Característiques de posició central i dispersió

Variables quantitatives

- **La mitjana $\bar{X}$:** És la suma de totes i cadascuna de les observacions dividida pel nombre total d'observacions.

$$\bar{X} = \frac{x_1 + \cdots + x_n}{n}$$

Per exemple, la mitjana de les notes 3,3,5,6,6,7,7,7 és

$$\bar{X} = \frac{3+3+5+6+6+7+7+7}{8} = 5.5$$

La mitjana també es pot calcular fent servir les freqüències

Nota	n_i	f_i	p_i	n_i	f_i	p_i
3	2	0.25	25%	2	0.25	25%
5	1	0.125	12.5%	3	0.375	37.5%
6	2	0.25	25%	5	0.625	62.5%
7	3	0.375	37.5%	8	1	100
Total	8	1	100			

Amb les freqüències absolutes:

$$\bar{X} = \frac{x_1 n_1 + \cdots + x_k n_k}{n} =$$

$$= \frac{3 \cdot 2 + 5 \cdot 1 + 6 \cdot 2 + 7 \cdot 3}{8} = 5.5$$

Amb les freqüències relatives:

$$\bar{X} = \frac{x_1 f_1 + \cdots + x_k f_k}{1} =$$

$$= 3 \cdot 0.25 + 5 \cdot 0.125 + 6 \cdot 0.25 + 7 \cdot 0.375 = 5.5$$

Amb els percentatges:

$$\bar{X} = \frac{x_1 p_1 + \cdots + x_k p_k}{100} =$$

$$= \frac{3 \cdot 25 + 5 \cdot 12.5 + 6 \cdot 25 + 7 \cdot 37.5}{100} = 5.5$$

Hem vist que el càcul de la mitjana requereix una variable quantitativa en escala numèrica (contínua o discreta).

Quan les dades estan agrupades en intervals: aleshores utilitzem les marques de classe de cada interval per a obtenir un valor aproximat per la mitjana. Per exemple,

Nota	x_i	n_i	f_i	p_i	N_i	F_i	P_i
$[0, 50)$	25	17	0.34	34%	17	0.34	34%
$[50, 70)$	60	22	0.44	44%	39	0.78	78%
$[70, 90)$	80	9	0.18	18%	48	0.96	96%
$[90, 100]$	95	2	0.04	4%	50	1	100
Total		50	1	100			

$$\bar{X} \simeq \frac{25 \cdot 17 + 60 \cdot 22 + 80 \cdot 9 + 95 \cdot 2}{50} = 53.1$$

mentre que el valor exacte obtingut a partir de les dades originals abans d'agrupar és

$$\bar{X} = 57.74$$

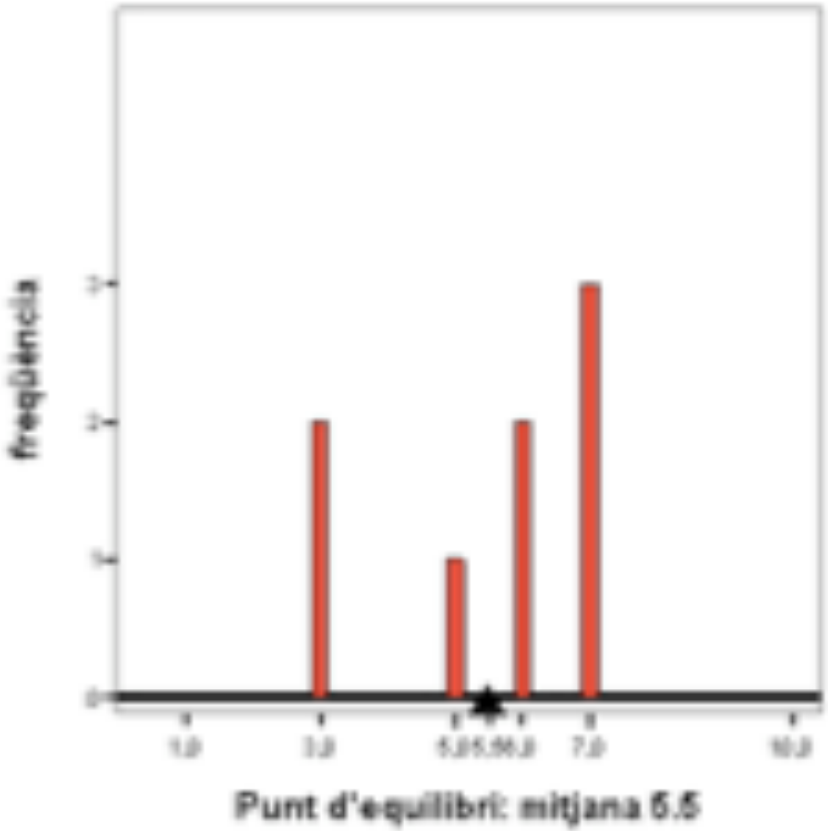
Propietats de la mitjana:

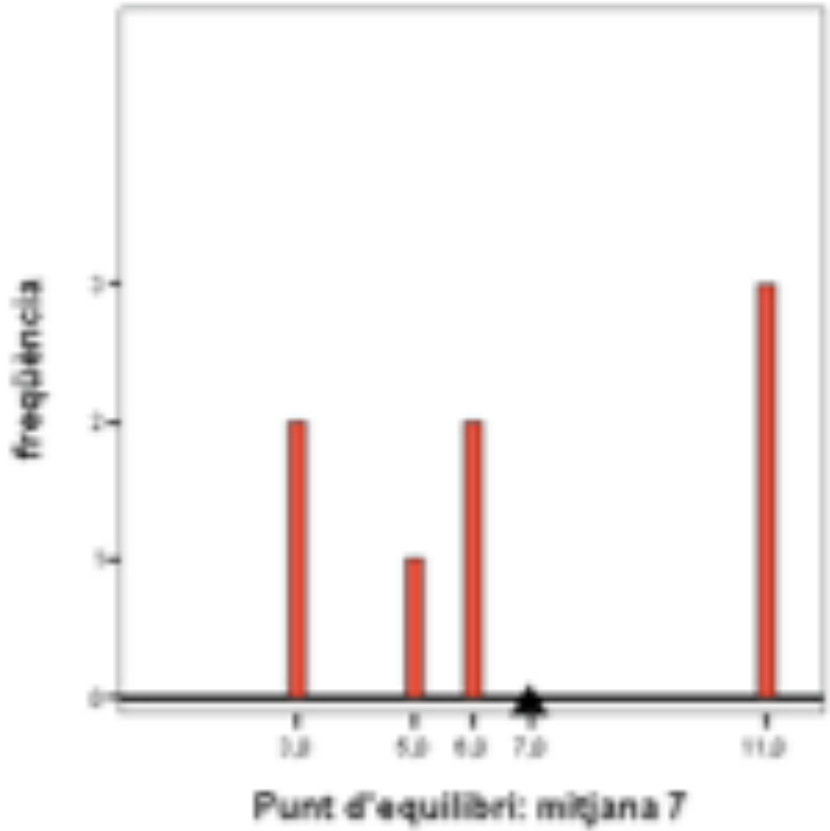
- La mitjana representa un punt central de la distribució de valors (centre de massa de la distribució).

- En el seu càlcul intervenen tots i cadascun dels valors de la variable.

- La mitjana es veu afectada pels valors extrems de la variable.

La **mitjana retallada al** 5% és el càlcul de la mitjana després d'eliminar el 5% dels valors més extrems (molt alts o molt baixos). D'aquesta manera s'elimina parcialment l'efecte dels valors extrems.


freqüència
1,0
3,0
5,0
5,5
6,0
7,0
10,0
Punt d'equilibri: mitjana 6.5


freqüència
3,0
5,0
6,0
7,0
11,0
Punt d'equilibri: mitjana 7

La mitjana ponderada s'obté quan a diferents casos o valors x_i se'ls assigna pesos diferents w_i, i és la suma de totes i cadscuna de les observacions miltiplicada pel seu pes i dividida per la suma de tots els pesos (un per cada observació)

$$\bar{X} = \frac{x_1 w_1 + \cdots + x_n w_n}{w_1 + \cdots + w_n} =$$

$$= \frac{x_1 w_1 n_1 + \cdots x_k w_k n_k}{w_1 n_1 + \cdots + w_k n_k}$$

Per exemple, imagineu que la nota final de l'assignatura s'obté a partir de les notes de tres examens 3, 6, 5 però que conten diferent: el primer conta un 30%, el segon un 30% i l´últim un 40%. Aleshores la nota final serà la mitjana ponderada

$$\bar{X} = \frac{3 \cdot 30 + 6 \cdot 30 + 5 \cdot 40}{30 + 30 + 40} = 4.7$$

Altres mesures de tendència central i dispersió que ja coneixem:

- **La Moda:** està pensada per variables nominals o numèriques que prenen pocs valors i és el valor amb la freqüència més gran.

  Si la variable està agrupada en intervals, aleshores l'interval modal és el que té la freqüència per unitat d'amplitud més gran.

  En l'exemple anterior l'interval modal és el $[50, 70)$.

- **La Mediana:** és el valor Md de la variable tal que la meitat dels casos menors o iguals que Md i l'altra meitat són més grans o iguals.

  Pel seu càlcul ens cal una escala ordinal o numèrica. I representa un punt central en la distribució de dades.

  En el seu càlcul només intervenen les posicions relatives dels diferents valors de la variable i no es vau afectada per valors extrems.

Recordeu que es calcula a partir dels percentages acumulats que es poden obervar en una taula de freqüències. Si una valor acumula exactament el 50%, es pren la mitjana entre aquest valor i el següent.

Si les dades venen agrupades en intervals, aleshores només podrem obtenir un càlcul aproximat i prendrem la marca de classe de l'interval que acumuli més d'un 50% de les dades.

- **Quartils, decils i percentils (o centils)**: es defineixen de la mateixa manera que per variables qualitatives ordinals (i de manera anàloga a la mediana).

  L'objectiu és posicionar la distribució donant uns pocs valors.

  Si tenim les dades agrupades en intervals, aleshores el valor aproximat és la marca de classe del primer interval que supera el percentage determinat.

  Observació: la informació proporcionada per una taula de freqüències sempre és molt més precisa que la donada pels quartils ja que obtenim els percentatges exactes (sobretot si la nostra variable pren pocs valors).

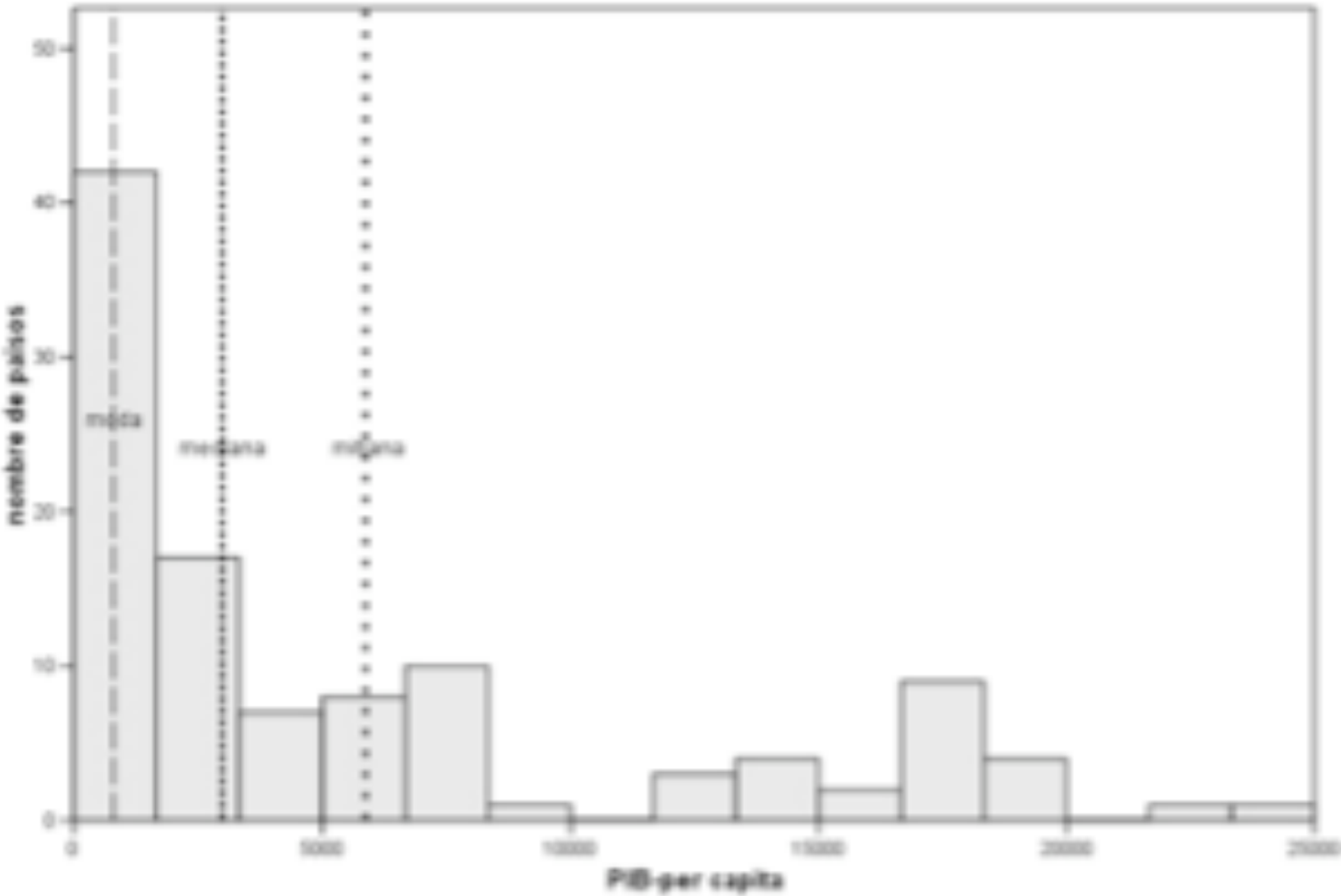


Figura 1.13: Moda, mediana i mitjana del *PIB per càpita*.

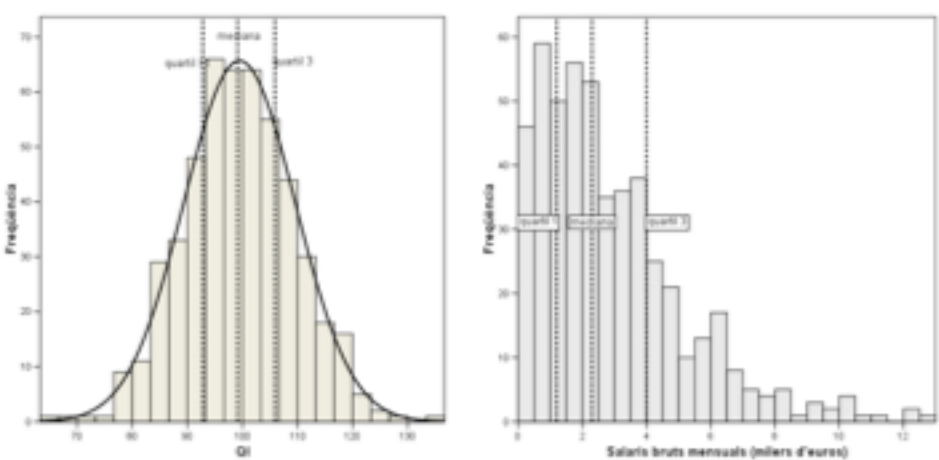


Figura 1.14: Els tres quartils: distribucions simètriques i asimètriques.

- El **rang** és la diferència entre els valors màxim i mínim que pren la variable

  $$R = x_{max} - x_{min}$$

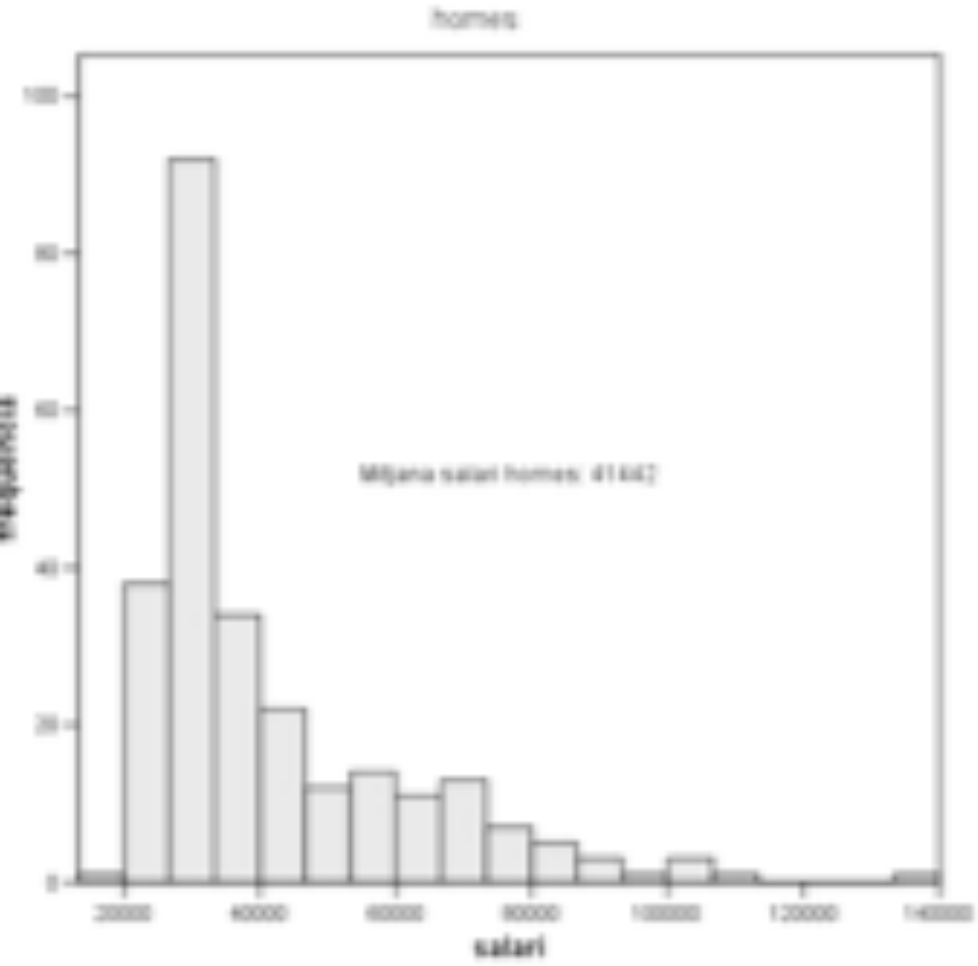
  Dins del rang es distribuixen el 100% dels casos.

- El **rang inter-quartil** és la diferència entre el tercer i el primer quartil,

  $$RI = Q_3 - Q_1$$

  Dins del rang interquartil es distribueixen el 50% dels casos.

Per exemple, en una empresa s'ha estudiat el salari dels empleats segons el sexe. I s'han obtingut els següents histogrames:



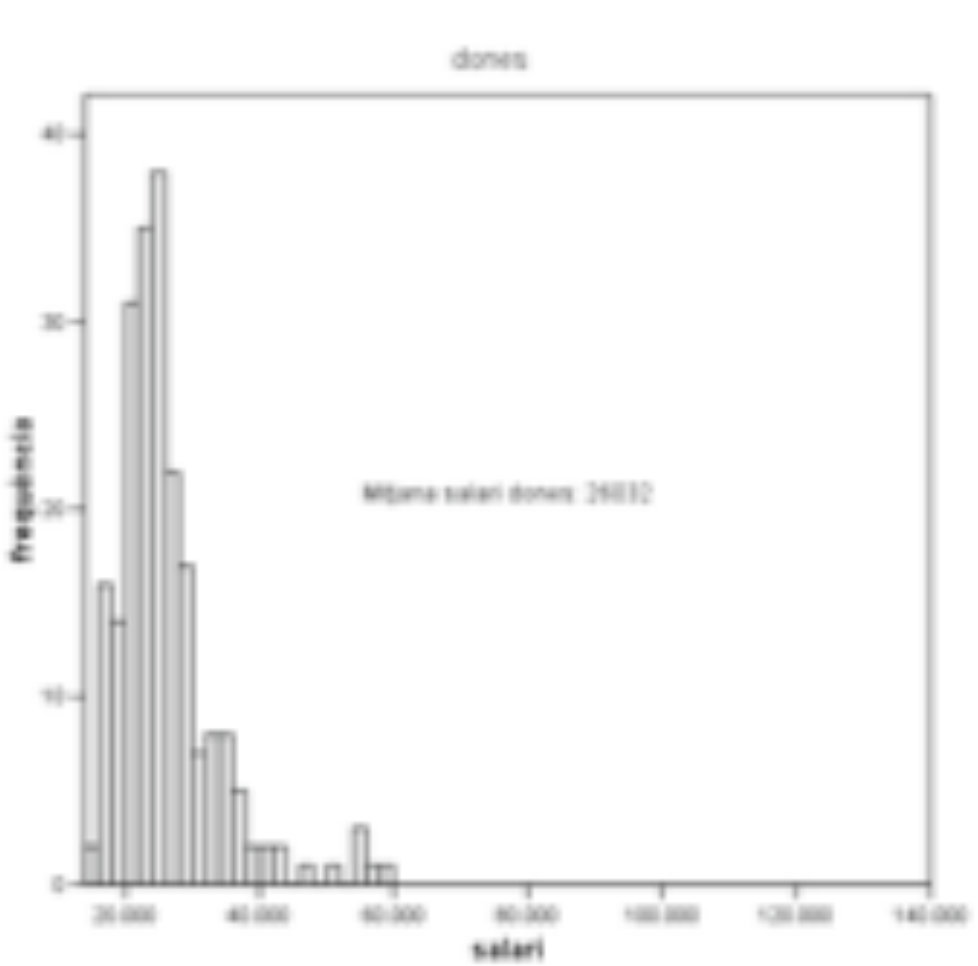


Figura 1.16: Dispersions diferents i grau de representació de la mitjana.

Els rangs pel salari dels homes són:

$$R = max - min = 135000 - 19650 = 115350$$

$$RI = Q_3 - Q_1 = 50725 - 28050 = 22675$$

$$Q_1 = 28050,\ Q_2 = 32850,\ Q_3 = 50725$$

Els rangs pel salari de les dones són:

$$R = max - min = 58125 - 15750 = 42375$$

$$RI = Q_3 - Q_1 = 28500 - 21487.5 = 7012.5$$

$$Q_1 = 21487.5,\ Q_2 = 24300,\ Q_3 = 28500$$

Observem una dispersió més gran en les dades corresponents al salari dels homes.

El diagrama de caixa

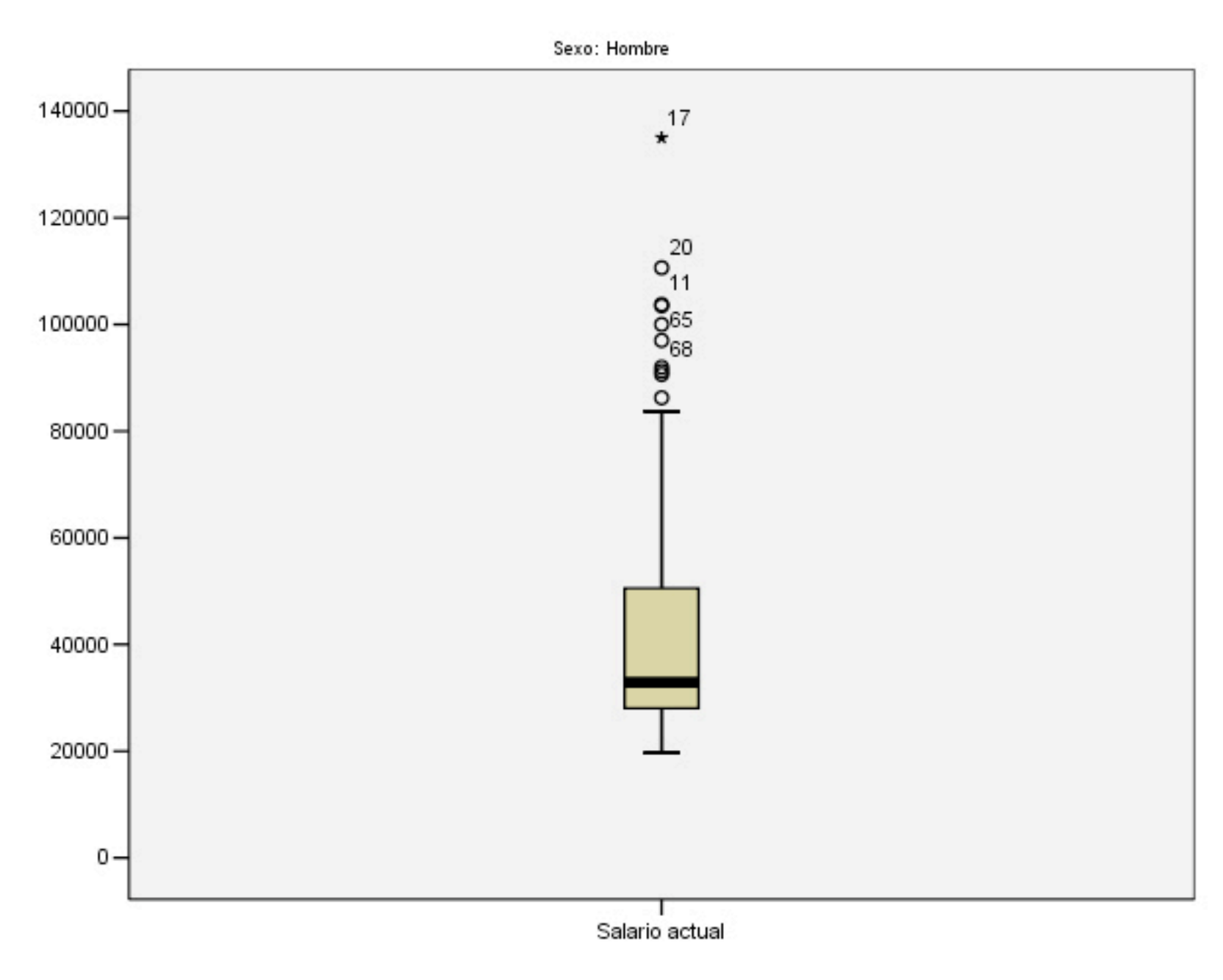
En el diagram de caixa es representen el valors màxim i mínim, els quartils (inclouen la mediana), el rang interquartil i el rang.

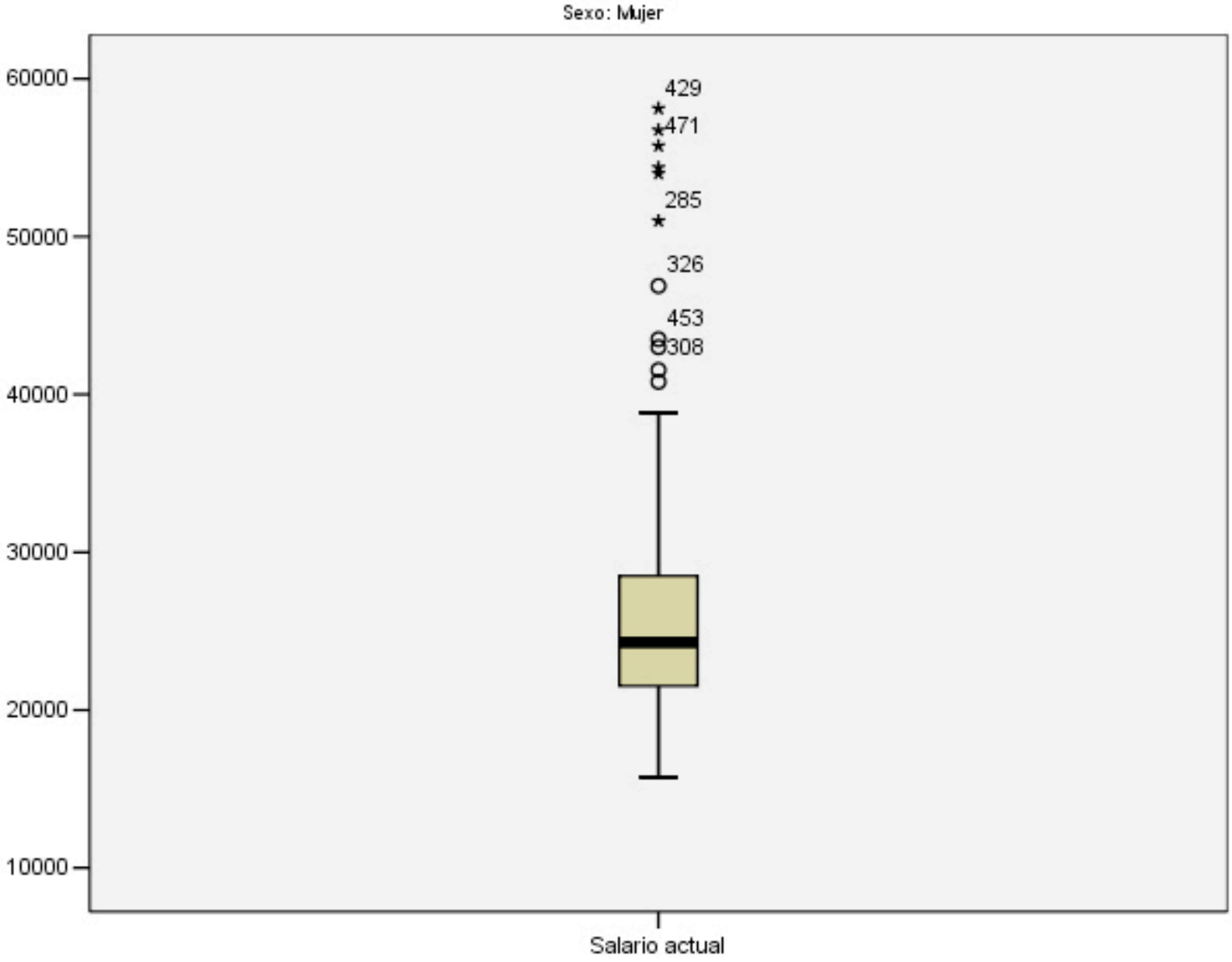
En l'eix vertical es representa l'escala de mesura de la variable.

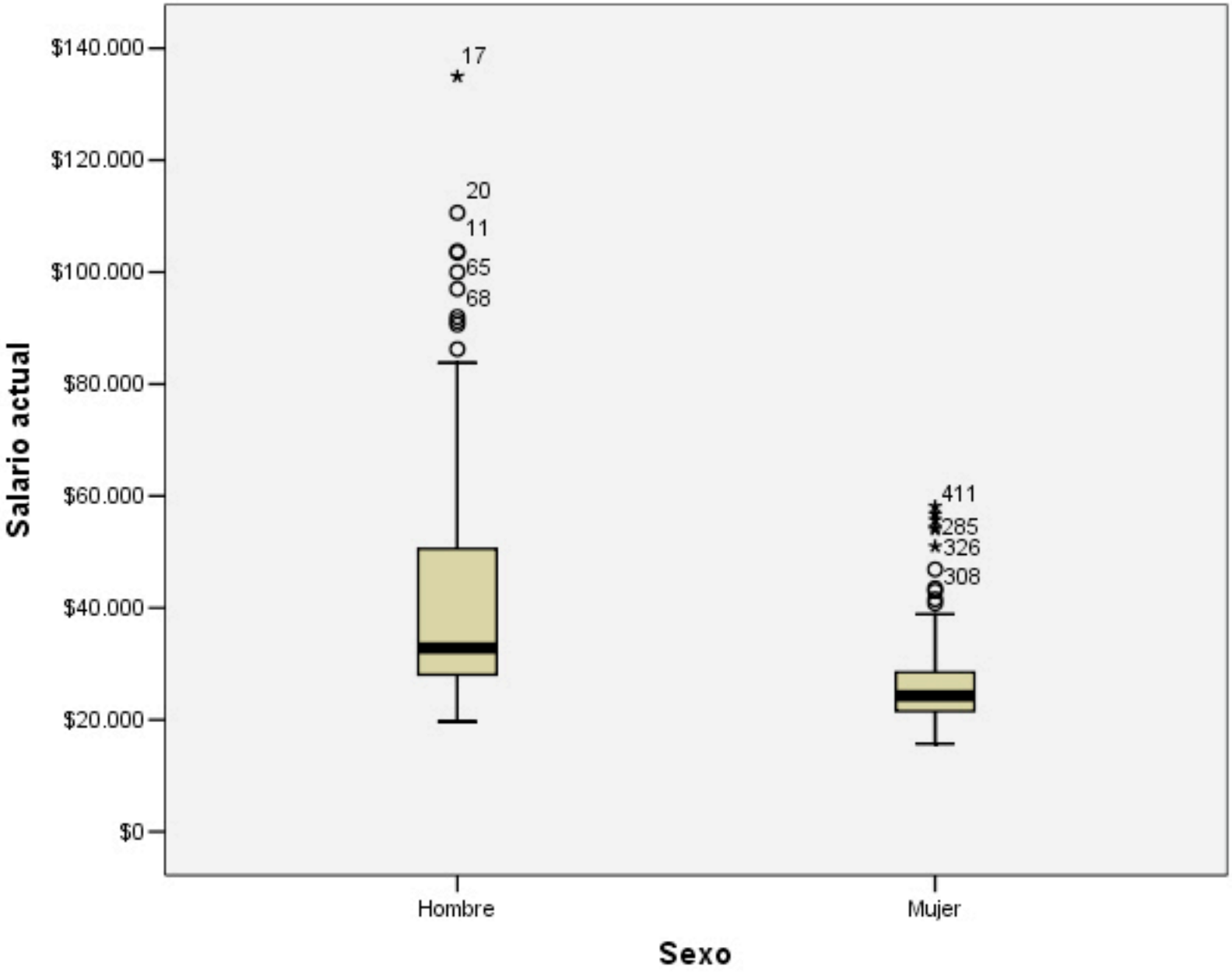
Una **caixa central** està delimitada pel primer i tercer quartil (conté el 50% de les observacions centrals), i una línia gruixuda a l'interior que correspon a la mediana (o segon quartil).

Les **patilles** es delimiten els valors que es troben a una distància inferior o igual a 1.5 vegades el rang interquartil.

Els valors que queden a una distància de les vores de la caixa superior a 1.5 vegades RI s'anomenen **outliers** i es marquen individualment amb un cercle. Els **ouliers severs** són els que es troben a una distància de la caixa superior a 3 vegades el rang interquartil RI.


Sexo: Hombre
140000
120000
100000
80000
60000
40000
20000
0
17
20
11
65
68
Salario actual


Sexo: Mujer
60000
50000
40000
30000
20000
10000
429
471
285
326
453
308
Salario actual


$140.000
$120.000
$100.000
$80.000
$60.000
$40.000
$20.000
$0
Salario actual
17
20
11
65
68
411
285
326
308
Hombre
Mujer
Sexo

Una altra característica de dispersió: la variància i la desviació típica

La **variància poblacional** σ^2 és la mitjana de les desviacions de tots i cadascun dels valors de la variable respecte la seva mitjana

$$\sigma^2 = \frac{(x_1 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n} =$$

$$= \frac{(x_1 - \bar{X})^2 n_1 + \cdots + (x_k - \bar{X})^2 n_k}{n}$$

La **variància mostral** s^2 és la mitjana de les desviacions de tots i cadascun dels valors de la variable respecte la seva mitjana

$$s^2 = \frac{(x_1 - \bar{X})^2 + \cdots + (x_n - \bar{X})^2}{n-1} =$$

$$= \frac{(x_1 - \bar{X})^2 n_1 + \cdots + (x_k - \bar{X})^2 n_k}{n-1}$$

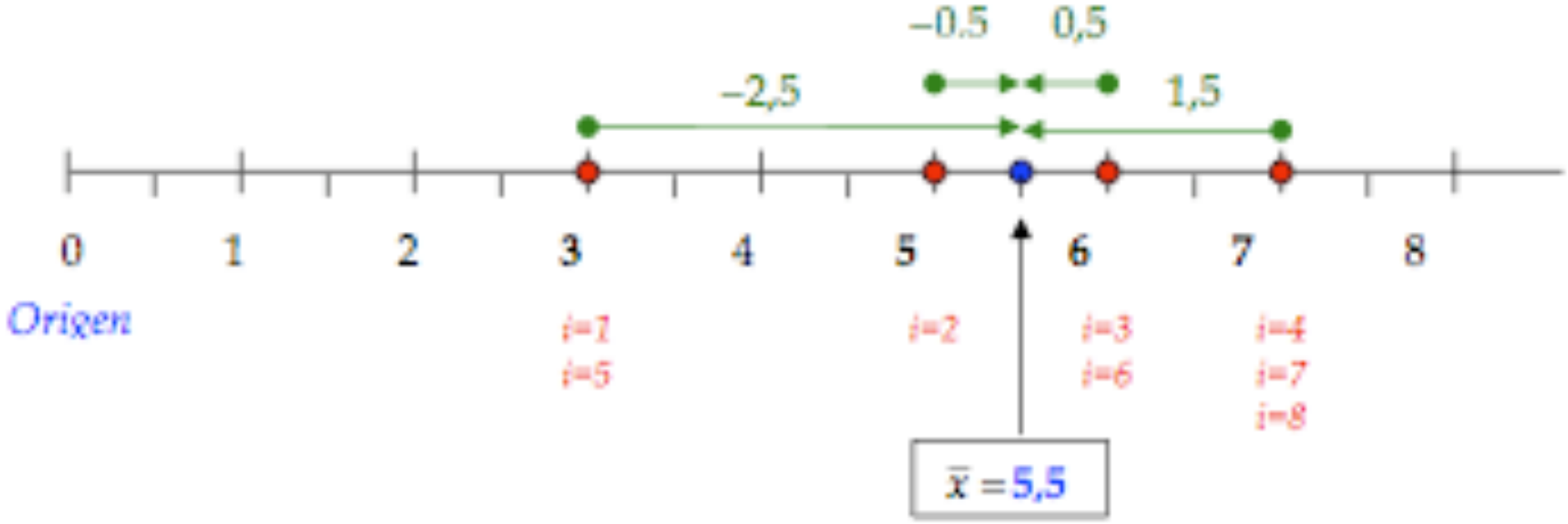
Per exemple, recordeu l'exemple de les notes

Nota	n_i	f_i	p_i	N_i	F_i	P_i
3	2	0.25	25%	2	0.25	25%
5	1	0.125	12.5%	3	0.375	37.5%
6	2	0.25	25%	5	0.625	62.5%
7	3	0.375	37.5%	8	1	100
Total	8	1	100			

$\bar{X} = 5.5$

Nota	n_i	$x_i - \bar{X}$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})^2 n_i$
3	2	-2.5	6.25	12.5
5	1	-0.5	0.25	0.25
6	2	0.5	0.25	0.5
7	3	1.5	2.25	6.75
Total	8			20

$$\sigma^2 = \frac{20}{8} = 2.5,\ s^2 = \frac{20}{7} = 2.857$$


−0.5
0,5
−2,5
1,5
0
1
2
3
4
5
6
7
8
Origen
i=1
i=5
i=2
i=3
i=6
i=4
i=7
i=8
$\bar{x} = 5{,}5$

Observeu que la variància (tant mostral com poblacional) té les unitats de l'escala de mesura de la variable elevades al quadrat. Per corregir-ho, considerarem la seva arrel quadrada.

La **desviació típica poblacional (resp. mostral)** σ **(resp.** s**)** és $\sigma = \sqrt{\sigma^2}$ (resp. $s = \sqrt{s^2}$).

En l'exemple anterior obtenim una desviació típica mostral $s = \sqrt{2.857} = 1.690$.

Una interpretació de la desviació típica

- Com a mínim un 75% dels casos es troben a l'interval $(\bar{X} - 2s, \bar{X} + 2s)$.

- Com a mínim un 89% dels casos es troben a l'interval $(\bar{X} - 3s, \bar{X} + 3s)$.

En l'exemple anterior veiem que un 75% dels casos o més es troba a l'interval

$$(5.5 - 2 \cdot 1.69, 5.5 + 2 \cdot 1.69) = (2.12, 8.88).$$

I un 89% dels casos o més es troba a l'interval

$$(5.5 - 3 \cdot 1.69, 5.5 + 3 \cdot 1.69) = (0.43, 10.57).$$

Per a poder comparar la dispersió de dues mostres o dues variables utilitzant la variància o la desviació típica, cal calcular el **coeficient de variació**.

$$CV = \frac{s}{\bar{X}}$$

Per exemple, volem compara la dispersió de les variables X=esperança de vida femenina, i Y=taxa de mortalitat infantil (per cada 1000 habitants), mesurades en tots els països del món. Hem obtingut

$$\bar{X} = 70.16,\ s_X = 10.57$$

$$\bar{Y} = 9.56,\ s_Y = 4.25$$

Observeu que $s_Y < s_X$, però si calculem els coeficients de variació

$$CV_X = \frac{10.57}{70.16} \simeq 0.15,\ CV_Y = \frac{4.25}{9.56} \simeq 0.44$$

veiem que Y presenta més dispersió relativa.

Característiques de forma: simetria i curtosi

Les característiques de forma s'observen en els histogrames o en els diagrames de caixa.

- **Simetria**: la simetria de la distribució ve determinada per la posició relativa entre la mitjana ($\bar{X}$), la mediana (Md) i la moda (Mo).

  Si $\bar{X} = Md = Mo$ direm que la distribució és **simètrica**.

  Si $Mo < Md < \bar{X}$ direm que la distribució té **asimetria positiva** o és **esbiaxada a la dreta**.

  Si $\bar{X} < Mo < Md$ direm que la distribució té **asimetria negativa** o és **esbiaxada a l'esquerra**.

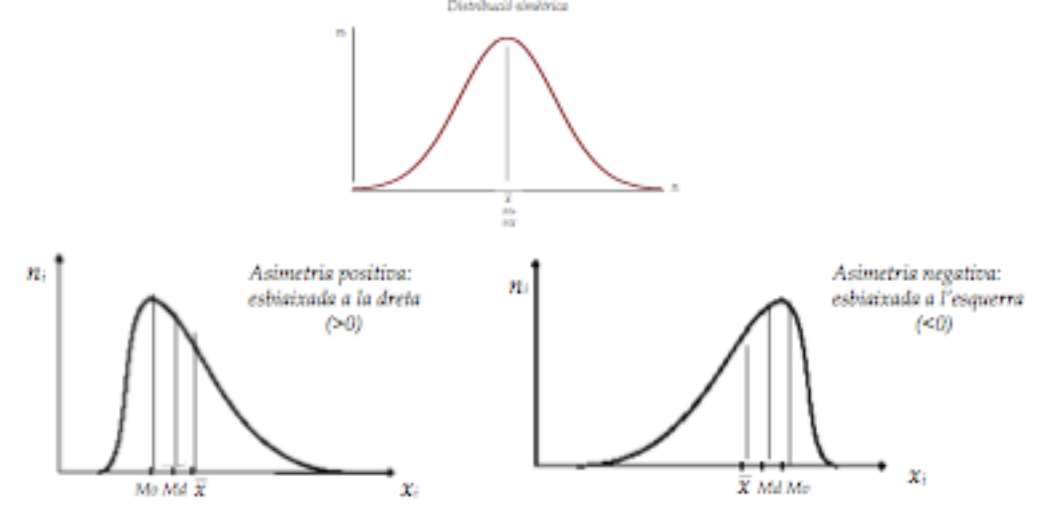


Figura 1.19: Simetria i asimetries en una distribució unimodal.

- **Curtosi**: només té sentit per distribucions simètriques i distingeix entre distribucions més apuntades o menys.

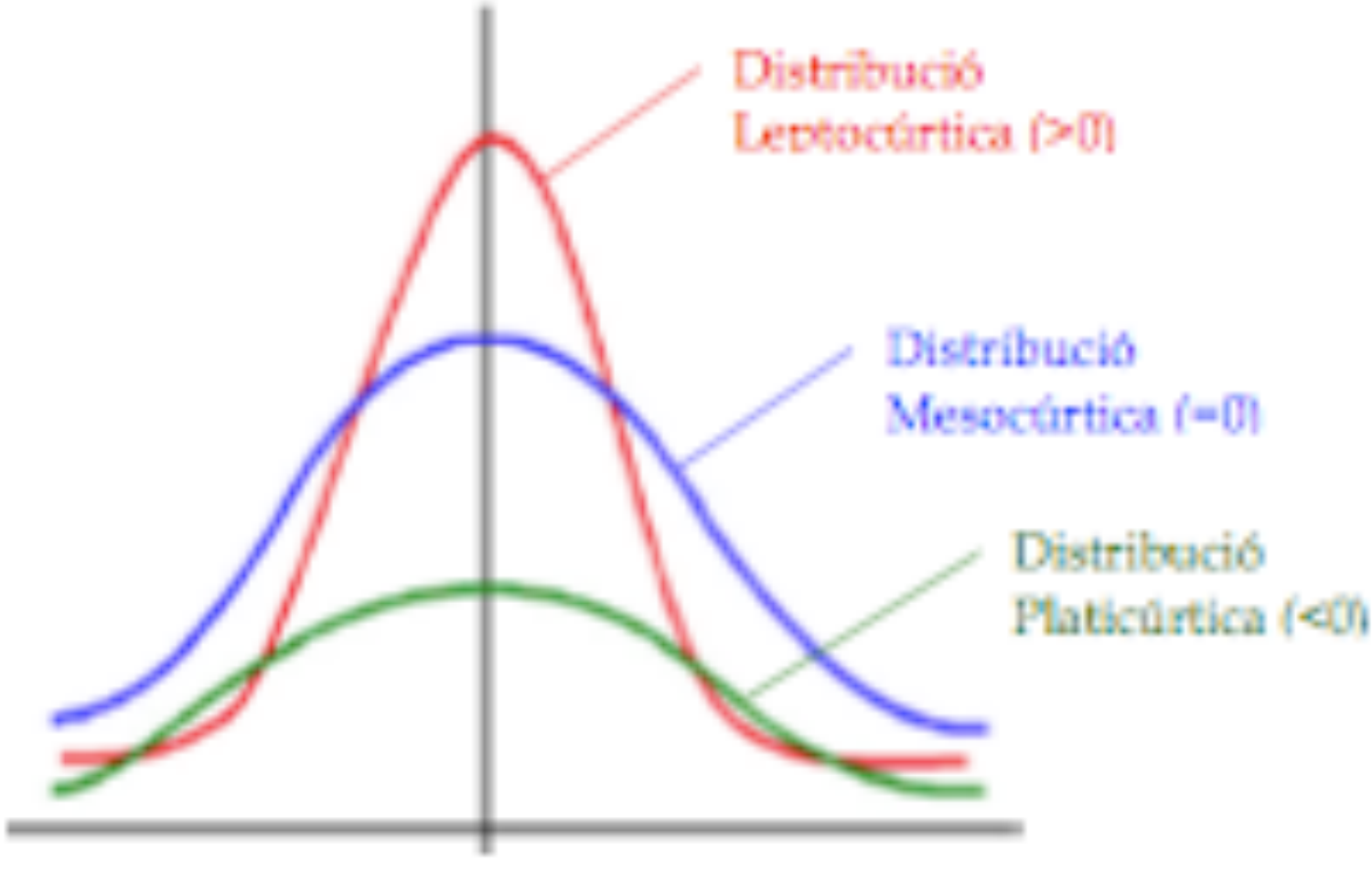


Figura 1.21: Curtosi plana o apuntada, respecte de la normal.