Estadística de la informació

Thierry Lafouge
Yves-François Le Coadic
Christine Michel

PID_00189702

Universitat Oberta de Catalunya

Índex

Introducció

"Hi ha tres classes de mentides: mentides, mentides maleïdes i estadístiques."

Benjamin Disraeli

Segons estigui en singular o en plural, el terme *estadística* significa dues coses diferents:

- En singular, l'estadística és el conjunt de tècniques d'interpretació matemàtica aplicades a l'anàlisi dels valors numèrics i, en concret, les aplicades als valors, l'estudi exhaustiu dels quals és impossible a causa de la seva gran quantitat i complexitat. El valor estadístic obtingut per a una variable és una estimació del valor real d'aquesta variable.

- En plural, les estadístiques signifiquen una col·lecció de valors numèrics relatius a una categoria de fets o d'objectes, com per exemple les estadístiques de consulta d'un servei d'Internet, de préstec de llibres, de registres en la biblioteca, de visites a un museu o d'audiència d'una emissió televisiva.

Nosaltres usarem el terme *estadística* bàsicament en el primer sentit. Una vegada recollits, els valors numèrics s'hauran d'analitzar per a ordenar-los i donar-los sentit. L'anàlisi pot ser senzillament descriptiva i pot explicar, per exemple, els usos que els usuaris fan de la informació o del sistema d'informació. En aquest cas, estaríem parlant de l'estadística unidimensional. L'anàlisi també pot ser interpretativa i dir què signifiquen aquests resultats. En aquest cas, estaríem parlant de l'estadística bidimensional, que descriu i mesura la relació entre dues variables.

Podeu consultar informació sobre estadística unidimensional en l'apartat 2, i sobre estadística bidimensional en l'apartat 3 d'aquest mateix mòdul.

La dimensió d'aquestes anàlisis serà diferent si el que es pretén és fer un treball consistent i de recerca en profunditat o bé una avaluació ràpida. En el primer cas, quan es busquin les relacions que permetin invalidar o confirmar les hipòtesis formulades en els valors, serà necessari treballar amb un gran nombre de variables. En el segon cas, només serà necessari una anàlisi amb dues o tres dimensions. La manera tradicional de procedir de l'estadística, que consisteix a confirmar les hipòtesis formulades, ha evolucionat d'una manera considerable gràcies a la generalització de les eines d'anàlisi estadística multidimensional (que a França encara es coneix com a *anàlisi de dades*), que, en particular gràcies a les eines infogràfiques, permeten formular hipòtesis que a continuació es verificaran amb altres mètodes, com per exemple les estadístiques d'exploració o extracció de dades (mineria de text, mineria de dades, mineria de web).

1. Escales i unitats de mesura estadística

En els sistemes físics o biològics, com que es disposa d'escales estandarditzades i d'unitats de mesura, com per exemple l'escala mètrica i el metre, no hi ha cap problema. Però hi ha nombroses variables dels sistemes d'informació que no es poden mesurar amb tanta precisió. Així, doncs, és important verificar que les mesures i els mètodes estadístics que es vulguin fer estiguin adaptats als tipus de dades recollides, ja que en depenen.

Tal com ja hem vist anteriorment, les escales de mesura poden ser de tres tipus: nominals, ordinals i cardinals.

Podeu consultar informació sobre escales de mesura en el subapartat 3.2.2 del mòdul 1, "La mesura de la informació".

Així, doncs, els processos estadístics apropiats dependran de l'origen de les dades analitzades. Això s'ha ignorat algunes vegades. En les figures 3, 4 i 5 següents resumim les correspondències entre tipus de dades i tipus de mesures i de mètodes estadístics, segons si s'analitza una, dues o més variables.

A més, també hi ha altres mètodes estadístics que fins ara han tingut una aplicació molt limitada en les ciències de la informació i que no abordarem en aquesta assignatura. En particular, no abordarem els mètodes estadístics que permeten posar de manifest una relació entre dues variables codificades a partir d'escales de naturalesa diferent.

Figura 3. Mètodes estadístics usats en el cas d'una escala nominal

Una variable	Dues variables	*n* variables
Proporció	Encreuament (independència)	Anàlisi factorial de les correspondències
Percentatge	Coocurrència	
Relació		
Taxa de creixement		
Moda		

Figura 4. Mètodes estadístics usats en el cas d'una escala ordinal

Una variable	Dues variables
Proporció	Correlació (Spearman)
Percentatge	
Relació	
Taxa de creixement	
Moda	
Mediana	

Figura 5. Mètodes estadístics usats en el cas d'una escala cardinal

Una variable	Dues variables	*n* variables
Proporció	Correlació lineal (Pearson)	Classificació ascendent jeràrquica
Percentatge		
Relació		
Taxa de creixement		
Moda		
Mediana		
Desviació típica		
Coeficient de variació		

1.1. Canvis d'escala

En l'estadística, sovint resulta útil passar d'una escala a una altra. Sempre és possible fer-ho després d'haver fet una mesura. Es transforma una escala cardinal en una escala ordinal, una escala ordinal en una escala nominal, i una escala cardinal en una escala nominal. En fer aquesta transformació, es perd precisió, però es pot guanyar significació.

1) Pas de l'escala cardinal a l'escala ordinal

Aquesta transformació es fa de manera natural. Un valor numèric indueix de manera natural a un ordre.

2) Pas de l'escala ordinal a l'escala nominal

Igual que en el cas anterior, aquest canvi es fa de manera natural. Les classes de l'escala nominal són les que estan induïdes pels diferents rangs o posicions dins de la sèrie.

3) Pas de l'escala cardinal a l'escala nominal

Moltes vegades aquesta operació resulta útil; consisteix a crear classes. Per a fer-ho es poden usar diverses estratègies. La més natural consisteix a crear classes de la mateixa amplitud. En aquest cas, n'hi ha prou d'escollir el nombre de classes que es volen crear. Aquesta tècnica és la més habitual i és la que exigeix el nombre mínim d'hipòtesis. La segona estratègia consisteix a basar-se en hipòtesis o coneixements previs que es vulguin validar.

Podeu consultar informació sobre la creació de classes en el subapartat 2.3 d'aquest mòdul.

Exemple

En fer una enquesta sobre la professió de periodista, s'ha preguntat l'edat a tots els periodistes. L'edat varia entre els 25 i els 59 anys; si creem unes classes d'amplitud 5, s'obtenen set classes: 25-29 anys, 30-34 anys, 35-39 anys, 40-44 anys, 45-49 anys, 50-54 anys i 55-59 anys.

La creació de classes a partir de la variable cardinal no és un exercici neutre. Quan volem buscar la relació entre dues variables, una codificada en una escala nominal i l'altra sobre una escala cardinal, aquesta operació és sovint necessària. La interpretació dels resultats dependrà d'aquesta transformació.

2. L'estadística unidimensional

L'estadística unidimensional ofereix mètodes i procediments que permeten resumir grans conjunts de valors numèrics d'una variable per a convertir-los en intel·ligibles i comunicar el que és essencial d'aquests valors.

2.1. La gestió de col·leccions (de documents, d'objectes) i de connexions (bibliometria, museometria, webmetria): proporció, percentatge, taxa

Les primeres mesures estadístiques unidimensionals permeten condensar o resumir grans quantitats d'informació sobre una variable i presentar un tema amb unes poques xifres significatives. Les més conegudes –proporció, percentatge, taxa, taxa de creixement, índex– són molt útils, ja que permeten estandarditzar grups de grandària diferent per a comparar-los.

2.1.1. La proporció, el percentatge de llibres en idiomes estrangers d'un centre de documentació (una variable)

La proporció és la relació quantitativa entre les parts d'un conjunt i el conjunt íntegrament.

Exemple

En el centre de documentació, els departaments de Lletres i de Ciències de la Informació tenen, respectivament, 440 títols en idiomes estrangers sobre un total de 1.980 títols, i 880 sobre un total de 2.360. Quin és el que proporcionalment té més llibres en idiomes estrangers?

Lletres → proporció = $\frac{440}{1.980} = 0,23$

Ciències de la informació → proporció = $\frac{880}{2.360} = 0,37$

El Departament de Ciències de la Informació, amb uns 2 llibres de cada 5, és el que té més llibres en idiomes estrangers. L'altre departament només en té 1 de cada 4.

El percentatge s'obté transportant a 100 el conjunt estudiat, la qual cosa equival a multiplicar les proporcions per 100. En general, s'usen més els percentatges que les proporcions.

En l'exemple anterior, el percentatge de llibres en idiomes estrangers és del 37% en el Departament de Ciències de la Informació i només del 23% en el Departament de Lletres.

És imprescindible indicar sempre la grandària de la mostra per a no adulterar les interpretacions, tal com mostra l'exemple següent.

Exemple

La comparació de dues col·leccions de quadres i escultures que tenen dos museus, A i B, dóna els resultats següents:

Taula 1. Els percentatges

	Museu A		Museu B	
	Quantitat	Percentatge	Quantitat	Percentatge
Quadres	150.000	60%	2.800	93%
Escultures	100.000	40%	200	7%

El museu B té un 33% (93% – 60%) de quadres més que el museu A, però aquest últim té 147.200 quadres més.

2.1.2. La taxa de catalogació i la taxa de creixement (dues variables)

La taxa és la relació quantitativa entre dues variables.

1) Taxa de catalogació

Exemple

Volem comparar les catalogacions diàries, fetes pels bibliotecaris, de llibres de ficció i de llibres de no-ficció: 150 títols per als primers i 300 títols per als segons. La taxa de catalogació és:

$$\text{Taxa} = \frac{150}{300} = 0{,}5$$

Per 1 títol de ficció que es cataloga, se'n cataloguen 2 de no-ficció; es dues vegades més ràpid catalogar un llibre de no-ficció.

No s'han de confondre les taxes amb les proporcions.

La catalogació diària total és de 450 llibres. La proporció dels llibres de ficció catalogats diàriament és:

$$\text{Proporció} = \frac{150}{450} = 0{,}33$$

La de llibres de no-ficció és de 0,66 (o 66,67%).

2) Taxa de creixement

La taxa de creixement (o de decreixement) és un tipus de taxa que resulta especialment interessant. Es calcula determinant la diferència entre el valor d'una variable al principi d'un període i el valor al final d'aquest període i dividint aquesta quantitat pel valor de la variable al principi del període.

Exemple

La taxa de creixement d'un servei en línia que ha passat de 5.000 connexions el 1994 a 15.000 el 1999 és:

$$\text{Taxa de creixement} = \frac{15.000 - 5.000}{5.000} = 2$$

En percentatge, el nombre de connexions ha crescut un 200% en cinc anys, la qual cosa equival a un 40% per any. El nombre de connexions s'ha multiplicat per 3, però això no significa que l'augment hagi estat del 300%.

2.2. L'avaluació dels serveis d'informació (infometria, mediametria, webmetria): factor d'impacte, índex, flux

2.2.1. El factor d'impacte (dues variables vinculades a la variable de temps)

El factor d'impacte d'una revista I_F mesura la freqüència amb la qual, durant un any determinat, l'article "mitjà" d'una revista apareix citat en els articles d'altres revistes. És la relació entre el nombre de vegades que se cita una revista durant un any i el nombre d'articles publicats durant els dos anys anteriors.

$$I_F(t) = \frac{U_{t-1}(t) + U_{t-2}(t)}{A(t-1) + A(t-2)}$$

en què

- t és l'any que s'estudia.
- $U_{t-1}(t)$ és el nombre de vegades que s'han citat l'any t els articles publicats en la revista l'any anterior $t - 1$.
- $A(t - 1)$ és el nombre d'articles publicats per la revista el mateix any.

Exemple

Els 96 i 98 articles publicats el 1997 i 1998 en la revista *Journal of the American Society for Information Science and Technology* (JASIST) van ser citats durant l'any 1999, respectivament, 127 i 130 vegades. Així, doncs, el factor d'impacte de la revista el 1999 és:

$$\frac{127 + 130}{96 + 98} = \frac{257}{194} = 1{,}32$$

Això vol dir que, de mitjana, un article publicat en aquesta revista serà citat 1,32 vegades durant els dos anys següents. En la taula següent es poden veure els valors del factor d'impacte de les revistes principals de ciències de la informació.

Taula 2. Factor d'impacte d'algunes revistes de ciències de la informació

	Factor d'impacte
J AM SOC INFORM SGI	1,325
ONLINE CDROM REVIEW	0,293
SCIENTOMETRICS	0,931
J INFORM SCI	0,659
DATA KNOWLEDGE	0,293
INFO PROCESS MANAG	0,773
SOC STUD SGI	0,296

Font: JCR®, 1999

2.2.2. L'índex dels preus de les subscripcions a revistes (dues variables vinculades a la variable de temps)

L'índex és la relació entre els valors observats d'una variable en dues dates diferents. Els índexs s'usen especialment en les ciències econòmiques i s'han d'usar amb moltes precaucions.

Exemple

La subscripció a *Library Review*, editada per MCB University Press, valia 21.152 francs el 1998 i 33.252,99 francs el 2000. L'índex del preu de la subscripció a aquesta revista és, doncs, prenent l'any 1998 com a base:

$$\text{Índex}_{1998/2000} = \frac{33.252,99}{21.152} = 1,57$$

Això vol dir que la subscripció ha pujat un 157% en dos anys.

2.2.3. El flux de consultes (dues variables, una és la de temps)

També apareixen, cada cop més sovint, noves variables dimensionades, com la variable de temps. Aquestes variables porten a la definició de magnituds noves, i també a la definició de mesures noves i d'unitats de mesura noves, com el flux, que mesura la velocitat d'un moviment. Així, la usabilitat mesura la quantitat de vegades que es consulta una revista concreta durant un temps determinat (en general un mes). Aquesta noció ens pot remetre en certa manera a la noció de valor d'ús, una noció metafòrica del valor de canvi dels economistes.

Exemple

Si la revista *Documentaliste* s'ha consultat 59 vegades durant els últims 18 mesos, direm que la seva usabilitat és:

$$U = \frac{59}{18} = 3,28 \text{ consultes per mes}$$

Una revista amb una usabilitat inferior a 0,1 (equivalent a una consulta durant un període de 10 mesos) no s'hauria de seguir tenint.

Gràcies a aquestes mesures d'índex, de factor d'impacte i de flux, un centre de documentació podrà definir millor la seva política de subscripcions a revistes.

2.3. L'ús del correu electrònic (webmetria): mitjana, mediana, moda, interval, desviació, desviació estàndard

2.3.1. Les diferents representacions gràfiques: distribució, taula, polígon de freqüències, histogrames de freqüències, corba de distribució

La representació gràfica del conjunt dels valors numèrics obtinguts es pot fer amb polígons, histogrames i corbes. Aquestes representacions permeten una visualització de les variacions de les variables estudiades. Es tracta d'un mitjà

per a estructurar aquests valors que permetrà formular hipòtesis i fer nous processaments.

La col·lecció dels valors d'un conjunt d'elements (individus, objectes, etc.), pel fet que cada valor està associat a cada element, constitueix una distribució (taula 3).

Exemple

Els 17 enginyers d'un laboratori d'investigació i desenvolupament, identificats d'A a Q, han fet el nombre de connexions següent al seu correu electrònic, respectivament, durant una setmana:

Taula 3. Distribució de les connexions dels enginyers

Enginyers	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**
Connexions	6	9	10	14	16	17	14	16	14	10	9	14	14	16	9	17	16

La freqüència és el nombre de vegades que apareix un mateix valor, per la qual cosa la distribució de les freqüències de les connexions dels enginyers es representa en una taula que conté els valors de la variable i la freqüència d'aquests valors (taula 4).

Taula 4. Distribució de les freqüències de les connexions dels enginyers

Connexions	**Enginyers**	**% d'enginyers**
6	1	5,88
9	**3**	**17,65**
10	2	11,76
14	5	29,41
16	4	23,53
17	2	11,76
Total	**17**	**100**

Això vol dir que 3 enginyers (que representen aproximadament un 18% del total d'enginyers) s'han connectat 9 vegades durant la setmana al seu correu (vegeu la segona línia de la taula). Amb aquesta nova taula es resumeix la informació de la primera taula, però s'ha perdut informació, ja que no sabem, per exemple, que són els enginyers B, K i O els que s'han connectat 9 vegades.

Per a construir la distribució de les freqüències també es poden reagrupar els valors en classes d'una amplitud igual. La variable *connexió* va de 6 a 17, per la qual cosa construirem, per exemple, unes classes amb una amplitud 3.

Taula 5. Distribució de les freqüències de les connexions dels enginyers (classes d'amplitud 3)

Connexions	**Enginyers**	**% d'enginyers**
[6 7 8]	1	5,88
[9 10 11]	5	29,41
[12 13 14]	5	29,41
[15 16 17]	6	35,29
Total	**17**	**100**

Si, en lloc de ser de 6 a 17, la variable fos d'1 a 100, l'amplitud de les classes seria més gran; 15 classes d'amplitud 6 permetrien una compressió suficient dels valors.

Podeu trobar informació sobre sobre distribució creuada en el subapartat 3.1 d'aquest mòdul.

Tot l'"art" de les estadístiques consisteix a acceptar una pèrdua d'informació per a obtenir així, com a contrapartida, un guany de significat. Aquesta operació d'"examen detingut" dels valors que hem fet es coneix normalment com a *distribució simple*, en oposició a la distribució creuada, que veurem a continuació.

Finalment, sovint resulta útil presentar les freqüències sota una forma acumulada. Aquest càlcul es fa directament a partir de la distribució de les freqüències. Se sumen els valors i es tornen a calcular els percentatges. Així, 11 enginyers es connecten com a màxim 14 vegades (vegeu la línia 4 de la taula 6).

Taula 6. Distribució de les freqüències acumulades de les connexions dels enginyers

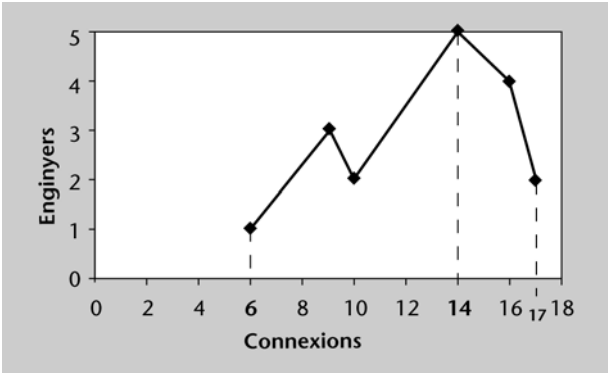
Connexions	Enginyers (acumulats)	% d'enginyers (acumulat)
6	1	5,88
9	4	23,53
10	6	35,29
14	**11**	**64,70**
16	15	88,23
17	17	100

Per a visualitzar les tres taules anteriors i resumir una gran quantitat de dades en una variable s'usen tres tipus de representacions gràfiques.

1) Polígon de freqüències

Utilitzant la taula 4, en l'eix horitzontal es representen els valors de la variable i en l'eix vertical els de les freqüències (vegeu la gràfica 1). D'aquesta manera es constata que el valor més freqüent del nombre de connexions és 14, que el nombre més baix de connexions és 6 i que el nombre més alt de connexions és 17.

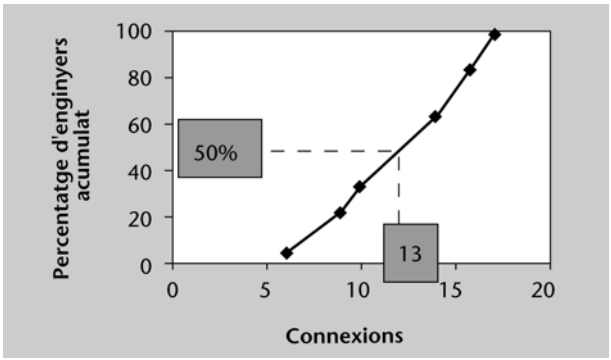
Gràfica 1. Polígon de les freqüències de connexions dels enginyers



2) Polígon de freqüències acumulades

A partir de la taula 6 es traça la gràfica 2, que demostra quin és el percentatge dels casos que queden per sota i per sobre d'un valor concret. Aquí, el 50% dels enginyers es connecten més de 13 vegades per setmana.

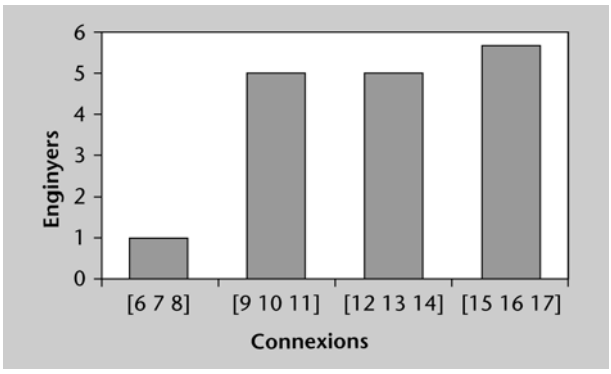
Gràfica 2. Polígon de les freqüències acumulades de connexions dels enginyers



3) Histograma de freqüències (diagrama de barres)

A partir de la taula 5 es traça la gràfica 3, que està formada per rectangles que tenen una amplària sobre l'eix horitzontal igual que l'amplitud de la classe (que aquí és de 3), i una altura igual que la freqüència corresponent. Aquesta representació no aporta més informació que el polígon de freqüències. Normalment s'usa quan la variable que s'estudia és nominal i l'eix horitzontal té una escala nominal.

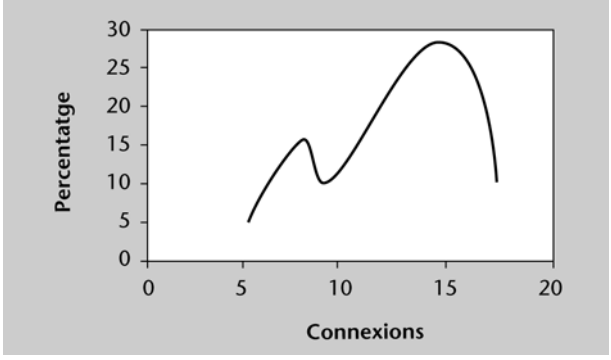
Gràfica 3. Histograma de les freqüències de connexions dels enginyers



4) Corba de distribució

Una vegada s'ha decidit representar els valors de la variable *connexió* en l'eix horitzontal i el percentatge d'enginyers en l'eix vertical, es fa un allisatge, manualment o informàticament. Així s'obté una corba contínua que es coneix com a *corba de distribució* (gràfica 4).

Gràfica 4. Distribució de connexions dels enginyers en percentatge



Les distribucions en les ciències de la informació

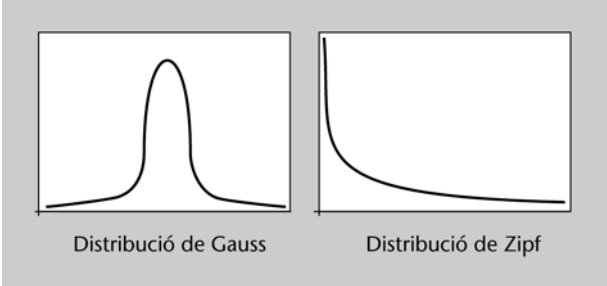
Podeu consultar informació sobre ciències de la informació en els subapartats 3.4 d'aquest mòdul i en els subapartats 2.1 i 3.3 del mòdul 3, "Matemàtica de la informació".

L'estudi anterior sobre els usos que els enginyers fan del correu ens ha portat de manera progressiva a reduir la nostra informació, primer amb l'ajuda de taules i després amb l'ajuda de gràfiques, per a arribar a una corba que es coneix com a *distribució*. En les ciències de la informació, les distribucions es poden classificar en dues grans categories: de Gauss i de Zipf.

Podeu consultar informació sobre distribució estadística en el subapartat 2.3.3 d'aquest mòdul i en l'annex 2, "Taules estadístiques".

a) Les distribucions de Gauss són distribucions conegudes com a *normals* perquè la majoria dels valors es troben prop del centre de la corba. Es representen mitjançant les famoses corbes de campana. Aquesta corba té algunes aplicacions en les ciències de la informació en estudiar els fenòmens coneguts com a "normals", encara que aquestes aplicacions no són justament les més habituals. Així, doncs, ens hem d'inclinar més aviat per les distribucions de Zipf.

Gràfica 5. Distribució de Gauss i de Zipf



Podeu consultar informació sobre distribució estadística en l'annex 3, "Càlcul de moments".

b) Les distribucions de Zipf són distribucions hiperbòliques (en *J* invertida). Són distribucions no gaussianes (no tenen moment, és a dir, no hi ha cap mitjana) que es troben amb freqüència en les ciències humanes i socials. En el camp de la lingüística, George Zipf va demostrar que si s'expliquen i ordenen les paraules d'un text segons la seva freqüència descendent, la freqüència és proporcional al rang. La paraula que ocupa la desena posició apareix deu vegades menys en el text que la paraula que ocupa el primer lloc. En les seves manifestacions no aleatòries, això es tradueix en el fet que a una causa (*input*) creixent de manera geomètrica correspon un efecte (*output*) creixent de manera aritmètica. Aquesta relació empírica s'ha observat en nombrosos autors i en nombrosos articles produïts (Lotka), en nombroses revistes i en nombrosos articles publicats (Bradford), en nombroses paraules i en nombroses circumstàncies en les quals apareixen (Zipf). Totes aquestes regularitats estadístiques han estat objecte de nombrosos estudis en ciències de la informació.

Moltes vegades, les distribucions de freqüències són massa importants per a permetre una bona representació gràfica. En aquests casos s'usen tres tipus de mesures "de condensació", que són les mesures de tendència central, les mesures de dispersió i les mesures de concentració.

2.3.2. Les mesures de tendència central

1) Mitjana (mesura cardinal)

Si tenim una variable *x*, la mitjana, que normalment es representa amb $\overline{x}$, és la suma de tots els valors de la variable *x* dividida pel nombre d'observacions, d'objectes o d'individus, representat per *N*.

Exemple

Si *x* és el nombre de connexions i *N* el nombre d'enginyers, la mitjana del nombre de connexions dels enginyers és igual a:

$$\frac{6+9+14+17+14+16+14+10+9+14+14+16+9+17+16}{17} = \frac{221}{17} = 13$$

Per a escriure aquestes expressions matemàtiques s'usen notacions indexades, i també símbols agafats de l'alfabet grec:

x_i és el valor de la variable per a l'individu *i*.

$\sum_{i=1}^{N} x_i$ és la suma dels *N* valors, en què *N* designa el nombre d'individus.

La mitjana, que normalment es representa amb $\overline{x}$, s'escriu, doncs:

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

És possible calcular la mitjana del nombre de connexions dels enginyers utilitzant la distribució de les freqüències anteriors (taula 4):

$$\frac{6 \cdot 1 + 9 \cdot 3 + 10 \cdot 2 + 14 \cdot 5 + 16 \cdot 4 + 17 \cdot 2}{1 + 3 + 2 + 5 + 4 + 2} = \frac{221}{17} = 13$$

Generalització:

Si la variable *x* estudiada adopta els valors sencers representats per *i* (*i* = 1, 2… *m*) i N_i designa les freqüències corresponents (és a dir, el nombre d'individus o d'observacions el valor de les quals és *i*), la mitjana es pot escriure:

$$\overline{x} = \frac{\sum_{i=1}^{m} iN_i}{\sum_{i=1}^{m} N_i}$$

En aquest cas, tenim $\sum_{i=1}^{m} N_i = N$ (*m* és el nombre de classes, i *N* és el nombre d'individus).

La **mitjana**, que és un indicador estadístic simple, ens pot induir a error, tal com mostra l'exemple següent.

Exemple

Un centre de documentació ha registrat el flux de préstecs següent:

- El primer any, 190 enginyers han agafat 5 obres cadascun i 10 investigadors han agafat 10 obres cadascun.

Un usuari del centre agafa, doncs, de mitjana:

$$\frac{190 \cdot 5 + 10 \cdot 10}{190 + 10}$$

És a dir, 5,25 obres a l'any.

- El segon any, 60 enginyers han agafat 6 obres cadascun i 140 investigadors han agafat 12 obres cadascun. Un usuari agafa, doncs, de mitjana, 10,2 obres a l'any.

  Si calculem les taxes de creixement, obtenim que:

  – El nombre d'obres que un enginyer ha agafat ha augmentat en $\frac{6-5}{5}$, és a dir, un 20%.

  – El nombre d'obres que un investigador ha agafat ha augmentat en $\frac{12-10}{10}$, és a dir, un 20%.

Ara bé, la variació del nombre mitjà d'obres que un usuari ha agafat és $\frac{10,2-5,25}{5,25}$, és a dir, un augment del 94%!

Tots els càlculs són correctes. No obstant això, sembla contradictori dir que la taxa mitjana de préstecs ha augmentat en un 94% i que, al mateix temps, les taxes de préstecs per als investigadors i els enginyers, que són els dos tipus d'usuaris del centre, han augmentat solament en un 20%. Això s'explica pel fet que el 94% d'increment s'ha calculat sobre les dues categories juntes, per la qual cosa les xifres no són comparables.

2) Mediana (mesura ordinal)

La mediana és el valor de la variable que divideix la distribució en dues parts iguals. És el valor central. Si el nombre total de casos és un nombre parell, el valor de la mediana s'obté calculant el valor mitjà dels dos valors centrals.

Exemple

En la distribució 2, 4, 6, 8 i 10, que té un nombre imparell de valors, la mediana és 6. Hi ha dos valors abans i dos després.

Per a la distribució de les connexions dels 17 enginyers (vegeu la taula 3), la mediana és igual a 14.

En els sis valors 6, 9, 10, 14, 16 i 17, la mediana és igual a $\frac{10+14}{2}$ = 12. La forma acumulada d'una distribució permet calcular la mediana d'una manera molt fàcil.

No s'ha de confondre la mediana amb la mitjana. En el primer exemple, la mediana i la mitjana tenen el mateix valor, que és 6. En la distribució 2, 4, 6, 8 i 20, per contra, la mediana continua essent igual a 6 però la mitjana és 8. La mediana es veu menys afectada pels valors extrems que la mitjana, que, per definició, dóna les mateixes oportunitats a tots els valors, que apareixen representats en proporció directa del seu "pes". En el sentit matemàtic, la mitjana és la mesura més representativa (el centre de gravetat de la distribució), però pot portar a interpretacions errònies en cas de distribucions asimètriques, que són freqüents en les ciències de la informació.

Exemple

En una distribució amb un nombre imparell de valors com la següent: 2.000, 4.000, 6.000, 8.000 i 10.000, la mediana i la mitjana són idèntiques i iguals a 6.000.

En una distribució amb un nombre imparell de valors com la següent: 2.000, 4.000, 6.000, 8.000 i 20.000, la mitjana és igual a 8.000 i la mediana a 6.000. Aquesta última és més representativa que la mitjana. Com que és possible ordenar els valors, és possible calcular la mediana (vegeu la gràfica 2).

3) Moda (mesura nominal)

La moda és el valor (o els valors) de la variable que es donen amb més freqüència. Així, la moda de la distribució de les connexions dels enginyers és igual a 14.

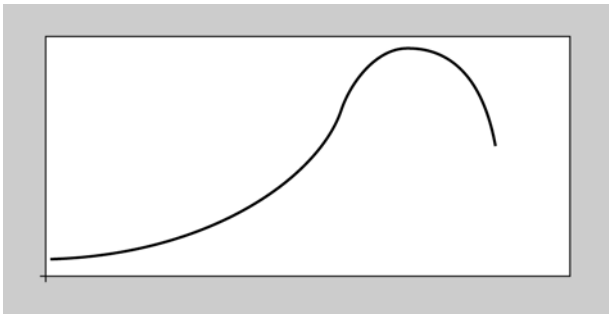
La moda no es única, tal com mostra la distribució següent, que té dues modes: 4 i 7.

2, 3, 4, 4, 4, 5, 6, 7, 7, 7

Distribucions unimodals i distribucions multimodals

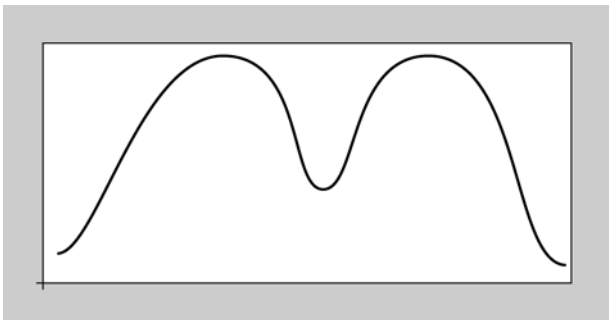
Una distribució amb una única moda es coneix com a *unimodal*.

Gràfica 6. Distribució unimodal



Una distribució amb diverses modes és una distribució multimodal. És molt freqüent i representa l'existència de grups diferents dins de la població estudiada.

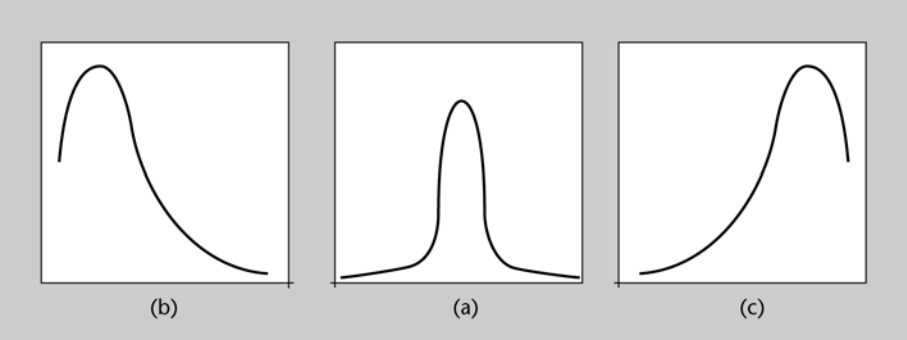
Gràfica 7. Distribució bimodal



Distribucions simètriques i distribucions asimètriques

Hi ha tres tipus de distribució unimodal: la distribució asimètrica negativa (*b*), la distribució simètrica (*a*) i la distribució asimètrica positiva (*c*):

Gràfica 8. Distribucions unimodals



- Per a una distribució asimètrica negativa (tipus *b*), tenim que moda < mediana < mitjana.

- En el cas de les distribucions simètriques, com les distribucions gaussianes (tipus *a*), la mitjana, la mediana i la moda tenen el mateix valor.

- Per a una distribució asimètrica positiva (tipus *c*), tenim que mitjana < mediana < moda.

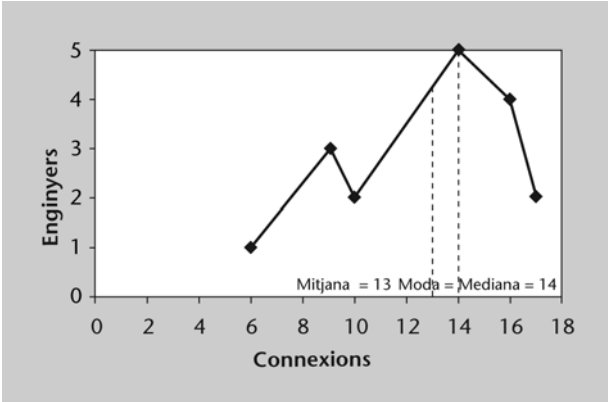
Exemple

Per al correu electrònic, tenim el següent:

Mitjana = 13 < moda = 14 = mediana

L'elecció de la mesura de tendència central no és indiferent en les distribucions no simètriques.

Gràfica 9. Mitjana, mediana, moda



2.3.3. Les mesures de dispersió

Les mesures de dispersió permeten veure com es reparteixen els valors en relació amb les mesures de tendència central. Així, doncs, podríem calcular la suma de les diferències de cada valor amb la mitjana:

$$\text{Dispersió} = \sum_{i=1}^{N}(x_i - \overline{x})$$

Aquest mètode no és pertinent, ja que la dispersió calculada d'aquesta manera és sempre nul·la, independentment dels valors de la variable. Nosaltres usarem tres mesures de dispersió: l'amplitud, la desviació mitjana i la desviació estàndard.

1) Amplitud dels valors d'una distribució

L'amplitud dels valors d'una distribució indica quina és la diferència entre el valor més alt i el valor més baix.

Exemple

L'amplitud de la distribució dels usuaris del correu de l'exercici anterior és:

$$17 - 6 = 11$$

2) Desviació mitjana

La desviació mitjana és la variació mitjana en relació amb el valor mitjà.

Si tenim una variable x, la desviació mitjana s'escriu:

$$\frac{\sum_{i=1}^{N}|x_i - \overline{x}|}{N}$$

en què N designa el nombre d'individus o objectes i $\overline{x}$ designa la mitjana $\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$.

Exemple

Durant sis setmanes s'han comptat les assistències de visitants setmanals següents al Museu de la Ciència: 1.200, 1.900, 1.600, 1.100, 2.000 i 1.800. El valor mitjà és de 1.600 visitants a la setmana. Les desviacions absolutes són:

$|1.600 - 1.200| + |1.600 - 1.900| + |1.600 - 1.600| + |1.600 - 1.100| + |1.600 - 2.000| + + |1.600 - 1.800|$

La desviació mitjana és:

$$\frac{400 + 300 + 0 + 500 + 400 + 200}{6} = \frac{1.800}{6} = 300$$

Direm que l'assistència mitjana és de 1.600 visitants a la setmana amb una desviació mitjana de 300 visitants d'una setmana a l'altra, és a dir, 1.600 ± 300.

3) Desviació estàndard (desviació típica) i variància

La desviació estàndard (o desviació típica), representada per σ_x, és la mitjana de dispersió més important. Permet quantificar la distribució dels individus en relació amb la mitjana. És més difícil de calcular que la desviació mitjana, però té algunes propietats matemàtiques interessants. Com que es tracta d'una mesura de dispersió estandarditzada, la desviació estàndard es pot usar per a comparar la igualtat o la desigualtat de dos o més grups. En el cas dels grups comparables, com més gran sigui la diferència de les desviacions estàndard, més gran serà la desigualtat.

La variància, representada per V, és el quadrat de la desviació estàndard.

Si tenim una variable x, la desviació estàndard σ_x s'escriu:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}}$$

En què N designa el nombre d'individus i $\overline{x}$ designa la mitjana.

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

La variància de σ_x^2 es coneix com a x.

Exemple

Pel que fa a l'ús del correu electrònic, la variància de la distribució de les connexions és igual a:

$$V = \frac{(6-13)^2 + (9-13)^2 + \ldots + (17-13)^2 + (16-13)^2}{17} = 9{,}17$$

La desviació estàndard és igual a: $\sigma_x = \sqrt{V} = 3{,}03$

Si la variable estudiada adopta valors enters representats per i (i = 1, 2... m) i N_i designa les freqüències corresponents (és a dir, el nombre d'individus o d'observacions que tenen com a valor i), la desviació estàndard es pot escriure:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{m} N_i (i - \overline{x})^2}{\sum_{i=1}^{m} N_i}}$$

en què m designa el nombre de classes.

Observació

Igual que en el cas de la mitjana, es pot escriure el valor de la desviació estàndard amb la desviació de les freqüències.

Exemple

En el cas del correu, això dóna:

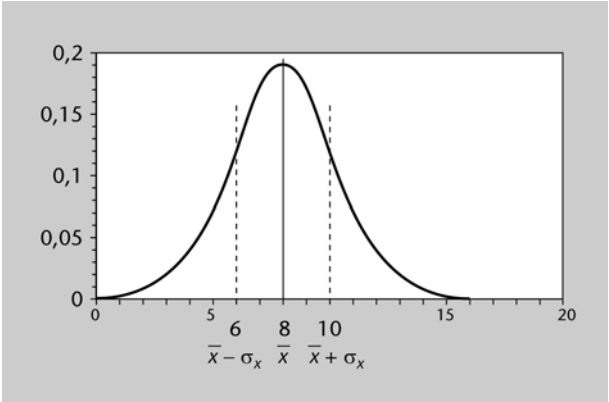
$$V = \frac{(6-13)^2 + 3\cdot(9-13)^2 + 2\cdot(10-13)^2 + \cdots 2\cdot(17-13)^2}{1+3+2+5+4+2} = 9{,}17$$

és a dir, $\sigma_x = \sqrt{V} = 3{,}03$

En el cas de les distribucions gaussianes (distribucions simètriques conegudes com a *normals* perquè la majoria dels casos queden en el centre de la corba), les dades de la mitjana $\overline{x}$ i de la desviació estàndard σ_x les caracteritzen per complet (vegeu la gràfica 10).

Podeu consultar informació sobre distribució estadística en el subapartat 2.3.1 d'aquest mòdul.

Gràfica 10. Corba de distribució de Gauss (corba de campana) amb una mitjana de 8 i desviació estàndard 2



La llei normal, o llei de Gauss, és una llei de distribució fonamental.

Si tenim una distribució amb una mitjana $\overline{x}$ i una desviació estàndard σ_x, dir que aquesta distribució segueix una llei de Gauss significa que la probabilitat *p* que el valor d'aquesta distribució es trobi entre $\overline{x} - t\sigma_x$ i $\overline{x} + t\sigma_x$, sigui *t* un nombre positiu que quantifiqui la desviació a la mitjana, és:

$$p = \frac{1}{\sigma\sqrt{2\pi}} \int_{\overline{x}-t\sigma_x}^{\overline{x}+t\sigma_x} e^{-\frac{1}{2}\left(\frac{\overline{x}-x}{\sigma_x}\right)^2} \cdot dx$$

Podeu consultar informació sobre les lleis de distribució en l'annex 2, "Taules estadístiques".

Així, per a una distribució gaussiana:

- El 68,3% dels efectius tenen un valor comprès dins de l'interval $[\overline{x} - \sigma_x, \overline{x} + \sigma_x]$.
- El 95,4% dels efectius tenen un valor comprès dins de l'interval $[\overline{x} - 2\sigma_x, \overline{x} + 2\sigma_x]$.
- El 99,7% dels efectius tenen un valor comprès dins de l'interval $[\overline{x} - 3\sigma_x, \overline{x} + 3\sigma_x]$.

Exemple

Com es reparteixen els enginyers usuaris del correu electrònic entorn de la mitjana de les connexions? La mitjana és igual a 13. Si prenem $\sigma_x = 3$, llavors $\overline{x} - \sigma_x = 10; \overline{x} + \sigma_x = 16$.

Taula 7. Distribució de les freqüències de les connexions dels enginyers (dispersió)

Connexions	Enginyers	% enginyers
6	1	5,88
9	3	17,65
10-14-16	**11**	**64,70**
17	2	11,76
Total	17	100,00

Encara que aquesta distribució no sigui gaussiana, aproximadament el 65% dels efectius es troben dins de l'interval $[\overline{x} - \sigma_x, \overline{x} + \sigma_x]$. Per a una distribució qualsevol, tenim el resultat següent:

Podem demostrar que, independentment de la naturalesa de la distribució d'una variable *x* amb una mitjana $\overline{x}$ i una desviació estàndard σ_x, el percentatge de l'efectiu concentrat entre els valors a un costat i a l'altre de la mitjana sempre és superior a $\frac{1}{t^2}$. És la desigualtat de Txebixhev, que s'escriu, per a totes les variables aleatòries *X* amb un valor esperat *E*(*X*).

$$\forall t > 0 \qquad P\left(\left|X - E(X)\right| \geq t\right) \leq \frac{Var(X)}{t^2}$$

Per exemple, si agafem $t = 2$, això ens permet escriure que com a mínim un 75% dels efectius es troben dins de l'interval $\overline{x} - 2\sigma_x$ i $\overline{x} + 2\sigma_x$.

Aquesta distribució dels efectius entorn de la mitjana, quantificada pels primers múltiples de la desviació estàndard, porta sovint a dir que aquesta última quantifica l'error que es comet si se substitueix el valor de la variable per la seva mitjana.

2.4. Dispersió de la literatura científica (infometria), homogeneïtat de les pràctiques (webmetria): distribució de rang, coeficient de variació

2.4.1. Les distribucions de rang

Tal com ja hem vist, les distribucions de Zipf són unes distribucions molt freqüents en les ciències de la informació. Desgraciadament, les mesures "de condensació", la mitjana i la desviació estàndard, les resumeixen molt malament.

Podeu consultar informació sobre distribució en les ciències de la informació en el subapartat 2.3.2 d'aquest mòdul.

Exemple

En un centre de documentació s'han analitzat quinze diaris durant un any. Per a cadascun s'ha comptabilitzat el nombre d'articles que tractaven sobre un tema concret d'investigació en curs. En la taula 8, els diaris es classifiquen per rang decreixent dels articles pertinents que contenen:

Taula 8. Nombre d'articles pertinents per diari

Diaris	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**M**	**N**	**O**
Articles	12	5	2	2	1	1	1	1	1	1	1	1	1	1	1

Els valors de les mesures "de condensació" són, llavors:

- mitjana = 2,13
- moda = 1
- mediana = 1
- desviació estàndard = 2,82

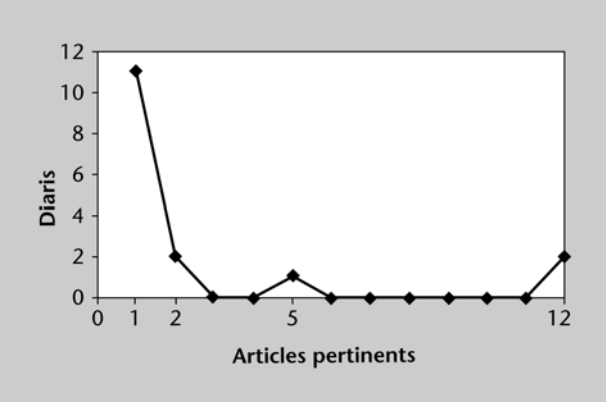
Tal com es pot constatar, la mitjana, la mediana i la desviació estàndard resumeixen malament la distribució. Només la moda és representativa. Així, doncs, construirem la distribució de les freqüències dels articles pertinents (taula 9) i després dibuixarem el polígon de les freqüències (gràfica 11).

Taula 9. Distribució de les freqüències dels articles en els diaris

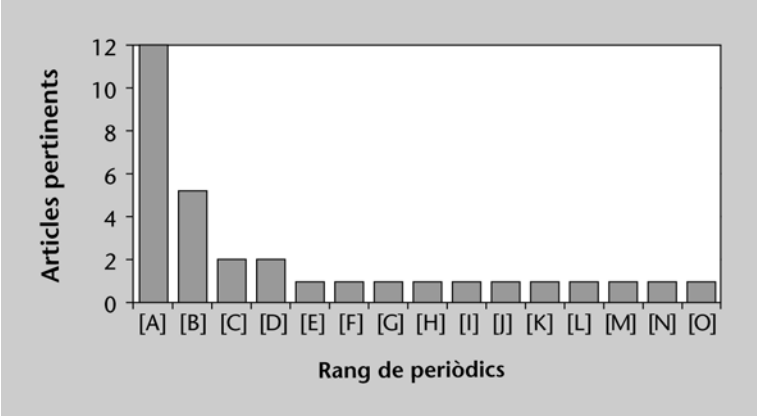
Articles pertinents	**Diaris**	**% de diaris**
1	11	73,33
2	2	13,33
5	1	6,67
12	1	6,67
Total	**15**	**100**

En lloc d'usar el nombre d'articles pertinents com una variable independent, s'usa el rang de classificació del diari. Per a això, en l'eix horitzontal es representa el rang de classificació dels diaris i en l'eix vertical el nombre d'articles pertinents (gràfica 12). Aquesta representació rang-freqüència, que es coneix com a *representació de Zipf*, s'usa molt en les ciències de la informació.

Gràfica 11. Polígon de les freqüències dels articles pertinents



Gràfica 12. Classificació dels diaris segons el nombre d'articles pertinents

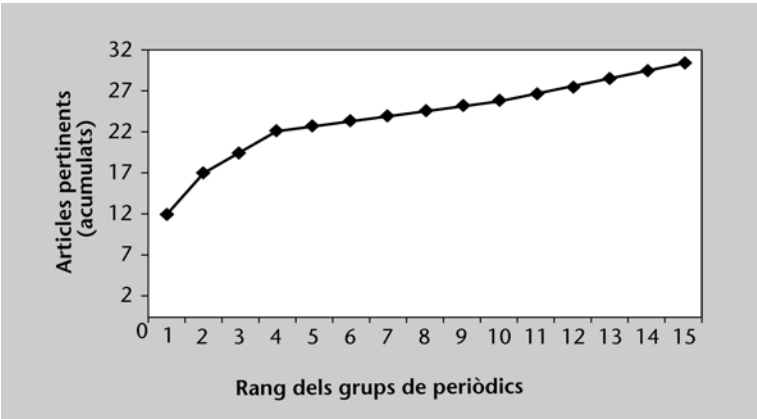


La representació rang - freqüència acumulada es coneix com a *representació de Bradford*. Després de construir la distribució de les freqüències acumulades (taula 10), es pot dibuixar la corba de distribució (gràfica 13).

Taula 10. Distribució de les freqüències acumulades del nombre d'articles pertinents

Grup de diaris	Freqüència acumulada	Rang
A	12	1
A,B	17	2
A,B,C	19	3
A,B,C,D	21	4
A,B,C,D,E	22	5
A,B,C,D,E,F	23	6
A,B,C,D,E,F,G	24	7
A,B,C,D,E,F,G,H	25	8
A,B,C,D,E,F,G,H,I	26	9
A,B,C,D,E,F,G,H,I,J	27	10
A,B,C,D,E,F,G,H,I,J,K	28	11
A,B,C,D,E,F,G,H,I,J,K,L	29	12
A,B,C,D,E,F,G,H,I,J,K,L,M	30	13
A,B,C,D,E,F,G,H,I,J,K,L,M,N	31	14
A,B,C,D,E,F,G,H,I,J,K,L,M,N,O	32	15

Gràfica 13. Classificació dels grups de diaris segons el nombre d'articles pertinents



2.4.2. Les mesures de concentració - el coeficient de variació

Basant-nos en les dues mesures característiques d'una distribució, que són la mitjana i la desviació estàndard, definim el coeficient de variació C_x:

Si tenim una distribució dels valors d'una variable x amb una mitjana $\bar{x}$ i una desviació estàndard σ_x, el seu coeficient de variació C_x és igual a la relació:

$$C_x = \frac{\sigma_x}{\bar{x}}$$

Quan volem comparar ràpidament dues distribucions, normalment és millor comparar-ne els coeficients de variació que no pas dibuixar les corbes de distribució. Aquest coeficient es pot presentar com un percentatge, encara que no és obligatori.

Exemple

En una enquesta entre els investigadors que pertanyen a dos camps científics diferents, s'ha calculat el temps que ha dedicat cadascun a usar el correu electrònic durant una setmana. S'ha calculat la mitjana $\bar{x}$ i la desviació estàndard σ_x per a cadascuna de les poblacions:

- Investigadors en ciències de l'enginyeria: $\bar{x}_1$ = 120 minuts, σ_1 = 20 minuts
- Investigadors en ciències humanes: $\bar{x}_2$ = 60 minuts, σ_2 = 4,8 minuts

Veient les desviacions estàndard, podríem estar temptats de concloure que, per als investigadors en ciències de l'enginyeria, la distribució del temps d'ús és aproximadament quatre vegades menys concentrada i que la seva població és molt més heterogènia que la dels investigadors en ciències humanes i socials.

Si calculem els coeficients de variació tenim:

$$C_1 = \frac{20}{120} = 16{,}7\% \; ; \; C_2 = \frac{4{,}8}{60} = 8\%$$

A la vista d'aquests coeficients, es constata que la distribució dels temps d'ús solament és dues vegades menys concentrada.

El valor del coeficient de variació C_1 no canvia.

Observació

Si tots els valors numèrics es multipliquen per una mateixa constant (és el cas d'un canvi d'unitats de mesura), llavors la mitjana i la desviació estàndard es multipliquen pel mateix valor i, per tant, el coeficient de variació no canvia. Per exemple, si passem a expressar el temps en hores i no en minuts, per al primer grup d'investigadors obtenim:

$\bar{x}_1 = 2$ hores $\sigma_1 = \frac{1}{3}$ hora

El coeficient de variació forma part d'un conjunt més ampli d'indicadors estadístics, coneguts com a *mesures de concentració*, que permeten resumir distribucions. Hi ha dos indicadors més, molt útils per a resumir les distribucions de Zipf: el primer està relacionat amb la distribució de Lotka i el segon amb la teoria de Shannon.

Podeu consultar informació sobre la distribució de Lotka en el subapartat 3.4 d'aquest mòdul, i sobre la teoria de Shannon en el subapartat 3.2 del mòdul 3, "Matemàtica de la informació".

2.5. En resum

Quan volem resumir un gran conjunt de valors numèrics d'una variable, i si aquesta variable té a veure amb objectes informatius o persones, pensarem de

manera natural a calcular en primer lloc els percentatges (la quantitat) i els fluxos (la quantitat per unitat de temps).

Tot seguit, en un segon moment, buscarem la manera de condensar aquest conjunt i de trobar-ne les característiques de centralitat, de dispersió i de concentració, que són la mitjana, la desviació estàndard i el coeficient de variació.

Les diferents representacions gràfiques seran una ajuda molt apreciable per a la interpretació. Cal considerar-les com unes ajudes visuals que complementen els textos escrits, però que en cap cas no els substitueixen. Tenen un doble objectiu: millorar la comprensió i permetre guanyar temps. Per a aconseguir el segon objectiu sense sacrificar el primer cal preparar molt bé les il·lustracions. Però llavors, hem de fer taules, diagrames o corbes? Si els valors mostren tendències pronunciades que produeixen una figura interessant, cal decidir-se per un diagrama o una corba. En cas contrari, n'hi ha prou amb una taula.

3. L'estadística bidimensional

Tal com ja hem vist, l'estadística unidimensional resumeix de diverses maneres un conjunt de valors numèrics d'una variable. No pretén arribar més lluny. L'estadística bidimensional és més atrevida i, en conseqüència, també més arriscada. Permet descobrir els vincles que hi ha entre dues variables.

Aquí, els procediments estadístics apropiats dependran també de la naturalesa nominal, ordinal o cardinal de les variables.

3.1. Edat dels lectors i préstecs de llibres (bibliometria): distribució creuada, independència de dues variables nominals

La tècnica usada per a posar de manifest un eventual vincle de dependència entre aquestes dues variables és la clàssica de la distribució creuada, una operació que es fa de manera habitual en processar enquestes.

Exemple

Una biblioteca municipal vol saber si l'edat dels lectors té alguna incidència sobre el nombre de préstecs de llibres al llarg de l'any. S'ha enquestat una mostra de 100 lectors, cadascun caracteritzat segons el grup d'edat al qual pertany i el nombre de llibres que ha pres en préstec al llarg de l'any. En la taula 11 apareixen els valors numèrics de les variables estudiades.

Observació

L'escala de la variable *préstec* es considera nominal, encara que aparegui representada amb nombres.

1) Els efectius observats (distribució creuada)

Taula 11. Els efectius observats

	Columna	1	2	3	4	5
Línia	Préstec / Edat	0	1	2	> 2	Total
1	< 21 anys	7	10	28	3	48
2	21-45 anys	2	9	8	2	21
3	46-60 anys	3	5	12	1	21
4	> 60 anys	2	3	3	2	10
5	Total	14	27	51	8	100

Què podem veure en aquesta taula? En el grup d'edat 21-45 anys, nou lectors han agafat un llibre (vegeu línia 2, columna 2). Aquí el que resulta interessant és intentar descobrir com varia una d'aquestes variables quan l'altra resta constant (i viceversa) i expressar aquests resultats en forma de percentatges. Així, doncs, calculem els percentatges per línia i per columna.

Observació

És evident que per a fer processaments d'aquest tipus és necessari usar algun programa informàtic d'estadística.

2) Els percentatges per línia

Calculem el percentatge de lectors per a cada grup d'edat. Així, en el grup d'edat de 21-45 anys, el 42,9% dels lectors ha agafat un llibre (vegeu la taula 12, línia 2, columna 2). El total de cada línia és, en conseqüència, 100%.

Taula 12. Distribució creuada - percentatges per línia

	Columna	**1**	**2**	**3**	**4**	**5**
Línia	**Préstec / Edat**	**0**	**1**	**2**	**> 2**	**Total %**
1	**<21 anys**	14,58	20,83	58,33	6,25	100
2	**21-45 anys**	9,52	42,86	38,10	9,52	100
3	**46-60 anys**	14,29	23,81	57,14	4,76	100
4	**>60 anys**	20,00	30,00	30,00	20,00	100

3) Els percentatges per columna

Calculem el percentatge de lectors per a cada modalitat de préstec. Així, tenim que el 37% dels llibres que s'han agafat una vegada els han agafat lectors de menys de 21 anys (línia 1, columna 2). El total de cada columna és, en conseqüència, 100%.

Taula 13. Distribució creuada - percentatges per columna

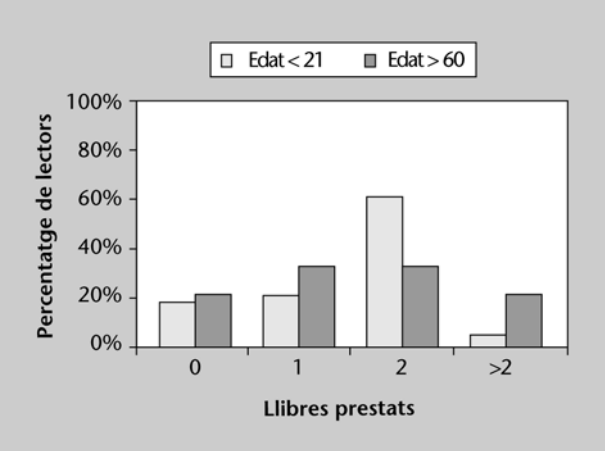
	Columna	**1**	**2**	**3**	**4**
Línia	**Préstec / Edat**	**0**	**1**	**2**	**> 2**
1	**< 21 anys**	50,00	37,04	54,90	37,50
2	**21-45 anys**	14,29	33,33	15,69	25,00
3	**46-60 anys**	21,43	18,52	23,53	12,50
4	**> 60 anys**	14,29	11,11	5,88	25,00
Total %	**100**	100	100	100	100

No ens hem de refiar dels percentatges dels efectius creuats, ja que sovint són una font d'errors d'interpretació.

4) Els histogrames

Per a comparar els perfils de les línies, a partir de les taules anteriors podem dibuixar uns histogrames semblants als que hem creat en el subapartat 2.3.1.

Gràfica 14. Comparació del nombre de préstecs de llibres fets pels lectors de menys de 21 anys i pels de més de 60 anys



Els resultats anteriors permeten representar els valors numèrics creuats amb vuit histogrames de freqüències: quatre per a les edats i quatre per als préstecs.

5) Els efectius teòrics

La pregunta que ens plantegem és la següent: les variables d'edat i préstecs són dependents l'una de l'altra? És a dir, l'edat té alguna influència sobre la quantitat de llibres que s'agafen? Per exemple, s'agafen més llibres a mesura que s'envelleix? Per resoldre aquest problema, suposarem que les dues variables són independents, és a dir, que l'edat no té cap efecte sobre els préstecs. Els efectius corresponents a aquest cas es coneixen com a *efectius teòrics*. Si hi ha aquesta independència, llavors ha de ser possible calcular els efectius creuats a partir dels totals de les línies (taula 14 B) i les columnes (taula 14 A) següents.

Taula 14. Total de línia i total de columna

Taula 14 A

Edat	
< 21 anys	48
21-45 anys	21
46-60 anys	21
> 60 anys	10
Total %	100

Taula 14 B

Préstecs				
0	**1**	**2**	**> 2**	**Total**
14	**27**	51	8	**100**

Efectius teòrics de lectors de menys de 21 anys que haurien d'haver agafat un llibre:

- 48 lectors tenen menys de 21 anys (vegeu la taula 14 A).
- 27 lectors de 100 agafen un llibre (vegeu la taula 14 B).

D'això es dedueix que 12,96 lectors de menys de 21 anys ($\frac{27}{100} \times 48$) haurien d'haver agafat un llibre. Aquest nombre és diferent de l'efectiu observat, que és igual a 10. Si fem aquest càlcul per a tots els valors de les dues variables, obtenim tots els efectius teòrics (vegeu la taula 15).

Taula 15. Efectius teòrics

	Columna	**1**	**2**	**3**	**4**	**5**
Línia	**Préstec / Edat**	**0**	**1**	**2**	**> 2**	**Total**
1	**< 21 anys**	6,72	12,96	24,48	3,84	48
2	**21-45 anys**	2,94	5,67	10,71	1,68	21
3	**46-60 anys**	2,94	5,67	10,71	1,68	21
4	**> 60 anys**	1,4	2,7	5,1	0,8	10
	Total	14	27	51	8	100

Veiem que:

- Els totals de les línies i de les columnes són els mateixos que els anteriors.
- Les línies 2 i 3 són idèntiques, la qual cosa ja era previsible, perquè els totals d'aquestes dues línies són idèntics.

6) Les desviacions cap a la independència

És possible calcular les desviacions entre els efectius observats i els efectius teòrics, que es coneixen com a *desviacions cap a la independència*.

Per als lectors de menys de 21 que agafen un llibre, la desviació és:

$$\text{Efectius observats} - \text{efectius teòrics} = 10 - 12{,}96 = -2{,}96$$

Taula 16. Desviacions cap a la independència

	Columna	**1**	**2**	**3**	**4**	**5**
Línia	**Préstec / Edat**	**0**	**1**	**2**	**> 2**	**Total**
1	**< 21 anys**	0,28	–2,96	3,52	–0,84	0
2	**21-45 anys**	–0,94	3,33	–2,71	0,32	0
3	**46-60 anys**	0,06	–0,67	1,29	–0,68	0
4	**> 60 anys**	0,6	0,3	–2,1	1,2	0

Aquestes desviacions poden ser positives o negatives. Un signe positiu significa que l'efectiu observat és superior a l'efectiu teòric. Les desviacions són de dues naturaleses diferents: les desviacions degudes a l'atzar i que, per tant, no són significatives, i les desviacions estructurals, que reflecteixen una dependència estadística de les variables i que, per tant, són significatives. Si totes les desviacions fossin nul·les, significaria que les variables són totalment independents.

7) Les proves estadístiques: la prova χ^2

Per a mesurar les desviacions cap a la independència s'usa la prova de χ^2 (khi quadrat).

Podeu consultar informació sobre les proves estadístiques en l'annex 1, "Les proves en estadística".

Càlcul de χ^2 per casella

Terminologia

El càlcul de χ^2 per casella es pot denominar, d'una manera més exacta, *contribució a χ^2 per a la casella corresponent*.

Per a cadascuna de les caselles de la taula calculem el quadrat de la desviació entre l'efectiu teòric i l'efectiu observat, que dividim per l'efectiu teòric:

$$\chi^2_{\text{per cas}} \text{ (línia1, columna2)} = \frac{(-2,96)^2}{12,96} = 0,68$$

Càlcul de χ^2 total

Sumem tots aquests valors per a obtenir la χ^2_{total} (coneguda com a $\chi^2_{\text{calculada}}$)

$$\chi^2_{\text{calculada}} = \sum \frac{\text{Desviació}^2}{\text{Efectiu teòric}} = 7,85$$

La comparació de χ^2_{total} amb el valor de χ^2 en la taula ens permet concloure que les desviacions entre les dues distribucions no són significatives; les dues variables es poden considerar independents l'una de l'altra, és a dir, que l'edat dels lectors no té cap incidència sobre el nombre de llibres que agafen al llarg de l'any.

Podeu consultar informació sobre les taules estadístiques en l'annex 2, "Taules estadístiques".

Taula 17. χ^2 per casella i total

	Columna	**1**	**2**	**3**	**4**
Línia	**Préstec** / **Edat**	**0**	**1**	**2**	**> 2**
1	**< 21 anys**	0,01	0,68	0,51	0,18
2	**21-45 anys**	0,3	1,96	0,69	0,06
3	**46-60 anys**	0,00	0,08	0,16	0,26
4	**> 60 anys**	0,26	0,03	0,86	1,8
	χ^2 total = 7,85				

No és χ^2 per casella, sinó χ^2 total, la qual cosa es pot comparar amb un valor llegit en la taula de χ^2. En el subapartat 4.2 veurem que l'anàlisi factorial de les correspondències ens permet representar visualment la proximitat o l'allunyament de les dues variables fent servir χ^2 per casella en una gràfica.

3.2. Usos d'una revista i vegades que se cita aquesta revista (infometria): correlació lineal de Pearson, regressió lineal (independència de dues variables cardinals)

3.2.1. Correlació lineal de dues variables

En analitzar la relació que hi pugui haver entre dues variables cardinals, el que es busca és veure si varien en el mateix sentit (correlació positiva) o en sentit invers (correlació negativa). Resulta interessant conèixer la força de les relacions que hi ha (expressada per una funció matemàtica), per a així poder predir un cert nombre de resultats.

Exemple

Hi ha alguna correlació entre l'impacte d'una revista científica (mesurat pel nombre de vegades que se cita en un any) i el seu ús (mesurat pel nombre d'articles demanats durant un any a un proveïdor de documents)? La qüestió és saber si una revista molt citada es fotocopia amb més freqüència que una revista poc citada. Per a les deu revistes més sol·licitades, els resultats són els següents:

Taula 18. Nombre de vegades que s'han citat les deu revistes més sol·licitades i articles fotocopiats de cadascuna

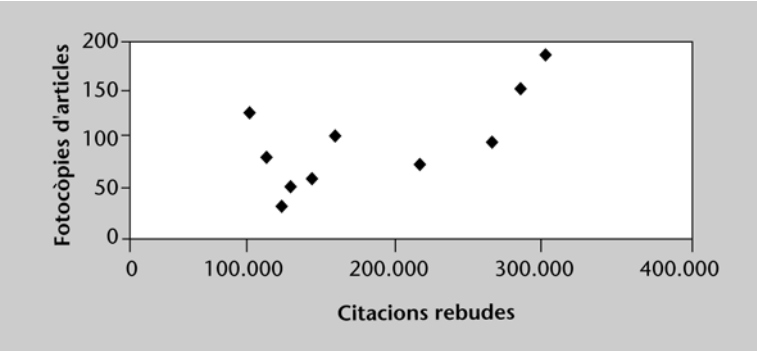
Títols de les revistes	Citacions	Articles fotocopiats
1. *Journal of Biological Chemistry*	293.024	184
2. *Proceedings of the National Academy of Science*	276.659	158
3. *Nature*	270.077	107
4. *Science*	221.696	79
5. *Journal of the American Chemical Society*	174.540	99
6. *Cell*	149.477	60
7. *Physical Review Letter*	132.797	50
8. *Physical Review B - Condensed Matter*	130.789	35
9. *Journal of Chemical Physics*	120.281	81
10. *New England Journal of Medicine*	118.106	115

El nombre de citacions s'ha obtingut de l'Informe de citacions de revistes (*Journal Citation Report*) de l'Institut d'Informació Científica (Institute for Scientific Information-ISI) de Filadèlfia.

Els valors de les dues variables, *X* per a les citacions i *Y* per a les fotocòpies, es representen en una gràfica, de la mateixa manera que dibuixem un punt sobre un mapa (latitud, longitud); una es mesura en l'eix horitzontal *X* (nombre de citacions) i l'altra en l'eix vertical *Y* (nombre de fotocòpies).

El punt resultant es troba en la intersecció de les dues perpendiculars que passen per aquests punts de mesura. D'aquesta manera, s'obté un núvol de punts que ens permet representar el vincle entre aquestes dues variables (gràfica 15).

Gràfica 15. Relació entre ús i citacions de les revistes



Per determinar si hi ha alguna correlació entre les dues variables, recorrem al coeficient de correlació r, que expressa la força de la relació (l'atracció) que hi ha entre aquestes. És un nombre que es calcula a partir de les dues variables *X* i *Y*.

- Per als valors cardinals, és el coeficient de correlació de Pearson, r_p.

$$r_p = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- Per als valors ordinals, és el coeficient de correlació de Spearman, r_s.

Podeu consultar informació sobre la correlació de rang de Spearman en el subapartat 3.3 d'aquest mòdul.

$$r_s = 1 - \frac{6 \cdot \sum d^2}{N\left(N^2 - 1\right)}$$

El càlcul de r_p es fa automàticament amb els programes informàtics d'estadística més habituals amb les fórmules següents:

$$\sigma_{XY} = \frac{1}{N} \sum_{i=N}^{i=N} (x_i - \bar{x})(y_i - \bar{y})$$

en què $\sigma_X = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}}$ és la desviació estàndard de *X*,

$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{y})^2}{N}}$ és la desviació estàndard de *Y*,

$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$ és la mitjana de *X* i $\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ és la mitjana de *Y*.

Els valors d'aquest coeficient es troben entre –1 i +1, per exemple:

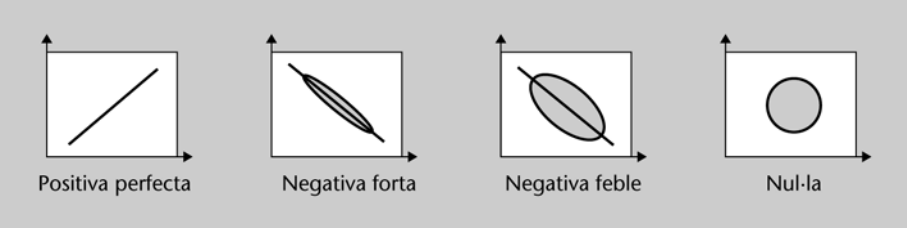
- $r_p = +1$ significa que hi ha una correlació positiva perfecta, és a dir, que la variable *Y* es manifesta linealment en funció de *X* i varia en el mateix sentit que *X*.

- $r_p = 0$ significa que no hi ha cap correlació.

- $r_p = -1$ significa que hi ha una correlació negativa perfecta, és a dir, que la variable *Y* es manifesta linealment en funció de *X* i varia en sentit invers a *X*.

La correlació també indica la força i la direcció de l'atracció. Així, doncs, la força d'una relació s'indica mitjançant el valor absolut del coeficient, però no pel seu signe. Les figures següents presenten exemples de correlacions.

Figura 6. Diferents tipus de correlació



Interpretació empírica:

0 a 0,5	correlació feble
0,5 a 0,7	correlació moderada
0,7 a 0,8	correlació important
0,8 a 1	correlació forta

Per a ser significatiu, el valor de r_p ha de ser més alt com més petita sigui la grandària *N* de la població. El grau d'importància es mesura amb una prova estadística.

Podeu consultar informació sobre les proves en estadística en l'annex 1, "Les proves en estadística".

Exemple

Amb els valors anteriors, el coeficient de correlació lineal entre usos i citacions s'obté de la manera següent:

$$N = 10$$

$$\left.\begin{array}{ll} \overline{x} = 188.744,6 & \sigma_x = 66.493,3 \\ \overline{y} = 96,8 & \sigma_y = 44,4 \end{array}\right\} \sigma_{xy} = 2.214.226,89 \quad r_p = 0,75$$

Així, doncs, hi ha una correlació lineal positiva important entre el nombre de demandes de fotocòpies d'articles d'una revista i el nombre de vegades que se cita aquesta revista.

De la mateixa manera que per a mesurar la força de la relació entre dues variables nominals hem usat una prova, que és la de χ^2, també es pot usar una prova per a mesurar la força de la correlació. La comparació del valor obtingut amb el valor llegit en la taula de r ens permet

Podeu consultar informació sobre les taules estadístiques en l'annex 2, "Taules estadístiques".

dir que les dues variables estan correlacionades positivament l'una amb l'altra, és a dir, que una revista és més sol·licitada com més vegades se cita (o al revés).

Podeu consultar informació sobre relacions no lineals en el subapartat 3.4 d'aquest mòdul.

Cal actuar amb prudència:

- La mostra representada per l'estudi està formada només pels deu títols més sol·licitats. El resultat pot ser diferent si es treballa amb deu títols agafats a l'atzar.
- En usar aquest coeficient es cometen habitualment nombrosos errors:
  - No hi ha transitivitat. Si una variable X es correlaciona amb una variable Y i aquesta es correlaciona amb una variable Z, això no implica que X i Z estiguin correlacionades.
  - El coneixement de r_p no és suficient per a certificar que dues variables es correlacionen; és necessari comprovar en la gràfica l'existència de la forma del núvol de punts.
  - Dues variables poden estar fortament relacionades entre elles malgrat tenir un coeficient de correlació lineal molt feble.

3.2.2. Regressió lineal (independència de dues variables cardinals)

Podem intentar expressar en termes matemàtics la relació lineal que hi ha entre Y i X, en què Y és la variable dependent i X la variable independent i explicativa.

Això ens porta a determinar l'equació d'una recta, coneguda com a *recta de regressió*. Aquesta recta s'obté dibuixant una línia recta que s'aproxima, en el millor dels casos, als diferents punts experimentals. Matemàticament, és la recta que minimitza la suma dels quadrats de les desviacions verticals (mètode dels mínims quadrats).

La forma general de l'equació d'aquesta recta és:

$$Y = a \cdot X + b$$

En què les constants a (pendent de la recta) i b (valor de Y quan X és nul) que determinen la posició de la recta estan determinades per les fórmules:

$$a = \frac{\sigma_{XY}}{\sigma_X^2}; \quad b = \bar{y} - a \cdot \bar{x}$$

Inversament, un cop tenim Y, podem calcular X:

$$X = c \cdot Y + d$$

on $c = \frac{\sigma_{XY}}{\sigma_Y^2}$; $d = \bar{x} - c \cdot \bar{y}$

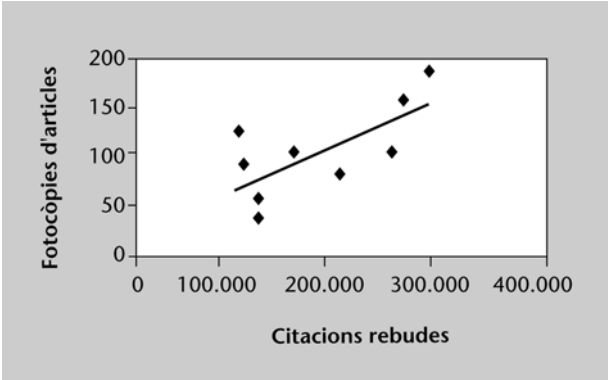
Aquestes dues rectes passen pel centre del núvol de punts amb les coordenades $\overline{x}, \overline{y}$, que són respectivament la mitjana de la variable X i la mitjana de la variable Y (vegeu la gràfica 18).

Exemple

Per a l'exemple anterior, l'equació de la recta de regressió és:

$$Y = 0{,}0005 \cdot X + 2{,}34$$

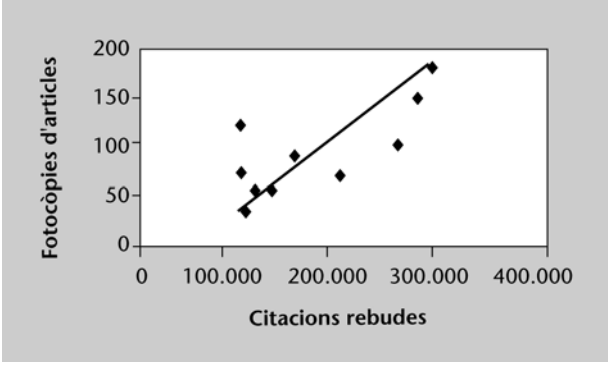
Gràfica 16. Recta de regressió: peticions de fotocòpies i citacions de la revista



Inversament, si el que volem és calcular X en funció de Y, l'equació de la recta de regressió és:

$$X = 1.122 \cdot Y + 80.129$$

Gràfica 17. Recta de regressió: citacions de la revista i peticions



Veurem que el quadrat del coeficient de correlació és igual al producte dels pendents de les dues rectes:

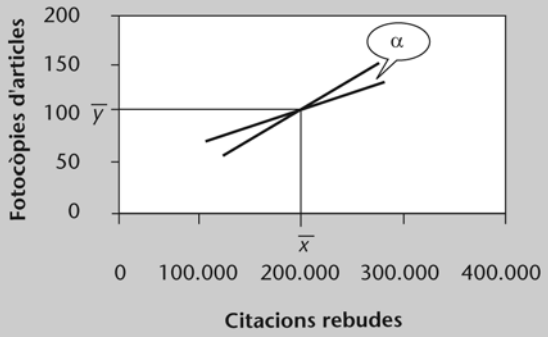
$$r_p^2 = a \cdot c$$

D'altra banda, el producte de $a \cdot c$ és el quadrat del cosinus de l'angle α (format per les dues rectes de regressió).

Quan $\alpha = 0$, les dues rectes es confonen i la correlació és perfecta (vegeu la gràfica 18):

$$\alpha = 0 \qquad \rightarrow \cos(\alpha) = 1 \qquad \rightarrow r_p = 1$$

Gràfica 18. Angle de les rectes de regressió



Al contrari que el coeficient de correlació, el càlcul de les constants de les rectes de regressió depèn de l'escala escollida per a les variables *X* i *Y*.

3.3. Ús de les cadenes de televisió per part dels estudiants (mediametria): correlació de rang de Spearman (independència de dues variables ordinals)

Amb freqüència ens trobem que només disposem d'un ordre de classificació dels individus, és a dir, de valors ordinals i no de valors cardinals. Per a mesurar la força de les associacions entre variables ordinals (rang, classificació) s'han proposat diverses solucions. Aquí estudiarem el coeficient de correlació de rang de Spearman, que s'escriu de la manera següent:

$$r_s = 1 - \frac{6 \cdot \sum d^2}{N\left(N^2 - 1\right)}$$

En què *d* designa la diferència de rang dins de la classificació per a les dues variables, i *N* el nombre d'individus. Aquest coeficient s'interpreta de la mateixa manera que el de Pearson, usant les mateixes proves.

Podeu consultar informació sobre les proves en estadístiques en l'annex 1, "Les proves en estadística".

Exemple

En un estudi sobre el temps lliure dels estudiants s'ha mesurat el temps que dediquen set d'ells a veure dues cadenes de televisió, A i B.

Taula 19. Temps dedicat a veure la televisió (en minuts)

	Temps cadena A	Temps cadena B	Rang A	Rang B	*d* = diferència de rang	d^2
Estudiant 1	53	34	3	6	3	9
Estudiant 2	91	43	1	5	4	16
Estudiant 3	49	73	4	3	–1	1

	Temps cadena A	Temps cadena B	Rang A	Rang B	d = diferència de rang	d^2
Estudiant 4	45	75	5	2	–3	9
Estudiant 5	38	93	6	1	–5	25
Estudiant 6	17	18	7	7	0	0
Estudiant 7	58	71	2	4	2	4

El coeficient de correlació és:

$$r_s = 1 - \frac{6 \cdot 64}{7\left(7^2 - 1\right)} = -0{,}143$$

Es constata que no hi ha cap correlació entre el temps dedicat a veure la cadena A i el temps dedicat a veure la cadena B.

3.4. Visites a museus (museometria), producció d'articles científics (llei de Lotka, cienciometria), préstecs de pel·lícules (mediametria): relacions no lineals

La relació lineal entre dues variables és la més simple de les relacions, però també hi ha altres relacions no lineals: exponencial, logarítmica, hiperbòlica i parabòlica. A continuació veurem que dues variables poden estar fortament vinculades per aquest tipus de relacions encara que tinguin un coeficient de correlació feble. D'altra banda, també poden tenir un coeficient de correlació molt fort però una relació que no sigui de tipus lineal.

Podeu consultar informació sobre funcions en l'apartat 3 del mòdul 3, "Matemàtica de la informació".

1) Visites a museus: relació hiperbòlica

Exemple

En una enquesta feta entre els visitants de museus s'han obtingut els resultats següents sobre el nombre de museus que han visitat durant l'any passat:

Taula 20. Distribució de les freqüències de visites a museus

Visites a museus	Visitants
1	1.000
2	125
3	37
4	15
5	8
6	4
7	3
8	2
9	2
10	1

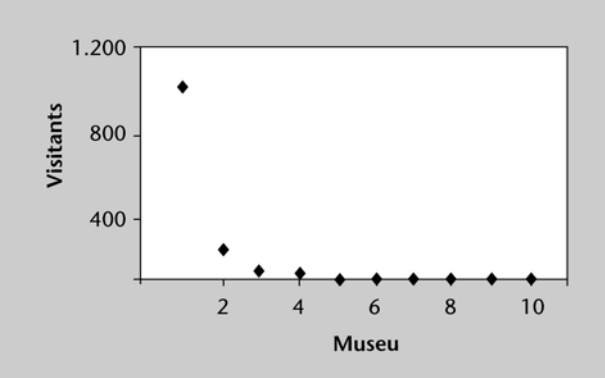
Si calculem el valor del coeficient de correlació de Pearson, obtenim el següent:

$$r_p = -0{,}59$$

La prova de r no ens permet determinar si hi ha una correlació lineal. Aquesta prova no hi escau. En realitat, la representació gràfica posa de manifest una relació hiperbòlica.

Podeu consultar informació sobre les proves en estadístiques en l'annex 1, "Les proves en estadística", i sobre la relació hiperbòlica en el subapartat 3.3 del mòdul 3, "Matemàtica de la informació".

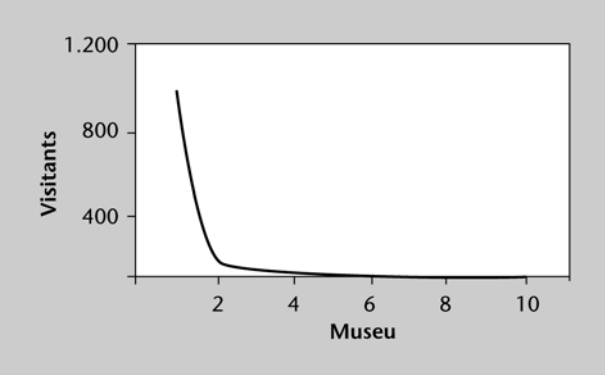
Gràfica 19. Relació hiperbòlica - observació



Això queda confirmat per un allisatge de tipus hiperbòlic (interpolació). L'equació de la corba és:

$$Y = \frac{1.000}{X^3}$$

Gràfica 20. Relació hiperbòlica - càlcul



Tal com ja hem explicat en parlar de les distribucions, en les ciències de la informació ens trobem molt sovint amb aquest tipus de relació. La no-linealitat és una constant de les lleis de la informació. En aquest cas, tenim que mil persones han visitat una vegada un museu en un any, mentre que només una ha fet deu visites. De manera general, es pot dir que hi ha uns pocs usuaris que usen molt els serveis i els productes d'informació. Per contra, són moltes les persones que els usen poc, o fins i tot que no els usen en absolut (absència d'ús).

2) Producció d'articles: relació hiperbòlica (llei de Lotka)

En el terreny de la investigació científica, la llei de Lotka és un bon exemple d'una llei hiperbòlica. El nombre *G* d'autors que han publicat *U* articles en un camp determinat i durant un període concret és igual a:

En l'exercici 32 s'ofereix una demostració de la llei de Lotka, apartat 4 del mòdul 3, "Matemàtica de la informació".

$$G_i = \frac{k}{U_i^2}$$

en què

G_i = nombre d'autors que han publicat U_i articles
U_i = nombre d'articles
k = constant

Verifiquem aquesta llei amb els valors numèrics obtinguts per Alfred Lotka.

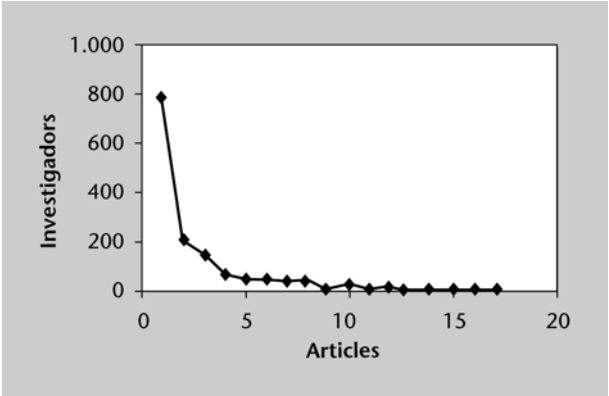
Bibliografia

A. J. Lotka (1926). "The frequency distribution of scientific productivity". *Journal of the Whashington Academy of Sciences* (núm. 16, pàg. 317-323).

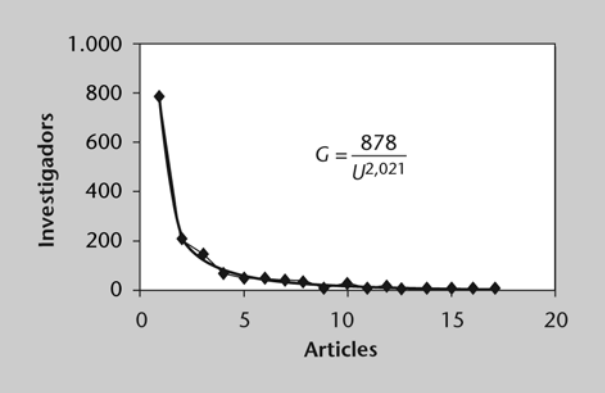
Taula 21. Articles publicats pels investigadors

Articles	Investigadors	Investigadors %
1	784	59,17
2	204	15,40
3	127	9,58
4	50	3,77
5	33	2,49
6	28	2,11
7	19	1,43
8	19	1,43
9	6	0,45
10	7	0,53
11	6	0,45
12	7	0,53
13	4	0,30
14	4	0,30
15	5	0,38
16	3	0,23
17	3	0,23
>17	16	1,21

Gràfica 21. Resultats de Lotka



Gràfica 22. Llei de Lotka



Per trobar l'equació de la corba, hem fet una regressió lineal després d'haver-hi linealitzat els valors (columnes 1 i 2 de la taula 21) amb la funció de logaritme.

Podeu consultar informació sobre la funció logarítmica en el subapartat 3.2 del mòdul 3, "Matemàtica de la informació".

Igual com abans, aquí constatem que hi ha uns pocs investigadors que publiquen molt i que, d'altra banda, hi ha molts investigadors que només publiquen uns pocs articles.

3) Préstecs en una videoteca: relació logarítmica

Exemple

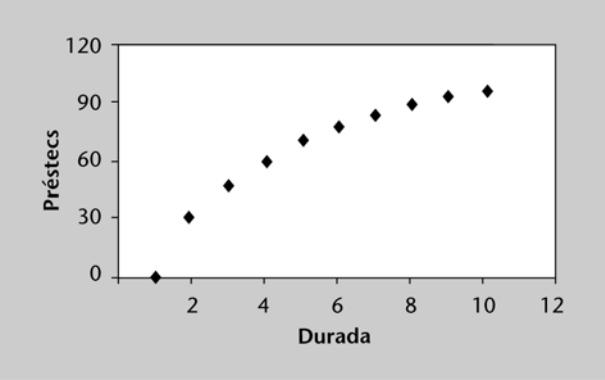
En una videoteca, el nombre acumulat de préstecs d'una pel·lícula al llarg de deu anys experimenta una variació de tipus logarítmic:

Taula 22. Distribució de les freqüències acumulades de préstecs d'una pel·lícula

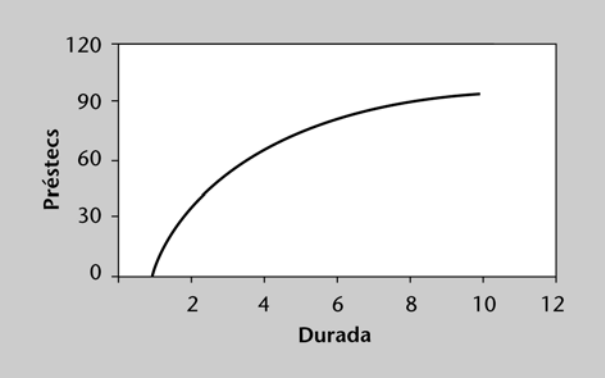
Durada (any)	Préstecs (acumulats)
1	0
2	30
3	47
4	60
5	69
6	77
7	84
8	90
9	95
10	100

Si, igual com abans, calculem el valor del coeficient de correlació, obtenim 0,95. La prova de r ens permet dir que hi ha una correlació molt significativa. No obstant això, l'examen gràfic del núvol de punts posa en relleu una relació realment forta entre les variables, encara que no de tipus lineal, sinó logarítmica.

Gràfica 23. Relació de tipus logarítmic - observació



Gràfica 24. Relació de tipus logarítmic - càlcul



Aquest aspecte queda confirmat per un allisatge de tipus logarítmic (interpolació), és a dir, per una regressió lineal feta sobre els punts que tenen en l'abscissa els logaritmes neperians de la durada i en l'ordenada les quantitats de préstecs.

Podeu consultar informació sobre la funció logarítmica en el subapartat 3.2 del mòdul 3, "Matemàtica de la informació".

L'equació de la corba és $y = 43{,}3 \cdot \ln(x)$, en què y és els préstecs acumulats, i x la durada.

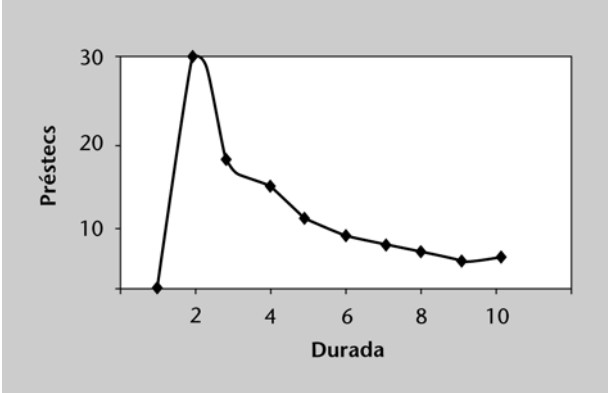
Observació

La variació dels préstecs acumulats en funció del temps és de tipus logarítmic, per la qual cosa observarem gràficament (vegeu la taula 23 i la gràfica 25) un decreixement ràpid en el temps del nombre de préstecs.

Taula 23. Distribució de les freqüències de préstecs de pel·lícules

Durada (any)	Préstecs
1	0
2	30
3	17
4	13
5	9
6	8
7	7
8	6
9	5
10	5

Gràfica 25. Freqüència dels préstecs de pel·lícules



Després d'un augment molt fort en el nombre de préstecs durant el primer any, aquest nombre disminueix molt ràpidament amb el temps. Aquest fenomen, conegut com a *obsolescència*, és molt habitual en el món de la informació.

El coeficient de correlació r s'ha d'usar correctament. Només mesura el caràcter lineal d'una relació. El seu ús està reservat als núvols de punts repartits a un costat i a un altre d'una recta.

3.5. Cartografia de la informació (infometria): coocurrència (paraules associades, cocitacions)

En el subapartat 3.1 hem estudiat un mètode que permet posar de manifest la relació entre dues variables nominals d'una extensió limitada. En el cas de variables nominals d'una extensió il·limitada, les tècniques anteriors de distribucions creuades no resulten adequades.

Vegem, per exemple, un conjunt d'articles científics caracteritzats, cadascun, per paraules diferents. *A priori* no coneixem ni aquestes paraules ni el seu nombre. Els primers tractaments simples que podem fer són establir la llista de les paraules usades i calcular-ne les freqüències (nombre d'aparicions), i tot seguit podem dirigir el nostre interès a la coocurrència de dues paraules, és a dir, al nombre de vegades que apareixen juntes en un text. Si associem les paraules d'aquesta manera, també associarem els interessos dels autors dels articles.

Tota aquesta informació es pot sintetitzar en una taula creuada en la qual les files i les columnes designen les paraules. En la diagonal de la taula trobem la freqüència de cadascuna de les paraules en el corpus d'articles, mentre que en les altres caselles veiem el nombre de coocurrències de cadascun dels parells de paraules. La taula simètrica que s'obté es coneix com a *taula de distàncies* (taula 24).

Taula 24. Taula de coocurrències

	Paraula 1	**Paraula 2**	**Paraula *i***	...	**Paraula *j***	...	**Paraula *n***
Paraula 1							
Paraula 2							
Paraula *i*			m_i		m_{ij}		
...							
Paraula *j*			m_{ij}		m_j		
...							
Paraula *n*							

m_i és el nombre d'ocurrències de cada paraula *i* en el conjunt dels articles científics, m_j és el nombre d'ocurrències de cada paraula *j* en el conjunt, i m_{ij} és el nombre de coocurrències de cada parell de paraules.

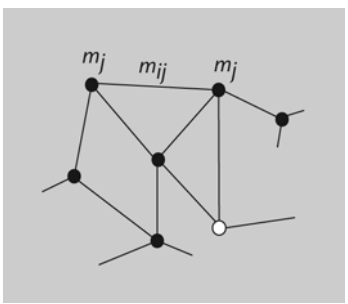
Hi ha diversos mètodes estadístics, sorgits en general de l'anàlisi multidimensional, que permeten processar aquestes taules de coocurrències. Aquí presentarem el mètode de les paraules associades i el mètode de les cocitacions.

Podeu consultar informació sobre estadística multidimensional en l'apartat 4 d'aquest mòdul.

3.5.1. L'anàlisi de les paraules associades (anàlisi de coparaules)

En aquest primer mètode, s'usen les paraules que han sorgit en indexar un article. El paper de les paraules com a operadors de l'autoestructuració dels camps científics i tècnics ja s'ha posat de manifest. Les paraules indiquen quins són els temes d'interès dins d'un camp d'investigació concret i en un moment concret. Si dues paraules apareixen simultàniament en un conjunt d'articles, els temes que representen estan associats. Per tant, els esquemes d'associació de paraules permeten posar en relleu les tendències de la investigació i els centres d'interès principals dels investigadors.

Figura 7. Xarxa d'associacions de paraules



Podeu consultar informació sobre el coeficient d'associació en el subapartat 5.3.4 del mòdul 3, "Matemàtica de la informació".

Per a construir la xarxa d'associacions de paraules, el primer pas consisteix a calcular el nombre d'ocurrències m_i de cada paraula *i* dins del conjunt d'articles, i el nombre de coocurrències m_{ij} de cada parell de paraules m_i i m_j. No obstant això, la coocurrència en si mateixa no permet mesurar la força de les associacions entre les paraules, ja que afavoreix les paraules que apareixen un gran nombre de vegades en relació amb les altres. Així, doncs, calculem un coeficient normalitzat (és a dir, els valors del qual es trobin entre 0 i 1), que creixerà amb el nombre de coocurrències, conegut com a *coeficient d'associació*, representat per E_{ij}. Un coeficient d'associació igual a 1 significa que les paraules *i* i *j* es troben sistemàticament juntes; un coeficient de 0 significa el contrari, és a dir, que mai no es troben juntes en un mateix document. Hi ha diversos mètodes de càlcul d'un coeficient d'associació. El coeficient usat en el mètode de les paraules associades és:

$$E_{ij} = \frac{m_{ij}^2}{m_i \cdot m_j}$$

Aquest coeficient varia entre 0 i 1. El seu valor és 0 si les paraules *i* i *j* no apareixen mai juntes i 1 en el cas contrari. En aquest cas, tenim la igualtat:

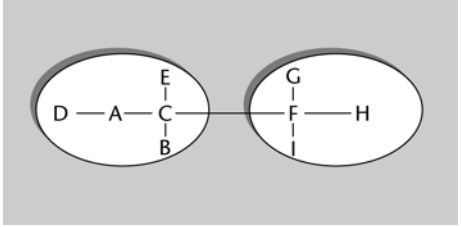
$$m_{ij} = m_i = m_j$$

Podeu consultar informació sobre la classificació jeràrquica en el subapartat 4.1 d'aquest mòdul.

Tot seguit, per a identificar els camps científics, usem una tècnica de clusterització que permet obtenir unes petites agrupacions (clústers) de paraules fortament associades. Un algorisme de classificació jeràrquica construeix aquests clústers, que són uns conjunts de paraules "properes les unes a les altres", la grandària de les quals es pot fixar. Aquesta limitació de la grandària *a priori* és un dels punts més delicats de l'aplicació d'aquest mètode.

En l'exemple següent, la grandària màxima de les agrupacions (clústers) s'ha fixat en 5. L'esquema següent (vegeu la figura 8) mostra dues agrupacions que contenen respectivament les paraules *A*, *B*, *C*, *D* i *E*, d'una banda, i *F*, *G*, *H* i *I* d'una altra. Les dues agrupacions estan vinculades entre elles per una associació externa entre les paraules *C* i *F*.

Figura 8. Formació d'agrupacions



Exemple

En la figura 9 veurem una gràfica de les paraules referents als revestiments ceràmics; els textos analitzats procedeixen d'un banc de patents i són els títols i els resums de setze mil patents d'aquest banc.

En la gràfica, el gruix dels traços entre les paraules és proporcional a la intensitat de la relació de les associacions. Per tant, cadascun dels clústers es caracteritza per les paraules del tema i les seves associacions internes i externes.

Aquesta classificació permet construir una representació cartogràfica original coneguda com a *diagrama estratègic*. La figura 10 representa el diagrama estratègic de revestiment ceràmic.

Els temes d'investigació es representen en un diagrama bidimensional definit per dos indicadors: la centralitat i la densitat.

Figura 9. Gràfica de revestiment ceràmic

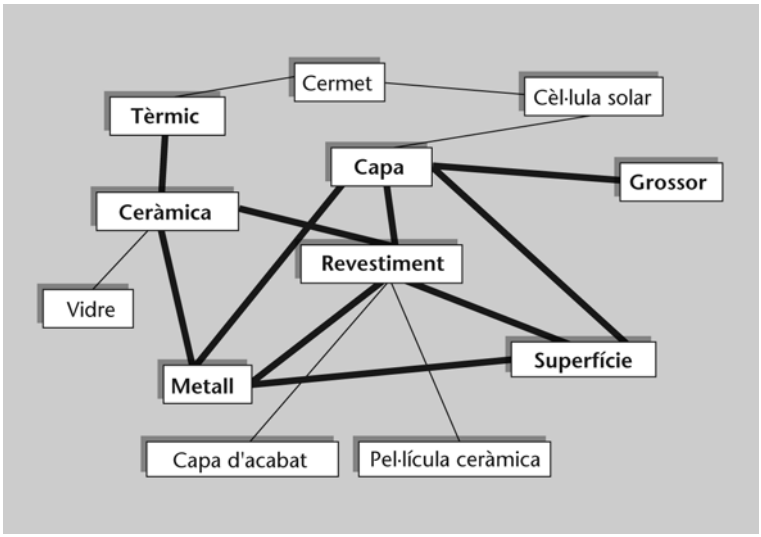
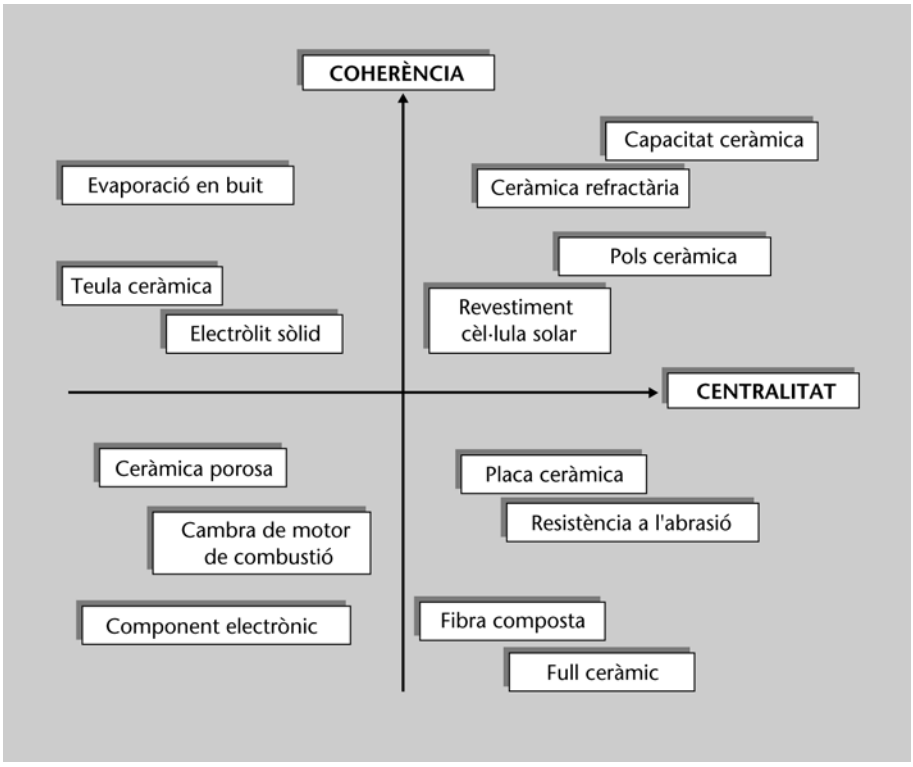


Figura 10. Diagrama estratègic de revestiment ceràmic



- La densitat d'un clúster és la mitjana de les seves associacions; és un indicador de la seva coherència.
- La centralitat d'un clúster és la mitjana de les seves associacions externes; indica si el clúster està més o menys aïllat.

Els temes que apareixen en el quadrant superior dret són els temes més importants de la investigació (molt estructurats, molt desenvolupats). Per contra, els que queden situats en el quadrant inferior esquerre són temes perifèrics i molt poc estructurats.

3.5.2. L'anàlisi de les cocitacions

El principi d'aquesta anàlisi és idèntic al de les paraules associades, però aquest mètode usa la citació en lloc de la paraula. La citació reflecteix una relació temàtica entre un text recent i els documents citats com a referència. La idea bàsica és que la freqüència de les vegades que se cita un document és l'índex de la seva importància. O, encara millor , la idea que l'estudi de les relacions entre documents molt citats permet posar de manifest unes relacions semàntiques molt fortes.

Igual que en l'exemple anterior de les paraules, explicarem les cocitacions i en construirem la matriu. Com que prenem el nom dels autors com a element de la citació, les fases d'una anàlisi d'aquest tipus són les següents:

- Selecció dels bancs de dades d'autors importants d'un tema d'investigació concret.
- Extracció de les citacions als bancs de citacions de l'ISI.
- Determinació del nombre de cocitacions dels autors estudiats.
- Construcció de la matriu de les cocitacions en la qual s'associa un valor a cada parell d'autors; un valor alt correspon a una taxa de cocitació important dels treballs d'aquests autors.
- Construcció, a partir d'aquesta matriu, d'una matriu de similitud; per a fer-ho podem escollir l'índex de correlació de Pearson, per exemple, per a cada parell d'autors.

Podeu consultar informació sobre la correlació de Pearson en el subapartat 3.2.1 d'aquest mòdul.

- Cartografia de la matriu de similitud. Partint de la matriu de similitud, l'objectiu de la cartografia és situar aquests autors dins d'un plànol, de tal manera que a una gran dissimilitud correspongui una distància important i al revés.

3.6. Anàlisi temporal dels serveis d'informació (museometria), variació estacional dels préstecs d'una biblioteca (bibliometria): sèrie cronològica

A causa del caràcter periòdic de les activitats d'informació, l'anàlisi de les sèries cronològiques (també conegudes com a *cròniques*), és a dir, de sèries de mesures espaiades en el temps, té una gran importància pràctica per als serveis d'informa-

ció. Els valors numèrics són les mesures d'una variable fetes durant uns períodes de temps concrets. El temps t és la variable independent representada en l'eix de les abscisses, i el valor mesurat corresponent, representat en l'eix de les ordenades, està representat per $y(t)$.

Si els intervals de temps són iguals, podem definir un període i assignar un nom genèric a la crònica:

- **Crònica horària**: nombre de visites que rep un lloc web cada hora.
- **Crònica diària**: nombre d'exemplars d'un periòdic que es publiquen cada dia.
- **Crònica mensual**: nombre de visites que rep un museu cada mes.
- **Crònica trimestral**: nombre de préstecs que fa una biblioteca universitària cada trimestre.

Perquè la interpretació sigui vàlida, la durada de les observacions ha de ser suficientment llarga en relació amb el període. És molt normal, per exemple, fer les cròniques mensuals al llarg de quatre o cinc anys.

Es pot representar la sèrie amb una recta de regressió lineal de la manera següent:

$$Y(t) = a \cdot t + b$$

Però el problema no sempre és tan senzill. La variació de la variable amb el temps no sempre és lineal. Algunes vegades és necessari fer uns processaments més complexos.

3.6.1. Sèrie cronològica simple: estudi de la freqüentació d'un museu

Exemple

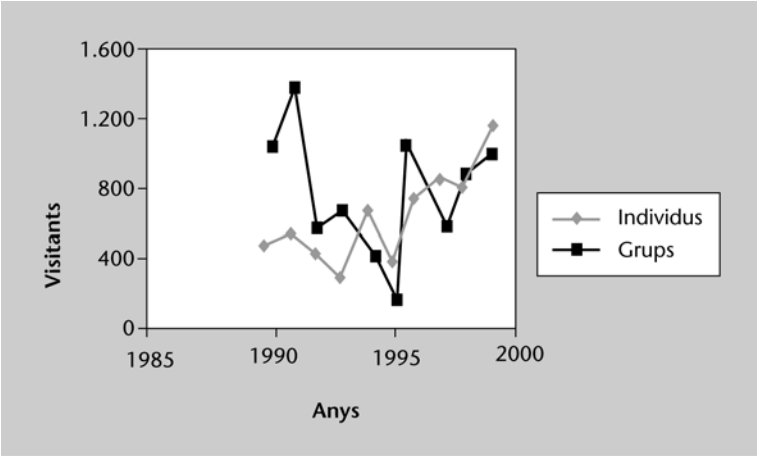
Les xifres següents descriuen la freqüentació d'un museu de 1990 a 1999 segons si els visitants van sols o en grup.

Taula 25. Freqüentació d'un museu: resultats

Any	Visitants individuals	Visitants en grup
1990	445	1.077
1991	543	1.384
1992	455	603
1993	305	682
1994	665	440
1995	394	146
1996	731	1.113
1997	840	493
1998	869	903
1999	1.132	1.006

Les variacions s'il·lustren en la gràfica següent:

Gràfica 26. Freqüentació d'un museu



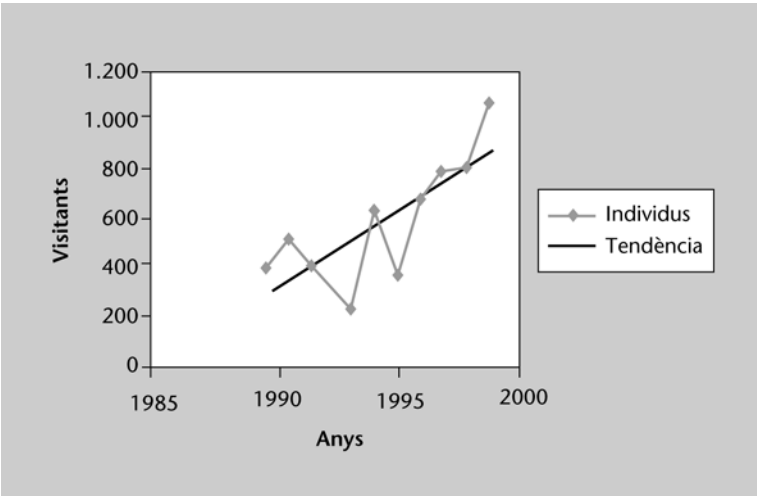
Tal com es pot veure, només la freqüentació individual sembla que experimenta una evolució creixent. L'equació de la recta de regressió determinada segons el mètode exposat abans és:

$$Y(t) = 69{,}07t - 136.436{,}92$$

En què Y és el nombre de visitants individuals i t el temps en anys.

La recta de regressió es representa en la gràfica 27.

Gràfica 27. Evolució de les visites individuals a un museu



3.6.2. Sèrie cronològica complexa: variacions estacionals dels préstecs en una biblioteca

Quan la regularitat de les observacions no és tan marcada com en el cas anterior, el que busquem és detectar si hi ha algun moviment estacional. L'aïllament del moviment estacional és especialment important, ja que la seva influència sobre els valors oculta sovint l'evolució o la tendència del moviment conjuntural. Per a posar-los de manifest, calculem els components de les

variacions característiques que observem a intervals regulars. Aquestes variacions són, per exemple:

- Una **tendència** (o *trend*), que és una variació al llarg d'un període de temps important i que correspon a una evolució llarga. Es representa amb $Tr(t)$. L'estudi anterior sobre la freqüentació d'un museu estudia la tendència.

- Una **variació estacional**, representada per $Sa(t)$, que posa de manifest evolucions degudes a fets que es repeteixen d'una manera periòdica. Per exemple, un periòdic veurà augmentar el seu tiratge el dia que publica la programació televisiva i un museu doblarà el nombre de visitants durant els mesos de vacances.

- Una **variació aleatòria** irregular, deguda a l'atzar; la seva característica és que no té cap periodicitat. Es representa amb $Al(t)$. Per exemple, un incendi en una part d'una biblioteca provocarà d'una manera puntual una disminució en l'afluència.

L'anàlisi matemàtica d'una crònica consisteix a distingir-ne i calcular-ne els diferents components. Per a això tenim a la nostra disposició diversos mètodes:

- Per a determinar la tendència usem mètodes de tipus regressió.
- Per a determinar el component estacional recorrem a altres mètodes que es presenten en obres especialitzades.

En general, s'admet que els tres components definits més amunt es relacionen segons un model, l'estructura del qual pot ser additiva, multiplicativa o mixta.

- Model additiu $Y(t) = Tr(t) + Sa(t) + Al(t)$
- Model multiplicatiu $Y(t) = Tr(t) \cdot Sa(t) + Al(t)$
- Model mixt $Y(t) = Tr(t)\ (a \cdot Sa(t) + 1) + bSa(t) + Al(t)$

(Per a $a = 0$, tenim el model additiu, i per a $b = 0$, el model multiplicatiu.)

Exemple

Les xifres següents descriuen el volum de préstecs d'una biblioteca municipal entre 1991 i 1995 (taula 26). La biblioteca només està oberta onze mesos a l'any, per la qual cosa no es registra cap valor per al mes de setembre. (Els valors usats en aquest exemple s'han adaptat de la tesi de Roland Ducasse (1978), *Méthode du traitement des données bibliométriques pour la gestion des systèmes d'information: application à l'analyse prévisionnelle de la demande d'ouvrage en bibliothèque*, Universitat de Bordeus 3.)

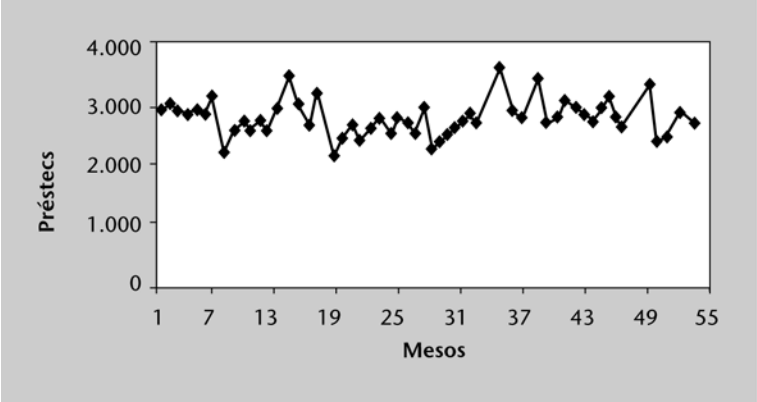
Taula 26. Volum de préstecs en la biblioteca entre 1991 i 1995

	1991	1992	1993	1994	1995
Gener	2.897	2.794	2.702	2.767	2.843
Febrer	2.923	2.713	2.848	2.857	2.807
Març	2.883	2.962	2.603	2.774	3.007

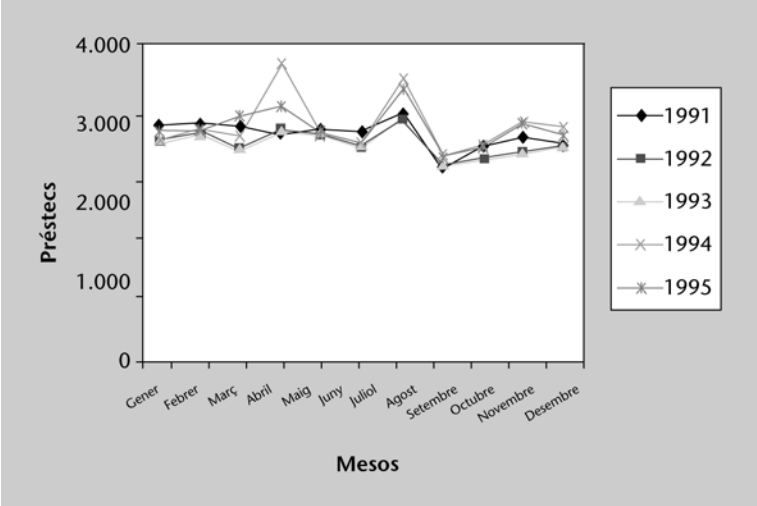
	1991	1992	1993	1994	1995
Abril	2.823	3.386	2.857	3.696	3.148
Maig	2.839	2.997	2.794	2.770	2.815
Juny	2.828	2.631	2.650	2.730	2.688
Juliol	3.070	3.161	3.022	3.498	3.381
Agost	2.374	2.201	2.416	2.552	2.509
Octubre	2.658	2.518	2.524	2.659	2.577
Novembre	2.765	2.714	2.583	2.936	2.955
Desembre	2.714	2.504	2.673	2.892	2.799

En la gràfica 28 es veu la tendència de l'evolució del préstec durant el transcurs dels cinc últims anys i, en la gràfica 29, la variació estacional.

Gràfica 28. Tendència del préstec de 1991 a 1995



Gràfica 29. Variació estacional del préstec



Què podem llegir en aquestes gràfiques?

La primera indica una disminució en els préstecs en els trenta primers mesos de 1991 a 1993 i, després, un increment en els dos últims anys. La segona posa de manifest els moviments estacionals, que aquí presenten un increment dels préstecs en els mesos d'abril i de juliol i una disminució considerable en el mes d'agost.

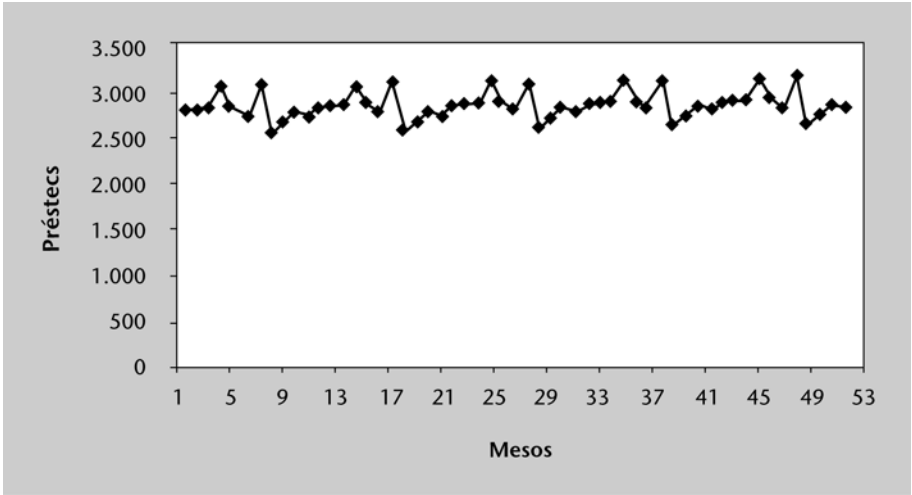
Ja que sembla un fenomen estacional, un ajust dels valors fet aplicant un esquema additiu, per exemple, donarà els valors següents:

Taula 27. Volum de préstecs en la biblioteca de 1995 a 1999: valors ajustats

	1991	1992	1993	1994	1995
Gener	2.750	2.775	2.800	2.826	2.850
Febrer	2.779	2.804	2.830	2.855	2.880
Mar5ç	2.795	2.820	2.846	2.871	2.896
Abril	3.131	3.157	3.182	3.205	3.232
Maig	2.793	2.817	2.843	2.868	2.893
Juny	2.655	2.680	2.705	2.730	2.755
Juliol	3.176	3.201	3.226	3.251	3.277
Agost	2.359	2.384	2.410	2.435	2.460
Octubre	2.537	2.561	2.587	2.612	2.637
Novembre	2.699	2.725	2.750	2.776	2.801
Desembre	2.666	2.641	2.717	2.742	2.767

La regularitat anual dels préstecs es mostra en la gràfica 30.

Gràfica 30. Tendència del préstec de 1991 a 1995: valors ajustats



L'interès de la construcció d'un model d'ajust és que permet una previsió satisfactòria de la demanda.

La predicció o la prospectiva fetes a partir d'aquestes corbes són qüestionables des d'un punt de vista matemàtic, ja que l'observació del passat no sempre ens aporta coneixements sobre el futur.

3.7. En resum

L'estadística bidimensional, que és la segona etapa dins del descobriment de les variacions de les magnituds informatives, ajuda a posar de manifest les relacions que hi ha entre dues d'aquestes magnituds.

Les magnituds poden ser diferents i es poden mesurar amb escales diferents. Si l'escala és cardinal, el que es busca és mesurar el grau de correlació que hi ha entre les variables estudiades; aquesta correlació pot ser lineal positiva o negativa i, llavors, les magnituds varien en el mateix sentit o en sentit invers; o també es poden donar correlacions no lineals de diversos tipus que han donat a les ciències de la informació algunes de les seves lleis més belles.

Les dues magnituds poden ser de la mateixa naturalesa. En aquest cas, el que buscarem és determinar-ne les coocurrències. Aquest és el cas, en els textos, de les paraules i de les citacions (o referències), les relacions de les quals cartografiarem per actualitzar les temàtiques informatives ocultes en aquests textos.

Finalment, acompassant la vida de les informacions, la variable de temps permet seguir les variacions estacionals, característiques de la periodicitat dels fenòmens informatius.

Tots aquests enfocaments ens parlen de manera interessant sobre la realitat de les activitats informatives. No obstant això, cal ser raonable i evitar deixar-nos portar per un optimisme excessiu que ens porti a veure en els valors numèrics obtinguts més del que poden significar realment.

4. L'estadística multidimensional

L'estadística multidimensional permet investigar les relacions que hi ha entre diverses variables. Aquesta estadística ofereix eines de les quals s'espera, abans de res, que tinguin una eficàcia pràctica, mentre que la seva justificació teòrica queda en un segon pla. En aquest sentit, s'ha produït una escissió en el si de la comunitat d'estadístics entre els "vells" i els "joves".

> "D'una banda hi ha els estadístics «vells», que han après i practicat l'estadística matemàtica clàssica, que busca formalitzar la inducció, seguint els estadístics anglosaxons dels anys 1900 a 1950. D'altra banda, hi ha els estadístics més joves, que sota la mateixa etiqueta d'*estadística* han après unes tècniques molt diferents, que es basen en una eina matemàtica purament algebraica i que pretenen descriure, reduir i classificar les observacions multidimensionals; aquests estadístics no estan gens interessats en la inducció i proclamen amb una gran convicció que l'estadístic ha d'aplicar les seves tècniques d'anàlisi sense fer cap hipòtesi sobre els fenòmens observats. Practiquen «l'anàlisi de dades»."
>
> G. Morlat

Podeu consultar informació sobre l'estadística probabilísta en l'apartat 5 d'aquest mòdul.

L'estadística multidimensional pot servir per a preparar raonaments probabilistes. Segurament és per aquest tipus d'estadística que la denominació de *matemàtiques del resum* resulta d'allò més pertinent. Podem distingir tres grans tipus de mètodes.

1) La classificació

La classificació crea tipologies dins del conjunt d'individus agrupant-los en classes, generalment disjuntes com, per exemple, la classificació ascendent jeràrquica.

2) L'anàlisi factorial

L'anàlisi factorial consisteix a processar grans taules de nombres, difícils de llegir, substituint-les per taules més senzilles que siguin una bona aproximació a les primeres. Els diferents mètodes d'anàlisi factorial (i, en particular, l'anàlisi factorial de les correspondències, que estudiarem a continuació) usen el mateix procediment: a partir d'un núvol de punts (els individus) proveïts de massa (els efectius) en un espai proveït d'una mètrica (que mesura la distància entre els individus) en el qual el gran nombre de dimensions no permet la visualització del núvol, es tracta de trobar els eixos d'inèrcia del núvol i d'obtenir visualitzacions dels plànols formats per les parelles d'eixos. Podríem resumir-ho dient que aquests mètodes d'anàlisi permeten "geometritzar les taules de nombres".

3) L'anàlisi relacional

Més recentment, l'anàlisi relacional reuneix un grup de tècniques d'anàlisi de valors numèrics de variables que permeten buscar una relació estructurada entre aquestes variables. Es fan comparacions per parells de valors. Els valors es processen en la seva forma bruta sense cap codificació prèvia, la qual cosa evi-

ta les pèrdues d'informació. Aquests mètodes ofereixen sovint uns resultats molt interessants quan es manipulen uns volums importants d'informació, la qual cosa és molt habitual en les ciències de la informació.

Tots aquests mètodes tenen com a objectiu extreure "significat" a partir de nombroses observacions efectuades. Els valors produïts per aquestes observacions s'estructuren en taules en les quals, en general, les files representen els individus o els objectes estudiats i les columnes les variables que els caracteritzen, que poden estar codificades segons diferents escales:

- En els mètodes classificatoris, es permuten les files quan es volen classificar els individus, i es permuten les columnes quan es volen classificar les variables.
- Les anàlisis factorials representen les línies (o les columnes) amb un núvol de punts en un espai multidimensional. Les tècniques d'anàlisi factorial permeten projectar aquest núvol de punts en espais llegibles de dues o tres dimensions. Per la seva banda, l'anàlisi factorial de les correspondències assigna un paper simètric a les línies i a les columnes.
- L'anàlisi relacional consisteix a permutar simultàniament les línies i les columnes; aquests mètodes, coneguts com *de seriació*, reparteixen al mateix temps les files i les columnes.

Observació

Els mètodes nous que usen les xarxes neuronals permeten projectar el núvol de punts tenint-ne en compte la densitat.

Els mètodes de cartografia de la informació (paraules associades, cocitacions), encara que usen tècniques de classificació, s'estudien en aquesta assignatura en el mòdul dedicat a l'estadística bidimensional.

Podeu consultar informació sobre cartografia de la informació en el subapartat 3.5 d'aquest mòdul.

4.1. Classificació de ciutats segons la seva producció científica i tècnica (cienciometria): classificació ascendent jeràrquica

Els mètodes de classificació tenen com a objectiu establir una divisió entre els objectes o els individus estudiats segons criteris de proximitat. Es construeix una divisió d'un conjunt d'objectes. La divisió d'aquest conjunt es construeix a partir del coneixement de les distàncies de dues en dues dels objectes.

Exemple

Suposem que decidim caracteritzar el dinamisme científic i tècnic de cinc grans ciutats franceses (París, Lió, Bordeus, Marsella, Rennes) a partir de l'anàlisi de la seva producció científica i tècnica durant un període concret: nombre d'articles referenciats en els grans bancs d'informació, nombre de patents registrades, nombre de projectes científics europeus i nombre de col·loquis internacionals organitzats pels docents i els investigadors d'aquestes ciutats.

Taula 28. Producció científica i tècnica de les ciutats de París i Bordeus

	Article	Patent	Projecte	Col·loqui
París	1.000	100	10	5
Bordeus	990	80	20	10

Observación

Igual que amb els indicadors de dispersió, hi ha múltiples maneres de calcular una distància. Les dues distàncies més conegudes són la distància euclidiana i la distància de χ^2 (que no s'ha de confondre amb la prova de χ^2). Aquí usarem la distància euclidiana.

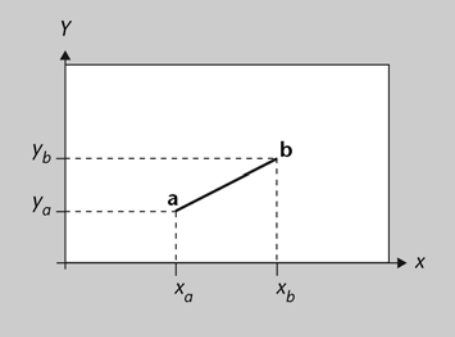
Podeu consultar informació sobre indicadors de dispersió en el subapartat 2.3 d'aquest mòdul.

Cadascuna d'aquestes ciutats es caracteritza, doncs, a partir de quatre nombres que serviran per a calcular la "distància" que mesura les diferències de projecció científica.

Quan tenim dos criteris *a* i *b*, la distància euclidiana és igual al teorema següent (molt conegut i anomenat *de Pitàgores* en la geometria clàssica):

$$d(a,b)=\sqrt{(x_a-x_b)^2+(y_a-y_b)^2}$$

Figura 11. Distància euclidiana (dos criteris)



Aquest teorema es generalitza en un espai multidimensional amb *n* criteris.

Exemple

Les dues ciutats de París i Bordeus es caracteritzen amb els dos conjunts de valors anteriors (vegeu la taula 28). Ens trobem en un espai de dimensió 4 en el qual tots els individus (aquí ciutats) tenen quatre coordenades: article, patent, projecte i col·loqui. La distància euclidiana entre París i Bordeus es calcula amb l'ajuda del teorema de Pitàgores generalitzat.

$$d(\text{París, Bordeus}) = \sqrt{(1.000-900)^2+(100-80)^2+(10-20)^2+(5-10)^2}$$

$$d(\text{París, Bordeus}) = \sqrt{100+400+100+25}$$

per tant, d(París, Bordeus) = 25

Veiem que la diferència entre el nombre d'articles publicats i el nombre de projectes europeus contribueixen de la mateixa manera al càlcul de la distància, encara que aquestes dues variables no tinguin comparació.

També hem calculat les deu distàncies euclidianes entre les cinc ciutats, i els seus valors es donen en una taula de doble entrada. Aquesta taula, evidentment simètrica, presenta zeros en la diagonal (taula 29). És la matriu de semblances.

Taula 29. Distàncies euclidianes entre les ciutats

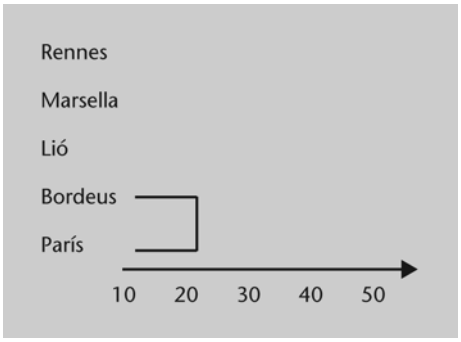
	París	Bordeus	Lió	Marsella	Rennes
París	0				
Bordeus	25	0			
Lió	60	50	0T		
Marsella	100	90	40	0	
Rennes	90	85	50	30	0

Ara classificarem aquestes ciutats agrupant en primer lloc les ciutats més properes, és a dir, aquelles per a les quals la distància és la més petita. Es tracta de París i Bordeus, i d'aquí sorgeix la primera classe de la classificació (Bordeus, París), que aïllem. En cada

etapa del càlcul, el procés de classificació de les ciutats es representa mitjançant un arbre conegut com a *dendrograma* (vegeu la figura 12). La divisió és, doncs:

{Rennes, Marsella, Lió, (Bordeus, París)}.

Figura 12. Dendrograma de la primera classe



A continuació calculem la distància entre cadascuna de les ciutats restants i la classe que acabem de crear. La tècnica aplicada per a calcular la propera classe i les següents es coneix com a *estratègia d'agregació de semblances* o *mètode del salt mínim* (i en anglès, amb el nom de *single linkage*), i consisteix a considerar que la distància entre un element i una classe és la distància més petita entre aquest element i tots els elements de la classe. Així, doncs, tenim:

d((París, Bordeus), Lió)

= Mínim (d(París, Lió),d(Bordeus, Lió))
= Mínim (60, 50) = 50

De la mateixa manera, tenim:

d((París, Bordeus), Marsella) = 90
d((París, Bordeus), Rennes) = 85

Els resultats apareixen representats en la taula 30:

Taula 30. Formació de la primera divisió

	(París,Bordeus)	Lió	Marsella	Rennes
(París, Bordeus)	0			
Lió	50	0		
Marsella	90	40	0	
Rennes	85	50	30	0

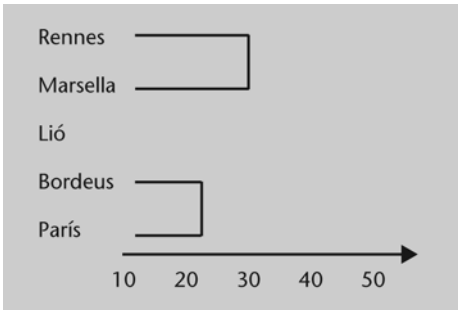
La distància més petita es dóna entre Rennes i Marsella, i és igual a 30. La divisió és, doncs:

{(Rennes, Marsella), Lió, (Bordeus, París)}

El dendrograma que il·lustra la formació de la segona classe es pot veure en la figura 13.

Les mesures de les distàncies es representen en la taula 31.

Figura 13. Dendrograma de la segona classe

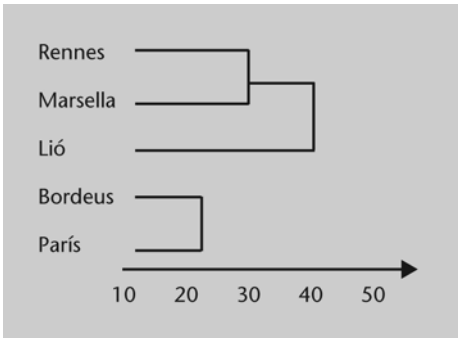


Taula 31. Formació de la segona divisió

	(París,Bordeus)	Lió	(Marsella,Rennes)
(París, Bordeus)	0		
Lió	50	0	
(Marsella, Rennes)	85	**40**	0

Per a la tercera divisió, els elements més propers són les classes (Marsella, Rennes) i Lió, i d'aquí surten el dendrograma de la figura 14 i la taula 32:

Figura 14. Dendrograma de la tercera classe



Taula 32. Formació de la tercera divisió

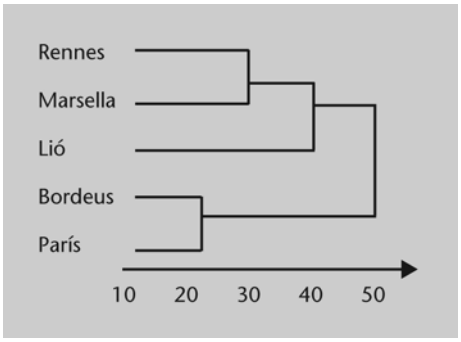
	(París,Bordeus)	(Lió,Marsella,Rennes)
(París, Bordeus)	0	
(Lió, Marsella, Rennes)	50	0

Així, doncs, la divisió completa és:

{(Rennes, Marsella, Lió), (Bordeus, París)}

El dendrograma final és el següent:

Figura 15. Arbre de classificació de les ciutats segons la seva producció científica



Basant-nos en criteris de producció científica, és possible agrupar les ciutats en classes segons la similitud de la seva producció d'articles, patents, projectes i col·loquis. Les ciutats de Rennes i de Marsella tenen un perfil de producció científica molt diferent del de les ciutats de Bordeus i París.

4.2. Tipologia de les produccions literàries científiques i tècniques segons les disciplines (infometria): anàlisi factorial de les correspondències

Aquesta anàlisi es fa descomponent la taula inicial de valors numèrics en una suma de taules, que són els productes de factors (mitjançant una factorització, del tipus ben conegut en àlgebra, que consisteix a factorizar una expressió algebraica), i d'això el nom d'*anàlisi factorial*. Aquest mètode és una prolongació de la distribució creuada de dues variables nominals.

Podeu consultar informació sobre distribució creuada en el subapartat 3.1 d'aquest mòdul.

Exemple

Tenim la taula següent T_1, que classifica, dins de quatre grans disciplines, la producció de literatura científica i tècnica dels investigadors d'una universitat. Aquest és un exemple d'una taula de contingència (vegeu la taula 33).

Bibliografia

Els valors, les taules i les gràfiques s'han adaptat del llibre de **Philippe Cibois** (2000). *Analyse factorielle*. París: PUF ("Que sais-je?", núm. 2095).

Taula 33. Els efectius observats (T_1)

	Article	Llibre	Patent	Total
Ciències de la matèria	13	2	5	20
Ciències de la vida	20	2	8	30
Ciències socials	10	5	5	20
Ciències de l'enginyeria	7	1	22	30
Total	**50**	**10**	**40**	**100**

Aquesta taula ens diu que el 50% de la producció científica dels investigadors consisteix en articles (fila 5, columna 1). Si apliquem aquest percentatge a les ciències de la matèria, es constata que, sobre una producció total de vint, només hi hauria d'haver deu articles:

$$\frac{20 \cdot 50}{100} = 10$$

Però resulta que n'hi ha tretze. Els científics dedicats a les ciències de la matèria produeixen bàsicament articles. Així, doncs, no hi ha independència entre la disciplina i el tipus de producció escollit; el respecte de la proporció mitjana correspon al que es coneix com *la situació d'independència*, que es representa en la taula T_2 (taula 34).

Taula 34. Els efectius teòrics (T_2)

	Article	Llibre	Patent	Total
Ciències de la matèria	10	2	8	20
Ciències de la vida	15	3	12	30
Ciències socials	10	2	8	20
Ciències de l'enginyeria	15	3	12	30
Total	50	10	40	100

Els efectius d'aquesta taula són el que es coneix com a *efectius teòrics*.

Podeu consultar informació sobre efectius teòrics en el subapartat 3.1.5 d'aquest mòdul.

El càlcul de les desviacions cap a la independència es fa calculant la diferència entre l'efectiu observat i l'efectiu teòric. D'aquesta manera, s'obté una tercera taula R_1 (taula

35), coneguda com a *taula de les desviacions cap a la independència*, que conté una informació més interessant per a la interpretació:

$$R_1 = T_1 - T_2$$

Taula 35. Les desviacions cap a la independència (R_1)

	Article	Llibre	Patent	Total
Ciències de la matèria	3	0	–3	0
Ciències de la vida	5	–1	–4	0
Ciències socials	0	3	–3	0
Ciències de l'enginyeria	–8	–2	10	0
Total	0	0	0	0

Aquests resultats s'interpreten de la manera següent:

- A les desviacions positives corresponen les opcions privilegiades

  Les ciències de la matèria i les ciències de la vida (desviacions de +3 i +5) produeixen, de mitjana, més articles que les altres dues disciplines. Les ciències socials (desviacions +3) produeixen més llibres que la mitjana de totes les disciplines, igual que les ciències de l'enginyeria produeixen més patents que la mitjana de totes les disciplines.

- A les desviacions nul·les corresponen les distribucions segons el percentatge mitjà

  La producció literària en ciències socials (desviació 0) consta de tants articles com la mitjana del conjunt de les disciplines.

- A les desviacions negatives corresponen dèficits

  Les ciències de l'enginyeria (desviació –8) produeixen menys articles que la mitjana de totes les disciplines.

En resum:

Atracció entre fila i columna	desviació > 0
Independència entre fila i columna	desviació = 0
Repulsió entre fila i columna	desviació < 0

En descompondre la taula inicial en una suma de dues taules hem començat una descomposició factorial:

$$T_1 = R_1 + T_2$$

Això vol dir que se substitueix una taula de valors numèrics per una suma de dues taules en la qual la segona presenta la característica particular de tenir uns marges que constitueixen un resum perfecte.

En efecte, conèixer els marges de la taula T_2 equival a conèixer la taula completa, ja que aquesta última s'obté mitjançant un simple producte de factors.

La taula de les desviacions encara no és una taula simple que es pugui resumir amb un producte de factors. Però podem continuar amb la descomposició factorial buscant dues taules, T_3 i T_4, que resumeixen R_1. Aquí hi ha la seva descomposició en una suma de dues taules "simples" que es poden resumir amb un joc de coeficients marginals que només cal multiplicar entre ells per a obtenir una casella de la taula:

$$\begin{bmatrix} 3 & 0 & -3 \\ 5 & -1 & -4 \\ 0 & 3 & -3 \\ -8 & -2 & 10 \end{bmatrix} = \begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ 2 & 2 & -4 \\ -4 & -4 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \\ -4 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & -2 \end{bmatrix} + \begin{bmatrix} 2 & -1 & -1 \\ 4 & -2 & -2 \\ -2 & 1 & 1 \\ -4 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ -1 \\ -2 \end{bmatrix} \times \begin{bmatrix} 2 & -1 & -1 \end{bmatrix}$$

$$R_1 = T_3 + T_4$$

Es verifica que:

$$T_1 = T_2 + T_3 + T_4$$

Hem descompost la primera taula en tres taules simples. Podríem pensar que amb aquesta descomposició compliquem el problema? La veritat és que no, en la mesura en què les tres taules T_2, T_3 i T_4 es poden obtenir multiplicant una fila per una columna.

Observació

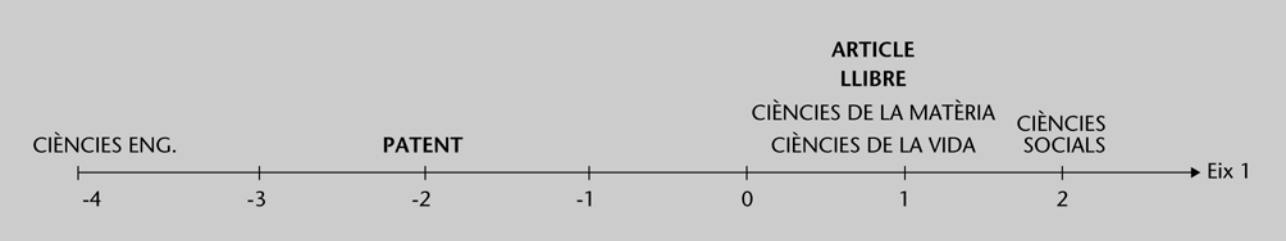
Les tècniques matemàtiques que permeten fer aquesta descomposició i calcular T_3 i T_4 procedeixen de l'àlgebra matricial, però no les expliquem.

La representació gràfica

L'interès d'aquesta descomposició es pot veure millor en la representació gràfica següent. Associem a cadascuna de les taules T_3 i T_4, sorgides de la descomposició final, un eix amb un sistema de coordenades. Sobre aquest eix es representen els valors dels marges de fila i de columna.

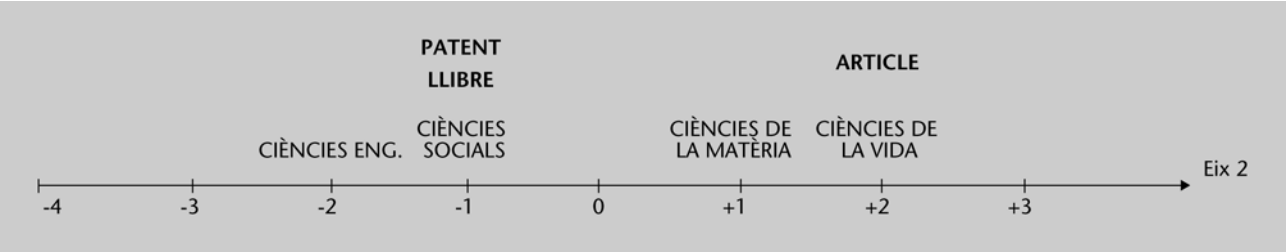
Els valors de T_3 es representen sobre un eix, identificat com a eix 1:

Gràfica 31. La literatura científica i tècnica segons les disciplines (eix 1)



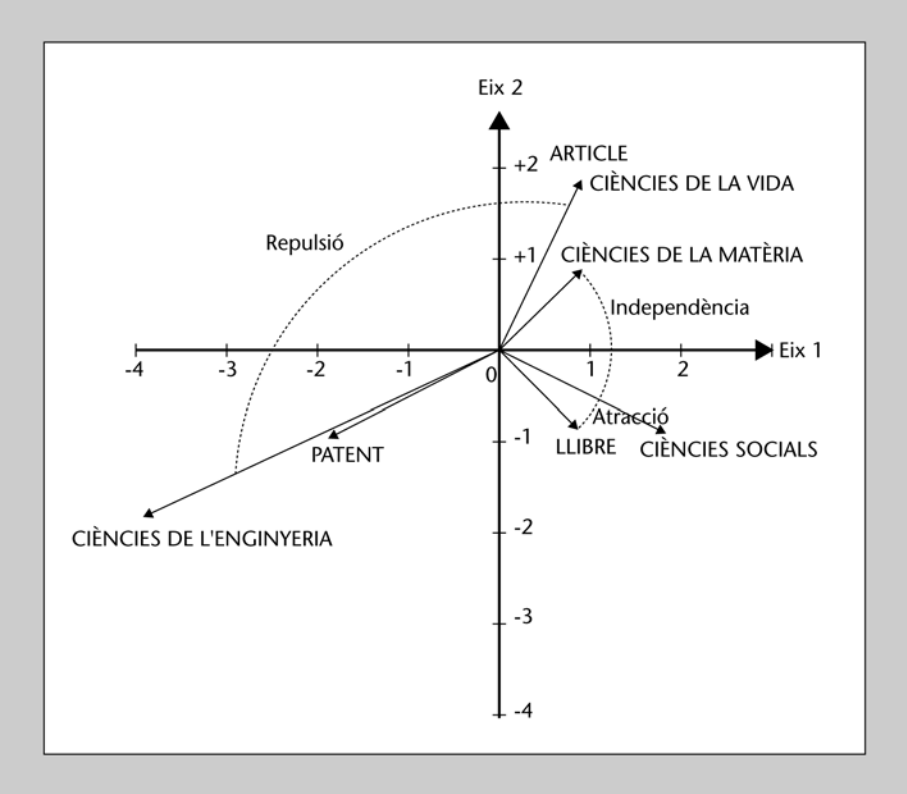
Els valors de T_4 es representen sobre un altre eix, identificat com a eix 2:

Gràfica 32. La literatura científica i tècnica segons les disciplines (eix 2)



Si dibuixem una gràfica-plànol que associï l'eix 1 i l'eix 2 (gràfiques 31 i 32), obtenim una representació dels tipus de la producció de literatura científica i tècnica segons les disciplines:

Gràfica 33. La literatura científica i tècnica segons les disciplines (plànol dels eixos 1 i 2)



Les aportacions de la representació planar

Si dibuixem les rectes que uneixen cadascun dels punts representatius i l'origen, aquestes rectes poden estar dues a dues:

- En **conjunció**, si entre elles es forma un angle nul o inferior a 90° (angle recte). Hi ha atracció entre la patent i les ciències de l'enginyeria i també entre el llibre i les ciències socials, que són les produccions privilegiades d'aquestes dues ciències.

- En **quadratura**, si entre elles es forma un angle recte. Hi ha independència entre el llibre i les ciències de la matèria, la qual cosa no significa que aquestes ciències no produeixin llibres, sinó que en produeixen igual que la mitjana de les disciplines.

- En **oposició**, si entre elles es forma un angle superior a 90°. Hi ha repulsió entre l'article i les ciències de l'enginyeria, la qual cosa significa que les ciències de l'enginyeria produeixen menys llibres que la mitjana.

Complement: càlcul de χ^2 per casella (taula 36).

Taula 36. χ^2 per casella (i el signe de la desviació cap a la independència)

	Article	Llibre	Patent
Ciències de la matèria	+0,9	0	–1,13
Ciències de la vida	+1,67	–0,33	–1,33
Ciències socials	0	+4,5	–1,13
Ciències de l'enginyeria	–4,27	–1,33	+8,33

Total = 24,92

Sumant les χ^2 per casella obtenim la χ^2 total, que és de 24,92. La lectura de la taula de χ^2 ens permet afirmar que les dues variables són depenents, és a dir, que tenim menys d'una possibilitat entre cent d'equivocar-nos si diem que les dues variables *tipus de publicació* i *disciplina científica i tècnica* són dependents una de l'altra. Els forts valors de χ^2 i el signe algebraic reflecteixen les atraccions (signe positiu) o les repulsions (signe negatiu) entre les files i les columnes.

Podeu consultar informació sobre taules estadístiques en l'annex 2, "Taules estadístiques".

Atracció → Patent / ciències de l'enginyeria
Atracció → Llibre / ciències socials
Atracció → Article / ciències de la vida

Repulsió → Article / ciències de l'enginyeria

Aquestes quatre associacions expliquen el 75% de χ^2.

La representació planar de la gràfica 33 pot induir a error. En efecte, es mostra que la dimensió necessària per a representar, en l'anàlisi factorial de les correspondències, l'espai d'una taula de n files i m columnes (n objectes en un espai de m dimensions o bé, al revés, m objectes en un espai de n dimensions) és igual al mínim de $(n - 1, m - 1)$. Així, doncs, en el nostre exemple, la representació en dues dimensions és exacta. Si el nombre de dimensions és superior a dues és possible que es donin errors de lectura relacionats amb les projeccions dels valors representats sobre els eixos. En general, els punts es projecten sobre els plànols i la lectura es facilita amb un coeficient que mesura la qualitat de la representació.

4.3. Xarxes de col·laboracions en l'àmbit industrial (cienciometria): anàlisi relacional

Avui dia, un nombre cada vegada més alt d'activitats humanes mobilitzen unes pràctiques de treball col·lectives, la qual cosa es veu afavorida per eines com el correu electrònic, el programari de grup i el *flowware*. És així com neixen diverses relacions estructurades, com les col·laboracions en el sector acadèmic, que és important descriure i mesurar.

Exemple

Durant els últims cinc anys, vuit laboratoris d'investigació industrial, identificats de la A a la H, han registratcent vuitanta-vuit patents. Això s'ha fet gràcies a col·laboracions amb catorze laboratoris universitaris, identificats de la M a la Z (vegeu la taula 37).

Taula 37. Patents registrades dins del marc de la col·laboració empreses-universitats

Empreses / Universitats	A	B	C	D	E	F	G	H
M	5	0	0	1	12	1	8	0
N	0	4	10	0	0	0	0	6
O	0	0	1	0	0	0	0	6
P	0	9	9	0	0	0	0	1
Q	0	1	0	0	0	0	0	3
R	0	12	1	0	0	0	0	0
S	3	0	0	7	8	0	2	0
T	0	2	4	0	0	0	0	1
U	0	2	2	0	0	0	0	2
V	1	0	0	2	3	4	5	0
W	0	0	0	3	9	1	0	0
X	0	1	9	0	0	0	0	8
Y	1	0	0	4	1	1	1	0
Z	0	2	1	0	0	0	0	8

Les dades recollides i ordenades en la taula anterior permeten veure d'una manera molt ràpida totes les patents que s'han registrat en el marc d'aquestes col·laboracions.

Després de calcular les distàncies entre els laboratoris industrials o els laboratoris universitaris, podríem, igual que en el subapartat 4.1, classificar aquests laboratoris. Però el que ens interessa aquí és posar de manifest una xarxa de col·laboracions entre les universitats i les empreses. Per a això, com que el nombre de no-col·laboracions és elevat (59%), permutem les files i les columnes d'aquesta taula. Així obtenim successivament les taules 38, 39 i 40 després de permutar les files 1 i 14, després les columnes 1 i 8 i, finalment, les files 7 i 12.

Taula 38. Taula 37 després de permutar les files 1 i 14

Empreses / Universitats	A	B	C	D	E	F	G	H
Z	0	2	1	0	0	0	0	8
N	0	4	10	0	0	0	0	6
O	0	0	1	0	0	0	0	6
P	0	9	9	0	0	0	0	1
Q	0	1	0	0	0	0	0	3
R	0	12	1	0	0	0	0	0
S	3	0	0	7	8	0	2	0
T	0	2	4	0	0	0	0	1
U	0	2	2	0	0	0	0	2
V	1	0	0	2	3	4	5	0
W	0	0	0	3	9	1	0	0

Universitats \ Empreses	A	B	C	D	E	F	G	H
X	0	1	9	0	0	0	0	8
Y	1	0	0	4	1	1	1	0
M	5	0	0	1	12	1	8	0

Taula 39. Taula 38 després de permutar les columnes 1 i 8

Universitats \ Empreses	H	B	C	D	E	F	G	A
Z	8	2	1	0	0	0	0	0
N	6	4	10	0	0	0	0	0
O	6	0	1	0	0	0	0	0
P	1	9	9	0	0	0	0	0
Q	3	1	0	0	0	0	0	0
R	0	12	1	0	0	0	0	0
S	0	0	0	7	8	0	2	3
T	1	2	4	0	0	0	0	0
U	2	2	2	0	0	0	0	0
V	0	0	0	2	3	4	5	1
W	0	0	0	3	9	1	0	0
X	8	1	9	0	0	0	0	0
Y	0	0	0	4	1	1	1	1
M	0	0	0	1	12	1	8	5

Taula 40. Taula 39 després de permutar les files 7 i 12

Universitats \ Empreses	H	B	C	D	E	F	G	A
Z	8	2	1	0	0	0	0	0
N	6	4	10	0	0	0	0	0
O	6	0	1	0	0	0	0	0
P	1	9	9	0	0	0	0	0
Q	3	1	0	0	0	0	0	0
R	0	12	1	0	0	0	0	0
X	8	1	9	0	0	0	0	0
T	1	2	4	0	0	0	0	0
U	2	2	2	0	0	0	0	0
V	0	0	0	2	3	4	5	1
W	0	0	0	3	9	1	0	0
S	0	0	0	7	8	0	2	3
Y	0	0	0	4	1	1	1	1
M	0	0	0	1	12	1	8	5

Després d'aquesta última permutació, en la taula 40 veiem dues xarxes de col·laboració entre les empreses H, B i C i les universitats Z, N, O, P, Q, R, X, T, O, d'una banda, i entre les empreses D, E, F, G, A i les universitats V, W, S, Y i M, d'una altra. Aquestes tècniques "d'exploració de dades", en les quals permutem les files i les columnes per construir blocs, són possibles gràcies a algorismes de seriació implantats en els programes informàtics d'estadística.

Observació

Els mètodes que permeten escollir de manera òptima les línies i les columnes que es permutaran estan relacionats amb els mètodes matemàtics d'anàlisi relacional. Com a exemple, només direm que si intentéssim crear totes les permutacions ($14! = 1 \cdot 2 \cdot 3 \cdot 4 \ldots 14$, el que es coneix com a 14 factorial) possibles sense cap estratègia, podríem crear $14! \cdot 8!$, és a dir, $35 \cdot 10^{14}$ permutacions.

4.4. En resum

Davant d'informacions cada vegada més nombroses i variades, de comunicacions i col·laboracions cada vegada més nombroses entre les persones, la qual cossa genera un veritable vertigen informatiu, ja no n'hi ha prou d'estudiar només una o dues variables ni de resumir uns pocs valors. Tenim al davant una gran profusió de variables i, en conseqüència, és necessari analitzar una gran profusió de valors numèrics. Resulta difícil triar les variables i les relacions que cal analitzar. És aquí on entra en joc l'estadística multidimensional.

En primer lloc, amb l'ajuda dels mètodes classificatoris, busquem classificar les variables de dues en dues per a dividir els grups d'individus o d'objectes que generen, comuniquen i usen aquestes informacions.

A continuació, gràcies a l'anàlisi relacional i als diferents mètodes d'anàlisi multidimensional, extraiem les relacions estructurades que hi ha entre aquestes diferents variables. Després deixem les representacions lineals (1D) o planars (2D) per a entrar en els espais tridimensionals (3D) a la cerca de núvols de punts, que posaran de manifest unes relacions riques de proximitat de les variables entre elles.

Però en aquests espais multidimensionals descobrirem, gràcies a l'anàlisi factorial, per exemple, altres proximitats, o bé posarem en relleu altres relacions estructurades entre aquestes nombroses variables.

5. L'estadística probabilista

Igual que qualsevol activitat que mobilitza moltes persones, l'ús d'una biblioteca, d'un museu o d'un servidor per part dels usuaris és el resultat d'un gran nombre de fets aleatoris, de fets que depenen de l'atzar, de la sort. La teoria de les probabilitats estudia aquestes situacions, aquests fets aleatoris.

Cada ús té els seus objectius propis; cada objecte portador d'informació, cada informació, té una audiència diferent. El comportament individual d'un usuari o el grau d'ús (la usabilitat) d'un objecte, d'una informació no es poden predir amb certesa. Però sí que es pot preveure el comportament mitjà d'un conjunt d'individus amb una precisió que augmenta amb el nombre d'individus i la durada de l'observació (diverses setmanes, mesos, anys).

En aquest cas, el càlcul de les probabilitats és molt important. Permet resumir un conjunt de fets aleatoris –com els usos quotidians d'un sistema d'informació, d'un objecte portador d'informació, d'una informació– en una magnitud numèrica, coneguda com a *probabilitat*, que expressa el caràcter probable, no cert, d'aquests fets o d'aquests usos. Aquesta magnitud permet predir l'ús mitjà d'objectes i d'informacions per part d'un conjunt de persones, la desviació mitjana en relació amb aquest ús i altres criteris bastant precisos i, en conseqüència, molt útils, per a fonamentar decisions, sempre que tinguem en compte centenars d'usuaris, d'objectes o d'informacions al llarg d'intervals bastant llargs.

Per exemple, un usuari que acudeix a la biblioteca hi pot fer diverses tasques i fer-ne diversos usos. Pot consultar un o diversos llibres abans de marxar, agafar en préstec un o diversos llibres, usar el catàleg informatitzat o fins i tot no fer res. Dir que la probabilitat que el proper usuari faci l'enèsim ús és igual al nombre P_n equival a dir que una fracció *Pn* dels propers usuaris farà l'enèsim ús. Així, doncs, veiem que les probabilitats, les proporcions i els percentatges estan relacionats.

- **Probabilitat i recompte**

El concepte de probabilitat, tal com es fa servir habitualment en la nostra societat, ha donat lloc a nombrosos debats. La concepció més antiga del concepte de probabilitat és la concepció objetivista heretada dels jocs d'atzar, en els quals el càlcul de les probabilitats es redueix sovint a una qüestió de recompte.

Exemple

Quan llancem una moneda, el resultat, una vegada ha caigut, és {CARA} o {CREU}. El conjunt dels fets elementals, és a dir, dels resultats possibles (conegut com a *univers*), està

constituït per dos elements: {CARA}, {CREU}. No podem preveure per endavant quin serà el resultat d'un llançament. En aquest cas, diem que els dos fets {CARA} i {CREU} són equiprobables. Com que cadascun té una possibilitat entre dues de produir-se, els atribuïm una probabilitat igual a $\frac{1}{2}$:.

$$P(\{CARA\}, \{CREU\}) = \frac{1}{2} + \frac{1}{2}$$

La suma de les probabilitats és igual a 1:

$$P(\{CARA\}, \{CREU\}) = 1$$

Suposem que fem un gran nombre de llançaments, representat per *N*. Cada vegada que es produeix el fet {CREU} ho anotem i acumulem aquesta quantitat en una variable *NBCREU*. Constatem experimentalment que la relació:

$$\frac{NBCREU}{N} \text{ tendeix a } \frac{1}{2}$$

a mesura que *N* és cada vegada més alt, és a dir, quan tendeix a l'infinit (representat per ∞).

Podeu consultar informació sobre el límit en matemàtiques en l'apartat 1 del mòdul 3, "Matemàtica de la informació".

La distribució de la probabilitat és una sèrie finita o infinita de nombres positius compresos entre 0 i 1, amb una suma igual a 1.

Sèries i funcions

Més endavant s'usaran alguns resultats sobre les sèries i les funcions, en l'apartat 2 i en l'apartat 3 del mòdul 3, "Matemàtica de la informació".

Cadascun dels nombres inferior o igual a 1 mesura la probabilitat d'un fet.

Una distribució de probabilitat és una sèrie de nombres p_i que compleix les dues propietats següents:

$$0 \leq p_i \leq 1 \quad \sum_{i=1}^{n} p_i = 1$$

En què *n* és un nombre enter positiu finit o infinit.

En aquest cas, es parla d'una distribució de probabilitat discreta finita o infinita.

- **Lleis de probabilitat**

Quan dibuixem les distribucions de freqüències de probabilitats, constatem que aquestes distribucions són regulars i que segueixen lleis. Aquestes lleis, conegudes com a *lleis de la probabilitat*, s'usen en molts camps. Més endavant trobarem les formulacions matemàtiques de les lleis de probabilitat usades en les ciències de la informació: la llei geomètrica, la llei binomial, la llei binomial negativa, la llei de Poisson i la llei de l'avantatge acumulat.

Podeu consultar informació sobre la representació gràfica de les distribucions en el subapartat 2.3.1 d'aquest mòdul.

5.1. Probabilitat de préstec d'una obra en una biblioteca (bibliometria): llei geomètrica, llei binomial negativa

Suposem que la circulació d'obres en una biblioteca és un procés aleatori. El fet elemental observat és:

"Una obra de la biblioteca s'ha deixat *i* vegades en un període determinat."

Si U_i és el nombre d'obres que s'han deixat *i* vegades, *i* varia entre 1 i *n*, i *n* designa el nombre màxim de préstecs d'una obra i, a partir d'aquestes dades, calculem el nombre total de préstecs, representat per *E:*

$$E = \sum_{i=1}^{n} U_i$$

La probabilitat que una obra es deixi *i* vegades és:

$$p_i = \frac{U_i}{E}$$

5.1.1. Llei geomètrica

Aquesta llei permet descriure els fenòmens d'ús en una biblioteca, i també els serveis d'Internet i audiovisuals (préstec, consulta, comanda, accés, escolta, visualització, etc.).

Exemple

Probabilitat de préstec d'una obra en una biblioteca

En una biblioteca, el nombre anual de préstecs d'una col·lecció de 289 obres és la següent:

Taula 41. Distribució dels préstecs, observats i calculats (llei geomètrica)

Préstecs	Obres	Obres llei geomètrica
0	228	208,08
1	31	59,60
2	16	17,25
3	7	4,87
4	3	1,27
5	4	0,40

Total = 289

Determinem simplement:

- Nombre d'obres actives: 61, sabent que hi ha 228 obres que no s'han agafat en préstec.
- Nombre total de préstecs: 116.
- Circulació mitjana d'una obra (també coneguda com a *rotació mitjana* o *usabilitat*): 0,4 préstecs/any.
- Desviació estàndard: 0,94.

La circulació mitjana calculada s'ha fet sobre el fons total. La taxa de rotació sobre el fons actiu és de $\frac{116}{61} = 1{,}9$, la qual cosa vol dir que les obres que circulen es deixen unes dues vegades a l'any.

Suposem que la distribució observada segueix una llei geomètrica.

Podeu consultar informació sobre la llei geomètrica en el subapartat 2.1 del mòdul 3, "Matemàtica de la informació".

Llei geomètrica $G(u)$

Suposem que un fet A es reprodueix en el temps amb una probabilitat constant:

$$p(A) = u$$

La probabilitat p_i que tingui lloc exactament i vegades $(0 \leq i < \infty)$ és:

$$p_i = (1-u)\, u^i \qquad i = 0,1,2\ldots$$

Obtenim una distribució de probabilitat:

$$\sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} (1-u)\, u^i = (1-u) \sum_{i=0}^{i=\infty} u^i = (1-u) \cdot \frac{1}{1-u} = 1$$

Hi ha diversos mètodes per a calcular el paràmetre u. El mètode més habitual és el dels moments. El valor dels paràmetres es calcula a partir dels moments observats. Si E designa el moment d'ordre 1, que és la mitjana, tenim:

Podeu consultar informació sobre el càlcul de moments en l'annex 3, "Càlcul de moments".

$$E = \frac{u}{1-u} \text{ és a dir } u = \frac{0,4}{1+0,4} = 0,29$$

La probabilitat que una obra es deixi i vegades s'escriu:

$$p_i = (1-u) \cdot u^i = 0,71 \cdot 0,29^i$$

en què i adquireix els valors 0, 1, 2, 3...

La col·lecció és de 289 obres, per la qual cosa els nombres teòrics d'obres deixades en préstec i vegades s'escriuen:

$$U_i \text{ (eòric)} = 0,71 \cdot 0,29^i \cdot 289$$

en què i adquireix els valors 0, 1, 2, 3...
Llavors, si $i = 2$,

$$U_2 = 0,71 \cdot 0,29^2 \cdot 289 = 17,25$$

Una mirada ràpida al càlcul dels efectius previstos per aquesta llei mostra que els resultats són mediocres (taula 41).

5.1.2. Llei binomial i llei binomial negativa

El resultat anterior no és gens sorprenent; en efecte, l'experiència demostra sovint que la llei geomètrica no sempre és una "bona llei" per a fer una previsió dels usos quan també es té en compte el no-ús. El no-ús es dóna quan la freqüència d'usos és zero. En aquests casos s'usa la llei binomial negativa.

Comencem per estudiar la llei binomial.

Llei binomial *B*(*n*, *u*)

Imaginem que una experiència porta a dos esdeveniments contraris representats per A i $\overline{A}$ (dues alternatives les probabilitats de les quals són constants –d'aquí el terme *binomial*); per exemple, la cara i la creu per al llançament d'una moneda.

Si l'esdeveniment A es produeix amb la probabilitat u, l'esdeveniment $\overline{A}$ es produeix llavors amb la probabilitat 1– u.

Repetirem aquesta experiència n vegades de manera independent i idèntica: la probabilitat p_i que l'esdeveniment A es produeixi i vegades ($0 \leq i \leq n$) és igual a $p_i = C_n^i u^i (1-u)^{n-i}$.

Aquesta llei de paràmetres n i u es coneix de vegades com a *llei de l'atzar* i s'escriu $B(n,u)$.

C_n^i designa el nombre de combinacions de i elements dins d'un conjunt de n elements (també parlem de *combinació sense repetició*). És el coeficient del binomi de Newton.

$$C_n^i = \frac{(n-i+1)(n-i+2)\ldots n}{2 \cdot 3 \ldots i} = \frac{n!}{(n-i)!i!}$$

Exemple: per a $n = 3$, $i = 1$, tenim $C_3^1 = \frac{3}{1}$; per a $n = 3$, $i = 2$, tenim $C_3^2 = \frac{2 \cdot 3}{2} = 3$; per a $n = 4$, $i = 2$, tenim $C_4^2 = 6$.

Nosaltres no usarem aquesta llei perquè implica que les obres es deixen d'una manera totalment aleatòria, la qual cosa no és veritat. En lloc d'això usarem la llei binomial negativa (generalització de la llei geomètrica), que es relaciona amb les lleis de la informació i que s'aplica millor.

Podeu consultar informació sobre les lleis de la informació en el subapartat 5.3 d'aquest mòdul.

Llei binomial negativa *Bn*(*r, u*)

Tenim una urna que conté una proporció $1 - u$ de boles vermelles. Procedim a extreure boles successivament, i les tornem a introduir en l'urna, fins a obtenir un total de r boles vermelles. La probabilitat p_i d'obtenir r boles vermelles després de i extraccions en què $i = r, r + 1, r + 2$ s'escriu:

$$u_i = C_{i-1}^{r-1}(1-u)^r u^{i-r}$$

Aquesta llei de probabilitat definida pels dos paràmetres r (nombre enter) i u (nombre comprès entre 0 i 1) es generalitza per a un nombre r qualsevol que no és necessàriament sencer. La definició usada aquí serà la següent. Si r designa un nombre positiu i u un nombre comprès entre 0 i 1, definim la llei binomial negativa representada com a $Bn(r,u)$ com a:

$$Bn(r,u)(0) = (1-u)^r$$

$$Bn(r,u)(i) = r(r+1)\ldots(r+i-1)\frac{(1-u)^r}{i!}u^i \quad i = 1,2,\ldots$$

Nota

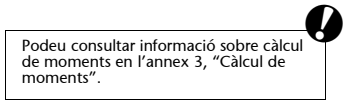
Si $r = 1$, $Bn(1,u)$ és geomètric $G(u)(i) = (1 - u) \cdot u^i$.

Posarem a prova aquesta llei binomial negativa amb els valors anteriors.

Taula 42. Distribució dels préstecs, observats i calculats (llei binomial negativa calculada amb els valors u i r no arrodonits)

Préstecs	Obres	Obres llei binomial negativa
0	228	222,34
1	31	40,26
2	16	14,66
3	7	6,24
4	3	2,84
5	4	1,35
	Total = 289	

El mètode usat per a calcular els paràmetres de la llei és el mètode dels moments. L'esperança E és:



Podeu consultar informació sobre càlcul de moments en l'annex 3, "Càlcul de moments".

$$E = \frac{r \cdot u}{(1-u)}$$

I la variància V és:

$$V = \frac{r \cdot u}{(1-u)^2}$$

Siguin:

$$u = 1 - \frac{E}{V} \text{ i } r = \frac{E^2}{V - E}$$

és a dir: $u = 0{,}55$ i $r = 0{,}33$.

Els nombres teòrics d'obres deixades en préstec s'escriuen:

$$U_i = p_i \cdot N$$

en què N és l'efectiu total i p_i la probabilitat que una obra s'agafi i vegades.

$$U_0(\text{teòrics}) = (1 - 0{,}55)^{0{,}33} \cdot 289$$

$$U_i(\text{teòrics}) = (1-0{,}55)^{0{,}33} \cdot 0{,}33(1+0{,}33)\ldots(0{,}33+i-1)(0{,}55)^i \cdot \frac{1}{i!} \cdot 289$$

en què i adopta els valors 1, 2, 3...

Tal com veiem en la taula 42 i en la gràfica 34, un ajust de tipus binomial negatiu és millor que un ajust de tipus geomètric.

Prova d'ajust

Podeu consultar informació sobre les proves en estadística en l'annex 1, "Les proves d'estadística".

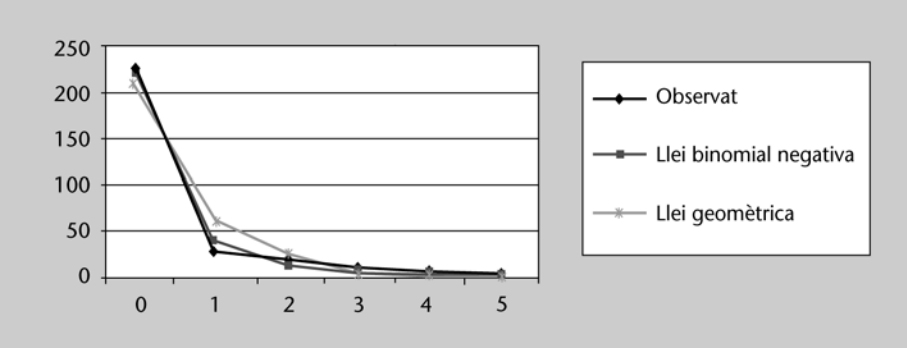
La prova de χ^2 es fa agrupant els efectius de les classes 4 i 5:

$$\chi^2_{\text{calculada}} = \frac{(228-222{,}34)^2}{222{,}34} + \frac{(31-40{,}26)^2}{40{,}26} + \frac{(16-14{,}66)^2}{14{,}66} + \frac{(7-6{,}24)^2}{6{,}24} + \frac{(7-4{,}19)^2}{4{,}19}$$

$$\chi^2_{\text{calculada}} = 1{,}70$$

Amb dos graus de llibertat (ddl = 5 – 2 – 1), arribem a la conclusió que la distribució dels préstecs és de tipus binomial negatiu. L'ajust és satisfactori.

Gràfica 34. Ajust dels préstecs mitjançant la llei binomial negativa



La circulació de les obres en una biblioteca, doncs, sovint es pot descriure amb una llei binomial negativa.

5.2. Extracció del vocabulari representatiu d'un conjunt d'informacions (infometria): llei de Poisson

En fer estudis d'indexació automàtica de documents textuals, és habitual recórrer a la teoria de les probabilitats. Hi ha dues situacions possibles.

5.2.1. El conjunt d'informacions és petit

Reprenguem una analogia molt coneguda per descriure aquest conjunt:

- El conjunt s'assimila a una urna representada per U.
- El document portador d'informació s'assimila a una bola d'aquesta urna.
- El document és una successió de paraules, per la qual cosa, per una paraula determinada, el document en qüestió conté aquesta paraula o no la conté, és a dir, que la bola és de color vermell si el document que representa conté aquesta paraula (èxit) o de color blanc si no la conté (fracàs).

Si M és una paraula concreta, tenim que:

- N és el nombre de boles de l'urna (grandària del conjunt).
- R és el nombre de boles vermelles (nombre de documents que contenen la paraula M) ($R \leq N$).
- $N - R$ representa, doncs, el nombre de boles blanques de l'urna (nombre de documents que no contenen la paraula M).

Bibliografia

D. Mulet-Marquis (1981). *Construction automatique d'un langage de représentation d'information spécialisé: structuration du vocabulaire*. Tesi. Universitat de Lió 1.

Des del punt de vista documental, la realització d'aquest esdeveniment és la selecció d'un conjunt de r documents en els quals apareix la paraula M dins d'un lot de grandària n agafat d'un fons documental de grandària N. Per a fer aquest càlcul usem tècniques de recompte.

Tenim que A i B són els esdeveniments següents:

A: conjunt de les parts de U que contenen exactament n boles $n \leq N$,

i B: conjunt de les parts de U que contenen exactament r boles vermelles $r \leq R$.

Amb l'ajuda del càlcul de recomptes, veiem que la probabilitat P_r d'obtenir r boles vermelles dins d'un lot de n boles és:

$$P_r = \frac{C_R^r}{C_N^n} \cdot C_{N-r}^{n-r}$$

Podeu consultar informació sobre la funció logarítmica en el subapartat 3.2 del mòdul 3, "Matemàtica de la informació".

Aquesta probabilitat és la probabilitat que tingui lloc l'esdeveniment *B* sabent que l'esdeveniment *A* ha tingut lloc; aquesta probabilitat s'escriu *P*(*B*/*A*).

El càlcul de les diferents probabilitats per als valors de *r* (*r* = 1, 2..., *R*) genera una llei que es coneix com a *llei hipergeomètrica*. Aquesta llei generalitza la llei binomial (en aquest cas, tenim en compte la grandària de l'urna *N*). En efecte, sota certes condicions, que es donen per a la majoria dels conjunts d'informacions, aquesta llei convergeix cap a una llei binomial de paràmetres *n* i *u* en què *u* és igual que la relació $\frac{R}{N}$, és a dir, representa la probabilitat d'aparició d'una paraula M en el fons documental.

5.2.2. El conjunt d'informacions és gran

Quan *N* creix (és el cas dels fons documentals), la llei binomial se substitueix per la llei de Poisson. En efecte, si usem el resultat anterior i els resultats teòrics que segueixen, podem demostrar que la llei binomial tendeix cap a una llei de Poisson de paràmetres $\frac{n \cdot R}{N}$. Totes les paraules segueixen una llei de probabilitat de Poisson.

LLei de Poisson *P*(*m*)

Aquesta llei, que devem al matemàtic Denis Poisson (1781-1840), es defineix de la manera següent:

$$p_r = \frac{m^r}{r!} e^{-m}$$

en què *r* = 0, 1, 2, 3... i *r*! = 1 · 2 · 3 · 4... *r*

e designa la funció exponencial.

Tenim efectivament una distribució de probabilitat perquè:

$$\sum_{r=0}^{\infty} p_r = \sum_{r=0}^{\infty} \frac{m^r}{r!} e^{-m} = e^{-m} \sum_{r=0}^{\infty} \frac{m^r}{r!} = e^{-m} \cdot e^{m} = 1$$

Aquesta llei es coneix normalment com a *llei de probabilitats de distribució dels esdeveniments rars* o *llei de Poisson de paràmetre m*.

Observació

Aquí ens hem interessat únicament per la presència o l'absència d'una paraula en el document. És possible generalitzar aquesta manera de procedir tenint en compte la freqüència de la paraula en el document.

Podeu consultar informació sobre la funció exponencial en el subapartat 3.1 del mòdul 3, "Matemàtica de la informació".

Relacions entre les diferents lleis de probabilitat

Convergència de la llei binomial vers la llei de Poisson

Quan $n \to \infty$ i $u \to 0$ de tal manera que la seva mitjana m té un límit finit $m = nu$, llavors la llei binomial de paràmetres n i u convergeix vers una llei de Poisson de mitjana m. Aquesta llei es coneix sovint com la *llei dels esdeveniments rars*.

Convergència de la llei binomial vers la llei normal

Quan $n \to \infty$, i si u no és ni gaire petit ni gaire gran, podem assimilar una llei binomial a una llei normal. La regla empírica que s'adopta sovint és la d'assimilar la llei binomial a la llei normal quan els productes nu i $n(1 - u)$ són superiors a un valor entre 15 i 20.

En l'annex 2 d'aquest mòdul hi ha la taula de la llei normal.
Podeu consultar informació sobre la llei normal en el subapartat 2.3, "Taules estadístiques".

5.3. Com més llegeixes, més llegiràs (cienciometria): llei dels avantatges acumulats (llei de Price)

Tal com acabem de veure, el model de l'urna s'usa amb freqüència per a simular els processos aleatoris; això és el que farem novament per a explicar la llei dels avantatges acumulats i la llei de Price:

1) Tenim una urna amb r boles vermelles (una bola vermella significa un èxit) i b boles blanques (una bola blanca significa un fracàs). En aquest cas, l'univers és el conjunt reduït als fets elementals: treure una bola blanca, treure una bola vermella. Si s'extreuen boles a l'atzar a intervals regulars i es tornen a introduir a l'urna, la mesura de la probabilitat d'extreure una bola vermella o blanca es manté constant, i tenim:

$$P(\text{"extreure una bola vermella"}) = \frac{r}{r+b};$$

$$P(\text{"extreure una bola blanca"}) = \frac{b}{r+b}$$

Anomenarem u la probabilitat d'un èxit.

Amb n extraccions aleatòries, la probabilitat de tenir i èxits, amb una i que varia entre 0 i n, és la llei binomial de paràmetres n, u, que s'escriu $B(n,u)$.

2) Imaginem que després de cada extracció s'introdueixen boles noves a l'urna. D'aquesta manera es modifiquen constantment les probabilitats d'èxit i de fracàs. Si després de cada extracció introduïm c boles del mateix color que l'extreta, es pot veure que en certes condicions s'obté una llei binomial negativa.

5.3.1. Llei dels avantatges acumulats

El 1976, per a explicar les produccions desiguals d'articles científics publicats pels investigadors, Price va anunciar la regla següent: com més articles es publiquen, més se'n publicaran; o, per dir-ho d'una altra manera, l'èxit genera èxit. Un lloc web o un museu que reben visites amb freqüència seran visitats més regularment que els que reben menys visites; un llibre que s'ha llegit moltes vegades serà llegit amb més freqüència que un llibre amb pocs lectors; i així successivament.

La regla enunciada consisteix a incrementar la probabilitat que l'èxit engendri nous èxits sense tenir en compte la influència dels fracassos. Els avantatges s'acumulen.

Bibliografia

D. S. Price (1976). "A general theory of bibliometric and other cumulative advantage process". *Journal of the American Society for Information Science* (vol. 27, núm. 5, pàg. 292-306).

La llei dels avantatges acumulats es formula matemàticament de la manera següent:

$$P_{u_i} = (k+1) \cdot Beta(u_i, k+2)$$

en què $u_i = 1,2\ldots$, k = constant, u_i = nombre d'èxits

La funció $Beta(a,b)$ és una llei de probabilitat contínua amb dos arguments a i b.

$$Beta(a,b) = \int_0^1 x^{a-1} \cdot (1-x)^{b-1}\, dx$$

En els llibres d'estadística es pot trobar una taula d'aquesta llei.

Aquesta formulació no és universal i moltes altres investigacions han conduït a proposar altres formulacions.

5.3.2. Llei de Price

Podeu consultar informació sobre la llei de Lotka en el subapartat 3.4 d'aquest mòdul.

A partir de la llei de Lotka, Price va formular empíricament la llei següent, coneguda amb el nom de *llei de Price*: en el camp de la investigació científica, el nombre d'autors més productius estarà determinat per l'arrel quadrada del nombre total d'autors.

Si N és el nombre total d'autors que publiquen en un camp científic concret, $\sqrt{N}$ dóna el nombre d'autors més productius, que són els responsables d'aproximadament la meitat de la producció d'articles científics. Això posa de manifest la influència que podrien tenir els col·legues invisibles. Aquesta llei és

imprecisa, ja que depèn de la grandària de la mostra. Una de les seves formulacions probabilistes simples és la següent:

$$p_i = \frac{1}{i(i+1)} \text{ si } i = 1,2\ldots$$

Encara que Price va demostrar que el seu model ofereix un marc d'interpretació probabilista per a les diferents lleis de la informació, el vincle entre els comportaments socials i la descripció estadística no sempre és clar. Per exemple, el fet de no publicar un article en un moment concret dins d'un camp no es pot considerar com un fracàs (vegeu el model de l'urna usat per Price), sinó millor com un no-fet. No obstant això, és necessari relativitzar aquesta crítica, ja que els models predictius rarament expliquen les causes. Pel que fa a això, vegeu l'analogia que hem fet en el primer mòdul entre la llei de la gravetat en física i la llei de Lotka (subapartat 4 del mòdul 1, "La mesura de la informació").

Llei de Gauss i llei de la informació

"Les informacions no són partícules."

Recordem una explicació de la llei normal (o llei de Gauss) amb què es va trobar el matemàtic H. Poincaré (1864-1912) al principi del segle XX. Imaginem que un fenomen macroscòpic és la suma d'un gran nombre de fets elementals microscòpics que, tal com el seu nom indica, són petits amb relació a la variable macroscòpica. Si les probabilitats dels fets microscòpics no són totes iguals, ni massa properes a 0 ni a 1, llavors aquest fenomen seguirà una llei normal (distribució gaussiana). Això significa que la llei normal només reflecteix la distribució a l'atzar d'uns efectes additius molt petits.

La llei anterior, la llei dels avantatges acumulats, resulta d'un procés en el qual un fet elemental serà molt més important que els altres en la construcció del fenomen estudiat:

"Citem un article perquè ja s'ha citat anteriorment."

"Publiquem un article perquè ja hem publicat diversos articles."

...

S'han proposat diversos arguments per a explicar aquest fenomen. Nombrosos treballs teòrics (vegeu Egghe) han demostrat la similitud, en certes condicions, dels diferents fenòmens observats ja descrits en aquesta assignatura: llei de Lotka (subapartat 3.4), llei de Bradford (subapartat 2.1 del mòdul 3, "Matemàtica de la informació"), llei de Zipf (subapartat 3.3 del mòdul 3, "Matemàtica de la informació"). La nostra disciplina no és l'única que presenta regularitats d'aquest tipus. Trobem aquestes similituds en biologia (i en particular en l'epidemiologia), en economia (llei de Pareto) i en geografia.

Observació

En la pràctica, moltes vegades es recorre a ajustaments de tipus hiperbòlic quan s'estudien fenòmens informacionals.

Podeu consultar informació sobre la freqüència de les paraules en un text en el subapartat 3.3 del mòdul 3, "Matemàtica de la informació".

Les distribucions hiperbòliques no tenen moments. De manera més concreta, es demostra que la distribució de probabilitat definida per la funció de densitat hiperbòlica: $f(x) = \frac{K}{x^{1+\alpha}}$ (K i α són constants positives que es volen calcular) en què x pertany a l'interval $[0, \infty]$ i té només moments d'ordre n en què n és estrictament inferior a α.

Podeu consultar informació sobre distribució estadística en l'annex 3, "Càlcul de moments".

5.4. Circulació d'obres (bibliometria): processos temporals (model de Morse)

Les distribucions estudiades no depenen directament del temps i es coneixen com a *estacionàries*. Les distribucions en les quals el temps és un paràmetre explícit es basen en la teoria dels processos estocàstics. Els processos markovians en són una il·lustració.

Un procés aleatori temporal és una sèrie de variables aleatòries. Aquest procés és de Markov si:

- *no té memòria*, és a dir, que per a determinar l'estat següent només té importància el coneixement de l'últim estat (axioma de Markov), o
- *és homogeni*, és a dir, que la probabilitat de transició d'un estat a un altre només depèn del temps transcorregut entre aquests dos estats i no de l'instant de la transició (axioma d'homogeneïtat).

Exemple

El model de Morse

Amb l'anàlisi dels processos markovians, Ph. Morse va construir un model que permet als bibliotecaris implantar un sistema d'anàlisi de les previsions de la circulació de les obres. Aquest model (conegut com a *model de Morse*) es basa en tres hipòtesis fonamentals:

Bibliografia

Ph. Morse (1968). *Library effectiveness*. Cambridge: The MIT Press.

- La circulació de les obres és un procés aleatori. És a dir, el préstec d'un llibre en una biblioteca no es pot predir amb certesa. No obstant això, sí és possible predir el comportament mitjà d'una classe de llibres.
- La circulació mitjana de les obres disminueix exponencialment amb el temps.
- Hi ha una correlació entre l'ús d'un llibre durant un període donat i el seu ús durant el període següent.

Morse va decidir analitzar el procés al llarg d'un any. Estipula que el coneixement de la circulació de l'any anterior és suficient per a conèixer la de l'any actual. El seu model permet tenir en compte els dos paràmetres clàssics d'una anàlisi biblioteconòmica de la demanda, que són:

- L'*obsolescència*, que reflecteix la disminució exponencial de la demanda d'una obra a mesura que transcorre el temps.
- La *resurgència*, que té en compte el cas de les obres que tornen a entrar en circulació després d'un període de no-ús.

Podeu consultar informació sobre la funció exponencial en el subapartat 3.1 del mòdul 3, "Matemàtica de la informació".

Morse observa que hi ha una relació lineal simple que permet conèixer la circulació mitjana M d'una classe d'obres durant un any, sabent que han circulat m vegades durant el període previ.

$$M(m) = a + bm \qquad (\alpha)$$

Els paràmetres a i b, calculats amb tècniques de regressió, tenen el significat següent:

Podeu consultar informació sobre tècniques de regressió en el subapartat 2.2 del mòdul 3, "Matemàtica de la informació".

- a representa la circulació mitjana de les obres més antigues de la classe en qüestió i
- b reflecteix la popularitat de les obres d'aquesta classe en comparació amb la que tenien abans.

Morse, a més, formula la hipòtesi que la probabilitat condicional $P(n/m)$, que és la probabilitat que una obra es deixi en préstec n vegades en un any, sabent que s'han prestat

Podeu consultar informació sobre probabilitat condicional en el subapartat 5.2 del mòdul 3, "Matemàtica de la informació".

m vegades l'any anterior, és del tipus de Poisson. Així, amb *m* fixat, la probabilitat *P*(*n*/*m*) és una distribució de Poisson de mitjana *a* + *bm:*

$$P(n/m) = \frac{1}{n!}(a+bm)^n e^{-(a+bm)}$$

En què *n* és un nombre enter que varia de 0 (obra que no s'ha deixat mai) a ∞ (en la pràctica, agafem el nombre màxim de vegades que s'ha deixat una obra de la classe estudiada).

Les previsions de circulació

Si la circulació mitjana d'una classe d'obres verifica la relació (α), llavors la circulació mitjana de la col·lecció al llarg de l'any *t* + 1 està vinculada a la circulació mitjana de l'any anterior per mitjà de la fórmula:

$$M(t+1) = a + b \cdot M(t) \qquad (\beta)$$

Se suposa que els paràmetres *a* i *b* són constants al llarg dels anys. Si fixem l'origen del temps en *t* = 1, llavors, segons la relació (β), la circulació mitjana de les obres de la col·lecció en el temps *t* = 2 serà:

$$M(2) = a + bM(1)$$

Mitjançant iteracions successives, obtenim, per a l'any $(t + 1)^{me}$, la circulació mitjana:

$$M(t+1) = a(1 + b + b^2 + \ldots + b^{t-1}) + b^t \cdot M(1).$$

Aquí reconeixem una sèrie geomètrica; així, doncs, podem escriure:

Podeu consultar informació sobre la sèrie geomètrica en el subapartat 2.1 del mòdul 3, "Matemàtica de la informació".

$$M(t+1) = a \cdot \frac{1-b^t}{1-b} + b^t \cdot M(1) \qquad (\gamma)$$

Aquesta relació permet conèixer la circulació de tot el conjunt d'obres en el moment *t* en funció de la circulació mitjana en l'instant inicial. En la pràctica, *b* és inferior a 1. Si suposem que *a* i *b* són constants, quan *t* creix, la circulació mitjana tendeix cap al valor límit $\frac{a}{1-b}$.

La relació (γ) també es pot escriure:

$$M(t+1) = \frac{a}{1-b} + \left(M(1) - \frac{a}{1-b}\right) b^t$$

Si $0 < b < 1$ i si $M(1) > \frac{a}{1-b}$, llavors la funció $t \to M(t+1)$ decreix amb el temps, que és la segona hipòtesi del model.

5.5. En resum

En el camp de la informació, a més del gran nombre de variables deterministes que tenien en compte els precedents estadístics, hi ha tot un món de variables

aleatòries, és a dir, que depenen de l'atzar, de la sort. És aquí on entra en joc l'estadística probabilista.

Algunes d'aquestes variables segueixen les grans lleis clàssiques de la probabilitat, que són la llei geomètrica, la llei binomial negativa i la llei de Poisson.

Però n'hi ha d'altres que no les segueixen, la qual cosa caracteritza l'especificitat dels processos informacionals i va induir els investigadors en ciències de la informació a formular un marc probabilista per a les noves lleis probabilistes, conegudes com a *lleis de la informació*, com la llei dels avantatges acumulats i la llei de Price.

Totes aquestes lleis, tant si són probabilistes tradicionals com probabilistes informacionals, només permeten usar variables estacionàries i estudiar processos estacionaris, és a dir, variables i processos que no depenen explícitament del temps. Però aquí també trobem variables i processos coneguts com a *estocàstics*, que depenen del temps, com per exemple els processos markovians. Els usarem per a elaborar models d'anàlisi preventiva de les activitats informacionals.

Activitats

Exercicis d'estadística unidimensional

1. Proporcions, percentatges, taxes (bibliometria)

Dels 150 títols que s'han deixat en la biblioteca durant aquesta setmana, 100 són llibres per a adults i 50 llibres infantils. Quines són les proporcions d'aquests dos tipus de literatura? Quins en són els percentatges? Calculeu la taxa de llibres per a adults / llibres infantils.

2. Representacions gràfiques (bibliometria)

Construïu el polígon de freqüències, el polígon de freqüències acumulades i l'histograma dels préstecs dels 46 llibres de la taula següent. La lectura de la taula és la següent: el primer llibre ha estat en préstec durant 18 dies, el segon durant 26 dies, etc.

Durada dels préstecs de 46 llibres (*en dies*):

18	4	13	12	12	15
26	7	16	11	16	10
15	9	9	6	6	6
13	14	12	9	31	22
24	8	35	28	26	27
21	10	13	16	31	9
5	20	29	38	31	
17	13	12	19	19	

3. Mesures de tendència central (mediametria)

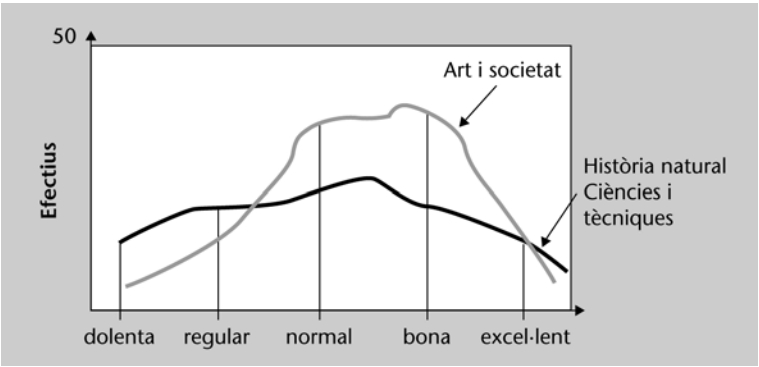
Calculeu la mitjana, la mediana i la moda de la distribució del nombre d'hores per setmana durant les quals els teleespectadors han estat veient la televisió. Els valors numèrics següents signifiquen que el primer teleespectador ha vist la televisió durant 15 hores al llarg de la setmana, el segon l'ha vista durant 27 hores, i així successivament.

Temps dedicat a veure la televisió (*en hores*):

15; 27; 17; 28; 19; 29; 20; 31; 20; 32; 20; 36; 23; 40; 23; 26.

4. Mesures de dispersió (mediametria)

S'ha preguntat a dues poblacions de conservadors de museus (una de museus d'art i de societat, l'altra de museus d'història natural i de museus de ciències i de tècniques) sobre la qualitat de les accions de formació que els proposen. Les distribucions de les respostes són les que apareixen representades a continuació:



Quin és el grup més dispers?, quin és el grup més homogeni?

5. El personal d'R+D (cienciometria)

Una enquesta feta pel Ministeri d'Investigació sobre la distribució del personal segons el tipus de laboratori ha donat els resultats següents:

Tipus de laboratori	Personal total (en milers)	Personal investigador (en milers)
Laboratoris acadèmics	93,3	56,8

Tipus de laboratori	Personal total (en milers)	Personal investigador (en milers)
Laboratoris públics d'investigació amb finalitats pròpies	50,4	18,7
Laboratoris militars	19,5	3,0
Laboratoris d'empreses	162,0	66,7
Laboratoris públics internacionals	2,2	2,2

Font: Informe de l'OST, 1998

(Personal investigador: directius, enginyers, investigadors, etc.)
(Personal total: personal investigador + administratiu + documentació)

a) Completeu la taula calculant la distribució en percentatge del personal total i del personal investigador (els percentatges s'han de donar amb una xifra significativa).
b) Resumiu la informació més important de la taula en poques línies.
c) Quin tipus d'indicador usaríeu per a calcular el personal de suport d'un investigador en cada tipus de laboratori. Classifiqueu els laboratoris a partir d'aquesta dada.

6. Qualitat dels noticiaris televisius (mediametria)

S'ha demanat a 36 teleespectadors que posin una nota entre 0 i 20 als noticiaris televisius de les 20.00 h de TF1 i de France 2. Els resultats obtinguts són els que apareixen en la taula següent.

1) Per a les dues cadenes:

a) Representeu el polígon de freqüències i el polígon de freqüències acumulades de les notes obtingudes.
b) Calculeu la mitjana, la mediana, la moda, la desviació estàndard i el coeficient de variació d'aquestes dues distribucions de notes.

A quines conclusions arribeu després de llegir aquestes diferents mesures de tendència central?

2) Torneu al punt (a) després d'haver creat prèviament classes d'amplituds variables.

Comenteu, en poques línies, els resultats més significatius d'aquest estudi.

Nre. de teleespectadors	TF1	France 2
1	6	8,5
2	7	12
3	8	10,5
4	9	20
5	9,5	17
6	6,5	12
7	12	15
8	9	15,5
9	8	16
10	8,5	13
11	11	12,5
12	11,5	12,5
13	12	2
14	13	6
15	9,5	18
16	8	6

Nre. de teleespectadors	TF1	France 2
17	7	5,5
18	12	6,5
19	19	7
20	9	7
21	10	0
22	12	1
23	13	16
24	14	15
25	9	3
26	8	3,5
27	6	5
28	2	5,5
29	9	9
30	7	5
31	12	6
32	15	14,5
33	6	7
34	7	7,5
35	11	11
36	5	6

Exercicis d'estadística bidimensional

7. Càlcul i interpretació d'un coeficient de correlació (infometria)

Un proveïdor de documents vol saber si hi ha alguna relació entre les revistes sol·licitades pels investigadors i el nombre total de vegades que se citen aquestes mateixes revistes en un any. L'ISI (Institute for Scientific Information) estableix cada any aquestes citacions i les presenta en el JCR (*Journal Citation Report*). Per a fer-ho, el proveïdor anota el nombre de sol·licituds per a les quinze revistes més usades el mes de gener de 1998 i consulta el nombre de vegades que s'han citat en l'any 1997.

Títols de les revistes	Nombre de sol·licituds	Total de citacions el 1997
1. *Journal of Biological Chemistry*	92	293.024
2. *Lancet*	85	100.526
3. *Proceedings of the National Academy of Science - USA*	79	276.659
4. *New England Journal of Medicine*	57	118.106
5. *Nature*	53	270.777
6. *Biochimica et Biophysica Acta*	52	69.557
7. *Chest*	52	19.456
8. *Biochemical and Biophysical Research Communications*	51	56.993
9. *Journal of Chromatography*	50	30.985
10. *Journal of the American Chemical Society*	49	174.540

Títols de les revistes	Nombre de sol·licituds	Total de citacions el 1997
11. *International Journal of Pharmaceutics*	48	4.373
12. *Circulation*	47	64.644
13. *Journal of Clinical Investigation*	46	72.249
14. *Febs Letters*	45	45.436
15. *Cancer Research*	45	82.343

Quines conclusions se'n poden treure?

8. Dibuixeu una recta de regressió lineal (cienciometria)

Quina relació hi ha entre l'import dels crèdits concedits a un laboratori d'investigació pel Ministeri d'Educació Superior i Investigació i el nombre d'articles que publica? Per a deu laboratoris, s'han obtingut els resultats següents:

Laboratori	Crèdits/laboratori (en milers d'euros)	Articles/laboratori
a	46	45
b	41	46
c	67	82
d	39	25
e	30	21
f	51	49
g	40	39
h	72	89
i	62	77
j	55	54

a) Representeu gràficament la relació entre els crèdits concedits i els articles publicats.
b) Dibuixeu la recta de regressió sabent que el pendent és d'1,66 i que talla l'eix *Y* en l'ordenada –30,8.
c) Compteu, gràficament i calculant-lo, el nombre d'articles per laboratori quan els crèdits concedits són de 80.000 euros.
d) Calculeu el coeficient de correlació que relaciona els crèdits i els articles. Traieu algunes condicions.

9. Distribució creuada (bibliometria)

Un bibliotecari vol identificar els gustos dels seus lectors, els quals ha classificat en tres categories: jove, adult i estudiant. El que li interessa són els temes de les obres que agafen en préstec. Els resultats que ha obtingut són els següents:

Tema de l'obra / tipus de lector	Jove	Adult	Estudiant
Ciències de la informació	5	45	23
Ciències humanes i socials	10	8	48
Ciències exactes i aplicades	25	9	73
Art, oci i esports	7	99	21
Literatura (excepte novel·les)	2	45	63
Història i geografia	34	38	93
Novel·les	65	576	49
Còmics	124	24	75

Quines conclusions es poden treure?

10. Taula de contingència: eficàcia d'un programa de documentació (infometria)

Per a posar a prova l'eficàcia d'un programa de documentació, es fan diferents consultes en un conjunt de mil documents. El resultat d'aquestes consultes porta a dividir els documents en quatre conjunts disjunts. El primer d'aquests conjunts està format pels documents extrets pertinents, el segon pels documents extrets no pertinents, el tercer pels documents no extrets pertinents, i el quart pels documents no extrets no pertinents.

Una de les preguntes proposades ha donat els resultats següents:

Documents	**Extrets**	**No extrets**
Pertinents	110	440
No pertinents	65	385

Què penseu d'aquests resultats?, el fet de ser pertinent i el fet de ser extret són independents per a un document? Justifiqueu la resposta.

11. Compra i préstec de llibres (bibliometria)

Normalment s'admet que les biblioteques tenen una funció pedagògica en la iniciació a la lectura. És solament després d'haver adquirit el gust per la lectura en aquestes instal·lacions quan moltes persones adquireixen el costum de freqüentar les llibreries. Per a verificar aquesta hipòtesi, s'ha fet una enquesta entre una mostra de 1.635 lectors. Cadascun ha respost les dues preguntes següents:

Pregunta 1 - Des de fa dos o tres anys, diríeu que...
compreu més llibres? C + +
compreu tants llibres com abans? C = =
compreu menys llibres? C – –

Pregunta 2 - Des de fa dos o tres anys, diríeu que...
agafeu en préstec més llibres? P + +
agafeu en préstec tants llibres com abans? P = =
agafeu en préstec menys llibres? P – –

Una distribució creuada ha donat els resultats següents:

	C + +	**C =**	**C – –**
P + +	86	183	346
P =	54	401	311
P – –	61	78	112

Quines conclusions se'n podrien treure?

12. Públic dels noticiaris televisius (mediametria)

Vegeu l'exercici 6.

S'ha demanat a trenta-sis teleespectadors que donin una nota entre 0 i 20 als noticiaris televisius de les 20.00 h de les cadenes TF1 i France 2. Aquests valors numèrics es completen caracteritzant els teleespectadors segons el seu nivell d'estudis; els teleespectadors amb estudis superiors s'identifiquen amb SUP, i els que no els tenen, amb NSUP.

a) Torneu als diferents punts de l'exercici 6.
b) Representeu el núvol de punts de les dues sèries de notes. Calculeu el coeficient de correlació d'aquestes dues sèries de notes. Quines conclusions en podeu treure?
c) Expliqueu en poques línies els resultats més significatius d'aquest estudi.

Nre. de teleespectadors	**TF1**	**France 2**	**Nivell d'estudis**
1	6	8,5	SUP
2	7	12	SUP
3	8	10,5	SUP
4	9	20	SUP

Nre. de teleespectadors	TF1	France 2	Nivell d'estudis
5	9,5	17	SUP
6	6,5	12	SUP
7	12	15	SUP
8	9	15,5	SUP
9	8	16	SUP
10	8,5	13	SUP
11	11	12,5	SUP
12	11,5	12,5	SUP
13	12	2	SUP
14	13	6	SUP
15	9,5	18	SUP
16	8	6	NSUP
17	7	5,5	NSUP
18	12	6,5	NSUP
19	19	7	NSUP
20	9	7	NSUP
21	10	0	NSUP
22	12	1	NSUP
23	13	16	NSUP
24	14	15	NSUP
25	9	3	NSUP
26	8	3,5	NSUP
27	6	5	NSUP
28	2	5,5	NSUP
29	9	9	NSUP
30	7	5	NSUP
31	12	6	NSUP
32	15	14,5	NSUP
33	6	7	NSUP
34	7	7,5	NSUP
35	11	11	NSUP
36	5	6	NSUP

Exercicis d'estadística multidimensional

13. Classificació de textos (infometria)

Amb l'ajut d'un motor de cerca, volem agrupar textos segons dues temàtiques: una temàtica de vigilància informativa i una temàtica d'anàlisi relacional. En la taula següent es presenta el recompte obtingut per a cadascuna d'aquestes temàtiques en els textos analitzats:

	Anàlisi	Relacional	Vigilància	Informativa
Text A	4	4	2	10
Text B	4	4	2	9
Text C	5	1	4	5
Text D	1	6	1	1
Text E	4	1	5	5

a) Amb la distància del cosinus, calculeu la matriu de semblances dels textos analitzats. La distància del cosinus per a dos textos A i B es calcula tal com s'indica a continuació:

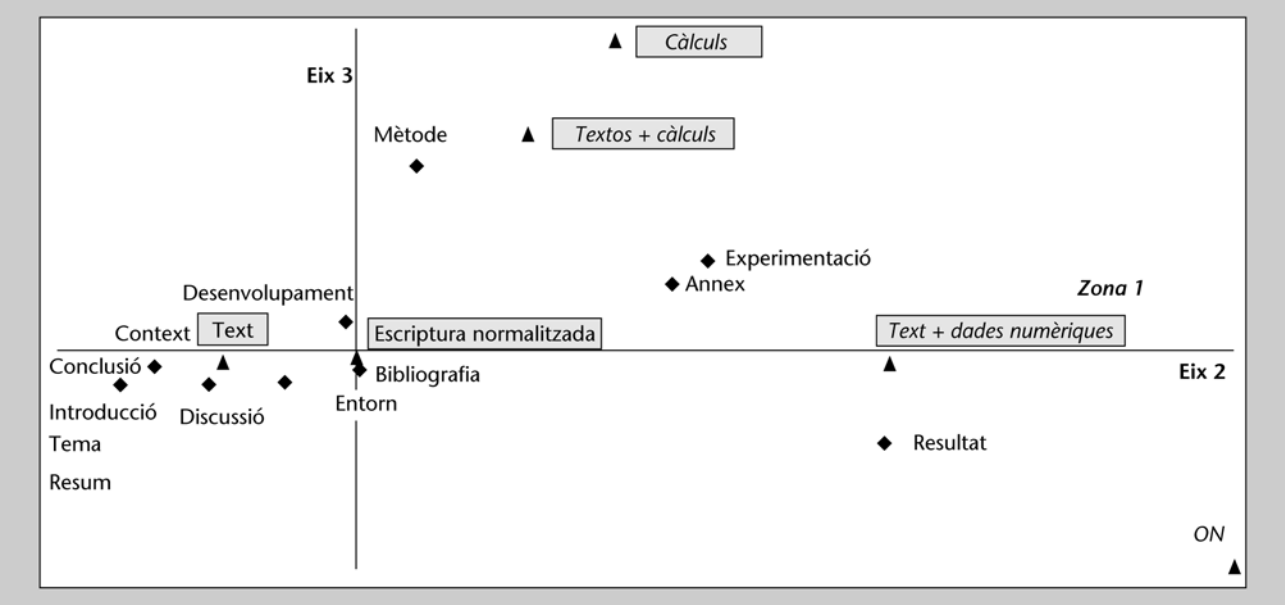
$$\cos(A,B) = 1 - \frac{\sum_{k=1}^{t} a_k b_k}{\sqrt{\sum_{k=1}^{t} (a_k)^2} \sqrt{\sum_{k=1}^{t} (b_k)^2}}$$

en què A i B són vectors de dimensió t.

b) A partir d'aquesta matriu, apliqueu un mètode de classificació que permeti agrupar els textos presentats.

14. Anàlisi factorial de les correspondències (infometria)

En analitzar les diferents produccions textuals (anàlisis del discurs), volem conèixer la relació entre les diferents formes textuals i paratextuals, que són el resum, la introducció, la descripció del context i del tema general de la problemàtica, la descripció de l'entorn de l'estudi, l'explicació del mètode, la descripció de l'experimentació, els resultats, la discussió, els desenvolupaments, la conclusió, la bibliografia, els annexos, etc. Les altres formes textuals amb les quals ens trobem fan referència als tipus de presentació dels valors i els resultats, és a dir, càlculs, dades numèriques, textos, dades numèriques + textos, textos + càlculs, escriptures normalitzades. D'un total de seixanta articles, extraiem aquestes diferències característiques i construïm (per encreuament) la taula de contingència Forma textual discursiva / Tipus de presentació. Després fem una anàlisi factorial de les correspondències (AFC), una part dels resultats de la qual, segons els eixos 2 i 3, es presenta en la gràfica següent.



a) Es poden veure dependències entre les formes textuals i els tipus de presentació a partir d'aquesta gràfica?
b) Ja que la representació planar de resultats multidimensionals pot comportar errors, es poden confirmar o invalidar aquests resultats a partir de la taula que presenta els χ^2 per casella de cada parell?

Taula 1. Encreuament de les formes textuals discursives i dels tipus de presentació

χ^2	Càlculs	Dades numèriques	Textos + càlculs	Textos + dades numèriques	Textos	Escriptures normalitzades	Total
Annex	3,584	0,285	0,392	10,067	1,311	1,355	16,994
Bibliografia	1,046	0,523	0,719	4,117	24,113	375,04	405,558
Conclusió	1,267	0,34	0,871	4,99	3,97	3,01	14,742
Context	2,218	1,109	0,148	6,847	4,913	5,267	20,502
Desenvolupament	0,505	0,998	0,287	0,166	0,191	4,741	6,888
Discussió	1,838	0,919	1,263	1,447	2,661	2,594	10,722

χ^2	Càlculs	Dades numèriques	Textos + càlculs	Textos + dades numèriques	Textos	Escriptures normalitzades	Total
Entorn	0,285	0,143	0,196	0,013	0,027	0,154	0,818
Experimentació	1,842	0,396	3,89	15,182	2,891	1,881	26,082
Introducció	1,394	0,697	0,958	5,489	4,367	3,311	16,216
Mètode	30,315	0,713	9,303	0,464	1,44	0,044	42,279
Resultat	0,497	49,13	0,101	46,615	8,719	4,741	109,803
Resum	0,919	0,459	0,632	3,618	2,878	2,182	10,688
Tema	0,254	0,127	0,174	0,998	0,794	0,602	2,949
Total	45,96	56,13	18,94	100,01	58,28	404,92	**684,24**

15. Anàlisi factorial de correspondències (bibliometria)

Després de fer una enquesta, una biblioteca ha enumerat els hàbits de lectura del seu públic.

Vegeu l'exercici 9.

Tema de l'obra / tipus de lector		Jove	Estudiant	Adult
	Codi	1	2	3
Ciències de la informació	10	5	23	45
Ciències humanes i socials	11	10	48	8
Ciències exactes i aplicades	12	25	73	9
Art, oci i esports	13	7	21	99
Literatura (excepte novel·les)	14	2	63	45
Història i geografia	15	34	93	38
Novel·les	16	65	49	576
Còmics	17	124	75	24

a) A partir d'aquests valors, feu una anàlisi factorial de les correspondències i representeu els dos primers eixos.
b) Una representació planar no és una font d'errors en aquest cas?
c) En observar la gràfica que s'obté, quines conclusions es poden treure?

Exercicis d'estadística probabilista

16. Ajust d'una distribució mitjançant una llei probabilista simple (bibliometria)

Una biblioteca registra el nombre de préstecs d'una col·lecció de quatre-centes sis obres en un any. Els resultats obtinguts són els següents:

Nombre de préstecs	Nombre d'obres
0	200
1	92
2	49
3	24
4	15
5	10
6	4
7	2
8	2
Igual o superior a 9	8

a) Calculeu la taxa mitjana de rotació d'una obra i la desviació estàndard associada. Farem el càlcul tenint en compte el fons mort.
b) Comproveu l'ajust d'aquesta distribució amb una llei geomètrica i amb una llei binomial negativa usant el mètode dels moments i tenint en compte el fons mort.
c) Dibuixeu les tres corbes en una gràfica.
d) Amb l'ajuda de la prova de χ^2, comproveu si l'ajust geomètric i l'ajust binomial negatiu són correctes.

17. Circulació de documents (bibliometria)

Un centre de documentació ha anotat el nombre de préstecs de dos-cents noranta documents el 1999 i el 2000.

1999/2000	0	1	2	3	4	5
0	195	17	9	3	2	1
1	24	8	3	1	0	0
2	2	1	3	1	1	1
3	5	3	0	0	0	0
4	1	0	1	0	1	1
5	1	2	0	2	0	1

La taula es llegeix de la manera següent:

- Fila 1, columna 2: dels 227 documents que no s'han deixat el 1999, 17 s'han deixat una vegada el 2000.
- Fila 2, columna 1: dels 228 documents que no s'han deixat el 2000, 24 s'han deixat una vegada el 1999.

a) Representeu les distribucions dels préstecs per als anys 1999 i 2000.
b) Calculeu la rotació mitjana dels documents en l'any 2000, sabent que han circulat j vegades el 1999. Designarem aquesta rotació mitjana per $M(j)$.
Representeu $M(j)$ i ajusteu la corba obtinguda mitjançant una recta de l'equació $M(j) = a \cdot j + b$. Interpreteu els coeficients a i b.
c) Formulem la hipòtesi que la distribució segueix una llei de Poisson de mitjana $a + b \cdot j$, és a dir:

$$P(i/j) = \frac{(a+bj)^i e^{-(a+bj)}}{i!}$$

(Probabilitat que un document circuli i vegades el 2000 sabent que ha circulat j vegades el 1999).

Calculeu la distribució teòrica dels préstecs del 2000. Què us sembla el resultat?

d) Volem preveure la circulació per al 2001. Per a això, suposem que la probabilitat que un document circuli i vegades durant el període 2001, sabent que ha circulat j vegades el 2000, està determinada per la matriu de transició prèvia (de tipus de Poisson i de mitjana $a + b \cdot j$).

18. Anàlisi del discurs (mediametria)

Una revista ha decidit fer un estudi retrospectiu sobre els articles que ha publicat aquests dos últims anys. Cada període correspon a polítiques editorials diferents executades per redactors gerents diferents. Els recomptes lexicomètrics han donat els resultats següents.

Període 1: 44.086 paraules, 7.599 paraules diferents
Període 2: 42.194 paraules, 6.483 paraules diferents

Ens interessem per tres adverbis que expressen judicis de certesa: *probablement*, *certament* i *evidentment*. Els recomptes han donat els resultats següents:

	Probablement	Certament	Evidentment
Període 1	36	16	10
Període 2	12	44	14
Total	48	60	24

a) Calculeu les probabilitats teòriques d'aparició d'una paraula qualsevol dels articles per a cada període. Què observem? Podíem preveure el resultat?
b) Per a cada adverbi i per a cada període, calculeu la desviació entre els efectius observats i els efectius teòrics. Què observem? Podíem preveure el resultat? A quines conclusions podem arribar?
c) Per saber si aquestes desviacions són significatives, formulem la hipòtesi que cada adverbi es distribueix segons una llei binomial, és a dir, que la seva distribució és independent del període editorial.

- Expliciteu aquestes lleis.
- Calculeu la desviació reduïda i traieu conclusions.

19. Ocupació d'una línia de transmissió (webmetria)

Com a part de les proves fetes en llocs web, s'han mesurat les taxes d'ocupació de les línies de transmissió de les xarxes i els temps de resposta d'una pàgina completa. Suposem que un servidor rep, al llarg de vint-i-quatre hores, N connexions de dos minuts, que és el temps necessari per a la lectura d'una pàgina concreta. Quines són les probabilitats d'ús del servidor si N és igual a 70, a 410, a 1.400, a 2.100 i a 3.200?

Solucionari

Solucions als exercicis d'estadística unidimensional

1.

Proporció de llibres per a adults = $\frac{100}{150} = 0{,}66$

Proporció de llibres infantils = $\frac{50}{150} = 0{,}33$

Percentatge de llibres per a adults = proporció de llibres per a adults multiplicada per 100 = 66%

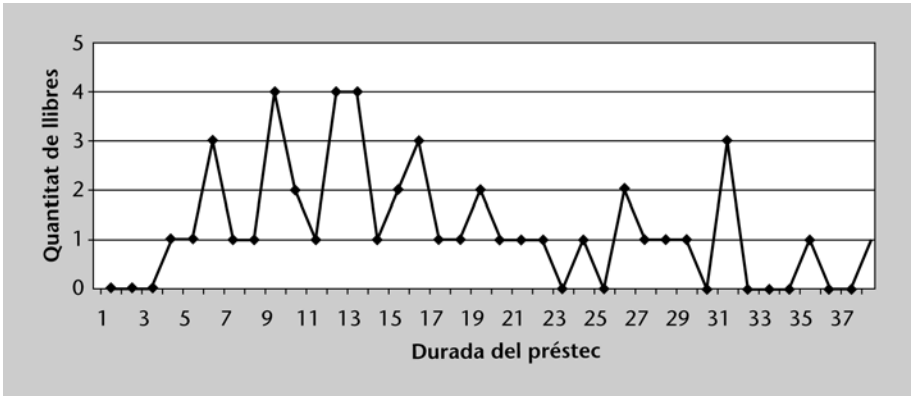
Percentatge de llibres infantils = proporció de llibres infantils multiplicada per 100 = 33%

Taxa de llibres per a adults / llibres infantils = $\frac{100}{50} = 2$

2. Per a dibuixar el polígon de freqüències és necessari comptar el nombre de vegades que un llibre s'ha deixat en préstec 1 dia, 2 dies, 3 dies, ... 38 dies (valor màxim), és a dir, cal comptar quantes vegades apareixen els valors 1, 2, 3, ..., 38. Aquests recomptes es poden fer a mà a partir de la taula o bé usant la funció de taules creuades que incorporen alguns fulls de càlcul.

D'aquesta manera s'obté la taula 1, en la qual es pot veure que hi ha 1 llibre que s'ha deixat 4 dies, 1 llibre que s'ha deixat 5 dies, 3 llibres que s'han deixat 6 dies... i 1 llibre que s'ha deixat 38 dies. Usant les funcions gràfiques del full de càlcul, dibuixem el polígon de freqüències (figura 1) i l'histograma de freqüències (diagrama de barres, figura 2). Hi ha altres tipus de representacions (sectors, àrees, etc.).

Figura 1. Polígon de les freqüències

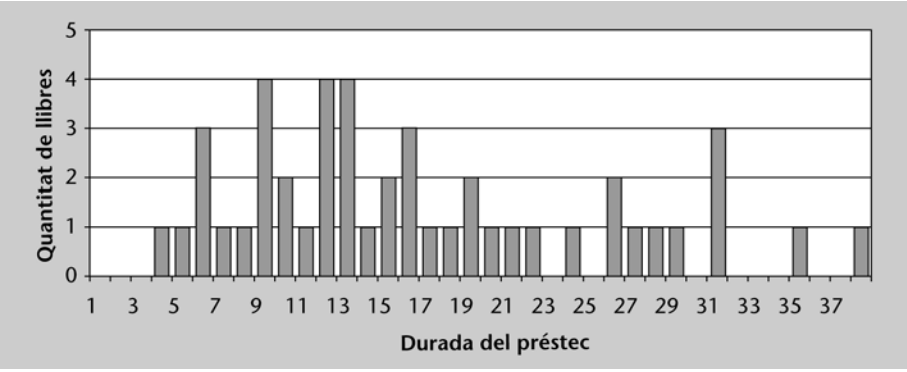


Taula 1. Distribució de les freqüències

Durada del préstec (en dies)	Nombre de llibres
1	0
2	0
3	0
4	1
5	1
6	3
7	1
8	1
9	4
10	2

Durada del préstec (en dies)	Nombre de llibres
11	1
12	4
13	4
14	1
15	2
16	3
17	1
18	1
19	2
20	1
21	1
22	1
23	0
24	1
25	0
26	2
27	1
28	1
29	1
30	0
31	3
32	0
33	0
34	0
35	1
36	0
37	0
38	1
Total	**46**

Figura 2. Histograma de les freqüències



Per a dibuixar el polígon de freqüències acumulades cal sumar, línia per línia, el nombre de llibres deixats. Els resultats són els que apareixen en la taula 2:

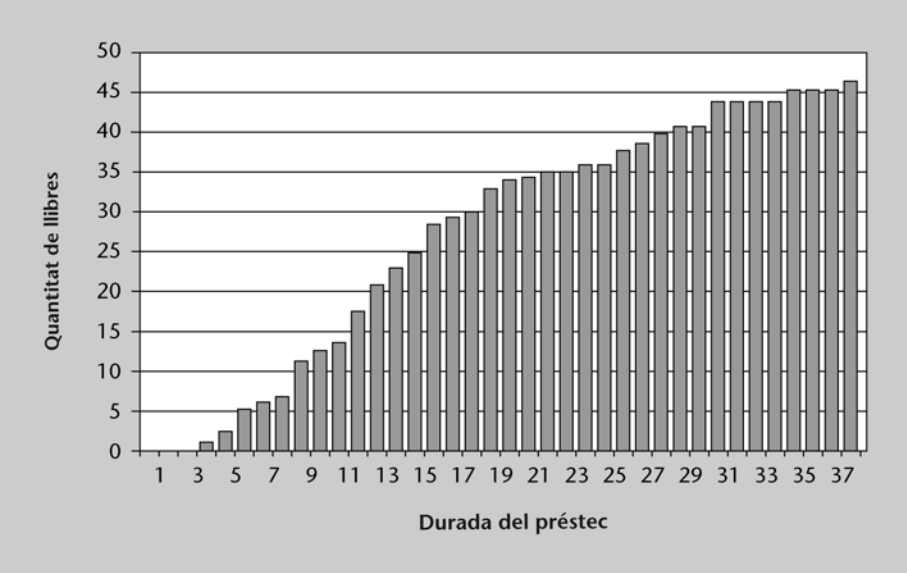
Taula 2. Distribució de les freqüències acumulades

Durada del préstec	Nombre de llibres	Nombre de llibres acumulats	Percentatges acumulats
4	1	1	2,2
5	1	2	4,3
6	3	5	10,9
7	1	6	13,0
8	1	7	15,2
9	4	11	23,9
10	2	13	28,3
11	1	14	30,4
12	4	18	39,1
13	4	22	47,8
14	**1**	**23**	**50,0**
15	2	25	54,3
16	3	28	60,9
17	1	29	63,0
18	1	30	65,2
19	2	32	69,5
20	1	33	71,7
21	1	34	73,9
22	1	35	76,1
24	1	36	78,3
26	2	38	82,6
27	1	39	84,8
28	1	40	87,0
29	1	41	89,1
31	3	44	95,7
35	1	45	97,8
38	1	46	100,0

Així, doncs, tenim 1 llibre que ha estat en préstec durant 4 dies, 2 llibres que han estat en préstec un màxim de 5 dies, 5 llibres que han estat en préstec un màxim de 6 dies..., 23 llibres (la meitat) un màxim de 14 dies... i 46 llibres que han estat en préstec un màxim de 38 dies.

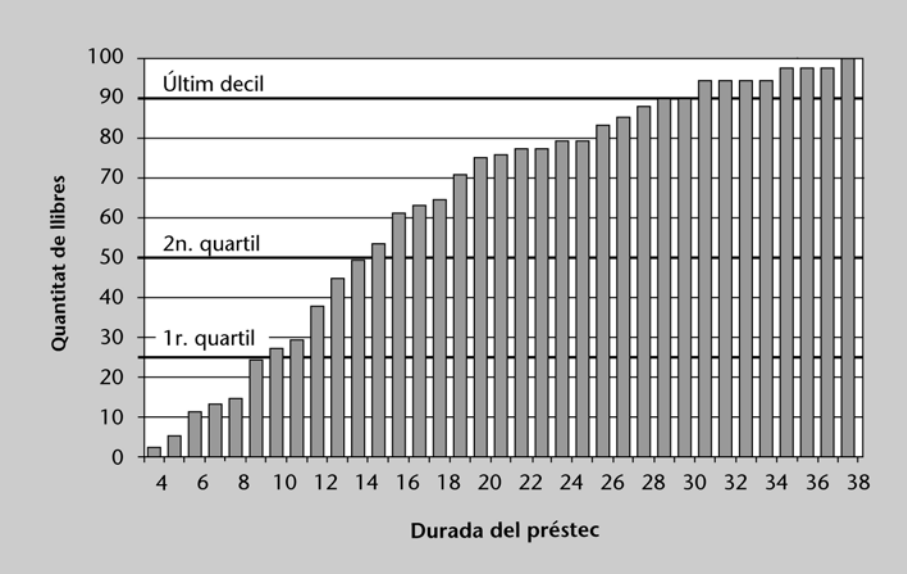
L'histograma de les freqüències acumulades apareix representat en la figura 3:

Figura 3. Histograma de les freqüències acumulades



En la figura 4 apareixen representats els percentatges acumulats de les freqüències dels préstecs:

Figura 4. Histograma dels percentatges acumulats de les freqüències



La lectura de la taula 2 i de l'histograma anterior ens permet arribar a les conclusions següents:

- El 25% dels llibres (primer quartil) ha estat en préstec menys de deu dies (fet que implica que les tres quartes parts han estat en préstec durant més de deu dies).
- El 50% dels llibres (segon quartil) ha estat en préstec un màxim de catorze dies.
- El 90% dels llibres ha estat en préstec menys de trenta-un dies, la qual cosa implica que només el 10% dels llibres (últim decil) ha estat en préstec més de trenta-un dies.

3.

a) Mitjana

$$\text{Mitjana} = \frac{15+27+17+28+19+29+20+31+20+32+20+36+23+40+23+26}{16} = 25{,}375$$

De mitjana, els teleespectadors han vist la televisió durant 25 hores i 22 minuts cada setmana.

b) Mediana

Per a determinar la mediana, és necessari classificar per endavant els elements en ordre creixent:

15 17 19 20 20 20 23 23 | 26 27 28 29 31 32 36 40
a | b

En aquest exemple, en què el nombre de valors és un nombre parell (16), la mediana es troba entre 23 i 26. En aquest cas, es calcula de la manera següent:

Mediana = $a + \frac{b-a}{2} = 24,5$, és a dir, 24 hores i 30 minuts.

c) Moda

Aquí només hi ha un valor modal: moda = 20 hores.

4. El grup més dispers és el grup dels conservadors de museus d'història natural i de museus de ciències i de tècniques. Les seves opinions sobre la qualitat de les accions de formació que els proposen estan molt més dividides. El grup més homogeni és el dels conservadors de museus d'art i de societat. La majoria consideren que les accions de formació tenen una qualitat mitjana o bona.

5.

a) Sumem cada columna, i això ens permet calcular l'efectiu i el percentatge de cada tipus de personal en els diferents laboratoris.

Tipus de laboratori	Personal total (en milers)	Personal total (en %)	Personal investigador (en milers)	Personal investigador (en %)
Laboratoris acadèmics	93,3	28,5	56,8	38,5
Laboratoris públics d'investigació amb finalitats pròpies	50,4	15,4	18,7	12,7
Laboratoris militars	19,5	6,0	3,0	2,0
Laboratoris d'empreses	162,0	49,5	66,7	45,3
Laboratoris públics internacionals	2,2	0,7	2,2	1,5
Total	327,4	100,0	147,4	100,0

b) 327.400 persones treballen en el camp de la investigació a França, de les quals 147.400 són investigadors. Gairebé la meitat d'aquestes 327.400 persones (162.000, és a dir, el 49,5%) treballen en l'àmbit de la investigació industrial. Una gran part dels investigadors (45,3%) també treballen en aquest sector.

c) La relació entre la quantitat de personal no investigador i la quantitat de personal investigador en cada laboratori és un bon indicador de la qualitat del lloc de treball dels investigadors.

Exemple: La taxa de personal de suport dels investigadors en els laboratoris públics és de $\frac{50,4-18,7}{18,7} = 1,70$. Això vol dir que, en aquest sector, cada investigador treballa amb 1,70 no-investigadors (personal administratiu, tècnic, operaris i personal de servei). La classificació dels laboratoris segons la seva taxa de personal de suport creixent és la següent:

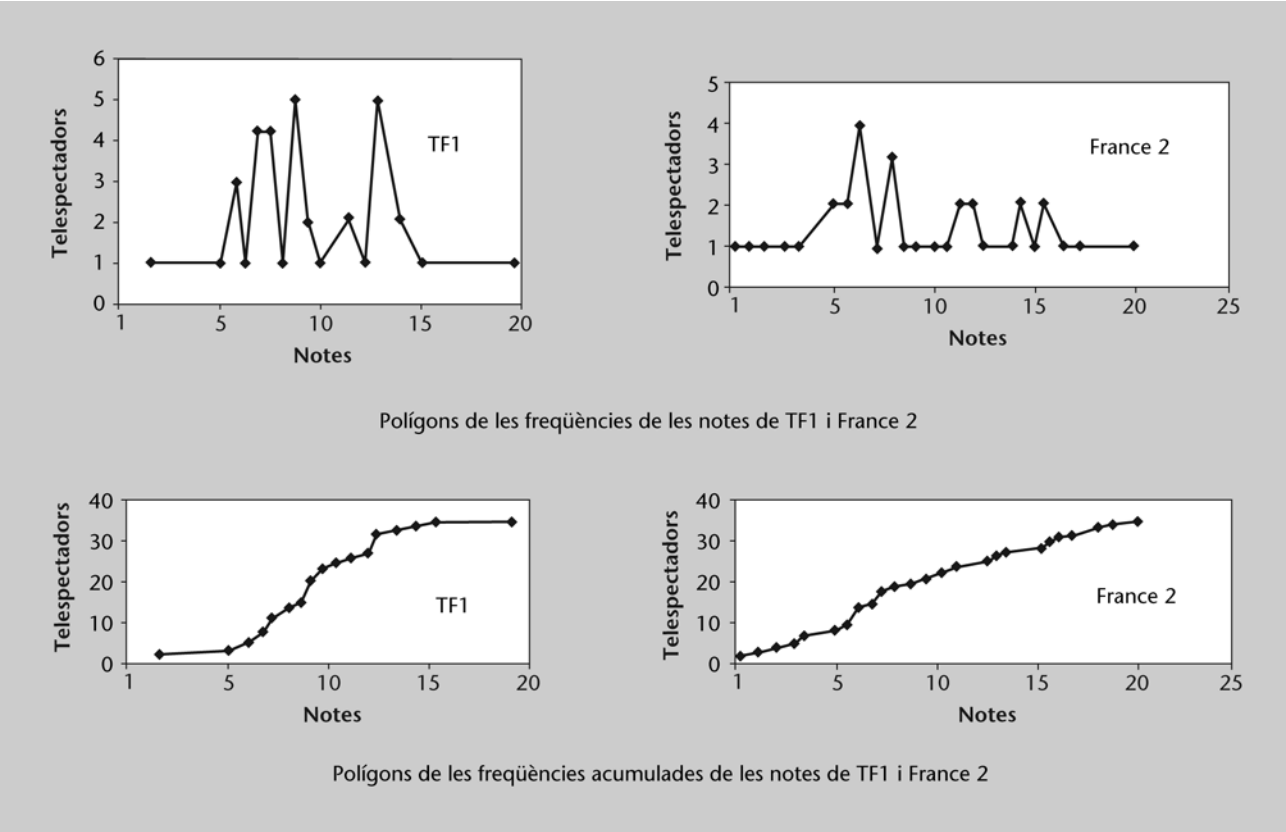
Laboratoris internacionals	0
Laboratoris acadèmics	0,64
Laboratoris d'empreses	1,43
Laboratoris públics	1,70
Laboratoris militars	5,55

6. Les distribucions de les freqüències de les notes obtingudes per TF1 i France 2 són les següents:

Taula 1. Distribució de les freqüències i de les freqüències acumulades

Notes TF1	Teleespectadors TF1	Acumulat teleespectadors TF1	Teleespectadors France 2	Acumulat teleespectadors France 2
0	0	0	1	1
1	0	0	1	2
2	1	1	1	3
3	0	1	1	4
3,5	0	1	1	5
4	0	1	0	5
5	1	2	2	7
5,5	0	2	2	9
6	3	5	4	13
6,5	1	6	1	14
7	4	10	3	17
7,5	0	10	1	18
8	4	14	0	18
8,5	1	15	1	19
9	5	20	1	20
9,5	2	22	0	20
10	1	23	0	20
10,5	0	23	1	21
11	2	25	1	22
11,5	1	26	0	22
12	5	31	2	24
12,5	0	31	2	26
13	2	33	1	27
14	1	34	0	27
14,5	0	34	1	28
15	1	35	2	30
15,5	0	35	1	31
16	0	35	2	33
17	0	35	1	34
18	0	35	1	35
19	1	36	0	35
20	0	36	1	36

1) a) Aquests resultats ens permeten representar els polígons de les freqüències i de les freqüències acumulades de les notes obtingudes per TF1 i France 2.



1) b)

	TF1	France 2
Mitjana	9,48	9,37
Mediana	9	8
Moda	9 i 12	6
Desviació estàndard	3,17	5,11
Coeficient de variació	33%	54%

Podem destacar que per a TF1 la distribució és bimodal, mentre que per a France 2 és unimodal asimètrica negativa. En efecte:

Moda	<	Mediana	<	Mitjana
6	<	8	<	9,37

Els dos noticiaris televisats tenen unes notes mitjanes molt similars, però amb un efectiu de notes baixes clarament més elevat i una dispersió més forta per a France 2.

2) Si creem unes classes d'amplitud 2 (oberta a l'esquerra i tancada a la dreta, excepte l'última classe, que està tancada a la dreta), obtenim deu classes.

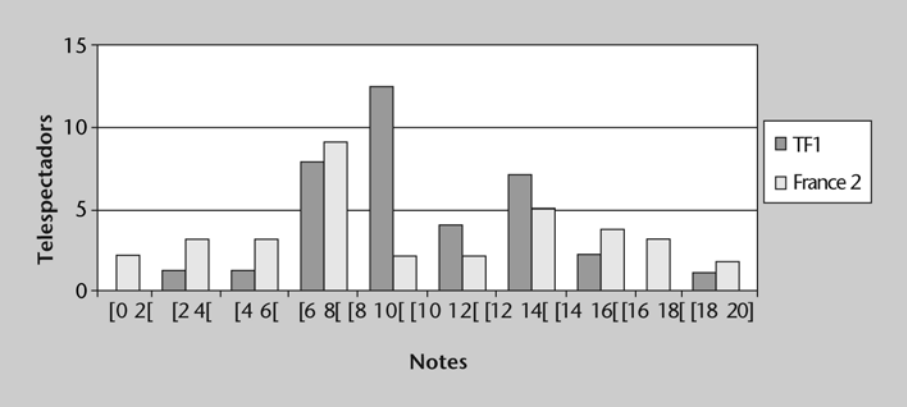
Taula 2. Distribució de les freqüències en classes d'amplitud 2

Notes	Nre. teleespectadors TF1	Nre. teleespectadors France 2
[0 2[	0	2
[2 4[	1	3
[4 6[	1	4
[6 8[	8	9
[8 10[	12	2

Notes	Nre. teleespectadors TF1	Nre. teleespectadors France 2
[10 12[	4	2
[12 14[	7	5
[14 16[	2	4
[16 18[	0	3
[18 20]	1	2

Per comparar les dues distribucions, dibuixem els dos histogrames en una mateixa gràfica.

Histograma de les freqüències (classes d'amplitud 2)



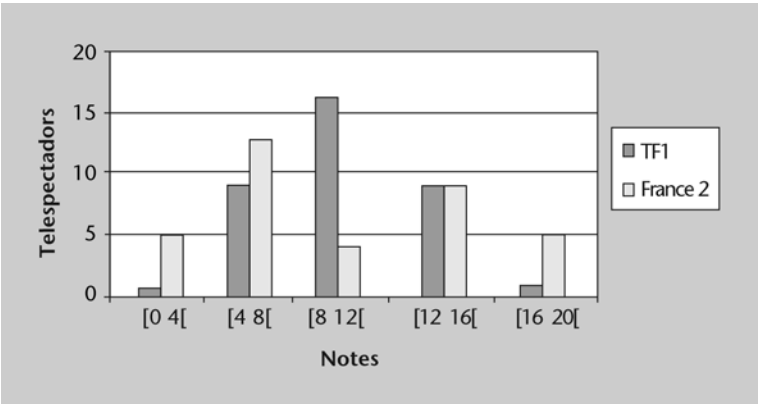
Es pot destacar que la distribució de les notes de TF1 és unimodal. Tal com es pot veure, la divisió en classes millora la llegibilitat dels resultats.

Si creem unes classes d'amplitud 4, obtenim cinc classes:

Taula 3. Distribució de les freqüències en classes d'amplitud 4

	TF1	France 2
[0 4[	1	5
[4 8[	9	13
[8 12[	16	4
[12 16[	9	9
[16 20[	1	5

La gràfica corresponent és la següent:

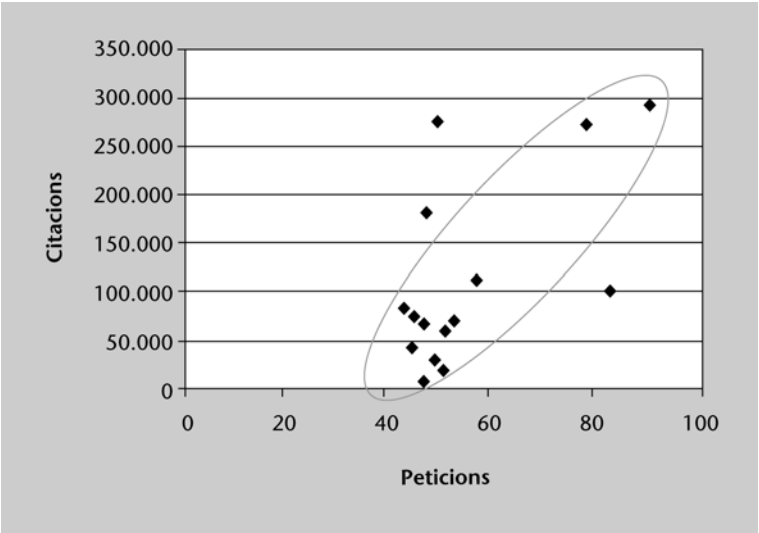


La distribució de TF1 és, doncs, unimodal simètrica.

Solucions als exercicis d'estadística tridimensional

7. Traçarem la gràfica que representa les diferents revistes en forma de punts de l'abscissa *nombre de sol·licituds* i de l'ordenada *nombre de citacions*. Els punts queden agrupats en un espai que té la forma d'un núvol allargat (punteigs).

Nombre de revistes demanades



La forma de núvol indica que hi ha una relació entre les sol·licituds i les citacions. Ara calcularem el coeficient de correlació lineal (coeficient de Pearson). Els programes informàtics d'estadística i els fulls de càlcul ofereixen funcions que el permeten calcular.

Amb quatre xifres significatives, el càlcul dóna un coeficient de correlació igual a $\mathbf{r}_p = 0{,}6335$. En la taula de **r**, amb un error del 5% i 13 graus de llibertat (que es calcula de la manera següent: nombre de revistes – 2), el coeficient **r** teòric és de 0,514. La **r** que hem calculat és superior. Així, doncs, podem arribar a la conclusió, amb un risc d'error del 5%, que les variables estan correlacionades. Com més vegades se cita una revista, més es demana i al revés.

Podeu consultar informació sobre taules estadístiques en l'annex 2, "Taules estadístiques".

Observació:

Per a calcular el coeficient de correlació, s'han de seguir els passos que indiquem a continuació:

1) Calcular les mitjanes del nombre de sol·licituds (56,73) i del nombre de citacions (111.977,86).

2) Calcular el quadrat de les desviacions de la mitjana (s_x i s_y), i també el producte de les desviacions de la mitjana (s_{xy}) per a cada línia. Els resultats s'anoten en les columnes s_{xy}, s_x i s_y.

Per exemple, per a la primera línia de la taula tenim el següent:

$s_{xy} = 6.384.893{,}64 = (92 - 56{,}73) \cdot (293.024 - 111.977{,}86)$
$s_x = 1.243{,}74 = (92 - 56{,}73)^2$
$s_y = 32.777.702.395 = (293.024 - 111.977{,}86)^2$

	Revistes	Demandes	Citacions	s_x	s_y	s_{xy}
1	*Journal of Biological Chemistry*	92	293.024	1.243,74	32.777.702.395	6.384.893,64
2	*Lancet*	85	100.526	799,00	131.145.250	323.706,10
3	*Proceedings of the National Academy of Science - USA*	79	276.659	495,80	27.119.875.676	3.666.899,90
4	*New England Journal of Medicine*	57	118.106	0,07	37.554.018	1.634,17
5	*Nature*	53	270.777	13,94	25.217.164.747	592.850,10
6	*Biochimica et Biophysica Acta*	52	69.557	22,40	1.799.529.929	200.792,10
7	*Chest*	52	19.456	22,40	8.560.295.811	437.936,84

	Revistes	Demandes	Citacions	s_x	s_y	s_{xy}
8	*Biochemical and Biophysical Research Communications*	51	56.993	32,87	3.023.335.562	315.246,57
9	*Journal of Chromatography*	50	30.985	45,34	6.559.844.451	545.351,97
10	*Journal of the American Chemical Society*	49	174.540	59,80	3.914.020.527	483.813,83
11	*International Journal of Pharmaceutics*	48	4.373	76,27	11.578.807.330	939.749,17
12	*Circulation*	47	64.644	94,74	2.240.494.934	460.716,30
13	*Journal of Clinical Investigation*	46	72.249	115,20	1.578.382.847	426.423,17
14	*Febs Letters*	45	45.436	137,67	4.427.820.019	780.757,90
15	*Cancer Research*	45	82.343	137,67	878.225.322	347.715,77

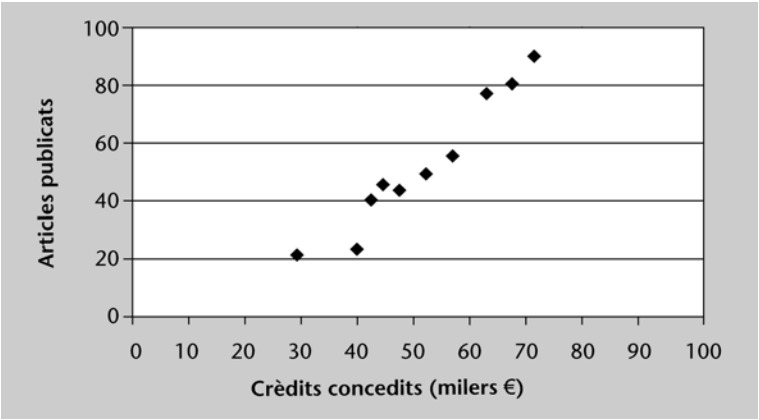
3) Càlcul de σ_{xy} com a mitjana dels elements de la columna s_{xy}, càlcul de σ_x i de σ_y com l'arrel quadrada de la mitjana dels elements de la columna s_x i de la columna s_y. Així, $\sigma_{xy} = 873.849,831$; $\sigma_x = 14,83$; $\sigma_y = 93.039$.

4) Finalment, amb quatre xifres significatives, càlcul del coeficient de correlació:

$$r_p = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = 0,6335$$

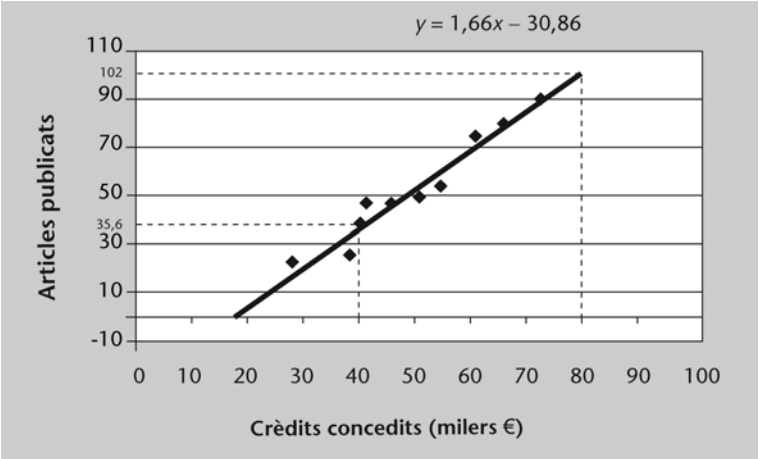
8.

a) Dibuixarem el núvol de punts que representa els laboratoris a partir dels crèdits concedits en l'abscissa i el nombre d'articles publicats en l'ordenada.



El núvol de punts té una forma allargada i segueix una recta amb un pendent positiu, per la qual cosa es pot suposar que hi ha una relació lineal bastant forta (punts relativament alineats) entre els crèdits concedits i els articles publicats, que evolucionen en el mateix sentit.

b) L'equació d'una recta té la forma de $y = ax + b$, per la qual cosa la recta de regressió té com a equació $y = 1,66x - 30,86$. Així, doncs, passa pels punts (0;–30,86) i (40;35,6).



c) Si dibuixem una perpendicular a l'abscissa en el valor 80, es determina gràficament el nombre d'articles publicats, que està lleugerament per sobre dels cent.

Per al càlcul directe, si en l'equació de la recta de regressió substituïm la variable *x* pel valor del crèdit de 80.000 euros, s'obté:

$$1{,}66 \cdot 80 - 30{,}86 = 101{,}94$$, és a dir, arrodonint, 102 articles.

d) El coeficient de correlació és igual a $\mathbf{r}_p = 0{,}9723$. Amb un marge d'error del 5% i un grau de llibertat de 8, la taula de **r** ens dóna 0,632. La **r** calculada és, doncs, superior. En conseqüència, podem arribar a la conclusió que tenim una correlació lineal positiva que és significativa amb un risc d'error del 5%; és a dir, que com més articles publiquen els laboratoris, més crèdits els concedeix el Ministeri d'Investigació, o bé, al revés, com més crèdits concedeix el Ministeri, més articles es publiquen.

9. La suma dels valors obtinguts en la línia o en la columna (taula 1) dóna un total de 1.561 lectors que han respost.

Taula 1. Taula dels efectius observats

Tema de l'obra / tipus de lector	Jove	Adult	Estudiant	Total
Ciències de la informació	5	45	23	73
Ciències humanes i socials	10	8	48	66
Ciències exactes i aplicades	25	9	73	107
Art, oci i esports	7	99	21	127
Literatura (excepte novel·les)	2	45	63	110
Història i geografia	34	38	93	165
Novel·les	65	576	49	690
Còmics	124	24	75	223
Total	**272**	**844**	**445**	**1.561**

Dir que no hi ha cap dependència entre les dues variables significa que es distribueixen de manera diferent una en relació amb l'altra. Així, si calculem el percentatge que representa cada categoria de lector en relació amb el total obtenim el següent:

Per als 272 lectors de la categoria "jove": 17,42%
Per als 844 lectors de la categoria "adult": 54,06%
Per als 445 lectors de la categoria "estudiant": 28,5%

S'obtindria una distribució equiprobable dels diferents tipus de lectors segons les categories de les obres si tinguéssim la mateixa distribució de lectors per a les obres de tots els temes. Així, doncs, per a les setanta-tres obres de ciències de la informació:

$73 \cdot 0{,}1742$, és a dir, 12,72 lectors "joves"
$73 \cdot 0{,}5406$, és a dir, 39,46 lectors "adults"
$73 \cdot 0{,}285$, és a dir, 20,81 lectors "estudiants"

Si fem el mateix càlcul per a totes les línies, obtenim la taula dels efectius teòrics:

Taula 2. Taula dels efectius teòrics

Tema de l'obra / tipus de lector	Jove	Adult	Estudiant	Total
Ciències de la informació	12,72	39,47	20,81	73
Ciències humanes i socials	11,50	35,68	18,81	66
Ciències exactes i aplicades	18,64	57,85	30,50	107
Art, oci i esports	22,13	68,67	36,20	127
Literatura (excepte novel·les)	19,17	59,47	31,36	110

Tema de l'obra / tipus de lector	Jove	Adult	Estudiant	Total
Història i geografia	28,75	89,21	47,04	165
Novel·les	120,23	373,07	196,70	690
Còmics	38,86	120,57	63,57	223
Total	**272**	**844**	**445**	**1.561**

Per a determinar si hi ha alguna dependència, és necessari calcular el χ^2 total, que és la suma de tots els χ^2 per casella. Es calculen de la manera següent:

$$\frac{(O-T)^2}{T}$$

O i *T* són els efectius observats i teòrics llegits en les taules. En la taula 3 trobarem els valors de χ^2 per casella, les contribucions de χ^2 de cada línia i de cada columna (total de línia i total de columna) i, en l'última cel·la ombrejada, la χ^2 total (suma de totes les contribucions de línia i de columna).

Taula 3. χ^2 per casella i χ^2 total

Tema de l'obra / tipus de lector	Jove	Adult	Estudiant	Total
Ciències de la informació	4,68	0,77	0,23	5,69
Ciències humanes i socials	0,19	**21,47**	**45,27**	66,94
Ciències exactes i aplicades	2,16	**41,25**	**59,20**	102,62
Art, oci i esports	10,34	13,40	6,38	30,12
Literatura (excepte novel·les)	15,37	3,52	**31,92**	50,82
Història i geografia	0,95	29,39	**44,91**	75,26
Novel·les	25,37	**110,38**	**110,90**	246,66
Còmics	**186,56**	**77,34**	2,05	265,96
Total	245,66	297,56	300,89	**844,11**

La χ^2 és de 844,11. El valor de la χ^2 llegit en la taula, amb el marge d'error del 5% i el nombre de graus de llibertat (ddl) de 14 (ddl = (8 – 1)(3 – 1) = 14), és de 23,68. La χ^2 calculada queda per sobre d'aquest valor amb escreix, per la qual cosa hi ha dependència entre el tema de les obres i el tipus de lector.

D'una manera més precisa, és possible explicar aquesta dependència observant els valors forts de χ^2 per casella (valors ombrejats). Determinem el sentit de les possibles relacions segons el signe de la diferència entre els efectius observats i els efectius teòrics. Una diferència positiva indica una atracció entre les dues variables, mentre que una diferència negativa indica una repulsió.

Així, les obres de ciències humanes i socials i les de ciències exactes i aplicades són molt poc sol·licitades pels adults però molt pels estudiants. Els còmics són bàsicament sol·licitats pels joves (són, a més, les seves lectures principals) i pels adults. I, finalment, les novel·les són molt sol·licitades pels adults i molt poc pels estudiants.

10. Per a posar a prova la dependència eventual entre el fet de ser pertinent i el fet de ser extret, usem la prova de χ^2 per a un document. Primer calcularem els efectius teòrics suposant que les dues variables nominals *ser extret sí/no* i *ser pertinent sí/no* són independents.

Taula 1. Efectius teòrics

Documents	Extrets	No extrets
Pertinents	96,25	453,75
No pertinents	78,75	371,25

A continuació calcularem la diferència entre els efectius observats i els efectius teòrics. Una diferència positiva indica una atracció, mentre que una diferència negativa indica una repulsió entre les dues variables.

Taula 2. Desviacions entre els efectius teòrics i els efectius observats

Documents	Extrets	No extrets
Pertinents	+13,75	–13,75
No pertinents	–13,75	+13,75

El resultat ens mostra que hi ha una forta dependència (+13,5) entre ser pertinent i ser extret, i també entre ser no pertinent i ser no extret.

Calculem la força de la relació de dependència; per a això calcularem χ^2 de la taula creuada després d'haver calculat les χ^2 per casella.

Taula 3. χ^2 per casella

Documents	Extrets	No extrets
Pertinents	1,96	0,42
No pertinents	2,4	0,51

La $\chi^2_{\text{calculada}}$ total és igual a 5,29.

La χ^2_{lu} en la taula amb un grau de llibertat 1 (ddl = $(2 - 1) \cdot (2 - 1)$) i un risc d'error del 5% dóna $\alpha = 5\%$ $\chi^2 = 3{,}84$.

Aquest resultat ens permet afirmar amb una probabilitat d'equivocar-nos inferior al 5% que *ser extret* i *ser pertinent* són depenents dos a dos.

Per afinar aquests resultats, calcularem el percentatge de la contribució de cada χ^2 per casella (taula 4).

Taula 4. Percentatge de les χ^2 per casella

Documents	Extrets	No extrets
Pertinents	37,13%	7,88%
No pertinents	45,38%	9,63%

El parell (extret, no pertinent) representa pràcticament la meitat (entorn del 45%) de la contribució total. No obstant això, el signe negatiu de la taula 2 indica que hi ha repulsió entre el fet de ser extret i el fet de ser no pertinent.

El parell (extret, pertinent) representa el 37% de la contribució total. El signe positiu de la taula 2 indica que hi ha una atracció entre el fet de ser extret i el fet de ser pertinent. Per contra, hi ha una repulsió entre ser extret i ser no pertinent.

Així, doncs, el programa de documentació és eficaç.

11. La taula dels efectius teòrics és:

Efectius teòrics

	C++	C=	C—
P++	75,74	249,47	289,79
P=	94,34	310,72	360,94
P—	30,91	101,81	118,27

La χ^2 total és 115,54. Amb 4 graus de llibertat (ddl = $(3 - 1) \cdot (3 - 1)$), podem afirmar que aquestes variables són depenents amb un risc d'error inferior a 1 sobre 1.000 ($\chi^2_{lu} = 18{,}47$).

Amb aquest càlcul tan senzill no es pot dir si el fet d'agafar cada vegada més llibres en préstec comporta el fet de comprar cada vegada més o menys llibres. És necessari un estudi més precís.

En primer lloc, calcularem la diferència entre els efectius observats i els efectius teòrics per veure si hi ha atracció o repulsió per a les modalitats de cada variable agafades de dues en dues.

Diferència entre els efectius observats i els teòrics

	C ++	**C =**	**C – –**
P ++	10,26	–66,47	**56,21**
P =	–40,34	90,28	–49,94
P– –	**30,09**	–23,81	–6,27

Hi ha atracció entre els dos parells (C + + P – –) i (C – – P + +).

Per quantificar la força de la relació, calculem χ^2 per casella.

Càlcul de χ^2 per casella

	C ++	**C =**	**C – –**
P ++	1,39	17,71	**10,90**
P =	17,25	26,23	6,91
P – –	**29,28**	5,57	0,33

Tots aquests càlculs es resumeixen en la taula següent:

	χ^2	% de χ^2	% de χ^2 acumulada	**Signe de la desviació**
C++P – –	29,28	**25,34%**	25,34%	+
C=P=	26,23	22,70%	48,04%	+
C=P++	17,71	15,32%	63,36%	–
C++P=	17,25	14,93%	78,29%	–
C– –P++	10,90	**9,43%**	87,72%	+
C– –P=	6,91	5,97%	93,70%	–
C=P– –	5,57	4,82%	98,52%	–
C++P++	1,39	1,19%	99,71%	+
C– –P– –	0,33	0,29%	100,00%	–

La dependència entre les dues variables de compra i préstec es deu, sobretot, a la dependència entre *com més compro* i *menys agafo en préstec* (25% de la χ^2). En efecte, la relació més forta procedeix de la contribució (C + + P– –). La segona contribució (22% de la χ^2 total) correspon a la constància de les pràctiques de compra i de préstec. La relació *com menys compro / més agafo en préstec* representa menys del 10% de la χ^2 total (concretament, el 9,43%). Per norma general, el gust per la lectura estimula la gent a comprar llibres.

12.

a) Distribucions de les freqüències i polígons de freqüències (vegeu la taula 1 en la pàgina següent).

La línia 9 (ombrejada) de la taula vol dir que la nota 6 ha estat posada:

- en el noticiari de TF1 per 1 teleespectador amb estudis superiors
- en el noticiari de TF1 per 2 teleespectadors sense estudis superiors
- en el noticiari de France 2 per 1 teleespectador amb estudis superiors
- en el noticiari de France 2 per 3 teleespectadors sense estudis superiors

A partir d'aquesta taula, construïm el polígon de freqüències per a les notes de TF1 i de France 2:

Taula 1. Distribució de les freqüències

Nota	TF1 SUP	TF1 NSUP	France 2 SUP	France 2 NSUP
0	0	0	0	1
1	0	0	0	1
2	0	1	1	0
3	0	0	0	1
3,5	0	0	0	1
4	0	0	0	0
5	0	1	0	2
5,5	0	0	0	2
6	**1**	**2**	**1**	**3**
6,5	1	0	0	1
7	1	3	0	3
7,5	0	0	0	1
8	2	2	0	0
8,5	1	0	1	0
9	2	3	0	1
9,5	2	0	0	0
10	0	1	0	0
10,5	0	0	1	0
11	1	1	0	1
11,5	1	0	0	0
12	2	3	2	0
12,5	0	0	2	0
13	1	1	1	0
14	0	1	0	0
14,5	0	0	0	1
15	0	1	1	1
15,5	0	0	1	0
16	0	0	1	1
17	0	0	1	0
18	0	0	1	0
19	0	1	0	0
20	0	0	1	0

Figura 1. Polígons de les freqüències per a les notes de TF1

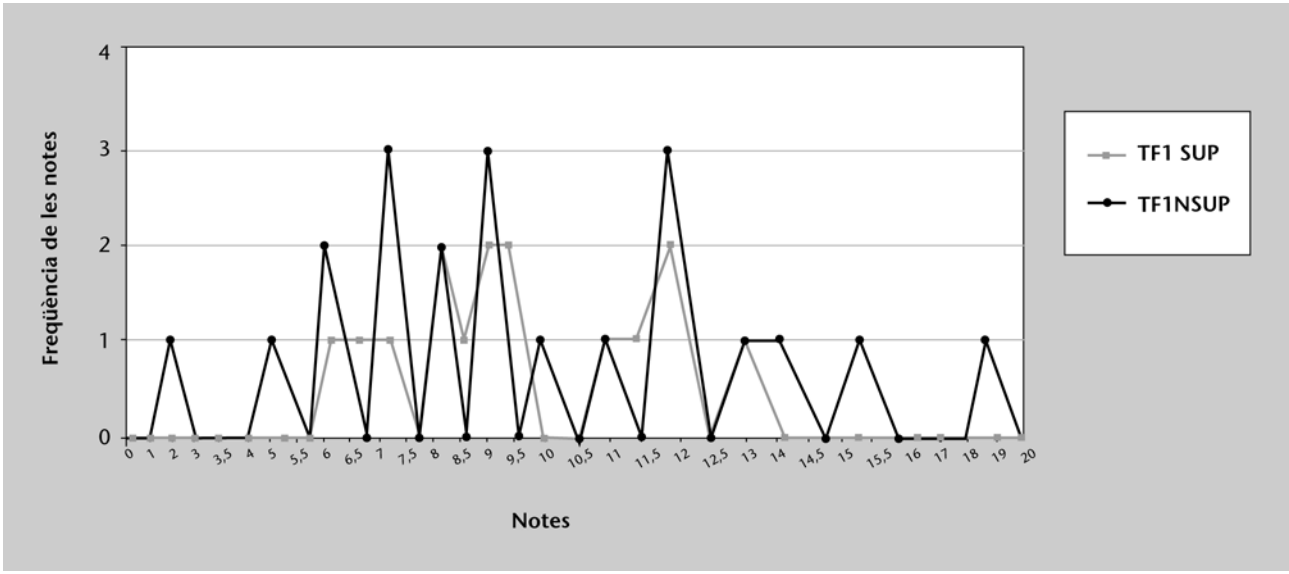
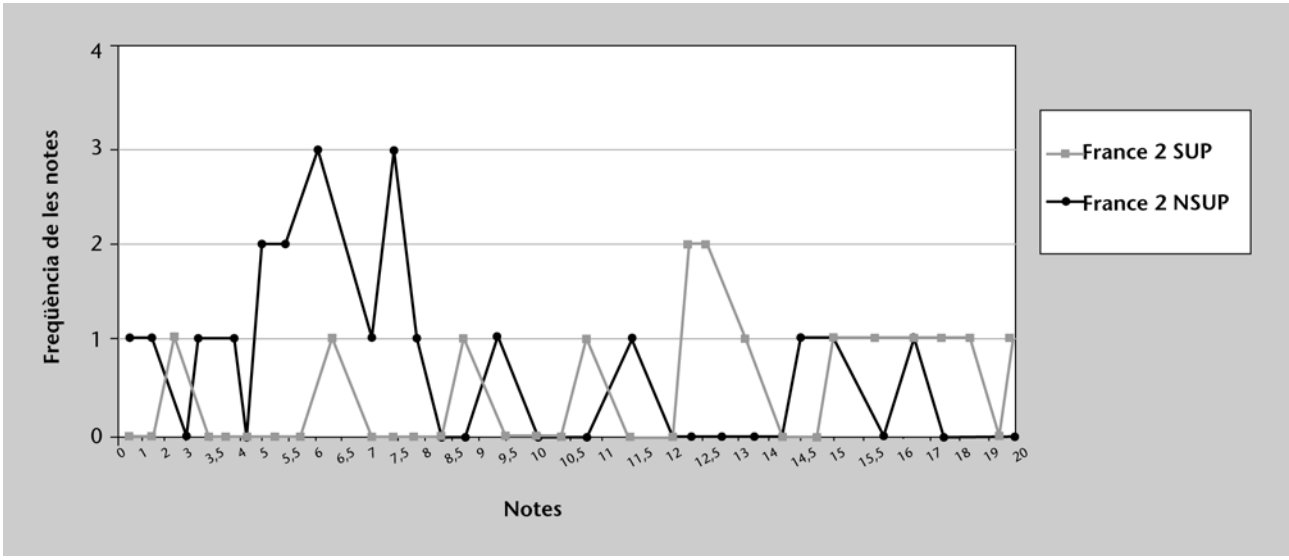


Figura 2. Polígons de les freqüències per a les notes de France 2



Amb aquestes gràfiques és difícil resumir visualment la informació. Així, doncs, és necessari construir la taula de les freqüències acumulades i dibuixar els polígons d'aquestes.

Taula 2. Distribució de les freqüències acumulades

Nota	Acumulat de TF1 SUP	Acumulat de TF1 NSUP	Acumulat de France 2 SUP	Acumulat de France 2 NSUP
0	0	0	0	1
1	0	0	0	2
2	0	1	1	2
3	0	1	1	3
3,5	0	1	1	4
4	0	1	1	4
5	0	2	1	6
5,5	0	2	1	8
6	1	4	2	11

Nota	Acumulat de TF1 SUP	Acumulat de TF1 NSUP	Acumulat de France 2 SUP	Acumulat de France 2 NSUP
6,5	2	4	2	12
7	3	7	2	15
7,5	3	7	2	16
8	5	9	2	16
8,5	6	9	3	16
9	8	12	3	17
9,5	10	12	3	17
10	**10**	**13**	**3**	**17**
10,5	10	13	4	17
11	11	14	4	18
11,5	12	14	4	18
12	14	17	6	18
12,5	14	17	8	18
13	15	18	9	18
14	15	19	9	18
14,5	15	19	9	19
15	15	20	10	20
15,5	15	20	11	20
16	15	20	12	21
17	15	20	13	21
18	15	20	14	21
19	15	21	14	21
20	15	21	15	21

La línia 17 (ombrejada) de la taula significa que han donat la nota 10 o inferior:

- al noticiari de TF1, 10 teleespectadors amb estudis superiors
- al noticiari de TF1, 13 teleespectadors sense estudis superiors
- al noticiari de France 2, 3 teleespectadors amb estudis superiors
- al noticiari de France 2, 17 teleespectadors sense estudis superiors

A partir d'aquesta taula, construïm els polígons de les freqüències acumulades per a les notes de TF1 i de France 2.

Figura 3. Polígons de les freqüències acumulades per TF1

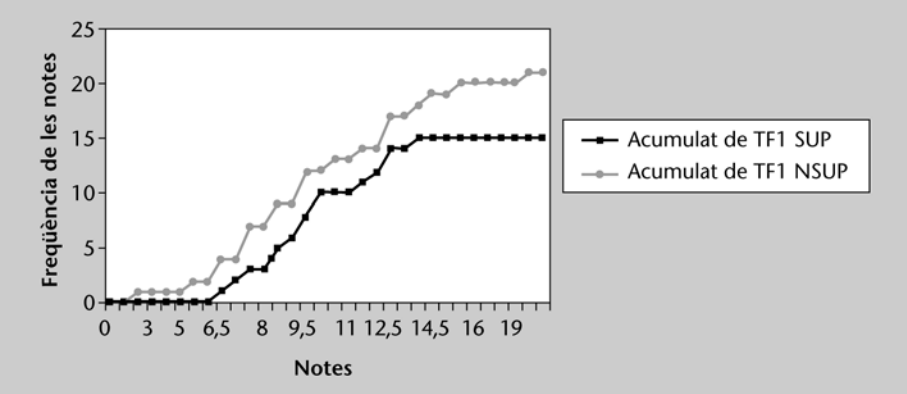
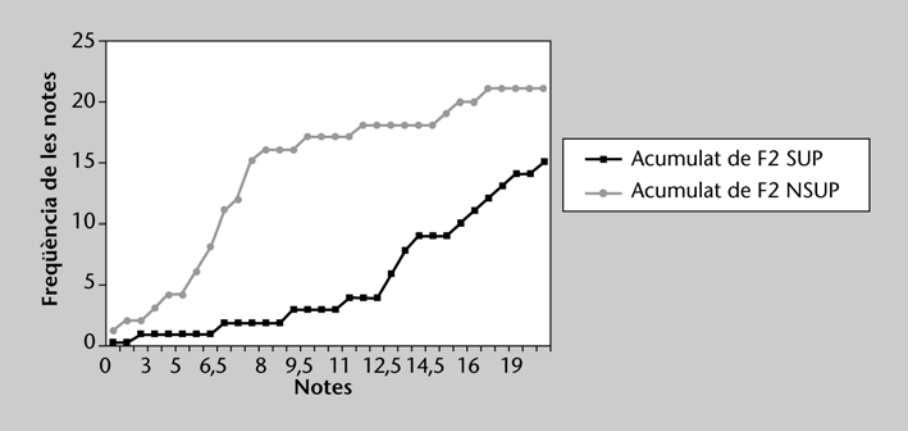


Figura 4. Polígons de les freqüències acumulades per France 2



Els polígons de les freqüències acumulades mostren que el noticiari televisiu de France 2 és més ben valorat pels teleespectadors amb estudis superiors, mentre que en el cas de TF1 no hi ha cap diferència important.

Ara determinarem les mesures de tendència central i les mesures de dispersió per a les dues categories de teleespectadors i per a les dues cadenes de televisió.

	TF1/SUP	TF1/NSUP	F2/SUP	F2/NSUP
Mitjana	9,37	9,57	12,7	7,08
Desviació estàndard	2,14	3,87	4,68	4,28
Mediana	9	9	12,5	6,25
Moda(es)	8-9-9, 5-12	7-9-12	12-12,5	6-7
Coeficient de variació	22,84%	40,44%	36,90%	60,45%

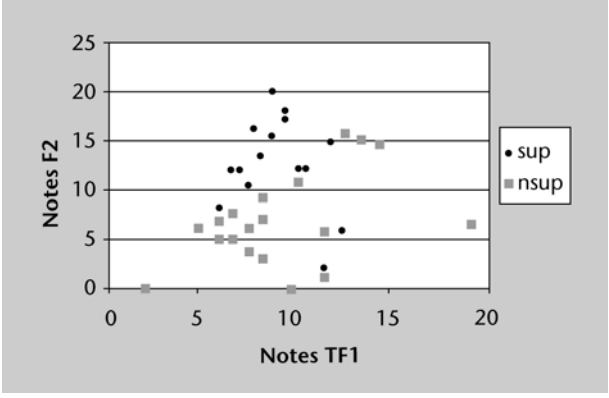
Per a TF1, els dos grups de teleespectadors (els que tenen estudis superiors i els que no en tenen) posen una nota mitjana equivalent. Per al noticiari de France 2, els teleespectadors amb estudis superiors posen una nota mitjana més alta que els que no n'han cursat.

Per a France 2, més de la meitat dels teleespectadors amb estudis superiors posen una nota superior a 12,5 (nota mitjana), mentre que la meitat dels teleespectadors sense estudis superiors posen una nota inferior a 6,25 (nota mitjana).

Per a TF1, la dispersió de les notes rebudes és més important en el grup dels que no han cursat estudis superiors. El coeficient de variació és del 40,44%, en lloc del 22,84% dels que han cursat estudis superiors.

b) Representació mitjançant núvols de punts de les dues sèries de notes.

Figura 5. Representació per núvols de punts de les dues sèries de notes



L'observació dels núvols de punts posa de manifest dues poblacions diferents: els teleespectadors amb estudis superiors donen una nota més favorable a France 2, mentre que els tele-

espectadors sense estudis superiors puntuen més favorablement TF1. El coeficient de correlació lineal és de 0,146. La prova de r amb 30 graus de llibertat dóna 0,349 (en realitat, ddl = 36 – 2 = 34; com que aquest valor no és a la taula, escollim un altre valor per fer la prova; no obstant això, sabem que el valor llegit en la taula serà superior a 0,349). Així, doncs, les dues variables són independents.

c) Les notes obtingudes per TF1 són menys disperses que les obtingudes per France 2. Així, doncs, la nota de TF1 no sembla discriminar els teleespectadors segons el nivell d'estudis. Per contra, la nota de France 2 separa d'una manera més clara els dos grups de teleespectadors. De mitjana, avaluen més favorablement France 2 que TF1. En canvi, les seves notes són més disperses.

Solucions als exercicis d'estadística multidimensional

13.

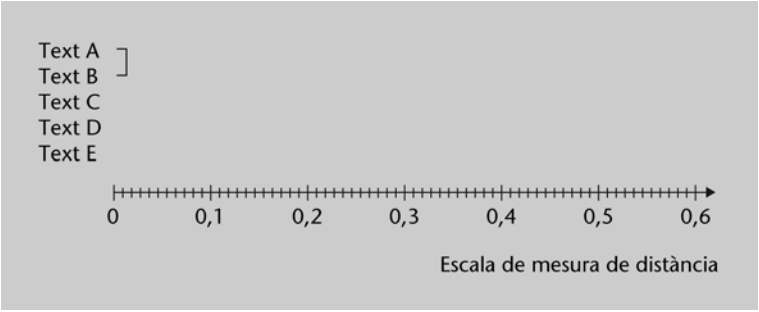
a) Com a exemple, el càlcul de la distància entre el text A i el text D és:

$$d(A,D) = 1 - \frac{4\cdot 1 + 10\cdot 1 + 2\cdot 1 + 4\cdot 6}{\sqrt{4^2+10^2+2^2+4^2}\cdot\sqrt{1^2+1^2+1^2+6^2}} = 1 - 0{,}549 = 0{,}451$$

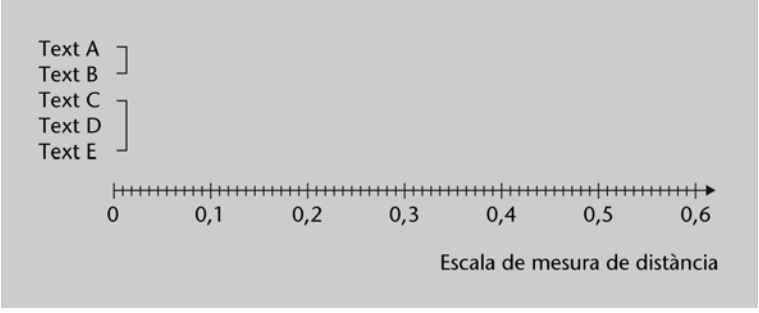
Procedint de la mateixa manera per als altres parells de text, calculem la matriu de semblances per als cinc textos:

	Text A	Text B	Text C	Text D	Text E
Text A	0	0,001	0,141	0,451	0,162
Text B	0,001	0	0,130	0,423	0,153
Text C	0,141	0,130	0	0,609	0,015
Text D	0,451	0,423	0,609	0	0,609
Text E	0,162	0,153	0,015	0,609	0

b) L'estratègia d'agregació escollida és la del salt mínim. La distància més feble es dóna entre el text A i el text B, que és de 0,001. Nosaltres els reunim i donem a aquesta classe el nom de [text A – B]. La nova divisió està formada, doncs, per [textos A – B], el text C, el text D i el text E.

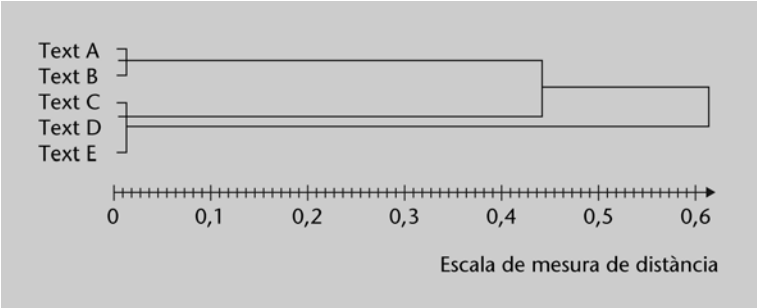


A continuació, la distància mínima es dóna entre [text A – B] i el text C. És de 0,13, la més petita de les distàncies entre (text A, text C) i (text B, text C). D'altra banda, la distància més petita es dóna entre el text C i el text E. És de 0,015. Com que és inferior a 0,13, agreguem el text C i el text E.



A continuació passem al text D. La distància més petita es dóna amb el text B, per la qual cosa l'agreguem amb el [text A – B].

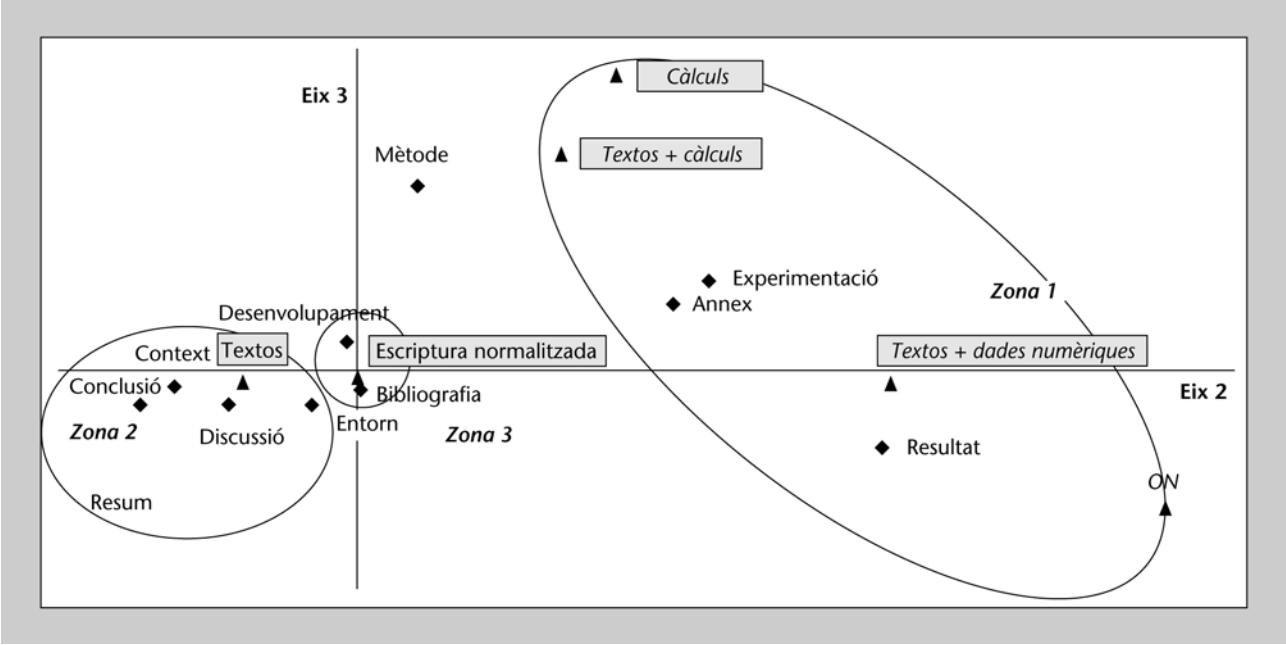
La classificació final és, doncs, [text A – B – D], [text C – E].



Els textos A, B i D tracten bàsicament del tema de l'anàlisi relacional, mentre que els textos C i E tracten principalment de la vigilància informativa.

14.

a) En la gràfica es posen de manifest tres zones.



- La zona 1 agrupa les parts purament científiques dels textos, és a dir, les que contenen els càlculs, les dades numèriques relatives a l'experimentació i als resultats i els annexos.
- La zona 2 agrupa les parts del discurs que es presenten essencialment en forma textual, com la presentació del context i de l'entorn, la introducció, etc.
- La zona 3 agrupa les parts normalitzades i codificades dels textos, com la bibliografia i els desenvolupaments.

Hi ha una gran oposició (repulsió) entre les formes discursives experimentals (les que acostumen a presentar les dades numèriques amb xifres i símbols) i els textos, mentre que les escriptures normalitzades es troben entre tots dos.

b) Examinem les χ^2 per casella dels elements d'aquestes tres zones.

Zona 1. Els valors més forts de les χ^2 per casella són els relatius a l'experimentació, el mètode i els resultats en la seva relació amb textos i dades numèriques, dades numèriques i càlculs, és a dir, amb les parts bàsicament matemàtiques. Així, doncs, la zona 1 es correspon molt bé amb una característica específica.

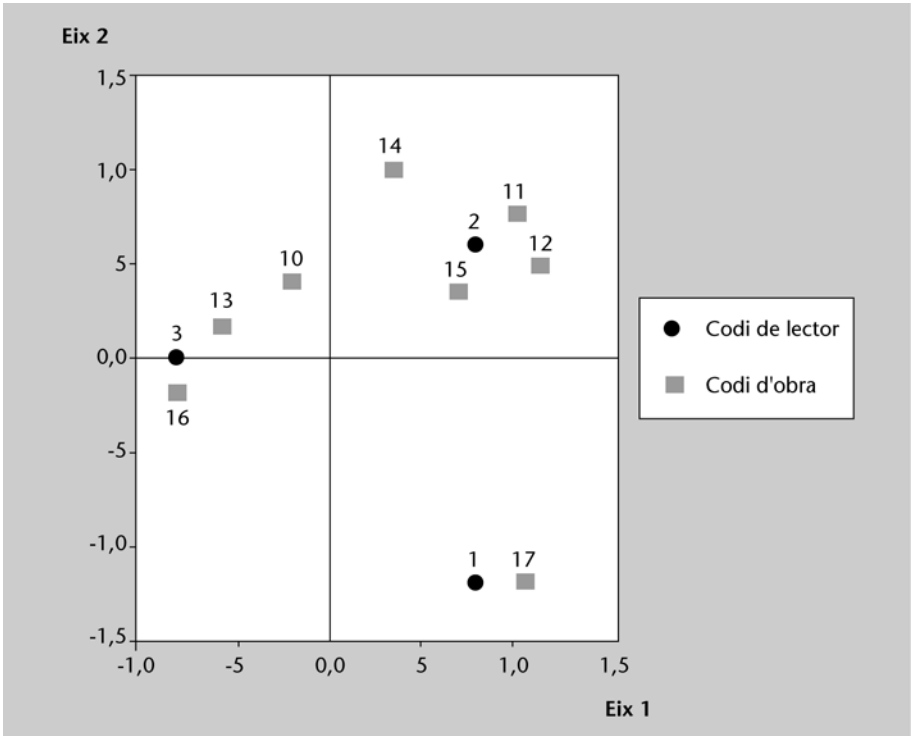
Zona 2. Les parts de text de la zona 2, que són el context, la conclusió, la introducció, etc., no tenen uns valors de χ^2 per casella especialment elevats en els seus vincles amb la presentació textual.

Zona 3. Finalment, quant a la zona 3, la bibliografia té una relació molt forta amb l'escriptura normalitzada (χ^2 per casella de 375), mentre que el desenvolupament no. Aquesta anomalia s'explica justament per la projecció sobre un plànol de variables definides en l'espai. La gràfica presenta els eixos 2 i 3, i l'eix 1 es troba en la vertical del punt d'origen. En efecte, el parell (bibliografia, escriptura normalitzada) se situa "en altura" al llarg de l'eix 1.

Així, doncs, la conclusió anterior a la qual havíem arribat és falsa; l'oposició entre les formes discursives es dóna entre les parts normalitzades i les parts matemàtiques. Les presentacions textuals no són especialment discriminants. En efecte, un paràgraf sempre conté caràcters alfabètics i, en conseqüència, es considera com a textual. Les presentacions discriminants són les presentacions matemàtiques i les presentacions normalitzades. Aquests grups es troben en dos eixos diferents: hi ha repulsió entre ells.

15.

a) En la gràfica següent, els temes de les obres es presenten amb quadrats i els tipus de públic amb cercles (per raons de comoditat de lectura de la gràfica, els nombres d'uns i altres s'indiquen en xifres, que són les que apareixen ombrejades en la taula de l'enunciat).



b) Les representacions gràfiques planars no acostumen a ser fiables, ja que només permeten donar compte de dues dimensions, i els efectes de projecció podrien induir a errors d'interpretació en els casos en què es maneja un gran nombre de dimensions. En aquest cas, la variable del públic només està definida per tres modalitats, per la qual cosa no podrem tenir més de dos eixos d'informació (3 – 1 = 2). En aquest cas, doncs, la representació planar no és cap font d'errors.

Vegeu el subapartat 4.2 d'aquest mòdul.

c) Veiem d'una manera molt clara que els temes de les obres s'agrupen entorn de cada tipus de lector. El tema "còmics" (codi 17) es troba proper al lector "jove" (codi 1). Els temes "novel·les" (codi 16) i "art" (codi 13) estan propers al lector "adult" (codi 3). Finalment, els temes "ciències humanes i socials" (codi 11), "ciències exactes i aplicades" (codi 12) i "història i geografia" (codi 15) estan propers als lectors "estudiants". Els temes "ciències de la informació" (codi 10) i "literatura" (codi 14) atreuen per igual els lectors "estudiants" i els lectors "adults", però molt poc els lectors "joves". El tema "ciències de la informació" està aïllat i atreu una mica més els adults que els estudiants.

Solucions als exercicis d'estadística probabilista

16.

a) La taxa mitjana de rotació és el nombre mitjà de vegades que es deixa una obra durant el període estudiat, és a dir, el quocient del nombre de préstecs en relació amb el nombre d'obres de la col·lecció.

Com que no coneixem el nombre de préstecs que els hem d'atribuir realment, no tindrem en compte les obres que s'han deixat en préstec més de nou vegades, per la qual cosa només considerarem 398 obres de la col·lecció. A més, com que hi ha menys de cinc obres que s'han deixat sis, set i vuit vegades, les agruparem.

- Tenint en compte el fons mort (200 obres no deixades), tenim un total de 398 obres:

  92 obres que s'han deixat 1 vegades, la qual cosa ens dóna 92 préstecs.
  49 obres que s'han deixat 2 vegades, la qual cosa ens dóna 98 préstecs.
  24 obres que s'han deixat 3 vegades, la qual cosa ens dóna 72 préstecs.
  15 obres que s'han deixat 4 vegades, la qual cosa ens dóna 60 préstecs.
  10 obres que s'han deixat 5 vegades, la qual cosa ens dóna 50 préstecs.

8 = 4 + 2 + 2 obres que s'han deixat més de 6 vegades, la qual cosa ens dóna un mínim de 48 préstecs.

Així, doncs, el total és 92 + 98 + 72 + 60 + 50 + 48 = 420 préstecs.

La taxa mitjana de rotació és, doncs:

$$\bar{x} = \frac{420}{398} = 1{,}05.$$

- La variància associada és:

$$V = 2{,}10.$$

La desviació estàndard és, doncs, d'1,45.

b) El mètode dels moments consisteix a dir que una distribució observada s'ajusta mitjançant una distribució teòrica concreta si les seves mitjanes (o moment d'ordre 1) són iguals.

Cas geomètric:

L'esperança (moment d'ordre 1) d'una llei geomètrica $G(u)$ és $E = \frac{u}{1-u}$. Així, doncs, $u = \frac{E}{1+E}$, és a dir, $u = 0{,}513$.

Podeu consultar informació sobre el càlcul de moments en l'annex 3, "Càlcul de moments".

Cas binomial negatiu:

L'esperança (moment d'ordre 1) d'una llei binomial negativa $Bn(r,u)$ és $E = \frac{ru}{1-u}$. I la seva variància (moment d'ordre 2 centrat) és $V = \frac{ru}{(1-u)^2}$.

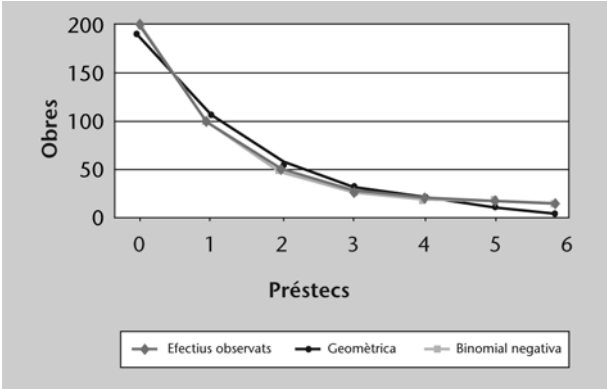
És a dir, que $1 - u = \frac{E}{V}$ i $r = \frac{E^2}{V-E}$.

Així, doncs, $u = 0{,}5$ i $r = 1{,}05$.

Els valors de les distribucions són:

Préstecs	Obres	Geomètric	Binomial negatiu
0	200	193,65	191,41
1	92	99,43	101,14
2	49	51,05	51,97
3	24	26,21	26,45
4	15	13,46	13,40
5	10	6,91	6,77
6	8	3,55	3,41

La gràfica corresponent és:



Les distribucions geomètrica i binomial negativa semblen pràcticament idèntiques a la distribució observada dels préstecs. Comprovem quina és la millor calculant χ^2.

Hipòtesi H0: la distribució observada s'ajusta mitjançant la llei geomètrica.

Podeu consultar informació sobre les proves en estadística en l'annex 1, "Les proves en estadística".

En aquest cas, el nombre de graus de llibertat és cinc (ddl = 7 – 1 – 1); si considerem una fiabilitat del 5%, la χ^2_{lu} de la taula és 11,07. La $\chi^2_{\text{calculada}}$ és igual a 8,17. La $\chi^2_{\text{calculada}}$ és inferior a la χ^2_{lu}. Podem acceptar la hipòtesi H0.

Hipòtesi H'0: La distribució observada s'ajusta mitjançant la llei binomial negativa.

En aquest cas, el nombre de graus de llibertat és quatre (ddl = 7 – 2 – 1); si considerem una fiabilitat del 5%, la χ^2_{lu} de la taula és 11,07. La $\chi^2_{\text{calculada}}$ és igual a 9,49. La $\chi^2_{\text{calculada}}$ és igual a 9,51. Així, doncs, la $\chi^2_{\text{calculada}}$ és superior a la χ^2_{lu}. No podem acceptar la hipòtesi H'0.

Com a conclusió, podem dir, amb un risc d'error del 5%, que la distribució dels préstecs d'obres segueix una llei de probabilitat geomètrica.

17.

a) La distribució estadística dels préstecs el 1999 descriu el nombre total d'obres deixades en préstec una vegada, dues vegades, etc., és a dir, el total de la 1a. fila, de la 2a. fila, etc. Els valors es donen en l'última columna de la taula 1 (i el mateix per a la distribució del 2000, sumant aquesta vegada les columnes). Les representacions gràfiques es presenten en les figures 1 i 2.

Taula 1. Càlcul de les distribucions de 1999 i 2000

1999/2000	0	1	2	3	4	5	Distribució 1999
0	195	17	9	3	2	1	227
1	24	8	3	1	0	0	36
2	2	1	3	1	1	1	9
3	5	3	0	0	0	0	8
4	1	0	1	0	1	1	4
5	1	2	0	2	0	1	6
Distribució 2000	228	31	16	7	4	4	290

Figura 1. Distribució dels préstecs el 1999

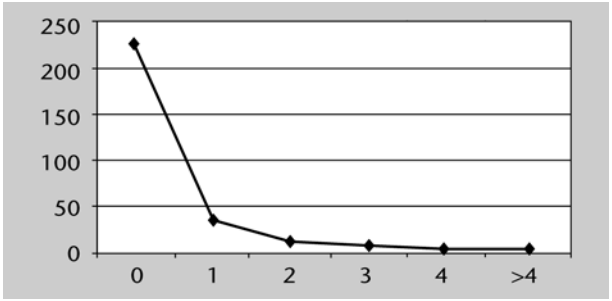
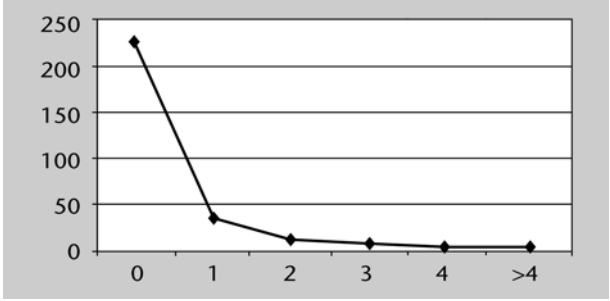


Figura 2. Distribució dels préstecs el 2000



b) Sabem que el 1999 hi ha 227 obres que mai no s'han deixat en préstec i que han generat

$$17 + (2 \cdot 9) + (3 \cdot 3) + (4 \cdot 2) + 5 = 57$$

préstecs el 2000, la qual cosa dóna una taxa de rotació

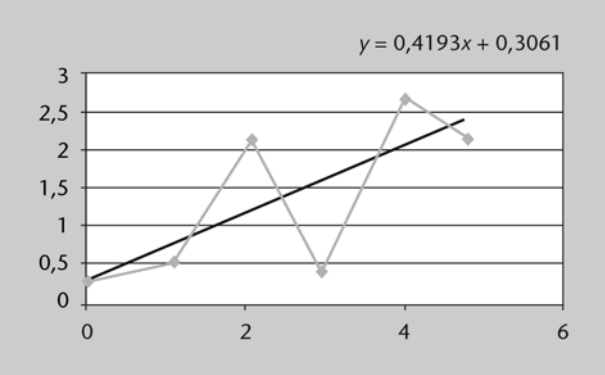
$$M(0) = \frac{57}{227} = 0{,}251$$

Calculem les altres taxes de rotació de la mateixa manera. Els resultats són els que apareixen en la taula 2 i en la figura 3.

Taula 2

Nombre de préstecs el 1999	Taxa mitjana de rotació el 2000
0	0,25
1	0,47
2	2,11
3	0,38
4	2,75
5	2,17

Figura 3. Taxa mitjana de rotació el 2000 sabent que les obres s'han deixat en préstec *j* vegades el 1999.



La corba més propera a tots els punts és la recta de l'equació $y = 0{,}3061 + 0{,}4193 \cdot x$. És a dir, si reprenem les notacions de l'enunciat:

$$M(j) = 0{,}3061 + 0{,}4193 \cdot j$$

El coeficient $a = 0{,}3061$ correspon a la taxa de rotació mínima del fons, independentment del període (la podem assimilar a la circulació mitjana de les obres més antigues). El coeficient $b = 0{,}4193$ correspon al creixement o al descens dels préstecs de les obres de la classe en el nou període; n'és la mesura de la popularitat.

Observació

Aquest tipus d'ajust és el més clàssic i el més senzill, però no és el millor. En efecte, no permet tenir en compte les ocurrències de cada categoria de préstec. Encara que la classe de no-préstecs és molt forta i la de cinc préstecs és molt feble, resulta interessant observar que aquestes dues classes es tracten de la mateixa manera.

c) Suposem que la distribució segueix una llei de Poisson del tipus:

$$P(i/j) = \frac{(0{,}3061 + 0{,}4193 \cdot j)^i \, e^{-(0{,}3061+0{,}4193 \cdot j)}}{i!}$$

Podem calcular cadascuna de les probabilitats $P(i/j)$. Per exemple, la probabilitat que un llibre deixat en préstec dues vegades el 1999 es deixi dues vegades el 2000 és:

$$P(2/2) = \frac{(0{,}3061 + 0{,}4193 \cdot 2)^2 \, e^{-(0{,}3061+0{,}4193 \cdot 2)}}{2!} = 0{,}21$$

El càlcul de cadascuna de les probabilitats $P(i/j)$ apareix en la taula 3 (els resultats s'han arrodonit a dos decimals).

Taula 3. Càlcul de les probabilitats $P(i/j)$

j/i	0	1	2	3	4	5
0	0,74	0,23	0,03	0,00	0,00	0,00
1	0,48	0,35	0,13	0,03	0,01	0,00
2	0,32	0,36	0,21	0,08	0,02	0,01
3	0,21	0,33	0,26	0,13	0,05	0,02
4	0,14	0,27	0,27	0,18	0,09	0,04
5	0,09	0,22	0,26	0,21	0,13	0,06

Amb els efectius de l'any 1999, podem calcular la distribució teòrica 1999/2000. Per exemple, en la distribució de 1999, sabent que hi ha 9 llibres que s'han deixat 2 vegades, podem preveure que hi haurà $9 \cdot 0,32 = 2,88$ llibres que es deixaran 0 vegades en el 2000, $9 \cdot 0,36 = 3,24$ llibres que es deixaran 1 vegada en el 2000, etc. Els resultats són els que apareixen en la taula 4:

Taula 4. Matriu de transició teòrica 1999/2000 i distribució teòrica de l'any 2000

1999/2000	0	1	2	3	4	5
0	167,98	52,21	6,81	0,00	0,00	0,00
1	17,28	12,60	4,68	1,08	0,36	0,00
2	2,88	3,24	1,89	0,72	0,18	0,09
3	1,68	2,64	2,08	1,04	0,40	0,16
4	0,56	1,08	1,08	0,72	0,36	0,16
5	0,54	1,32	1,56	1,26	0,78	0,36
Distribució teòrica 2000	190,92	73,09	18,10	4,82	2,08	0,77

Si sumem cada columna obtenim la distribució teòrica de l'any 2000. La distribució teòrica és molt propera a la distribució del 2000 que apareix en la taula 1 per als valors superiors a 3.

d) Amb tot rigor, a partir dels efectius observats el 2000 hauríem de dibuixar la gràfica de la taxa mitjana de rotació per al 2001, sabent que el 2000 s'han deixat en préstec j obres (tal com hem fet en la figura 3), i així podríem determinar els nous paràmetres del model de Poisson que permet calcular les probabilitats $P(i/j)$ de transició 2000/2001. Suposarem que les variacions són tan fiables que és perfectament possible conservar els paràmetres calculats a partir de la transició observada per al 1999/2000. Així, aplicant els càlculs de probabilitats de la taula 3 a la distribució del 2000, obtenim la matriu de transició 2000/2001 (taula 5). Per exemple, sabent que el 2000,n'hi va haver 16 que es van deixar 2 vegades, calculem que el 2001 hi haurà $5,12 = 16 \cdot 0,32$ llibres que es deixaran en préstec 0 vegades, el 2001 hi haurà $5,76 = 16 \cdot 0,36$ llibres que es deixaran 1 vegada, el 2001... Els resultats apareixen en la taula 5. Igual que abans, sumant les xifres de les columnes obtenim la distribució teòrica de previsió per a l'any 2001.

Taula 5. Matriu de transició teòrica 2000/2001 i distribució teòrica de previsió per a l'any 2001

2000/2001	0	1	2	3	4	5
0	168,72	52,44	6,84	0,00	0,00	0,00
1	14,88	10,85	4,03	0,93	0,31	0,00
2	5,12	5,76	3,36	1,28	0,32	0,16
3	1,47	2,31	1,82	0,91	0,35	0,14
4	0,56	1,08	1,08	0,72	0,36	0,16
5	0,36	0,88	1,04	0,84	0,52	0,24
Previsions de distribució per al 2001	191,11	73,32	18,17	4,68	1,86	0,70

18.

a) Suposem que totes les paraules tenen la mateixa probabilitat d'aparèixer; el nombre total de paraules en els dos períodes és de 86.280, per la qual cosa la probabilitat d'aparició d'una paraula durant el període 1 és:

$$p = \frac{44.086}{86.280} = 0,51$$

D'això deduïm que la probabilitat de no-aparició és $q = 1 - p = 0,49$. De la mateixa manera, la probabilitat d'aparició d'una paraula en el període 2 és de $\frac{42.194}{86.280} = 0,49$. La probabilitat de no-aparició és de 0,51.

Constatem que la probabilitat d'aparició (o de no-aparició) d'una paraula en el període 1 (o el període 2) és igual a la probabilitat de no-aparició (o d'aparició) d'una paraula durant el període 2 (o el període 1).

Aquest resultat era previsible. En efecte, si PARAULA1, PARAULA2, PARAULA designen, respectivament, el nombre de paraules del corpus de textos dels períodes 1, 2 i tots dos junts, tenim:

$$p = \frac{\text{PARAULA1}}{\text{PARAULA}} = \frac{\text{PARAULA1}}{\text{PARAULA1} + \text{PARAULA2}} = \frac{\text{PARAULA1} + \text{PARAULA2} - \text{PARAULA2}}{\text{PARAULA1} + \text{PARAULA2}} =$$
$$= 1 - \frac{\text{PARAULA2}}{\text{PARAULA1} + \text{PARAULA2}} = 1 - \frac{\text{PARAULA2}}{\text{PARAULA}}$$

b) L'efectiu teòric de cada adverbi per a cada període és proporcional a la probabilitat d'aparició d'una paraula.

Per exemple, per a l'adverbi *probablement*, tenim un efectiu teòric de $48 \cdot 0,51$, és a dir, aproximadament 24.

El càlcul de les desviacions entre els efectius observats i teòrics dóna el resultat següent:

Desviació observada/teòrica	**Probablement**	**Certament**	**Evidentment**
Període 1	11,47	–14,66	–2,26
Període 2	–11,47	14,66	2,26

Constatem que la suma de les desviacions és nul·la per a cada adverbi. Aquest resultat era previsible; en efecte, si n_1 i n_2 designen el nombre d'ocurrències d'un adverbi per als dos períodes, el resultat obtingut en la pregunta 1 ens permet escriure:

$$n_1 - p(n_1 + n_2) + n_2 - (1 - p)(n_1 + n_2) = 0.$$

Com que la desviació és positiva (o negativa), això significa que l'efectiu observat és superior (o inferior) a l'efectiu teòric i que, per tant, l'adverbi és més (o menys) freqüent en aquest període i serà característic d'aquest. El problema que es planteja és saber a partir de quin valor aquesta desviació és significativa. Ho serà si la distribució no és aleatòria.

c) Tenim les hipòtesis següents:

- H_0^1 : l'adverbi *probablement* es distribueix a l'atzar.
- H_0^2 : l'adverbi *certament* es distribueix a l'atzar.
- H_0^3 : l'adverbi *evidentment* es distribueix a l'atzar.

Si suposem que la distribució és aleatòria, el més natural és escollir una distribució que tingui com a mitjana l'efectiu teòric anterior, és a dir, $n \cdot u$, en què n és el nombre d'ocurrències de l'adverbi i u la probabilitat d'èxit. Així, doncs, la distribució binomial $B(n,u)$ escollida té com a paràmetres n i u.

Adverbi *probablement*:

$n = 48$, $u = 0,51$, període 1
$n = 48$, $u = 0,49$, període 2

Adverbi *certament*:

$n = 60$, $u = 0{,}51$, període 1
$n = 60$, $u = 0{,}49$, període 2

Adverbi *evidentment*:

$n = 24$, $u = 0{,}51$, període 1
$n = 24$, $u = 0{,}49$, període 2

Calculem la mitjana per a cadascuna d'aquestes distribucions, que és igual al producte $n \cdot u$ (destacarem que la mitjana és igual que l'efectiu teòric calculat prèviament), i la desviació estàndard és igual a $\sqrt{nu(1-u)}$. Tot seguit, determinem el valor de l'efectiu centrat reduït (conegut simplement com a *desviació reduïda*) mitjançant la fórmula:

$$\frac{\text{Efectiu} - nu}{\sqrt{nu(1-u)}}$$

Així, doncs, tenim:

Desviació reduïda	Probablement	Certament	Evidentment
Període 1	3,31	–3,79	–0,92
Període 2	–3,31	3,79	0,92

Els paràmetres de la llei binomial (n prou gran i p pròxim a 0,5) permeten fer una aproximació i assimilar la llei binomial a la llei normal.

Segons la taula de la llei normal, amb un risc d'error de l'1% ($\alpha = 0{,}01$), podem dir que les desviacions no superen el valor de 2,576 ($t_\alpha = 2{,}576$). Els valors absoluts 3,31 i 3,79 dels adverbis *probablement* i *certament* són superiors; per tant, les seves distribucions no segueixen la de la llei normal i no estan vinculades a l'atzar. Així, doncs, podem rebutjar H_0^1 i H_0^2. El valor absolut de 0,92 per a l'adverbi *evidentment* és inferior a 2,576; com que aquesta distribució respecta les regles de la llei normal, no rebutgem H_0^3. Així, doncs, els adverbis *probablement* i *certament* tenen un paper particular en el discurs dels períodes 1 i 2, mentre que *evidentment* no té cap paper. Els valors positius ens permeten precisar que el primer període editorial de la publicació està marcat significativament per la presència de l'adverbi de judici *probablement*, mentre que el segon període està marcat per l'adverbi *certament*.

Podeu consultar informació sobre la taula de la llei normal en l'annex 2, "Taules estadístiques".

19. El nombre de connexions durant el temps t de dos minuts és una variable aleatòria C, el valor mitjà de la qual és:

$$C = \frac{N \cdot t}{T}$$

T és el temps de funcionament del servidor, per la qual cosa tenim $24 \cdot 60 = 1.440$ minuts.

Per a $N = 70$, $C = 0{,}10$
Per a $N = 410$, $C = 0{,}57$
Per a $N = 1.400$, $C = 1{,}94$
Per a $N = 2.100$, $C = 2{,}92$
Per a $N = 3.200$, $C = 4{,}44$

Aquests valors són petits en comparació de N, per la qual cosa podem usar la llei de Poisson amb una bona aproximació. Un usuari podrà usar el servidor amb la condició que no hi hagi cap altre usuari en la línia, és a dir, si $r = 0$ (l'expressió de la fórmula de Poisson en el curs, ja esmentada).

El càlcul dóna:

Per a $N = 70$, $P = 0{,}91$: tenim un 91% de possibilitats d'interrogar el servidor després de 2 minuts d'espera.

Per a $N = 410$, $P = 0{,}57$: tenim un 57% de possibilitats d'interrogar el servidor després de 2 minuts d'espera.

Per a $N = 1.400$, $P = 0,14$: tenim un 14% de possibilitats d'interrogar el servidor després de 2 minuts d'espera.

Per a $N = 2.100$, $P = 0,05$: tenim un 5% de possibilitats d'interrogar el servidor després de 2 minuts d'espera.

Per a $N = 3.200$, $P = 0,01$: tenim un 1% de possibilitats d'interrogar el servidor després de 2 minuts d'espera.