

CFGS DOCUMENTACIÓ I ADMINISTRACIÓ SANITÀRIES

MP 9. Validació i explotació de dades

UF1. Estadística i epidemiologia

NF 1. Estadística

ESTADÍSTICA UNIVARIANT

 Institut Xifra Girona	CFGS DOCUMENTACIÓ I ADMINISTRACIÓ SANITÀRIES	MP9: VALIDACIÓ i EXPLOTACIÓ DE BASES DE DADES
Data de realització: 18/11/2020 Realitzat per: Margarita Pérez		UF1 NF1. Estadística univariant

ESTADÍSTICA UNIVARIANT

OBJECTIUS

- Adquirir amb claredat els conceptes de població, mostra, caràcter estadístic i variables estadístiques.
- Saber resumir una sèrie de dades mitjançant una taula de freqüències.
- Representar de manera gràfica una sèrie de dades estadístiques.

1. INTRODUCCIÓ A L'ESTADÍSTICA

L'estadística és la ciència que ens proporciona una metodologia per a recollir, organitzar, resumir i analitzar dades, i fer estimacions a partir d'aquestes.

L'estadística es divideix en dues grans branques:

- **Estadística descriptiva:** s'encarrega de la recollida i descripció de les característiques bàsiques d'un conjunt d'observacions amb la finalitat de descobrir les regularitats o característiques existents en els seus elements. L'estudi es realitza sobre un grup donat sense treure conclusions sobre un grup més gran.
- **Inferència estadística o estadística inductiva:** s'encarrega d'arribar a una conclusió per als elements d'una població a partir de les dades obtingudes a partir d'una mostra. La mostra sobre la que es fa l'estudi és representativa de la població i per tant podem treure conclusions la població de la que procedeix i generar models matemàtics predictius.

2. CONCEPTES BÀSICS

POBLACIÓ: conjunt de tots els elements que compleixen una determinada característica. Els elements de la població es diuen individus o unitats estadístiques.

MOSTRA: qualsevol subconjunt de la població.

CARÀCTERS ESTADÍSTICS: cadascuna de les característiques que pot presentar un element de la població.

MODALITAT: són les diferents situacions possibles d'un caràcter. Les modalitats d'un caràcter han de ser incompatibles o excloents, de forma que cada element de la població presenti només una de les modalitats dels caràcter (alt o baix, prim o gras, sa o malalt)

Es diferencien dos tipus de caràcters estadístics:

- **Variables estadístiques** o **caràcters quantitatius**: són aquells que mesuren una quantitat (pes, edat, alçada ...).

Les variables estadístiques poden ser **discretes** o **contínues**:

- o **Variables estadístiques discretes**: una variable és discreta quan pot prendre un nombre finit de valors o infinit numerable (nombre de famílies, habitants d'un municipi, malats de fibromiàlgia, nombre de llits d'hospital..).
 - o **Variables estadístiques contínues**: una variable és contínua quan pot prendre tots els valors possibles dintre d'un cert **interval** de la recta real.
-
- **Atributs** o **caràcters qualitius**: són aquells que no es poden mesurar (professió, estat civil, nivell d'estudis ...); a vegades, si interessa, els atributs es poden transformar en caràcters del tipus:
 - o **Atributs nominals**: utilitzem nombres per designar categories (exemple: home = 0, dona = 1); l'ordre i la magnitud dels valors és irrellevant; s'utilitza per poder treballar les dades de forma automatitzada (sexe en el CMBD); grups sanguinis (A / B / AB / 0).
 - o **Atributs ordinals**: les observacions s'anomenen v. ordinals si l'ordre dins les categories és important (exemple: a oncologia s'utilitzen per designar l'estat dels pacients sotmesos a proves clíniques: 1 = dany fatal, 2 = dany sever, 3 = dany moderat, 4 = dany menor).

Exercicis

1. Què és l'estadística descriptiva? I la inferencial? Posa un exemple de cada tipus d'estadística i explica perquè has triat cadascuna.
2. Classifica els següents caràcters en **atributs** (nominals o ordinals) o **variables** (discreta o contínua): pacients donats d'alta al servei d'urgències, sexe, professió, intervencions quirúrgiques urgents, estat civil, dies d'estada, nivell de colesterol en sang, consum de cigarretes en majors de 18 anys, temperatura corporal.
3. Indica les diferents modalitats del caràcter "intenció de vot a les eleccions autonòmiques a Catalunya" i "pronòstic d'una malaltia".

3. FORMES D'OBSERVAR UNA POBLACIÓ

OBSERVACIÓ EXHAUSTIVA: es mesuren tots els elements de la població investigada.

OBSERVACIÓ PARCIAL: per limitacions de temps o per qüestions econòmiques no és possible recollir totes les dades de la població:

Subpoblació: els elements que la formen reuneixen unes característiques especials que no són generals a la població.

Mostra: és un subconjunt d'elements escollits a l'atzar a partir dels quals es dedueixen les característiques de la població.

OBSERVACIÓ MIXTA: es realitza observant exhaustivament aquells elements que són bàsics per la investigació i els restants, a partir d'una mostra.

CENS: s'utilitza per les investigacions estadístiques amb observació exhaustiva (cens electoral).

ENQUESTA: s'utilitza per les investigacions estadístiques realitzades amb observació parcial o mixta a partir de mostres i/o subpoblacions.

Exercicis

5. Indica el **tipus d'observació** que realitzaries en els següents estudis i raona la resposta:

- a) Nivell d'acceptació d'una normativa als treballadors de l'hospital de la Vall d'Hebron
- b) Grau de satisfacció dels usuaris del Sistema Nacional de Salut
- c) Nombre de naixements a Espanya l'any 2009
- d) Èxit del transplants realitzats durant el període 2000-2004
- e) Seguiment d'una vaga en el col·lectiu sanitari

6. Classifica els **caràcters estadístics** següents i raona la resposta:

- a) Nombre de pacients atesos al servei de traumatologia el mes de febrer
- b) Concentració de plom en una mostra d'aigua
- c) Nombre de suïcidis als EEUU en un any específic
- d) Pronòstic dels pacients afectats per una malaltia
- e) Quantitat d'avortaments previs d'una dona abans del seu primer fill
- f) Concentració de glucosa dels malalts diabètics abans de començar el tractament amb insulina

3.1 MOSTREIG

Quan es realitza un estudi estadístic és pràcticament impossible estudiar tots els individus que formen la població. Es necessiten molts recursos econòmics i temps per realitzar aquesta tasca. Afortunadament, es pot realitzar una selecció acurada d'individus d'una població, un mostreig i realitzar l'estudi únicament sobre ells. Si els individus seleccionats representen la seva població d'origen, els resultats que s'obtinguin es podran aplicar o inferir al grup d'origen.

Una altre avantatge de realitzar l'estudi sobre la mostra és la relacionada amb la validesa de l'estudi. Les mesures tendeixen a ser de major qualitat ja que es poden dedicar més esforços a estudiar un menor nombre d'individus. També, al poder invertir més recursos en l'estudi de cada individu, poden mesurar-se altres variables, el que no és possible fer en tota la població.

Ens els estudis estadístics s'utilitza la terminologia següent per referir-se al procés de selecció dels individus objecte de l'estudi:

Població: conjunt d'individus que tenen en comú una o varies característiques i que l'investigador tria per al seu estudi. Exemple: estudiants d'infermeria, pacients amb diabetis, les dones embarassades, ...

Mostra: conjunt d'individus que s'escollin de la població per ser estudiats. La mostra triada ha de ser representativa de la població de la que procedeix, és a dir, ha de compartir totes les característiques que defineixen la població.

Mida de la mostra: nombre d'individus que formen la mostra. La mida s'obté a partir de mètodes estadístics tenint en compte la mida de la població i la freqüència de la variable objecte d'estudi:

- Quant més homogènia sigui la mostra menor serà la mida de la mostra necessària i quant més heterogènia major serà la mida de la mostra requerida.
- Quant major sigui la freqüència de la característica menor serà la mida de la mostra. Pel contrari, si és una característica amb poca presència a la població necessitarem una mostra major per poder observar-la.

Mostreig: procés de selecció de la mostra d'estudi a partir de la població.

Unitat de mostreig: grup d'individus de la població de on es trien els participants a la mostra. Pot ser la població completa o grups específics de la població.

Quina informació ens proporciona la mida de la població i de la mostra?

Imaginem que volem realitzar l'estudi d'una determinada malaltia entre els 560 pacients atesos en un centre sanitari per la qual cosa hem triat una mostra de 28 pacients. Què significa triar a 28 de 560 pacients? Quina proporció de la població estem entrevistant? I a l'hora d'obtenir conclusions sobre la població, quants pacients de la població total representa cada un dels de la mostra?

Per calcular la proporció pacients que estem entrevistant, dividim la grandària de la mostra entre el de la població: $28/560 = 0,05$, el que vol dir que estem passant l'enquesta al 5% de la població.

Ara anem a calcular a quants individus representa cada un dels elements de la mostra. Fem el pas contrari, dividim el número d'individus de la població entre els de la mostra: $560/28 = 20$, el que voldria dir que cada un dels elements de la mostra representa a 20 pacients de l'hospital.

Els dos conceptes que acabem de veure tenen la següent definició formal:

1. **Factor de elevació:** es el quocient entre la mida de la població y la mida de la mostra, N/n .
Representa el número de elements que hi ha a la població por cada element de la mostra.
2. **Factor de mostreig:** es el quocient entre la mida de la mostra y la mida de la població, n/N .
Si es multiplica per 100, obtenim el percentatge de la població que representa la mostra.



3.1.1 TIPUS DE MOSTREIG

El mostreig és el pas decisiu en qualsevol estudi estadístic. Si els individus a estudiar no estan ben escollits, les conclusions obtingudes sobre ells no es podran aplicar a la població d'origen. La condició bàsica que ha de complir la mostra es que sigui **representativa de la població**, és a dir, que tots els individus de la mostra presentin les mateixes característiques que defineixen la població de procedència. Si no fos així, es produirien biaixos a l'estudi que afectaria la seva validesa externa.

Hi ha diferents mètodes de mostreig:

PROBABILÍSTIC	NO PROBABILÍSTIC
Aleatori simple	Intencional
Aleatori sistemàtic	Bola de neu
Aleatori estratificat	Discrecional
Per conglomerats	

MOSTREIG PROBABILÍSTIC

Tots els individus de la població tenen la mateixa probabilitat de ser escollits per pertànyer a la mostra. És el tipus més encertat, ja que ens assegura que la mostra és representativa de la població, al no intervenir cap criteri de selecció que pugui esbiaixar la mostra.

Els centres oficials com l'INE, utilitzen sempre mostrejos aleatoris.

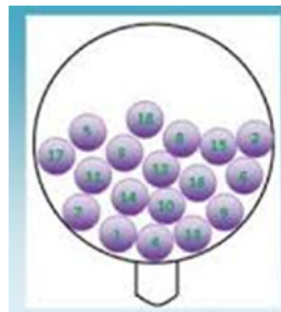
La mostra es pot triar de diferents maneres:



Mostreig aleatori simple

És el mètode que produeix mostres més representatives.

S'utilitza quan la mida de la població no és molt elevada i es disposa per avançat de d'un llistat sencer de tots els membres de la població. A cada individu de la llista se l'assigna un número i, a través d'un procediment a l'atzar, es tria el nombre d'individus necessaris (exemple: taula de nombres aleatoris).



Exemple

Imaginem que volem fer un estudi sobre el nivell de "burnout" en infermeres d'un determinat hospital. Necessitem un llistat de totes les infermeres que hi treballen i numerem consecutivament a totes les infermeres. (N=400)

El programa informàtic ens proporciona el nombre de subjectes en funció de la mida de la mostra ($n=100$). En concret ens proporciona els 100 nombres aleatoris, entre el 1 i el 400, que hem de seleccionar i triem com a mostra a les infermeres a les que els corresponen aquest números

Mostreig aleatori sistemàtic

Igual que al cas anterior es necessita tenir a tots els individus de la població en un llistat. Consisteix en seleccionar individus segons una regla o procés periòdic.

Primer es calcula una **constant de mostreig, K**, dividint el nombre total d'individus de la població, N, entre el nombre de persones que han de formar la mostra. A continuació es tria un nombre a l'atzar, **r**, compres entre 1 i K. A partir del llistat es tria el que ocupa la posició r, i a partir d'aquí es van comptant K individus i es van seleccionant els membres de la mostra.

És molt important que el llistat es faci sense cap criteri dirigit, per evitar l'esbiaix de la mostra.

En aquest exemple **K = 3 i r = 3**



Exercici

Volem conèixer el grau de satisfacció de les dones que acudeixen a les urgències ginecològiques d'un determinat hospital. Per portar-ho a terme, triarem una mostra de 100 dones sobre les 400 dones ateses l'any anterior. Imaginem que:

$$N / n = 400 / 100 = 4 = K \text{ i } r = 3$$

Escriu els 10 primers elements de la mostra

Mostreig aleatori estratificat

S'utilitza quan la característica objecte d'estudi no es distribueix de forma homogènia a la població, però existeixen grups o estrats en que sí es presenta d'aquesta manera. Aquest grups o estrats tenen alguna característica comú però són mútuament excloents.



Una vegada que coneixem la distribució de la variable per estrats, els individus de la mostra de cada estrat s'escolliran segons diferents criteris:

- Simple: seleccionem el mateix nombre de participants a cada estrat.
Inconvenient: la mostra no és proporcional a cada estrat
- Proporcional: la mida de cada una de les mostres és proporcional a la mida de cada estrat. En aquest cas, la proporció de subjectes a la mostra és similar a la població
- Òptim: la selecció de la mostra es fa en funció de la homogeneïtat o heterogeneïtat de la característica a la població.

Exercici

Volem conèixer el nivell de “burnout” o resposta a l'estrès, en una població de metges de l'hospital Vall d'Hebron, en el conjunt de professionals i en els grups diferenciats per sexes. Imaginem que $N = 1.000$ i volem aconseguir una mostra de $n = 100$. Si el nombre d'homes és de 700 i el de dones 300:

Com podríem triar la mostra mitjançant el mostreig estratificat, seguint el model simple i proporcional?

b. Mostreig aleatori per conglomerats

A diferència dels models anteriors, en què la unitat de mostreig era la població completa, en el mostreig per conglomerats els individus de la mostra es seleccionen a partir de grups preseleccionats de la població. Cadascun d'aquests grups es diu conglomerat (exemple: hospitals, col·legis, universitats ...).

Aquest sistema s'utilitza quan la població de referència és molt dispersa o quan no hi ha un llistat dels individus.

En primer lloc, l'investigador selecciona grups o conglomerats i de cada grup selecciona els subjectes individuals, ja sigui per mostreig aleatori simple o mostreig aleatori sistemàtic. L'investigador també pot optar per incloure a tot el conglomerat, no només a un subconjunt.

El conglomerat més utilitzat en la investigació és un conglomerat geogràfic. Per exemple, un investigador desitja estudiar el rendiment acadèmic dels estudiants secundaris a Espanya.

La diferència principal entre el mostreig per conglomerats i el mostreig estratificat és la inclusió de conglomerats o estrats.

En el mostreig aleatori estratificat, el mostreig es realitza en tots els estrats de la població, mentre que en el mostreig per conglomerats l'investigador només selecciona aleatòriament un nombre de conglomerats de la col·lecció de conglomerats de tota la població. Per tant, només es realitza una mostra sobre una quantitat de conglomerats. Els altres queden sense representació.

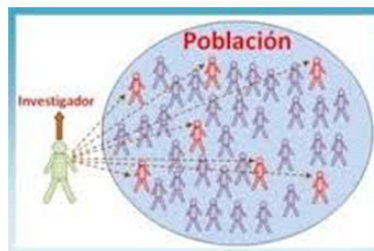


MOSTREIG NO PROBABILÍSTIC

En aquest tipus de mostreig, tots els individus no tenen la mateixa probabilitat de formar part de la mostra. És per això que és més difícil realitzar inferències estadístiques al ser la mostra no representativa de la població. S'utilitza quan es fan estudis exploratoris o quan no hi ha recursos suficients per realitzar un estudi probabilístic.

Mostreig intencional

S'escull als individus de la mostra a partir d'un grup de fàcil accés.



Mostreig en bola de neu

S'escull un nombre d'individus inicial per realitzar l'estudi i se'ls demana que busquin altres persones conegudes, i així successivament.

Es realitza sobre poblacions en les quals no es coneix als seus individus o és molt difícil accedir-hi.



Mostreig discrecional

L'investigador selecciona directament la mostra d'estudi a partir d'un grup d'individus que li poden ser d'utilitat, segons el seu propi criteri.

ACTIVITATS

1. Utilitzant una taula de nombres aleatoris, tria 15 elements d'una població numerada de l'1 al 89. Pots utilitzar aquest enllaç :
<https://ca.calcprofi.com/generador-de-nombres-aleatoris-en-linia.html>
2. Dignes de quina forma triaries una mostra de 50 alumnes del teu institut, per mostreig aleatori simple, sistemàtic i estratificat (cada estrat una classe, o un nivell).
3. Dels 500 directors de complexos turístics de les nostres Illes, 300 corresponen a complexos d'20 o menys habitacions, 150 a complexos d'entre 20 i 50 habitacions i finalment 50 corresponen a complexos de més de 50 habitacions
Si pretens fer una enquesta a una mostra de mida 50, com la prendries? Seria indiferent l'aspecte estadístic que haguessis de estudiar?
4. Un hospital disposa d'un llistat de pacients, organitzats per àrees d'atenció (neurologia, traumatologia,). Dignes quins tipus de mostreig podrien realitzar-se, i com els faries.
5. Per realitzar una enquesta sobre el consum d'un producte en una ciutat, es va prendre una mostra de forma que de cada barri es consultava a un nombre de persones proporcional a la superfície ocupada pel barri. Et sembla un mètode fiable ? Escriu un comentari.
6. Un majorista d'aliments, vol enviar mostres dels seus productes, a una mostra de supermercats. Tria de les 5 grans cadenes de supermercats, una mostra de cada, i mana els seus productes per posar-los a prova. Quin tipus de mostreig està utilitzant?

4. FASES DEL PROCÉS ESTADÍSTIC

1. FORMULACIÓ DE L'OBJECTIU: determinar el nivell d'albúmina en sang.
2. SELECCIÓ DE LES VARIABLES RELLEVANTS: concentració d'albúmina en sang
3. RECOLLIDA DE DADES, CLASSIFICACIÓ I TABULACIÓ
4. REPRESENTACIÓ GRÀFICA
5. SÍNTESI I ANÀLISI DE LA INFORMACIÓ
6. INTERPRETACIÓ DELS RESULTATS

4.1 RECOLLIDA DE DADES, CLASSIFICACIÓ I TABULACIÓ

La recollida de dades cal fer-la de la manera més acurada possible, ja que aquesta fase del procés estadístic influeix de manera decisiva a la resta d'etapes.

Una vegada obtingudes una sèrie de dades passem a ordenar-les i classificar-les: en el cas que el caràcter sigui qualitatiu ordenarem les dades per cadascuna de les modalitats del caràcter estudiat. Per contra, si el caràcter és quantitatiu llavors ordenarem les dades de menor a major.

La tabulació consisteix simplement en ordenar, en una taula de freqüències o taula estadística, les dades recollides per tal de tenir una visió global.

4.1.1. ELABORACIÓ DE TAULES ESTADÍSTIQUES.

Què hem de fer amb el conjunt de dades corresponents a un estudi estadístic?

Sempre és aconsellable procedir de manera sistemàtica i ordenada i seguir els següents passos:

- Ordenació de dades
- Agrupació de dades: si la variable és contínua o disposem de gran quantitat de dades (v. discreta) és aconsellable agrupar-les en classes, categories i intervals, procurant que totes elles tinguin una mateixa amplitud.
- Recompte de freqüències: cal comptar el nombre de vegades que es repeteix cada valor o qualitat de la variable.
- Elaborar la taula estadística:

x_i	n_i	n_{ia}	p_i	P_i	p_{ia}	P_{ia}

Modalitats (x_i): són cadascun dels valors que pot prendre una variable estadística.

Freqüència absoluta (n_i): nombre de vegades que es repeteix una modalitat.

Freqüència relativa d'un valor de la variable o atribut és el quocient entre la freqüència absoluta i el nombre total de dades (**N**). Es pot expressar en tant per 1 (**p_i**) o en tant per cent (**P_i**). En el primer cas, la suma de les diferents p_i serà igual a 1 i en el segon cas, la suma de les diferents P_i serà igual a 100.

Freqüència absoluta acumulada (n_{ia}) d'una certa modalitat és la suma de la freqüència absoluta d'aquesta modalitat i les anteriors a ella.

Freqüència relativa acumulada d'una certa modalitat és la suma de les freqüències relatives d'aquesta modalitat i les anteriors. Les representem per **p_{ia}** i **P_{ia}** .

Exercicis

7. Construeix la taula de freqüències a partir de les dades següents:

a) Les qualificacions obtingudes pels 25 alumnes d'una classe en una prova de C4 són:

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

b) Es va preguntar el grup sanguini a 42 persones que treballen a una empresa i es van obtenir les dades: (treballa les modalitats en l'ordre: A, B, AB i 0)

A, AB, A, 0, 0, 0, AB, A, A, B, B, A, 0,

A, AB, 0, B, AB, A, B, B, 0, 0, A, B, B,

A, AB, 0, 0, A, 0, 0, A, B, AB, AB, 0, B, AB, 0, A

8. S'ha fet un estudi sobre el nombre de parts en una població de 273 dones. Les dades agrupades són:

X_i	n_i
0	87
1	61
2	47
3	25
4	28
≥ 5	25

a) Completa la taula estadística

b) Respon :

Quantes dones han tingut un part?

Quin tant per cent de dones ha tingut 2 o menys parts?

Quantes dones han tingut 3 o més parts?

Quin tant per cent de dones ha tingut 4 parts?

Quantes dones han tingut menys de 3 parts?

9. Contesta les següents preguntes:

- Es pot fer un estudi estadístic sense triar una mostra? Posa'n un exemple.
- Quin valor ha de donar la suma de la columna de les ni? I de les pi? I de les Pi? I la nia de la última modalitat del caràcter estudiat? Per què?

4.1.2. DISTRIBUCIONS DE FREQÜÈNCIES AGRUPADES

- Correspon al cas en que el nombre de valors de la variable estadística és molt gran o es tracta d'una variable contínua.
- En aquest cas els valors s'agrupen en **intervals** anomenats **classes**.
- En aquestes distribucions la informació es fa més manejable però hi ha una pèrdua d'informació.

Regles per a formar distribucions de freqüències agrupades:

- Determinar el rang o recorregut de la distribució (**R**) : és la diferència entre el valor més gran i el més petit de la variable.

$$R = X_n - X_m$$

- Fixar el nombre d'intervals (**K**, és un nombre natural): es recomana formar entre 5 i 20 intervals per reduir la pèrdua d'informació. Per fer-ho calcularem l'arrel quadrada del nombre total de dades N:

$$K = \sqrt{N}$$

- Calcular l'amplitud dels intervals (**A**):

$$A = X_n - X_m / K$$

És convenient, triar sempre un nombre més gran perquè totes les dades estiguin incloses Exemple: si ens dona 2,8 utilitzarem l'amplitud 3.

- Determinar el límit inferior i el límit superior del primer interval (**L_i**, **L_s**).
- Esbrinar la marca de classe o punt mig de cada interval (**X_j**)

Exercicis

10. Un alumne anota els minuts que tarda en arribar a l'institut. Les dades són les següents:

9	8.3	8.9	9.2	8.4	8.7	8.5	9	8.5	8.6	8.2	9.2	9.3	9.1	8.7
8.9	8.3	9.4	8.7	8.4	8.5	8.8	8.7	8.5	8.6	8.1	8.5	8.8	8.7	8.6
8.5	8.2	9.8	9.2	9.6	8.9	9.5	8.7	8.6	8.5	8.8	8.6	8.3	8.3	8.9
8.2	8.5	9	8.8	9.1	8.6	8.6	8.5	8.7	8.4	9	9.3	8.7		

Construeix una taula amb les freqüències absolutes i relatives i els percentatges, agrupant les dades en 6 intervals.

4.2 REPRESENTACIÓ GRÀFICA

La informació recollida en un estudi estadístic se sol representar mitjançant gràfics per facilitar-ne una lectura global i ràpida.

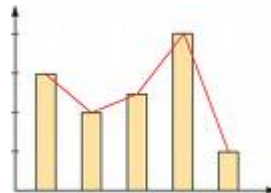
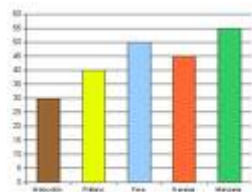
Es poden construir molts gràfics, però si volem ser objectius, hem de tenir en compte que, en cada cas, sempre n'hi ha un més apropiat que l'altre. No hem d'escollir el gràfic que ressalti els nostres interessos o el que doni més rellevància a allò que volem destacar, sinó el que sembli més apropiat segons el tipus de dades.

Els gràfics que s'utilitzen més sovint són:

Diagrames de barres: idoni per representar gràficament les dades d'una variable quantitativa discreta, quan la variable té pocs valors diferents, o per un caràcter qualitatiu. Les modalitats de la variable es representen a l'eix X i s'aixeca una barra d'altura igual a la freqüència de cada modalitat en l'eix Y. Si els valors de la variable els situem a l'eix horitzontal, dibuixarem línies o barres verticals; si els situem a l'eix vertical, les línies o barres seran horitzontals.

Aquestes barres han de tenir la mateixa amplada per a totes les dades i estar separades de forma que no es percebi continuïtat; si no fos així, ens donarien una visió errònia del conjunt.

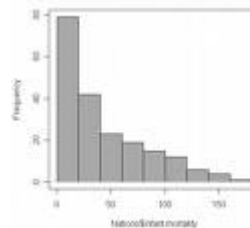
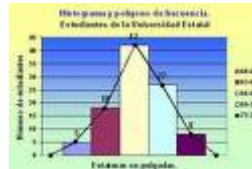
Podem treballar amb freqüències absolutes o relatives, acumulades o no. En cada cas s'obtidran gràfics diferents.



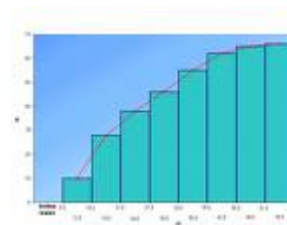
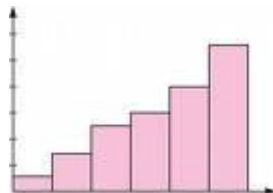
Si en lloc de dibuixar barres, unim amb una línia poligonal els punts que tenen com a abscissa el valor de la variable i com a ordenada la freqüència corresponent, obtenim el **polígon de freqüències**.

També es pot representar simultàniament una variable en dues situacions distintes: **diagrama de barres juxtaposades** i **diagrama de barres apilades**.

Histograma: s'utilitza per representar gràficament les dades d'una variable contínua (o quantitativa discreta amb moltes modalitats), que han estat agrupades en intervals de classes. A un eix (X) representen les classes, tenint en compte l'amplitud dels intervals, i a l'altre eix (Y) aixequem un rectangle d'àrea igual a la freqüència de la classe. Els rectangles són adjacents, és a dir, units.



De vegades és més útil representar las freqüències acumulades:



Polígons de freqüència: en les variables quantitatives o en les qualitatives ordinals es poden representar polígons de freqüència en lloc d'histogrames; es construeix a partir d'un punt en el centre de cada interval (o centre de la barra) de manera que l'altura del punt sigui igual a la freqüència o freqüències relatives dels intervals. A continuació s'uneixen els punts amb línies rectes.

És el gràfic més adient per representar l'evolució d'un caràcter en el temps.

Quan es representa la freqüència acumulativa, es denomina ogiva.

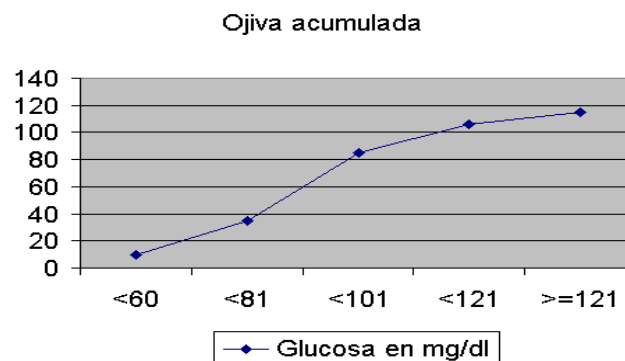


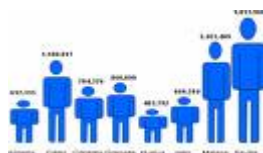
Diagrama de sectors: es fa servir per a dades qualitatives o per a quantitatives discretes, és a dir, quan la variable té pocs valors diferents. Les dades es representen en un cercle

bidimensional o tridimensional, on a cada dada li correspon un sector circular d'amplitud proporcional a la freqüència absoluta corresponent. Per calcular les diferents amplituds , cal tenir en compte que al total de dades li corresponen 360 °:

$$\text{amplitud del sector (}^\circ\text{)} = \frac{\text{pi}}{\text{total}} \times 360^\circ$$



Pictograma: és el gràfic que obtenim si substituïm les barres per dibuixos relacionats amb el valor que pren la variable.



4.3 MESURES DE CENTRALITZACIÓ, POSICIÓ I DISPERSIÓ

En un treball estadístic, a part de tenir la taula de freqüències i la representació gràfica de les variables quantitatives, resulta útil disposar d'uns valors que ens resumeixen com es distribueixen les dades i uns altres que ens informin com estan d'agrupades o disgregades aquestes a la distribució.

Els paràmetres de centralització són valors que informen sobre el valor central al voltant del qual es distribueixen les dades. Els més habituals són: la mitjana aritmètica, la mediana i la moda.

Els paràmetres de dispersió: són valors que informen sobre el grau de separació o concentració de les dades. Els més habituals són: el rang, la variància i la desviació estàndard.

4.3.1 MESURES DE CENTRALITZACIÓ

MITJANA ARITMÈTICA

S'anomena mitjana aritmètica d'una distribució estadística el quocient entre la suma de tots els valors x_i observats i el nombre total d'observacions, N . Es representa per $\bar{x}$.

$$\bar{x} = \frac{\sum_{i=1}^N x_i \times n_i}{N}$$

Exemple: la mitjana aritmètica de les dades que s'observen a la taula adjunta és 2,82

x_i	n_i	$x_i \cdot n_i$
1	8	8
2	7	14
3	9	27
4	4	16
5	5	25
6	1	6
	N = 34	96

$$\bar{x} = \frac{96}{34} = 2.82$$

En el cas de dades agrupades en intervals o classes, prendrem com a x_i les marques de cadascuna de les classes.

Pot utilitzar-se com a mesura de centralització per a variables discretes o contínues, però no per a variables ordinals o nominals.

És un estadístic molt sensible (no robust) a la presència de dades atípiques o extremes. En aquestes casos és preferible calcular la mitjana truncada [media recortada al 5%] = mitjana de la variable després de suprimir el 5% dels valors superiors i el 5% dels valors inferiors. D'aquesta manera la mitjana que resulta és més robusta davant la possible presència de dades atípiques.

Es considera el centre de gravetat de les dades.

MODA (MO)

La moda d'una distribució estadística és la modalitat o modalitats de freqüència més alta, la que es repeteix més vegades. En el cas de variable contínua, considerarem per moda la marca de l'interval de major freqüència.

Si aquest valor és únic, parlem de distribució unimodal. Si té dues, tres o més modes, parlarem de distribució bimodal, trimodal etc. Si totes les freqüències són iguals, la distribució no té moda.



Es pot considerar una mesura de tendència central que pot emprar-se per a tot tipus de dades.

Exemple: a la taula següent sobre el nombre de parts la moda és 0 part.

x_i	n_i
0	87
1	61
2	47
3	25
4	28
≥ 5	25

MEDIANA (M)

S'anomena mediana d'una distribució estadística, un cop les dades ordenades, al valor de la variable tal que la freqüència absoluta dels valors més petits que ell és igual a la freqüència absoluta dels valors més grans que ell (aquest valor deixa per sota d'ell la meitat de les dades, el 50%, i per sobre, l'altra meitat).

Es pot utilitzar com a mesura resum per a les observacions ordinals, així com per les discretes i contínues.

Com es calcula la mediana?

En el cas d'una **variable discreta**:

- Si el nombre total de dades és imparell, existeix una dada central, que és la que ocupa el lloc central $\frac{n+1}{2}$, és a dir, la dada $x_{\frac{n+1}{2}}$.

Exemple: la mediana del conjunt $\{3,4,5,7,8,9,21\}$ és $M = 7$

- Si el nombre total de dades és parell, es fa la mitjana aritmètica entre els valors dels llocs $\frac{n}{2}$ i $\frac{n}{2} + 1$, és a dir, entre $x_{\frac{n}{2}}$ i $x_{\frac{n}{2} + 1}$.

Si tenim les dades en una taula, només cal buscar quin és el valor que correspon al 50% de les freqüències acumulades.

Exemples:

La mediana del conjunt $\{3,4,5,7,8,9\}$ és $M = \frac{5+7}{2} = 6$

Si les dades les tenim disposades en una taula hem d'esbrinar el valor que es correspon amb la f. absoluta acumulada igual a la meitat del nombre de dades $200 / 2 = 100$. Hi ha dos valors

centrals: 11 i 17. La mediana és $M = \frac{11+17}{2} = 14$

X_i	n_i	n_{ia}
8	40	40
10	40	80
11	20	100
17	48	148
21	52	200

Si les dades estan agrupades o la variable és contínua hem de utilitzar l'expressió:

$$M = L_i + c \times \frac{\frac{N}{2} - n_{ia-1}}{n_{im}}$$

On:

L_i és el límit inferior de la classe on es troba la mediana

n_{im} és la f. absoluta de la classe on es troba la mediana

n_{ia-1} és la f. absoluta acumulada de d'interval anterior

c és l'amplitud d'interval

$N/2$ és la meitat de les dades

Exemple:

classes	n_i	n_{ia}
0-5	9	9
5-10	7	16
10-15	7	23
15-20	8	31
20-25	14	45
25-30	5	50

La classe mediana de la distribució de la taula és 15-20. Per tant:

$$L_i = 15$$

$$n_{im} = 8$$

$$n_{ia-1} = 23$$

$$c = 5$$

Substituint tenim:

$$M = 15 + 5 \cdot \frac{50 - 23}{8} = 16.25$$

Indica el centre geomètric de la distribució. És convenient quan les dades són asimètriques. És un estadístic robust a la presència de dades atípiques.

4.3.2 MESURES DE POSICIÓ

QUARTILS (Q)

Són tres valors de la variable que divideixen el conjunt de dades en 4 parts, amb el mateix nombre d'elements, quan aquestes dades estan ordenades.

En el cas que la variable sigui discreta calcularem els quartils de la següent manera:

El primer quartil (Q_1) és el valor que deixa per sota d'ell el 25% de les dades, o el 75% per sobre. Ocupa el lloc $\frac{n+1}{4}$. El segon quartil (Q_2) deixa per sota d'ell el 50% de les dades i coincideix amb la mediana M . El tercer quartil (Q_3) deixa per sota d'ell el 75%, o el 25% per sobre; és el valor que ocupa el lloc $\frac{3(n+1)}{4}$.

També es pot mirar en les f. acumulades en tant per u.

PERCENTILS (P)

Quan el conjunt de dades queda dividit en 100 parts iguals. Tenim 99 mesures, que representem per $P_1, \dots, P_{99}$.

Els quartils i els percentils es calculen de la mateixa manera que la mediana. Són mesures de posició, no de centralització.

4.3.3 MESURES DE DISPERSIÓ

RANG

S'anomena rang d'una distribució estadística la diferència entre el valor màxim i el valor mínim d'aquesta.

El rang és molt sensible a valors extrems. Per evitar aquest inconvenient, s'utilitzen altres mesures en què s'eliminen aquests valors extrems: el rang interquartílic i el rang entre percentils:

Rang interquartílic: és la diferència entre el tercer i el primer quartil:

$$Q = Q_3 - Q_1$$

Una dada x es considera **atípica** si:

$$X < Q_1 - 1.5Q \quad \text{o} \quad X > Q_3 + 1.5Q$$

Rang entre el percentils n i $100 - n$: és la diferència entre els percentils

$$P = P_{100-n} - P_n$$

DESVIACIÓ MITJANA (D_M)

És la mitjana aritmètica dels valors absoluts de les diferències entre les observacions i la mitjana.

$$d_m = \frac{\sum_{i=1}^N |x_i - \bar{x}| \cdot n_i}{N}$$

La raó per la qual prenem el valor absolut de les diferències és que aquestes poden ser tant positives com a negatives (segons es tracti de valors per sobre o per sota de la mitjana) i per tant, si no prenguéssim el valor absolut, en sumar-les s'anul·larien les unes amb les altres.

Exemple: en els quatre controls de M9 de la UF1, un alumne ha obtingut les notes: 3, 6, 5 i 8. La seva mitjana $\bar{x} = 5.5$. Comprovarem que la suma de les desviacions és zero i calcularem la desviació mitjana.

x_i	$x_i - \bar{x}$	$ x_i - \bar{x} $
3	-2.5	2.5
5	-0.5	0.5
6	0.5	0.5
8	2.5	2.5
suma	0	6

Desviació mitjana $d_m = 6 / 4 = 1.5$

VARIÀNCIA (σ^2)

És la mitjana aritmètica dels quadrats de les diferències entre les dades i la mitjana.

$$\sigma^2 = \frac{\sum_{i=1}^n x_i^2 \cdot n_i}{N} - \bar{x}^2$$

DESVIACIÓ TIPUS O ESTÀNDARD (σ)

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

És l'arrel quadrada positiva de la variància. Té l'avantatge sobre la variància que s'expressa en les mateixes unitats que els valors de la variable, mentre que la variància s'expressa amb aquestes unitats al quadrat. Per tant és més fàcil d'interpretar la informació que dona la desviació tipus.

Com més grans són la variància o la desviació típica, les dades se separen més de la mitja, és a dir, hi ha més dispersió.

En les distribucions en forma de campana, quan N és gran, es compleix aproximadament que :

- Entre $\bar{x} - s$ i $\bar{x} + s$ hi ha al voltant d'un 68% de les dades
- Entre $\bar{x} - 2s$ i $\bar{x} + 2s$ hi ha al voltant d'un 95% de les dades
- Entre $\bar{x} - 3s$ i $\bar{x} + 3s$ hi ha al voltant d'un 100% de les dades

COEFICIENT DE VARIACIÓ (CV)

És el quocient entre la desviació tipus i la mitjana aritmètica.

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

Aquesta raó s'utilitza per comparar la dispersió entre diferents distribucions, atès que una mateixa desviació tipus pot indicar diferents dispersions segons la magnitud de les dades.

També se la denomina variabilitat relativa. Es freqüent mostrar-la en percentatges.

Exemple: si la mitjana és 80 y la desviació tipus 20 llavors $CV=20/80=0,25=25\%$ (variabilitat relativa).

Si el pes té $CV=30\%$ y la altura té $CV=10\%$, els individus presenten més dispersió en pes que en altura.

No s'ha d'utilitzar quan la variable presenta valors negatius on el valor 0 és una quantitat fixada arbitràriament.

Per exemple: $0^{\circ}\text{C} \neq 0^{\circ}\text{F}$

COEFICIENT D'ASIMETRIA

El coeficient d'asimetria d'un conjunt de dades $x_1, x_2, \dots, x_n$, de mitjana i desviació estàndard:

$$CA = \frac{1}{N} \frac{\sum (x_i - \bar{x})^3}{\sigma^3}$$

En cas que les dades estiguin repetides, hem de multiplicar les diferències entre cada valor de la variable i la mitjana aritmètica per les seves freqüències

- Si $CA \sim 0$ la distribució és simètrica

- Si $CA \gg 0$ la distribució presenta un biaix a la dreta
- Si $Ca \ll 0$ la distribució presenta un biaix a l'esquerra

Resum sobre la mitjana i la mediana; influència dels valors atípics

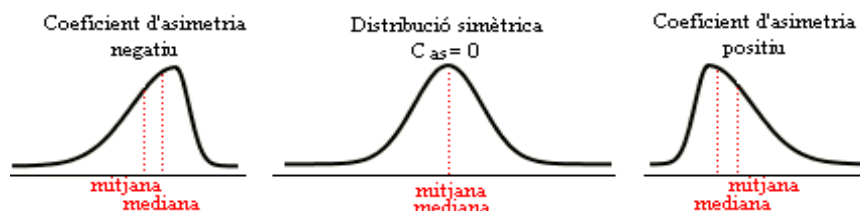
Per a la utilització de la mediana i de la mitjana com a paràmetres de centralització, hem de tenir ben present les idees següents:

- La mitjana és un paràmetre centralitzador, el centre de gravetat de la distribució.
- La mediana presenta limitacions pel fet de no tenir en compte totes les dades de la distribució (el canvi d'una dada no té per què fer canviar el valor d'aquests paràmetre).

Precisament pel que acabem de comentar, la mediana és un paràmetre rellevant en cas que la distribució presenti dades singulars, justament el contrari de la mitjana, que es veu fortament influïda si ens trobem amb dades singulars.

Visualització de la mitjana i la mediana: la influència dels valors atípics

Una manera intuïtiva de mesurar el grau de simetria d'una variable numèrica és la de comparar els valors de la mediana i la mitjana. Efectivament, si la distribució és totalment simètrica, la mediana i la mitjana coincideixen i, en canvi, la distribució difereix més d'un model simètric com més distanciades estiguin la mediana i la mitjana, de tal manera que "la cua més allargada" es presenta cap al cantó de la distribució on es trobi la mitjana.



Com que la variable estandarditzada s'expressa sense unitats, això mateix succeeix amb el valor del coeficient d'asimetria.

Quan la distribució és simètrica, $C_{as} = 0$, perquè els sumands positius es compensen amb els negatius; si el valor de C_{as} és proper a 0, es pot considerar un grau de simetria alt per a la distribució estudiada i, doncs, una versemblança del model normal si ens fixem només en la simetria.

En canvi, si el C_{as} té un valor positiu gran, és que predominen els sumands positius en el sumatori que ens porta al valor de C_{as} ; aquests sumands corresponen als valors de la variable superiors a la mitjana. En aquest cas, en què predominen els valors allunyats de la mitjana cap a la dreta es diu que **la distribució és esbiaixada cap a la dreta** o que presenta "una cua" cap a la **dreta**.

Semblantment, si el coeficient d'asimetria té un valor negatiu gran en valor absolut es diu que *la distribució és esbiaixada cap a l'esquerra* i, en aquest cas, el perfil presenta "una cua" cap a l'esquerra.

DIAGRAMES DE CAIXA

S'anomena diagrama de caixa i bigotis, una representació gràfica que es pot elaborar després de localitzar les dades extremes d'una distribució estadística i calcular els valors dels quartils, amb la finalitat de fer visual la posició relativa d'aquests valors i, doncs, donar una idea de la dispersió que presenta la distribució de les dades al llarg del rang.

Com a primera idea, per construir un diagrama de caixa dibuixarem un eix graduat amb els valors de la variable i, referit a aquest eix, un rectangle de base igual al segment que separa el primer i el tercer quartil, partit en dos per una línia feta pel valor de la mediana. A banda i banda del rectangle es dibuixen uns segments anomenats cues o bigotis. El bigoti esquerre queda determinat pel valor mínim i el Q_1 i el dret pel tercer quartil i el valor màxim, excepte que hi hagin valors atípics. Vegeu un exemple de diagrama de caixa:

Es consideren valors atípics els menors a $Q_1 - 1.5 \cdot (Q_3 - Q_1)$ i els majors a $Q_3 + 1.5 \cdot (Q_3 - Q_1)$. Això vol dir que la longitud màxima dels bigotis és una vegada i mitja la longitud de la caixa ($Q_3 - Q_1$). Quan hi ha valors atípics aquests es representen fóra del diagrama de Caixa, mitjançant asterisc o punts.

El gràfic de l'exemple (que recull les notes d'un grup d'alumnes en un examen) mostra una major concentració de notes a la zona central de la distribució perquè la longitud de la caixa és més petita que la de les cues. Cues més llargues no vol dir que hi ha moltes dades lluny de la mediana, sinó que les dades lluny de la mediana són més disperses.

També ens mostra que es tracta d'una distribució lleugerament esbiaixada cap a l'esquerra: la "cua" d'aquest costat és sensiblement més llarga que la de la dreta, però aquesta tendència

queda matisada pel fet que les dues parts de la caixa central tenen la mateixa llargada (i això indica un repartiment ben simètric del 50 % de dades centrals entorn a la mediana).

La longitud de la caixa central, ens explica com de juntes estan les dades centrals i per això, com ja hem dit, pot ser entès com un paràmetre de dispersió i s'empra també per caracteritzar les dades atípiques.

L'ús fonamental dels diagrames de caixa no el trobem en la descripció d'una variable sinó com a element intuïtiu de **contrast entre distribucions de variables estadístiques**.

Efectivament, si imaginem juxtaposats dos diagrames de caixa relatius a variables anàlogues (per exemple, les notes d'un altre examen en el mateix grup classe que l'anterior), veurem que efectivament això permet una comparació global i ràpida entre la distribució de dades de dues variables anàlogues definides sobre la mateixa població. Semblantment, si hom ha de comparar dues o més distribucions d'una mateixa variable mesurada sobre poblacions diferents (o sobre diferents grups estadístics en una mateixa població), la visualització simultània dels corresponents diagrames de caixa sol aportar una informació molt valuosa.

A l'enllaç que hi ha a continuació pots observar com es construeix el diagrama de caixa:

http://www.caib.es/sites/racoeducatiudelibestat/ca/diagrama_caixa-36557/