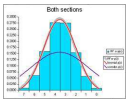
Estadística





L'Estadística és la part de les matemàtiques que tracta del recompte, ordenació i classificació de les dades que s'han obtés a partir de l'observació i estudi d'una població, per poder comparar i traure coclussions.

L'estudi estadístic té les següents fases:

- Arreplegada de dades.
- Organització i ordenació de les dades.
- Anàlisi de les dades.
- Obtenció de conclusions.

Arreplegada de dades. Conceptes generals d'estadística.

Població és conjunt format per tots els elements que són objecte d'estudi.

Mostra és un subconjunt representatiu de la població, que serveix per inferir característiques de tota la població.

Individu és cada un dels elements de la pobació o de la mostra.

Variable estadística (X) és cada una de les característiques de la població que volem estudiar.

Variable qualitativa es refereix a una característica que no pot ser mesurada amb números.

Variable quantitativa és la que pren valors númerics.

Variable discreta sols pren valors aïllats.

Variable contínua pot prendre qualsevol valor entre dos valors donats de la variable.

Organització i ordenació de les dades. Taula de freqüències. Gràfics estadístics.

Dada estadística (X_i) cada un dels valors que pren la variable estadística.

Distribució de freqüències

És una ordenació en forma de taula de les dades estadístiques, assignant-li a cada dada la seua freqüència. S'emplea en variables estadístiques discretes.

Freqüència absoluta (f_i) número de vegades que es dóna una dada estadística.

$f_1 + f_2 + f_3 + \cdots + f_n = N$ o també $\sum_{i=1}^{n} f_i = N$ on N representa el número total de dades.

Freqüència relativa (n_i) és el valor del quocient $n_i = \frac{f_i}{N}$. Evidentment $\sum_{i=1}^{n} n_i = 1$.

$\sum sumatori$; $\cap$ *intersecció*; $\cup$ *unió*; $\emptyset$ *conjunt buit*; $\in$ *pertany*; $\forall$ *per a tot*; $\rightarrow$ *aleshores*
$\subseteq$ *inclós en*; $\sqcap$ *Producte*; $/$ *de forma que*; $\leftrightarrow$ *si i sols si*; $\exists$ *existeix*; $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$; Estadística

Freqüència acumulada(F_i) és la suma de les freqüències absolutes de totes les dades iguals o inferiors a x_i.

Freqüència relativa acumulada(N_i) és el valor del quocient $N_i = \frac{F_i}{N}$.

Distribució de freqüències agrupades

És una ordenació en forma de taula de les dades agrupades. S'utilitza quan la variable és contínua o quan pren un gran número de valors.

Els valors s'agrupen en intervals que tenen la mateixa amplitud anomenats ***classes***.

A cada classe se li assigna la seua freqüència.

Cada ***classe*** és un interval tancat per l'esquerre i obert per la dreta$[L_{i-1}, L_i[$. És convenient que hi haja entre 5 i 15 intervals o classes.

Marca de classe és el punt mig de cada interval i és el valor que representa a tot l'interval per calcular els paràmetres estadístics.

$$X_i = \frac{L_{i-1} + L_i}{2}$$

Gràfics estadístics

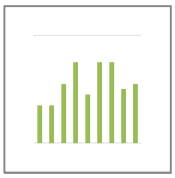
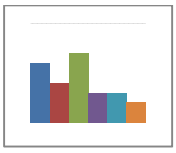


Diagrama de barres

S'utilitza en variables estadístiques discretes. Les dades es representen a l'eix X i en cada dada es representa una barra d'ltura igual a la seua freqüencia.



Histograma de freqüencies

S'utilitza en variables estadístiques contínues. En cada interval classe es representa un rectangle que té d'altura la frqüencia de cada classe.

Polígon de freqüencies

S'utilitza també en variables estadístiques contínues. Es construeix unint els punts mitjans de la base superior de cada rectangle de l'histograma.

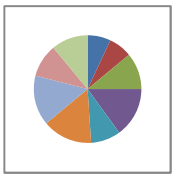


Diagrama de sectors

Com el seu nom indica cada dada esta representada en un sector circular l'angle del qual és proporcional a la freqüencia de la dada.

Anàlisi de les dades. Paràmetres estadístics.

A les següents expressions matemàtiques tindrem:

x_i representa el valor de la dada "i-essima" o de la marca de classe correspondent.

f_i representa la freqüència de la dada x_i.

Mesures de centralització

Mitjana aritmètica($\bar{x}$)

$\bar{x} = \frac{\sum f_i \cdot x_i}{N}$ on $N = \sum f_i$ que és el número d'individus de la població.

Moda(M_o)

És la dada de major freqüència. Una distribució pot ser bimodal (dos modes), trimodal (tres modes), ...

Mediana(M_e)

Si ordenem les dades de major a menor, la mediana (M_e) és el valor que està al mig. Si el número d'individus és un número parell, la mediana és valor mitjà dels dos termes centrals.

Mesures de dispersió

Recorregut o Rang

És la diferència entre el major valor de les dades i el menor valor de les dades.

Desviació mitjana(DM)

Aquest valor ens dóna informació sobre la dispersió de les dades respecte de la mitjana aritmètica $\bar{x}$. $DM = \frac{\sum f_i \cdot |x_i - \bar{x}|}{N}$

Millor informació que la deviació mitjana DM ens la proporciona la desviació típica. Per calcular la desviació típica hem de calcular la variància (*Var*).

Variància($\sigma_x{}^2$=*Var*) $\sigma_x{}^2$=*Var*$= \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum f_i \cdot x_i{}^2}{N} - \bar{x}^2$

Desviació típica(σ_x) $\sigma_x = +\sqrt{\sigma_x{}^2}$

Com més gran siga la deviació típica, més disperses estan les dades respecte de la mitjana arimètica $\bar{x}$.

Coeficient de variació(*CV*) *CV*=$\frac{\sigma_x}{\bar{x}}$ Com més s'aproxima a 0, menys dispersió hi ha.

Com més menut siga el coeficient de variació, més agrupades estan les dades al voltant de la mitjana aritmètica $\bar{x}$.

Mesures de posició

Percentils o centils

El percentil k, P_k, és el valor que deixa per baix d'ell el k % de la població.

Quartils

Quartil 1, Q_1, és el valor que deixa per baix d'ell el 25 % de la població. $Q_1 = P_{25}$

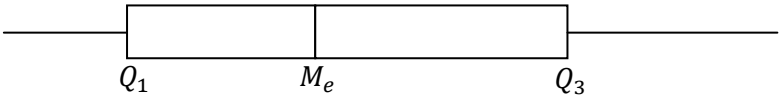
Quartil 2, Q_2, és el valor que deixa per baix d'ell el 50 % de la població. $Q_2 = P_{50}$=Mediana.

Quartil 3, Q_3, és el valor que deixa per baix d'ell el 75 % de la població. $Q_3 = P_{75}$

Recorregut interquartílic

El recorregut interquartílic=$Q_3 - Q_1$.

Diagrames de caixa



Conclusions. Inferència estadística.

Per què una mostra

Si vullguèrem estudiar l'altura de tots els valencians, de tots els espanyols o de tots els europeus és evident que el·legiríem una mostra representativa de la població i no sen's ocorreria medir a tots els ciutadans.

Si estem estudiant el percentatge de tornillos defectuosos que fabrica una empresa, també recorrirem a una mostra, seria impossible estudiar-los tots.

Si estem fent un sondeig del vot en les pròximes el·leccions, és necessari recorrer a una mostra ja que enquestar a tots els votants és quasi impossible.

Grandària de la mostra i nivell de confiança

Si el·legim bé la mostra, és licit pensar quels paràmetres que obtenim de la distribució presenten un cert error. Hem de presentar les dades obtingudes de l'estudi estadístic diguent el nivell de confiança que ens mereixen. Aquest grau de confiança es dóna amb un percentatge o probabilitat.

Com més gran siga la mostra, si s'ha fet bé, el nivell de confinça deuria ser major.

Com més representativa siga la mostra de tota la població, més gran serà el nivell de confiança.

Mostra aleatòria

Per el·legir bé la mostra hi ha tota una disciplina estadística anomenada mostratge, però hi ha una condició necessaria per a que siga representativa i fiable al màxim: els individus de la mostra han de ser ecollits al atzar.

Exercicis proposats

En els següents exemples s'estudien dues variables, una discreta i l'altra continua.

Busca, en la teua vida quotidiana, variables estadístiques que consideres interessants i fes un estudi similar al que s'ha fet als exemples.

Exemple de variable estadística discreta

Les notes de matemàtiques en selectivitat de 100 alumnes foren:

5-4-6-7-4-3-6-6-7-1-4-3-2-8-9-7-5-6-7-1-4-6-8-3-4-7-9-8-4-9-7-6-5-3-2-5-7-8-9-6-7-6-4-4-3-5-6-8-7-5
8-2-3-1-5-4-7-9-2-4-1-9-9-6-7-7-3-4-2-2-6-7-7-4-3-4-8-9-1-5-4-6-3-8-9-1-9-7-8-3-6-9-3-2-1-6-4-5-6-8

Taula de freqüències

x_i	f_i	F_i	$n_i = \frac{f_i}{N}$	N_i	$x_i \cdot f_i$	$\lvert x_i - \bar{x}\rvert \cdot f_i$	$x_i^2 \cdot f_i$
1	7	7	07%	07%	7	30'31	7
2	7	14	07%	14%	14	23'31	28
3	11	25	11%	25%	33	25'63	99
4	15	40	15%	40%	60	19'95	240
5	9	49	09%	49%	45	2'97	225
6	15	64	15%	64%	90	10'05	540
7	15	79	15%	79%	105	25'05	735
8	10	89	10%	89%	80	26'70	640
9	11	100	11%	100%	99	40'37	891
	$N = 100$		1		533	204'34	3405

Diagrama de barres

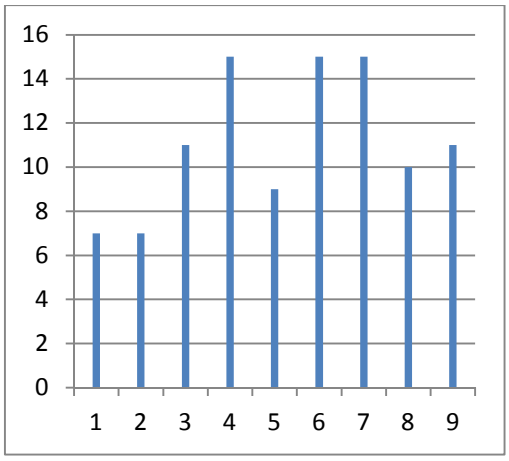
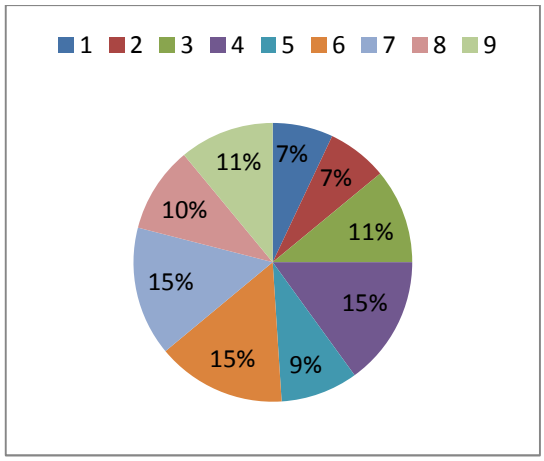


Diagrama de sectors



Mesures de centralització

Moda $M_o = 4, 6\ i\ 7$ (trimodal) **Mediana** $M_e = 6$ **Mitjana aritmètica** $\bar{x} = 5'33$

Mesures de dispersió

Recorregut $R = 9 - 1 = 8$ **Desviació mitjana** $DM = \frac{\sum f_i \cdot |x_i - \bar{x}|}{N} = \frac{204'34}{100} = 2'04$

Variancia $\sigma_x{}^2 = V = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum f_i \cdot x_i{}^2}{N} - \bar{x}^2 = \frac{3405}{100} - (5'33)^2 = 34'05 - 28'4089 = 5'6411$

Desviació típica $\sigma_x = +\sqrt{\sigma_x{}^2} = 2'375$ **Coeficient de variació** $CV = \frac{\sigma_x}{\bar{x}} = \frac{2'375}{5'33} = 0'4456$

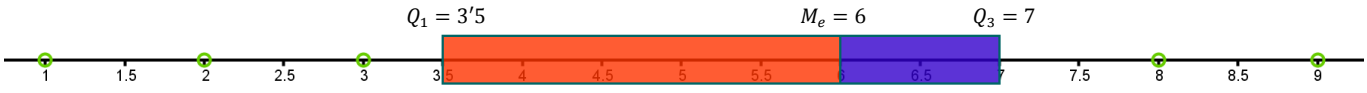
Mesures de posició

Quartils $Q_1 = \frac{3+4}{2} = 3'5$ $Q_2 = M_e = 6$ $Q_3 = 7$

Percentils $P_{10} = 2$ $P_{30} = 4$ $P_{40} = \frac{4+5}{2} = 4'5$ $P_{70} = 7$ $P_{90} = 9$

Recorregut interquartílic $Q_3 - Q_1 = 7 - 3'5 = 3'5$

Diagrames de caixa



Exemple de variable estadística contínua

L'altura, expressada en cm, de 50 alumnes de l'IES és:

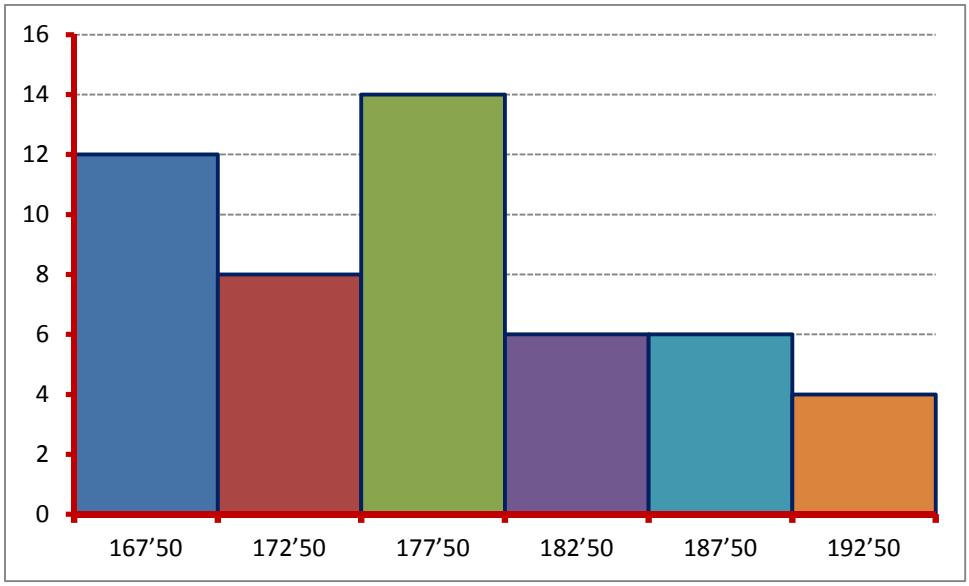
167-172-170-169-168-183-194-191-186-175-165-166-176-182-184-166-171-173-178-185-183-165-193-189-178

176-167-185-186-165-168-175-177-178-178-180-170-169-172-182-173-189-194-171-169-176-178-179-175-176

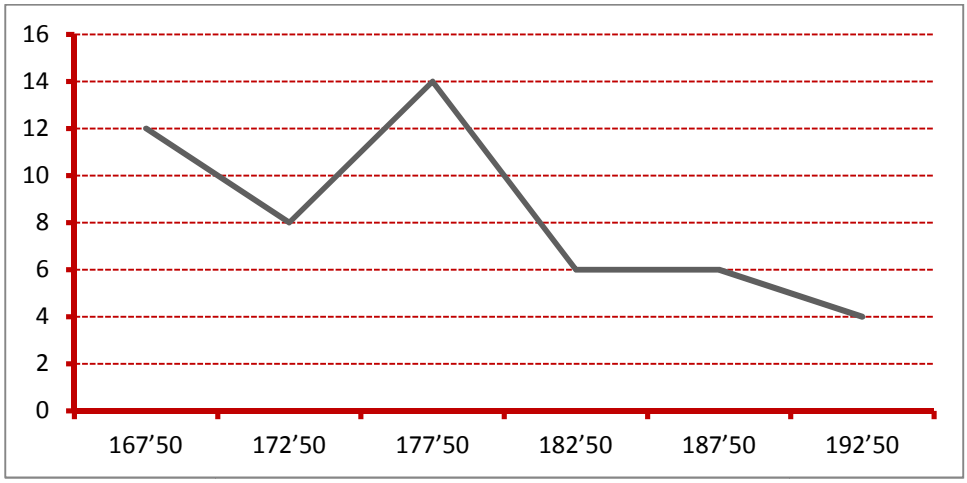
L'altura està entre 165 cm i 195 cm. Considerarem les dades agrupades en 6 intervals de longitut 5 cm.

Classe	x_i	f_i	F_i	$n_i = \frac{f_i}{N}$	N_i	$x_i \cdot f_i$	$\lvert x_i - \bar{x}\rvert \cdot f_i$	$x_i^2 \cdot f_i$
[165,170[	167'50	12	12	24%	24%	2010	117'60	336675'00
[170,175[	172'50	8	20	16%	40%	1380	38'40	238050'00
[175,180[	177'50	14	33	28%	68%	2485	2'80	441087'50
[180,185[	182'50	6	40	12%	80%	1095	31'20	199837'50
[185,190[	187'50	6	46	12%	92%	1125	61'20	210937'50
[190,195[	192'50	4	50	8%	100%	770	60'80	148225'00
		N=50		100%		8865	312'00	1574812'50

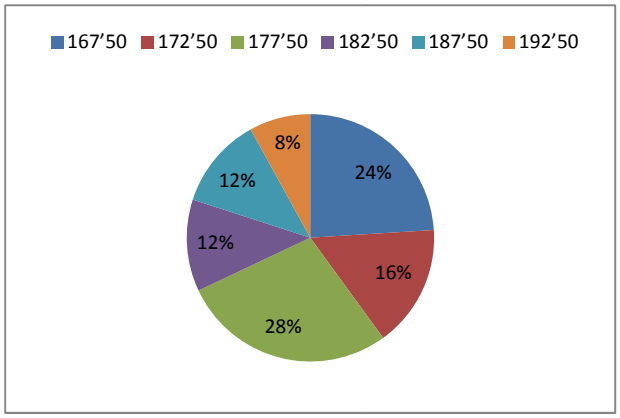
Histograma de freqüències

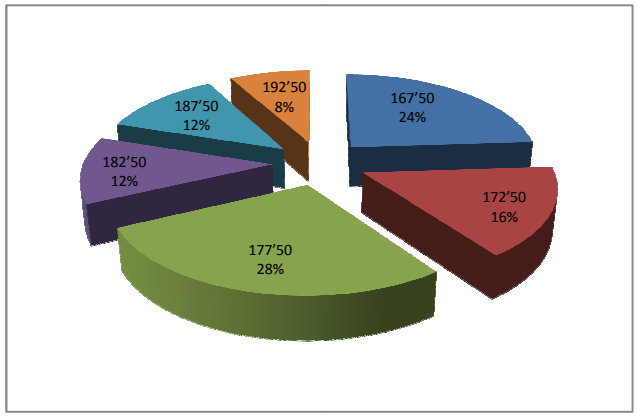


Polígon de freqüències



Digrama de sectors





Mesures de centralització

Classe Modal $[175,180[$ **Mediana** $M_e = 177'50$ **Mitjana aritmètica** $\bar{x} = 177'30$

Moda $M_o = 177'50$

Mesures de dispersió

Recorregut $R = 195 - 165 = 30$ **Desviació mitjana** $DM = \frac{\sum f_i \cdot |x_i - \bar{x}|}{N} = \frac{312'00}{50} = 6'24$

Variancia $\sigma_x{}^2 = V = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum f_i \cdot x_i{}^2}{N} - \bar{x}^2 = \frac{1574812'50}{50} - (177'30)^2 = 60'96$

Desviació típica $\sigma_x = +\sqrt{\sigma_x{}^2} = 7'81$ **Coeficient de variació** $CV = \frac{\sigma_x}{\bar{x}} = \frac{7'81}{177'30} = 0'044$

Mesures de posició

Quartils $Q_1 = 172'50$ $Q_2 = M_e = 177'50$ $Q_3 = 182'50$

Recorregut interquartílic $Q_3 - Q_1 = 182'50 - 172'50 = 10$

$\sum$ *sumatori*; $\cap$ *intersecció*; $\cup$ *unió*; $\emptyset$ *conjunt buit*; $\in$ *pertany*; $\forall$ *per a tot*; $\rightarrow$ *aleshores* $\subseteq$ *inclós en*; $\sqcap$ *Producte*; $/$ *de forma que*; $\leftrightarrow$ *si i sols si*; $\exists$ *existeix*; $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$; *Estadística*