Estadística *Descriptiva*
Datos y Estimaciones

Probabilidades y Estadística (C) - 2019 - Parte 2

Otra vez alfajores

Una fábrica de alfajores tiene dos sedes: una en Quilmes ($Y = 0$) y la otra en Pilar ($Y = 1$). Cada sede empaqueta sus alfajores en cajas con CUATRO unidades. Denotemos con X a la variable aleatoria que indica el número alfajores defectuosos en una caja. En el siguiente link podrá acceder a datos que se obtienen al examinar cajas al azar producidas por la fábrica. Es decir, realizaciones del vector (X, Y).

`https://probac2019.shinyapps.io/genero_alfajores/`

▸ Link

Indique cuantos datos quiere e incluya su número de libreta, para que le demos *sus datos*.

Estimaciones

1. ¿Cuántas cajas se examinaron?
   Con los datos obtenidos :
2. ESTIME la probabilidad de que una caja provenga de la sede Quilmes.
3. ESTIME la probabilidad de que una caja sea producida por la sede Quilmes y tenga 3 alfajores defectuosos.
4. ESTIME la función de probabilidad puntual conjunta del vector (X, Y) y las marginales.
5. ESTIME la esperanza y la varianza de X.

Algunos datos

- Considere los siguientes datos, generados por una distribución F.

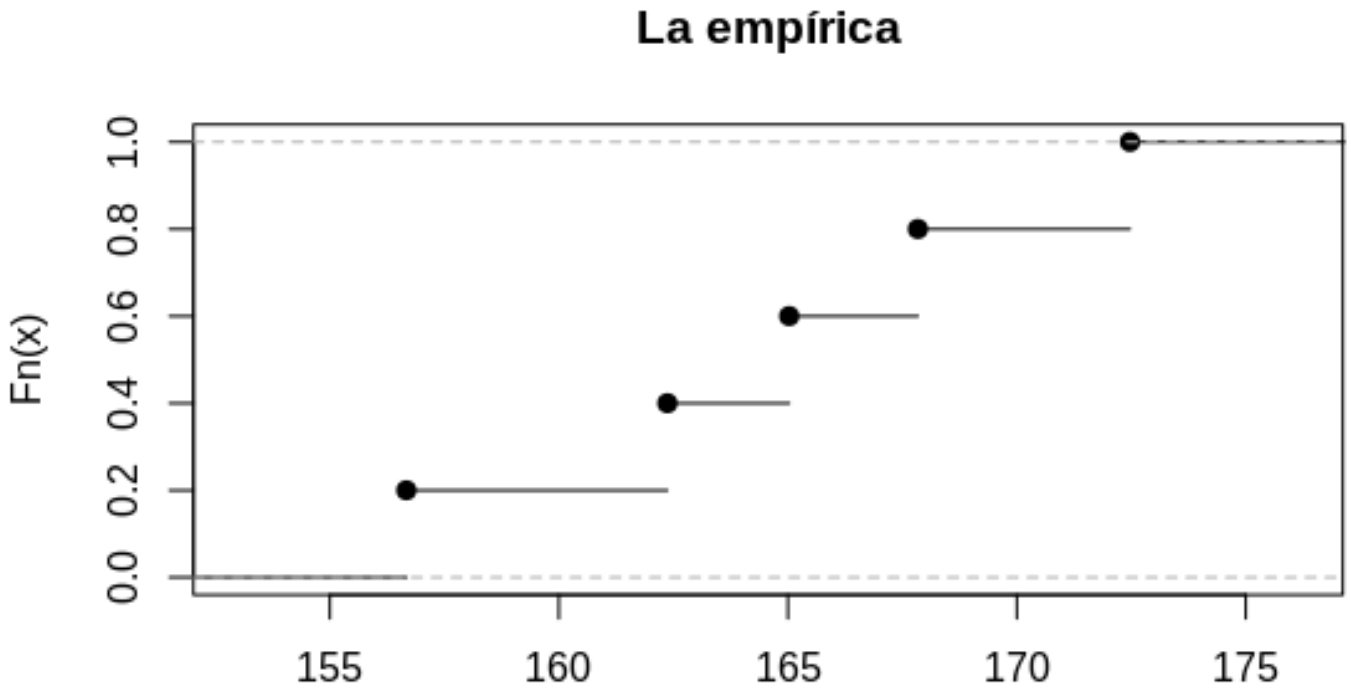
$$165.03\ ,\ 162.37\ ,\ 156.67\ ,\ 167.84\ ,\ 172.47$$

- Estime $F(160)$.
- Estime $F(168)$.
- Proponga una fórmula para estimar $F(t)$

$$\widehat{F}(t) = \ldots\ldots\ldots$$

- Grafique la función $\widehat{F} : \mathbb{R} \to [0, 1]$.

La empírica



valores	156.67	162.37	165.03	167.84	172.47
puntual	1/5	1/5	1/5	1/5	1/5

Distribución Empírica

- Datos - Observaciones generadas con F:

$$x_1, \ldots, x_n$$

- Acumulada F. $F(t) = \mathbb{P}(X \leq t)$.
- Estimación de la acumulada:

$$\widehat{F}_{n,obs}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \leq t} \quad \texttt{mean(datos<=t)}$$

Distribución Empírica

- Datos - Observaciones generadas con F:

$$x_1, \ldots, x_n$$

- Acumulada F. $F(t) = \mathbb{P}(X \leq t)$.
- Estimación de la acumulada:

$$\widehat{F}_{n,obs}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \leq t} \quad \texttt{mean(datos<=t)}$$

- $\widehat{F}_{n,obs}$ ES una función de distribución acumulada (de discreta).
- $\widehat{F}_{n,obs}$ asigna peso $1/n$ a cada valor $x_1, \ldots, x_n$.

valores	x_1	$\ldots$	$\ldots$	$\ldots$	x_n
puntual	$1/n$	$1/n$	$1/n$	$1/n$	$1/n$

Medidas de resumén - *Posición* - Mongo *muestral*

- Media $\overline{x} = n^{-1} \sum_{i=1}^{n} x_i$
- Mediana $\widetilde{x}$:
  - n impar $x_{(\frac{n+1}{2})}$
  - n par: $\frac{1}{2}\{x_{(n/2)} + x_{(n/2+1)}\}$
- Percentil α: *Buscamos el dato que ocupa la posición* $\alpha(n+1)$. *Si este número no es entero seinterpolan los dos adyacentes.*
- Percentil bis: $\widehat{F}^{-1}(\alpha)$ o $x_{([\alpha n])}$
- Cuartiles: Primero: $Q_1 = \widehat{F}^{-1}(0.25)$ o $x_{([0.25n])}$; Segundo: Q_2 mediana; Tercero: $Q_3 = \widehat{F}^{-1}(0.75)$ o $x_{([0.75n])}$
- Media α- podada:

$$\overline{x}_\alpha = \frac{x_{([n\alpha]+1)} + \cdots + x_{(n-[n\alpha])}}{n - 2[n\alpha]}$$

Medidas de resumén - *Dispersión*

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

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

- Distancia intercuartil. $Q_3 - Q_1$
- MAD: medianda$\{|x_i - \widetilde{x}|\}$

Histograma (*de Probabilidad*)

- datos: $x_1, \ldots, x_n$.
- datos ordenados: $x_{(1)} \leq \ldots \leq x_{(n)}$.
- rango: [min(datos)=$x_{(1)}$, max(datos)=$x_{(n)}$]
- Divida el rango de valores observados en intervalos (cells - bins)
- Cantidad de intervalos=K (típicamente $K = \sqrt{n}$)
- Grafique una constante sobre cada intervalo.
- El área sobre cada intervalo debe representar la **frecuencia relativa de datos en ese intervalo**.

Histograma (*de Probabilidad*)

```
> rend.A
 238 237 235 220 233 203 228 220 221 215 218
 217 232 225 209
```

Histograma (*de Probabilidad*)

```
> rend.A
 238 237 235 220 233 203 228 220 221 215 218
 217 232 225 209

> sort(rend.A)
 203 209 215 217 218 220 220 221 225 228
  232 233 235 237 238
```

Hagamos el histograma correspondiente a los siguientes intervalos:

$$[203, 215] \quad (215, 230] \quad (230, 238]$$

Histograma (*de Probabilidad*)

```
> rend.A
 238 237 235 220 233 203 228 220 221 215 218
 217 232 225 209

> sort(rend.A)
 203 209 215 217 218 220 220 221 225 228
  232 233 235 237 238
```

Hagamos el histograma correspondiente a los siguientes intervalos:

$$[203, 215] \quad (215, 230] \quad (230, 238]$$

$$[203, 215]$$
$$(215, 230]$$
$$(230, 238]$$

Histograma: AREA=FRECUENCIA RELATIVA

1. datos: $x_1, \ldots, x_n$
2. datos ordenados: $x_{(1)} \leq \ldots, \leq x_{(n)}$
3. $I_1, \ldots I_K$, K intervalos que particionan $[x_{(1)}, x_{(n)}]$
4. Graficamos una constante sobre cada intervalo de forma tal que el area del rectángulo coincida con la frecuencia relativa en el intervalo:

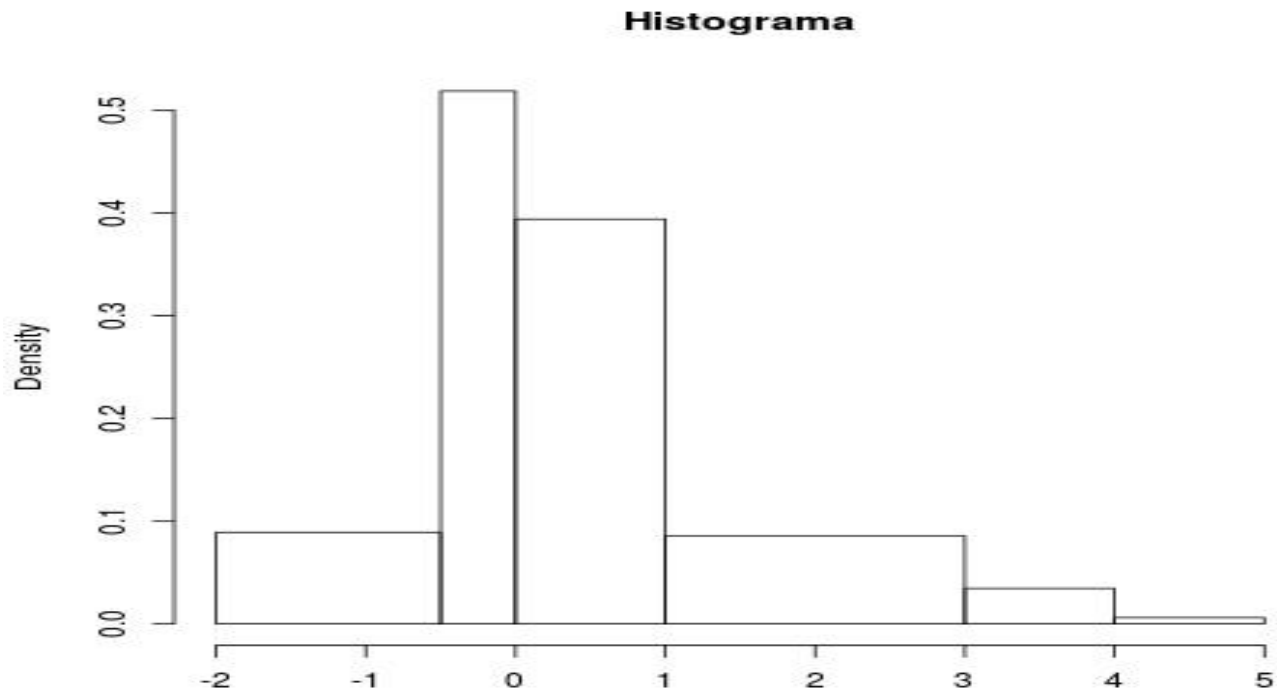
$$\text{Altura sobre } I_j \times \text{longitud } (I_j) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \in I_j}$$

$$\text{Si } I_j = (a, b), \text{ entonces longitud } (I_j) = |I_j| = b - a$$

$$\text{Altura sobre } I_j = \frac{1}{|I_j|} \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{x_i \in I_j}$$

Histograma: AREA=FRECUENCIA RELATIVA

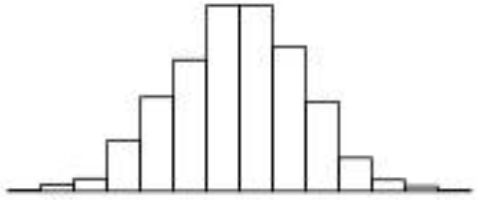
int.:	$[-2,-0.5)$	$[-0.5,0)$	$[0,1)$	$[1,3)$	$[3,4)$	$[4,5)$
frec.:	174	337	512	223	45	8
fre. rel.:	$174/n$	$337/n$	$512/n$	$223/n$	$45/n$	$8/n$



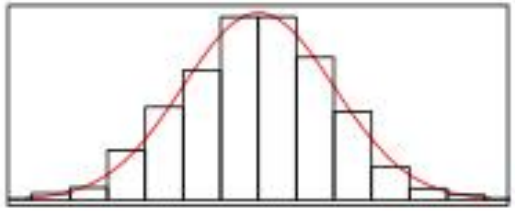
en R `hist(datos, prob=TRUE)`

Histogramas- Media & Mediana

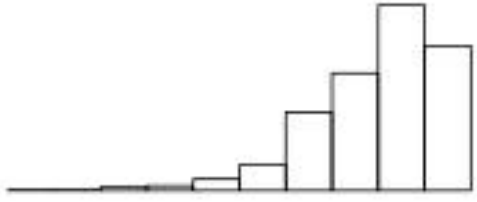
Acampanado



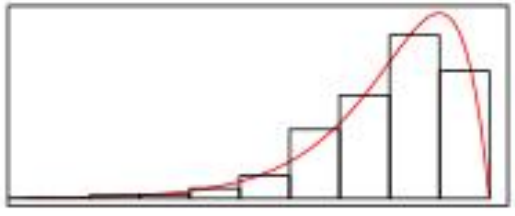
Acampanado



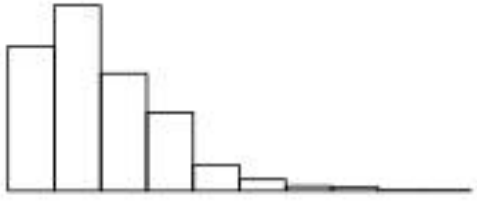
Colas pesada a Iquierda



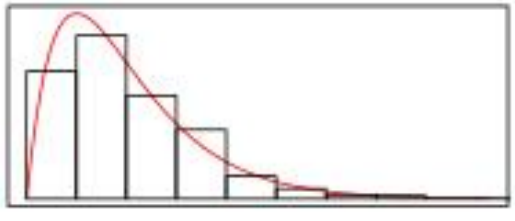
Colas pesada a Iquierda



Colas pesada a Derecha



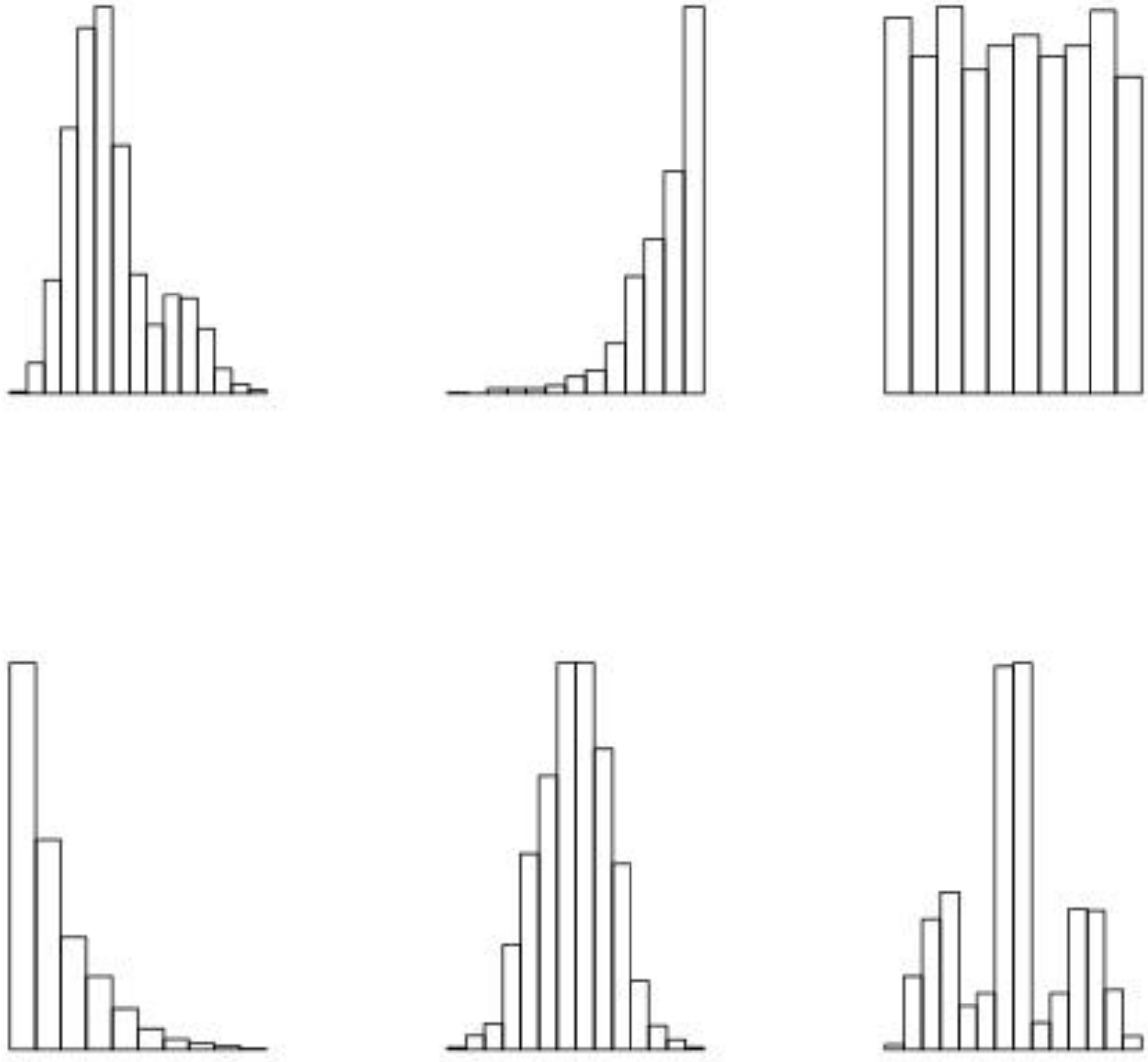
Colas pesada a Derecha



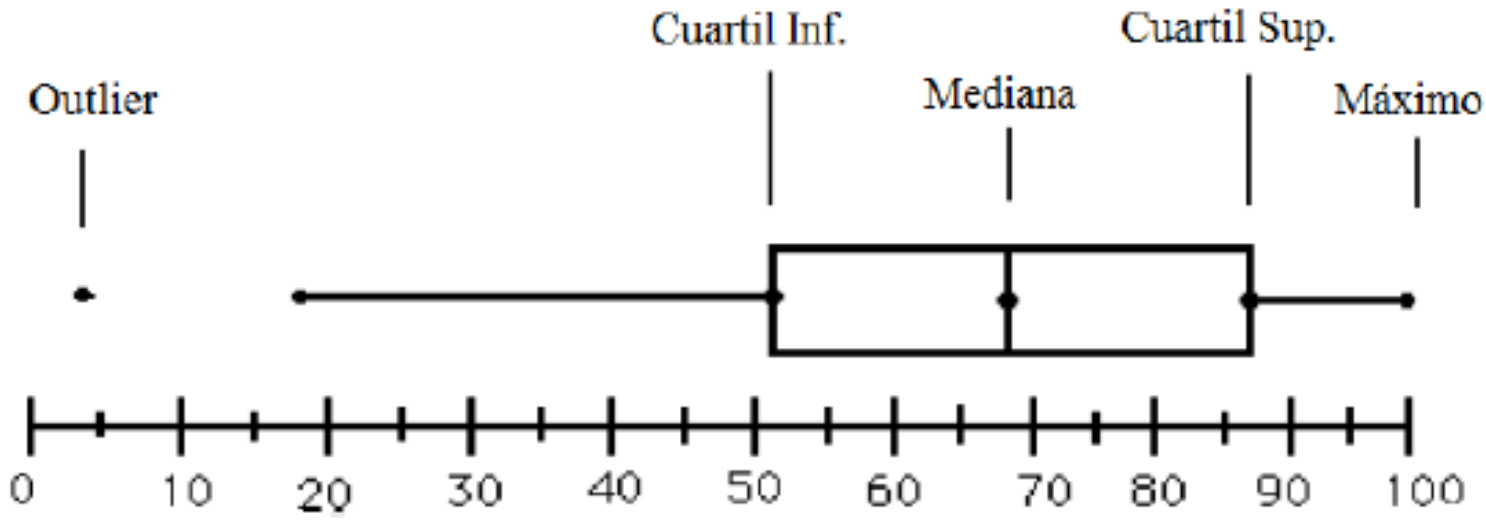
Compare los siguientes comandos

```
hist(rend.A)

hist(rend.A, prob=T)

hist(rend.A, breaks=4, prob=T)

hist(rend.A, breaks=c(200,215,230,240), prob=T)
```

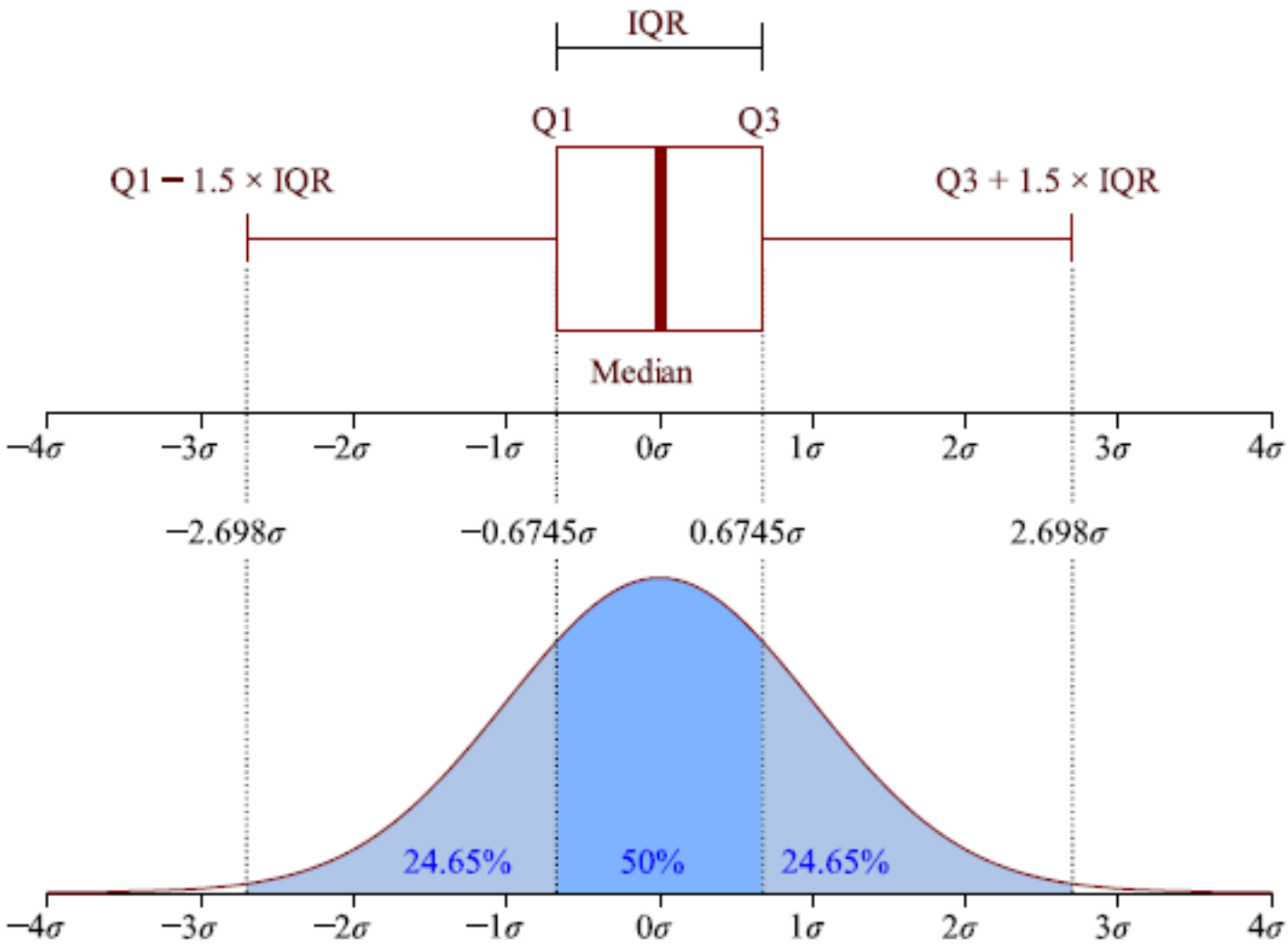
Histogramas - Moda: *pico*



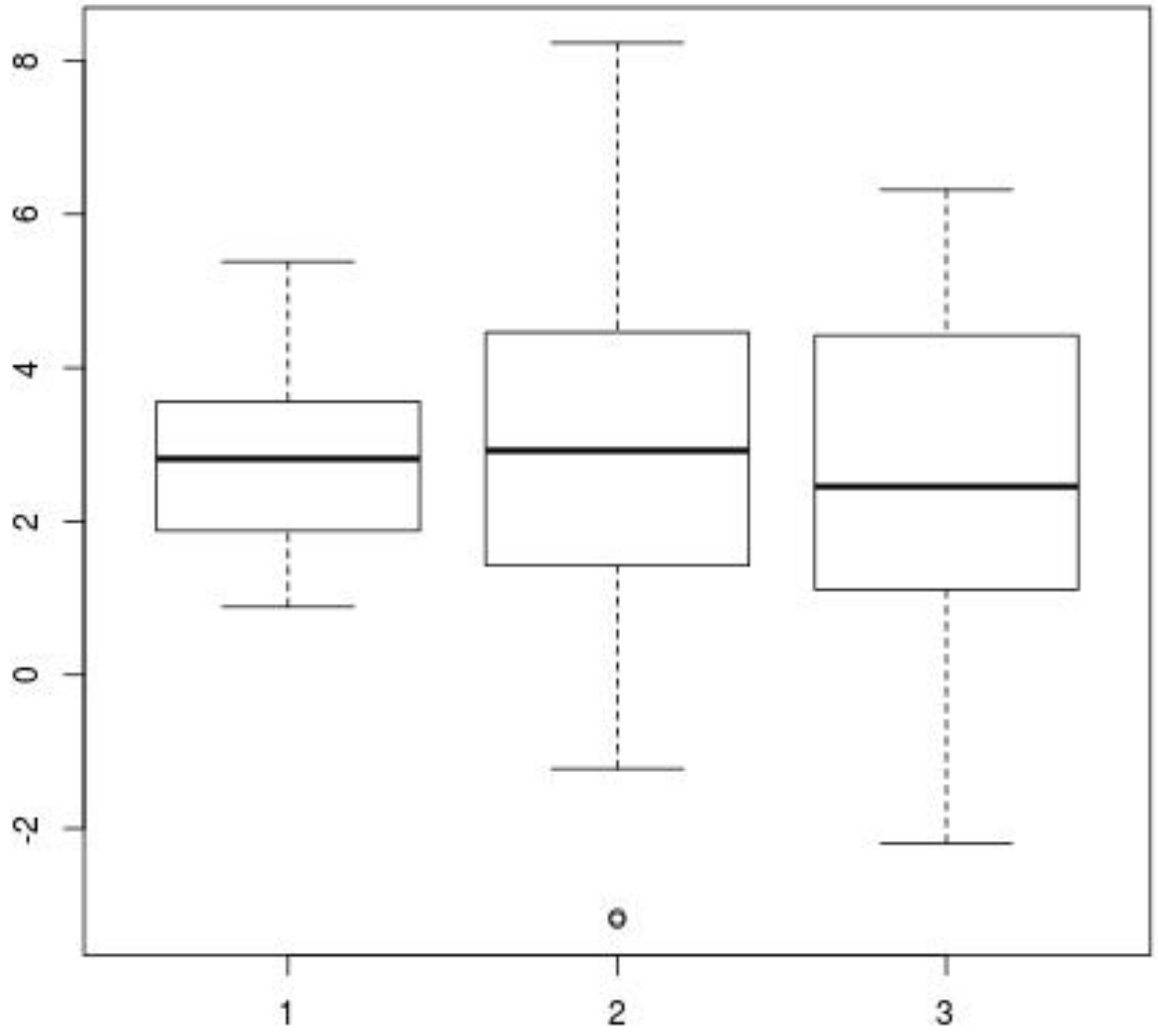
Boxplot - en R boxplot(datos)



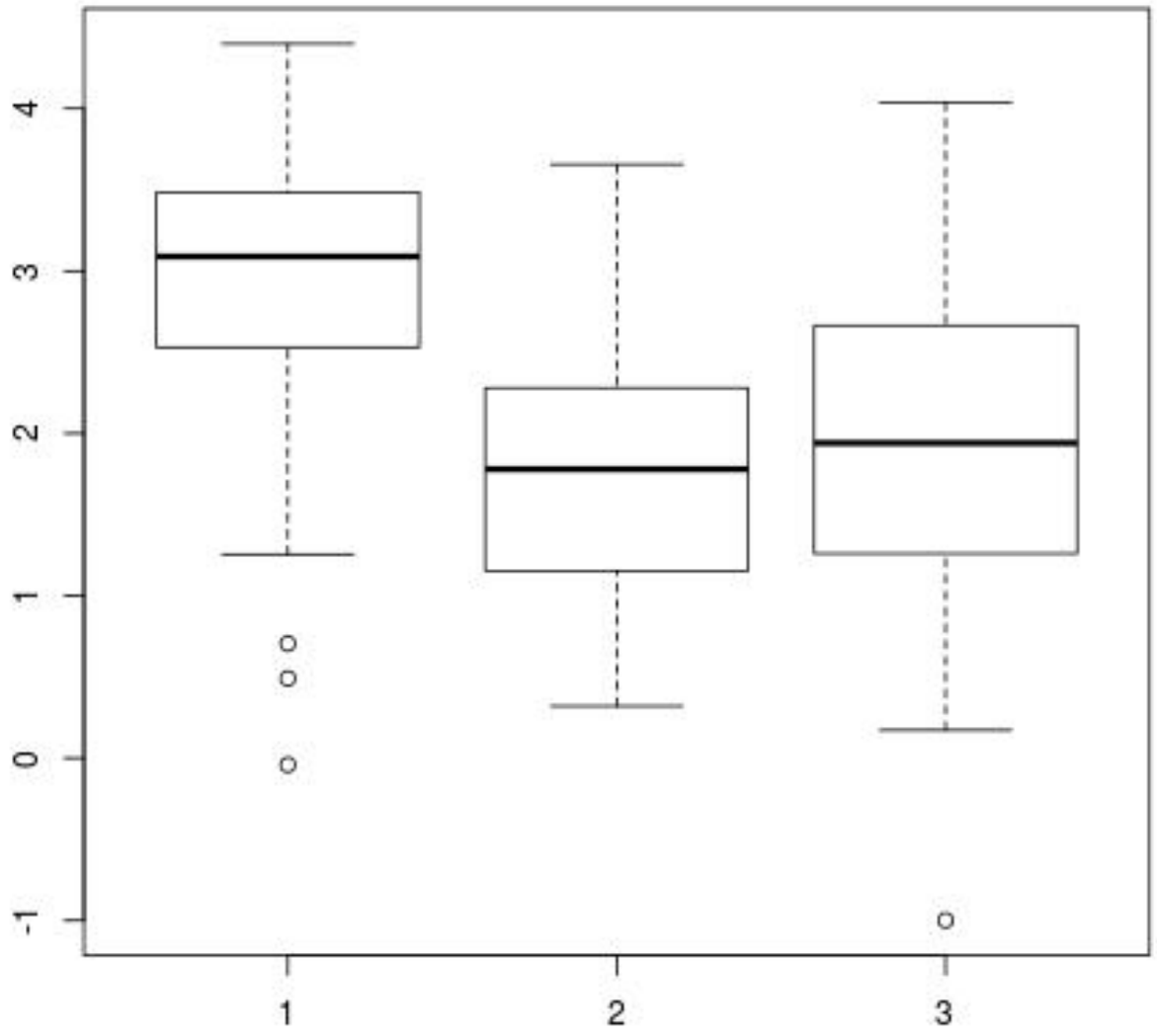
Datos normales (Gracias Daniela!)



Boxplot - en R boxplot(datos)



Boxplot - en R boxplot(datos)



Teorema Central del Límite

shiny TCL

Medidas de resumen

- n datos: $x_1, x_2, \ldots, x_n$
- datos ordenados $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ - en R: sort(datos)

mean(datos)	$\overline{x}$
median(datos)	$\tilde{x}$
quantile(datos, alfa)	$x_{([n\alpha])}$
quantile(datos,0.25)	Q_1 primer cuartil
quantile(datos,0.75)	Q_3 tercer cuartil
min(datos), max(datos)	$x_{(1)}$, $x_{(n)}$
mean(datos,trim = alfa)	$\overline{x}_\alpha = \{x_{([n\alpha])} + \ldots + x_{(n-[n\alpha])}\}/(n - 2[n\alpha])$
var(datos)	$s^2 = \sum(x_i - \overline{x})^2/(n-1)$
sd(datos)	$\sqrt{s^2}$
IQR(datos)	$d_I = Q_3 - Q_1$
mad(datos)	mediana$(\lvert x_i - \tilde{x}\rvert)$