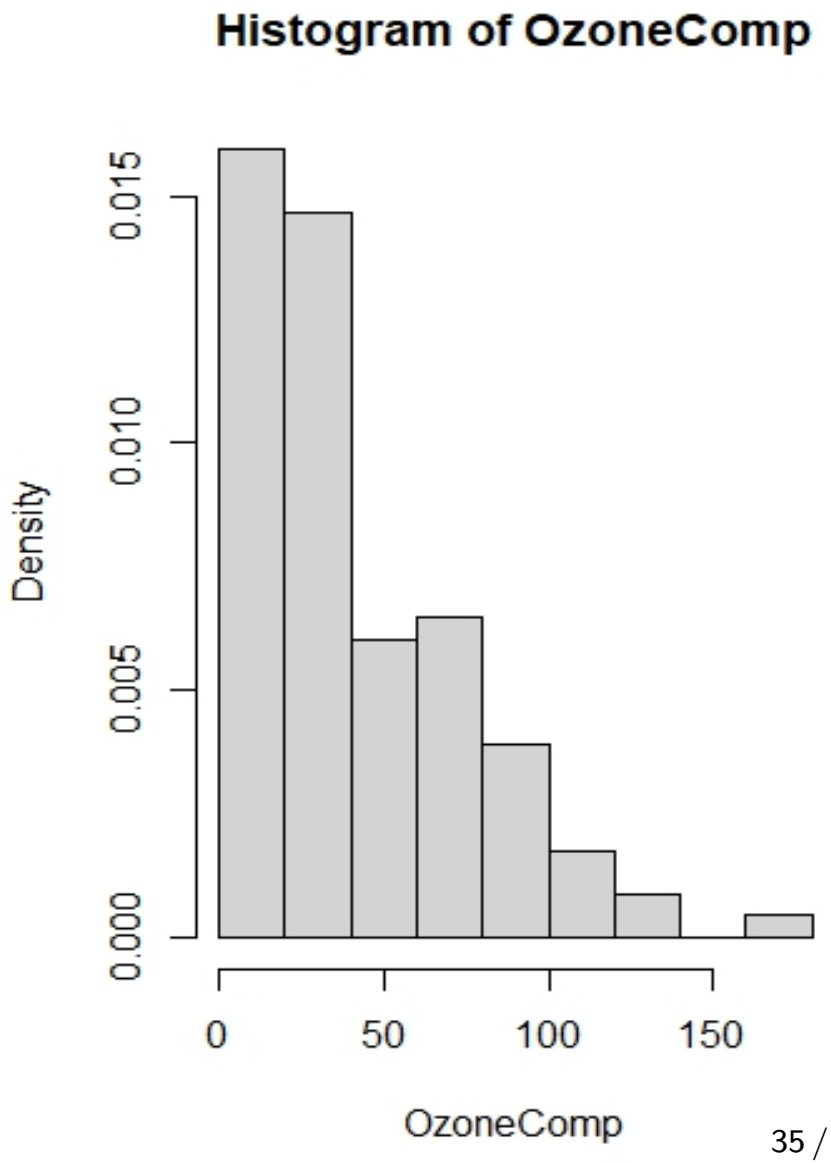
Airquality - Histograma





Airquality - Histograma





Ejercicios de la práctica que pueden hacer

Práctica 2: - Ej. 6. y 7.