

CAL RECORDAR-1 (Bloc 2)

ESTADÍSTICA

INTRODUCCIÓ

Per tots es conegut que la paraula "estadística" prové del mot "estat", atès que va ser la funció principal dels Governes dels Estats per establir registres de població, naixements, morts, collites, tributs...

És molt difícil poder precisar sobre la data històrica exacta dels orígens de la estadística. A Internet es pot trobar molta informació i diverses opinions sobre aquest fet històric.

Avui dia, podem afirmar sense cap mena d'error, que l'estadística té molta importància a la nostra societat. Fa relativament pocs anys, ningú s'hagués atrevit a dir-ho. L'estadística es considerada com l'eina més utilitzada per mesurar, analitzar i prendre decisions sobre nombroses disciplines (medicina, economia, sociologia, política...) i fets que afecten al món actual.

Malgrat tot, moltes persones encara pensen que la estadística es redueix a un conjunt de dades recollides en taules i gràfics publicats als diaris.

En un futur proper, es preveu un gran desenvolupament de l'estadística degut a les noves teories estadístiques aplicades com poden ser, teoria de les decisions, teoria de l'estimació, teoria de jocs, teoria de cues, anàlisi multivariant... i amb l'ajut de la tècnica i els moderns ordinadors.

L'estudi de l'estadística es divideix en dues parts:

- **Estadística descriptiva** es limita a recollir, ordenar i classificar les dades obtingudes corresponents a la variable estadística objecte d'estudi i mitjançant uns paràmetres definits adequadament, realitzar deduccions directament d'aquestes observacions.
- **Estadística inferencial** és la part de l'estadística que treballa amb mostres i a partir d'aquestes mostres, intenta deduir (inferir) característiques de tota la població. S'utilitza, quan no és possible observar la característica objecte d'estudi en tots els individus de la població (bé sigui, per motius econòmics, per manca de temps o per qualsevol altra causa). En aquests casos s'escull una part de la població anomenada mostra. Perquè els resultats siguin fiables, cal que la mostra sigui representativa. Hi ha diferents mètodes per fer-ho o tècniques de mostreig i s'ha de saber triar en cada cas, la més adient.

CONCEPTES BÀSICS

Per fer qualsevol estudi estadístic, cal conèixer els següents conceptes:

- **Població**: conjunt dels elements que són objecte d'estudi.
- **Individu**: cadascun dels elements de la població.
- **Variable estadística**: és la característica o propietat de la població que és objecte del nostre estudi. Aquesta pot ser,
 - **Qualitativa**: expressa una qualitat, per tant, no es pot mesurar i no pot prendre valors numèrics, com per exemple la professió, el sexe, el color dels ulls...
 - **Quantitativa**: és quan la variable estadística (característica d'estudi) es pot comptar o mesurar, com per exemple, edat dels alumnes d'una escola, el seu pes, la seva talla, les notes de matemàtiques obtingudes al final d'un semestre en un determinat grup...

També es poden classificar en:

- **Discretes**: quan únicament la variable estadística pot prendre un nombre finit de valors (nombre de fills de 100 famílies, nombre de diaris venuts a Barcelona cada dia durant un més...)
- **Contínues**: quan la variable estadística pot prendre, al menys teòricament, tots els valors possibles dins d'un interval (el pes, la talla ...)

CONCEPTE DE FREQUÈNCIA

Per fer el recompte i organització de les dades, s'utilitzen les taules de freqüències i els principals conceptes relacionats amb elles són:

- **Freqüència absoluta**: és el nombre d'individus que presenta cada valor de la variable. Es representa per n_i
- **Freqüència relativa**: és el quocient entre la freqüència absoluta i el nombre total d'individus que hi ha a la mostra. Es representa per f_i

$$f_i = \frac{n_i}{n}$$

on n és el nombre total d'individus.

- **Percentatge**: és el tant per cent d'individus que presenta cada valor de la variable. Es representa per p_i i es calcula multiplicant per 100 la freqüència relativa.

$$p_i = f_i \cdot 100$$

- **Freqüència absoluta acumulada** del valor x_i : és la suma de les freqüències absolutes, un cop hem ordenat del més petit al més gran els valors de la variable, fins a la que ocupa la posició i . Es representa per N_i .

$$N_i = \sum_{j=1}^i n_j$$

- **Freqüència relativa acumulada** del valor x_i : és la suma de les freqüències relatives, un cop hem ordenat els valors, fins a la que ocupa la posició i . Es representa per F_i .

$$F_i = \sum_{j=1}^i f_j$$

- **Percentatge acumulat** del valor x_i : és la suma dels percentatges, un cop hem ordenat els valors, fins al que ocupa la posició i . Es representa per P_i .

$$P_i = \sum_{j=1}^i p_j$$

Cal saber que no tots els autors utilitzen els mateixos símbols (lletres) per representar les freqüències. Com per exemple:

Llibre	n_i	f_i	p_i	N_i	F_i	P_i
WEB	f	fr	%	F	Fr	%a

Per tant, és molt important, tenir clar els diferents conceptes de freqüències i saber-les d'utilitzar independentment del símbol que les representa.

GRÀFICS ESTADÍSTICS

S'utilitzen per representar la informació continguda a les taules estadístiques per fer-la més clara i evident.

Segons sigui la naturalesa de la variable estadística d'estudi, hi ha de diversos tipus.

Si les variables són qualitatives o quantitatives discretes, el més utilitzats són:

- [Diagrama de barres](#)
- [Pictogrames](#)
- [Diagrama de sectors](#)

Si és tracta de representar variables quantitatives contínues, el més emprats són:

- [Histogrames](#)
- [Polígons de freqüències](#)

En estudis demogràfics i socials, s'utilitzen les [piràmides de població](#).

Per representar dades referents a una regió, s'utilitzen els [cartogrames](#).

PARÀMETRES ESTADÍSTICS

El principal objecte dels paràmetres estadístics és permetre resumir i sintetitzar en uns pocs valors (els paràmetres pròpiament dits) el gran número de dades de la població d'estudi, proporcionant una descripció el més completa i aproximada possible de tota la població.

Tot procés de síntesi comporta una pèrdua d'informació, però és molt millor treballar amb uns pocs valors (paràmetres) amb una significació molt precisa i acurada que amb la totalitat de les dades.

En estadística descriptiva s'utilitzen dos tipus de paràmetres:

§ PARÀMETRES DE CENTRALITZACIÓ

Paràmetres de centralització: Són valors que tendeixen a situar-se cap al mig del conjunt de dades ordenades, al voltant dels quals es distribueixen els altres. Els més utilitzats són: [la mitjana aritmètica \(mitjana\)](#), [la mediana](#) i [la moda](#).

-La **mitjana aritmètica**, o simplement **mitjana**, es calcula sumant tots els valors i dividint pel número total de dades. Es representa per \bar{x}

Es calcula mitjançant l'expressió:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{N}$$

Propietats de la mitjana:

1. Si sumem una constant a tots els valors d'una variable, la seva mitjana també augmenta en aquesta constant.
2. Si multipliquem tots els valors de la variable per una constant, la mitjana també queda multiplicada per aquesta constant.

-La **moda** és el valor de la variable que té una freqüència absoluta més gran. Pot ser que hi hagi més d'una moda i es parla aleshores de distribució bimodal, trimodal...segons siguin 2, 3..., els valors de la variable que assoleixin aquesta freqüència més gran.

-La **mediana** és el valor central de la distribució, un cop hem ordenat els valors de la variable en ordre creixent. Si el nombre de valors és parell, llavors és la mitjana dels dos valors centrals.

§ PARÀMETRES DE DISPERSIÓ

Paràmetres de dispersió: per fer un estudi complet d'una variable estadística, no solament cal saber les mesures de centralització, també s'ha de conèixer si les dades estan agrupades o no al voltant dels valors centrals.

Això s'anomena *dispersió* i als paràmetres que mesuren aquestes desviacions s'anomenen *paràmetres de dispersió*.

Els més utilitzats són: el **recorregut** o **rang**, la **variància** i la **desviació estàndard** (típica o tipus).

- El **recorregut** o **rang** és la mesura de dispersió que s'obté restant el valor màxim i el mínim de la variable estadística.

-La **variància** és la mitjana dels quadrats de les desviacions respecte de la mitjana. Es representa per s^2 i també per σ^2 . Les expressions per calcular-la són:

$$s^2 = \sigma^2 = \frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N} = \frac{(x_1 - \bar{x})^2 \cdot n_1 + \dots + (x_n - \bar{x})^2 \cdot n_n}{n_1 + \dots + n_n}$$

Però, per fer càlculs, és millor utilitzar la següent fórmula alternativa:

$$s^2 = \sigma^2 = \frac{\sum_i x_i^2 \cdot n_i}{N} - \bar{x}^2 = \frac{x_1^2 \cdot n_1 + \dots + x_n^2 \cdot n_n}{n_1 + \dots + n_n} - \bar{x}^2$$

La variància sempre és positiva, ja que és una suma de quadrats, i un nombre qualsevol elevat al quadrat sempre és positiu.

Té l'inconvenient que no ve expressada en les mateixes unitats que les dades, donat que les desviacions estan elevades al quadrat.

- La **desviació estàndard (típica o tipus)** és l'arrel quadrada positiva de la variància. El seu símbol és s i també σ

$$s = \sigma = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 \cdot n_i}{N}} = \sqrt{\frac{(x_1 - \bar{x})^2 \cdot n_1 + \dots + (x_n - \bar{x})^2 \cdot n_n}{n_1 + \dots + n_n}}$$

$$s = \sigma = \sqrt{\frac{\sum_i x_i^2 \cdot n_i}{N} - \bar{x}^2} = \sqrt{\frac{x_1^2 \cdot n_1 + \dots + x_n^2 \cdot n_n}{n_1 + \dots + n_n} - \bar{x}^2}$$

Atès que la variància és positiva, sempre podem calcular-ne l'arrel quadrada.

Està expressada en les mateixes unitats que les dades.

Propietats de la variància i de la desviació típica:

1. Si a tots els valors d'una variable els sumem una constant, la variància i la desviació típica no es modifiquen.
2. Si tots els valors d'una variable els multipliquem per una constant positiva, la variància queda multiplicada pel quadrat de la constant i la desviació típica queda multiplicada per aquesta constant.