

Estadística: nocions bàsiques

Conceptes

L'**estadística** és la ciència que es dedica a recopilar, analitzar, interpretar i representar dades relatives a fenòmens de la realitat.

En cada estudi estadístic hi ha almenys una característica que s'estudia: per exemple, l'altura dels nois de 15 anys de Catalunya. La característica que s'estudia s'anomena **variable estadística**. En l'exemple anterior, els nois de 15 anys de Catalunya constitueixen la **població** que és objecte d'estudi. Cada noi de 15 anys de Catalunya és un **individu** de la població.

De vegades els individus que componen la població estudiada són persones, però de vegades no. Per exemple, podem estudiar si les peces que fabrica una màquina són defectuoses o no són defectuoses. Aquí la població està formada per les peces que fabrica la màquina i els individus de l'estudi no són persones, sinó que són les peces que fabrica la màquina.

Hi ha diversos **tipus de variables estadístiques**:

- **Qualitatives**: s'expressen mitjançant una qualitat (que no és un número), per exemple, el **sexe** (home o dona), el **color dels ulls** (marrons, blaus, negres, etc.), l'**estat civil** (casat o solter), etc.
- **Quantitatives**: s'expressen mitjançant una quantitat (que és un número). Hi ha dos subtipus:
 - o **Contínues**: són aquelles en què la variable pot prendre qualsevol valor entre dos valors donats, per exemple, l'**altura** (entre 1,65 m i 1,66 m pot haver 1,653 m, 1,6541 m, etc.) o el **pes** d'una persona.
 - o **Discretes**: són aquelles en què la variable no pot prendre qualsevol valor entre dos valors donats, per exemple, el **nombre de fills** d'una parella (entre 3 i 4 no pot haver 3,235 fills), el **nombre de germans**, el **nombre d'electrodomèstics** venuts en una botiga cada any, el **nombre d'espectadors** dels partits de futbol, etc.

De vegades resulta molt costós (en temps i diners) fer un estudi a tota la població i només es fa l'estudi en un subconjunt d'ella (en una part seva). En aquests casos es diu que l'estudi s'ha fet en una **mostra** (que haurà de ser **representativa** de tota la població, és a dir, la composició que haurà de tenir la mostra s'haurà d'assemblar a la de la població).

Recollida, organització i representació de les dades

Després de recollir les dades de l'estudi estadístic convé ordenar-les o organitzar-les en taules per tal de poder calcular després certs paràmetres estadístics o bé per representar-les mitjançant gràfics diversos.

Exemple: En una aula es mira el color dels ulls dels alumnes i s'obtenen els resultats següents:

M	M	N	B	N	M	M	M	N	B	B	N	N	M	M
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Aquí M = "marrons", N = "negres" i B = "blaus".

Si fem el recompte per a aquesta variable qualitativa, posant-ho en forma de taula tenim que:

Color	Recompte	n_i
M	I I I I I I I	7
N	I I I I I	5
B	I I I	3
Totals →		$N=15$

Aquí $N = 15$ és el nombre total de dades i n_i representen les **frequències absolutes** de cada color d'ulls (el nombre de vegades que surt cada color d'ulls).

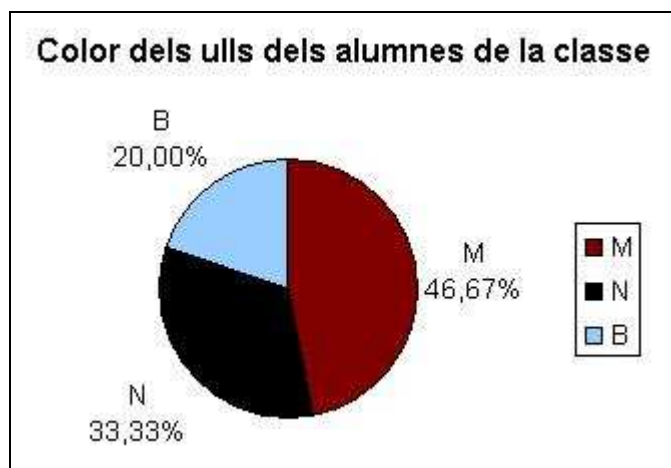
A la taula anterior podem afegir una columna (la de les freqüències relatives): $f_i = n_i/N$, que contenen els tants per un de cada color d'ulls respecte al total. Si les multipliquem per 100 el que tenim són **percentatges**.

Color	Recompte	n_i	f_i	Percentatge
M	I I I I I I I	7	$7/15=0,4667$	46,67%
N	I I I I I	5	$5/15=0,3333$	33,33%
B	I I I	3	$3/15=0,2$	20%
Totals →		$N=15$	1	100%

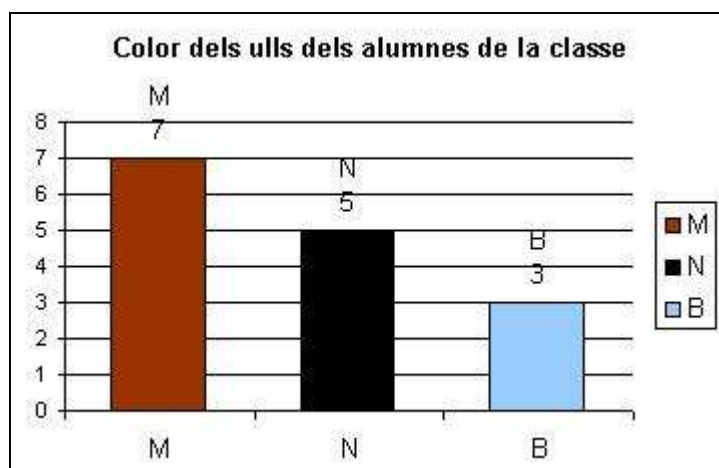
Si ens interessa fer un gràfic de sectors (també anomenat diagrama de sectors o diagrama circular) convé afegir una columna a la dreta amb l'angle que correspon a cada sector. L'angle es pot calcular fent una senzilla regla de tres, tenint en compte que tota la circumferència sencera serien 360° o bé usant que $\text{angle}(^\circ) = f_i \cdot 360$.

Color	Recompte	n_i	f_i	Percentatge	Angle ($^\circ$)
M	I I I I I I I	7	$7/15=0,4667$	46,67%	$168,0^\circ$
N	I I I I I	5	$5/15=0,3333$	33,33%	$120,0^\circ$
B	I I I	3	$3/15=0,2$	20%	$72,0^\circ$
Totals →		$N=15$	1	100%	360°

Amb un transportador d'angles dibuixaríem un sector de la mida corresponent i podríem fer el **gràfic de sectors**:



La representació gràfica també es podria fer en un **diagrama de barres** posant en l'eix y les freqüències absolutes:



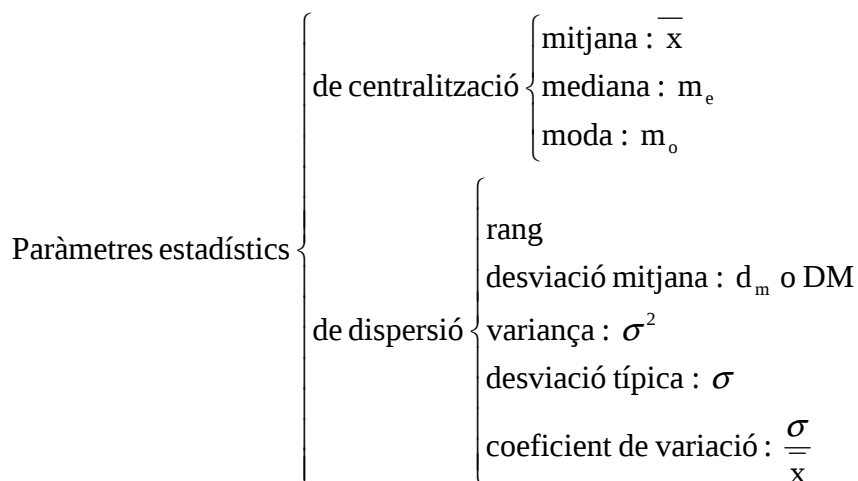
Un altre exemple:

Suposem que anem preguntant casa per casa quants televisors hi ha en cada habitatge i obtenim els resultats següents:

2	2	1	0	1	1	1	3	2	1	2	1	2	0	1	2	2	1	2	2
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

La variable, en aquest cas, és quantitativa. Això té una repercussió immediata, que és que dóna peu a fer-hi més manipulacions. S'hi poden fer més coses, s'hi poden calcular més **paràmetres estadístics**.

Els paràmetres estadístics són uns números que ens proporcionen informació sobre com estan distribuïdes les dades de l'estudi estadístic. En particular, tenim dos tipus de paràmetres estadístics, que se subdivideixen així:



Cal dir que els **paràmetres de centralització** ens diuen al voltant de quins valors s'agrupen les dades de la nostra distribució estadística. El més conegut és la mitjana (moltes vegades els alumnes han de calcular la seva nota **mitjana** d'exàmens).

Els **paràmetres de dispersió** ens informen sobre si les dades estan molt separades o estan molt properes entre ells. El més fàcil de calcular és el **rang**, ja que rang = $x_{\text{màx.}} - x_{\text{mín.}}$ (és la diferència entre el valor màxim i el mínim).

Tornem al nostre exemple, el del nombre de televisors en cada casa i anem a ordenar les dades de menor a major. Ho podem fer de cap o amb un programa informàtic, com l'Excel, usant la icona d'ordenar .

0	0	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Amb la ordenació feta tenim molt guanyat. Per una banda podem buscar quin és el valor que es queda enmig (que s'anomenaria **mediana**, m_e). En el nostre cas no seria únic, ja que hi ha dos valors que es queden enmig, que són el 1 i el 2.

									1	2										
--	--	--	--	--	--	--	--	--	---	---	--	--	--	--	--	--	--	--	--	--

Quan passa això, que hi ha dos valors, es diu que la mediana és el promig dels dos valors. En el nostre cas:

$$m_e = \frac{1+2}{2} = 1,5 \text{ televisors}$$

Si comptem posicions, aquests 1 i 2 ocupen les posicions 10a i 11a.

Quan el nombre de dades és petit és fàcil trobar quin valor o quins valors es queden enmig, però quan hi ha molts, és més difícil. En aquest cas s'utilitzen dues formuletes per saber quin valor és la mediana:

- Si el nombre de dades N és parell (com abans), els dos valors que es queden enmig són els que ocupen les posicions $\frac{N}{2}$ i $\frac{N}{2}+1$. Podem veure que en el nostre cas funcionava: les posicions buscades eren: $\frac{20}{2}=10$ i $\frac{20}{2}+1=10+1=11$ (la 10a i la 11a, com havíem vist abans).
- Si el nombre de dades N és imparell només hi ha un valor que es queda enmig, que és el que ocupa la posició $\frac{N+1}{2}$. Per exemple, si hi hagués 123 dades, el que ocuparia la posició central seria el que ocuparia la posició $\frac{123+1}{2}=\frac{124}{2}=62$, o sigui, el 62è.

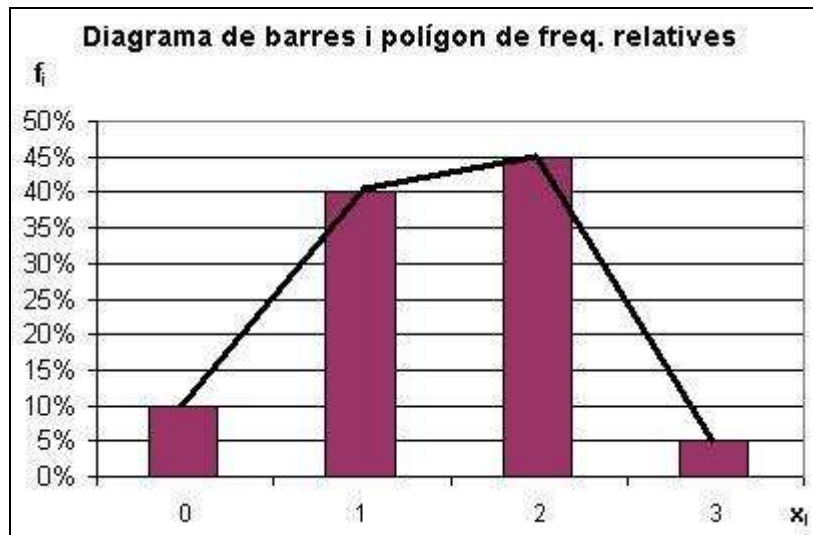
Tornant a les dades ordenades:

0	0	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	3
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

També veiem que ens faciliten organitzar la informació en forma de taula:

x_i	n_i	f_i	Percentatge
0	2	$2/20=0,1$	10%
1	8	$8/20=0,4$	40%
2	9	$9/20=0,45$	45%
3	1	$1/20=0,05$	5%
Total $\rightarrow$	$N=20$	1	100%

A partir de les freqüències absolutes (n_i) o de les relatives (f_i) podem fer un **diagrama de barres** acompanyat del seu **polígon de freqüències** (el polígon de freqüències seria la línia que uniria els punts centrals superiors de cada barra).



Ara ens podem plantejar: quin és el nombre de televisors que més apareix en l'estudi? És 2. Dos televisors en cada casa és el més freqüent. Direm, doncs, que la **moda** d'aquesta distribució de dades estadístiques és $m_o=2$ televisors.

Com podríem calcular la **mitjana** ($\bar{x}$) del nombre de televisors que hi ha en cada casa? Una manera fàcil de fer-ho seria:

$$\bar{x} = \frac{0 \cdot 2 + 1 \cdot 8 + 2 \cdot 9 + 3 \cdot 1}{20} = \frac{29}{20} = 1,45 \text{ televisors}$$

A continuació veurem una manera també senzilla, però usant la taula de freqüències (i afegint-hi una columna):

x_i	n_i	$x_i \cdot n_i$
0	2	0·2
1	8	1·8
2	9	2·9
3	1	3·1
Totals →	N=20	$\Sigma x_i \cdot n_i = 29$

Aquí $\Sigma x_i \cdot n_i$ significa que hem de multiplicar cada x_i per n_i i fer la suma (el símbol Σ s'anomena “sumatori” i vol dir que s’ha de fer la suma per a tot el que té a la seva dreta).

La fórmula que serveix per calcular la **mitjana** és la següent:

$$\bar{x} = \frac{\sum x_i \cdot n_i}{N}$$

En el nostre cas $\bar{x} = \frac{\sum x_i \cdot n_i}{N} = \frac{29}{20} = 1,45$ televisors (igual que amb l’altre mètode).

Bé, fins ara hem calculat els tres **paràmetres de centralització** següents:

$$m_e = 1,5 \text{ televisors}$$

$$m_o = 2 \text{ televisors}$$

$$\bar{x} = 1,45 \text{ televisors}$$

Les dades s’agrupen al voltant d’aquests paràmetres.

Ara ens interessaria calcular els **paràmetres de dispersió** que apareixien algunes pàgines enrere. Recordem que aquests paràmetres ens informaven sobre si les dades estan molt separades o molt juntes.

Ja hem dit que el paràmetre de dispersió més fàcil de calcular era el **rang**, ja que és la diferència entre el valor de x més gran i el més petit:

$$\text{Rang} = x_{\text{màx}} - x_{\text{mín}} = 3 - 0 = 3 \text{ televisors.}$$

Els altres paràmetres de dispersió tenen fórmules més complicades i, per calcular-los, ens hem d’enrecordar que $\bar{x} = 1,45$ i hem d’afegir més columnes a la taula que hem usat abans:

x_i	n_i	$x_i \cdot n_i$	$ x_i - \bar{x} \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$
0	2	0·2	$ 0 - 1,45 \cdot 2 = 2,9$	$(0 - 1,45)^2 \cdot 2 = 4,205$
1	8	1·8	$ 1 - 1,45 \cdot 8 = 3,6$	$(1 - 1,45)^2 \cdot 8 = 1,62$
2	9	2·9	$ 2 - 1,45 \cdot 9 = 4,95$	$(2 - 1,45)^2 \cdot 9 = 2,7225$
3	1	3·1	$ 3 - 1,45 \cdot 1 = 1,55$	$(3 - 1,45)^2 \cdot 1 = 2,4025$
Totals →	N=20	$\Sigma x_i \cdot n_i = 29$	$\sum x_i - \bar{x} \cdot n_i = 13$	$\sum (x_i - \bar{x})^2 \cdot n_i = 10,95$

[Recordeu que $|\dots|$ significa “fer el valor absolut” i vol dir “deixar igual el nombre de dins si és positiu” o “convertir el nombre de dins en positiu si era negatiu”.]

Fórmules per al càlcul dels paràmetres de dispersió:

$$\text{Desviació mitjana} = d_m = DM = \frac{\sum |x_i - \bar{x}| \cdot n_i}{N} = \frac{13}{20} = 0,65 \text{ televisors}$$

$$\text{Variança} = \sigma^2 = \frac{\sum (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{10,95}{20} = 0,5475 \text{ televisors}^2 \text{ (atenció a la unitat!)}$$

Desviació típica o desviació estàndard = σ = és l'arrel quadrada de la variança = $\sqrt{0,5475} = 0,7399$ televisors.

Coefficient de variació = $CV = \frac{\sigma}{\bar{x}} = \frac{0,7399}{1,45} = 0,5103$, que en tant per cent és 51,03% (bastant elevat!).

Exercici per practicar:

Observa les dades següents:

3	2	2	4	0	3	2	2	1	0	1	3	2	2	2	4	3	2	3	1	1	0	2	1	1
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Corresponen al nombre d'assignatures que han suspès els alumnes d'una classe.

Amb aquesta variable estadística es vol que facis el mateix que s'ha fet amb la variable estadística del nombre de televisors que hi havia en cada casa. Has de fer tots els passos, igual que en els exemples anteriors (l'ordenació, el recompte, la taula, el diagrama de barres i el polígon de freqüències, el càlcul dels paràmetres de centralització i dispersió amb les fórmules i la taula quan convingui, etc.).

Observacions finals

En aquest dossier no s'han explicat què son les **freqüències absolutes acumulades** ni les **freqüències relatives acumulades**.

També cal dir que, com a gràfics estadístics, a més del **diagrama de barres** i el **de sectors**, vistos en pàgines anteriors, també existeixen els següents: **histograma**, **pictograma**, **gràfic de dispersió** (o **de núvol de punts**), **piràmide de població**, **cartograma**, **diagrama de línies**, **gràfic evolutiu** i **comparatiu**, **diagrama de caixa**...

D'aquests continguts no fem una descripció, atès que aquest dossier només conté nocions bàsiques.