L'ESTADÍSTICA

1.DEFINICIONS

L'**estadística** és una ciència que recopila, analitza , interpreta i representa dades relatives a fenòmens de la realitat. S'utilitza quan cal tenir dades que ens puguin ajudar a extreure conclusions sobre el fenomen que volem estudiar. Qualsevol fenomen pot ser expressat mitjançant dades i per tant l'estadística ens ajuda a analitzar-les I ens dóna resposta a preguntes com ara:

- Quin és el percentatge de dones a Catalunya?
- Quin va ser el programa de TV més vist en una certa franja horària?
- Quin percentatge d'aturats hi ha a Espanya?
- Quina ha estat l'evolució dels preus dels pisos a Barcelona els darrers 10 anys?

És a dir l'estadística està present a la nostra vida quotidiana i tenir unes nocions d'aquesta ciència ens ajudarà a poder fer una interpretació crítica de molts fenòmens que ens envolten, que llegim a la premsa i que ens preocupen .

Comencem per definir els primers mots usats en estadística.

La **població** és el conjunt de tots els individus o elements que tenen unes característiques comunes i sobre les que volem fer el nostre estudi.

En moltes ocasions no és possible analitzar a tots els individus de la població per fer-ne l'estudi degut a què s'empraria massa temps i el cost seria massa alt. En aquests casos s'escull una part de la població , seleccionant-la de forma aleatòria , per tal de fer l'anàlisi. Aquesta part de la població es coneix com a **mostra** i aquesta ha de ser representativa de la població , és a dir ha de mantenir les característiques de la població d'estudi.

Li diem **variable** a la característica que volem estudiar i cada una de les dades que recollim sobre la variable és una **observació**.

Per exemple:si volem fer un estudi sobre el salari mensual de tots els treballadors de Catalunya i per això es demana a 1000 individus que treballen a Catalunya quin és el seu sou. La població està formada per totes les persones que treballen a Catalunya. La mostra la formarien els 1000 individus enquestats , la variable d'estudi és el salari que cobren i les observacions serien els salaris concrets de cada individu preguntat: 1200€ , 900€, 1150€, ….

2. TIPUS DE DADES

Les dades recollides poden ser de molts tipus. Si les classifiquem segons si les podem o no comptar aquestes poden ser:

2.1.Dades quantitatives: si fan referencia a observacions de variables que poden mesurar-se i que poden representar-se de forma numèrica. Dins d'aquest tipus encara podem distingir:

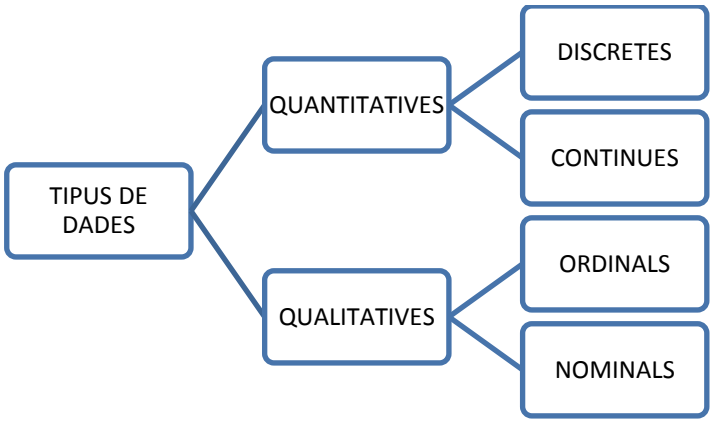
2.1.1 Dades quantitatives discretes: prenen un nombre finit de valors generalment enters. Per exemple: si estudiem el nombre de germans que tenen els alumnes de GES les respostes podran ser 0 , 1, 2, 3, ….però ningú respondrà tinc 2.3 germans o 1.25 germans; aquestes dades doncs diríem que són de tipus quantitatiu discret.

2.1.2 Dades quantitatives contínues: poden prendre qualsevol valor dins d'un interval [a,b]. Per exemple : si estudiem l'alçada dels nens de 12 anys , i la nostra precisió a l'hora de mesurar fos prou bona entre 150 i 151 cm podríem trobar qualsevol valor entremig 150.2 ; 150.5; 150.7; ….

2.2.Dades qualitatives: responen a variables que només poden prendre un nombre fint de valors i que no són numèriques, sinó atributs. Dins d'aquestes encara podríem destacar:

2.2.1 Dades qualitatives ordinals: són aquelles que malgrat no ser numèriques podrien ser ordenades. Per exemple: les notes(E,N, B, S, I); la satisfacció per alguna cosa (Molt, Força, Poc, Gens….)

2.2.2 Dades qualitatives nominals: responen a atributs que no poden ser ordenats d'una manera natural. Per exemple: la nacionalitat de les persones que arriben a l'aeroport del Prat (espanyola, francesa, anglesa, …); el color del jersei més venut d'un cert model (vermell, negre, blau, …).



3. LES TAULES. DADES EN NET.

Un cop recollides les dades, per tal de poder-ne fer un anàlisi i treure'n conseqüències, caldrà tenir-les ordenades de manera que sigui fàcil estudiar-les. Pensem que en general en un estudi estadístic es recullen moltes dades , centenars o potser milers i si no les tenim ordenades seria un caos. Això és el que es coneix com posar les "dades en net".

La manera d'ordenar aquestes dades és en una taula. Aquestes taules tenen diverses columnes , entre parèntesi posarem la notació abreujada. La mida de la mostra és a dir el nombre total d'observacions recollides l'escriurem N.

- **Valor de les dades (x_i)**: es posen les observacions recollides , sense repetir. Si les dades són quantitatives s'ordenen de petita a gran.
- **Freqüència absoluta (n_i)**: és el nombre de vegades que apareix la dada x_i, entre les observacions. La suma de totes les freq. absolutes és la mida de la mostra N.($\sum_{i=1}^{k} ni = N$)
- **Freqüència absoluta acumulada (N_i):** ens diu quantes dades hi ha inferiors o iguals a la x_i. S'obtindrà sumant (endarrera) la columna de les n_i. L'última N_i serà N.
- **Freqüència relativa (f_i):** es calcula dividint la freqüència absoluta n_i entre el nombre total de dades. Totes les f_i són positives i inferiors a 1. La suma de tota la columna dóna 1.
- **Freqüència relativa acumulada(F_i):** acumula les freqüències relatives . La darrera F_i sempre és 1.

Exemples.

Es vol fer un estudi sobre el nombre de germans que tenen els estudiants d'ESO. Per fer-ho s'han recollit unes dades i han estat ordenades a la següent taula.

x_i	n_i	N_i	f_i	F_i
0	8	8	0.16	0.16
1	21	29	0.42	0.58
2	13	42	0.26	0.84
3	5	47	0.1	0.94
4	3	50	0.06	1
SUMA	50		1	

Analitzem cada columna.

Fixem-nos que a la **primera columna** tenim els **valors de les observacions** , és a dir que de les persones enquestades algunes han respost que no tenen cap germà, altres 1, altres 2, 3 i 4.

Què ens diu la **segona columna**, la columna de les **freqüències absolutes**? Doncs que 8 dels preguntats tenen 0 germans, que 21 individus dels enquestats han respost que tenen 1 germà,que 13 tenen 2 germans….Observem que si sumem aquesta segona columna , dóna 50. Aquest 50 és la N, és la mida de la mostra, ens diu que en total tenim 50 dades.

La **tercera columna** ens dóna l'acumulació de les freqüències absolutes. El primer nombre és un 8 això vol dir que 8 individus tenen 0 germans o menys. El 29 que obtenim de 8+21 ens indica que 29 individus tenen un germà o menys. El 42 l'obtenim sumant 8+21+13 ens indica quants dels enquestats tenen 2 germans o menys, i així seguiríem . Observem que l'últim valor és 50 , el nombre total de dades.

La **quarta columna** ens dóna les freqüències relatives. És a dir el 0.16 s'obté de dividir 8/50. El 0.42 s'obté de dividir 21/50. El 0.26 l'obtenim de dividir 13/50….Moltes vegades la freqüència relativa es dóna en tant per cent, senzillament multiplicant per 100 aquests valors. Així donaria 16% no tenen germans, el 42 % tenen un germà, el 26% tenen dos germans….

La **cinquena columna** acumula l'anterior. El primer és 0.16, el segon 0.16+0.42=0.58; el tercer surt de sumar 0.26 més, …. En tant per cent voldria dir que el 16% tenen 0 germans o menys, el 58% tenen un germà o menys , el 84% tenen 2 o menys germans…

Val a dir que les taules no són sempre iguals. A vegades poden canviar l'ordre de les columnes, se'n poden afegir d'altres, l'important és que llegim a la capçalera de la taula quina informació ens dóna cada columna i ho sapiguem interpretar.

També molt a sovint amb dades quantitatives continues les dades les agrupem amb intervals. En aquest cas la primera columna en lloc de contenir el valor de les dades discretes són els intervals. La resta de columnes acostumen a ser les mateixes que per les dades discretes, és a dir les freqüències absolutes i relatives i si s'escau les freqüències acumulades. També sol aparèixer una columna amb la **marca de classe (x_i) dels intervals** que són els punts mitjos dels intervals. Per calcular la marca de classe d'un interval senzillament hem de sumar els dos extrems i dividir-ho entre 2. Així la marca de classe de l'interval [a, b] es calcularia $\frac{a+b}{2}$. Per posar un exemple numèric, la marca de classe de l'interval [12, 15] seria $\frac{12+15}{2} = 13.5$.

En general els intervals es posen tancats al límit inferior i oberts al límit superior d'aquesta manera: [a, b[o [a, b) qualsevol d'aquestes notacions ens indica que acceptem valor més grans o iguals a a però més petits que b.

És a dir , si considerem l'interval [12, 15[dins aquest interval està inclòs el valor 12, el 12.5; el 13; el 13.6; el 14; el 14.9 , ……, però el 15 ja no.

Veiem l'exemple d'una **taula amb dades agrupades en intervals** per tal d'analitzar-ne cada element.

Aquesta taula recull les 50 observacions referides a l'alçada en cm de 50 nens de 13 anys.

Intervals	**x_i=marca de classe**	**n_i**	**N_i**	**f_i en %**	**F_i en %**
[140-150[	145	10	10	20%	20%
[150-160[	155	21	31	42%	62%
[160-170[	165	12	43	24%	86%
[170-180[	175	5	48	10%	96%
[180-190[	185	2	50	4%	100%

Comencem per analitzar la **variable** estudiada : és **l'alçada dels nens de 13 anys**. Es tracta doncs d'una variable de tipus **quantitativa continua**. Per fer l'estudi s'ha agafat una **mostra** de **50 nens** d'aquesta edat.

Observem ara detingudament cada columna de la taula.

A la primera columna tenim els **intervals**. Observem que els intervals estan oberts per l'extrem superior. Així doncs una alçada de 150 cm exactament estaria considerada en el segon interval i no en el primer.

A la segona columna les **marques de classe** , és a dir els punts mitjos dels intervals , observem: $\frac{140+150}{2} = 145; \frac{150+160}{2} = 155; \frac{160+170}{2} = 165;$ etc…

A la tercera columna hi tenim les **freqüències absolutes**. Cal destacar que en el cas de les dades agrupades perdem precisió, és a dir sabem que 10 de les alçades estan entre 140 i 150 cm , però no sabem quins són els valors exactes, no sabem si són 141 o 145 o 148. De la mateixa manera sabem que 21 de les alçades estan entre 150 i 160 cm, etc. Observem que la suma de tota aquesta columna és 50, és a dir la mida de la mostra.

A la columna de les **freqüències absolutes acumulades** anem sumant les freqüències anteriors. L' última N_i és 50 , és a dir el nombre total de dades.

La cinquena columna recull les **freqüències relatives (en %)** , per calcular-les es divideix la freq absoluta entre el nombre total de dades i en aquest cas com les tenim

en % , aquesta divisió la multipliquem per 100. Així la 1ª freq relativa surt de fer $\frac{10}{50} \cdot 100 = 20\%$. La següent surt de fer el càlcul fer $\frac{21}{50} \cdot 100 = 42\%$, etc. Això s'interpretaria dient que el 20% dels nois de la mostra tenen una alçada entre 140 i 150 cm; el 42% tenen una alçada entre 150 i 160 cm. Fixem-nos que la suma de totes les freqüències relatives sumen 100 si estan expressades en tant per cent o 1 quant estan expressades en tant per 1.

Per acabar la columna de les **freqüències relatives acumulades** en % es calcula sumant les freqüències relatives sense acumular. Així 62= 20+42; 86= 20+42+24;..i la darrera és 100% doncs ja les tenim totes acumulades. Aquesta columna s'interpretaria dient : el 20% dels nens de la mostra tenen una alçada de menys de 150cm , el 42% dels nens de la mostra tenen una alçada inferior a 160 cm , etc.