TEMA 6. ESTADÍSTICA UNIDIMENSIONAL

Estadística

L' estadística s'ocupa de recollir i ordenar dades referides a diversos fenòmens per després analitzar-les, interpretar-les i, en alguns casos, fer previsions

Fases:

a) Planificació de l'estudi i de les fases
b) Recollida de dades
c) Expressió de les dades en taules i gràfics
d) Anàlisi i interpretació a través del càlcul de paràmetres matemàtics (constants numèriques que ens donen informació de les característiques de l'estudi)
e) Previsió de resultats en relació a la característica indicada

Elements:

- Població: conjunt format per tots els elements de l'estudi
- Mostra: part de la població que forma l'estudi
- Individu: cada element individual que forma part de l'estudi
- Mida o grandària de la mostra (N): nombre d'individus que formen la mostra
- Variable: característica estudiada

Ex:

Es vol fer un estudi estadístic en relació al pes dels alumnes matriculats a 4t d'ESO a Catalunya. Per realitzar l'estudi s'han seleccionat 300 alumnes de diferents centres escolars.

Elements:

Població: tots els alumnes matriculats a 4t d'ESO a Catalunya
Mostra: els 300 alumnes als que es pesarà
Individu: cada alumne que forma part de la mostra
Mida: 300
Variable estadística: pes

Fases:

D'una correcta planificació i recollida de dades dependrà la fiabilitat dels resultats.

En la planificació caldrà determinar en quins instituts es fa l'estudi, com s'escolleixen, és important que corresponguin a localitats petites o grans ciutats? s'ha de pesar al mateix nombre de noies que de nois? pot alguna d'aquestes coses modificar els resultats?.

En la recollida de dades haurem de discutir com es recullen les dades: es pregunta als alumne el seu pes o es pren el pes als alumnes?, si és així: qui ho fa?, a quina hora?, amb roba o sense?, amb diferents bàscules?, en quina unitat?, fins a quina xifra decimal es pren la mesura?, ...

Tipus de variables

a) Qualitatives. Els valors de les variables no es pot associar amb un nombre. Exemple: sexe (home o dona)
b) Quantitatives. Els valors s'associen amb un nombre. Hi ha de dos tipus:
    - Discretes. Només poden prendre determinats valors. Exemple: el nombre de fills pot ser 0, 1, 2, … però no 3,27
    - Contínues. Entre dos nombres la variable pot prendre infinits valors. Exemple: l'alçada entre 1,70 m i 1,80 m potser 1,74m, 1,654 m, …

Després de recollir les dades, tal com em indicat abans, aquestes s'expressaran en taules i gràfics. En relació a aquests processos cal diferenciar entre:

- variables qualitatives o quantitatives discretes;
- variables quantitatives contínues.

Taules de freqüències per variables qualitatives o quantitatives discretes

Les dades recollides es recompten i es recullen els resultats en taules.

A les taules tenim:

- freqüència absoluta (n_i). És el nombre de vegades que apareix un determinat valor o tipus de la característica estudiada;

- freqüència relativa (f_i). És el resultat de la divisió entre la freqüència absoluta i la mida de la mostra: $f_i = \frac{n_i}{N}$. La suma de les freqüències relatives ha de ser 1;

- percentatge (%). És el resultat de multiplicar la freqüència relativa per 100.
% = $f_i \cdot 100$. La suma de percentatges ha de ser 100.

En el cas de variables quantitatives a més hi ha:

- freqüència absoluta acumulada (N_i) de una dada x_i és la suma de les freqüències absolutes dels valors més petits o iguals a aquesta dada

$$N_i = n_1 + n_2 + \ ... + n_i$$

- freqüència relativa acumulada (F_i) de una dada x_i és la suma de les freqüències relatives dels valors més petits o iguals a aquesta dada

$$F_i = f_1 + f_2 + ... + f_i$$

Ex:
Les notes de matemàtiques en un grup de 20 alumnes han estat:

2	5	2	7	4	4	3	8	6	2
5	9	4	4	3	7	8	3	9	5

X 100

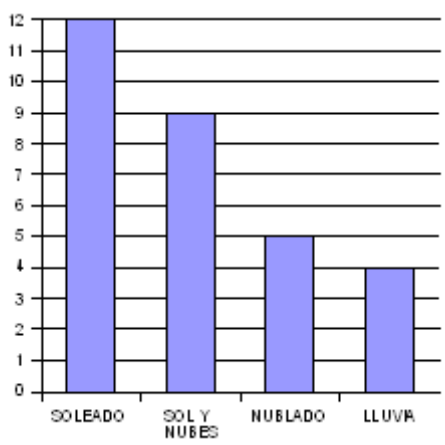
x_i	n_i	N_i	$f_i = \frac{n_i}{N}$	F_i	$\%=f_i \cdot 100$
2	3	3	0,15 = 3/20	0,15	15
3	3	6	0,15 = 3/20	0,30	15
4	4	10	0,20 = 4/20	0,50	20
5	3	13	0,15 = 3/20	0,65	15
6	1	14	0,05 = 1/20	0,70	5
7	2	16	0,10 = 2/20	0,80	10
8	2	18	0,10 = 2/20	0,90	10
9	2	20	0,10 = 2/20	1,00	10
∑	20		1,00		100

<u>Gràfics estadístics per variables qualitatives i quantitatives discretes</u>
Una altre forma d'ensenyar els resultats de les dades recollides són les representacions gràfiques. Les més importants són:

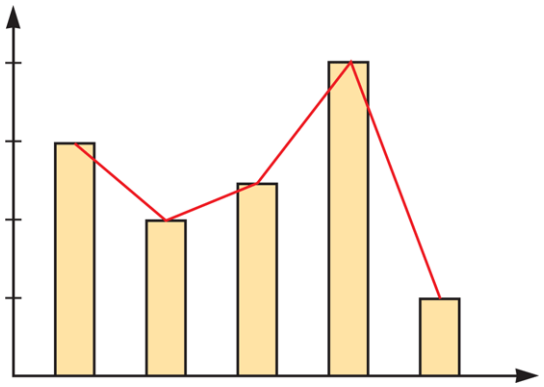
a) diagrama de barres
b) diagrama de sectors
c) diagrama de línies
d) pictogrames

a) Diagrama de barres

A l'eix horitzontal representem els valors de la variable i a l'eix vertical les freqüències
Ex: Estat del temps a una població durant un mes



Quan la variable aleatòria és quantitativa, podem unir amb línies els extrems superiors de les barres i obtenir un **polígon de freqüències**.
Ex:

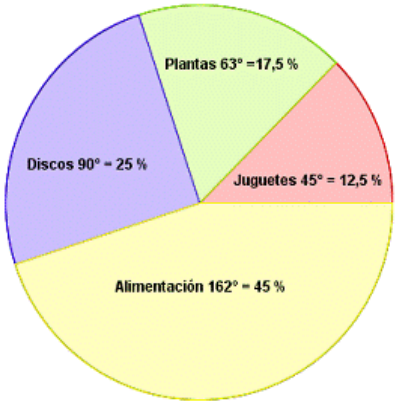


b) Diagrama de sectors
Les dades es representen en un cercle dividit en sectors que representen els valors de la variable. L'angle de cada sector és proporcional a la freqüència de la dada que representa
Per calcular l'angle del sector circular s'aplica la fórmula:

$$\text{Angle} = f_i \cdot 360^o = \frac{n_i}{N} \cdot 360^o$$

Ex:
En una botiga s'han venut joguines per valor de 125 €, plantes per 175€, música per 250€ i alimentació per 450€

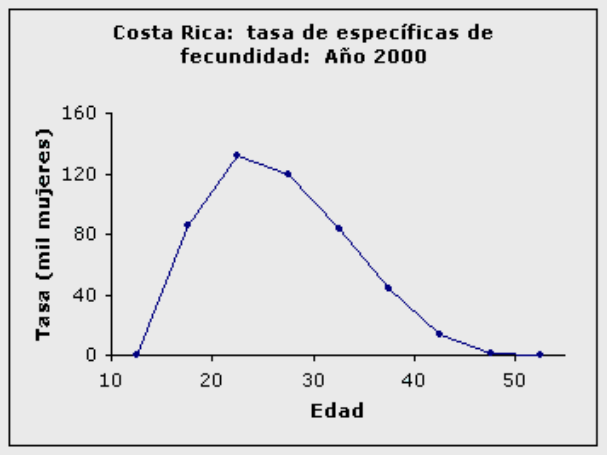
x_i	n_i	$f_i = \frac{n_i}{N}$	Angle= $f_i \cdot 360^o$
Joguines	125	0,125	45^o=0,125·360^o
Plantes	175	0,175	63^o=0,175·360^o
Música	250	0,250	90^o=0,250·360^o
Alimentació	450	0,450	162^o=0,450·360^o
$\sum$	1000	1,000	360^o



c) Diagrama de línies

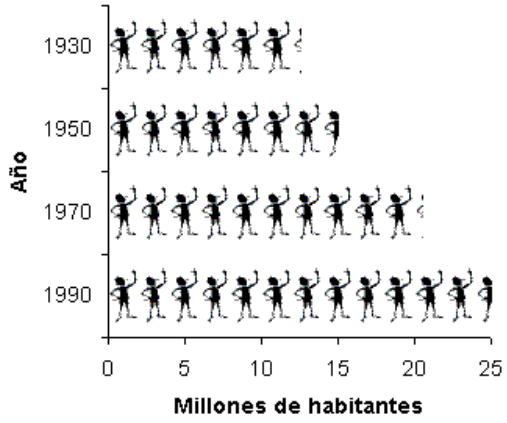
S'utilitza per a variables qualitatives quan volem representar l'evolució d'una variable en relació al temps

Ex:



d) Pictogrames

S'utilitzen figures relacionades amb la variable estudiada.



Taules de freqüències per variables quantitatives contínues

Quan treballem amb variables contínues ens podem trobar amb una gran varietat de resultats que podem ajuntar en grups sense que les conclusions de l'estudi quedin alterades. Aquests grups de nombres s'anomenen intervals. De vegades, segons el nombre de dades que continguin, els intervals de la taula de freqüències poden tenir amplituds diferents.

Ex:
Estem fent un estudi en relació al temps que necessiten els alumnes del nostre centre en desplaçar-se a peu fins a l'institut des de casa seva. Per això demanem que, durant una setmana, mesurin el temps amb ajut d'un rellotge i ens donin el promig dels resultats obtinguts en minuts amb una xifra decimal. Així un company ens dona la dada 2,5 minuts que resulta ser la més petita, mentre altre ens dona com a temps promig 18,2 minuts que és la dada més gran. Podem preveure que la quantitat de resultats diferents que tindrem serà enorme i que el que més ens convé es establir intervals:

De 0 a 5 minuts justos	→	[0 , 5]
de 5 a 10 minuts justos	→	(5 , 10]
de 10 a 15 minuts justos	→	(10 , 15]
de 15 a 20 minuts	→	(15 , 20)

que ens faran més fàcil la recollida de dades.

Per ser rigorosos el nombre d'intervals que em de fer queda determinat amb la fórmula:

$$N^o\ d'intervals = \sqrt{nombre\ de\ dades}$$

Taules de freqüències
En el cas de fer estudis amb dades agrupades en intervals cal trobar la **marca de classe** que serà el nombre a partir del qual farem els càlculs.

$$Marca\ de\ classe = \frac{suma\ dels\ extrems\ d'un\ \text{int}erval}{2}$$

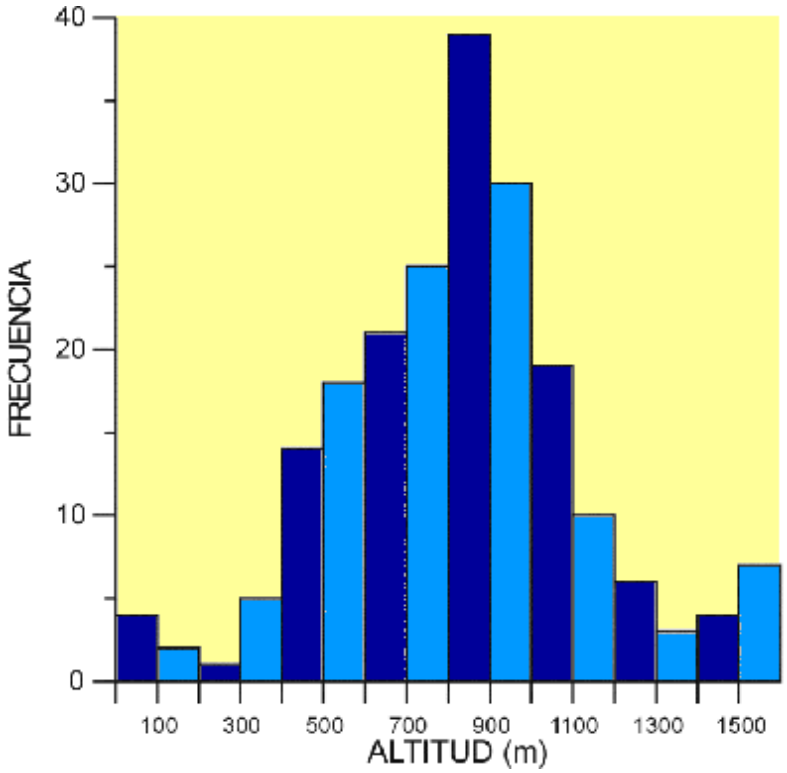
Ex:

x_i	marca de classe	n_i	N_i	$f_i = \frac{n_i}{N}$	F_i	$\%=f_i \cdot 100$
[0,5]	2,5	4	4	0,20	0,20	20
(5,10]	7,5	8	12	0,40	0,60	40
(10,15]	12,5	5	17	0,25	0,85	25
(15,20]	17,5	3	20	0,15	1,00	15
$\sum$				1,00		100

Gràfics estadístics per variables quantitatives contínues

La representació gràfica d'aquestes variables és l' **histograma** on l'àrea del rectangle és proporcional a la freqüència de la dada.

Ex: Nombre d'escarbats *Pseudolucanus barbarossa* (Fabricius, 1801) en la Península Ibèrica segons l'altitud (metres)



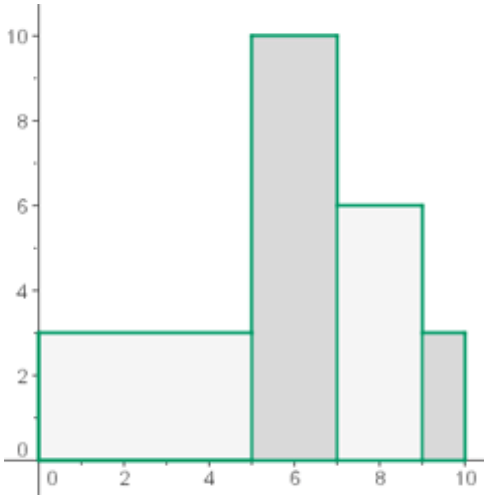
Per construir histogrames amb intervals de diferent amplitud i per tal que l'àrea del rectangle representi la freqüència corresponent, i pugi ser comparable amb les altre, hem de calcular les altures dels rectangles

$$h_i = \frac{f_i}{a_i}$$

on a_i és l'amplitud de l'interval.

Ex: En un grup de 50 alumnes s'han obtingut les següents qualificacions

	[0, 5)	[5 , 7)	[7 , 9)	[9, 10)
f_i	15	20	12	3
h_i	3	10	6	3



Paràmetres estadístics

Després de recollir les dades i expressar els resultats obtinguts en taules o gràfiques resulta interessant calcular una sèrie de paràmetres o de nombres que ens expliquen més coses de la característica estudiada i que ens deixa comparar els resultats de dos estudis semblants.

Aquests nombres característics o paràmetres poden ser:

- De centralització. Ens indiquen al voltant de quin valor central es distribueixen les dades.
- De dispersió. Ens diuen com estan de concentrats o dispersos al voltant d'un valor central les dades obtingudes.
- De posició. Divideixen el grup de dades en grups amb el mateix nombre de dades.
- D'asimetria. La distribució de les dades és simètrica si valors que es troben a la mateixa distància d'un valor central tenen les mateixes freqüències.

i) Paràmetres de centralització:

a) Mitjana aritmètica ($\bar{x}$). És el valor promig de les dades.

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \ldots + x_n \cdot f_n}{N}$$

Ex: En un curs hi ha 8 nois i 6 noies. Les notes finals dels nois han estat: 2,3,4,5,6,6,7 i 7, i les de les noies: 3,4,5,6,7 i 8.

La mitjana aritmètica dels nois serà $\bar{x}_1 = \frac{2 \cdot 1 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 2 + 7 \cdot 2}{8} = 5$

La mitjana aritmètica de les noies serà $\bar{x}_2 = \frac{3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1 + 6 \cdot 1 + 7 \cdot 1 + 8 \cdot 1}{6} = 5{,}5$

I la de tota la classe $\bar{x} = \frac{2\cdot1 + 3\cdot2 + 4\cdot2 + 5\cdot2 + 6\cdot3 + 7\cdot3 + 8\cdot1}{14} = 5{,}2$

Ex: Les qualificacions obtingudes per 88 persones en relació al seu rendiment en el treball han estat

Puntuació	[38,44)	[44,50)	[50,56)	[56,62)	[62,68)
Nº de treballadors	12	15	20	25	16

Per calcular la mitjana aritmètica utilitzarem la marca de classe,

$$\bar{x} = \frac{41\cdot12 + 47\cdot15 + 53\cdot20 + 50\cdot25 + 65\cdot16}{88} = 51{,}67$$

Ex: Un professor fa tres controls parcials i un final. El segon parcial val el doble del primer, el tercer el doble del segon, i el final el triple del tercer. Un alumne té en el primer parcial un 10, en el segon un 7, en el tercer un 5 i en el final un 4,25. Quina serà la seva nota final?

Es tracta d'una mitjana ponderada

$$\bar{x} = \frac{10\cdot p + 7\cdot2\,p + 5\cdot4\,p + 4{,}25\cdot12\,p}{p + 2\,p + 4\,p + 12\,p} = 5$$

Observacions:

- Els valors exageradament grans o petits en els extrems de la distribució fan que la mitjana variï de forma considerable;
- Quan les dades d'una distribució es troben agrupades en classes i alguna d'elles està oberta, no és possible calcular la mitjana i es fan servir altres paràmetres.

b) Mediana (Me). És el valor de la variable que ocupa la posició central quan ordenem les dades quantitatives de menor a major.
En el cas de dades sense agrupar en intervals la mediana és el valor central (N és un nombre senar) o la semisuma dels valors centrals (N és un nombre parell).

Observacions:
- La mediana s'utilitza com a mesura central en les distribucions asimètriques o quan les dades es troben agrupades en classes i alguna és oberta.

c) Moda (Mo). És el valor de la variable que té una major freqüència

Ex:

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

x_i	n_i	$f_i = \frac{n_i}{N}$	$x_i \cdot f_i$
0	4	0,20 = 4/20	0 = 0·4
1	8	0,40 = 8/20	8 = 1·8
2	6	0,30 = 6/20	12 = 2·6
3	2	0,10 = 2/20	6 = 3·2
∑	20	1,00	26

$\bar{x} = \frac{26}{20} = 1{,}3$ Me = 1 Mo = 1

Observacions:

- Una distribució pot no tenir moda o tenir més d'una, en aquest cas parlem de distribució bimodal, trimodal, …
- En la moda no intervenen tots els valors de la distribució;
- Si be és un paràmetre de centralització pot trobar-se proper als extrems de la distribució;
- En el cas de treballar amb intervals, per calcular la moda i la mediana assenyalarem l'interval corresponent però en realitat hi ha unes fórmules per dades agrupades (veure http://www.vitutor.com/estadistica.html)

• Relació entre mitjana aritmètica, mediana i moda.

Si quan es construeix el polígon de freqüències s'observa que la distribució és simètrica o lleugerament asimètrica, es dona la següent relació:

$$Me - Mo = 3 (x - Me)$$

de forma que, amb un cert error, podem aconseguir uns paràmetres en funció dels altres.

ii) Paràmetres de posició:

a) Quartils, Q_1, Q_2 i Q_3 són mesures que divideixen el conjunt de dades ordenades en quatre parts iguals, en cada una de les quals es troba el 25% de les dades;

b) Percentils o centils, P_k, són mesures que divideixen el conjunt de dades en cent parts iguals. Així $P_{25} = Q_1$, $P_{50} = Q_2 = Me$ i $P_{75} = Q_3$.

A http://www.vitutor.com/estadistica.htm trobarem les fórmules per dades agrupades en intervals.

Ex: Donada la taula

x_i	n_i	N_i
1	11	11
2	27	38
3	4	42
4	18	60
$\sum$	60	

Calculem els tres quartils:

Q_1 és la dada que ocupa la posició 15 $Q_1 = 2$
Q_2 = Me és la dada que ocupa la posició 30 $Q_2 = 2$
Q_3 és la dada que ocupa la posició 45 $Q_3 = 4$

Si volem calcular P_{65} em de saber en quina posició es troba la dada 65 si la mida de la mostra és 100

65 % de 60 és 39

P_{65} és la dada que ocupa la posició 39 $P_{65} = 3$,
es a dir, el 65 % de les dades és menor o igual a 3

iii) Paràmetres de dispersió:

a) Rang o recorregut (R), és la diferència entre el valor més gran i més petit de la variable.

R = valor màxim – valor mínim

Observacions:

- El grau de representació que tenen els paràmetres centrals és més gran quant més petit és el rang;
- S' aplica en procediments de control de qualitat, verificació de longitud, pes i volum;
- El rang té l'inconvenient que només depèn dels valors extrems. Per donar més estabilitat es fan servir:

Rang interquartílic: $Q = Q_3 - Q_1$
Rang entre percentil: $P = P_{90} - P_{10}$

b) Desviació mitjana (dm), és la mitjana aritmètica dels valors absoluts de les desviacions de cada dada. La desviació de una dada x_i és la seva diferència amb la mitjana.

$$dm = \frac{\sum_{i=1}^{n} \left| x_i - \bar{x} \right|}{N} = \sum_{i=1}^{k} \left| x_i - \bar{x} \right| \cdot f_i$$

c) Variància (σ^2). És la mitjana dels quadrats de les desviacions.

$$\sigma^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{n} x_i^{\,2} \cdot f_i}{N} - \bar{x}^{\,2}$$

Propietats:
- el valor de la variància és un nombre positiu o zero;
- la variància és zero quan tots els valors són iguals.

Observacions:
- la variància, igual que la mitjana aritmètica, és un paràmetre molt sensible als valors extrems;
- la variància no s'expressa en les mateixes unitats que les dades ja que les desviacions es troben elevades al quadrat.

d) Desviació típica (σ), és l'arrel quadrada positiva de la variància.

$$\sigma = s = +\sqrt{\sigma^2}$$

Propietats:
- el valor de la desviació típica és sempre positiu o zero, en el cas que els valors siguin tots iguals.

Observacions:
- Com en el cas de la mitjana aritmètica i la variància, és molt sensible als valors extrems;
- Quant més petita és la desviació típica més gran és la concentració de dades al voltant de la mitjana aritmètica.

e) Coeficient de variació de Pearson (CV), és el quocient de la desviació típica i la mitjana.

$$CV = \frac{\sigma}{\bar{x}}$$

Observacions:
- De dues distribucions, la que té un coeficient de variació més petit presenta una menor dispersió relativa, es a dir, la seva mitjana aritmètica és més representativa;
- S'utilitza per comparar el grau de dispersió de dues distribucions que es donen en unitats diferents.

Ex: Dues regions, A i B, amb la mateixa població, tenen com a renda mitjana 600 i 560 € respectivament i desviacions típiques de 180 i 140 €. Quina renda mitjana és més representativa?

$$CV_A = \frac{180}{600} = 0{,}3 \qquad CV_B = \frac{140}{560} = 0{,}25$$

És més representativa la B

f) Puntuacions típiques. Sigui *X* la variable estadística que pren valors $x_1, x_2, \ldots, x_n$ (puntuacions directes), s'anomenen puntuacions típiques als valors:

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Ex: Les notes d'un alumne en relació a dos test han estat

Test A : 50 punts
Test B : 32 punts

mentre que la mitjana aritmètica i la desviació típica dels tests són:

Test A: $\overline{x_A} = 45$ $s_A = 6$
Test B: $\overline{x_B} = 26$ $s_B = 2$

En quin dels dos test ha obtingut millor resultat l'alumne respecte al grup?

$$z_A = \frac{50-45}{6} = \frac{5}{6} \qquad z_B = \frac{32-26}{2} = 3$$

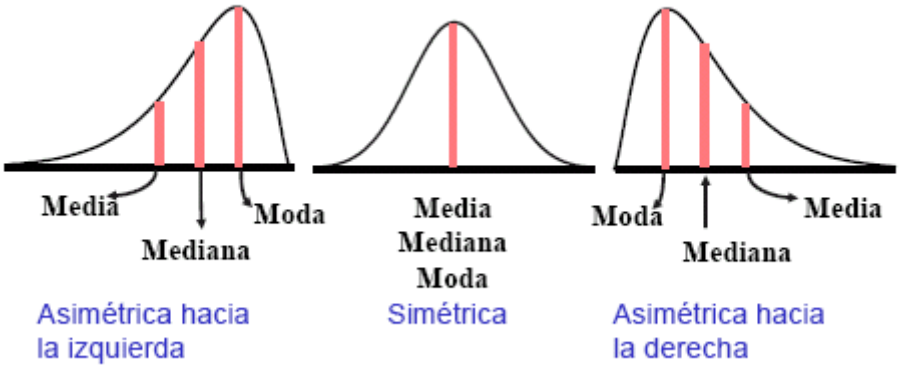
Si be la nota A ha estat més alta, la nota del test B és millor respecte al grup ja que està més per sobre que la mitjana que en el cas A.

Observacions:
- Es fan servir en ciències socials;
- Les puntuacions típiques es refereixen a puntuacions obtingudes per a cada individu del grup, mentre que la desviació típica fa referència a tot el grup.

iv) Paràmetres d'asimetria

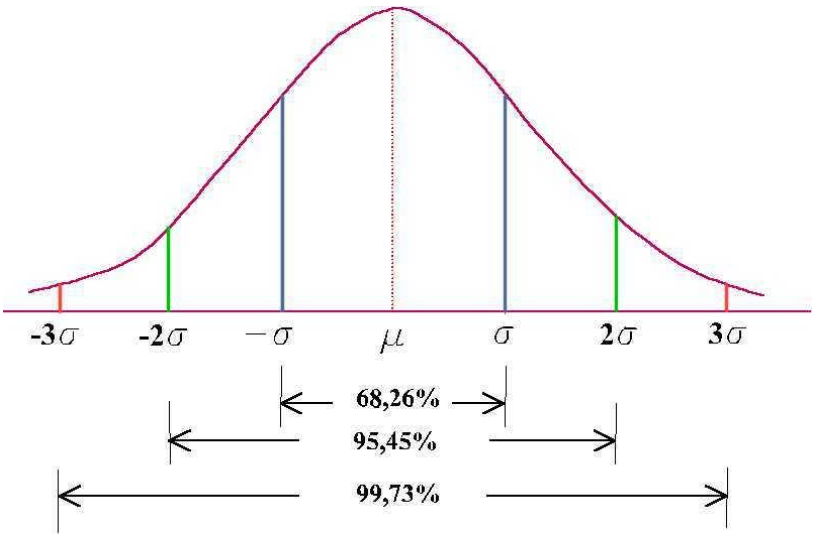
En una distribució simètrica x = Mo = Me.
Les distribucions que no són simètriques poden ser

Asimètriques a la dreta o positives: x > Me > Mo
Asimètriques a l'esquerra o negatives: x < Me < Mo



Es pot saber si una distribució és simètrica a través de la seva representació gràfica, mitjançant una taula o calculant les mesures de centralització.
La representació gràfica simètrica per excel·lència correspon a una **distribució normal** i s'anomena **Campana de Gauss**. En aquests gràfics el valor màxim correspon a la mitjana aritmètica i es compleix l'anomenada *Desigualtat de Txebixeff,*

el 68 % de les dades es troben a l'interval ($\bar{x}-\sigma$, $\bar{x}+\sigma$)
el 95 % de les dades pertanyen a l'interval ($\bar{x}-2\sigma$, $\bar{x}+2\sigma$)
i el 99 % es troben en ($\bar{x}-3\sigma$, $\bar{x}+3\sigma$)



En el cas de no complir-se diem que és una **dada atípica**.

Ex: En un estudi sobre la superfície d'explotacions agrícoles, s'han obtingut les dades següents:

Superfície (ha)	[0,5)	[5,20)	[20,50)	[50,100)	[100,300)
Percentatge (%)	54,2	26,8	4,5	4,4	4,5

Estudiem la concentració de dades.

Taula de freqüències

Superfície	x_i	f_i	x_i^2	$x_i \cdot f_i$	$f_i \cdot x_i^2$
[0,5)	*2,5*	*54,2*	*6,25*	*135,5*	*338,75*
[5,20)	*12,5*	*26,8*	*156,25*	*335,0*	*4187,50*
[20,50)	*35*	*10,1*	*1225*	*353,5*	*12372,50*
[50,100)	*75*	*4,4*	*5625*	*330,0*	*24750,00*
[100,300)	*200*	*4,5*	*40000*	*900,0*	*180000,00*
Total		*100,0*	*47012,5*	*2054,0*	*221648,75*

$$\bar{x} = \frac{2054{,}0}{100} = 20{,}54\ ha$$

$$\sigma = \sqrt{\frac{221648{,}75}{100} - (20{,}54)^2} = 42{,}36$$

La desviació típica és molt gran, pràcticament el doble de la mitjana, el que indica que les dades es troben molt disperses.
Si ara calculem els intervals:

$(\bar{x} - \sigma\ ,\ \bar{x} + \sigma\ )$ = (-21,82 , 62,90)
$(\bar{x} - 2\sigma\ ,\ \bar{x} + 2\sigma\ )$ = (-64,18 , 105,26)
$(\bar{x} - 3\sigma\ ,\ \bar{x} + 3\sigma\ )$ = (-106,54 , 147,62)

Observem que no és una dada típica. Si calgués ajustar la distribució, per tal que fos més simètrica eliminaríem l'últim interval.

El coeficient d'asimetria de Pearson mesura el grau d'asimetria **d'una distribució unimodal i campanoide**

$$A_p = \frac{\bar{x} - Mo}{\sigma}$$

Si

$A_p > 0$ asimetria positiva
$A_p < 0$ asimetria negativa
$A_p = 0$ simetria

Aquest coeficient no és gaire bo per mesurar asimetries lleus.

Ex: Una vacuna antitetànica es va administrar a 42 persones. A les 5 hores de la seva injecció es va prendre la temperatura dels vacunats

Temperatura (ºC)	37,0	37,2	37,5	38,0	38,1	38,5	39,0
Nº de persones	1	5	15	6	10	5	0

x_i	f_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
37,0	*1*	*37,0*	*1369,00*
37,2	*5*	*186,0*	*6919,20*
37,5	*15*	*562,5*	*21093,75*
38,0	*6*	*228,0*	*8664,00*
38,1	*10*	*381,0*	*14516,10*
38,5	*5*	*192,5*	*7411,25*
30,0	*0*	*0*	*0*
$\sum$	*42,0*	*1587,0*	*59973,30*

$\bar{x} = 37{,}78$

$\sigma = 0{,}78$ $\qquad A_p = \frac{37{,}78 - 37{,}5}{0{,}78} = 0{,}36$

$Mo = 37{,}5$

Asimetria positiva