

Unitat 1

RECOLLIM DADES

11

UNITAT 1 RECOLLIM DADES

19. ESTADÍSTICA

Matemàtiques, Ciència i Tecnologia

què treballaràs?

En aquesta unitat aprendràs a organitzar les dades recollides per fer un estudi estadístic i coneixeràs les relacions entre estadística i probabilitat.

En acabar aquesta unitat has de ser capaç de:

- Explicar i reconèixer els conceptes estadístics de població, mostra i variable.
- Distingir entre les variables qualitatives i les variables quantitatives.
- Reconèixer les variables quantitatives discretes i les variables quantitatives contínues.
- Precisar els diferents tipus de freqüències estadístiques que poden ésser estudiades a partir d'un recull de dades.
- Agrupar les dades en taules de freqüències.
- Calcular la freqüència absoluta, la freqüència relativa, la freqüència absoluta acumulada i la freqüència relativa acumulada.
- Distingir entre probabilitat teòrica i probabilitat experimental.
- Calcular la probabilitat experimental a partir de la població sencera i a partir d'una mostra de la població.

1. Població, mostra i variable

Abans d'endinsar-nos en el tractament estadístic de dades, és necessari que ens familiaritzem amb alguns termes estadístics.

Població

Una població? Com Terrassa, Manresa o Berga? Sí, això són poblacions en la nostra manera de parlar habitual, però, a l'hora de fer estudis estadístics, ens caldrà redefinir aquest terme.

En estadística, anomenem **població** la totalitat de persones, animals o coses que són objecte d'un estudi estadístic.

Alguns exemples de poblacions són:

- Els estudiants que hi ha en una classe
- Els nens d'entre cinc i deu anys que viuen a Manresa
- Els ordinadors d'una escola
- Els arbres d'una plaça de la ciutat
- Els gats de la ciutat

Mostra

Moltes vegades no és factible estudiar tota la població sencera perquè, atesa la seva grandària, comportaria massa despesa i hauríem de dedicar-hi molt de temps. Aquest és el cas de les enquestes sobre la intenció de vot. Entrevistar tothom seria molt costós i, per aquest motiu, es treballa amb una mostra.

Una **mostra** és una part de la població.

Així doncs, els investigadors seleccionen sovint un subgrup de la població, anomenat mostra, per tal de fer els seus estudis. Una mostra mai no és perfecta, ja que es tracta d'una part de la població. Malgrat tot, amb una bona mostra, agafada a l'atzar però de forma representativa, disposarem d'una bona aproximació a la realitat de la població sencera.

Què significa que una mostra sigui representativa? Continuem amb l'exemple de les enquestes sobre la intenció de vot. A l'hora de triar la mostra, caldrà que el percentatge d'homes i dones en la mostra sigui el mateix que el que hi ha a la població sencera. Serà necessari considerar el mateix pel que fa al nivell d'estudis, la situació econòmica, etc. D'aquesta manera la mostra s'ajustarà millor a la realitat de la població sencera i els resultats de l'enquesta seran molt més fiables.

La mida de la mostra dependrà de la mida de la població i del tipus de variable que estudiem.

Variable

Anomenem **variable** allò que estudiem en la població.

Les variables poden ésser qualitatives o quantitatives.

Són les que ens proporcionen estudis sobre qualitats. Els valors d'aquest tipus de variables són noms o nivells. Fem una ullada a alguns exemples:

Variable	Valors
Professió del pare	Electricista, professor, secretari, escriptor, arquitecte, polític, etc.
Estat civil	Solter/a, casat/ada, vidu/dua, divorciat/ada, separat/ada
Color dels ulls	Blau, negre, verd, gris, marró
Tipus de matalàs	D'aire, d'aigua, de plomes, futon, de làtex, etc.
Llengua materna	Anglès, català, espanyol, àrab, francès, etc.

Variables quantitatives

Són les que proporcionen estudis numèrics, de quantitats. N'hi ha de dos tipus: discretes i contínues.

Les **variables discretes** són les que tenen valors concrets i periòdics, sense valors intermedis. Vegem-ne alguns exemples:

Variable	Valors
Treballadors d'una fàbrica	1, 2, 3, 4, 5, 6,...
Nombre de germanes	0, 1, 2, 3, 4, 5, 6,...
Diaris venuts al quiosc	0, 1, 2, 3, 4, 5, 6,...
Gent que ha assistit al cinema	0, 1, 2, 3, 4, 5, 6,...
Nombre de visitants al museu Dalí, avui	0, 1, 2, 3, 4, 5, 6,...

Si considerem el primer exemple, és a dir, els treballadors d'una fàbrica, fixa't que no hi ha valors intermedis. No hi poden haver 3,4 treballadors.

Les **variables contínues** Són les que tenen valors sense discontinuïtats i, per tant, infinits. Vegem-ne alguns exemples:

Variable	Valors
Pressió sanguínia	Infinits
Quantitat de nicotina a les cigarretes	Infinits
Alçada dels ximpanzés	Infinits
Pes dels nounats	Infinits
Consum d'aigua per persona	Infinits

Entens per què els valors són infinits? Si considerem l'alçada dels ximpanzés, l'alçada d'un ximpanzé podria ésser 123,2 cm i la d'un altre podria ésser 123,3 cm. Malgrat la poca diferència de mida entre els dos ximpanzés, podríem trobar ximpanzés d'una alçada entre la d'aquests dos: 123,221 cm, 123,2475 cm, etc.

Tanmateix, en alguns estudis, les variables contínues es converteixen en discretes. Com? Si considerem el mateix exemple de l'alçada dels ximpanzés, podem redefinir la variable i considerar l'alçada dels ximpanzés en centímetres sense decimals, aproximant la part decimal al valor més proper sense decimals. Així, per exemple, faríem:

Alçada real	Alçada arrodonida sense decimals
119,2 cm	119 cm
121,6 cm	122 cm
116,4 cm	116 cm

- Fes les activitats d'aprenentatge 1, 2 i 3

2. Una eina molt útil: les taules de freqüències

Quan hem de manejar un gran nombre de dades amb les quals volem fer un estudi estadístic, és molt útil ordenar-les per tal de facilitar els càlculs numèrics i disminuir la possibilitat de cometre errades. Amb aquesta finalitat, tot seguit aprendràs a construir taules de freqüències, les quals poden ésser de diferents tipus segons la manera de representar les dades.

A partir d'ara haurem de realitzar moltes operacions amb nombres decimals. És per aquest motiu que cal establir uns criteris a l'hora d'expressar aquestes quantitats decimals:

- En general, deixarem dues xifres decimals significatives. Això vol dir que cal posar dues xifres decimals a partir de la primera que no sigui un zero.
Exemples: 0,58 0,034 0,0032
- Abans d'escriure el segon decimal significatiu cal fixar-se en el tercer. Si el tercer és 5 o més de 5, apujarem un dígit el segon decimal. Si és 4 o menys de 4, deixarem el segon decimal tal com està. D'aquesta manera minimitzarem l'error produït per escriure només dos decimals significatius.

Vegem uns exemples:

El nombre 0,8245 el deixarem així: 0,82

El nombre 0,03271 el deixarem així: 0,033

El nombre 0,01924 el deixarem així: 0,019

El nombre 0,2963 el deixarem així: 0,3

El nombre 0,5555 el deixarem així: 0,56

- Si un nombre decimal només té tres xifres decimals significatives, les posarem totes tres. Així doncs, si dividim 1 entre 8 o 1 entre 16:

$1 : 8 = 0,125$

Deixarem les tres xifres i no posarem 0,13

$1 : 16 = 0,0625$

Deixarem les tres xifres i no posarem 0,063

16 Freqüència absoluta

Una taula de freqüències bàsica mostra els valors de la variable que estudiem i el nombre de vegades que succeeix cada valor, és a dir, la seva **freqüència absoluta** (n_i). És una manera d'organitzar les dades, fent un llistat de tots els valors possibles de la variable en una columna i la freqüència en què es produeixen en una altra columna.

Per tal d'entendre millor com fer taules, considerarem un exemple.

S'analitzen 20 peixeres i s'estudia quantes espècies d'algues unicel·lulars conté cada una. Els resultats obtinguts són els següents:

3	5	2	3	6	3	5	2	6	1
1	2	7	5	5	3	6	6	3	0

Per construir una taula de freqüències, primer cal fer una columna a l'esquerra amb els valors de la variable enregistrats (nombre d'espècies d'algues unicel·lulars), des del més petit fins al més gran. Després farem una altra columna a la dreta, on escriurem la freqüència amb què es produeix cada un dels valors.

Nombre d'espècies d'algues unicel·lulars	n_i
0	1
1	2
2	3
3	5
4	0
5	4
6	4
7	1
$N = 20$	

Fixa't que hem escrit tots els valors de la variable, des del més petit fins al més gran, fins i tot els que no succeeixen.

La suma de la columna de les freqüències absolutes ha d'ésser igual al nombre de dades recollides (N), en el cas de l'exemple, el nombre de peixeres analitzades.

Freqüència relativa

Malgrat que la freqüència absoluta ens dona una bona informació sobre la població estudiada, el valor de la freqüència absoluta no ofereix molta informació tot sol. Imagina't que, després d'haver analitzat un determinat nombre de peixeres, dic que hi ha 5 peixeres que contenen 3 espècies d'algues unicel·lulars. Això són moltes peixeres si el total de peixeres analitzades és de 20, però no en són gaires si el total de peixeres analitzades és de 1.800. És per això que ens cal un altre paràmetre: la freqüència relativa.

La **frequència relativa** (f_i) d'un valor de la variable s'obté dividint la freqüència absoluta del valor de la variable entre el total de la població N , en el nostre exemple el nombre de peixeres analitzades.

Nombre d'espècies d'algues unicel·lulars	n_i	f_i
0	1	$1 : 20 = 0,05$
1	2	$2 : 20 = 0,1$
2	3	$3 : 20 = 0,15$
3	5	$5 : 20 = 0,25$
4	0	$0 : 20 = 0$
5	4	$4 : 20 = 0,2$
6	4	$4 : 20 = 0,2$
7	1	$1 : 20 = 0,05$
	$N = 20$	Total $f_i = 1$

Fixa't que la suma de les freqüències relatives ha d'ésser sempre 1. Tanmateix, si les freqüències relatives d'alguns valors de la variable no són exactes, la suma de totes les freqüències relatives pot ésser lleugerament superior o inferior a 1.

Ara, si dic que la freqüència relativa del valor 3 espècies d'algues unicel·lulars és 0,25, queda molt clara la proporció del total de peixeres analitzades que contenen 3 espècies, sense que ens calgui saber si el total de peixeres analitzades és 20, 800 o 20.000.

Si multipliquem per 100 la freqüència relativa d'un valor de la variable, n'obtenim el seu percentatge, que és la forma popular d'expressar la freqüència relativa:

Nombre d'espècies d'algues unicel·lulars	n_i	f_i	Percentatge
0	1	0,05	5 %
1	2	0,1	10 %
2	3	0,15	15 %
3	5	0,25	25 %
4	0	0	0 %
5	4	0,2	20 %
6	4	0,2	20 %
7	1	0,05	5 %
	$N = 20$	Total $f_i = 1$	Total % = 100

Fixa't que la suma dels percentatges de tots els valors de la variable ha d'ésser 100. Malgrat que, en general, no utilitzarem els percentatges en l'elaboració de les taules, sí que els utilitzarem de vegades a l'hora de fer diagrames. L'operació de calcular els percentatges a partir de les freqüències relatives la farem directament quan elaborem el diagrama on vulguem que apareguin. Només caldrà multiplicar per cent les freqüències relatives o, el que és el mateix, passar la coma dos llocs cap a la dreta.

18 Freqüència acumulada

La freqüència acumulada expressa la freqüència d'un valor determinat de la variable sumada a la de tots els valors inferiors. Podem calcular la freqüència absoluta acumulada i la freqüència relativa acumulada.

La **freqüència absoluta acumulada** (N_i) d'un valor de la variable és el nombre de vegades que succeeix el valor en consideració o un d'inferior. Es calcula sumant les freqüències absolutes dels valors igual o inferiors al que estem considerant. També es pot calcular sumant a la freqüència absoluta del valor en consideració, la freqüència absoluta acumulada del valor immediatament inferior.

Podem calcular la **freqüència relativa acumulada** (F_i) seguint els mateixos procediments, però amb les freqüències relatives. Però hi ha una tercera manera: dividir la freqüència absoluta acumulada entre el total de dades (N).

Així doncs, la taula que inclou totes les freqüències estudiades és:

Nombre d'espècies d'algues unicel·lulars	n_i	N_i	f_i	F_i
0	1	1	0,05	0,05
1	2	3	0,1	0,15
2	3	6	0,15	0,3
3	5	11	0,25	0,55
4	0	11	0	0,55
5	4	15	0,2	0,75
6	4	19	0,2	0,95
7	1	20	0,05	1
		$N = 20$		Total $f_i = 1$

Si agafem, per exemple, el valor de la variable 3 espècies, podem comprovar com les tres maneres d'obtenir la freqüència relativa acumulada són equivalents:

1. Sumar les freqüències relatives dels valors iguals o inferiors al valor de la variable en consideració:

$$F_i = 0,05 + 0,1 + 0,15 + 0,25 = 0,55$$

2. Sumar la freqüència relativa del valor en consideració amb la freqüència relativa acumulada del valor immediatament inferior:

$$F_i = 0,25 + 0,3 = 0,55$$

3. Dividir la freqüència absoluta acumulada del valor en consideració entre el total de dades (N):

$$F_i = \frac{11}{20} = 0,55$$

Nosaltres utilitzarem aquest mètode en la resolució de les activitats.

Les freqüències acumulades només es poden obtenir si es treballa amb variables quantitatives, atès que en les qualitatives no té sentit establir-hi una ordenació de menor a major en els valors de la variable.

La utilització de les freqüències acumulades és molt habitual en els estudis temporals com, per exemple, l'increment de l'IPC (Índex de Preus al Consum) durant un any.

Les taules de freqüències ens seran molt útils per dibuixar gràfiques i calcular els termes mitjans. No sempre ens caldran totes les freqüències, per la qual cosa només en calcularem les que ens facin falta, segons l'exercici.

- Fes les activitats d'aprenentatge 4, 5, 6 i 7

3. La probabilitat

Malgrat que l'estudi a fons de la probabilitat es duu a terme en el mòdul comú de Genètica, ara tractarem alguns aspectes relacionats amb l'estadística que no es desenvolupen prou en l'esmentat mòdul de Genètica.

La probabilitat teòrica

Quan llancem una moneda enlaire, hi ha dos resultats possibles, anomenats **esdeveniments**: cara i creu. Els dos tenen la mateixa probabilitat de produir-se. Diem que cara i creu són **equiprobables**. En casos com aquest en què els esdeveniments possibles d'un experiment aleatori tenen la mateixa probabilitat p de produir-se, aquesta probabilitat es pot calcular a partir de la fórmula:

$$p = \frac{\text{Nombre casos favorables}}{\text{Nombre casos possibles}}$$

En el cas de la moneda, la probabilitat que surti creu és:

$$p(\text{creu}) = \frac{1}{2} = 0,5$$

Si tirem un dau, la probabilitat que surti un sis és:

$$p(\text{sis}) = \frac{1}{6} = 0,17$$

Això és així perquè només una de les sis cares és un sis.

Si al tirar el mateix dau volem saber la probabilitat que surti un nombre parell:

$$p(\text{parell}) = \frac{3}{6} = 0,5$$

Això és així perquè, de les sis cares del dau, tres tenen un nombre parell (2, 4 i 6) i, per tant, són favorables per tal que es produeixi l'esdeveniment.

Fixa't que la probabilitat d'un esdeveniment ha d'ésser sempre un valor entre 0 i 1. Mai més gran que 1, ja que el nombre de casos favorables no pot sobrepassar mai el nombre de casos possibles. Malgrat que la probabilitat també es pot expressar com a fracció o en forma de percentatge, nosaltres l'expressarem sempre com a valor entre 0 i 1. Això farà que la probabilitat s'expressi amb un nombre

decimal, excepte en els casos de probabilitat 0 (esdeveniments impossibles) o de probabilitat 1 (esdeveniments segurs).

La probabilitat experimental

Hi ha moltes situacions en què la probabilitat teòrica no pot ésser determinada, ja que els esdeveniments possibles no són equiprobables ni se'n pot determinar la seva probabilitat de manera teòrica. Aleshores, l'única manera de conèixer la probabilitat és fent un estudi estadístic per determinar-la. Es tracta de determinar la freqüència relativa (f_i) de l'esdeveniment del qual en volem conèixer la probabilitat.

Càlcul de la probabilitat a partir de la població sencera

Vegem-ho a través de dos exemples.

Exemple 1

Es volia conèixer la probabilitat que, entrevistant a l'atzar un dels 660 treballadors d'una empresa, aquest tingués estudis universitaris.

Per fer-ho, es va fer un estudi amb els 660 treballadors de l'empresa, que revelà que 99 treballadors tenien estudis universitaris. Aleshores es va poder calcular la probabilitat que entrevistant-ne un a l'atzar en tingués, calculant-ne la seva freqüència relativa:

$$p(\text{Estudis Universitaris}) = f_i(\text{Estudis Universitaris}) = \frac{n_i}{N} = \frac{99}{660} = 0,15$$

Exemple 2

Un estudi sobre un model de cotxe que ha sortit al mercat recentment indica que, de les 5.000 unitats venudes fins ara, 375 han patit una avaria electrònica durant el primer mes de funcionament.

Podem calcular la probabilitat que un comprador d'aquests cotxes hagi patit un problema electrònic durant el primer mes:

$$p(\text{avaría}) = f_i(\text{avaría}) = \frac{n_i}{N} = \frac{375}{5.000} = 0,075$$

En aquests dos exemples, l'estudi agafa la totalitat de la població, però en altres casos això no és possible atesa la grandària de la població. En aquests casos cal treballar amb mostres.

Càlcul de la probabilitat a partir d'una mostra

Vegem-ho també a través de dos exemples.

Exemple 1

Al Canadà es volia saber l'eficiència d'un nou antibiòtic per combatre una malaltia. Atès que no es podia fer un estudi que agafés totes les persones del país, es va fer una prova amb 800 pacients. Els resultats demostraren que només en 44 pacients l'antibiòtic fou ineficaç. La resta de pacients van superar la malaltia

gràcies a l'antibiòtic. Amb aquestes dades es determinà que la probabilitat que un malalt que arribés a un hospital es guarís seguint un tractament amb el nou antibiòtic era:

$$p_o(\text{Guariment}) = f_i(\text{Guariment}) = \frac{n_i}{n} = \frac{756}{800} = 0,945$$

Anomenem p_o la probabilitat calculada a partir d'una mostra de la població i n al nombre d'individus que hi ha a la mostra.

Cal entendre, però, que, en aquest cas, la fiabilitat de la probabilitat obtinguda depèn de la mostra que s'hagi agafat. Cal que sigui una **mostra representativa** de tota la població.

D'una banda cal que a la mostra hi siguin representats en la proporció que correspongui, els diferents grups específics que hi ha a la població i que poden comportar-se de manera diferent respecte a la variable. En el cas de l'exemple, caldria tenir en compte l'alimentació, l'edat, etc.

D'altra banda cal que la mostra tingui una mida prou gran. Si en comptes de fer l'estudi amb 800 malalts, s'hagués fet amb 200, el resultat de la probabilitat de guariment hagués estat menys fiable. Si l'estudi s'hagués fet amb 2.000 malalts, la probabilitat de guariment obtinguda hagués estat més fiable. Sempre hi haurà un **marge d'error**. Cal que sigui el més baix possible sense haver de fer estudis amb mostres excessivament grans, que encaririen molt els estudis.

S'han fet diversos estudis per determinar quina ha d'ésser la mida de la mostra perquè els resultats siguin aplicables al conjunt de la població amb un determinat marge d'error. Fruit d'aquests estudis hom disposa de diverses fórmules matemàtiques per calcular quina ha d'ésser la mida de la mostra. Nosaltres només estudiarem una fórmula molt simplificada:

$$n = \frac{1}{e^2} \quad \begin{array}{l} n = \text{mida de la mostra} \\ e = \text{error en tant per u} \end{array}$$

Per poder aplicar aquesta fórmula cal que es compleixin aquestes dues condicions:

- Que la població sigui molt gran.
- Que s'accepti un nivell de confiança de 0,95 (95 %). Això significa que hi ha un 95 % de possibilitats que la probabilitat real (la que en resultaria d'entrevistar tota la població sencera) estigui dins del marge d'error, és a dir, dins de l'interval entre $p_o - e$ i $p_o + e$.

Vegem un exemple de càlcul de la mida d'una mostra per entendre millor el funcionament d'aquesta fórmula.

Exemple 2

Es vol calcular la mida que ha de tenir una mostra de la població de Catalunya per poder determinar, en una entrevista a l'atzar, la probabilitat que l'entrevistat practiqui un esport, amb un error inferior al 6 % i amb un nivell de confiança de 0,95. També es vol calcular quina hauria d'ésser la mida de la mostra si només s'accepta un error inferior al 4 %.

Apliquem la fórmula per a un error inferior al 6 %:

$$n = \frac{1}{0,06^2} = \frac{1}{0,0036} = 277 \text{ persones}$$

Apliquem ara la fórmula per a un error inferior al 4 %:

$$n = \frac{1}{0,04^2} = \frac{1}{0,0016} = 625 \text{ persones}$$

Fixa't que si volem que l'error sigui més petit, cal que la mostra sigui més gran. En tots dos casos (i sempre que s'utilitzi aquesta fórmula) hi ha un nivell de confiança de 0,95. Això vol dir que hi ha un 95 % de possibilitats que la probabilitat real p , la que obtindríem entrevistant a tota la població de Catalunya estigui dins de l'interval comprès entre la probabilitat obtinguda a partir de la mostra p_0 i el marge d'error e acceptat (cap amunt i cap avall). Complementàriament hi ha un 5 % de possibilitats que p s'allunyi de p_0 més enllà del marge d'error e acceptat. Sempre que treballem amb mostres cal acceptar la possibilitat que p s'allunyi de p_0 més enllà del marge d'error e acceptat. Aquesta probabilitat es pot disminuir augmentant la mida de la mostra, però no es pot reduir a zero. Reduir-la a zero implicaria treballar amb tota la població i no amb una mostra.

Hi ha diferents fórmules per calcular la mida de la mostra, segons les circumstàncies d'estudi: si hi ha estudis previs, si la població és petita, etc. Algunes d'aquestes fórmules són força complexes. Nosaltres només hem intentat fer una primera aproximació al treball amb mostres.

- **Fes les activitats d'aprenentatge 8, 9 i 10**