Cuinant cuinant... la millor recepta estadística[1]

Maria T. Sanz
Emilia López-Iñesta
Departament de Didàctica de la Matemàtica
Universitat de València

[1] Aquest treball s'ha realitzat amb l'ajuda del projecte UV-SFPIE_PID19-1095187, Kit-Est: eines estadístiques per a alumnat de grau en mestre d'educació primària, en el marc de Projectes d'innovació educativa i millora de la qualitat docent a la Universitat de València, en la convocatòria del 2019-2020.

Avui dia, les dades estadístiques envaeixen els mitjans de comunicació i les xarxes socials, dades que consumeixen xiquetes, xiquets, joves i adults. Sembla que tot receptor hauria de tenir, per tant, una cultura estadística suficient per a entendre els gràfics, les taules i les xifres que irrompen a casa (Batanero, 2002). Això ens porta a convenir que aquests coneixements s'han d'adquirir als centres educatius i, per tant, les i els docents han de tenir la formació necessària perquè això passe.

La nostra pregunta és: el professorat té el coneixement estadístic necessari per a poder produir, consumir i comunicar dades estadístiques per a intervenir en la vida quotidiana, més específicament en la realitat d'una aula? Coneixen els i les docents l'estadística com a eina per a la millora de l'ensenyament, i per tant, de l'aprenentatge de l'alumnat?

El cas és que, tot i que ara per ara, a causa del desenvolupament de la societat de la informació, l'estadística es considera un coneixement imprescindible per a la participació ciutadana i la presa de decisions, com apunta Alsina (2016), aquesta part de les matemàtiques escolars ha rebut, fins ara, més poca atenció que la resta. En definitiva, tot i considerar-se l'estadística un contingut imprescindible per a aconseguir, juntament amb molts altres factors, que els estudiants siguen ciutadans ben informats i consumidors intel·ligents, avui dia continua sense haver-se assolit, en termes generals, el desenvolupament eficaç del seu contingut.

Hi ha estudis que mostren que, en l'actualitat, alguns d'aquests continguts s'ensenyen de forma deficitària i en dissonància amb el currículum (Alsina i Vásquez, 2016). S'hi poden adduir diverses raons: *a*) l'escassa formació inicial i permanent de professorat; *b*) els qüestionats enfocaments d'ensenyament amb els quals es presenten els continguts estadístics en els llibres de text i, finalment, *c*) les decisions del professorat d'afavorir altres parts del currículum en detriment de l'estadística per causa, sovint, de falta de temps.

A més, en l'àmbit universitari podem notar que la situació no és molt diferent. Així, en particular, a la Universitat de València, en el Grau de Mestre/a en Educació Infantil i Primària, únicament hi ha una assignatura obligatòria per a tots els discents, Matemàtiques per a Mestres, de 90 hores i, en el cas dels titulats de primària, tenen l'assignatura de Didàctica de la Geometria, la Mitjana i la Probabilitat i l'Estadística, de 60 hores.

Aquest dèficit en la formació estadística ens porta a pensar en la necessitat d'un suport didàctic, ja que les i els docents no sols haurien de dominar els coneixements estadístics específics que indica el currículum per a cada curs, sinó que també haurien de tenir una alfabetització estadística com qualsevol ciutadà, que en termes de Gal (2002) serien els següents:

a) La capacitat per a interpretar i avaluar críticament la informació, els arguments basats en dades o els fenòmens estocàstics que les persones poden trobar en diversos contextos, incloent-hi els mitjans de comunicació, però sense limitar-se a aquests mitjans.
b) La capacitat de discutir o opinar sobre aquestes informacions estadístiques quan siga rellevant.

Aquesta alfabetització s'hauria de complementar amb les eines estadístiques necessàries per a potenciar la millora del seu quefer professional. Tal com indica Estrada (2007), la o el "docent com a professional necessitarà el coneixement estadístic per a la selecció i l'ús adequat d'eines útils en les seues pròpies anàlisis i presa de decisions. Dia a dia, el professor ha de prendre decisions en què hi ha una incertesa associada i, per això, necessita coneixements de probabilitat i sobre els errors freqüents en els judicis probabilístics. El professor dissenya i usa instruments d'avaluació i pren dades sobre els coneixements dels alumnes, però aquestes dades són sempre

limitades i estan subjectes a variabilitat aleatòria, i per això ha d'aplicar inferència a partir d'una mostra de dades específiques. El professor ha d'analitzar la distribució d'una certa capacitat o competència en la seua classe a fi de detectar casos atípics, tant d'alumnes destacats com d'aquells que necessiten alguna ajuda; ha de comparar les seues dades amb els paràmetres nacionals o amb altres grups de cursos passats o d'altres companys".

Amb tot això, el propòsit d'aquest llibre és acostar els coneixements estadístics al professorat de tots els nivells, des de l'ensenyament primari fins al superior, perquè dispose d'eines objectives per a la millora de la pràctica docent, a la manera d'un mestre cuiner entre els fogons.

A fi de complir aquest objectiu, el nostre text es divideix en tres capítols. El primer, anomenat *Preparem tots els ingredients* en referència a la pràctica culinària, mostra que el primer pas en estadística és conèixer els elements amb els quals cal treballar, exactament igual que en una cuina s'han de conèixer amb detall tots els estris i ingredients amb els quals hem de funcionar per a fer la millor de les receptes. El segon capítol fa referència a la mà d'obra, *Cuinem les dades*: una vegada coneguts els elements essencials, ens podem disposar a planificar la recepta i començar amb les mescles més bàsiques. Així doncs, en aquest capítol centrem l'atenció en el disseny d'un bon experiment dins de les aules i en l'estadística descriptiva, és a dir, la recollida i el tractament de la informació necessària per a l'emplatat final. Finalment, el capítol 3, *Comparem les nostres receptes*: per a poder comparar bones receptes han d'estar cuidades al detall i hem de poder reconèixer-hi la més mínima diferència per a identificar quina és la que despunta sobre les altres. Així doncs, en aquest últim capítol analitzarem amb detall tot el que s'ha obtingut abans i tractarem de fer-hi comparacions, activitat que en aquest camp del coneixement es denomina estadística inferencial.

Cal fer notar que a més de tractar-se d'un llibre que recull la part teòrica d'aquesta disciplina, adaptada a futur professorat de nivell infantil, primària i secundària, també té suport visual i apartats denominats **PRÀCTICA**.

El suport visual s'aconsegueix gràcies al projecte UV-SFPIE_PID19-1095187 que sustenta aquest treball en el marc Projectes d'innovació educativa i millora de la qualitat docent a la Universitat de València. És un recull de tres vídeos disponibles al canal YouTube del SFPIE de la Universitat de València. El primer dels vídeos (figura 1) es pot veure en l'enllaç següent: https://www.youtube.com/watch?v=6-Ja07Nmdco. S'hi inclou una explicació del projecte d'innovació docent, on una de les autores i coordinadora d'aquest llibre exposa la importància que té que el futur professorat d'educació primària conega i use eines estadístiques.

Figura 1. Vídeo explicatiu d'aquest llibre: https://www.youtube.com/watch?v=6-

Ja07Nmdco

En la part de **PRÀCTICA**, el lector ha de tractar de posar en pràctica allò que ha llegit fins al moment. Aquesta part s'aborda a través d'un projecte. L'elaboració de projectes estadístics a l'aula és un mètode que ajuda a tractar continguts, en particular continguts estadístics, en un context pròxim a l'alumnat, a la seua vida diària, als objectes i elements d'ús quotidià, a la seua situació sociodemogràfica, econòmica o a l'estudi de situacions que li desperten interès.

El desenvolupament del projecte el determinen tres fases:

- Comencem: fase de descobriment de les idees prèvies i motivació pels nous aprenentatges.
- Investiguem: conté les experiències manipulatives, d'investigació, intercanvi verbal i raonament matemàtic que permetran a l'alumne identificar i interpretar continguts a fi d'aplicar-los a les seues tasques.
- Comuniquem: aquesta última fase inclou activitats de comunicació i exposició de resultats.

Aquestes fases permeten a l'alumnat treballar activament en l'àmbit formatiu de manera que, a través de la investigació i la realització de tasques, s'acoste al coneixement estadístic.

Referències bibliogràfiques

Alsina, Á. (2016). "La estadística y la probabilidad en educación primaria. ¿Dónde estamos y hacia dónde queremos ir?" *Uno. Revista de Didáctica de las Matemáticas,71*, 46-52.

Alsina, Á. i Vásquez, C. (2016). "La probabilidad en educación primaria. De lo que debería enseñarse a lo que se enseña". *Uno. Revista de Didáctica de las Matemáticas*, *71*, 46-52.

Batanero, C. (2002). *Los retos de la cultura estadística.* Conferència inaugural de les Jornades Interamericanes de l'Ensenyament de l'Estadística. Recuperat el 16 de juliol del 2012 d'ací: <www.ugr.es/~batanero/ARTICULOS/CULTURA.pdf>.

Estrada, A. (2007). "Evaluación del conocimiento estadístico en la formación incial del profesorado". *Revista Uno, 45*, <https://docplayer.es/31385355-Evaluacion-del-conocimientoestadistico-en-la-formacion-inicial-del-profesorado.html>.

Gal, I. (2002). "Adults' Statitistical Literacy: Meanings, Components, and Responsabilities". *International Statistical Review, 70* (1), 1-51.

Capítol I. Preparem tots els ingredients[2]

"Quan els clients coneixen les tècniques necessàries per a fer alta cuina, valoren els plats que els servim i en gaudeixen molt més". Pedro Subijana

Segons Pedro Subijana: "Quan els clients coneixen les tècniques necessàries per a fer alta cuina, valoren els plats que els servim i en gaudeixen molt més". De la mateixa manera, quan les i els docents coneixen les entranyes d'allò que tracten d'ensenyar a l'alumnat, gaudeixen i transmeten aquests coneixements des del cor i amb la seguretat del saber. A més, són capaços de trobar la utilitat d'allò que ensenyen i, anant més enllà de l'ensenyament, passen a l'aplicabilitat.

És per això que en aquest capítol centrem l'atenció en el coneixement de tots els conceptes estadístics que una o un bon docent hauria de conèixer per dues raons fonamentals: per a enriquir la metodologia pròpia i per a transmetre'ls, en la mesura que les i els discents siguen capaços de comprendre'ls. Així, aquest primer capítol és útil tant per a la docència com per a la millora de la docència. A més, tot allò que s'aprenga serà fonamental per a poder analitzar la intervenció docent.

Un exemple es pot trobar al repositori de vídeos del SFPIE de la Universitat de València (<https://www.youtube.com/watch?v=uKANAmbA1AE>). S'hi presenta un dels millors projectes desenvolupat per l'alumnat realitzat en la pràctica docent que s'inclou en aquest llibre, on es fa una explicació de l'estadística descriptiva (aquest capítol) i inferencial (capítol II) emprada per a analitzar les dades que van obtenir en el seu projecte. L'objectiu era analitzar el perfil de l'alumnat que cursa el Grau de Mestre/a en Educació Infantil i Primària en relació amb la via d'accés al grau i l'opció/prioritat d'elecció del grau a l'hora de presentar la sol·licitud d'entrada i matrícula a la universitat.

I.1. Primeres idees

I.1.1. Per què l'estadística?

Fins ara s'ha parlat únicament de l'estadística com una eina fonamental per al personal docent, però no hem definit el concepte d'estadística. La Reial Acadèmia Espanyola indica entre les diverses accepcions d'aquest mot:

> "[...] 3. *f.* Estudi de les dades quantitatives de la població, dels recursos naturals i industrials, del tràfic o de qualsevol altra manifestació de les societats humanes.
>
> [...] 5. *f.* Branca de la matemàtica que usa grans conjunts de dades numèriques per a obtenir-ne inferències basades en el càlcul de probabilitats."

Així, en l'accepció número 3 parla únicament de dades quantitatives. Algú ha sentit mai parlar de dades qualitatives? I en l'accepció número 5 parla de dades numèriques; i les no numèriques?, així com d'inferències, i les descriptives? Aquestes preguntes que es plantegen fan pensar a ampliar la definició d'estadística de la manera següent:

> L'estadística és una branca de les matemàtiques que estudia els mètodes i els procediments per a recollir, classificar, resumir, trobar regularitats i analitzar dades, sempre que la variabilitat i incertesa siguen una causa intrínseca; així com per a fer inferències a partir d'aquestes dades a fi d'ajudar a la presa de decisions i, si escau, formular prediccions.

Així, l'estadística tracta d'aconseguir una aproximació a la realitat, la qual sempre és molt més complexa i rica que el model que podem abstraure. Si bé aquesta ciència és ideal per a descriure processos quantitatius, té seriosos problemes per a explicar *el perquè* qualitatiu de les coses.

En general, podem parlar de dues branques diferents en estadística: l'estadística descriptiva i l'estadística inferencial,

- "L'estadística descriptiva té com a fi presentar resums d'un conjunt de dades i posar de manifest les característiques d'aquestes dades mitjançant representacions gràfiques. Les dades s'usen per a fins comparatius i no s'hi fan servir principis de probabilitat. L'interès se centra a descriure el conjunt de dades i no es planteja estendre les conclusions a altres dades diferents o una població." (Batanero i Godino, 2002).
- "La inferència estadística, per contra, estudia els resums de dades amb referència a un model de tipus probabilístic. Se suposa que el conjunt de dades analitzades és una mostra d'una població i l'interès principal és predir el comportament de la població a partir dels resultats de la mostra." (Batanero i Godino, 2002).

Cal fer notar que en aquest capítol únicament es tracten els elements necessaris per a iniciar un estudi estadístic, ja sigui descriptiu (capítol II) o inferencial (capítol III).

I.1.2. Sobre què es pot fer estadística?

Així com a la cuina, per a començar una recepta i el seu procés cal saber el fi últim quin és, per a poder començar un estudi estadístic cal un objecte d'estudi, és a dir, determinar què es vol estudiar. Així, per exemple, l'objecte d'estudi en una cuina pot ser:

ES1. La classe de ganivet que usen en les cuines espanyoles per a tallar cada tipus d'aliment.

ES2. Determinar si afegir un cert aliment a una recepta, per exemple, afegir pinya a la *pizza* barbacoa, en millora el gust (avaluat per comensals).

No obstant això, en una aula de futur professorat ens podria interessar:

ES3. El nombre d'estudiants que tenen coneixements sobre conceptes estadístics abans de començar el temari d'estadística.

o

ES4. Determinar si la metodologia basada en projectes fa que el procés d'ensenyament i aprenentatge siga millor (quant a adquisició i aplicació del coneixement) en l'àrea d'estadística en comparació d'un mètode tradicional.

PRÀCTICA

Pr. I.0.1. Delimita el teu tema de recerca (el teu projecte). Per a fer-ho, traça l'objecte d'estudi. Mira al voltant i pensa sobre què vols extraure informació. Pots pensar en una pregunta, en un interrogant que et faces sobre una qüestió concreta. Aquesta activitat es fa en grups de quatre persones i és l'activitat desenvolupada al llarg de tot el llibre.

Pr. I.0.2. Imagina't que estàs en una aula d'infantil o primària: què podries voler investigar? En grups de quatre o cinc persones tracteu de donar-ne quatre exemples.

Una vegada decidit què es vol estudiar / investigar, cal pensar en les característiques que té que són d'interès per a l'estudi. Això ens fa arribar a la definició del concepte *variable estadística*, que justament és qualsevol característica (numèrica o no) que ens interessa estudiar i les dades sobre la qual es poden extraure amb algun mètode (una enquesta, mitjançant observació, etc.). Aquestes variables són de diversa mena i es poden classificar de la forma següent (esquema en la figura I.0.1).

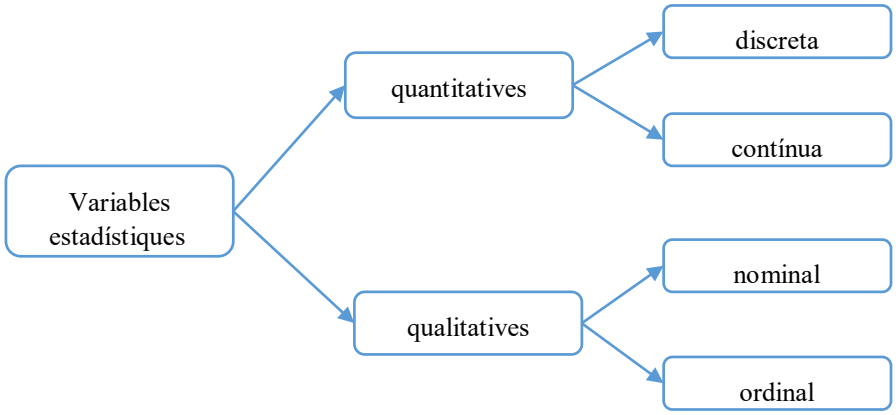


Figura I.0.1. Esquema del tipus de variables en un estudi estadístic.

- **Variables quantitatives**: són les que responen a la pregunta *quant?*
  Es poden expressar numèricament (és a dir, sempre prenen un valor numèric). Al seu torn es divideixen en:

  - ***Variables contínues***: poden prendre qualsevol valor (enter o no) dins d'un rang determinat de valors. També es pot dir que els valors de les variables contínues estan definits en un interval: per exemple, la "nota numèrica en una prova de matemàtiques" pren valors entre [0, 10]. Això implica que la variable *NOTA* pot prendre valors decimals.
  - ***Variables discretes***: només poden prendre uns certs valors concrets (habitualment nombres enters). Un exemple el tenim en la variable *nombre de germans/anes*. Es poden tenir 0, 1, 2... germans/anes (valors enters). No es pot tenir *un germà i mig* ;-).

- **Variables qualitatives o categòriques**: responen a la pregunta *de quin tipus?* Poden prendre qualsevol valor, numèric o de qualsevol altre tipus. Cadascun dels possibles valors que pot prendre aquesta mena de variables s'anomenen categories. Al seu torn, les variables qualitatives es divideixen en:

  - ***Variables ordinals***: són les variables de tipus qualitatiu en què les possibles respostes admeten una ordenació lògica o una graduació. Per exemple, la variable *interès per les matemàtiques* pot prendre valors en una escala des de molt interessat fins a molt poc interessat.
  - ***Variables nominals***: són les variables de tipus qualitatiu a les respostes possibles que NO admeten cap tipus d'ordenació lògica. Per exemple, la variable *aprovar* pot prendre valors *sí* o *no*.
    Cal tenir present que ací s'inclouen les variables denominades dicotòmiques, és a dir que només accepten dues categories: per exemple, resposta a una pregunta de tipus test amb opcions *sí* / *no*.

Nota: les variables solen indicar-se amb la lletra X. Per exemple, estudiarem la variable X = alçada en centímetres d'alumnes de la nostra classe. Les lletres minúscules es reserven per a valors concrets o específics de la variable. Per exemple, $x = 1{,}55$ cm. No obstant això, el més habitual és que hi haja més d'un valor. Així, en una aula on hi ha 50 persones, tenim $x_1 = 1{,}55$ cm, $x_2 = 1{,}71$ cm, ..., $x_{50} = 1{,}76$ cm.

Per als exemples que s'han marcat com a possibles objectes d'estudi, es poden delimitar algunes variables com les que es presenten a la taula I.0.1.

Taula I.0.1. Variables estadístiques dels objectes d'estudi ES1, ES2, ES3 i ES4.

Tipus de variable	**ES1**	**ES2**	**ES3**	**ES4**
Variable quantitativa discreta	-	-	1. Edat 2. Nombre d'estudiants amb coneixements sobre...	1. Edat 2. Nombre d'estudiants amb coneixements sobre

			(mitjana, moda, variància, ...)	(mitjana, moda, variància, ...)
Variable quantitativa contínua	-	1. Quantitat de pinya afegida	1. Temps que triguen a reconèixer el concepte (temps que triguen a respondre a cada pregunta)	1. Temps que triguen a reconèixer el concepte (temps que triguen a respondre a cada pregunta)
Variable qualitativa ordinal	-	-	-	-
Variable qualitativa nominal	1. Tipus de ganivet	1. *Pizza* barbacoa 2. *Pizza* barbacoa amb pinya 3. Tipus de comensal	1. Sexe 2. Via d'accés al Grau de Mestra/e	1. Nombre d'estudiants que treballen amb metodologia basada en projectes. 2. Nombre d'estudiants que treballen amb metodologia basada en projectes.

PRÀCTICA

Pr. I.0.3. Determina totes les variables que es poden tractar en la investigació proposada i classifica-les segons s'ha vist anteriorment.

I.1.3. On es pot fer estadística?

Una vegada determinat què es vol estudiar, amb totes les variables delimitades, hem de tractar d'aconseguir dades. Però d'on s'obtenen aquestes dades? Aquestes dades s'obtenen de la població.

Una població (o univers) és el conjunt total d'objectes que són d'interès per a un problema donat. Els objectes poden ser persones, animals, productes fabricats, etc. Cadascun rep el nom d'element (o individu) de la població.

Així doncs, la població relativa a l'estudi ES1 són tots els cuiners que hi ha al món, la població relativa a l'estudi ES2 són totes les persones que mengen *pizza* barbacoa, i la població d'ES3 i ES4 són les i els futurs docents.
Però, tal com s'ha delimitat en l'objecte d'estudi, ens interessa conèixer algunes propietats d'aquesta població a través de l'estudi d'algunes variables, tal com s'han delimitat en l'apartat anterior. De quina manera coneixerem aquestes característiques? La resposta és investigant els individus o elements de la població objecte d'estudi. Per a fer-ho, es poden diferenciar dos casos segons que es dispose d'una població finita o d'una població infinita.

Si la població és finita, el millor procediment és la inspecció de cada individu. Un estudi estadístic efectuat sobre tota una població es denomina cens. No obstant això, la majoria dels objectes d'estudi impliquen poblacions infinites o poblacions finites molt grans, cosa que demana molt de temps i un cost elevat per a poder investigar tots i cadascun dels elements que la formen. Això obliga a haver de seleccionar, per procediments adequats, un subconjunt d'elements de la població, denominada **mostra**, que ha de ser representativa d'aquesta població (figura I.0.2).

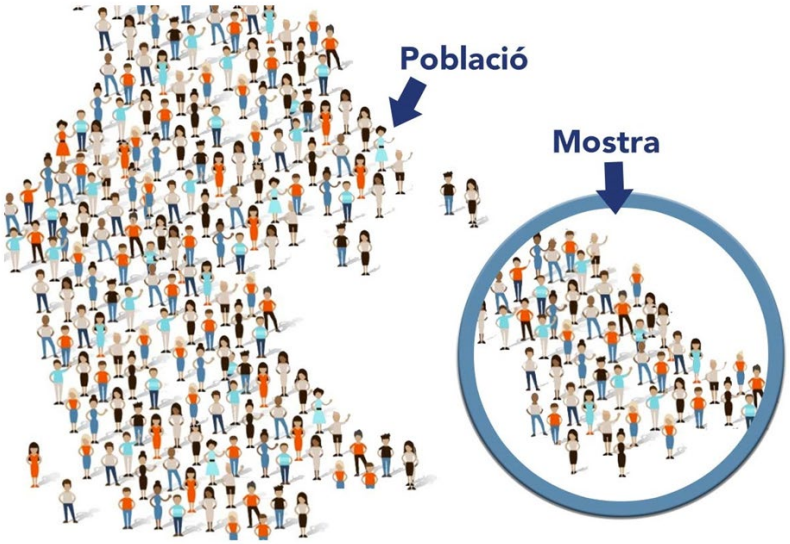


Figura I.0.2. Conceptes de població i mostra
font: <https://www.questionpro.com/es/tama%C3%B1o-de-la-muestra.html>

Per exemple, en el cas de l'estudi ES1, quan parlem de cuineres o cuiners, no podem preguntar a totes les cuineres i cuiners del món. Així doncs, la mostra es podria delimitar a Espanya i, ací, preguntar a 10 cuiners/eres de cada comunitat autònoma. Pel que fa a l'ES2, actuaríem de forma anàloga a ES1, tot i que en referir-nos a *pizza* potser ens interessa fer l'estudi a Itàlia, on es pressuposen experts en la matèria. En el cas dels estudis ES3 i ES4, actuaríem de forma anàloga. Per exemple, com que som docents o alumnes de la Universitat de València (Estudi General), l'estudi es podria reduir a alumnes del Grau de Mestra/e de la UVEG i d'ací prendre una mostra representativa d'alumnes de 2n curs d'aquest grau.

El mostratge o mostreig és, per tant, una eina de recerca o investigació científica que té per funció bàsica determinar quina part d'una certa població ha d'examinar-se amb la finalitat de fer inferències sobre aquesta població (la inferència la veurem en el capítol III).

La mostra ha de ser una representació adequada de la població en la qual es reproduïsquen de la millor manera els trets essencials d'aquesta població que són importants per a la investigació. Perquè una mostra siga representativa i, per tant, útil, ha de reflectir les similituds i diferències trobades en la població, és a dir exemplificar les característiques d'aquesta població.

Els errors més habituals que es cometen en aquest procés són:

- Arribar a conclusions molt generals a partir de l'observació de només una part de la població (error de mostratge).

- Arribar a conclusions sobre una població molt més gran que aquella de la qual originalment s'ha pres la mostra (error d'inferència).

Els diversos mètodes de selecció de mostres representatives d'una població es coneixen amb el nom de mètodes de mostratge i es divideixen en dos grans grups: 1. Mostratge probabilístic; 2. Mostratge no probabilístic.

A continuació detallem cada mostratge.

1. Mostratge probabilístic

Són els mostratges en què tots els individus o elements de la població (per exemple, futurs mestres d'educació primària) tenen la mateixa probabilitat de ser triats i, per tant, de formar part d'una mostra. Totes les possibles mostres de mida n tenen la mateixa probabilitat de ser seleccionades (mostres de grandària $n = 10$ futurs professors). Només aquests mètodes de mostratge probabilístic ens asseguren la representativitat de la mostra extreta i són, per tant, els més recomanables.

1.1. Mostratge aleatori simple

Es tracta d'un procediment poc útil per a poblacions molt grans. El procés és el següent:

a) S'assigna un número a cada individu o element de la població.

b) A través d'algun mitjà mecànic aleatoritzat es trien tants individus o elements com calga a fi de completar la mida de mostra.

Exemple: volem obtenir una mostra de 10 estudiants de la nostra classe en què hi ha aproximadament 50 persones. Tot l'alumnat es posa un dorsal amb un número que l'identifica (hi ha 50 dorsals). S'introdueixen en una bossa boletes o paperets amb números de l'1 al 50, es remenen i, sense mirar, se'n trauen 10. Els 10 números s'han obtingut de manera aleatòria.

En la figura I.0.3 s'ha fet un mostratge aleatori simple d'una població de mida 20 i se n'ha obtingut una mostra de mida $n = 9$.

Figura I.0.3. Mostratge aleatori simple

Font: <https://www.questionpro.com/blog/es/como-realizar-un-muestreo-aleatorio-simple/>

1.2. Mostratge aleatori sistemàtic

Es realitza el procés anterior, però una vegada obtinguda la primera dada, a partir d'ací se seleccionen els elements successius. Així, si aleatòriament s'obté el 38, els elements següents de la mostra són el 39, 40, 41... fins a completar la mida de la mostra requerida.

L'error que es pot cometre és la introducció d'homogeneïtat, és a dir, si per exemple en l'estudi d'ES3 ens interessara el sexe, potser un dels sexes podria estar aglutinat a partir d'un cert número i, per tant, la mostra no seria representativa.

En el cas de poder disposar de tots els elements d'una població de manera ordenada, es pot aplicar aquest tipus de mostratge denominat sistemàtic. Es denomina així perquè la tècnica per a seleccionar els elements de la mostra és *sistemàtica*.

Considerem el cas més senzill de mostratge anomenat *mostreig sistemàtic uniforme de pas k*.

Quins passos cal seguir per a l'obtenció de la mostra?

Tenim una població de mida $N = 9$ i volem obtenir-ne una mostra de mida $n = 3$ com s'indica en la figura I.0.4.

Cal seguir els passos següents:

1. Aconseguir una llista ordenada dels N elements de la població (en el nostre cas, podem imaginar que les persones de la figura I.0.4 porten un dorsal amb un número de l'1 al 9).

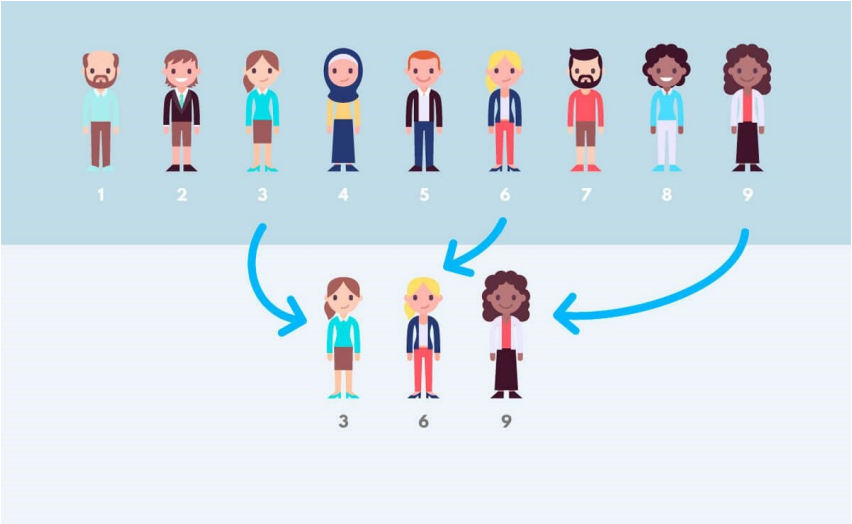


Figura I.0.4. Mostratge sistemàtic

Font: <https://www.questionpro.com/blog/es/como-realizar-un-muestreo-sistematico/>

2. Pensar o determinar quina és la mida de la mostra n que es necessita ($n = 3$).
3. Calcular la mida del salt sistemàtic k mitjançant la fórmula $k = N/n$. També es pot denominar k coeficient d'elevació. En el nostre exemple, $k = 9/3 = 3$.
4. Triar un nombre aleatori A que prenga un valor entre 1 i k (en el nostre exemple ja sabem que $k = 3$). Aquest nombre A serveix per a obtenir el primer element que forme part de la mostra de grandària $n = 3$. El valor de A es pot obtenir, per exemple, mitjançant un full de càlcul o una calculadora triant un nombre aleatori entre 1 i 3, mitjançant una taula de nombres aleatoris o a través d'altres mitjans més complexos. Nota: si k no és un nombre enter, es pot emprar l'arredoniment a l'enter més pròxim a N/n.

5. Finalment, quedaria la selecció dels n-1 individus restants prenent els elements segons aquesta successió: $A + k$, $A + 2k$, $A + 3k$, etc.
6. En la mostra de mida 3 hi haurà els elements 3, 6, 9 (els individus que porten els dorsals amb aquests números).

1.3. Mostratge aleatori estratificat

Consisteix a considerar categories típiques diferents entre si (estrats) que posseeixen gran homogeneïtat pel que fa a alguna característica (es pot estratificar, per exemple, segons la professió, el municipi de residència, el sexe, l'estat civil, etc.). El que es vol amb aquesta classe de mostratge és assegurar-se que tots els estrats d'interès estaran representats adequadament en la mostra. En la figura I.0.5 s'ha estratificat per sexe i edat.

Figura I.0.5. Estrats definits
Font: <https://www.questionpro.com/blog/es/como-hacer-un-muestreo-estratificado/>

La distribució de la mostra segons els diversos estrats s'anomena assignació i pot ser de diferents tipus: *a*) Assignació simple: a cada estrat li correspon igual nombre d'elements mostrals; *b*) Assignació proporcional: la distribució es fa d'acord amb el pes (mida) de la població en cada estrat; *c*) Assignació òptima: es té en compte la previsible dispersió dels resultats, de manera que es considera la proporció i la desviació típica. Té poca aplicació perquè la desviació no se sol conèixer.
A partir d'això, s'aplicaria la tècnica del mostratge aleatori simple per a cada estrat considerat.
Considerem l'exemple següent:[3]
Suposem que volem estudiar el rendiment acadèmic estudiantil en una certa província després de la implantació d'una nova llei educativa. Per a fer-ho, seleccionarem 600 estudiants.
Se sap que hi ha 10.000 xiquetes i xiquets escolaritzats distribuïts de la manera següent: 6.000 en col·legis públics, 3.000 en col·legis concertats i 1.000 en privats no concertats.
Volem que els tres estrats estiguen representats d'acord amb:

[3] Text adaptat d'ací:
<https://proyectodescartes.org/iCartesiLibri/materiales_didacticos/EstadisticaProbabilidadInferencia/Muestreo/2MuestreoProbabilisticoTiposdeMuestreo/index.html>

a) Assignació simple
b) Assignació proporcional
Solució:
Hi ha tres estrats que corresponen a col·legis públics, col·legis privats concertats i col·legis privats no concertats. Cal veure ara quina és la distribució de les i els 600 estudiants de la mostra segons l'assignació emprada en el mostratge.
a) Els tres estrats han de tenir el mateix nombre d'elements (en aquest cas 200) perquè és una assignació simple. Com se seleccionen aquests 200 individus en cada estrat? Emprant un mostratge aleatori simple.

b) Per a l'assignació proporcional
Hem de tenir en compte la proporció/la representació de cada estrat sobre la mida de la població $N = 10.000$. Per a fer-ho, fem els càlculs següents:

- Proporció de col·legis públics: 6.000/10.000 = 0,60
- Proporció de col·legis privats concertats: 3.000/10.000 = 0,30
- Proporció de col·legis privats no concertats: 1.000/10.000 = 0,10

Per a saber la mida de cada estrat en la mostra, es multiplica la proporció calculada de cada estrat per la grandària mostral.

- Col·legis públics: 0,60 × 600 = 360 subjectes
- Col·legis privats concertats: 0,30 × 600 = 180 subjectes
- Col·legis privats no concertats: 0,10 × 600 = 60 subjectes Fem la comprovació 360 + 180 + 60 = 600 estudiants. A tall de resum, es pot compondre la taula següent:

Estrat	*Població*	*Proporció*	*Mostra proporcional*
Col·legis públics	6.000	0,60	360
Col·legis privats concertats	3.000	0,30	180
Col·legis privats NO concertats	1.000	0,10	60

1.4. Mostratge aleatori per conglomerats

En el mostratge per conglomerats, la mostra és un grup d'elements de la població que formen una unitat, la qual denominem conglomerat. Les unitats hospitalàries, els departaments universitaris, els barris, els districtes, una caixa d'un producte determinat, etc., són conglomerats naturals. A vegades es poden emprar conglomerats no naturals com, per exemple, les marques de ganivet en ES1 o els grups d'estudiants per aules en ES3.
El mostratge per conglomerats consisteix a seleccionar aleatòriament un cert nombre de conglomerats (el necessari per a assolir la mida mostral establida) i investigar després tots els elements que pertanyen als conglomerats escollits. És útil especialment quan la població està dispersa perquè estalvia costos en temps, desplaçaments i econòmics.
La figura I.0.6 mostra un exemple en què se selecciona a l'atzar una mostra de quatre consultoris de salut (conglomerats) dels nou que hi ha en un districte.

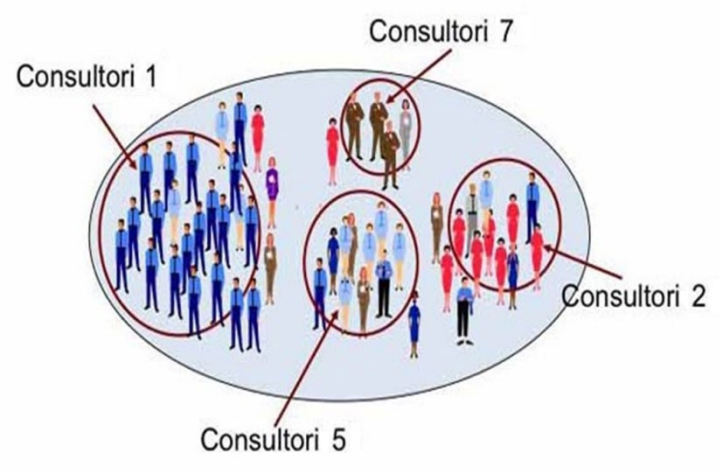


Figura I.0.6. Mostratge per conglomerats
Font: <https://bit.ly/2Y5KKYg>

2. Mostratge no probabilístic

Per a estudis exploratoris,[4] el mostratge probabilístic és excessivament costós i s'acudeix a mètodes no probabilístics. Aquests mètodes tenen el desavantatge que no serveixen per a fer generalitzacions (estimacions inferencials sobre la població) perquè, com que no tots els subjectes de la població tenen la mateixa probabilitat de ser escollits, no es té la certesa que la mostra extreta siga representativa. Tot i això, en general els subjectes se seleccionen seguint determinats criteris i procurant, en la mesura del possible, que la mostra siga representativa. Els mètodes més emprats són:

2.1. Mostratge per quotes

Es basa generalment en un bon coneixement dels estrats de la població o dels individus més *representatius* o *adequats* per als fins de la investigació. Manté, per tant, semblances amb el mostratge aleatori estratificat, però no té el caràcter d'aleatorietat d'aquell.

En aquesta classe de mostratge es fixen unes *quotes* que consisteixen en un nombre d'objectes que reuneixen unes determinades condicions, per exemple, en ES1: 20 ganivets fabricats a Espanya i amb mànec de fusta. En ES3: 30 estudiants entre 18 a 20 anys de la Facultat de Magisteri de la Universitat de València. Una vegada determinada la quota, es trien els primers que es troben que complisquen aquestes característiques.

2.2. Mostratge intencional o de conveniència

El personal investigador selecciona de manera directa i intencionada els individus de la població. El cas més freqüent d'aquest procediment empra com a mostra els individus als quals es té fàcil accés, per exemple, en ES1 podríem delimitar ganivets que s'usen en els restaurants d'un poble determinat, i en ES3 podríem delimitar les i els estudiants dels grups del docent que fa l'enquesta.

2.3. Bola de neu

[4] Es refereix a estudis en què l'objectiu és fer una primera aproximació al problema que es vol estudiar i conèixer. La investigació de tipus exploratori es fa a fi de conèixer la qüestió que es tractarà, cosa que ens permet *familiaritzar-nos* amb una cosa que fins al moment desconeixíem.

Es localitzen uns quants individus que ens condueixen a d'altres i aquests a uns altres, i així fins a aconseguir una mostra d'individus suficient.
Aquest mètode s'usa si la mostra per a l'estudi és petita o si té unes característiques que no són molt habituals. És una tècnica de mostratge que funciona en cadena. Per exemple, per a estudiar una determinada malaltia rara que només es presenta en una persona per milió. Un equip d'investigadors interessats a trobar-hi cura podria emprar aquesta classe de mostratge perquè no és fàcil trobar pacients a qui s'haja diagnosticat aquesta malaltia. Si les o els pacients pertanyen a alguna associació, pot ser que un pacient duga a un altre pacient, etc. D'aquesta manera, a través d'una cadena, un individu de la mostra porta als següents i així successivament.

2.4. Mostratge discrecional

A criteri del personal investigador, els elements es trien a partir d'allò que es pensa que poden aportar a l'estudi. S'empra quan un nombre molt limitat d'individus posseeix el tret que es vol estudiar. De vegades és l'única tècnica de mostratge viable per a obtenir informació d'un grup molt concret de persones.
Un exemple pot ser el personal que fa expedicions per la península Antàrtica.[5] No és fàcil trobar dades o investigadors que participen en aquesta mena d'expedicions o que duguen a terme aquestes investigacions.

Finalment, de segur que apareix el dubte sobre quina mida cal escollir perquè siga representativa: hi ha fórmules que la delimiten, però no és objecte d'aquest llibre; el nostre focus d'interès és l'aplicació de l'estadística en la futura docència de mestres, tant en el procés d'ensenyament i aprenentatge com en la millora d'aquest procés. Així, parlem de grups d'estudiants d'educació infantil o primària de com a màxim 25 alumnes (ràtio actual delimitada pel Ministeri d'Educació, Cultura i Esport (LOMCE)).

PRÀCTICA

Pr. I.0.4. Defineix la població del teu estudi.

Pr. I.0.5. Defineix la mostra del teu estudi i escull el mètode que empraràs.

Pr. I.0.6.[6] Suposem que, en un institut, els 330 alumnes d'ESO estan repartits en grups de la manera següent:

Sexe / Curs	*1r d'ESO*	*2n d'ESO*	*3r d'ESO*	*4t d'ESO*
Xics	46	50	36	28
Xiques	54	40	44	32

Hem de seleccionar-ne una mostra estratificada per sexe i nivell de mida 50 i de mida 120. Per a fer-ho, els elements corresponents a cada curs i sexe s'han de calcular de manera proporcional a la importància que tenen en tot el centre.

[5] L'estadística Ana Justel, investigadora de la Universitat Autònoma de Madrid, explica en l'enllaç següent quant costa aconseguir cada dada: <https://wp.bcamath.org/news/es/2018/05/30/ana-justel-ya-van-7-campan%cc%83as-en-la-antartica-y-tengo-muy-claro-lo-mucho-que-cuesta-un-dato/>

[6] Text pres i adaptat d'ací:
<http://recursostic.es/descartes/web/materiales_didacticos/muestreo_poblaciones_ccg/tipos_muestreo.htm>

I.2. Recull de dades

En aquest apartat delimitem la construcció de les eines que ens serviran per a recollir informació.

Els mètodes o les eines mitjançant els quals podem obtenir les dades necessàries són:

- **Cercar dades ja publicades per fonts governamentals industrials o individuals**. Ací cal fer esment de fonts com ara l'Institut Nacional d'Estadística (INE), que té una web on es pot obtenir diversa informació sobre dades d'Espanya (<https://www.ine.es/>) i l'Oficina d'Estadística de l'Ajuntament de València (<http://www.valencia.es/ayuntamiento/estadistica.nsf>). Una altra institució similar d'àmbit europeu és l'EUROSTAT, l'Oficina Estadística de la Unió Europea, (<https://ec.europa.eu/eurostat>). En l'àmbit internacional cal esmentar el Banc Mundial, que a través del seu web proporciona accés a les dades que gestiona (<https://data.worldbank.org/>). Un altre portal on podem buscar dades és Kaggle (<https://www.kaggle.com/>), que conté dades de tota classe; per exemple, dades sobre estadístiques educatives del Banc Mundial (<https://www.kaggle.com/theworldbank/education-statistics>).
  Sense importar la font emprada, cal distingir entre el recol·lector original de les dades i l'organització o els individus que les compilen en taules i diagrames. El recol·lector de dades (el banc de dades) és la font primària, mentre que el compilador de les dades (nosaltres) és la font secundària.

- **Dissenyar un experiment que ens proporcione les dades necessàries**
  Un segon mètode per a obtenir dades és l'experimentació. En un experiment[7] s'exerceix un control estricte sobre el tractament aplicat a les persones participants. Per exemple, en el cas de l'estudi ES4, es tracta de comprovar si la metodologia basada en projectes dona millors resultats en el procés d'ensenyament i aprenentatge que la metodologia tradicional. L'equip investigador determinaria quins participants de l'estudi seguirien una metodologia o l'altra i tractaria de delimitar perfectament cadascuna. Per a fer-ho caldria qüestionaris abans i després d'efectuar el procés d'ensenyament i aprenentatge.

- **Fer una enquesta**
  Un tercer mètode per a aconseguir dades és fer una enquesta. Simplement es formulen preguntes respecte a opinions, actituds, comportament i altres qüestions.
  Després, les respostes s'editen, es codifiquen i es tabulen a fi d'analitzar-les.
  Un exemple d'això seria la manera com s'aconsegueixen dades en les investigacions ES1 i ES2.

Per al segon i tercer mètode calen qüestionaris. Un qüestionari consisteix en un conjunt de preguntes respecte a una variable o més que es volen mesurar. El contingut de les preguntes d'un qüestionari pot ser tan variat com els aspectes que mesure. I bàsicament, podem parlar de dos tipus de preguntes: tancades i obertes.
Les preguntes tancades contenen categories o alternatives de respostes que han sigut delimitades. És a dir, es presenten als subjectes les possibilitats de respostes i ells han de circumscriure's a aquestes possibilitats. Poden ser dicotòmiques (dues alternatives de resposta)

[7] Un experiment es pot definir com una prova que consisteix a provocar un cert fenomen en unes condicions determinades per tal d'analitzar-ne els efectes o de verificar una hipòtesi o un principi científic.

o incloure diverses alternatives de resposta. Les preguntes obertes permeten a l'enquestat escriure lliurement sobre allò que la pregunta li requereix.

El disseny d'enquestes és habitual en l'àmbit de les ciències socials i parteix de la premissa que, si hi ha res que vulguem saber sobre el comportament de les persones, la millor manera, la més directa i simple, és preguntar-los-ho directament. Es tracta, per tant, de requerir informació a un grup socialment significatiu de persones sobre els problemes en estudi per tal que, mitjançant una anàlisi de tipus quantitatiu, es puguen extraure conclusions de les dades recollides. L'enquesta és un mètode de treball relativament econòmic i ràpid. Si es disposa d'un equip d'entrevistadors i codificadors convenientment entrenat, és fàcil arribar ràpidament a una multitud de persones i obtenir una gran quantitat de dades en poc de temps. El cost, per als casos simples, és sensiblement baix.

PRÀCTICA

Pr. I.0.7. Per parelles, tracteu de presentar exemples d'estudis que es poden resoldre amb enquestes i exemples d'estudis que es poden resoldre amb qüestionaris.

Pr. I.0.8. Determina quines preguntes de les següents seleccionaries per a una enquesta i quines per a un qüestionari.

1. Sexe

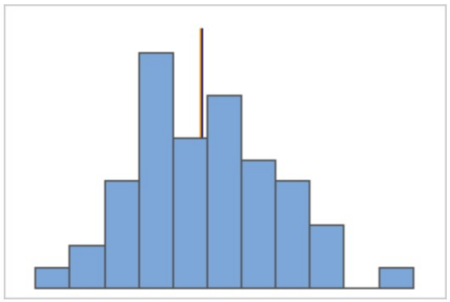
Home Dona Un altre

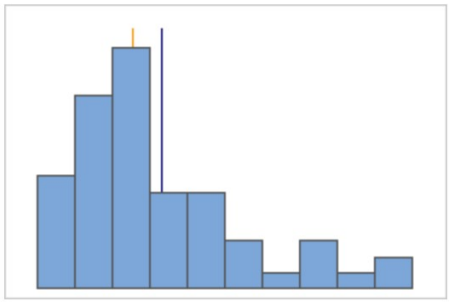
2. Edat: (resposta numèrica)

3. Quin dels termes següents expressa millor el significat següent?: "És una part o un grup representatiu d'una població". Tria l'opció que cregues que és correcta i que correspon a la definició indicada.

a) Mostra *b*) Recerca *c*) Individus *d*) Variable

4. En quin dels dos histogrames diries que hi ha presència de valors atípics? Indica l'opció que et semble correcta.





Font: <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/storedescriptive-statistics/interpret-the-statistics/interpret- the-statistics />

a) El de l'esquerra *b*) El de la dreta
Explica per què: _____

Pr. I.0.9. En el grup amb el qual fas el projecte, delimiteu com voleu recollir la informació.

I.2.1. Delimitem les preguntes

Ara és el moment de dissenyar les preguntes que s'adeqüen a l'objectiu previst. Ara bé, tot i que no és imprescindible, és molt útil fer un estudi previ sobre allò que es vol estudiar. Per exemple, en el cas d'ES3 i ES4, seria útil saber de quins coneixements es parteix. Com que tractem amb alumnes de grau, se sap que tots han passat per 4t d'educació secundària obligatòria i, per tant, hauríem d'anar al currículum i investigar sobre els coneixements tractats en l'àrea d'estadística.

Una vegada fet aquest estudi previ, ens centrem a formular les preguntes i, per a fer-ho, se segueixen els criteris següents amb l'objectiu de ser clars, no sols en el concepte, sinó també en la forma en què s'exposa:

- Claredat en les preguntes i llenguatge senzill.
- Cal evitar preguntes que puguen incitar determinades respostes.
- Preguntes tan curtes com siga possible.
- Les preguntes no s'han de redactar en forma negativa.
- Cal emprar un ordre lògic en les preguntes i que aquest ordre no condicione les respostes.
- No n'hi ha d'haver moltes.

PRÀCTICA

Pr. I.0.10. Dissenyeu les preguntes del vostre projecte.

I.2.2. Recollim dades

Tot i que hi ha diverses modalitats de recollida de dades, nosaltres ens centrarem únicament en dues: *a*) de forma personal; *b*) per correu electrònic o plataformes com ara Aula Virtual, Google Drive o similar.

La decisió d'escollir una forma o l'altra depèn del tema tractat, del temps, dels recursos econòmics i de la població objecte d'estudi.

Pel que fa a la modalitat personal, es pot observar que es controla el nombre de respostes, es pot cooperar amb les persones entrevistades i ajudar a resoldre problemes imprevistos, perquè l'entrevistador està present i permet evitar la influència d'agents externs ja siguen humans o material d'ajuda. No obstant això, hi ha inconvenients com el temps, ja que és un procés lent i costós, l'investigador ha de controlar el seu paper per tal de no modificar la resposta dels entrevistats i, finalment, a vegades és difícil accedir a certes poblacions.

Pel que fa a fer un qüestionari a través d'un giny electrònic, els desavantatges indicats es converteixen en punts positius, però no es poden controlar tots els avantatges esmentats.

PRÀCTICA

Pr. I.0.11. Selecciona la modalitat per a la investigació i adapta les preguntes fetes anteriorment. Després, recull les dades.

I.2.2. Organitzem les dades

Les dades es *porten* a una matriu de dades, com ara un full de càlcul, i a partir d'ací s'organitzen en taules de freqüències. Aquestes taules són diferents (en les primeres columnes) segons la natura de les variables que s'estudien, qualitatives o quantitatives, discretes o contínues. La figura I.0.7 mostra un exemple de variables quantitatives.

	A	B	C	D	E	F	G	H	I	J
1	**Revisión del Padrón municipal 2007. Datos a nivel nacional, comunidad autónoma y provincia.**									
2	**00.- Nacional**									
3										
4	**Población por edad (grupos quinquenales) y sexo**									
5	Unidades:Personas									
6										
7										
8		Ambos sexos	Varones	Mujeres	Varones	Mujeres			Varones	Mujeres
9	Total	45.200.737	22.339.962	22.860.775	-22.339.962	22.860.775		0-4	-1.152.780	1.084.747
10	0-4	2.237.527	1.152.780	1.084.747	-1.152.780	1.084.747		05-09	-1.087.410	1.027.248
11	05-09	2.114.658	1.087.410	1.027.248	-1.087.410	1.027.248		10-14	-1.092.802	1.035.845
12	10-14	2.128.647	1.092.802	1.035.845	-1.092.802	1.035.845		15-19	-1.198.595	1.132.282
13	15-19	2.330.877	1.198.595	1.132.282	-1.198.595	1.132.282		20-24	-1.457.797	1.397.048
14	20-24	2.854.845	1.457.797	1.397.048	-1.457.797	1.397.048		25-29	-1.926.676	1.809.995
15	25-29	3.736.671	1.926.676	1.809.995	-1.926.676	1.809.995		30-34	-2.084.538	1.937.683
16	30-34	4.022.221	2.084.538	1.937.683	-2.084.538	1.937.683		35-39	-1.959.174	1.847.382
17	35-39	3.806.556	1.959.174	1.847.382	-1.959.174	1.847.382		40-44	-1.832.087	1.774.602
18	40-44	3.606.689	1.832.087	1.774.602	-1.832.087	1.774.602		45-49	-1.636.907	1.622.680
19	45-49	3.259.587	1.636.907	1.622.680	-1.636.907	1.622.680		50-54	-1.375.080	1.387.797
20	50-54	2.762.877	1.375.080	1.387.797	-1.375.080	1.387.797		55-59	-1.242.564	1.284.711
21	55-59	2.527.275	1.242.564	1.284.711	-1.242.564	1.284.711		60-64	-1.103.584	1.176.897
22	60-64	2.280.481	1.103.584	1.176.897	-1.103.584	1.176.897		65-69	-869.799	969.665
23	65-69	1.839.464	869.799	969.665	-869.799	969.665		70-74	-903.141	1.090.612
24	70-74	1.993.753	903.141	1.090.612	-903.141	1.090.612		75-79	-702.414	949.641
25	75-79	1.652.055	702.414	949.641	-702.414	949.641		80-84	-444.042	717.031
26	80-84	1.161.073	444.042	717.031	-444.042	717.031		85 y más	-270.572	614.909
27	85 y más	885.481	270.572	614.909	-270.572	614.909				

Figura I.0.7. Padró municipal del 2007
Font: Institut Nacional d'Estadística

L'estadística tracta d'organitzar dades. El primer que s'ha de fer quan ens donen dades és tractar de resumir la informació a través de taules de freqüències que tenen diverses columnes. Així doncs, per a les variables qualitatives o quantitatives discretes, les primeres columnes han de ser de la manera següent:

1. Si delimitem, per exemple, la variable X = sexe d'ES3, que és variable qualitativa nominal dicotòmica i que té els valors (x_1 = dona, x_2 = home). Escriurem X_i o x_i per a fer referència als diversos valors de la variable X que s'estudia.

x_i
dona
home

2. Si delimitem, per exemple, la variable edat d'ES3 com a variable quantitativa discreta:

x_i
19
20
21
22
24
27

En el cas de les variables quantitatives contínues, s'empra la distribució de freqüències agrupades o taula amb dades agrupades en intervals. Els valors s'agrupen en intervals que tinguen la mateixa amplitud anomenats classes. A cada classe s'assigna la freqüència corresponent. Així, el primer que s'ha de delimitar és el nombre d'intervals per considerar; és recomanable que estiga entre 6 i 15. Cal fer avinent que, com menys quantitat d'intervals, si el rang (valor màxim - valor mínim) és molt gran, es perd molta informació de la mostra.

Després s'ha de calcular l'amplitud de l'interval o amplitud de la classe (I), el rang ja calculat s'incrementa fins a arribar a un nombre divisible pel nombre d'intervals triat. Finalment, es formen els intervals tenint present que el límit inferior d'una classe pertany a l'interval, però que el límit superior no pertany a l'interval, sinó que es compta en l'interval següent.

3. Si delimitem, per exemple, la variable edat d'ES3 com a variable quantitativa contínua.
Edat màxima = 27
Edat mínima = 19
Rang = valor màxim - valor mínim = 27 - 19 = 8→8→I = 8/6 = 1,33≈2
Les classes són: [19, 21 [; [21, 23 [; [23, 25 [; [25, 27 [; [27, 29 [; [29, 31 [

classes	x_i
[19, 21 [	20
[21, 23 [	22
[23, 25 [	24
[25, 27 [	26
[27, 29 [	28
[29, 31 [	30

En general, la taula de freqüències és una ordenació en forma de taula de les dades estadístiques en què s'assigna a cada dada la freqüència corresponent. Hi ha quatre tipus de freqüències:

- *Freqüències absolutes.* Les freqüències absolutes (*fa*) d'una variable es defineixen com el nombre de vegades que un determinat valor apareix en un estudi estadístic. La suma d'aquestes freqüències és igual al nombre total de dades (*N*).
- *Freqüències relatives.* D'altra banda, les freqüències relatives (*fr*) es defineixen com el quocient entre la freqüència absoluta d'un valor determinat (*fa*) i el nombre total de dades (*N*).
- *Freqüència acumulada.* Es defineix (*Fa*) com la suma de les freqüències absolutes de tots els valors inferiors o iguals al valor considerat.
- *Freqüència relativa acumulada.* La freqüència relativa acumulada (*Fr*) és el quocient entre la freqüència acumulada (*Fa*) d'un determinat valor i el nombre total de dades (*N*).

Suposem que tenim les dades sobre sexe i edat de 82 persones. Les 82 persones tenen una edat compresa entre 19 i 27 anys.

Dades sexe (D, dona; H, home "D" "D" "D" "D" "D" "D" "D" "H" "H""D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "H" "H" "H" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "D" "H" "H" "H" "H" "H"

Dades edat: 19 21 19 19 19 24 20 20 20 20 24 20 19 19 19 19 19 19 19 19 19 19 19 19 19 20 20 20 2019 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 19 20 20 20 20 19 19 19 19 19 19 19 19 19 19 19 19 19 19 20 20 20 20

20 20 20 20 21 21 27

A fi de no tenir 82 dades desordenades, les resumim en una taula amb les columnes *fa*, *fr*, *Fa* i *Fr*

Les taules anteriorment iniciades quedarien de la manera següent,

1. Per a la variable *sexe*, farem un recompte de les vegades que es repeteixen les categories dona (D) i home (H) i a partir d'ací es construeix la taula.

x_i	*fa*	*fr*	*Fa*	*fr*
dona	72	0,878	72	0,878
home	10	0,122	82	1
	$N = 82$	1		

2. Per a la variable *edat* considerada quantitativa discreta.

x_i	*fa*	*fr*	*Fa*	*fr*
19	51	0,622	51	0,622
20	21	0,256	72	0,878
21	3	0,037	75	0,915
22	4	0049	79	0,963
24	2	0,024	81	0,988
27	1	0,012	82	1
	82	1		

3. Per a la variable *edat* considerada quantitativa contínua (ací s'han de determinar intervals)

classes	x_i	*fa*	*fr*	*Fa*	*fr*
[19, 21 [	20	72	0,878	72	0,878
[21, 23 [	22	7	0,085	79	0,963
[23, 25 [	24	2	0,024	81	0,988
[25, 27 [	26	0	0,000	81	0,988
[27, 29 [	28	1	0,012	82	1,000
[29, 31 [	30	0	0,000	82	1
		82	1		

PRÀCTICA

Pr. I.0.12. Totes les variables de la teua recerca s'han d'organitzar en taules. Recorda que, segons la naturalesa de les variables, tenen un perfil o altre.

I.3. Ús de programari que facilite l'organització de dades i l'anàlisi posterior

Abans hem ressenyat el full de càlcul com a matriu de dades on es posen totes i cadascuna de les dades obtingudes. Però sovint, en el cas que es tinguen dades, amb un simple paper i llapis ja en tindrem prou per a organitzar la informació.

Val a dir que hi ha programes estadístics que ens faciliten tant l'organització de dades com l'anàlisi posterior. El més popular d'aquests programes és l'SPSS, però requereix una llicència i el pagament corresponent. En aquesta línia, hi ha altres programes gratuïts com, per exemple, l'R-Commander. I, finalment, hi ha el mateix full de càlcul, tant en Linux com en Microsoft, que ens permet fer estadística al nivell que s'emmarca en aquest llibre.

Ací explicarem el full de càlcul.

I.3.1. El full de càlcul per a organitzar taules de freqüències

Per a crear taules de freqüència de variables qualitatives o quantitatives discretes, hem d'emplenar la primera columna (xi), on s'inclouen totes les categories (k) de la variable objecte d'estudi.

A continuació hem d'emplenar la columna de freqüències absolutes, després del recompte de cadascuna de les dades en la categoria pertinent. La casella blava és N.

xi	***fa***	***fr***	***Fa***	***Fr***
Categoria 1	n_1	$= n_1/N$	$= n_1$	$= n_1/N$
Categoria 2	n_2	$= n_2/N$	$= n_{1+}\, n_2$	$= n_1/N + n_2/N$
Categoria 3	n_3	$= n_3/N$	$= n_{1+}\, n_{2+}\, n_3$	$= n_1/N + n_2/N + n_3/N$
…	…	…	…	…
…	…	…	…	…
Categoria k	n_k	$= nk/N$	$= n_{1+}\, n_{2+}\, n_3 + \ldots + n_k$	$= n_1/N + n_2/N + n_3/N + n_4/N$
	$=$ suma $(n_1, n_2, n_3, \ldots, n_k)$	1	…	…

I.3.1. El full de càlcul per a crear taules amb dades agrupades

En aquest cas, igual que hem explicat més amunt, l'única diferència amb la taula de freqüències és la construcció de la columna classe (Ci), i després d'això de la marca de classe (xi), la resta, les freqüències, es calculen igual que en la taula de freqüències.

Així, ens centrem a mostrar les dues primeres columnes d'aquesta nova taula.

Ci	***xi***
$[l_1, l_2[$	$= (l_1 + l_2)/2$
$[l_2, l_3[$	$= (l_2 + l_3)/2$
$[l_3, l_4[$	$= (l_3 + l_4)/2$
…	…
…	…
$[l_{k-1}, l_k[$	$= (l_{k-1} + l_k)/2$

Capítol II. Cuinem les dades[8]

“La creativitat no arriba en minuts ni en hores: la creativitat arriba en el moment que ha d’arribar”
Ferran Adrià

El segon capítol fa referència a la mateixa mà d’obra, “Cuinem les dades”, una vegada sabem com cal recollir dades i organitzar-les, procés detallat en el capítol I, ja podem disposar-nos a planificar la nostra recepta i començar amb les mescles més bàsiques.

Així doncs, en aquest capítol centrarem l’atenció en el tractament de la informació de manera descriptiva, és a dir, analitzarem cada variable del nostre estudi tant de forma gràfica com numèrica.

A fi facilitar la comprensió de la part teòrica, s’hi inclou l’anàlisi descriptiva d’un estudi que es va fer el curs 2019-2020 amb alumnes de segon del Grau de Mestres en Educació Infantil i Primària. Aquest estudi tenia per objecte avaluar els coneixements previs dels alumnes en el camp de l’estadística. Els participants havien de contestar un qüestionari amb 15 preguntes el detall de les quals està disponible en l’annex II.1.

A més a més, es pot veure el vídeo desat al repositori del canal de YouTube del SFPIE de la Universitat de València (<https://www.youtube.com/watch?v=uKANAmbA1AE>). En aquest vídeo, que presenta un dels millors projectes desenvolupats per l’alumnat realitzat en la pràctica docent que s’inclou en aquest llibre, es fa una explicació de l’estadística descriptiva (aquest capítol) i inferencial (capítol II) emprada per a analitzar les dades obtingudes en el projecte. L’objectiu era analitzar el perfil de l’alumnat que cursa el Grau de Mestre/a en Educació Infantil i Primària en relació amb la via d’accés al grau i l’opció / prioritat d’elecció del grau a l’hora de presentar la sol·licitud d’entrada i matrícula a la universitat.

Contingut

II.1. Anàlisi gràfica de dades

Una vegada construïda la taula o matriu de dades, s'ha d'explorar a fi de buscar-hi generalitats o informació atípica o anormal. En aquest primer apartat es tracta de resumir la informació en gràfics. Els gràfics ens permeten presentar la informació que donen les dades de manera resumida i gràfica, fàcil d'entendre. A més, permeten la detecció de possibles errors i dades atípiques que poden formar part de la mateixa mostra o que s'hagen produït per mala gestió (transcripció o trasllat de la informació) de les dades.

Els gràfics poden ser univariats, bivariats i multivariats, segons el nombre de variables que tinguen. Ara, doncs, detallarem els gràfics univariats per tipus de variable, ja que en aquest apartat es vol fer un resum de cada variable estudiada per separat. En el capítol III presentarem els gràfics bivariats i multivariats, gràfics que donen compte de les relacions entre les variables de l'estudi.

II.1.1. Variables qualitatives

Pel que fa a la representació gràfica de les variables qualitatives, destaquem dos tipus de gràfics perquè són els que s'usen més sovint.

II.1.1.1. Diagrama de sectors

El primer, el diagrama de sectors, s'usa per a visualitzar de forma senzilla les freqüències relatives de les variables. En els gràfics de sectors es divideix una figura, habitualment de forma circular, de manera que l'àrea que correspon a cada possible resposta de la variable és proporcional a la freqüència relativa de la variable. Però el més complicat, si volem representar-lo amb llapis i paper, és com s'obté aquesta proporcionalitat. Tracem el diagrama a partir de les dades de la variable *Sexe* del qüestionari.

Per a la variable *Sexe* tenim la taula de freqüències absolutes següent:

x_i	fa
dona	72
home	10
	$N = 82$

Així doncs, la mida mostral, 82, representa tota la forma circular, amb la qual cosa es pot formar la raó següent:

$$\frac{82\ estudiants}{360^o}$$

Per a saber quina part de la forma circular correspon a les dones, cal establir la proporció següent:

$$\frac{82\ estudiants}{360^o} = \frac{72\ estudiants}{x^o}$$

D'ací s'obté que per als homes $x \approx 316^\circ\ 360^\circ - 316^\circ = 44^\circ$
Aquestes dades es representen amb l'ajuda del transportador d'angles.

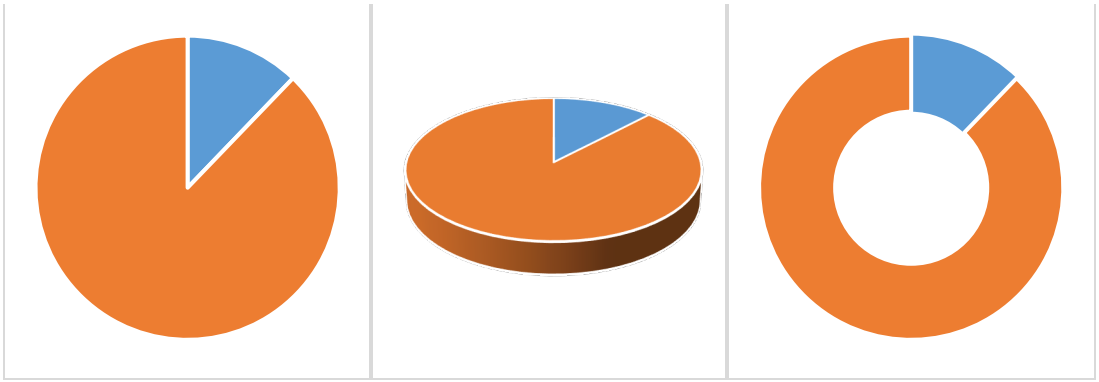
Si decidim usar un programa informàtic, aquesta representació es pot adornar amb etiquetes a l'interior o a l'exterior del gràfic. A més, sol ser habitual incloure per a cada categoria de la

variable la freqüència relativa (o, si es vol, absoluta de la variable). En qualsevol cas, tots aquests ornaments de la representació, com també altres detalls (color, forma del gràfic...) són complements que faciliten la visualització dels resultats i que els programes habituals d'estadística solen incorporar. L'elecció de com s'ha de personalitzar aquesta classe de gràfics és una decisió personal que depèn del detall que volem que incloga la representació final.

En l'annex II.1 tenim, per exemple, la pregunta **1. Sexe**, que es podria representar amb un diagrama de sectors perquè és una variable qualitativa. Així, amb l'opció de full de càlcul amb l'Excel, si marquem l'opció següent:

Insereix → Gràfics recomanats → Tots els gràfics → Circular

Tenim els diagrames circulars següents, on *Home* apareix de color blau i *Dona* de color taronja.



II.1.1.2. Gràfic de barres

El segon tipus de representacions gràfiques que tractem són els gràfics de barres. En aquest tipus de gràfic es representa una barra vertical (o horitzontal si es prefereix) per a cadascuna de les categories de la variable d'altura proporcional a la freqüència, bé absoluta o relativa.

En aquest cas, quan es tracta de fer un gràfic amb llapis i paper, només s'ha de determinar quina freqüència s'ha de representar en l'eix d'ordenades (eix Y), i cal tenir en compte que les barres verticals no poden estar unides perquè no es tracta d'una variable contínua (vegeu més avant la diferència amb l'histograma).

Igual que els diagrames de sectors, els gràfics de barres se solen personalitzar a gust de l'usuari de manera que la configuració siga tan il·lustrativa com siga possible. Els gràfics de barres solen ser preferibles als diagrames de sectors perquè, segons s'ha pogut comprovar, l'ull humà està particularment entrenat per a comparar longituds i no per a comparar àrees. Però atesa la popularitat d'aquests últims gràfics en la bibliografia, cal saber interpretar-los i ser conscients del possible ús d'aquests gràfics.

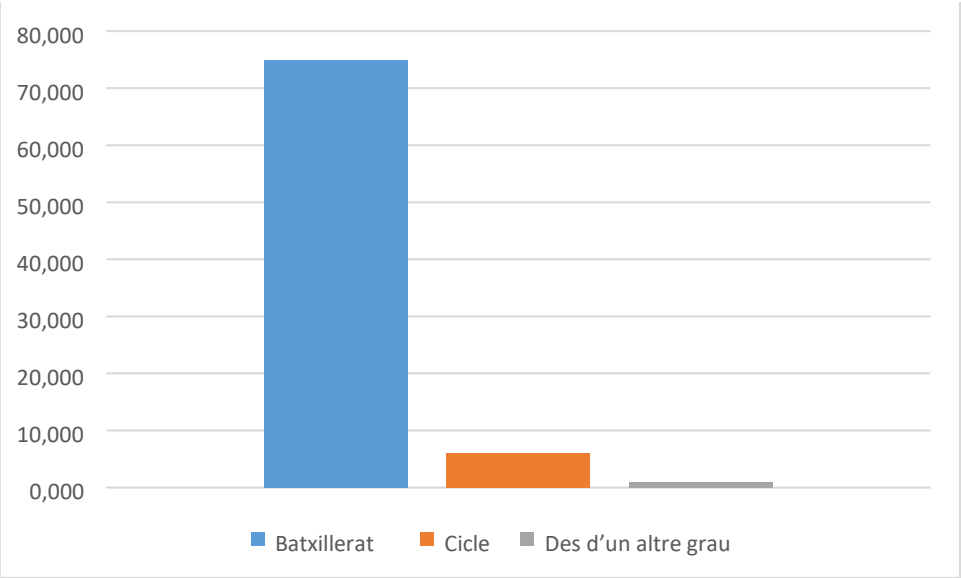
En aquest cas, representarem una altra variable qualitativa del qüestionari que estem utilitzant a manera d'exemple, **3. Accés a la Facultat de Magisteri**.

Per a la variable *Accés* tenim la taula de freqüències absolutes següent:

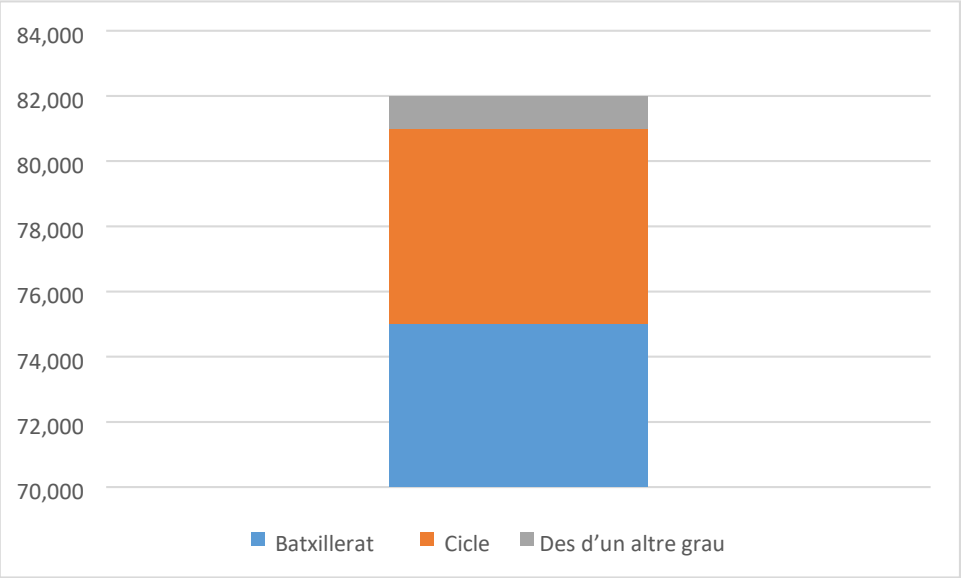
x_i	fa
Batxillerat	75,000
Cicle	6,000
Des d'un altre grau	1,000
	$N = 82$

Així, amb l'opció de full de càlcul amb l'Excel, si marquem l'opció següent: Insereix → Gràfics recomanat → Tots els gràfics → Columna

I dins d'aquesta tenim *columna agrupada*



O *columnes apilades*



La diferència entre columna agrupada i columnes apilades és que en la primera es veu clarament el nombre d'individus que es té de cadascuna de les categories definides per a la variable, però no passa igual amb la segona, que es podria considerar una variant del diagrama circular però en forma rectangular, i, així mateix, és més interessant quan es representen freqüències relatives.

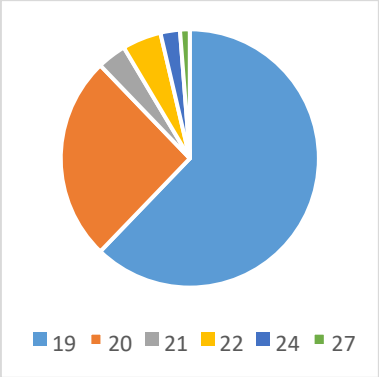
PRÀCTICA

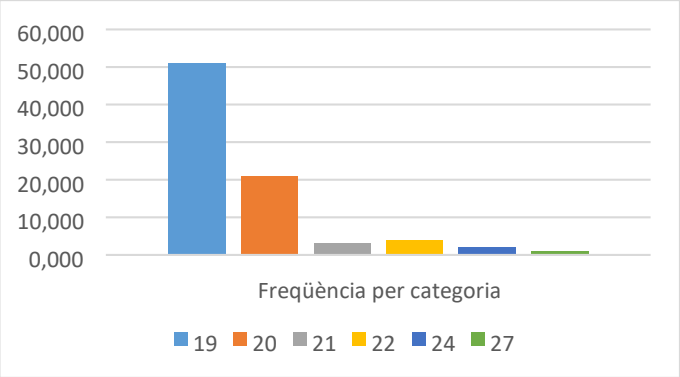
Pr. II.0.1. Representa les variables qualitatives del teu projecte tant en un diagrama circular o diagrama de barres amb columnes apilades, com en un diagrama de barres amb columna agrupada. A quina conclusió arribes en vista dels gràfics?

I.1.2. Variables quantitatives

Les variables quantitatives es poden classificar en discretes i en contínues. Per al primer cas, els gràfics poden ser els mateixos que els que s'han explicat per a les variables qualitatives perquè cada marca de classe es pot considerar una categoria, per la no continuïtat destacada en les quantitatives discretes.

Per al cas de la variable *Edat* del qüestionari, que tant es pot considerar discreta com contínua, presentem a continuació els dos gràfics associats en el cas de ser considerada discreta.

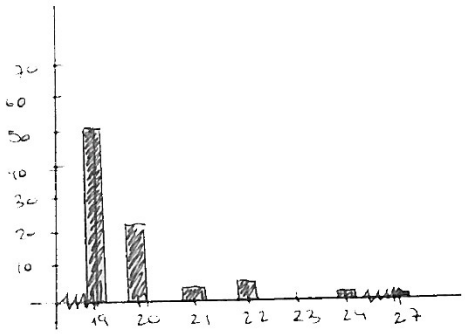




En el cas del diagrama de barres, cal tenir en compte les consideracions següents:

- No es presenta un gràfic que comença per 0 en l'eix d'abscisses, que representa la recta real.
- No es presenta la divisió de la recta real per les marques de classe considerades.

Així, quan tracem el gràfic amb llapis i paper, cal tenir presents les consideracions anteriors i ens queda un gràfic com el següent:



Per al cas de les variables quantitatives contínues, es presenten a continuació els dos gràfics corresponents.

PRÀCTICA

Pr. II.0.2. Representa les variables quantitatives discretes del teu projecte en un diagrama de barres amb columna agrupada. A quina conclusió arribes en vista dels gràfics?

II.1.2.1. Histograma

L'histograma aglutina les dades de cada marca de classe i es representa de la mateixa manera com s'ha fet en el diagrama de barres, però tenint en compte la continuïtat d'aquestes variables,

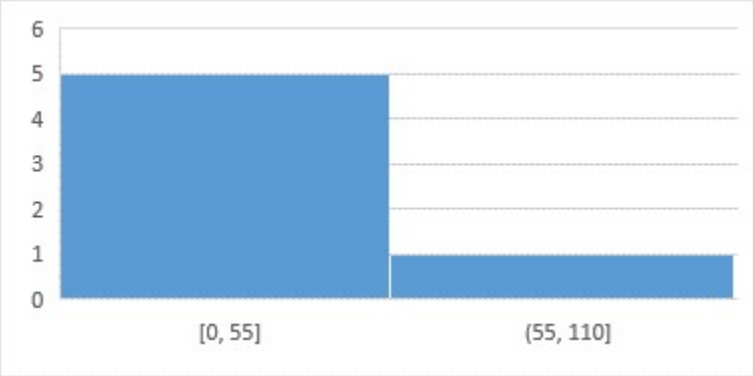
cosa que implica que les columnes verticals estan totes unides. L'eix d'ordenades (o eix Y) representa les freqüències absolutes o relatives; i l'eix d'abscisses (o eix X) representa la recta real, que és delimitada per cada marca de classe.
Per a la variable *Edat*, considerada quantitativa contínua, tenim la taula de freqüències absolutes següent:

classes	x_i	fa
[19, 21 [	20	72
[21, 23 [	22	7
[23, 25 [	24	2
[25, 27 [	26	0
[27, 29 [	28	1
[29, 31 [	30	0
		$N = 82$

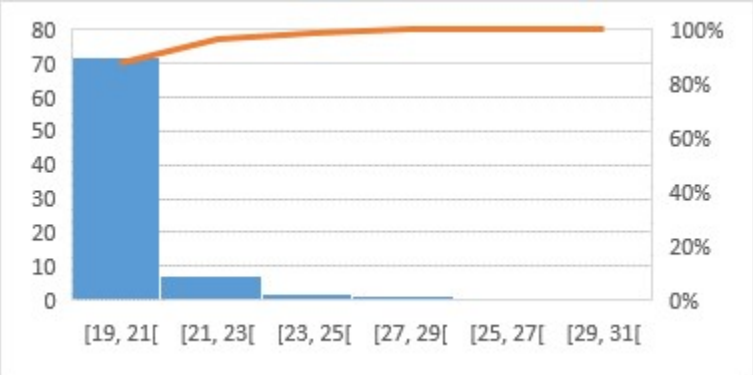
Ací cal fer esment que en l'eina del full de càlcul:

Insereix → Gràfics recomanats → Tots els gràfics → Histograma

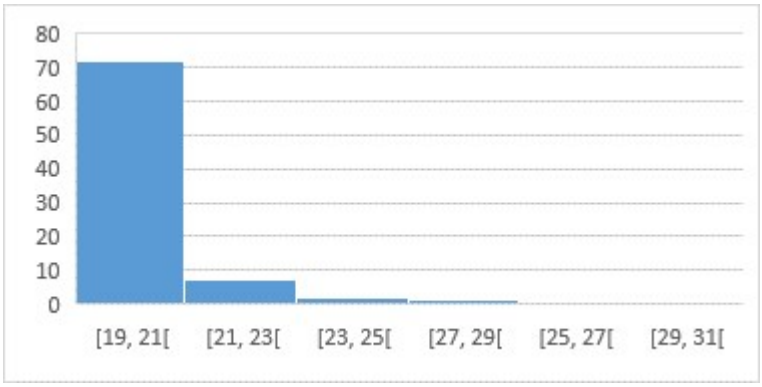
Hi ha dues opcions, *histograma* o *Pareto*. Si triem l'opció *histograma*, observem que no es respecten les nostres categories per interval:



No obstant això, si escollim l'opció de Pareto, tenim el gràfic següent:

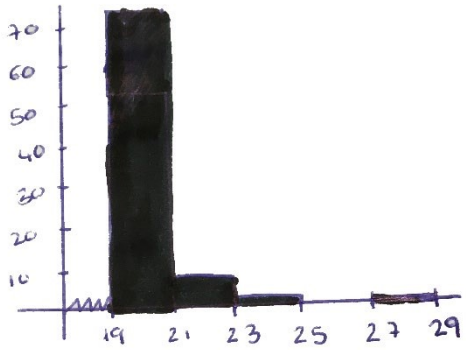


del qual es pot eliminar l'eix de la dreta i la línia superior i queda de la manera següent:

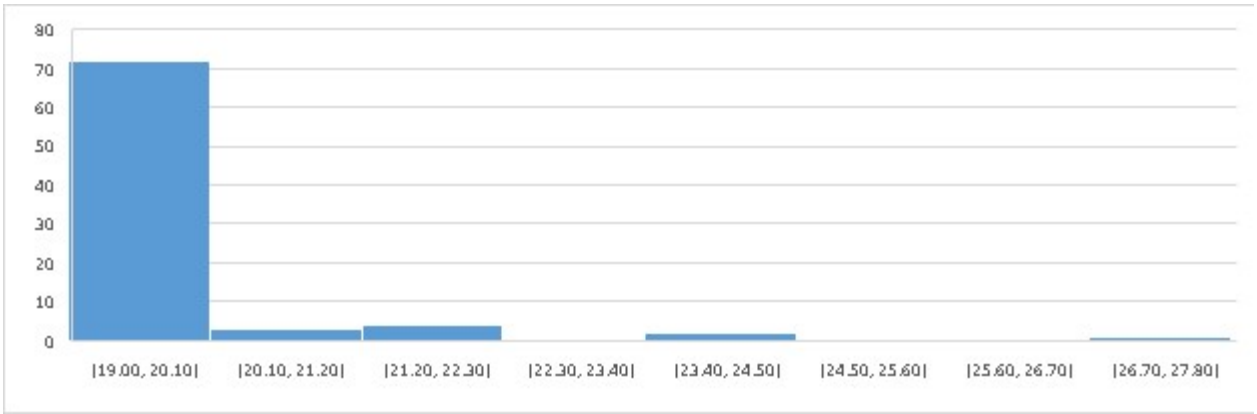


En aquests gràfics cal tenir en compte les consideracions següents:

- No es presenta un gràfic que comença pel 0 en l'eix d'abscisses, que representa la recta real.
- No es presenta la divisió de la recta real pels intervals considerats.



Hi ha la possibilitat que el mateix full de càlcul cree l'histograma, però si seleccionem les dades completes i no les taules de freqüències construïdes. Val a dir que en aquesta opció no podem controlar el nombre d'intervals generats i, així mateix, cal tenir present el que hem dit més amunt sobre l'inici en 0 i la divisió de la recta real.



II.1.2.1. Diagrama de caixes

Així i tot, si busquem una representació encara més esquemàtica de com es distribueixen les dades, podem optar pel diagrama de caixes. A la part central d'aquest diagrama apareix una caixa en què els extrems estan delimitats pel primer i tercer quartil, mentre que la mitjana apareix com una línia que divideix la caixa anterior. Al seu torn, els denominats bigots de la caixa apareixen units per un segment que travessa la caixa anterior i que permet fer-se una idea aproximada de la franja de les dades. Hi ha diversos criteris per a representar els bigots, però el que estudiarem en aquest curs és el que es detalla a continuació: el bigot inferior representa o bé

1,5 vegades el rang interquartílic (RIC) per davall del primer quartil o bé el valor mínim si aquest no és un valor atípic; i el bigot superior representa o bé 1,5 vegades el RIC per damunt del tercer quartil o bé el valor màxim si aquest no és un valor atípic. Si hi ha valors atípics en el conjunt de dades, es representen mitjançant punts aïllats fora del diagrama.

Però què són els valors atípics?

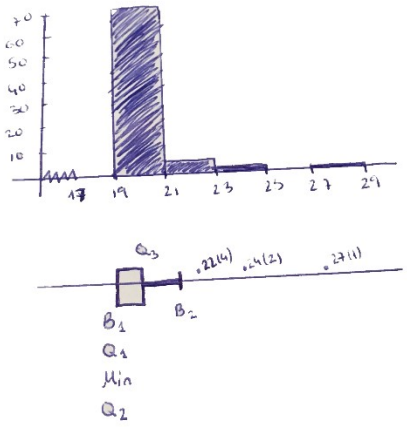
Un valor més extrem (*outlier* o atípic) és un valor en un conjunt de dades que és molt diferent dels altres valors. És a dir, els valors atípics són valors excepcionalment allunyats del centre. En la majoria dels casos tenen influència en la mitjana, però no en la mediana ni en la moda. Per tant, els valors atípics són importants per l'efecte que tenen en la mitjana.

No hi ha una regla per a identificar els valors atípics. Però alguns llibres consideren que un valor és atípic si és més gran que 1,5 vegades el valor del rang interquartílic més enllà dels quartils, és a dir, fora de la franja determinada pels bigots.

Tots els càlculs de les diverses mesures es detallen en l'apartat següent. Per tant, considerant ja conegudes les dades següents:

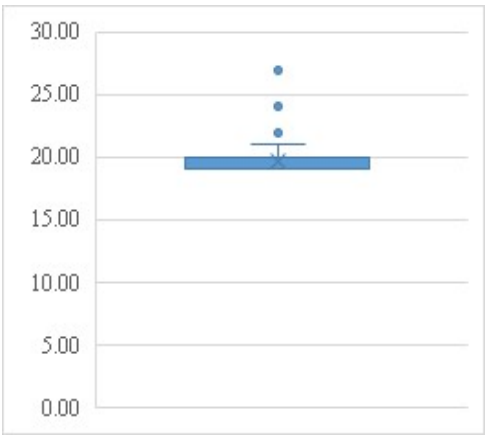
$$\text{mínim} = 19; \quad \text{màxim} = 27; \quad Q1 = 19; \quad Q2 = 19; \quad Q3 = 20$$
$$\text{mitjana} = 19{,}695; \quad \text{RIC} = 1; \quad bigot1 = 19; \quad bigot2 = 21{,}5$$

Ací simplement mostrem el resultat gràfic d'aquests valors:

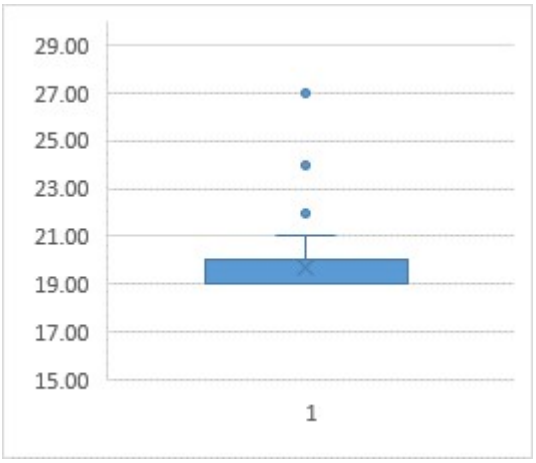


A més, si fem servir el full de càlcul, hem de seleccionar totes les dades (no les taules de freqüències construïdes) i,

Insereix → Gràfics recomanats → Tots els gràfics → Caixes i bigots



A fi de tractar de veure millor la informació, i com que sabem que les dades comencen en 19 anys, modifiquem el gràfic perquè la informació de l'eix d'ordenades (o eix Y) comence per 15, per exemple.



PRÀCTICA

Pr. II.0.3. Representa les variables quantitatives contínues del teu projecte en un histograma i en un diagrama de caixes. A quines conclusions arribes en vista dels gràfics?

II.2. Anàlisi numèrica de dades

II.2.1. Variables qualitatives

Per a les variables quantitatives, únicament s'hi pot incloure, a més de les freqüències obtingudes en la taula de freqüències, la moda (mesura de centralització).

La moda d'una variable és el valor que es repeteix més vegades, el valor que té més freqüència absoluta. Quan la variable que volem estudiar a penes pren valors repetits, aquest estadístic és de poca utilitat (quan la variable en estudi és quantitativa contínua se sol parlar de l'interval o rang que més valors conté com la moda).

II.2.2. Variables quantitatives

Per a resumir numèricament variables de tipus quantitatiu tenim un ventall d'eines bastant ampli. Una vegada creades les taules de freqüència, s'han de calcular mesures de resum específiques d'aquesta classe de variables i que, a grans trets, es poden classificar de la manera següent:

- Mesures de centralització
  Resumeixen la localització al voltant de la qual es distribueixen les dades. En aquest curs tractarem la mitjana, la moda i la mediana.
- Mesures d'ordre o posició
  Informen sobre diverses característiques de les dades a partir de l'ordenació dels valors observats. Les mesures d'ordre que estudiarem són els percentils i quartils.
- Mesures de dispersió

Resumeixen la variabilitat que presenten les dades al voltant d'algun dels estadístics de centralització. Estudiarem com a mesures de dispersió el rang, el rang interquartílic (RIC), la variància i la desviació típica.

- Mesures de forma
  Informen sobre el comportament de la distribució de les dades (simetria). En aquest curs les comentarem únicament en l'àmbit gràfic.

II.2.2.1. Mesures de centralització

Les mesures de centralització ens informen sobre la localització al voltant de la qual se situen els valors de la variable en estudi. Hi ha diversos estadístics que ens informen sobre aquest valor, entre els quals destaquem:

- Mitjana aritmètica
  La mitjana aritmètica o simplement mitjana d'un conjunt de mesures és la mesura de tendència central més emprada i coneguda. Aquesta mesura se simbolitza per $\bar{x}$ quan representa la mitjana mostral i per μ (lletra grega minúscula) quan representa la mitjana poblacional, és a dir, de tota la població, $\bar{x}$.
  Es calcula com la suma de tots els valors de la mostra (o població) dividits pel nombre de casos. La fórmula per a calcular-la és la següent:

$$\frac{\sum_{i=1}^{N} x_i}{N}$$

  Però en el cas que tinguem la taula de freqüències i coneguem totes les vegades que es repeteix cada x_i, llavors la fórmula és:

$\frac{\sum_{i=1}^{k} x_i \cdot fa_i}{N}$, en que *k* es refereix a el nombre de categories.

  Per al cas de les variables quantitatives contínues representa la marca de classe x_i.
  Com a exemple, prenem *Edat* del qüestionari amb el qual treballem:

$$\bar{x} = \frac{19 \cdot 51 + 20 \cdot 21 + 21 \cdot 3 + 22 \cdot 4 + 24 \cdot 2 + 27 \cdot 1}{82} = 19.695\ anys$$

- Mediana
  La mediana, *M*, d'un conjunt de valors és el valor que hi ha al punt mitjà o centre, una vegada s'han ordenat de més petits a més grans.
  Si els mesuraments d'un conjunt de dades s'ordenen de menys a més valor i *N* és imparell, la mediana correspon al mesurament amb l'ordre (*N* + 1) / 2. Si el nombre de mesuraments és parell, *N* = parell, la mediana s'escull com el valor de *M* a la meitat de les dues mesures centrals, és a dir com el valor central entre el mesurament amb rang *N* / 2 i el que té rang (*N* / 2) + 1.
  Quan el conjunt de dades és petit, les dades es poden ordenar i comptar directament. En el cas d'una mostra gran, la taula de freqüències (*Fa*) ens indica aquesta ordenació.
  Segons l'exemple que portem entre mans, tenim *N* = 82. Com que es tracta d'un nombre parell, hem de buscar els valors que es troben en la posició, 82/2 = 41, i la posició 82/2 + 1 = 41 + 1 = 42.
  Així, vegem en primer lloc com es calcularia en el cas de considerar-la variable quantitativa discreta i després en el cas de considerar-la contínua. Com que les taules de freqüència són diferents, haurem de veure com cal actuar en cada situació.

- Quantitativa discreta
  La taula següent és la que correspon a la variable:

xi	*fa*	*fr*	*Fa*
19	51	0,622	51
20	21	0,256	72
21	3	0,037	75
22	4	0,049	79
24	2	0,024	81
27	1	0,012	82
	82	1	

  L'element que hi ha en la posició 41, i també en la 42, és el 19, ja que la *Fa* indica que hi ha 51 elements que corresponen a l'edat 19 anys. Per tant:

$$M = \frac{19 + 19}{2} = 19 \text{ anys}$$

- Quantitativa contínua
  La taula següent correspon a la variable:

classes	*xi*	*fa*	*fr*	*Fa*
[19, 21 [	20	72	0878	72
[21, 23 [	22	7	0,085	79
[23, 25 [	24	2	0,024	81
[25, 27 [	26	0	0,000	81
[27, 29 [	28	1	0,012	82
[29, 31 [	30	0	0,000	82
		82	1	

  En aquest cas, els elements que hi ha en les posicions 41 i 42 són el 20, ja que en aquest cas tenim 72 individus amb edats compreses entre 19 i 21 anys i, com per a fer els càlculs no puc prendre un interval, prenc la marca de classe que és 20 anys. Així, la mediana és la següent:

$$M = \frac{20 + 20}{2} = 20 \text{ anys}$$

Observacions

1. La mitjana és molt sensible a l'existència de valors extrems de la variable (particularment alts i baixos): com que totes les observacions intervenen en el càlcul de la mitjana, l'aparició d'una observació extrema fa que la mitjana es desplace en aquesta direcció.

2. Si considerem una variable discreta, per exemple el nombre de fills en les famílies de la ciutat de València, el valor de la mitjana pot no pertànyer al conjunt de valors possibles de la variable; Per exemple, x = 2,5 fills per família.

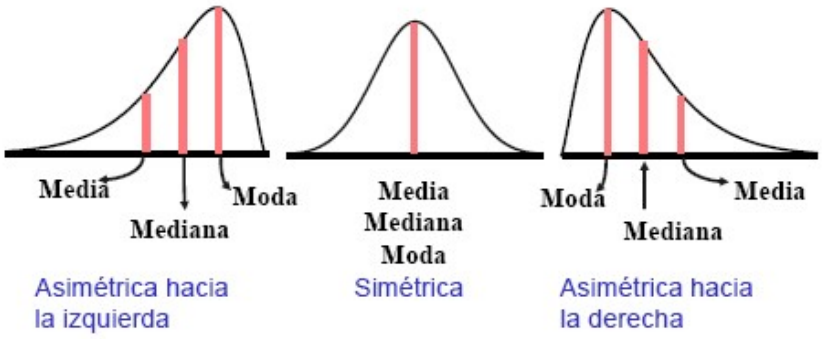
Les tres mesures de tendència central, la mitjana, la mediana i la moda, no són útils per igual per a obtenir una mesura de tendència central. Per contra, cadascuna d'aquestes mesures té característiques que fan que usar-les siga avantatjós en certes condicions però no en d'altres.

La mediana sol ser la mesura preferida quan es fa servir una escala ordinal, que són les situacions en què el valor assignat a cada cas no té cap més significat que indicar l'ordre entre els casos. Per exemple, saber en una classe quins alumnes resten dins del 50% amb millors notes i quins dins del 50% amb pitjors notes.

També se sol preferir la mediana quan uns pocs valors extrems distorsionen el valor de la mitjana (observació número 1 anterior). Per exemple, si tinc nou alumnes que han tret un 7 en un examen de matemàtiques i només un alumne ha tret un 1, la mediana em pot donar a entendre que la majoria té una nota de 6,4, però això no és real.

La moda en certes condicions pot ser la més apropiada, per exemple quan es vol informació ràpida i quan la precisió no és un factor especialment important. En certs casos només aquesta mesura té sentit, per exemple en un equip de futbol que porte l'estadística per jugador (escala ordinal) de la quantitat de passades que fa per partit, això per a detectar quin jugador és el millor distribuint la pilota. En aquest cas la mitjana i la mediana no tindrien significat, només la moda.

Un aspecte interessant de les tres mesures és el comportament respecte a la simetria que pren una distribució. Quan les distribucions són simètriques, sense biaix, el cas de la distribució normal que té forma de campana, "la mitjana, la mediana i la moda coincideixen". Si la distribució és asimètrica amb biaix positiu, hi ha més dades cap a l'esquerra de la mitjana, llavors "la mitjana és més gran que la mediana i aquesta és més gran que la moda". Si passa el contrari, el biaix és negatiu, llavors "la mitjana és més petita que la mediana i aquesta és més petita que la moda".



Font: <https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwiv4qruuJfmAhUEExoKHYiQB1cQjB16BAgBEAM&url=http%3A%2F%2Fanasep96.blogspot.com%2F&psig=AOvVaw377PvH_PexRaCPalP90d-& ust = 1575392586940344>

PRÀCTICA

Pr. II.0.4. Calcula les mesures de centralització per a les variables del teu projecte i elabora conclusions sobre el resultat.

II.2.2.2. Mesures d'ordre o posició

Aquestes mesures indiquen, com expressa el nom, l'ordre o la posició d'una observació entre els valors d'una variable quantitativa. Per a calcular aquestes mesures hem d'ordenar de forma ascendent els valors de la mostra, tal com s'ha fet per al càlcul de la mitjana, que es pot considerar una mesura de posició.

- Mínim i màxim
  Són el valor més petit i més alt, respectivament, del conjunt de dades.
- Percentil
  El percentil és el valor que compleix que el p% de les observacions de la mostra són inferiors a ell (i per tant els altres són superiors).
  Els percentils al 25%, 50% i 75% reben noms concrets a causa de la importància que tenen (i es representen per $P25 = Q1$, $P50 = Q2 = M$ i $P75 = Q3$). Aquests percentils es denominen quartils (primer, segon i tercer quartil respectivament) perquè divideixen la mostra en quatre parts de la mateixa mida. Si hi reflexionem un poc, podem adonar-nos que ja hem definit el segon quartil, perquè aquest estadístic no és més que el valor que és superior al 50% de les observacions de la variable, i aquesta propietat era la condició que havia de complir necessàriament la mediana. Per tant, quan parlem del percentil al 50%, del segon quartil o de la mediana d'una variable, ens referim exactament a la mateixa quantitat.
  Per a fer el càlcul d'aquests estadístics, actuem de manera anàloga a l'operació feta en el cas de la mediana, ara percentil 50 o quartil 2. L'única cosa que canvia és que per a aquests casos s'ha d'obtenir la posició de l'element que es trobe en el p% que indique el que es vulga calcular.
  Per exemple, per a calcular el valor del $Q1$, que es tracta del $P25$, fem l'operació següent:

$$\frac{100\%}{82} = \frac{25\%}{x} \rightarrow x = 20{,}5$$

Així, es calcula com el valor que es troba en la posició 20 més el que està en la posició 21 i es divideix entre 2.

Per al cas discret: $P_{25} = \frac{19+19}{2}$ = 19 anys

Per al cas continu: $P_{25} = \frac{20+20}{2}$ = 20 anys

Per al cas del $Q3$, que es tracta del $P75$, es calcula de la manera següent:

$$\frac{100\%}{82} = \frac{75\%}{x} \rightarrow x = 61{,}5$$

Així, es calcula com el valor que es troba en la posició 61 més el que està en la posició 62 i es divideix entre 2.

Per al cas discret: $P_{75} = \frac{20+20}{2}$ = 20 anys

Per al cas continu: $P_{75} = \frac{20+20}{2} = 20$ anys

PRÀCTICA

Pr. II.0.5. Calcula les mesures d'ordre per a les variables del teu projecte i elabora conclusions sobre el resultat.

II.2.2.3. Mesures de dispersió

Els estadístics de dispersió en general ens informen sobre la variabilitat de les dades, és a dir si són més disperses o, per contra, si s'agrupen de forma més o menys precisa al voltant d'un cert valor.

Algunes mesures de dispersió importants són les següents:

- Rang
  El rang és la diferència entre el màxim i el mínim valor de la variable. És la mesura de dispersió més fàcil de calcular, però també és la més inestable perquè està molt influïda per valors extrems atípics. Com més gran és el rang, més gran és la dispersió de les dades d'una distribució. És adequada per a mesurar la variació de conjunts petits de dades.
- Rang interquartílic
  El rang interquartílic es defineix com la diferència entre el tercer i primer quartil.

$$RIQ = Q3 - QQ1$$
$$RIQ = P75 - P25$$

  El principal avantatge que té el rang interquartílic respecte al rang és que aquest últim se sol veure bastant afectat per la presència de qualsevol valor anòmal (valor atípic o *outlier*), mentre que el rang interquartílic és bastant menys sensible a aquesta mena d'observacions. Per tant, a vegades sol ser preferible usar el rang interquartílic en lloc de la franja com a mesura de dispersió de les dades.

- Desviació típica
  La desviació típica pot ser mostral, s, o poblacional, σ. Resumeix la distància que sol haver-hi entre cada observació i la mitjana. En el càlcul d'aquesta mesura de dispersió, a diferència del rang i del rang interquartílic, en què únicament s'inclouen dues observacions (o bé el màxim i mínim o bé el primer i tercer quartil), intervenen tots i cadascun dels valors.
  És la mesura de dispersió més àmpliament usada i la més estable perquè depèn de tots els valors de la distribució.
  S'interpreta com la distància que es desvia de la mitjana un conjunt de valors. Aquest valor es representa gràficament com un interval $\bar{x} \pm s$.
  Es calcula mitjançant l'expressió següent:

$$s = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot fa_i}{N-1}},$$ en què k es refereix al nombre de categories.

$$s = \frac{(19-19{,}7)^2 \cdot 51 + (20-19{,}7)^2 \cdot 21 + (21-19{,}7)^2 \cdot 3 + (22-19{,}7)^2 \cdot 4 + (24-19{,}7)^2 \cdot 2 + (27-19{,}7)^2 \cdot 1}{82-1}$$
$$= 1.33\ anys$$

Per tant, es considera que la mitjana es desvia 1,33 anys, l'interval següent [19,7-1,33; 19,7 + 1,33], és a dir, que la mitjana oscil·la entre els valors [18,37; 21,03].

- Variància

  La variància és el quadrat de la desviació típica. La interpretació no és tan clara com en el cas de la desviació típica; simplement hem de conèixer quins valors més grans de la variància corresponen a mostres que tenen més variabilitat.

  $$s^2 = 1{,}77$$

- Coeficient de variació

  El coeficient de variació és una mesura de dispersió que es defineix com el quocient entre la desviació típica i la mitjana multiplicat per 100.

$$CV = \frac{s}{\bar{x}} \cdot 100$$

Aquest coeficient permet comparar la variabilitat de diverses mostres en una mateixa variable o la variabilitat que hi ha entre variables diferents. Per tant, el farem servir en el capítol III.

La justificació d'aquest indicador és que habitualment les variables amb valors més grans (la mitjana és més gran) són també les variables amb més dispersió (la desviació típica és més gran). Quan calculem el quocient de la desviació típica i la mitjana, estem anul·lant aquest efecte i, per tant, el coeficient de variació ens permet comparar la variabilitat de variables mesurades en escales o unitats diferents.

PRÀCTICA

Pr. II.0.6. Calcula les mesures de dispersió per a les variables del teu projecte i elabora conclusions sobre els resultats.

Pr. II.0.7. Quan s'analitzen les notes (sobre 100) d'individus de dues mostres diferents de la mateixa mida, s'obté que la mitjana i desviació per a ambdues són 44,5 i 14,9, respectivament. Assenyala l'opció que et semble correcta.

a) Només és possible si les dades (les respostes dels individus de les dues mostres) són exactament les mateixes.
b) Això és possible encara que les dades no siguen les mateixes.
c) Cap de les anteriors.

Pr. II.0.8. Quan s'analitzen les notes (sobre 100) d'individus de dues mostres diferents de la mateixa mida, s'obté que la mitjana per a ambdues és 44,5, però les desviacions típiques són 6 i 14,9, respectivament. Indica l'opció que et semble correcta.

a) Això indica que les notes de la primera mostra, que té una desviació típica de 6, estan poc disperses i més pròximes al valor de la mitjana. No obstant això, per a l'altre cas, desviació de 14,9, les notes estan més disperses.

b) Això indica que els individus no han estudiat molt i que en el primer grup hi algú que no ha arribat a 0,6 (6/100); i que en el segon hi ha algú que no ha arribat a 1,49 (14,9 / 100).

c) La diferència en les desviacions típiques no em dona informació rellevant perquè el que ens interessa és la mitjana, i ja sabem que és la mateixa, cosa que indica que les notes no han sigut molt bones.

Pr. II.0.9. La figura següent presenta el nombre de persones que hi ha en cada interval d'edat. La mida i la mitjana tenen el mateix valor per a les dues mostres, però la desviació típica és diferent. Els histogrames són diferents perquè... Indica l'opció que et semble correcta.

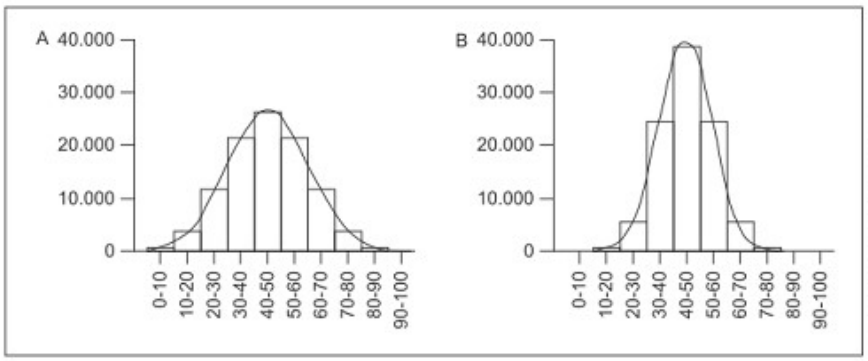
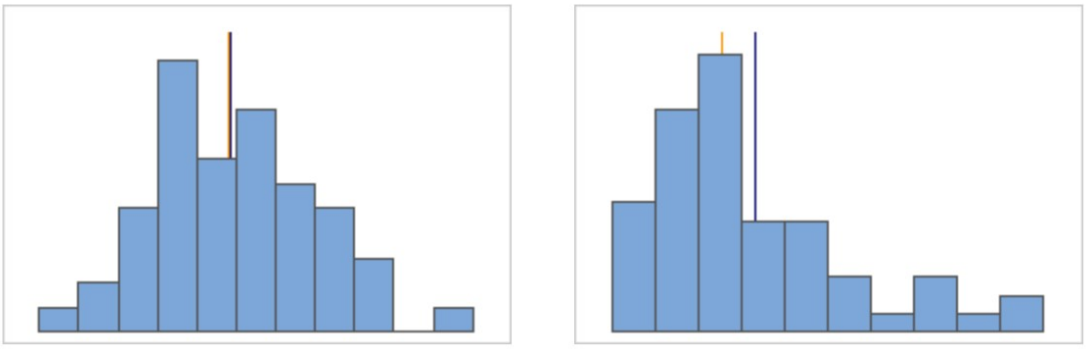


Figura 1. Histogramas de frecuencias de la edad en unas poblaciones hipotéticas de 100.000 individuos. En ambas poblaciones la media es 44,5; las desviaciones estándar son 14,9 en A y 10,0 en B. Las líneas continuas muestran la distribuciones teóricas correspondientes a curvas gaussianas con las mismas medias y desviaciones estándar.

Font: <http://formacion.intef.es/pluginfile.php/49844/mod_imscp/content/4/desviacin_tpica_o_desviacin_estndar.html>

a) Aquests gràfics són impossibles; haurien d'haver eixit iguals.
b) Els gràfics són possibles i indiquen que la població B té més dispersió.
c) Els gràfics són possibles i indiquen que la població A té més dispersió.

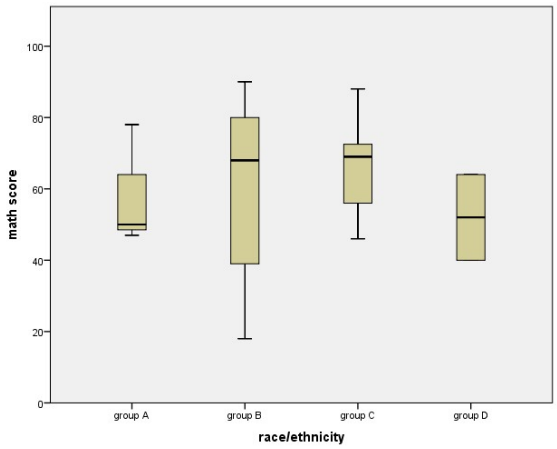
Pr. II.0.10. En quin dels dos histogrames diries que hi ha presència de valors atípics? Assenyala l'opció que et semble correcta.



Font: <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/store-descriptive-statistics/interpret-thestatistics/interpret- the-statistics />

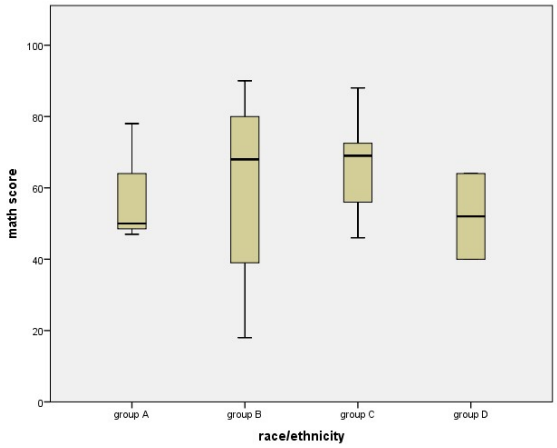
a) El de l'esquerra
b) El de la dreta

Pr. II.0.11. El gràfic següent de caixes i bigots mostra informació sobre les notes de matemàtiques (de 0 a 100) de quatre grups diferents. Pel que fa a la mitjana... assenyala l'opció que et semble correcta.



a) No hi ha diferències entre les mitjanes dels quatre grups.
b) El grup B té una mitjana molt més alta que els altres grups.
c) En aquesta classe de gràfics no es pot determinar res sobre la mitjana.

Pr. II.0.12. El gràfic de caixes i bigots següent mostra informació sobre les notes de matemàtiques (de 0 a 100) de quatre grups diferents. Pel que fa als valors extrems, atípics i de dispersió, assenyala l'opció que et semble correcta.



a) El grup B té dades atípiques perquè els extrems estan més allunyats.
b) No hi ha valors atípics en cap grup, però sí que es pot ressaltar que el grup B ha tingut més dispersió en les notes perquè els valors extrems estan allunyats.
c) Els grups A, C i D són els millors perquè no tenen valors extrems allunyats.

Annex II. 1

L'enquesta següent tracta sobre coneixements estadístics

SOCIODEMOGRÀFICA

1. Sexe
Home
Dona
Un altre

2. Edat

3. Via d'accés a la Facultat de Magisteri
Batxillerat
Cicle
Des d'un altre grau

3.1. Indica la modalitat de batxillerat
Arts
Ciències Socials
Ciències
Humanitats

Indica la família professional
Activitats físiques i esportives
Fabricació mecànica
Química
Administració i gestió
Hostaleria i turisme
Sanitat agrària
Imatge personal
Seguretat i medi ambient
Arts gràfiques
Imatge i so
Serveis socioculturals i a la comunitat
Arts i artesanies
Indústries alimentàries
Tèxtil, confecció i pell
Comerç i màrqueting
Informàtica i comunicacions
Transport i manteniment de vehicles
Edificació i obra civil
Instal·lació i manteniment
Vidre i ceràmica
Electricitat i electrònica
Fusta, moble i suro
Energia i aigua
Maritimopesquera

CONCEPTES TEÒRICS

4. Els conceptes de mostra i població signifiquen el mateix per a tu? Indica l'opció que et semble correcta.

Sí

No

5. Quin dels termes següents expressa millor el significat següent?: "És una part o un grup representatiu d'una població". Tria l'opció que cregues que és correcta i que correspon a la definició indicada.

a) Mostra

b) Recerca

c) Individus

d) Variable

6. Quina de les mesures següents representa millor el conjunt d'un conjunt de dades d'una mostra? S'obté a partir de la suma de tots els valors dividida entre la mida de la mostra.

a) La mitjana mostral

b) La moda

c) La mediana

d) El percentatge

7. És una mesura de dispersió per a variables quantitatives que expressa la variació existent entre les dades, de gran utilitat en estadística descriptiva. Indica l'opció que et semble correcta.

a) Mesures de tendència central

b) Desviació estàndard

c) Variància

d) Coeficient de variació

8. Els punts o valors *outlier*, també denominats atípics o estranys, dins d'una mostra... indica l'opció que et semble correcta.

a) No modifiquen els meus valors de mitjana i desviació estàndard en cap cas.

b) Sí que els considere; modifiquen el valor de la mitjana, però no el de la desviació estàndard.

c) Sí que els considere; modifiquen tant el valor de la mitjana com el de la desviació estàndard.

d) Depèn del context del problema per al qual s'han obtingut les dades. Per tant, el fet de considerar-los o no depèn de qui soluciona el problema.

e) Totes / cap de les anteriors.

9. Els punts o valors *outlier*... Indica l'opció que et semble correcta.

a) Sempre han de ser eliminats de la mostra perquè no la representen.

b) No s'han d'eliminar mai de la mostra. Els representem en els gràfics, però no per al càlcul de la mitjana i la desviació estàndard.

c) No s'han d'eliminar de la mostra perquè la variabilitat de les dades és representada per ells. Per tant, emprarem altres estadístics per a tractar d'analitzar la mostra donada.

d) Caldria estudiar el comportament de la mostra tant en els casos en què sí que es consideren com en els casos en què no es consideren i analitzar-ne les conseqüències.

CONCEPTES AMB LA PRÀCTICA

10. Quan analitzem les notes (sobre 100) d'individus de dues mostres diferents de la mateixa mida, s'obté que la mitjana i desviació per a ambdues són 44,5 i 14,9, respectivament. Indica l'opció que et semble correcta.

a) Només és possible si les dades (les respostes dels individus de les dues mostres) són exactament les mateixes.

b) Això és possible, encara que les dades no siguen les mateixes.

c) Cap de les anteriors.

11. Quan analitzem les notes (sobre 100) d'individus de dues mostres diferents de la mateixa mida, s'obté que la mitjana per a ambdues és de 44,5, però les desviacions típiques són 6 i 14,9, respectivament. Indica l'opció que et semble correcta.

a) Això indica que les notes de la primera mostra, que té una desviació típica de 6, estan poc disperses i més pròximes al valor de la mitjana. No obstant això, per a l'altre cas, desviació de 14,9, les notes estan més disperses.

b) Això indica que els individus no han estudiat molt, que en el primer grup hi algú que no ha arribat a 0,6 (6/100) i que en el segon algú no ha arribat a 1,49 (14.9 / 100).

c) La diferència en les desviacions típiques no em dona informació rellevant perquè el que ens interessa és la mitjana, i ja sabem que és la mateixa, cosa que indica que les notes no han sigut molt bones.

12. La figura següent presenta el nombre de persones que hi ha en cada interval d'edat. La mida i la mitjana tenen el mateix valor per a les dues mostres, però la desviació típica és diferent. Els histogrames són diferents perquè... Indica l'opció que et semble correcta.

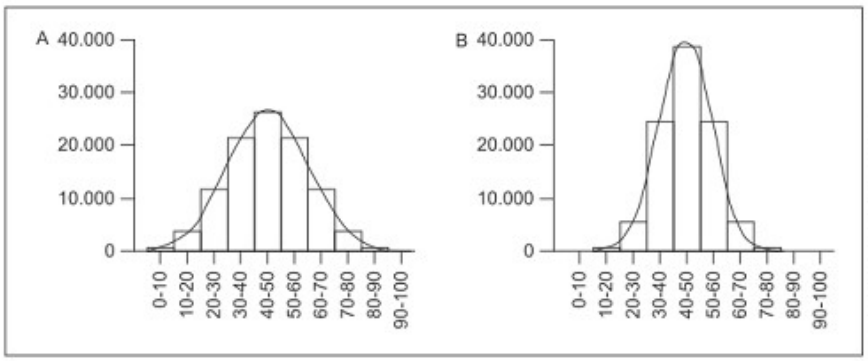


Figura 1. Histogramas de frecuencias de la edad en unas poblaciones hipotéticas de 100.000 individuos. En ambas poblaciones la media es 44,5; las desviaciones estándar son 14,9 en A y 10,0 en B. Las líneas continuas muestran la distribuciones teóricas correspondientes a curvas gaussianas con las mismas medias y desviaciones estándar.

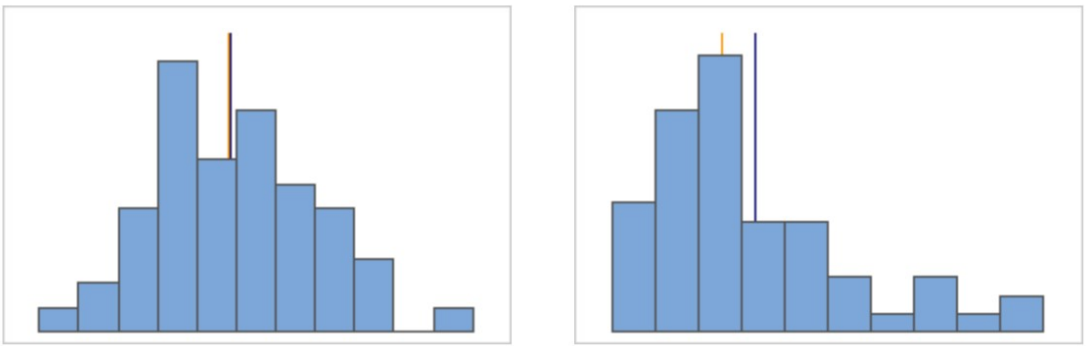
Font: <http://formacion.intef.es/pluginfile.php/49844/mod_imscp/content/4/desviacin_tpica_o_desviacin_estndar.html>

a) Aquests gràfics són impossibles; haurien d'haver eixit iguals.

b) Els gràfics són possibles i indiquen que la població B té més dispersió.

c) Els gràfics són possibles i indiquen que la població A té més dispersió.

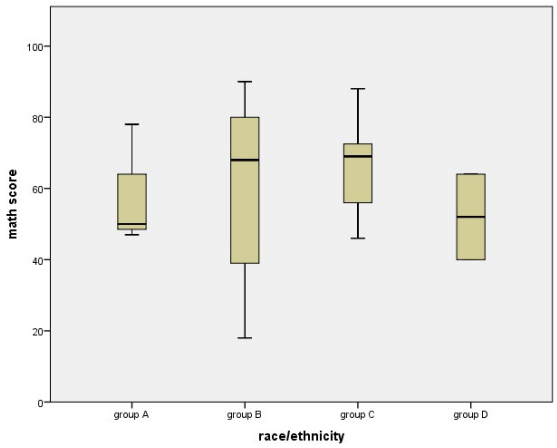
13. En quin dels dos histogrames diries que hi ha presència de valors atípics? Indica l'opció que et semble correcta.



Font: <https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/store-descriptive-statistics/interpret-the-statistics/interpret- the-statistics />

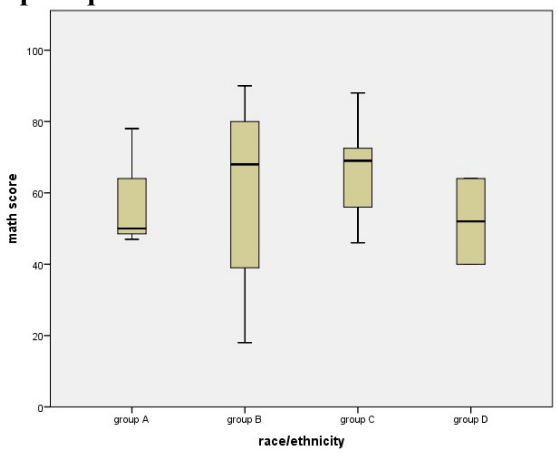
a) El de l'esquerra
b) El de la dreta

14. El gràfic de caixes i bigots següent mostra informació sobre les notes de matemàtiques (0 a 100) de quatre grups diferents. Pel que fa a la mitjana... Indica l'opció que et semble correcta.



a) No hi ha diferències entre les mitjanes dels quatre grups.
b) El grup B té una mitjana molt més alta que els altres grups.
c) En aquest tipus de gràfics no es pot determinar res sobre la mitjana.

15. El gràfic de caixes i bigots següent mostra informació sobre les notes de matemàtiques (0 a 100) de quatre grups diferents. Pel que fa als valors extrems, atípics i dispersió, indica l'opció que et semble correcta.



a) El grup B té dades atípiques perquè els extrems estan més allunyats.
b) No hi ha valors atípics en cap grup, però sí que es pot ressaltar que el grup B ha tingut més dispersió en les notes perquè els valors extrems estan allunyats.
c) Els grups A, C i D són els millors perquè no tenen valors extrems allunyats.

Capítol III. Comparem les nostres receptes[9]

"La cuina és per a divertir-se: un fet cultural i, també, un espai lúdic on podem jugar i interpretar"
Joan Roca

Finalment, en el capítol 3 comparem les receptes. Per a poder comparar bones receptes, han d'estar cuidades al detall i reconèixer fins a la més mínima diferència per a identificar quina és la que despunta sobre les altres. Així doncs, en aquest capítol analitzarem amb detall tot el que s'ha obtingut abans i tractarem de fer comparacions, cosa que es denomina estadística inferencial.

El que s'ha estudiat fins ara han sigut diverses característiques o variables d'una mostra extreta d'una determinada població. Què més es pot fer amb aquestes dades? La resposta és clara, es podria investigar si hi ha relacions entre les variables estudiades de la mostra.

Si continuem amb l'exemple del capítol II (vegeu l'enquesta en l'annex II.1) es podria determinar si hi ha relació entre el sexe i respondre correctament a la tasca 4. O si el tipus d'accés a la facultat de Magisteri pot estar relacionat amb més puntuació (entenent per puntuació posar un 1 a totes les respostes correctes i sumar).

Ja sabem que podem relacionar quan tenim una mostra, però l'anàlisi inferencial únicament es basa en això? La resposta és no.

Com a futurs docents, us podria interessar observar si una certa metodologia millora els resultats d'un determinat coneixement. Per exemple, si l'enquesta de l'annex II.1 es passara a l'alumnat al principi de curs i al final, i a grups diferents, els uns amb una metodologia tradicional i els altres amb metodologia basada en projectes, es podria observar d'una banda si cada grup ha millorat i en quina mesura i, de l'altra, es podrien relacionar els grups a fi de veure quin ha obtingut millors resultats i, per tant, poder concloure quina metodologia funciona millor. En l'enllaç següent hi ha un exemple visual d'una situació d'aula https://www.youtube.com/watch?v=uK8c0pu38mw, on una alumna de 4t curs del Grau de Mestre en Educació Primària exposa una investigació extreta del seu treball final de grau, en què detalla la manera com es pot aplicar l'estadística per a determinar si una metodologia implementada en una aula és efectiva o no.

III.1. Anàlisi inferencial, una mostra

Ja s'ha estudiat que, dins d'una mostra, les variables poden ser variables qualitatives o quantitatives. Així, es pot comprovar la hipòtesi de si hi ha relació (són dependents) entre dues variables qualitatives (sexe i accés a la Facultat de Magisteri), entre dues de quantitatives (edat i puntuació de l'enquesta) o entre una variable qualitativa i una de quantitativa (sexe i puntuació de l'enquesta).

Així, en aquesta part determinarem cadascuna de les relacions esmentades des de la perspectiva gràfica i des de la perspectiva numèrica. Atès que ja tenim coneixements sobre estadística descriptiva, que són la base per a emprendre aquesta mena d'estudis, procedim a fer els gràfics i els càlculs amb un full de càlcul, en la mesura que es puga i que faciliten la tasca a l'estudiantat.

III.1.1. Dues variables qualitatives

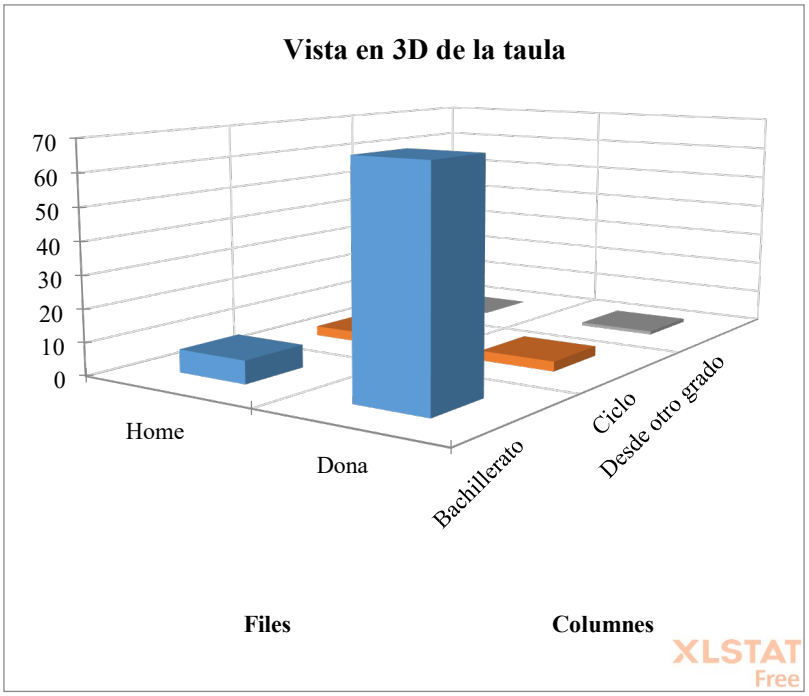
Per a emprendre aquesta anàlisi, es procedeix a construir una taula de contingència. En la primera fila hi ha les categories d'una de les variables qualitatives per estudiar (accés a la Facultat de Magisteri) i la primera columna té les categories de l'altra variable (sexe). Introduïm en les caselles la freqüència absoluta de cada situació. Per exemple, a la casella (2, 2) s'ha de d'introduir el nombre de persones enquestades que són home i que han contestat *batxillerat*; en la casella (4,3) s'escriu el nombre d'enquestades dones i que hi han accedit des d'un altre grau.

	Batxillerat	*Cicle*	*Des d'un altre grau*	*Total*
Home	7			
Dona			1	
Total				

Una vegada completada la taula de contingència amb les freqüències absolutes, els totals generats es calculen a partir de la suma de les dades d'una fila o d'una columna.

	Batxillerat	*Cicle*	*Des d'un altre grau*	*Total*
Home	7	3	0	10
Dona	68	3	1	72
Total	75	6	1	82

La representació dels valors permet observar gràficament que dones i batxillerat són les més habituals, però no podem apreciar l'existència o no de relació entre les variables, tret que ens fixem en les distribucions condicionades. És a dir, si les dades es distribueixen de la mateixa manera, per exemple, per a dones que per a homes, cosa que sí que s'observa, significa que hi ha una possible relació.



Per a poder determinar això farem ús de la probabilitat entre successos: diem que dos successos són independents si la probabilitat de la intersecció, és a dir que tinguen lloc tots dos, és igual a la multiplicació de les probabilitats de cadascun.

$$P(A \cap B) = P(A) \cdot P(B)$$

A la nostra taula:

$$P(Home \cap Batxillerat) \neq P(Home) \cdot P(Batxillerat)$$

$$P(Home \cap Cicle) \neq P(Home) \cdot P(Cicle)$$

Si fem totes les combinacions, podem obtenir que cap succés no és independent i, així, podem concloure que tots els successos estan relacionats. Per tant, hi ha una relació entre sexe i accés a la Facultat de Magisteri.

Finalment, per a un estadístic hi ha una prova estadística que permet dir, sense fer totes les proves, si aquestes dues variables són o no dependents, i per a fer-ho s'empra la prova khi quadrat de Pearson, que es pot usar quan les freqüències absolutes de cada columna són més grans de 5.

Prova d'independència entre les files i columnes (khi quadrat):

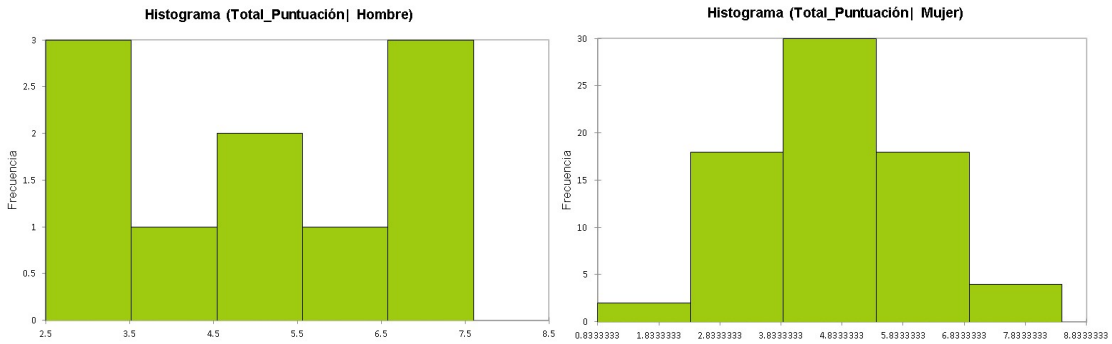
Khi quadrat (valor observat)	8,721
Khi quadrat (valor crític)	5,991
GL	2
Valor *p*	**0,013**
Alfa	0,05

Atès que el valor *p*[10] és més petit que la significació considerada (alfa),[11] s'ha demostrat que hi ha dependència entre les variables, és a dir, que estan relacionades.

III.1.2. Variable quantitativa *vs.* variable qualitativa

L'anàlisi d'aquesta classe d'associació comporta comparar les distribucions condicionals d'una variable per als diversos valors que pren l'altra. Normalment, se sol prendre la quantitativa (puntuació de l'enquesta) com a condicionada i la qualitativa o categòrica (sexe) com a condicionant, si bé les conclusions a què arribaríem serien les mateixes si es fera a l'inrevés. Si no hi ha diferències entre les distribucions condicionals, això indica que no hi ha associació entre ambdues variables.

El primer que farem és el gràfic, tal com s'ha presentat amb la resta de l'anàlisi. En aquest cas dibuixem dos histogrames, l'un que representa la puntuació de les dones i l'altre la puntuació dels homes, i s'observen diferències clares en la distribució de les dades en cadascun. Per tant, hi ha indicis d'associació entre les variables, és a dir, que estan relacionades.



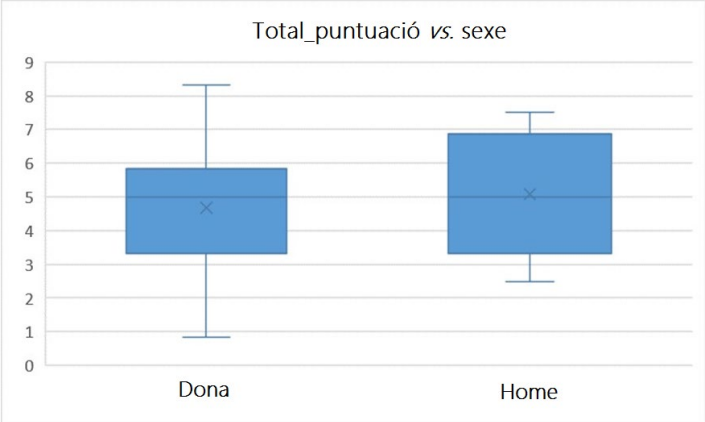
Després calculem la part numèrica descriptiva de la puntuació de l'enquesta relativa al sexe:

Variable	*N*	*Mínim*	*Màxim*	*Mitjana*	*Desv. típica*
Total_puntuació | Home	10	2,500	7,500	5,083	1,776
Total_puntuació | Dona	72	0,833	8,333	4,676	1,471

I amb aquestes dades traçarem els diagrames de caixes associats. Aquests gràfics ens permeten comparar el grau d'encavalcament (coincidència) de les distribucions condicionals. En general, com més gran és l'encavalcament, més petita és la relació entre les dues variables i viceversa, com més petit és l'encavalcament, més gran és la relació. En l'exemple que presentem ací es pot determinar un encavalcament, de manera que es posa de manifest que no hi ha relació.

[10] El valor *p* és un valor de probabilitat que varia entre 0 i 1. El valor *p* ens mostra la probabilitat d'haver obtingut el resultat que hem obtingut suposant que la hipòtesi nul·la *H*0 és certa. Se sol dir que valors alts de *p* no permeten rebutjar la *H*0, mentre que valors baixos de *p* sí que permeten rebutjar la *H*0.

[11] Es defineix com la probabilitat de prendre la decisió de rebutjar la hipòtesi nul·la quan aquesta és veritable (decisió coneguda com a error de tipus I o *fals positiu*).



Si passem ara a la part numèrica, si observem la mitjana es podria intuir que no hi ha grans diferències entre les puntuacions per a homes i per a dones. Per a determinar numèricament si hi ha aquestes diferències o no entre les puntuacions relatives al sexe, es procedeix a observar la diferència de mitjanes per a mostres independents, el valor *p* és més gran que la significació, de manera que numèricament no es poden confirmar diferències significatives entre les puntuacions d'homes i dones.

Prova t per a dues mostres independents / Prova bilateral:

Interval de confiança per a la diferència entre les mitjanes al 95%:

[-0,605; 1,420]

Diferència	0,407
t (valor observat)	0,800
| *T* | (valor crític)	1,990
GL	80
Valor *p* (bilateral)	**0,426**
Alfa	0,05

Cal assenyalar que per a aplicar la prova t s'han de complir unes certes hipòtesis en les dades i la mostra ha de ser superior a 30 individus. No entrem en la determinació de les hipòtesis perquè, en situacions d'aula, el nombre d'alumnes no supera els 30 segons la llei vigent. És per això que, en situació d'aula s'ha d'aplicar una prova no paramètrica: la prova U de Mann-Whitney.

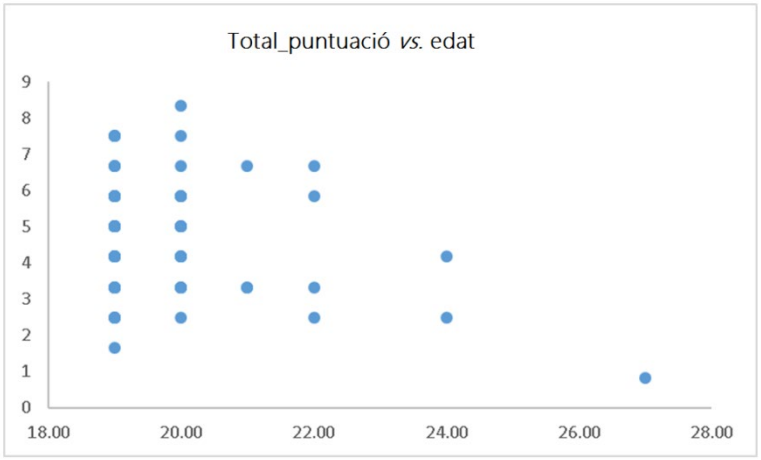
III.1.3. Dues variables quantitatives

Igual que en els casos anteriors, l'existència de correlació o associació entre dues variables quantitatives és determinada per la presència de diferències en les distribucions condicionals d'una variable per als diversos valors de l'altra.

En aquest cas, l'associació gràfica s'analitza a través d'un diagrama de dispersió, observant la disposició del núvol de punts que representa la distribució conjunta de les dues variables. Un

aspecte rellevant de l'anàlisi de la correlació entre dues variables quantitatives és que la presència d'aquesta correlació es pot plantejar d'acord amb diversos models o patrons, per exemple, en forma de línia recta o en forma curvilínia. Així, la manera d'avaluar la intensitat de la correlació sol consistir a analitzar l'ajust del núvol de punts al model d'associació que es considere que representa més adequadament la distribució conjunta de les dues variables.

En l'exemple que considerem es podria determinar que per a edats inferiors s'observa més puntuació i s'observa una relació possible lineal inversa (com menys edat, més puntuació) o fins i tot quadràtica. Tot i que és veritat que no hi ha claredat en aquest aspecte perquè hi ha puntuacions baixes en les primeres edats estudiades.



Pel que fa a la part numèrica, en aquest cas es mesura a través de les correlacions i la que *més sona* és el coeficient de correlació de Pearson. Però hem de tenir clar que aquest coeficient únicament ens servirà per a relacions lineals; es podria observar una relació quadràtica i el coeficient de correlació de Pearson no ho ha de manifestar en cap cas.

III.2. Anàlisi inferencial, dues o més mostres

Per a aquest cas, com que es tracta d'observar si hi ha diferències entre els resultats de diverses mostres, dues o més, el primer que s'ha de determinar és si les dades que es tenen de cadascuna de les mostres corresponen als mateixos individus o no. Per exemple, si tinc alumnes d'una classe i els faig una prova, els aplique una metodologia i els torne a posar la mateixa prova (abans del test i després), ací les dades són dels mateixos individus, abans i després d'una actuació, de manera que direm que les mostres estan relacionades. En el cas que tinga un grup a què aplique una metodologia i a un altre grup li n'aplique una altra, llavors parlem de mostres no relacionades.

Per al segon exemple, l'anàlisi seria exactament la mateixa que la determinada en l'apartat III.1.3. No obstant això, per al primer exemple, quan parlem de mostres relacionades, el procés gràfic es manté, però no el numèric, ja que en fer el test s'hauria d'usar per a mostres aparellades, i tenint en compte, a més, que per als casos d'aula, tal com hem dit en l'apartat III.1.3, no sempre es compleixen les hipòtesis per a aplicar aquest test, i per tant, s'haurà d'aplicar la prova de Wilcoxon.

En el cas que tinguem no únicament dues mostres, sinó més, per exemple, més classes amb les quals es poden fer comparacions, la part gràfica seguiria el mateix procés mentre que les proves de la part numèrica es modificarien: l'ANOVA s'aplicaria en el cas de mostres no relacionades (Kruskal-Wallis si no es té prou mostra) i la prova de Friedman per a mostres relacionades.

PRÀCTICA

Pr. III.0.1. En el projecte que tens entre mans tracta d'estudiar, de forma gràfica i numèrica, les possibles relacions de dues variables qualitatives.

Pr. III.0.2. En el projecte que tens entre mans tracta d'estudiar, de forma gràfica i numèrica, les possibles relacions de dues variables quantitatives.

Pr. III.0.3. En el projecte que tens entre mans tracta d'estudiar, de forma gràfica i numèrica, les possibles relacions entre una variable qualitativa i una variable quantitativa.