Matemàtica de la informació

Thierry Lafouge
Yves-François Le Coadic
Christine Michel

PID_00189703

Universitat Oberta de Catalunya

Índex

Introducció

"Poloni: Què llegiu, senyor meu?
Hamlet: Paraules, paraules, paraules."

William Shakespeare
La tragèdia de Hamlet, príncep de Dinamarca

En les ciències de la informació, els recomptes d'objectes informatius es van fer ja des del començament, i van obrir la via a l'ús de sèries matemàtiques. També es van constatar les relacions entre dues magnituds per a veure la manera en què qualsevol variació de la primera comporta una variació corresponent en la segona; es diu que una està en funció de l'altra. Tots aquests desenvolupaments no són independents: es basen en la teoria de conjunts.

Abans de començar a estudiar-los, explicarem, tal com vam fer per als nombres en general per a l'estadística, les nocions de nombre infinitament petit i de nombre infinitament gran.

1. Precisió i límit en matemàtiques: el zero i l'infinit

El nombre, l'objecte privilegiat dels primers matemàtics, és l'element principal de tots els processos matemàtics. Entre altres coses, ens ha permès fer recomptes, és a dir, comptar. Fins aquest moment hem usat diferents tipus de nombres: els nombres enters positius i negatius i el número zero, els nombres fraccionals i els nombres decimals.

1.1. La precisió

Podem conèixer sempre el valor d'un nombre amb precisió? La resposta és *no*. En efecte, un nombre es coneix amb precisió a partir del moment en què coneixem tot el conjunt de les xifres del seu desenvolupament decimal (si treballem en base 10). Des de l'antiguitat ja se sap que hi ha nombres el desenvolupament decimal dels quals no coneixem amb exactitud, com $\sqrt{2}$ i π (pi), que és 3,1416 amb una precisió de quatre decimals, i que s'ha calculat amb una precisió d'uns dos milions de xifres després de la coma. Així doncs, podem considerar que π s'ha definit amb una infinitat de xifres després de la coma. Què significa aquesta infinitud i què significa la noció matemàtica d'infinit?

Exemple

Agafem el número trencat $\frac{1}{3}$:

- Amb una precisió de 3 decimals, el seu desenvolupament decimal és 0,333.
- Amb una precisió de 6 decimals, el seu desenvolupament decimal és 0,333333.
- Amb una precisió de 10 decimals, el seu desenvolupament decimal és 0,3333333333.

En realitat, la precisió màxima possible consistiria a escriure una sèrie sense fi de xifres 3 després de la coma, la qual cosa en la pràctica és impossible. Per aquesta raó, els matemàtics han introduït un objecte matemàtic conegut com a *infinit* i representat per ∞.

Així, doncs, podem escriure $\frac{1}{3}$ com una suma infinita de nombres:

$$0,3=\frac{3}{10'}\ 0,03=\frac{3}{100}=\frac{3}{10^{2'}}\ 0,003=\frac{3}{10^{3'}}\ 0,0003=\frac{3}{10^{4}}\ldots$$

És a dir:

$$\frac{1}{3}=\frac{3}{10}+\frac{3}{10^2}+\frac{3}{10^3}+\ldots+\frac{3}{10^i}+\ldots=\sum_{i=1}^{\infty}\frac{3}{10^{i'}}$$

És a dir, la suma de $i = 1$ a ∞ de $\frac{3}{10^i}$.

Quins són els límits d'aquests recomptes?

1.2. El límit

Quan fem la suma de nombres U_i (aquí, per motius de senzillesa, suposarem que tots són positius), si i és cada vegada més gran i tendeix a l'infinit, el resultat:

- pot tendir cap a un límit infinitament gran,
- pot tendir cap a un límit finit,
- o bé no tenir cap límit.

Aquesta suma és l'addició d'una quantitat infinita de nombres, però això no implica que ella mateixa sigui infinita.

$$U_1 + U_2 + U_3 + U_4 + \ldots + U_i + U_{i+1} + \ \ldots = \sum_{i=1}^{\infty} U_i$$

- **Límit infinitament gran:** si els U_i són cada cop més i més grans, la suma és ella mateixa més gran i diem que tendeix a l'infinit.

  Exemple

  Considerem la suma de 2^i, en què i varia d'1 a l'infinit.

  $$\sum_{i=1}^{\infty} 2^i = 2 + 4 + 8 + 16 + \ldots$$

  Correspon a un nombre infinitament gran.

- **Límit finit:** a la inversa, si els U_i són cada cop més petits quan i és cada cop més gran, la suma de tots aquests nombres U_i pot ser finita.

  Exemple

  Tornem a l'exemple anterior. i és cada cop més gran i $\frac{3}{10^i}$ és cada cop més petit. En realitat, es converteix en infinitament petit, és a dir, que s'aproxima a 0 sense arribar mai a aquest valor. En aquest cas direm que $\frac{3}{10^i}$ tendeix a 0 quan i tendeix a l'infinit, o bé que el límit de $\frac{3}{10^i}$ quan i tendeix a l'infinit és 0. S'usen indistintament les dues notacions següents:

  $$\frac{3}{10^i} \xrightarrow[i \to \infty]{} 0 \quad \text{o bé} \quad \lim_{i \to \infty} \left(\frac{3}{10^i} \right) = 0$$

L'anterior és un exemple d'una sèrie de nombres que tendeixen a un nombre finit; diem que és convergent i que convergeix a 0.

D'altra banda, sabem que:

$$\lim_{n \to \infty} \sum_{i=1}^{n} \left(\frac{3}{10^i} \right) = \frac{1}{3}$$

En l'altre cas, direm que:

$$\sum_{i=1}^{n} 2^i \xrightarrow[n\to\infty]{} \infty \text{ o bé que } \lim_{n\to\infty}\left(\sum_{i=1}^{n} 2^i\right) = \infty$$

Aquesta sèrie de nombres que tendeixen a un nombre infinit es diu que és divergent.

Tornarem a trobar aquests dos conceptes d'infinit i de límit en els estudis de les sèries i de les funcions. Ens permetran desenvolupar les bases de l'anàlisi numèrica, que són el càlcul diferencial i el càlcul integral, que usarem en l'apartat 4.

Observació

Si U_i no és sempre positiu, és possible que la suma dels U_i tendeixi a 0, un límit infinitament petit, que representa el conjunt buit, l'absència de quantitat. El 0 és també, tal com ja hem vist, l'origen d'una escala cardinal de qualsevol mesura, positiva o negativa.

2. Les sèries

Els recomptes d'objectes informatius són innombrables. Ens donen un conjunt de nombres $x_1, x_2, \ldots, x_i, \ldots x_n$ que ja ens han servit per a fer estudis probabilistes. Aquest conjunt de nombres, conegut com a *distribució* en estadística, es diu *sèrie* en matemàtiques.

Podeu consultar informació sobre els estudis probabilistes en l'apartat 5.

Exemple

Sèrie de dades sincrònica o asincrónica en telecomunicacions.

Tenim un terminal origen A que transmet cap al terminal destinació B un conjunt de n caràcters $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_i \ldots, \mathbf{c}_n$. Cadascun d'aquests caràcters es representa amb un conjunt de m símbols binaris (**m** depèn de la codificació escollida: en la codificació ASCII és igual a 7).

$$c_i = \left(d_i^1, \ldots, d_i^m\right) \text{ amb } \; d_i^j \in \{0,1\}$$

Després de la conversió paral·lel-sèrie, el terminal origen pot lliurar, sota la forma d'un senyal elèctric bivalent, els seus caràcters reunits uns amb altres en una successió regular en el temps, i d'aquesta manera formar una sèrie de símbols binaris. Si la sèrie de dades binàries no s'interromp, es diu que és sincrònica, i asincrónica en el cas contrari.

Una sèrie és una successió de nombres o d'expressions matemàtiques formada segons una llei coneguda de la qual considerem la suma.

Amb x_i com a iè terme de la sèrie, la suma $S_n = x_1 + x_2 + \ldots + x_i + \ldots + x_n$ dels primers termes de la successió és una sèrie:

$$S_n = \sum_{i=1}^{n} x_i$$

Una sèrie és convergent si tendeix cap a un nombre S (anomenat *límit de la sèrie*) quan n tendeix a l'infinit.

$$\lim_{n\to\infty} S_n = S = \sum_{i=1}^{\infty} x_i$$

Exemple

Si tenim la sèrie següent: $x_1 = \frac{1}{2}$; $x_2 = \frac{1}{2\cdot 2}$; $x_3 = \frac{1}{2\cdot 2\cdot 2}$, el terme iè s'escriu en aquest cas com $x_i = \frac{1}{2^i}$. La suma dels dos primers termes és $\frac{1}{2} + \frac{1}{4} = \frac{3}{4} = 0{,}75$. La suma dels tres

primers termes és igual a $\frac{7}{8}$, és a dir, 0,87... La suma dels cinc primers termes és $\frac{29}{32}$, és a dir, 0,91... La suma dels deu primers termes és 0,999.

Destacarem que la suma s'aproxima al valor 1 sense arribar-hi mai. Aquesta sèrie és convergent i el seu límit és 1.

Els dos tipus de sèries que són especialment importants en les ciències de la informació són les sèries geomètriques i les sèries hiperbòliques.

2.1. Dispersió de la literatura científica (infometria): sèrie geomètrica (llei de Bradford)

El 1934, un documentalista britànic, Samuel C. Bradford, va enunciar una llei (que porta el seu nom) que permet que els gestors d'un servei de documentació gestionin la seva col·lecció de publicacions científiques.

Bibliografia

S. C. Bradford (1934). "Sources of Information on Specific Subjects". *Engineering* (pàg. 85-86).

Si, dins d'un camp científic concret, es classifiquen les publicacions per ordre decreixent d'articles publicats, hi ha un nombre q (superior a 1) i un nombre r de publicacions tals que, si agrupem les revistes considerant les r primeres, després les rq següents, després les rq^2 següents... (és a dir, una progressió geomètrica), es pot observar que cada grup de revistes conté el mateix nombre d'articles. El nombre de revistes per grup és una sèrie geomètrica:

$$r, rq, rq^2 \ldots rq^i$$

Molt habitual, la sèrie geomètrica és tal que cada terme es calcula multiplicant el terme anterior per una constant.

Si x_1 és el primer terme de la sèrie:

$$x_2 = q \cdot x_1,\ x_3 = q \cdot x_2\ (= q^2 \cdot x_1), \ldots,\ \text{i}\ x_n = q \cdot x_{n-1}$$

en què n adopta els valors 2, 3...

La constant q es coneix com a *raó de la sèrie*. El coneixement de la sèrie està totalment determinat pel seu primer terme, que aquí és x_1, i la seva raó és q. El terme enèsim d'una sèrie geomètrica es pot escriure de la manera següent:

$$x_n = x_1 \cdot q^{n-1} \qquad n = 1, 2\ldots$$

Exemple

Si tornem a l'exemple anterior, tenim que:

$$x_1 = \frac{1}{2}\ ;\ \ x_2 = x_1 \cdot \frac{1}{2}\ ;\ \ x_3 = x_2 \cdot \frac{1}{2}\ \ldots$$

- Si q és diferent d'1, demostrem que el càlcul de la suma dels n primers termes és:

$$S_n = \sum_{i=1}^{n} x_i = x_1 \cdot \frac{1-q^n}{1-q}$$

- Si $q < 1$, q^n convergeix cap a 0 si n tendeix a l'infinit, la sèrie S_n és convergent i té S com a límit:

$$S = \frac{x_1}{1-q}\text{, és a dir, } S = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1 \text{ (vegeu l'exemple anterior).}$$

- Si $q \geq 1$, la sèrie és divergent, és a dir, no admet cap límit.

Tornem ara a l'exemple de Bradford:

- r, el primer terme, representa la literatura essencial del camp científic que es considera.
- q, la raó de la sèrie, es coneix com a *factor de Bradford*. És característica del camp estudiat.

Suposem que la literatura essencial d'un camp està formada per vint revistes i que el factor de Bradford q és igual a 2. Aquestes vint revistes contenen, per exemple, cent articles sobre el tema. Si volem consultar el triple d'articles, és a dir, tres-cents, llavors serà necessari recórrer a cent quaranta revistes en lloc de vint, és a dir, set vegades més publicacions:

$$20 + 20 \cdot 2 + 20 \cdot 2^2 = 140$$

Com es poden determinar els paràmetres r i q? Ho farem usant els resultats experimentals que Bradford va obtenir analitzant revistes de geofísica aplicada aparegudes entre 1928 i 1931 (vegeu la taula 43).

Taula 43. Articles publicats en les revistes de geofísica aplicada (segons Bradford)

Revistes	Articles	Revistes acumulades (rang)	Articles acumulats	Logaritme decimal del rang	
1	93	1	93	0,00	
1	86	2	179	0,30	
1	56	3	235	0,48	
1	48	4	283	0,60	
1	46	5	329	0,70	**1r. grup**
1	35	6	364	0,78	
1	28	7	392	0,85	
1	20	8	412	0,90	
1	17	9	429	0,95	

Revistes	Articles	Revistes acumulades (rang)	Articles acumulats	Logaritme decimal del rang	
4	16	13	493	1,11	
1	15	14	508	1,15	
5	14	19	578	1,28	
1	12	20	590	1,30	
2	11	22	612	1,34	
5	10	27	662	1,43	**2n. grup**
3	9	30	689	1,48	
8	8	38	753	1,58	
7	7	45	802	1,65	
11	6	56	868	1,75	
12	5	68	928	1,83	
17	4	85	996	1,93	
23	3	108	1.065	2,03	**3r. grup**
49	2	157	1.163	2,20	
169	1	326	1.332	2,51	

- En la la. columna veiem el nombre de revistes.
- En la 2a. columna apareix el nombre d'articles publicats en aquestes revistes; per exemple, hi ha quatre revistes que han publicat setze articles.
- En la 3a. columna tenim els valors acumulats de la 1a. columna.
- En la 4a. columna, els valors acumulats de la 2a. columna.
- En la 5a. columna hi ha el logaritme decimal de la 3a. columna (el logaritme decimal es defineix més endavant).

Podeu consultar informació sobre el logaritme decimal en el subapartat 3.2 d'aquest mòdul.

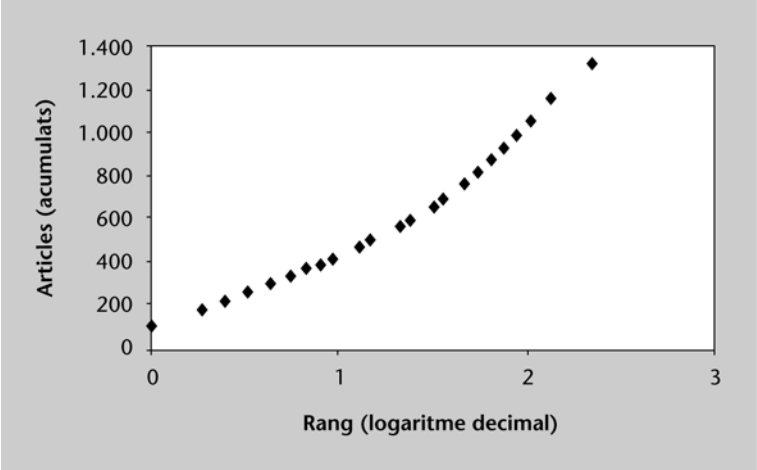
Destacarem que les columnes 3 i 4 contenen les sèries calculades a partir de les sèries de les columnes 1 i 2, i que hi ha un nombre petit de revistes que produeixen molts articles de geofísica aplicada i un gran nombre de revistes que produeixen molt pocs articles dins d'aquest mateix camp. Tal com ja havíem vist, aquest tipus de distribució és molt habitual en les ciències de la informació. De manera més exacta, podem destacar tres grups de revistes:

Podeu consultar informació sobre la distribució en les ciències de la informacion en el subapartat 3.4 del mòdul 2, "Estadística de la informació.

- Un primer grup que conté les revistes que han publicat més de quatre articles a l'any.
- Un segon grup que conté les revistes que han publicat menys de quatre articles a l'any.
- Un tercer grup que conté les revistes que han publicat com a màxim un article a l'any.

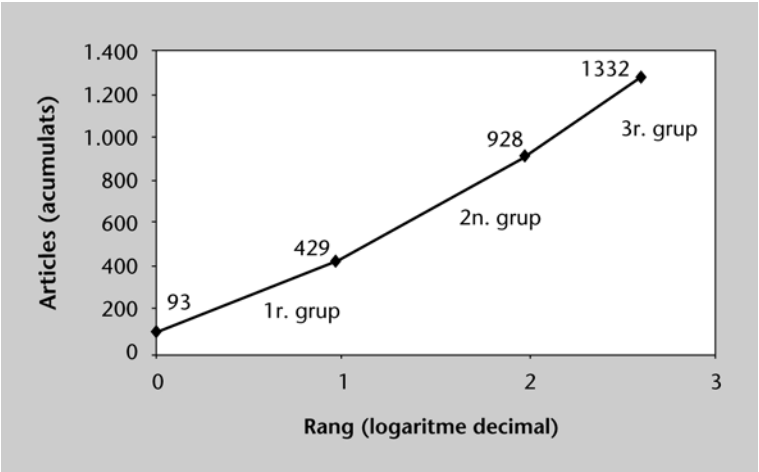
Si dibuixem la corba de variació del nombre d'articles segons el rang de la revista, constatem que és una recta en gran part de la seva longitud (gràfica 35). Això representa una proporcionalitat entre el nombre d'articles i el logaritme del nombre de revistes.

Gràfica 35. Classificació de les revistes segons el nombre d'articles publicats



En aquesta corba, es posen de manifest els tres grups de revistes (gràfica 36).

Gràfica 36. Els tres grups de Bradford



El primer grup inclou 9 revistes que han produït 429 articles, el segon conté 59 revistes (68 – 9) que han produït 499 articles (928 – 429) i el tercer conté 258 revistes (326 – 68) que han produït 404 articles (1.332 – 928).

Així, doncs, comprovem que el nombre de revistes segueix una sèrie geomètrica amb $r = 9$, $rq = 59$ i $rq^2 = 258$. Si suposem que r val 9, llavors q es troba entre 5,35 i 6,55 (en l'exercici 20 veurem una il·lustració d'aquests resultats).

2.2. Distribució de l'ús d'un lloc web (webmetria): sèrie hiperbòlica

La sèrie hiperbòlica (coneguda també com a *sèrie de Rieman*, matemàtic, 1826-1866) s'usa habitualment per a analitzar els fluxos informatius observats en les fases de producció i d'ús de la informació.

Els termes de la successió que formen la sèrie s'escriuen de la manera següent:

$$\frac{k}{1^a}; \frac{k}{2^a} \ldots\ldots\ldots\ldots \frac{k}{i^a} \ldots\ldots\ldots\ldots$$

en què a i k són dues constants positives o nul·les, i a es coneix com a *raó de la sèrie*.

La suma S_n dels n primers nombres s'escriu:

$$S_n = \sum_{i=1}^{n} \frac{k}{i^a}$$

Demostrem que:

Si $a \leq 1$, llavors la sèrie és divergent.
Si $a > 1$, llavors la sèrie és convergent.

Observació

No hi ha cap fórmula senzilla que permeti, igual que per a una sèrie geomètrica, calcular la suma dels n primers termes.

La noció de límit per a una sèrie no és una noció intuïtiva. En efecte, podríem pensar, erròniament, que és suficient que el terme de rang n de la sèrie sigui cada cop més petit i tendeixi a 0 quan n tendeix a l'infinit perquè la sèrie convergeixi. L'exemple següent demostra el contrari:

- Si $a = 1$, sabem que la sèrie és divergent. O bé, per a $a = 1$ i $k = 1$, tenim:

$$x_1 = \frac{1}{1};\ x_2 = \frac{1}{2};\ x_3 = \frac{1}{3} \ldots\ldots\ldots\ldots\ x_n = \frac{1}{n}$$

- Si x_n tendeix a 0 quan n tendeix a l'infinit, podem demostrar que $\sum_{i=1}^{n} x_i$ divergeix quan n tendeix a l'infinit.

1) Ús d'un lloc web

Exemple

En la taula 44 es presenten les estadístiques de freqüentació mensual d'un lloc web. L'última fila agrupa el nombre de visitants que es connecten més de deu vegades al lloc; per als càlculs només tindrem en compte el valor 10.

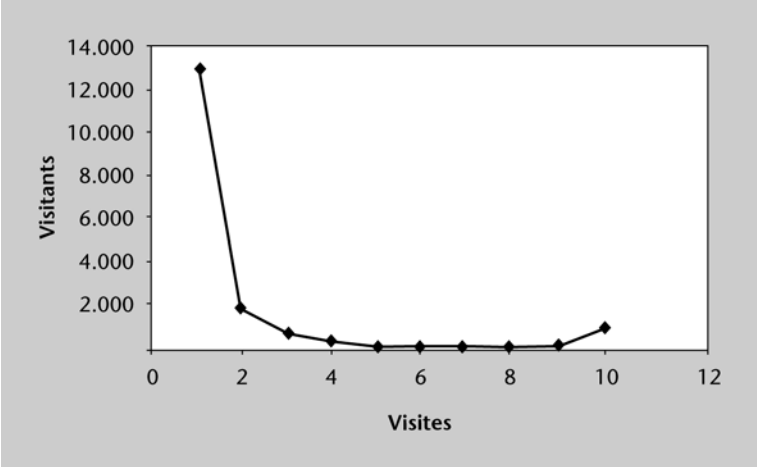
Taula 44. Estadístiques de freqüentació mensual d'un lloc web

Visites U_i	Visitants G_i	Logaritme neperià del nombre de visites $\ln(U_i)$	Logaritme neperià del nombre de visitants $\ln(G_i)$
1	12.749	0	9,45
2	1.411	0,69	7,25
3	574	1,10	6,35
4	313	1,39	5,75
5	240	1,61	5,48
6	151	1,79	5,02
7	116	1,95	4,75
8	92	2,08	4,52
9	74	2,20	4,30
>10	597	2,30	6,39

La representació gràfica de les dues primeres columnes mostra una distribució hiperbòlica del nombre de visitants en funció del nombre de visites (gràfica 37), és a dir, que hi ha una relació del tipus:

$$G_i = \frac{k}{U_i^a} k > 0, a > 0$$

Gràfica 37. Freqüentació mensual d'un lloc web



Fem una regressió lineal per calcular els valors dels paràmetres k i a. Això exigeix una transformació logarítmica dels valors:

Podeu consultar informació sobre la regressió lineal en el subapartat 3.2.2 del mòdul 2 "Estadística de la informació".

Podeu consultar informació sobre la transformació logarítmica en el subapartat 3.2 d'aquest mòdul.

$$y = \frac{k}{x^a} \Leftrightarrow \ln(y) = \ln(k) - a \cdot \ln(x) \Leftrightarrow Y = B + A \cdot X$$

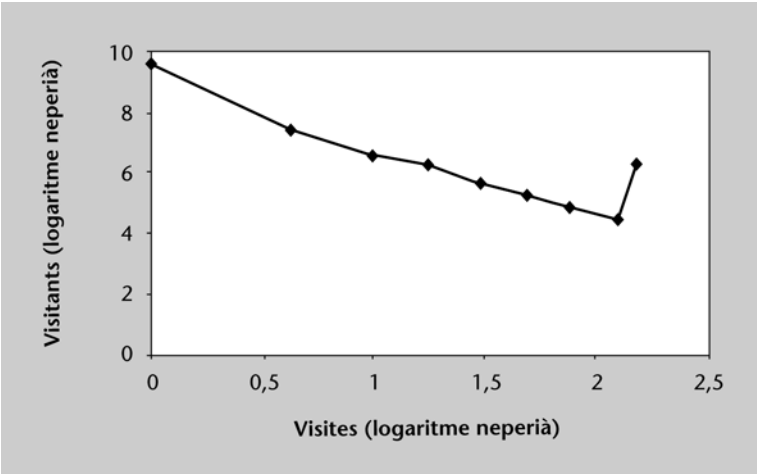
La regressió lineal aporta dos valors, A i B, que permeten calcular k i a:

$$k = e^B \quad \text{i} \quad a = -A$$

Hi ha una relació lineal de la forma $I = B + A \cdot X$ entre el logaritme neperià de la freqüència del nombre de visites i el logaritme neperià del nombre de visitants?

Efectivament, en la representació gràfica veiem que els punts de la corba estan alineats (gràfica 38).

Gràfica 38. Freqüentació d'un lloc web (representació logarítmica)



La singularitat del primer punt procedeix del fet que hem fet un truncament per als valors superiors a 10.

Càlcul de l'equació de la recta de regressió per a un nombre de visites comprès entre 1 i 10:

$$G_i = \frac{k}{u_i^a}, \text{ amb la qual cosa } \ln(G_i) = \ln(k) - a \cdot \ln(U_i)$$

El càlcul de la regressió lineal ens dóna:

$$\ln(G_i) = 8{,}71 - 1{,}84 \cdot \ln(U_i), \text{ si } i \text{ varia d'1 a 10.}$$

Ara podem calcular els coeficients de la sèrie hiperbòlica:

$$a = 1{,}84 \text{ i } \ln(k) = 8{,}71, \text{ i llavors, } k = e^{8{,}71} = 6.063{,}24$$

La relació hiperbòlica és, doncs, $G_i = \dfrac{6.063{,}24}{u_i^{1{,}84}}$. [1]

Càlcul de l'equació de la recta de regressió per a un nombre de visites comprès entre 1 i 9:

Obtenim que $\ln(G_i) = 9{,}08 - 2{,}25 \cdot \ln(U_i)$, en què i varia d'1 a 9.

En aquest cas, tenim $a = 2{,}25$ i $\ln(k) = 9{,}08$, i llavors, $k = e^{9{,}08} = 8.777{,}97$.

La relació hiperbòlica és, doncs, $G_i = \dfrac{8.777{,}97}{u_i^{9{,}08}}$. [2]

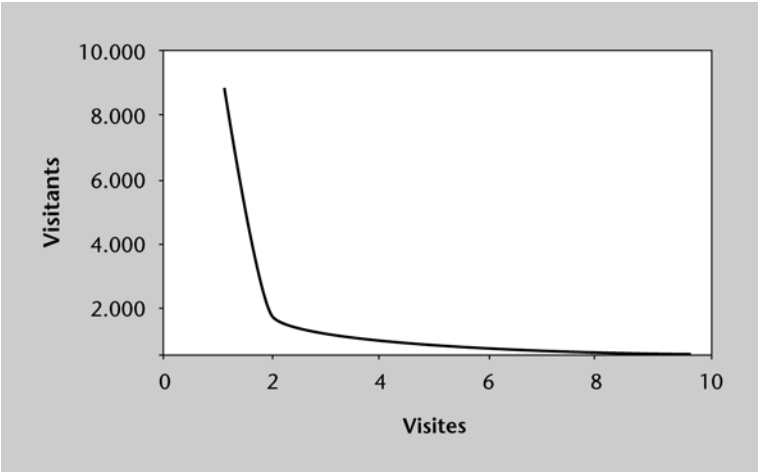
Les relacions [1] i [2] donen dues estimacions teòriques del nombre de visitants (taula 45).

Taula 45. Freqüentacions calculades i observades d'un lloc web

Visites	Visitants (teòric) relació [1]	Visitants (teòric) relació [2]	Visitants (observat)
1	6.063	8.778	12.749
2	1.694	1.845	1.411
3	803	741	574
4	473	388	313
5	314	235	240
6	224	156	151
7	169	110	116
8	132	82	92
9	106	63	74
>10	95		597

Els resultats dels dos càlculs són diferents. Veiem que la relació [2] (gràfica 39) dóna uns resultats millors que la relació [1] per als nombres de visites elevats, però són mediocres per als nombres de visites baixos. Això és perquè no coneixem el desglossament de l'efectiu dels 597 visitants que han fet un mínim de deu visites.

Gràfica 39. Freqüentació d'un lloc web: sèrie hiperbòlica (relació [2])



2) Flux d'informació: llei de Lotka

Aquesta llei, que ja hem comentat en els mòduls 1 i 2, va ser formulada per Alfred Lotka el 1926. Després d'estudiar durant deu anys l'índex acumulat dels autors que apareixien a *Chemical Abstracts*, va constatar una relació simple entre el nombre d'autors d'articles científics i el nombre d'articles que havien publicat.

D'una manera més precisa, la fracció d'autors G que han publicat U articles durant un temps concret i sobre un tema concret es reparteix així:

$$G = \frac{k}{U^2}$$

És a dir, $\frac{k}{U_1^2}$ $\frac{k}{U_2^2}$ ... $\frac{k}{U_i^2}$ considerant la sèrie $U_i = i$ per a $i = 1, 2\ldots$

Treballant sobre les fraccions d'autors, tenim, per al conjunt d'autors:

$$\sum_{i=1}^{\infty} \frac{k}{U_i^2} = \frac{k}{U_1^2} + \frac{k}{U_2^2} + \ldots + \frac{k}{U_i^2} = k \cdot S = 1$$

A partir de l'anterior, sabem que la sèrie $\sum_{i=1}^{\infty} \frac{k}{U_i^a}$ és convergent, ja que $a = 2$.

Ara bé, podem demostrar que $\sum_{i=1}^{\infty} \frac{1}{U_i^2} = \frac{\pi^2}{6}$.

Aquesta igualtat ens permet calcular la constant k:

$$1 = \sum_{i=1}^{\infty} \frac{k}{U_i^2} = k \cdot \sum_{i=1}^{\infty} \frac{1}{U_i^2} = k\frac{\pi^2}{6}, \text{ és a dir, } k = \frac{6}{\pi^2} \approx 0,6$$

Per a una població, deduïm que:

- $\frac{0,6}{1^2} \cdot 100$ = 60% dels investigadors publiquen un article sobre un tema concret durant un període determinat.

- $\frac{0,6}{2^2} \cdot 100$ = 15% dels investigadors publiquen dos articles sobre un tema concret durant un període determinat.

- $\frac{0,6}{3^2} \cdot 100$ = 6% dels investigadors publiquen tres articles sobre un tema concret durant un període determinat i així successivament.

2.3. En resum

Aprendre matemàtiques implica, abans de res, aprendre a explicar. Els recomptes d'informació són molt nombrosos. Ofereixen conjunts de nombres, coneguts com a *sèries*, les sumes dels quals constitueixen en certs casos sèries matemàtiques. Les sèries geomètriques i les sèries hiperbòliques han creat, des del començament, marcs d'anàlisi dels processos d'informació com, per exemple, el de la dispersió de la literatura o el de la circulació dels fluxos d'informació observats durant les fases d'ús, de comunicació i de producció de la informació. En el primer cas parlem de la famosa llei enunciada per Bradford i que porta el seu nom; en l'altre cas, es tracta de la no menys famosa llei de Lotka.

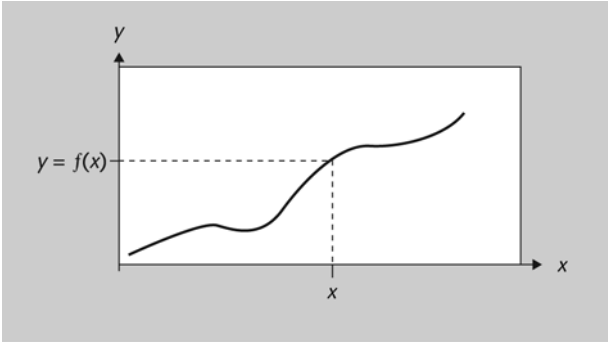
3. Les funcions

El 1749, Leonhard Euler (matemàtic, 1707-1783) parlava, per a definir una funció matemàtica, d'una "quantitat variable que depèn d'una altra quantitat variable". Actualment, preferim dir que la variable y està en funció de la variable x si la dada del valor de x fixa el valor de y. Atès que el valor de y està determinat per complet pel valor de x, escrivim de manera general:

$$y = f(x)$$

$f(x)$ simbolitza una expressió matemàtica en la qual intervé x i pot estar representada per una taula de valors, una corba o simplement una fórmula. Quan és objecte d'una representació gràfica en un sistema de dos eixos Ox i Oy perpendiculars, el punt de coordenades x i y està associat a un valor de x i al valor corresponent de $y = f(x)$. El conjunt dels punts permet dibuixar "la corba" $y = f(x)$, que representa l'evolució del fenomen estudiat (figura 16).

Figura 16. Representació gràfica de la "corba" $y = f(x)$



Quan la funció f es pot expressar a partir de combinacions algebraiques finites de x, parlem d'una funció algebraica. Les funcions algebraiques principals són:

- Les funcions lineals, que s'escriuen de la manera següent:

$$f(x) = a_0 + a_1 x$$

Aquest tipus de funció s'ha usat per a fer regressions lineals.

Podeu consultar informació sobre la regressió lineal en el subapartat 3.2 del mòdul 2 "Estadística de la informació".

- Les funcions polinòmiques, que s'escriuen de la manera següent:

$$f(x) = a_0 + a_1 x + \ldots a_p x^p$$

- Les funcions hiperbòliques, que s'escriuen de la manera següent:

$$f(x) = \frac{a^0}{x^p}$$

en què p és un nombre enter i a_0, a_1, ..., b_0, b_1, ... són qualsevol nombre.

Aquí no presentarem aquestes funcions, ja que s'estudien àmpliament en l'ensenyament secundari. Ens interessarem per unes altres tres funcions no algebraiques que es troben en les ciències de la informació: les funcions exponencial, logarítmica i de potència. Veurem que es poden calcular amb l'ajuda de les sèries (vegeu el subapartat anterior) en les quals cada terme de la successió es defineix per una funció polinomial elemental.

3.1. Augment de la producció científica (cienciometria), obsolescència de la informació (infometria): funció exponencial

3.1.1. Propietats de la funció exponencial

La funció exponencial es coneix algunes vegades com a *funció de creixement natural*, ja que molts processos naturals, com el creixement d'un bosc o d'una població, varien de manera exponencial.

Exemple

Suposem que la població d'una espècie es dobla cada any. Si en el moment inicial (representat com a $t_0 = 0$) la mida de la població és G, serà $2G$ l'any següent (és a dir, en el moment $t_1 = 1$) i $4G$ al cap de dos anys (és a dir, en el moment $t_2 = 2$). La mida de la població $G(t)$ és:

$$G(t) = 2^t G$$

El creixement de la població es coneix com a *exponencial de base 2*.

Observació

La successió de nombres $(G(t_0), G(t_1) \ldots G(t_i) \ldots)$,, si i varia de 0 a l'infinit, és una successió geomètrica de raó 2.

Si t es converteix en una variable contínua, generalitzem aquest tipus de distribució. Per exemple, podem calcular la mida de la població al cap d'un any i mig:

$$G(1{,}5) = 2^{1{,}5} \cdot G$$

Podeu consultar informació sobre la successió geomètrica en el subapartat 2.2 d'aquest mòdul.

Les funcions exponencials de base a, representades com a $x \to y = a^x$, en què a és un nombre estrictament positiu, són definides i contínues en l'interval: $]-\infty, +\infty[$.

Les seves propietats principals són:

$$a^0 = 1$$

Per a qualsevol valor de x,	$a^x > 0$
Per a qualsevol valor de a 0 < a < 1:	a^x és una funció decreixent
Per a qualsevol valor de a a > 1:	a^x és una funció creixent
Per a qualsevol valor de x, i per a qualsevol valor de z, tenim:	$a^{x+z} = a^x \cdot a^z$
Per a qualsevol valor de x, i per a qualsevol valor de z, tenim:	$a^{x-z} = \frac{a^x}{a^z}$
Per a qualsevol valor de x, i per a qualsevol valor de m, tenim:	$(a^x)^m = a^{xm}$

Quan a és igual a $e = 2{,}72828...$, coneguda com a *constant d'Euler*, la funció exponencial es coneix com *de base e*. També es coneix senzillament com a *funció exponencial* i s'escriu:

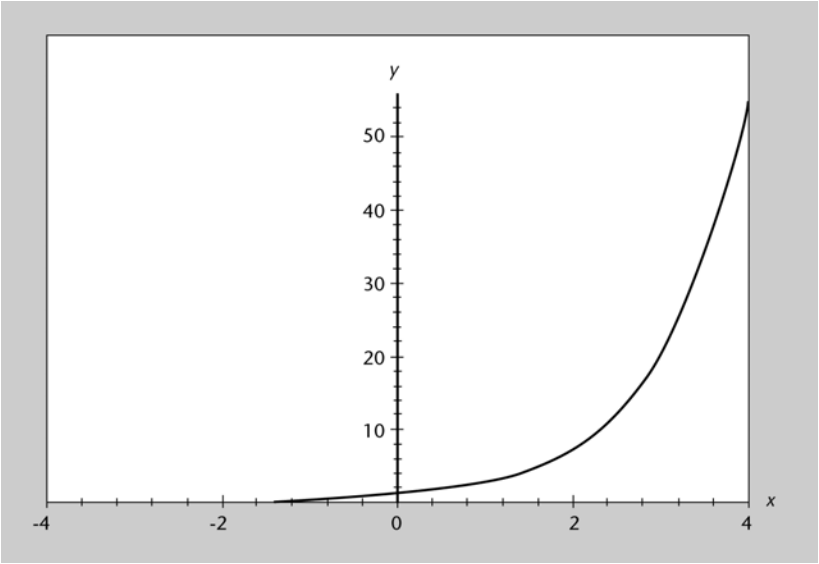
$$\exp(x) = e^x$$

El càlcul i la representació són els següents:

Taula 47. Càlcul de la funció exponencial

Valor de x	exp(x)
–4	0,018
–3	0,050
0	1
1	2,718
2	7,389
3	20,086
4	54,598

Gràfica 40. Gràfica de la funció exponencial $y = \exp(x)$



3.1.2. Creixement de la producció de revistes científiques electròniques a Internet: exponencial creixent

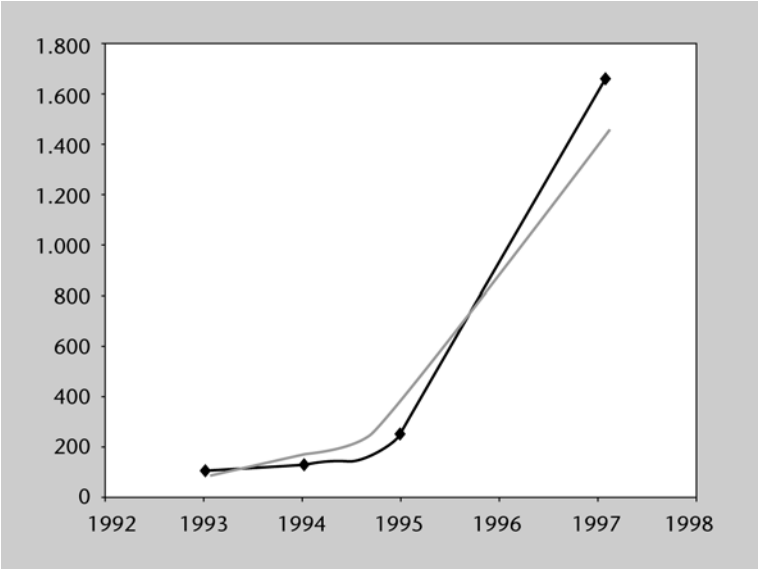
Tenim $C(0)$ com a nombre de revistes científiques electròniques a Internet en el moment $t_0 = 0$. Si hi ha un creixement exponencial, el nombre de revistes en el temps t serà, en conseqüència:

$$C(t) = C(0)e^{at}$$

en què a és un nombre positiu superior a 1.

La representació gràfica és la següent:

Gràfica 41. Creixement de les revistes electròniques



Què representa e^a?

És la taxa de creixement de la població de revistes:

$$e^a = \frac{C(t+1)}{C(t)} \text{ es representa com a } b.$$

No depèn del temps.

Així, doncs, escrivim $C(t) = C(0)\, b^t$, en què b és superior o igual a 1: així, doncs, és una funció creixent.

Obtenim la mateixa expressió que l'obtinguda per al creixement de la població amb $b = 2$ i $C(0) = G$.

Podeu consultar informació sobre les variables vinculades a la variable de temps en el subapartat 2.1.2 del mòdul 2, "Estadística de la informació".

Observació

Igual que per a les sèries cronològiques, l'opció de l'escala de temps és important.

Podem definir la funció exponencial amb la sèrie:

Independentment del valor de x, escriurem:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} = 1 + x + \frac{x^2}{2!} + \ldots \frac{x^n}{n!} + \ldots$$

$$e = e^1 = \sum_{i=0}^{\infty} \frac{1}{i!} = 2{,}72828\ldots.$$

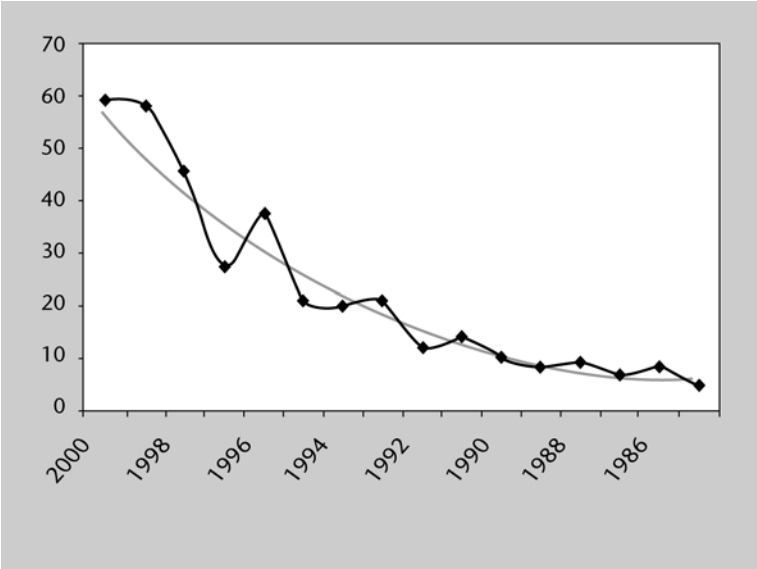
3.1.3. Obsolescència de la informació: exponencial decreixent

Paral·lelament a aquest creixement ràpid del nombre de publicacions, hi ha una obsolescència igualment ràpida de l'estoc d'informacions disponibles. Això significa que si les referències a la literatura del passat es distribueixen de manera aleatòria, sense cap relació amb la data de publicació, la majoria remeten a treballs recents, ja que hi ha més articles disponibles que es poden citar:

$$C(t) = C(0)e^{-at}$$

en què a és un nombre positiu superior a 1.

Gràfica 42. Obsolescència de la informació



Els estudis sobre la vida mitjana de la literatura científica ofereixen elements que permeten aclarir aquest tipus de qüestions.

La vida mitjana de la literatura és el temps en el qual s'ha citat la meitat de la literatura activa. Els estudis d'obsolescència de les diferents literatures han mostrat variacions importants d'aquesta característica: 4,6 anys en física, 7,2 anys en psicologia, 10,5 anys en matemàtiques. De la mateixa manera, si coneixem el nombre total de vegades que s'ha citat una revista, la vida mitjana d'aquesta revista mesura el nombre d'anys en què s'han fet el 50% de les cita-

cions. Com a exemple, a continuació es donen les vides mitjanes per a algunes revistes de ciències de la informació:

Taula 48. Vida mitjana de les revistes de ciències de la informació (any 1999)

Revistes	Vides mitjanes (anys)
J AM SOC INFORM SCI	6,8
SOC STUD SCI	9,6
SCIENTOMETRICS	5,1
INFORM PROCESS MANAG	6,8
J INFORM SCI	6,2

Font: JCR.

3.2. Codificació de la informació (infometria) (teoria de Shannon): funció logarítmica

3.2.1. Propietats de la funció logarítmica

Quan representem fenòmens físics d'una dimensió molt gran o molt petita, sovint usem coordenades logarítmiques en lloc de les coordenades lineals clàssiques. Igual que per als nombres, normalment s'usa la base 10.

La propietat destacable de la funció logarítmica és que transforma un producte de dos nombres en una suma de dos nombres; el logaritme d'un producte de dos nombres positius és igual a la suma dels logaritmes d'aquests dos nombres.

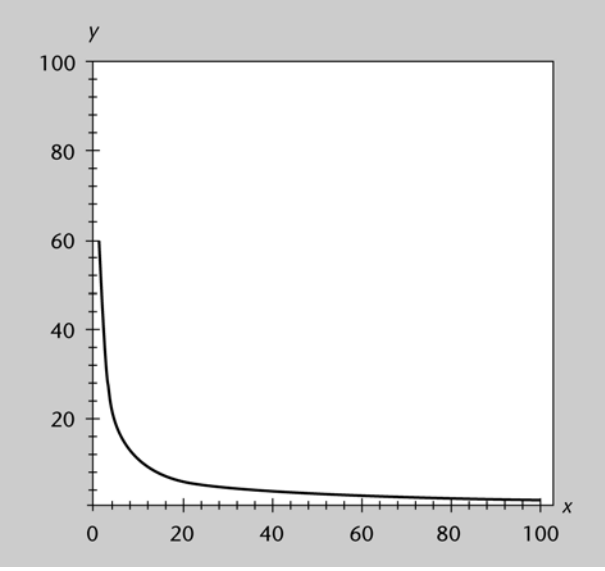
$$\log (A \cdot B) = \log (A) + \log (B)$$

Aquesta propietat ja s'ha usat en els subapartats 3.4.2 i 3.4.3 del mòdul 2 "Estadística de la informació", i 2.1 d'aquest mòdul.

Exemple

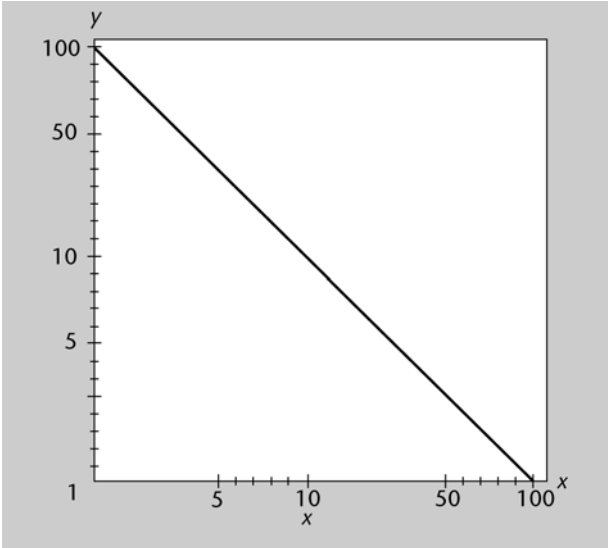
Agafem la funció $g(x) = \frac{1}{x}$ representada en la figura següent:

Gràfica 43. Gràfica de la funció $y = \frac{1}{x}$ en coordenades lineals



Aquesta funció decreixent tendeix a 0. En la gràfica és pràcticament impossible llegir les coordenades dels punts d'abscissa superiors a 20. El canvi de coordenades de lineals a logarítmiques en base 10 (en l'abscissa i en l'ordenada) amplifica els efectes. Les coordenades dels punts de la corba es poden llegir ara molt més fàcilment.

Gràfica 44. Gràfica de la funció $y = \frac{1}{x}$ en coordenades logarítmiques



Vegem ara d'una manera més precisa com es defineixen les funcions logarítmiques.

Les funcions logarítmiques de base a, que es representen com a $x \to y = \log_a(x)$, en què a és un nombre estrictament positiu, són definides i contínues en l'interval $]\,0, \infty[$.

Propietats principals:

$\log_a(1) = 0$.
Si $x < 1$, llavors $\log_a(x) < 0$ i si $x > 1$ llavors $\log_a(x) > 0$.
Si $0 < a < 1$, llavors $\log_a(x)$ és una funció decreixent.
Si $a > 1$, llavors $\log_a(x)$ és una funció creixent.
Independentment dels valors de x i z positius, tenim:

$$\log_a(xz) = \log_a(x) + \log_a(z)$$

Independentment dels valors de x i z positius, tenim:

$$\log_a\left(\frac{x}{z}\right) = \log_a(x) - \log_a(z)$$

Independentment del valor de x positiu i del valor de m, tenim:

$$\log_a(x^m) = m \cdot \log_a(x)$$

Hi ha tres funcions logarítmiques molt usades:

- El logaritme de base 10, conegut també com a *logaritme decimal* i representat per *log*.

- El logaritme de base *e* (en què *e* és la constant d'Euler –vegeu el subapartat anterior), conegut també com a *logaritme neperià* i representat per *ln*.
- El logaritme de base 2 (unitat de mesura del senyal).

La funció logarítmica de base *a*, representada per $\log_a(x)$, és la recíproca (o funció inversa) de l'exponencial de base *a*. En efecte:

$$y = a^x \Leftrightarrow x = \log_a(y) \text{ en què } y > 0$$

i inversament

$$y = \log_a(x) \Leftrightarrow x = a^y \text{ en què } x > 0$$

Entre el logaritme de base *a* i el logaritme neperià tenim les relacions següents:

$$\log_a(x) = \frac{\ln(x)}{\ln(a)}$$

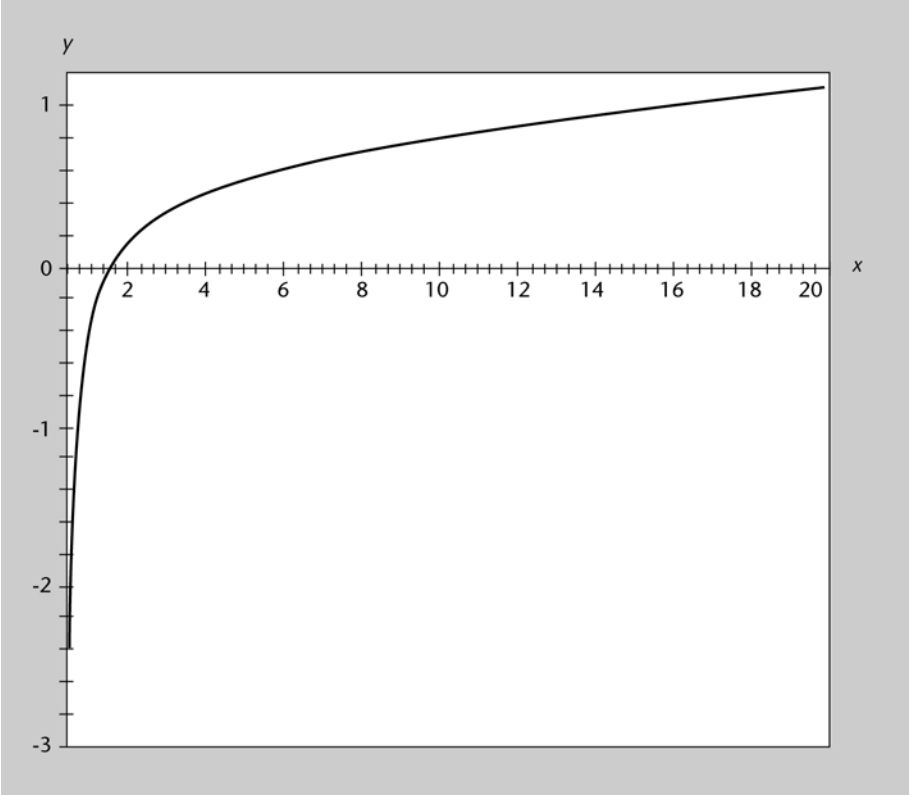
$$a^x = e^{x\ln(a)} \log_a(a) = 1 \text{ i } \ln(e) = 1$$

El càlcul i la representació del logaritme decimal són els següents:

Taula 49. Càlcul del logaritme decimal

Valor de *x*	log (*x*)
0,001	–3
0,01	–2
0,1	–1
1	0
2	0,301
3	0,477
4	0,602
5	0,699
6	0,778
7	0,845
8	0,903
9	0,955
10	1
20	1,301

Gràfica 45. Gràfica del logaritme decimal $y = \log(x)$



Igual que per a la funció exponencial, la funció logarítmica es defineix per la suma de la sèrie.

Independentment del valor de x que verifica que $|x| \leq 1$:

$$\ln(1+x) = \sum_{i=1}^{\infty} \frac{x^i}{i!}(-1)^{i+1} = \underset{n\to\infty}{\text{Límit}}\left(x - \frac{x^2}{2} + \frac{x^3}{3!} + \ldots (-1)^{n+1}\frac{x^n}{n!} \right)$$

3.2.2. Mesura de la quantitat d'informació estadística

La funció logarítmica de base 2, també coneguda com a *teoria matemàtica de la "comunicació" de Claude Shannon*, té un paper important en la teoria de la informació estadística en l'aplicació per a mesurar la probabilitat de transmissió d'un senyal electrònic. També s'ha anomenat moltes vegades, erròniament, *teoria de la informació*. Recordem l'interès d'aquesta teoria amb l'exemple següent.

Exemple

Si busquem un tresor i sabem que està amagat en una de les vuit sales d'un castell (amb les sales numerades del 0 al 7), la informació que busquem és, doncs, el número de la sala. Per a ajudar-nos a trobar-lo, el geni maligne que el custodia accepta respondre preguntes del tipus: "el nombre de la sala és inferior a n (n és un nombre comprès entre 0 i 7)?".

Després de diverses proves, ens adonem que cal plantejar com a mínim tres preguntes per a trobar el resultat. Per exemple:

- Pregunta 1: el número de la sala és inferior a 4? Resposta 1: no
- Pregunta 2: el número de la sala és inferior a 6? Resposta 2: no
- Pregunta 3: el número de la sala és inferior a 7? Resposta 3: no

La sala en la qual es troba el tresor és, doncs, la sala 6. És evident que hi ha altres seqüències de tres preguntes que ens poden permetre trobar la sala.

Suposem ara que hem numerat les habitacions amb les xifres equivalents escrites en codi binari, amb la qual cosa tindrem:

- Sala 0: 000; sala 1: 001; sala 2: 010; sala 3: 011;
- Sala 4: 100; sala 5: 101; sala 6: 110; sala 7: 111;

Podem identificar la sala plantejant les tres preguntes següents:

- Pregunta 1: la primera xifra del número de la sala és inferior a 1?
- Pregunta 2: la segona xifra del número de la sala és inferior a 1?
- Pregunta 3: la tercera xifra del número de la sala és inferior a 1?

No hi ha cap altra seqüència de preguntes possible.

Així, doncs, podem identificar una sala responent tres vegades seguides a pregunta 0 o 1, que no és res més que la pregunta anterior, a la qual hem respost amb sí o no.

En el cas d'un conjunt E de k elements, en què $k = 2^n$ i n és un nombre enter, veiem que localització d'un element qualsevol de E necessita n unitats d'informació estadística elemental. Diem que n mesura la quantitat d'informació estadística necessària, que representem per $I(E)$:

$$I(E) = \log_2(k)$$

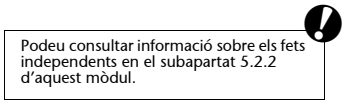
Verifiquem que:

$$I(E) = \log_2(2^n) = n \cdot \log_2(2) = n$$

D'una manera més general, la mesura de la quantitat d'informació en el sentit estadístic es basa en el postulat següent:

"Com més improbable és un fet, més gran és la seva quantitat d'informació estadística."

Si $P(A)$ designa la probabilitat de realització del fet A, llavors $I(A)$ és la quantitat d'informació estadística del fet A. Si A i B són dos fets independents, es dedueix que la quantitat d'informació estadística del fet A i del fet B és la suma de la quantitat d'informació estadística de A i de B:



$$I(A \cap B) = -\log\big(P(A \cap B)\big) = -\log\big(P(A) \cdot P(B)\big) = -\big(\log(A) + \log(B)\big) = I(A) + I(B)$$

En el cas dels senyals elèctrics, Shannon mesura aquesta quantitat amb la fórmula:

$$I(A) = -\log_2(P(A))$$

La base 2 del logaritme es justifica perquè la unitat usada per a mesurar aquesta quantitat d'informació estadística en el cas dels senyals elèctrics (mesura del senyal) és el bit, és a dir, un valor binari 0 o 1. A més, si suposem que la quantitat d'informació estadística es caracteritza per un nombre positiu, el signe menys és necessari perquè $P(A)$ és inferior a 1 i, en conseqüència, $\log_2(P(A))$ és negatiu.

Podeu consultar informació sobre el valor binari en el subapartat 3.2.1 del mòdul 1, "La mesura de la informació".

Exemple

Tornem a l'exemple del principi. El conjunt E està format per vuit sales, amb la qual cosa la probabilitat de trobar la sala és $\frac{1}{8}$. Llavors, $I(A) = -\log_2\left(\frac{1}{8}\right)$. Ara bé, $8 = 2^3$, per la qual cosa tenim:

Podeu consultar informació sobre l'estadística probabilista en l'apartat 5 del mòdul2, "Estadística de la informació".

$$I(A) = -\log_2\left(\frac{1}{2^3}\right) = 3 \cdot \log_2(2) = 3$$

La quantitat d'informació estadística necessària per a resoldre el problema de quin és el número de la sala és, doncs, tres; ens tornem a trobar amb el nombre mínim de preguntes elementals que cal plantejar.

Aquesta mesura de la informació estadística ha portat a la definició d'una funció coneguda com a *entropia*, que s'ha agafat de la termodinàmica, que mesura la imprevisibilitat mitjana d'un missatge emès. Aquesta funció d'entropia també es pot usar com a índex de diversitat per a caracteritzar les distribucions.

3.3. Freqüència de les paraules en un text (infometria) (llei de Zipf): funció de potència

3.3.1. Propietats de la funció de potència

Les funcions polinòmiques simples són molt conegudes:

$$y = x^m$$

en què l'exponent m és un nombre enter positiu o negatiu.

x^m significa que tenim:

- m vegades el producte de x si m és un enter positiu; és la funció de potència.
- m vegades la inversa d'aquest producte si m és un enter negatiu; és la funció hiperbòlica. Independentment del valor de m com a nombre enter positiu, tenim:

$$y = x^{-m} = \frac{1}{x^m}$$

En les ciències de la informació s'acostuma a anomenar *funció hiperbòlica* qualsevol funció de potència que tingui un exponent negatiu, tant si és enter com si no. En efecte, el que caracteritza diversos fenòmens de la informació són els comportaments de naturalesa hiperbòlica, és a dir, que el producte de potències fixes de les variables és constant:

$$F(x) \cdot x^n = \text{constant}$$

En les seves manifestacions discretes, això es reflecteix en el fet que a una causa que creix de manera geomètrica correspon un efecte que creix de manera aritmètica.

Podem generalitzar aquest tipus de funció quan m (ara representat per a) és un nombre qualsevol i x un nombre positiu. En aquests casos, es parla de funcions de potència representades així:

$$y = x^a = e^{a \log (x)}$$

en què a és un nombre qualsevol.

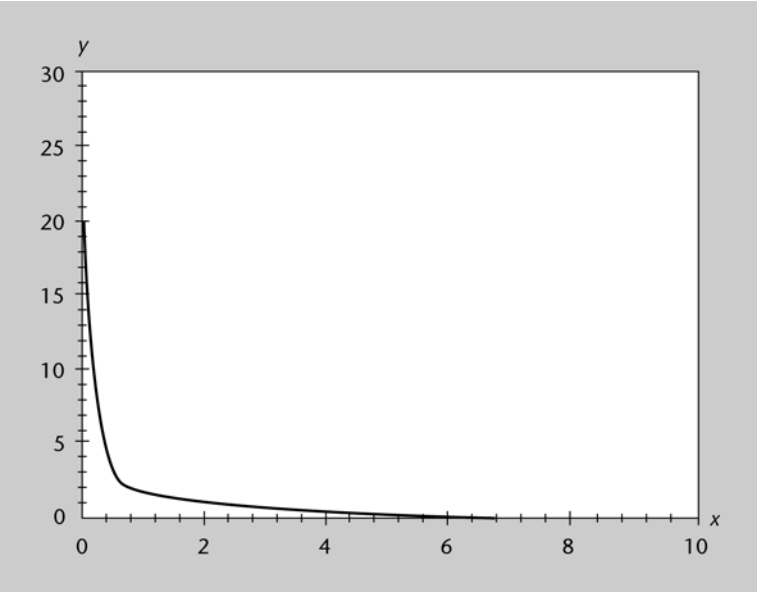
Podeu consultar informació sobreles distribucions hiperbòliques en el subapartat 2.2 d'aquest mòdul.

El càlcul i la representació de la funció de potència d'exponent $a = -1{,}5$ són els següents:

Taula 50. Càlcul de la funció de potència $y = \frac{1}{x^{1,5}}$

Valor de *x*	$y = \frac{1}{x^{1,5}}$
0,1	31,623
0,2	11,180
1	1
2	0,354
3	0,192
4	0,125
5	0,089
6	0,068
7	0,054
8	0,044
9	0,037
10	0,032

Gràfica 46. Gràfica de la funció de potència $y = \frac{1}{x^{1,5}}$



Les propietats següents són una conseqüència de les propietats de les funcions logarítmiques i exponencials.

Les funcions de potències, representades com $x \to x^{\alpha} = e^{\alpha \log(x)}$ en què a és un nombre qualsevol, són definides i contínues en l'interval:

$$]0, +\infty[.$$

Propietats principals:

Independentment del valor de x positiu, x^a es una funció positiva.
Si $a > 0$, x^a és una funció creixent.
Si $a < 0$, x^a és una funció decreixent.

Independentment dels valors de x i z positius, tenim:

$$x^a \cdot z^a = (xz)^a$$

Independentment del valor de x positiu i per a qualsevol valor de m, tenim:

$$\frac{x^a}{z^a} = \left(\frac{x}{z}\right)^a$$

Independentment del valor de x positiu i per a qualsevol valor de m, tenim:

$$(x^a)^m = x^{am}$$

3.3.2. Ley de Zipf

El nombre d'ocurrències de tot objecte en un conjunt, com per exemple un llibre d'una col·lecció o una paraula d'un text, obtingut per recompte es coneix –tal com ja hem vist– com a *freqüència*. Si ordenem els objectes en funció de la seva freqüència decreixent, els podem atribuir un rang. Diversos objectes que tinguin la mateixa freqüència tindran uns números d'ordre consecutius. Les propietats de les corbes (rang, freqüència) s'han observat i estudiat en camps molt variats. En els anys cinquanta, George Zipf es va interessar per la freqüència de les paraules en els textos. Va observar una relació constant, de tipus hiperbòlic, entre la freqüència i el rang de les paraules:

Bibliografia

G. K. Zipf (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley. [Reimpressió (1965). Nova York: Hafner].

Podeu consultar informació sobre les propietats de les corbes en el subapartat 2.4 del mòdul 2, "Estadística de la informació".

$$rang \cdot freqüència = constant \ (representat\ per\ k)$$

Tornem a trobar una funció de naturalesa hiperbòlica amb l'exponent m igual a 1:

$$U(r) = \frac{k}{r}$$

En què U representa la freqüència i r el rang.

En realitat, la relació entre rang i freqüència és del tipus de potència inversa d'exponent $b \geq 0$:

$$U(r) = \frac{k}{r^b}$$

Exemple

Recompte de paraules de la versió intermèdia (febrer de 2001) de la part "curs" del llibre original *Éléments de estatistique et de mathématique de l'information*.

El total de paraules d'aquest text és 27.938 i el nombre de paraules diferents és 3.622. A continuació trobarem les cinquanta primeres paraules classificades, d'una banda per ordre de freqüència decreixent i, de l'altra, per ordre alfabètic. El 49% de les paraules tenen com a freqüència 1 (en lexicometria, aquestes paraules de freqüència 1 es coneixen com a *hàpax*, la seva proporció en els textos és relativament constant i es troba entorn del 50%).

Taula 51. Les cinquanta primeres paraules classificades per ordre de freqüència decreixent i per ordre alfabètic

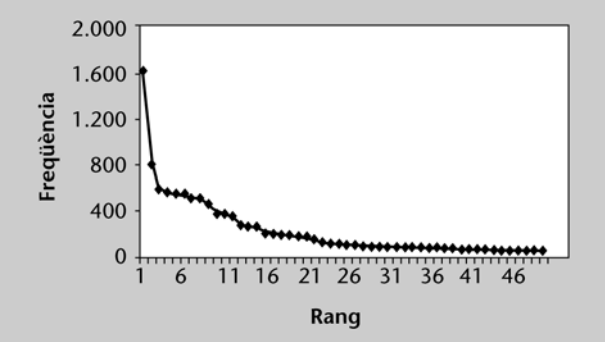
Classificació per freqüència creixent		**Classificació per ordre alfabètic**	
paraula	**freqüència**	**paraula**	**freqüència**
de	1.588	a	193
la	796	A	72
des	589	à	385
d	561	abondamment	1
1	551	abonnement	3
est	550	abonnements	3

Classificació per freqüència creixent		Classificació per ordre alfabètic	
paraula	**freqüència**	**paraula**	**freqüència**
les	518	abord	5
et	517	abordée	1
le	465	abordons	1
une	389	abscisse	2
à	385	absence	3
un	364	absolue	1
en	285	absolues	1
que	277	absolus	1
on	273	Abstracts	1
par	219	abstraits	1
dans	213	Academy	1
nombre	210	accélération	1
qui	210	accepte	2
sont	194	accepter	1
a	193	accompagnant	1
du	167	accordent	1
deux	141	Accoupler	1
pour	132	accroît	1
La	128	accru	1
ou	123	achats	1
ce	116	acheteur	1
plus	111	acquisition	1
On	109	actes	1
sur	107	actif	1
entre	104	actifs	1
qu	104	action	2
pas	100	activité	3
n	98	activités	2
loi	96	actualiser	1
il	95	Actuellement	1
mesure	93	actuellement	1
information	91	adaptée	1
Les	89	adaptées	1
nous	86	adaptés	1
Le	85	additifs	1
moyenne	84	additionnent	1
au	83	additive	1

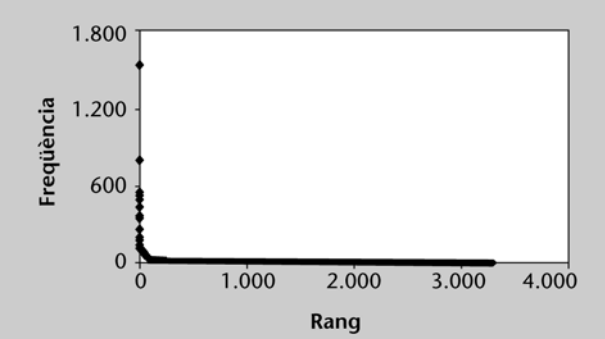
Classificació per freqüència creixent		Classificació per ordre alfabètic	
paraula	**freqüència**	**paraula**	**freqüència**
ces	81	additivité	2
type	74	admet	1
peut	73	administration	1
A	72	administrée	1
ont	72	adopte	1
paragraphe	70	adopter	1
ensemble	69	advantage	1

Els traçats de la relació (rang, freqüència) per a les cinquanta primeres paraules i per a les 3.622 paraules del text, són els que apareixen en les gràfiques següents:

Gràfica 47. Corba (rang, freqüència) de les cinquanta primeres paraules



Gràfica 48. Corba (rang, freqüència) de les 3.622 paraules



Per a determinar la constant k i l'exponent b característics del vocabulari del text que estem estudiant, fem un ajust amb una tècnica de regressió lineal.

D'aquesta manera transformem la relació hiperbòlica en una relació lineal:

$$u = \frac{k}{r^b} \Leftrightarrow \ln(u) = \ln(k) - b\ln(r)$$

El càlcul de la regressió dóna k = 6.308,3 i b = 1,09.

3.3.3. Equivalència entre la llei de Zipf i la llei de Lotka

La llei de Lotka i la llei de Zipf són equivalents si a = 1 i b = 1

$$U = \frac{k_1}{r} \Leftrightarrow G = \frac{k_2}{U^2}$$

en què k_1 i k_2 són constants.

Això significa que, en termes de probabilitat, la llei de Zipf es confirma únicament si la funció relativa a la freqüència d'aparició de les paraules és de la forma $\frac{k}{x^2}$.

Si *N* designa el nombre de paraules (o el nombre d'autors) suposadament finit, podem escriure:

$$G(u) = \int_u^{\infty} N \cdot \frac{k_2}{x^2} dx$$

$\frac{k_2}{x^2}$ és la distribució (funció de densitat) de les paraules (dels autors).

G(*u*) representa exactament el rang de les paraules (o dels autors) que tenen exactament *u* com a nombre d'ocurrències.

$$G(u) = \int_u^{\infty} N \cdot \frac{k_2}{x^2} dx = N \cdot \left[-\frac{k_2}{x} \right]_u^{\infty} = \frac{k_1}{u}$$

La llei de Lotka i la llei de Zipf només són vàlides de manera aproximada. Encara que la llei de Zipf no té cap aplicació significativa en lingüística, sí que és una realitat incontestable en les ciències de la informació. De manera més general, es coneix com a funció zipfiana la funció de potència definida per la relació:

Podeu consultar informació sobre les distribucions en ciències de la informació en l'apartat 2 del mòdul 2, "Estadística de la informació".

$$y = \frac{k}{x^{1+a}}$$

en què *k* i *a* són constants positives.

3.4. En resum

Més que la correlació, és la funció la que autentifica una regularitat matemàtica entre dues variables: diu que l'una depèn de l'altra. Posar de manifest una regularitat, és a dir, una relació quantitativa constant, és l'esperança que acaricia tot quantitivista. En les ciències de la informació, les grans funcions matemàtiques són la funció exponencial, la funció logarítmica i la funció de potència. La funció exponencial descriu l'"augment natural" de la quantitat d'informació o la "disminució natural" de la seva actualitat. La funció logarítmica, en la seva versió de base 2, té un paper molt important tant en la teoria matemàtica de la transmissió dels senyals elèctrics de Shannon com en la mesura estadística de la improbabilitat d'un fet. Finalment, la funció de potència (coneguda més habitualment com a *funció hiperbòlica*) permet una bona mesura de la freqüència d'aparició de les paraules en un text.

4. Les equacions

Una equació és una igualtat entre magnituds conegudes i desconegudes que només és vàlida per a certs valors de les magnituds desconegudes, anomenades *solucions de l'equació*. La resolució de les equacions, que són els objectes principals de l'àlgebra i de l'anàlisi, té com a objectiu la cerca d'aquest valor numèric en el cas de les equacions d'una incògnita, o diversos valors numèrics en el cas de les equacions de diverses incògnites.

Terminologia

En matemàtiques usarem el terme *arrel* més que *solució*.

En el cas d'una equació algebraica de primer grau d'una incògnita, quan totes les magnituds amb conegudes, ens trobem davant una equació de la forma:

$$a \cdot x = -b$$

Que s'escriu més senzillament:

$$a \cdot x + b = 0$$

Aquesta equació només té una solució:

$$x = -\frac{b}{a} \text{ si } a \text{ és diferent de } 0.$$

Quan l'equació fa intervenir una altra magnitud i les seves derivades successives amb la magnitud depenent (vegeu l'annex 4), l'equació es coneix com a *diferencial*. Les diferencials es representen per *dy* i *dx*.

Podeu obtenir informació sobre les equacions diferencials a l'annex 4, "Derivada i integral, exemple de resolució d'equacions diferencials".

Exemple

Equació del temps de reacció per a triar una funció en un menú en una pantalla tàctil:

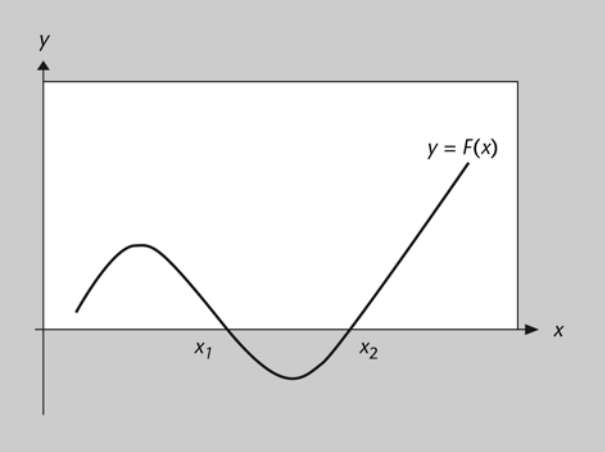
$$t = c + k \cdot \log_2(b)$$

en què
t és el temps en segons,
b és el nombre d'alternatives,
c i *k* són constants característiques de l'usuari.

4.1. Preu de les visites (museometria): equació algebraica

Hi ha equacions algebraiques de diversos tipus. Una equació d'una incògnita es defineix per la igualtat $F(x) = 0$, en què F és una funció. Resoldre aquesta equació consisteix a buscar els valors de x que compleixen la igualtat anterior. En la figura següent, x_1 i x_2 són les solucions de l'equació $F(x) = 0$.

Figura 17. Solucions de l'equació $F(x) = 0$



Una equació amb diverses incògnites es defineix per la igualtat $F(x,y) = 0$, en què F és una funció de x i de y.

Les equacions més conegudes són les equacions algebraiques en les quals F és una funció algebraica.

Podeu consultar informació sobre la funció algebraica en l'apartat 3 d'aquest mòdul.

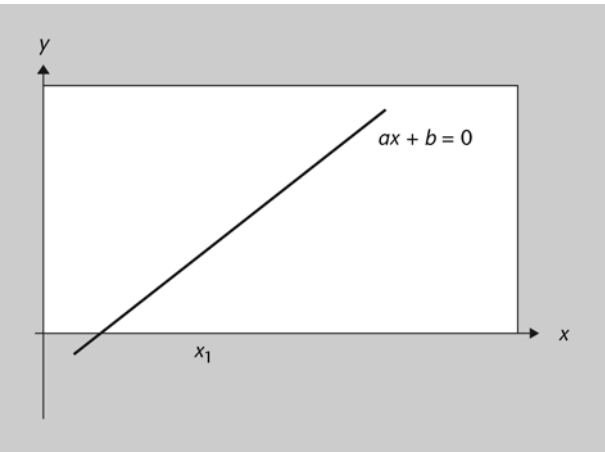
4.1.1. Equació de primer grau amb una incògnita

Una equació de primer grau amb una incògnita x s'escriu, tal com ja hem vist:

$$a \cdot x + b = 0$$

En què a i b són unes constants qualssevol. L'única solució que admet és $x = -\frac{b}{a}$ si a $\neq 0$.

Figura 18. Solucions de l'equació $ax + b = 0$



Exemple

Un museu vol establir el preu de les visites en grup (cinc persones per grup com a màxim). L'administrador decideix que el preu de la visita en grup y ha de ser dues vegades més car que el preu de les visites individuals x.

L'equació corresponent és:

$$y = 2 \cdot x, \text{ és a dir, } y - 2 \cdot x = 0$$

Si fixem el preu de les visites individuals a $x = 5$ euros, llavors haurem de demanar $y = 10$ euros per a una visita en grup.

4.1.2. Equació de segon grau amb una incògnita

La incògnita és sempre x, que apareix elevada al quadrat. Així, doncs, l'equació corresponent és:

$$a \cdot x^2 + b \cdot x + c = 0$$

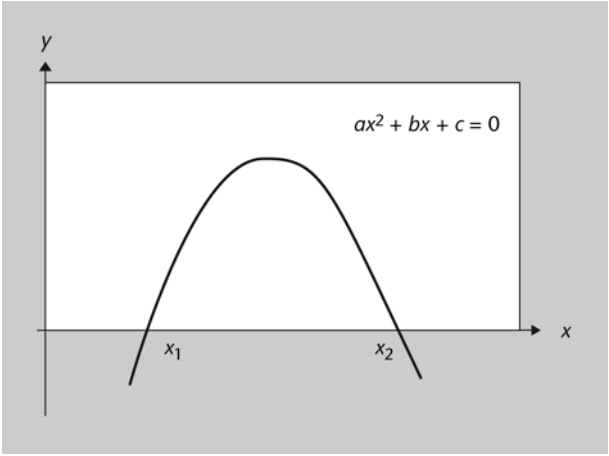
en què a, b i c són unes constants qualssevol.

Si $b^2 - 4 \cdot ac \geq 0$, aquesta equació admet dues solucions.

Si a és diferent de 0, tenim:

$$x_1 = \frac{-b + \sqrt{b^2 - 4 \cdot ac}}{2 \cdot a} \text{ i } x_2 = \frac{-b - \sqrt{b^2 - 4 \cdot ac}}{2 \cdot a}$$

Figura 19. Solucions de l'equació $ax^2 + bx + c = 0$



4.1.3. Equació de primer grau amb dues incògnites

Una equació de primer grau amb dues incògnites s'escriu:

$$a \cdot x + b \cdot y + c = 0$$

en què a, b i c són unes constants qualssevol. Aquesta equació pot admetre diverses solucions.

Per a determinar els valors únics de x i y que verifiquen una situació concreta és necessari tenir dues equacions. El sistema següent és un sistema de dues equacions de primer grau amb dues incògnites, x i y:

$$\begin{cases} a \cdot x + b \cdot y + c = 0 \\ d \cdot x + e \cdot y + f = 0 \end{cases}$$

Si $a \cdot e - b \cdot d \neq 0$, aquest sistema admet dues solucions:

$$\begin{cases} x = \dfrac{b \cdot f - c \cdot e}{a \cdot e - b \cdot d} \\ y = \dfrac{d \cdot c - a \cdot f}{a \cdot e - b \cdot d} \end{cases}$$

Exemple

En l'exemple anterior, per a fixar el preu de les entrades l'administrador ha de tenir en compte una condició suplementària, que és que les entrades han d'aportar 5.000 euros a l'any perquè el museu pugui tenir una gestió equilibrada. Durant aquests cinc últims anys hi ha hagut, de mitjana, 500 visites de grups i 400 visites individuals a l'any. A més de l'equació $y = 2x$, els valors x i y han de complir l'equació $400 \cdot x + 500 \cdot y = 5.000$.

El sistema és, doncs:

$$\begin{cases} 400 \cdot x + 500 \cdot y = 5.000 \\ y = 2 \cdot x \end{cases}$$

La solució és:

$$x = \frac{5.000}{400 + 1.000} = 3,57 \quad \text{i } y = 2 \cdot 3,57 = 7,14$$

Per tant, l'administrador decideix que el preu per a un visitant individual serà de 3,6 euros i que el d'un grup serà de 7,15 euros.

En molts casos, les solucions de les equacions no es poden expressar en forma algebraica i llavors la resolució es fa usant el càlcul diferencial, com és el cas del mètode de resolució següent, conegut com a *mètode de Newton*:

F és una funció que es pot derivar dues vegades i continua en l'interval $[a,b]$, de manera que $\dfrac{dF}{dx}(x) \neq 0$ per a qualsevol valor de x dins de l'interval $[a,b]$. Si hi ha una solució s de l'equació $F(x) = 0$ dins de $[a,b]$, llavors hi ha una successió x_{i+1} de primer terme x_0 dins de $[a,b]$ que convergeix cap a la solució s i definida per:

$$x_{i+1} = x_i - \frac{F(x_i)}{\dfrac{dF}{dx}(x_i)}$$

Podeu consultar informació sobre el mètode de Newton en l'annex 4 "Derivada i integral, exemple de resolució d'equacions diferencials".

Per tant, com més gran sigui el nombre d'iteracions i en el càlcul x_{i+1}, amb més precisió es coneixerà la solució de l'equació.

Aquest és el tipus de mètode iteratiu que usen els programes informàtics de matemàtiques per a resoldre les equacions.

Podeu consultar informació sobre la precisió i el límit en matemàtiques en l'apartat 1 d'aquest mòdul.

4.2. Rumors i comunicació de les informacions (mediametria): equació diferencial

En l'estudi dels comportaments socials, els processos de comunicació d'informacions ocupen un lloc important. S'han estudiat les comunicacions d'idees, informacions, rumors, opinions, tècniques, etc.

Exemple

Actualment corren per Internet els rumors següents:

- A Internet es pot trobar tot.
- La informació que es troba a Internet és fiable.
- És impossible trobar informació a Internet.

S'han demostrat els papers diferents que tenen els mitjans de comunicació en el moment d'adoptar una innovació. La comunicació interpersonal (coneguda com a *contagiosa*) té un paper d'influència, mentre que la comunicació dels mitjans de comunicació (coneguda com a *irradiant* o de *font constant*) té un paper més informatiu.

Les dinàmiques de la comunicació estan governades per equacions diferencials que s'expressen de manera determinista o bé de manera estocàstica.

4.2.1. El model determinista de la comunicació mediatizada

Si G és la mida de la població que es considera, quin és, en el moment t, el nombre de persones $g(t)$ que, gràcies a un mitjà de comunicació, han adoptat una informació concreta?

En l'interval de temps Δt (conegut com a *increment*), el nombre de persones noves que hauran adquirit la informació (conegudes com les A) serà $g(t + \Delta t) - g(t)$. Aquest nombre és proporcional al següent:

- $G - g(t)$, el nombre de persones a les quals encara cal informar en el moment t dins de la població (conegudes com les *no-A*).
- Δt, l'interval de temps.
- α, el coeficient relatiu de comunicació mediatizada.

Tots els models de comunicació social parteixen de la hipòtesi d'una barreja perfecta de la població en la qual es difon la innovació, és a dir, que

cada individu rep la informació, la innovació, de la mateixa manera que tots els altres. Això comporta que α sigui constant. Així, doncs:

$g(t+\Delta t) - g(t) = \alpha \cdot \Delta t \cdot (G - g(t))$, coneguda com a *diferencial de G*, és a dir, $\frac{g(t+\Delta t) - g(t)}{\Delta t} = \alpha \cdot \left(G - g(t)\right) \cdot$

Per a intervals de temps cada cop més petits (és a dir, $\Delta t \to 0$), passem dels increments a la derivada, és a dir, a l'equació diferencial següent:

$$\frac{dg(t)}{dt} = \alpha \cdot \left(G - g(t)\right)$$

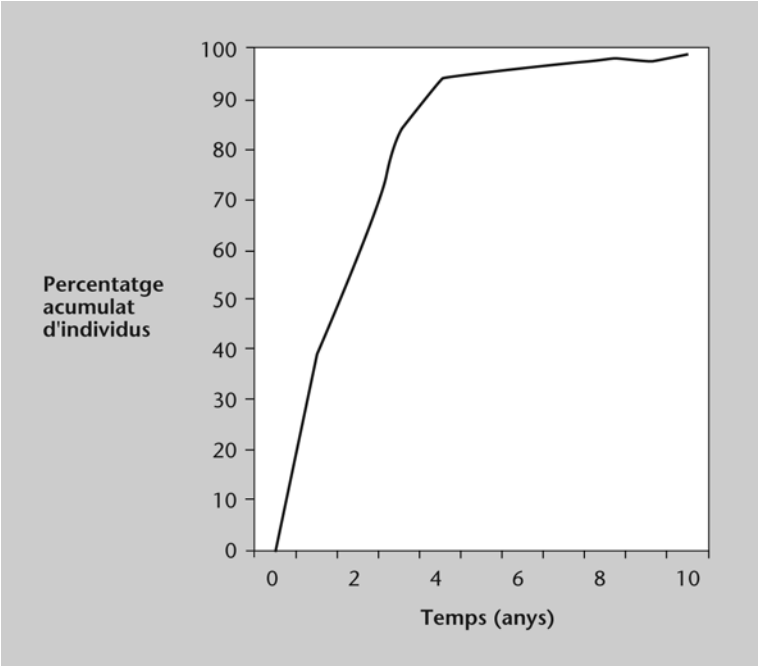
La resolució dóna el resultat següent:

$$g(t) = -G \cdot e^{-\alpha t} + G$$

Podeu consultar informació sobre la derivada i la integral en l'annex 4 "Derivada i integral, exemple de resolució d'equacions diferencials".

Gràficament, això es reflecteix amb una corba exponencial amb una forma característica de *J* invertida, la qual cosa fa aparèixer un fenomen de saturació; és a dir, que la velocitat de comunicació decreix de manera contínua amb el temps (figura 20).

Figura 20. Comunicació mediatizada (irradiació)



4.2.2. El model determinista de la comunicació interpersonal

De la mateixa manera, és possible conèixer el nombre de persones $g(t)$ que, comunicant-se entre elles, han adoptat una informació nova en el moment *t*.

Durant l'espai de temps Δt, el nombre de persones que, pel fet de participar en un procés de comunicació interpersonal, hauran adquirit aquesta informació serà:

$$g(t + \Delta t) - g(t)$$

Aquest nombre és proporcional al següent:

- $G - g(t)$, el nombre de persones a les quals encara cal informar en el moment t dins de la població.
- $g(t)$, el nombre de persones que coneixen la informació en el moment t i que són susceptibles de difondre-la.
- Δt, l'interval de temps.
- β, el coeficient de comunicació interpersonal.

Igual que en el cas anterior, aquí partim de la hipòtesi d'una barreja perfecta, és a dir, que cada individu difon la informació de la mateixa manera que els altres. Així, doncs, β es considera com a constant. Llavors:

$$g(t + \Delta t) - g(t) = \beta \cdot \Delta t \cdot g(t) \cdot (G - g(t))$$

i

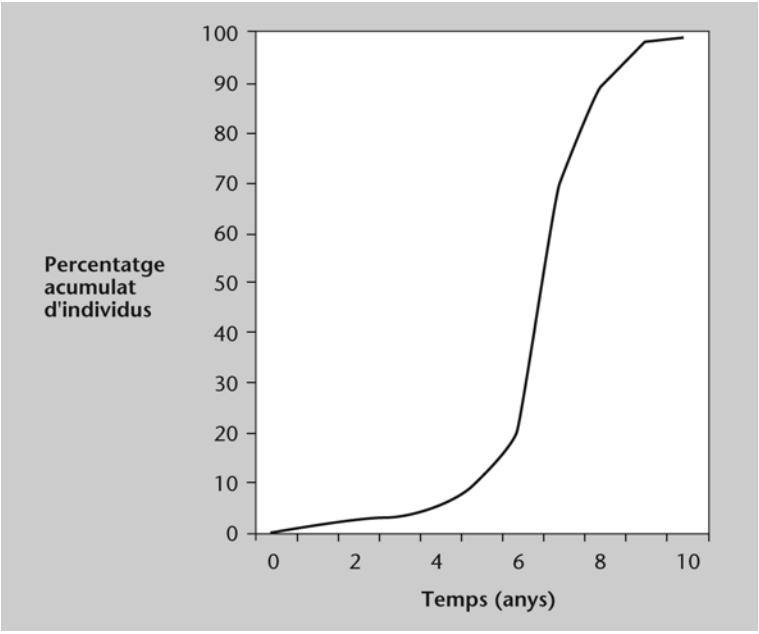
$$\frac{dg(t)}{dt} = \beta \cdot g(t) \cdot \left(G - g(t)\right)$$

La resolució d'aquesta equació dóna el resultat següent:

$$g(t) = \frac{G}{1 + (G-1) \cdot e^{-\beta G t}}$$

La seva representació gràfica és una corba en *S* o corba de creixement logístic (figura 21):

Figura 21. Comunicació interpersonal (contagi)



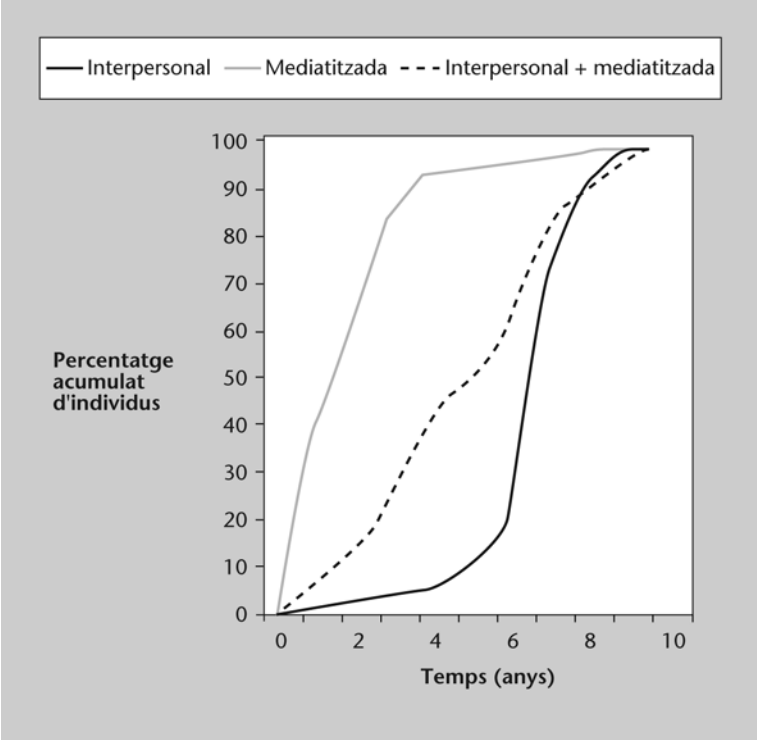
4.2.3. El model determinista de la comunicació

Els dos processos de comunicació sempre estan actius, però d'una manera més o menys variable, per la qual cosa el procés global de comunicació de la informació es desenvoluparà en el temps seguint una dinàmica intermèdia. L'equació diferencial és, doncs:

$$\frac{dg}{dt} = \alpha(G-g) + \beta \cdot g(G-g)$$

Gràficament, la corba resultant també es troba entre les dues corbes anteriors:

Figura 22. Dinàmica de la comunicació de la informació



Observació

En el moment inicial ($t = 0$), suposem que:

- Per a la comunicació mediatitzada, ningú no coneix la informació.
- Per a la comunicació interpersonal, com a mínim una persona coneix la informació.

Però, en realitat, la barreja perfecta no existeix. No hi ha igualtat entre els que tenen (els A) i els que no tenen (els no-A). Les estructures socials són incompletes: "Els amics dels meus amics també són els meus amics". A causa d'aquestes interaccions privilegiades, la comunicació tindrà lloc dins d'una població reduïda.

4.3. En resum

A la cerca de la bella incògnita: la solució de l'equació que descriu la igualtat entre magnituds conegudes i magnituds desconegudes o bé la igualtat entre les seves derivades successives. En el primer cas parlem de les equacions algebraiques d'una o diverses incògnites. Trobarem moltes d'aquestes equacions, la resolució de les quals és més complexa a mesura que creix el nombre d'incògnites. En el segon cas descobrim les interessants propietats de les equacions

diferencials quan es tracta de seguir en el temps l'avenir de les informacions. Aquestes descriuen, en particular i amb gran bellesa, els processos de comunicació, que pot ser comunicació interpersonal de tipus contagiós i comunicació mediatizada de tipus irradiant.

5. Els conjunts

Georg Cantor (matemàtic, 1845-1918) va ser el primer que va definir la noció de conjunt com "qualsevol col·lecció d'objectes ben diferenciats de la nostra percepció o la nostra ment", i va elaborar els primers elements de la teoria de conjunts. Un conjunt és, doncs, una col·lecció d'elements, en nombre finit o infinit, susceptibles de tenir certes propietats i de tenir certes relacions entre ells o amb altres elements d'altres conjunts. En el camp de la informació, les col·leccions, diverses i nombroses, són conjunts fetitxes, objectes dignes de conservar-se. La teoria de conjunts i la lògica (aquí parlem de lògica formal o matemàtica) estan estretament relacionades.

5.1. Localització de la informació (infometria): lògica clàssica booleana

La lògica clàssica booleana, el nom de la qual procedeix del matemàtic George Boole (1815-1864) (també coneguda com a *lògica matemàtica*), identifica, en els conjunts finits, tres relacions de dependència gràcies als operadors booleans Y, O i NO. Aquests tres operadors permeten fer les operacions amb conjunts més importants, que són, respectivament, la intersecció, la unió i la diferència.

Y	(producte lògic) enllaça els components d'una frase.
O	(suma lògica) enllaça els termes sinònims o gairebé sinònims.
NO	(negació lògica) elimina els termes.

Si tenim un conjunt E, direm que A és una part (o un subconjunt) de E si tots els elements de A pertanyen a E. Considerem tots els subconjunts de E i anomenem $\mathcal{P}(\mathrm{E})$ aquest nou conjunt, també conegut com a *conjunt de les parts de E*. Tenim l'equivalència en què A és una part de E, és a dir:

$$A \subset E \Leftrightarrow A \in \mathcal{P}(\mathrm{E})$$

Si $E = \{a, b, c\}$, tenim

$$\mathcal{P}(\mathrm{E}) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b\ c\}, \{a, b, c\}\}$$

$\emptyset$ designa el conjunt buit, és a dir, el conjunt que no conté cap element.

És necessari distingir l'element a i el conjunt reduït a l'element a, que es representa per $\{a\}$.

El cardinal d'un conjunt E finit és el seu nombre d'elements i es representa amb | |.

Si $|E| = n$, llavors $\left|\mathcal{P}(\mathrm{E})\right| = 2^n$.

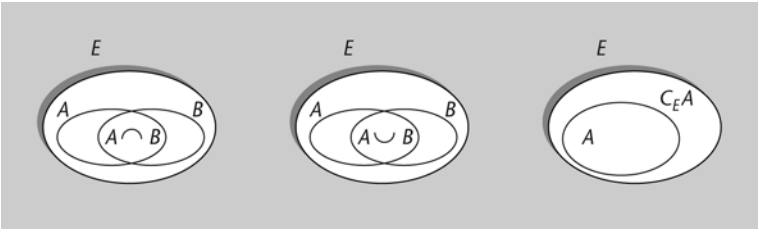
A i B són dos subconjunts de E. Les tres operacions fonamentals entre conjunts són la intersecció (producte lògic), representada per $\cap$; la unió, representada per $\cup$ (suma lògica); i la diferència, representada per $\overline{A}$ o per $C_E A$ si es tracta de la diferència de A. Aquestes operacions defineixen tres subconjunts nous:

$A \cap B$:
$x \in A \cap B \Leftrightarrow x \in A$ i $x \in B$: és el conjunt dels elements que pertanyen a A i a B.

$A \cup B$:
$x \in A \cup B \Leftrightarrow x \in A$ o $x \in B$: és el conjunt dels elements que pertanyen a A o a B.

$\overline{A}$ o $C_E A$:
$x \in \overline{A} \Leftrightarrow x \notin A$: és el conjunt dels elements que pertanyen a E però no a A.

Figura 23. Operacions sobre conjunts



Agafem exemples de l'ús d'aquests operadors en l'escriptura "d'equacions de cerca" documental.

Exemple

- Busquem documents que tracten del tema de l'estadística de la informació i la matemàtica de la informació. En la lògica booleana, escriurem l'equació de cerca booleana següent:

  (estadística O matemàtica) Y informació.

- L'operador O enllaça dos sinònims que tenen la mateixa importància: estadística i matemàtica; l'operador Y enllaça les frases: estadística de la informació, matemàtica de la informació.

- Si el que busquem són documents que tractin del tema de "les aplicacions bibliomètriques dins del camp de la innovació, però excloent els documents que tracten del tema de la vigilància tecnològica", escriurem, seguint la lògica booleana:

  (bibliometria O cienciometria O infometria)
  Y innovació Y NO vigilància tecnològica.

- Hi ha un altre operador equivalent, que és l'operador EXCEPTE. Llavors escrivim:

(bibliometria O cienciometria O infometria)
Y innovació EXCEPTE vigilància tecnològica.

L'operador O usat aquí és l'*o* lògic i no l'*o* exclusiu usat en el llenguatge habitual.

5.2. Cerca documental (bibliometria): probabilitat condicional

5.2.1. Probabilitat i esdeveniment

1) Probabilitat i esdeveniment elemental

Exemple

Tenim l'experiència que consisteix a llançar a l'atzar un dau cúbic amb les sis cares numerades de l'1 al 6. En llançar el dau es poden produir sis esdeveniments diferents: pot sortir un 1, un 2... o un 6. Si llancem una moneda, els resultats possibles estan constituïts per dos fets. Direm que tots els fets possibles d'aquesta experiència es poden descriure per mitjà dels esdeveniments elementals {1}, {2}... {6}.

Un observador pot estar interessat en un resultat concret d'aquesta experiència, com per exemple si el nombre que surt és parell o bé si és més gran que 4. En el primer cas seran els esdeveniments {2} o {4} o {6}, i en el segon cas seran els esdeveniments {5} o {6}.

Convencionalment, representarem l'esdeveniment "el nombre que surt és parell" amb {2, 4, 6} i el fet "el nombre que surt és més gran que 4" amb {5, 6}.

Podeu consultar informació sobre l'estadística probabilista en l'apartat 5 del mòdul 2, "Estadística de la informació".

2) Probabilitat condicional i esdeveniment independent

Si E designa el conjunt de tots els esdeveniments elementals, $\mathcal{P}(E)$ és el conjunt de tots els subconjunts de E i representa tots els esdeveniments possibles.

Hi ha dos esdeveniments que són singulars:

- L'esdeveniment segur, és a dir, el que es produeix sempre; es representa amb E.
- L'esdeveniment improbable, és a dir, l'esdeveniment que no es produeix mai; es representa amb ϕ.

Tenim dos esdeveniments A i B de manera que $A \in \mathcal{P}(E), B \in \mathcal{P}(E)$. Si la realització de l'esdeveniment A comporta la de B, això significa que A implica B. En un context de conjunts, direm més aviat que A està inclòs en B i ho representarem amb $A \subset B$.

Anomenem *probabilitat sobre* $(E, \mathcal{P}(\mathrm{E}))$ qualsevol aplicació P de $\mathcal{P}(E)$ en [0,1] que tingui les propietats següents $P(E) = 1$.

Per a tota successió O_n de fets de $\mathcal{P}(E)$ dos a dos disjunts (o incompatibles, és a dir, que no es poden produir simultàniament) tenim:

$$P\left(\bigcup_n O_n\right) = \sum_n P(O_n)$$

(axioma complet de les probabilitats)

Per a qualsevol esdeveniment A de $\mathcal{P}(E)$, definim l'esdeveniment contrari de A, representat per $\bar{A}$ (complementari de A). Tenim la relació següent:

$$P\left(\bar{A}\right) = 1 - P(A)$$

Això significa que la probabilitat d'un esdeveniment determina la probabilitat de l'esdeveniment oposat o contrari.

Exemple

Tornem a l'exemple de les monedes; suposarem que llancem dues monedes i que ens interessa l'esdeveniment "ha sortit creu en els dos llançaments". Busquem la relació del nombre de casos favorables sobre el nombre total de resultats possibles. Si P1 i F1 són els esdeveniments de creu i cara de la primera moneda i P2 i F2 els esdeveniments de creu i cara per a la segona moneda, el conjunt dels esdeveniments possibles associats al llançament de les dues monedes és {{P1, P2}, {P1, F2}, {F1, P2}, {F1, F2}}. Com que tots els esdeveniments són equiprobables, tenim:

$$P(\{\text{P1,P2}\}) = \frac{1}{4}$$

Escriurem $P(A \cap B)$ per a designar la probabilitat que els esdeveniments A i B es produeixin simultàniament. Quan els esdeveniments són independents, és a dir, no tenen cap vincle objectiu, $P(A \cap B)$ és igual al producte de les probabilitats $P(A)$ i $P(B)$, és a dir:

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

Així, en el nostre exemple:

$$P(P1 \cap P2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

ja que se suposa que els dos fets són independents.

Ja hem usat aquesta propietat quan hem estudiat la independència de dues variables nominals. Els esdeveniments eren la presència o l'absència d'una característica de les variables. La probabilitat de realització de cada esdeveniments es calculava en funció de les freqüències d'aparició d'aquestes característiques. Els efectius teòrics creuats corresponents a una situació d'independència de les variables, i en conseqüència a la realització simultània dels esdeveniments relatius a cadascuna, es calculaven amb el producte de les probabilitats dels esdeveniments considerats.

Podeu consultar informació sobre la independència de dues variables nominals en els subapartats 3.1 i 4.2 del mòdul 2 "Estadística de la informació".

El formalisme de les probabilitats condicionals completa aquesta noció d'esdeveniment simultani quan no ens trobem en un cas d'independència, sinó en un cas en el qual la realització d'un esdeveniment condiciona la de l'altre.

La probabilitat de realització d'un esdeveniment *A*, si sabem que l'esdeveniment *B* (de probabilitat no nul·la) s'ha produït, es coneix com a *probabilitat de A condicionada per B* o *probabilitat de A coneixent B*. Això s'escriu i es defineix:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Així, direm que dos esdeveniments són independents si tenim les relacions:

$$P(A \cap B) = P(A) \cdot P(B) \text{ o } P(A/B) = P(A)$$

5.2.2. Mesura del rendiment dels sistemes documentals: relació i precisió

Una pregunta plantejada a un sistema d'informació documental permet dividir el banc de documents en quatre conjunts:

- els documents extrets pertinents,
- els documents extrets no pertinents,
- els documents no extrets pertinents i
- els documents no extrets no pertinents.

a, *c*, *b*, *d* són els cardinals d'aquests conjunts. Els resultats es presenten en una taula de doble entrada (coneguda com a *taula de contingència*) del tipus:

Taula 51. Taula de contingència

	Extret	No extret
Pertinent	a	b
No pertinent	c	d

La resposta del sistema es caracteritza per dues proporcions:

- La relació (representada per *R*) és la proporció de documents pertinents extrets respecte als documents pertinents.
- La precisió (representada per *PR*) és la proporció de documents pertinents extrets respecte als documents extrets.

$$R = \frac{a}{a+b}$$

$$PR = \frac{a}{a+c}$$

Es pot donar una interpretació probabilista d'aquestes mesures. Si representem amb *Ex* i *Per* els esdeveniments aleatoris "ser extret" i "ser pertinent", el formalisme de les probabilitats condicionals ens permet escriure les igualtats següents:

$$R = P(Ex/Per) = P(Per \cap Ex)\frac{1}{P(Per)}$$

$$PR = P(Per/Ex) = P(Per \cap Ex)\frac{1}{P(Ex)}$$

La precisió, doncs, és la probabilitat que un document extret sigui pertinent. La relació és, en conseqüència, la probabilitat que un document pertinent sigui extret.

5.2.3. Formació de paraules en un idioma: lingüística quantitativa

Les probabilitats condicionals s'usen des de fa molt temps en lexicometria. En efecte, ja hem comentat que l'ordre d'aparició de les lletres en les paraules d'un idioma està vinculat a les lletres prèvies. Així, per exemple, en francès, la lletra *w* no precedeix mai a una altra *w* ni a una *z*. Aquesta característica s'ha quantificat amb l'ajuda de les probabilitats condicionals per a una gran part de les seqüències de lletres. Per exemple, podem escriure:

$$P(a/b\ u)$$

que és la probabilitat que aparegui el caràcter *a* sabent que precedeix el caràcter *b*, i que aquest precedeix al seu torn el caràcter *u*. Aquest tipus de tractament lexicomètric té múltiples usos per a l'anàlisi de textos i d'idiomes.

5.3. Proximitat de dos documents (infometria): coeficient d'associació

A i *B* són dos conjunts de *n* i *m* elements:

$$A = \{A_1, A_2, \ldots A_n\} \text{ i } B = \{B_1, B_2, \ldots B_m\}$$

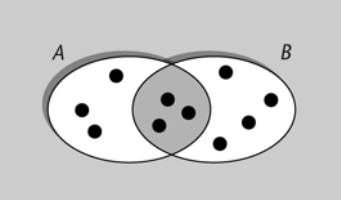
Anomenem *similitud* una mesura matemàtica de comparació d'aquests dos conjunts: com més gran sigui la similitud entre *A* i *B*, més s'assemblaran. I, inversament, una similitud de 0 indica que els conjunts són totalment disjuntos. El contrari de la similitud és la distància.

Hi ha diverses maneres de calcular similituds i distàncies entre conjunts.

5.3.1. Mesura nominal de comparació de conjunts

Si tots els elements de A i de B són diferents i de naturalesa equivalent, el mètode més senzill consisteix a considerar que la similitud de dos conjunts és el nombre d'elements que tenen en comú , és a dir, el cardinal de $A \cap B$ (representat per la zona ombrejada de la figura següent) representat com a $|A \cap B|$.

Figura 24. Intersecció de dos conjunts



Així, si A i B són documents i A_n i B_n les paraules d'aquests documents, com més paraules en comú tinguin A i B, més s'assemblaran.

No obstant això, aquest tipus de mesura és problemàtic perquè està directament vinculat al nombre d'elements de A i de B. Qualsevol comparació de distància o de similitud es converteix en impossible.

Exemple

A, B i C són els documents formats per les paraules següents:

- A: *difús*, *impossible*, *matemàtica*, *comparació*, *conjunts*, *inversament*, *idèntics*
- B: *conjunts*, *inversament*, *idèntics*
- C: *conjunts*, *inversament*, *idèntics*

Tindrem: $|A \cap B| = 3 = |B \cap C| = 3$.

La similitud entre A i B és, doncs, idèntica a la similitud B i C, la qual cosa és absurda, perquè B i C són totalment idèntics, mentre que A i B no.

Per a resoldre aquest problema, cal normalitzar les mesures entre 0 i 1. Així, s'obté una similitud d'1 si els dos conjunts són idèntics, mentre que una similitud de 0 indica que els conjunts són disjuntos, és a dir, que no tenen cap element en comú. (I inversament, s'obté una distància de 0 si els dos conjunts són idèntics, mentre que una distància d'1 indica que els conjunts són disjunts.)

Les tres mesures principals de similitud usades en el camp de la informació són:

- El coeficient de Jaccard:

$$\mathrm{Jac}(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} = \frac{|A \cap B|}{|A \cup B|}$$

- El coeficient de Dice:

$$\mathrm{Dice}(A,B) = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

- El cosinus:

$$\cos(A,B)=\frac{|A\cap B|}{\sqrt{|A|\cdot|B|}}$$

D'una banda, podem destacar que aquestes mesures està normalitzades (adopten valors compresos entre 0 i 1) i, d'altra banda, es distingeixen pels valors dels denominadors.

Exemple

Si tornem a l'exemple anterior i fem el càlcul del coeficient de Jaccard, tindrem:

$$\text{Jac}(A,B)=\frac{3}{7}=0,43,\quad \text{Jac}(B,C)=1$$

I, així doncs, Jac(A,B) < Jac(B,C).

La similitud entre A i B és més petita que entre B i C. De la mateixa manera, per al coeficient de Dice, tenim:

$$\text{Dice}(A,B)=\frac{6}{10}=0,6 \text{ i } \text{Dice}(B,C)=1$$

Així doncs, tenim Dice(A,B) < Dice(B,C).

Finalment, per al coeficient del cosinus, els resultats són:

$$\cos(A,B)=\frac{3}{\sqrt{21}}=0,65 \text{ i } \cos(B,C)=1$$

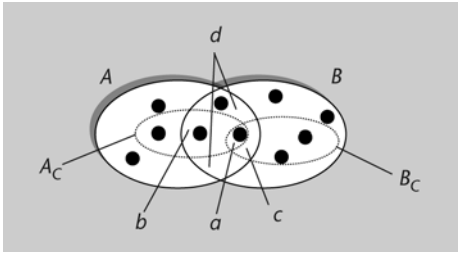
També tenim cos(A,B) < cos(B,C).

5.3.2. Mesura nominal de comparació de conjunts segons un criteri

Ara suposem que hem escollit un criteri concret i que volem comparar els conjunts A i B segons la presència d'elements que compleixin aquest criteri en els dos conjunts.

Descompondrem A i B en quatre subconjunts: A_C i B_C, que són els conjunts formats pels elements que compleixen el criteri, i $A_{\overline{C}}$ i $B_{\overline{C}}$, que són els conjunts formats pels elements que no el compleixen.

Figura 25. Elements d'intersecció



Les observacions es resumeixen en la taula de contingència següent:

Taula 52. Taula de contingència de dues variables

		B	
		B_C	$B_{\overline{C}}$
A	A_C	*a*	*b*
	$A_{\overline{C}}$	*c*	*d*

$a = |A_C \cap B_C|$, que representa el nombre d'elements comuns a A i B que compleixen el criteri.

$b = |A_C \cap B_{\overline{C}}|$, que representa el nombre d'elements de A que compleixen el criteri, i el nombre d'elements de B que no el compleixen.

En les ciències de la informació, aquesta taula s'usa per a determinar les mesures de rendiment dels sistemes d'informació, que són la relació i la precisió.

5.3.3. Mesura nominal de dos conjunts segons diversos criteris

Quan s'escullen diversos criteris, i no solament un, per a descriure les variables A i B, la taula de contingència és més gran (taula 53).

Taula 53. Taula de contingència de les variables dels criteris

		B							
		C_1	C_2	C_3		C_j			C_n
A	C_1	F_{11}	F_{22}	…	…	F_{1j}	…	…	F_{1n}
	C_2	F_{21}							…
									…
									…
	C_i	F_{i1}	…	…	…	F_{ij}			…
		…							…
		…							…
	C_n	F_{in}	…	…	…	…	…	…	F_{nn}

En cada cel·la de la taula, el valor F_{ij} mesura el nombre d'elements de A i B que compleixen alhora els criteris C_i i C_j. Els mètodes d'anàlisi de dades permetran estudiar les dependències entre els dos conjunts.

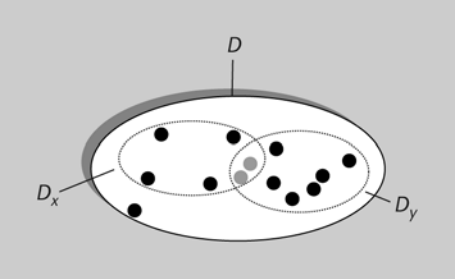
Podeu consultar informació sobre dependències entre dos conjunts en el subapartat 3.1del mòdul 2 "Estadística de la informació".

5.3.4. Comparació de conjunts segons la proximitat de dos criteris

Ara suposem que la comparació dels conjunts es basa en la presència comuna, o la coocurrència, de dos criteris en els elements dels conjunts. En les investigacions d'informació, aquest tipus d'estudi es fa, entre altres, dins del marc de l'emparellament pregunta-document i per a construir mapes temàtics.

D és una col·lecció de documents i x i y són dues paraules presents en alguns d'aquests documents; la intersecció dels dos subconjunts de documents es representa de la manera següent:

Figura 26. Intersecció de conjunts definida segons dos criteris



D_x i D_y són els dos subconjunts de documents que contenen les paraules x i y.

Amb aquestes notacions, els coeficients de similitud anteriors s'escriuen:

- El coeficient de Jaccard:

$$\mathrm{Jac}\left(D_x, D_y\right) = \frac{\left|D_x \cap D_y\right|}{\left|D_x \cup D_y\right|}$$

- El coeficient de Dice:

$$\mathrm{Dice}\left(D_x, D_y\right) = \frac{2\left|D_x \cap D_y\right|}{\left|D_x\right| + \left|D_y\right|}$$

- El cosinus:

$$\cos\left(D_x, D_y\right) = \frac{\left|D_x \cap D_y\right|}{\sqrt{\left|D_x\right|} \cdot \sqrt{\left|D_y\right|}}$$

Aquests coeficients ens permeten definir una mesura de proximitat entre les paraules x i y. Dues paraules seran més properes com més gran sigui el nombre de documents en què es trobin simultàniament.

Exemple

Tornem als documents A, B i C anteriors i calculem la proximitat de les paraules.

x = conjunts, y = idèntics, i z = distància

D_x, D_y i D_z són els tres subconjunts de documents que contenen les paraules x, y i z.

D_x = {A,B,C}, D_y = {A,B,C}, D_z = {A}

Calculem el coeficient Jaccard:

$$\mathrm{Jac}(D_x, D_y) = 1, \quad \mathrm{Jac}(D_z, D_y) = \frac{1}{3}$$

Les paraules *x* i *y* són més pròximes que *y* i *z*, ja que sistemàticament sempre es troben juntes en tots els documents.

Observació

El coeficient del cosinus elevat al quadrat és justament el coeficient usat en les anàlisis de les paraules associades, i mesura la força d'associació entre dues paraules.

Podeu consultar informació sobre l'anàlisi de les paraules associades en el subapartat 3.5 del mòdul 2, "Estadística de la informació".

5.4. Similitud entre pregunta i resposta (infometria): espai vectorial

Com es pot mesurar la proximitat de dos conjunts definits a partir de diversos criteris? Un dels possibles models de descripció dels conjunts és el dels espais vectorials, desenvolupat per Salton.

Bibliografia

G. Salton; M. J. McGill (1984). *Introduction to modern information retrieved*. Nova York: McGraw-Hill.

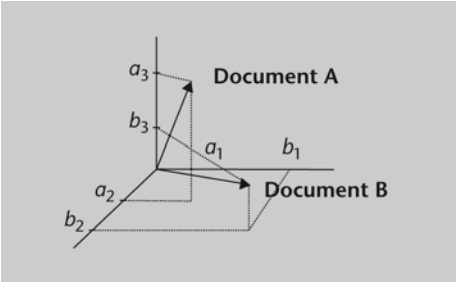
5.4.1. El model vectorial

D és un conjunt de documents i *M* és el conjunt de *m* paraules $\{M_1, M_2, \ldots M_i, \ldots M_m\}$ presents en els documents. Cada document es representarà sota la forma d'un vector amb *m* components:

$$\text{Document A: } \vec{A} = [a_1\ a_2 \ldots a_m]$$
$$\text{Document B: } \vec{B} = [b_1\ b_2 \ldots b_m]$$

En un espai de tres dimensions, els documents es representaran, doncs, de la manera següent:

Figura 27. Representació vectorial dels documents A i B en un espai de tres dimensions



Els valors a_1 i b_3 són els "pesos" de les paraules M_i i M_j presents en els documents *A* i *B*. Quantifiquen la manera com *A* i *B* estan representats per aquestes dues paraules.

Aquest tipus de model s'ha usat per a calcular la proximitat d'una pregunta (formada per *m* paraules) i d'un document, i per a calcular la proximitat de dos documents.

5.4.2. Càlcul de la proximitat de dos documents

Per a determinar aquesta proximitat, calculem el cosinus de l'angle que formen els dos vectors dels documents entre ells.

El cosinus o el coeficient de Salton:

$$\cos\left(\vec{A},\vec{B}\right)=\frac{\vec{A}\cdot\vec{B}}{\left\|\vec{A}\right\|\cdot\left\|\vec{B}\right\|}=\frac{\sum_{k=1}^{m}a_k b_k}{\sqrt{\sum_{k=1}^{m}(a_k)^2}\sqrt{\sum_{k=1}^{m}(b_k)^2}}$$

$\vec{A}\cdot\vec{B}$ és el producte escalar dels vectors $\vec{A}$ i $\vec{B}$, i $\left\|\vec{A}\right\|$ i $\left\|\vec{B}\right\|$ designen la norma euclidiana dels vectors $\vec{A}$ i $\vec{B}$ (vegeu l'exercici 35).

Aquesta fórmula, inicialment adaptada per als espais vectorials, s'ha transformat per a aplicar-la als conjunts nominals. En efecte, el producte escalar dels vectors $\left(\vec{A}\cdot\vec{B}\right)$ se substitueix pel cardinal de la seva intersecció $\left(|A\cap B|\right)$. Si a_i i b_j adopten únicament els valors 1 i 0 per especificar si les paraules M_i i M_j estan presents en el document, llavors $\sum_{k=1}^{m}a_k b_k$ és igual a les quantitats de paraules comunes als documents A i B.

Observació

Les mesures de les similituds anteriors de Jaccard i de Dice es poden escriure:

- El coeficient de Jaccard

$$\text{Jac}\left(\vec{A},\vec{B}\right)=\frac{\sum_{k=1}^{m}a_k b_k}{\sum_{k=1}^{m}a_k+\sum_{k=1}^{m}b_k-\sum_{k=1}^{m}a_k b_k}$$

- El coeficient de Dice

$$\text{Dice}\left(\vec{A},\vec{B}\right)=\frac{2\sum_{k=1}^{m}a_k b_k}{\sum_{k=1}^{m}a_k+\sum_{k=1}^{m}b_k}$$

5.4.3. Càlcul de la proximitat d'una pregunta i d'un document

A és un document representat pel vector:

$$A=\sum_{k=1}^{m}a_k\overrightarrow{M}_k$$

Q és una pregunta plantejada al sistema; aquesta és una llista de paraules del conjunt M, que es pot escriure:

$$Q=\sum_{k=1}^{m}q_k\overrightarrow{M}_k$$

en què q_k val 0 o 1 segons si la paraula M_k està present o no en la pregunta.

El càlcul, amb algun dels coeficients anteriors, de la proximitat de la pregunta amb cada document del corpus, permet classificar els documents per ordre de pertinència.

El càlcul de la proximitat de Q i A amb el coeficient del cosinus és:

$$\cos(Q,A) = \frac{\sum_{k=1}^{m} q_k a_k}{\sqrt{\sum_{k=1}^{m} (q_k)^2}\sqrt{\sum_{k=1}^{m} (a_k)^2}}$$

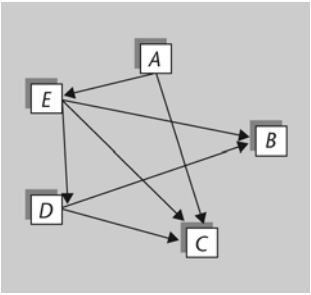
5.5. Mapes dels vincles entre llocs web (webmetria): sociogrames, gràfiques

En el subapartat 2.4 ja hem vist que les estadístiques multidimensionals es poden usar per a analitzar grans volums de valors numèrics i generar cartografies que permeten agregar elements propers o bé detectar xarxes de col·laboració. Les tècniques matemàtiques sorgides de la teoria de les gràfiques permeten treure altres tipus de conclusions, en particular sobre els diferents vincles possibles entre els elements. Aquests mètodes tenen diverses aplicacions, que són especialment útils per a analitzar l'estructura de la "tela" que s'estudia aquí com un espai físic en el qual ens desplacem.

Exemple

Tenim cinc llocs web representats per A, B, C, D i E, i considerats com a pertinents en relació amb una temàtica concreta. Una vegada identificats els vincles hipertextuals que els relacionen, podem construir la gràfica, composta per cinc nusos (nombre de llocs web) i tantes fletxes (aquí vuit) com vincles de navegació hi hagi entre els llocs.

Figura 28. Gràfica dels vincles hipertextuals de cinc llocs web



Quan el nombre de nusos creix, ja no podem analitzar manualment aquestes estructures. En aquests casos, usem una representació matricial, és a dir, una taula amb tantes files i columnes com nusos, i plena de valors 1 o 0 en funció de la presència o l'absència de vincles entre ells.

La lectura es fa de la manera següent:

- En la primera fila, els uns indiquen que el lloc A assenyala només cap als llocs C i E.
- En la primera columna, tots els zeros indiquen que cap lloc no assenyala cap a A.

Taula 54. Taula dels llocs web assenyalants-assenyalats

		Assenyalats				
	Assenyala cap a	***A***	***B***	***C***	***D***	***E***
Assenyalants	***A***	0	0	1	0	1
	B	0	0	1	0	0
	C	0	0	0	0	0
	D	0	1	1	0	0
	E	0	1	1	1	0

Podem extreure la matriu "assenyalants-assenyalats" següent (representada per PP):

Taula 55. Matriu (representada per PP)

0	0	1	0	1
0	0	1	0	0
0	0	0	0	0
0	1	1	0	0
0	1	1	1	0

Destacarem que aquesta matriu és quadrada (té tantes línies com columnes). Té zeros en la diagonal, la qual cosa significa que aquí no tenim en compte els vincles de navegació interns del lloc. A més, a causa de l'orientació del vincle de navegació (si *A* assenyala cap a *C*, *C* pot no assenyalar cap a *A*), aquesta matriu no és simètrica en relació amb la diagonal.

Com podem aprofitar aquesta matriu per a conèixer l'estructura dels llocs web de la Xarxa?

Sumant els valors de les files i de les columnes, podem determinar quins són els llocs que assenyalen més o menys cap a altres o que són més assenyalats per altres. Per exemple, el lloc més citat és *C*, mentre que aquest és el que menys assenyala.

Si fem el producte matricial de la matriu PP per ella mateixa, determinem el nombre de vincles indirectes de longitud 2 entre cada parell de nusos. Per exemple, anem de *E* a *C* per dos camins; el primer és directe, mentre que el segon passa per *B*; aquest últim és de longitud 2, ja que segueix dos vincles.

El producte de les dues matrius X i Y (X té tantes files com columnes té Y) és una tercera matriu Z. El seu nombre de files és el de X, el seu nombre de columnes és el de Y i els seus valors es calculen de la manera següent:

Y

Y_{11}	Y_{12}	Y_{13}	Y_{14}
Y_{21}	Y_{22}	Y_{23}	Y_{24}
Y_{31}	Y_{32}	Y_{33}	Y_{34}

X

X_{11}	X_{12}	X_{13}
X_{21}	X_{22}	X_{23}

Z_{11}	Z_{12}	Z_{13}	Z_{14}
Z_{21}	Z_{22}	Z_{23}	Z_{24}

Z

$$Z_{11} = X_{11} \cdot Y_{11} + X_{12} \cdot Y_{21} + X_{13} \cdot Y_{31}$$
$$Z_{12} = X_{11} \cdot Y_{12} + X_{12} \cdot Y_{22} + X_{13} \cdot Y_{32}$$
$$Z_{13} = X_{11} \cdot Y_{13} + X_{12} \cdot Y_{23} + X_{13} \cdot Y_{33}$$
$$\ldots$$

En el nostre exemple, el producte matricial de PP per ella mateixa serà PP^2. Els seus valors es donen a continuació:

Taula 56. PP^2: producte matricial de PP per ella mateixa

0	1	1	1	0
0	0	0	0	0
0	0	0	0	0
0	0	1	0	0
0	1	2	0	0

Així, doncs, hi ha un camí indirecte de longitud 2 per a anar de A a B (passant per E), un de A a C (passant per E)... i dos de E a C (passant per D i per B).

De la mateixa manera, fent el producte matricial $PP^3 = PP \times PP \times PP$ obtenim el nombre de vincles de longitud 3, etc.

D'altra banda, la suma de PP i PP^2 permet conèixer el nombre de camins de longitud 1 o de longitud 2 entre dos nusos. Els zeros permeten posar de manifest els llocs que no estan directament vinculats ni tampoc indirectament per camins de longitud 2.

Taula 57. PP + PP^2: nombre de vincles directes o de longitud 2 entre els llocs

0	1	2	1	1
0	0	1	0	0
0	0	0	0	0
0	1	2	0	0
0	2	3	1	0

Així, tenim un vincle de longitud 1 i un vincle de longitud 2 entre el lloc *A* i el lloc *C*.

5.6. En resum

Last but not least, ja que gràcies als desenvolupaments anteriors, els conjunts (i la seva teoria) són uns objectes omnipresents en el sector de la informació que usem tan sovint i sense adonar-nos que això ha marcat les professions del sector. La col·lecció de llibres, d'objectes museístics, el fons documental, els arxius audiovisuals i actualment els arxius electrònics són uns conjunts informacionals sobre els quals funcionen els dispositius que apliquen lògiques matemàtiques, com la lògica booleana, les estructures geomètriques i algebraiques, com els espais vectorials i les gràfiques. Identifiquem, busquem, comparem, vinculem les informacions gràcies a aquestes lògiques, gràcies a aquestes estructures.

Conclusió

Entre l'estadística i la matemàtica, però sense inclinar-nos excessivament envers l'una ni l'altra, hem volgut explorar els universos desconeguts de la informació. Una informació infinitament creixent, una informació infinitament ràpida, una informació infinitament complexa; com la podem comprendre per a controlar-ne millor la producció, la comunicació i l'ús? A més, les tècniques que la produeixen, les tècniques que la memoritzen i les tècniques que la vehiculen depassen tots els dies els límits de l'infinitament petit i els límits de l'infinitament gran.

En primer lloc, l'hem abordada en les seves dimensions més modestes gràcies a l'estadística unidimensional i a les sèries que formen. Després l'hem estudiada en la seva bidimensionalitat. En efecte, sovint es correlaciona amb alguna cosa i encara més sovint en funció d'alguna cosa. Sempre hi ha algú que la demana i sempre hi ha algú que la produeix. Finalment, encara que no per això menys important, l'hem vista en la seva multidimensionalitat, bàsicament la dels agregats informacionals que forma.

Les regularitats ocultes que segueix es van desvelant a poc a poc; són immenses en comparació de les que s'han pogut descobrir fins ara. Totes són el resultat de les temptatives, audaces en el seu temps, de recomptes, de classificacions i de normalitzacions que han estat abordades pels professionals dels diferents sectors de la informació, de les biblioteques, dels centres de documentació, dels museus, dels arxius i dels mitjans de comunicació tradicionals. Avui dia, aquestes temptatives són a càrrec dels mateixos professionals que elaboren i gestionen les versions cada cop més electròniques d'aquests serveis d'informació.

També hem intentat fer un primer pas en la direcció d'una relació més profunda de l'eina estadística i matemàtica en les ciències de la informació. Els desenvolupaments actuals de les activitats científiques, tècniques i industrials en els diferents sectors de la informació i de la cultura permeten preveure un ús més intensiu d'aquesta eina, però també seria millor que es descobrissin nous mètodes, noves lleis i noves tècniques estadístiques i matemàtiques més benadaptades a l'objecte *informació*.

Al costat de les diverses cultures que la informació ha anat incorporant fins ara, les ciències de la informació integren una cultura que potser molt pocs esperaven, que és la cultura matemàtica.

Activitats

Exercicis sobre les sèries

20. Gestió de la informació en un laboratori d'investigació (infometria)

El gestor del centre de documentació d'un laboratori d'investigació especialitzat en l'estudi dels mitjans de comunicació analitza la cobertura del tema per part de les publicacions científiques i professionals. Calcula en 1.000 el nombre de publicacions dins del camp que estudia. D'aquestes, n'analitza 164 detalladament i troba 396 articles sobre els mitjans de comunicació. Després classifica aquestes 164 publicacions segons el nombre decreixent d'articles que contenen i crea tres grups de publicacions, de manera que el nombre d'articles sobre els mitjans de comunicació sigui equivalent en cada grup (que ell fixa en 132 articles):

1r. grup: 10 publicacions (G_1) que contenen 127 articles.
2n. grup: 35 publicacions (G_2) que contenen 133 articles.
3r. grup: 119 publicacions (G_3) que contenen 136 articles.

a) Demostreu que aquesta distribució segueix la llei de Bradford i calculeu el nucli d'aquesta literatura sobre els mitjans de comunicació i el factor de Bradford.
b) Quantes publicacions cal analitzar si volem doblar el nombre d'articles?
c) Quants articles hem d'esperar trobar si analitzem les 1.000 publicacions existents?

21. Llei de Bradford (infometria)

Un bibliotecari ha comptat totes les revistes que apareixen en un repertori bibliogràfic de geofísica aplicada. Després ha classificat aquestes revistes segons el nombre d'articles publicats sobre un tema concret de geofísica i ha obtingut la classificació següent:

Núm. de revista	Nre. d'articles	Rang
Núm. 1	26	1
Núm. 2	20	2
Núm. 3	19	3
Núm. 4	10	4
Núm. 5	6	5
Núm. 6	6	5
Núm. 7	5	7
Núm. 8	5	7
Núm. 9	4	10
Núm. 10	4	10
Núm. 11	4	10
Núm. 12	3	13
Núm. 13	3	13
Núm. 14	3	13
Núm. 15	3	13
Núm. 16	2	17
Núm. 17	2	17
Núm. 18	2	17
Núm. 19	2	17
Núm. 20	2	17
Núm. 21	2	17

Núm. de revista	Nre. d'articles	Rang
Núm. 22	2	17
Núm. 23	1	18
Núm. 24	1	18
Núm. 25	1	18
Núm. 26	1	18
Núm. 27	1	18
Núm. 28	1	18
Núm. 29	1	18
Núm. 30	1	18
Núm. 31	1	18
Núm. 32	1	18
Núm. 33	1	18
Núm. 34	1	18
Núm. 35	1	18
Núm. 36	1	18
Núm. 37	1	18
Núm. 38	1	18
Núm. 39	1	18
Núm. 40	1	18
Núm. 41	1	18
Núm. 42	1	18
Núm. 43	1	18
Núm. 44	1	18
Núm. 45	1	18
Núm. 46	1	18

Volem calcular els paràmetres de la llei de Bradford d'aquesta col·lecció.

a) Després de posar aquests valors en una taula de tres columnes (nombre d'articles, nombre de revistes que els contenen i nombre total d'articles), dividiu la col·lecció en un nombre arbitrari de grups, cadascun amb el mateix nombre n d'articles. Les classes de revistes, organitzades per productivitat decreixent, es designen amb l'índex k. Per a cada classe, calculeu G^k, el nombre de revistes, i r_k, el nombre mitjà d'articles per revista.

b) Si N és el nombre total de classes, calculeu el coeficient q per verificar la relació de Bradford:

$$G^k = q^{k-1} \cdot G^1 = q^{k-1} \cdot r \quad \text{k} = 1 \ldots \text{N} \qquad [1]$$

c) Si $R(k)$ és la suma acumulada dels articles fins a la classe k, demostreu que tenim la relació:

$$R(k) = km \qquad [2]$$

Quin és el valor de m?

d) Si J_k és la suma acumulada de totes les revistes incloses fins a la classe d'ordre *k*, demostreu que podem escriure la relació:

$$J_k = \left(\frac{q^k - 1}{q - 1}\right) \cdot G^1 \qquad [3]$$

e) Tenim que $t = \frac{G^1}{q-1}$ i $j = \frac{m}{\log q}$. Demostreu que podem escriure la relació:

$$R(k) = j \log\left(\frac{J_k}{t} + 1\right) \qquad [4]$$

Substituint *t* pel seu valor i usant el resultat [3], demostreu que tenim la relació següent:

$$R(k) = j \cdot \log J_k + B_k \text{ amb } B_k = j \cdot \log\left(\frac{q^k}{G^1}\frac{q-1}{q^k-1}\right) \qquad [5]$$

f) Dibuixeu la corba experimental i la corba teòrica. Quines conclusions es poden treure?

22. Ús de la llei de Bradford (infometria)

Aquest exercici és idèntic a l'anterior, excepte pel fet que la bibliografia que s'ha reunit és sobre biologia teòrica.

Rang	Nre. d'articles	Núm. de revista
1	66	Núm. 1
2	50	Núm. 2
3	20	Núm. 3
4	15	Núm. 4
4	15	Núm. 5
6	13	Núm. 6
6	13	Núm. 7
8	11	Núm. 8
8	11	Núm. 9
10	10	Núm. 10
10	10	Núm. 11
12	8	Núm. 12
12	8	Núm. 13
14	6	Núm. 14
14	6	Núm. 15
16	5	Núm. 16
16	5	Núm. 17
18	4	Núm. 18
18	4	Núm. 19
20	3	Núm. 20
20	3	Núm. 21
20	3	Núm. 22
20	3	Núm. 23

Rang	Nre. d'articles	Núm. de revista
20	3	Núm. 24
20	3	Núm. 25
20	3	Núm. 26
20	3	Núm. 27
29	2	Núm. 28
29	2	Núm. 29
29	2	Núm. 30
29	2	Núm. 31
29	2	Núm. 32
29	2	Núm. 33
29	2	Núm. 34
29	2	Núm. 35
29	2	Núm. 3ó
29	2	Núm. 37
37	1	Núm. 38
37	1	Núm. 39
37	1	Núm. 40
37	1	Núm. 41
37	1	Núm. 42
37	1	Núm. 43
37	1	Núm. 44
37	1	Núm. 45
37	1	Núm. 46

a) Calculeu les constants q i m que confirmen els dos resultats següents:

$$G^k = q^{k-1}G^1 \quad \mathrm{k} = 1 \cdots \mathrm{N} \qquad [1]$$

$$R(k) = km \qquad [2]$$

b) Dibuixeu la corba experimental i la recta teòrica ($R(k)$, $\log(J_k)$).

23. Peces visitades i durada de les visites (museometria)

En fer estudis de les visites d'exposicions a museus, s'ha constatat que, per a les visites de durada curta (inferiors a dues hores), el nombre de peces (elements d'una exposició) visitades pels visitants variaven de la manera següent:

Nombre de visitants (V)	18	32	50	72	98	128	162	200
Nombre de peces visitades (E)	3	4	5	6	7	8	9	10

a) Representeu gràficament la relació entre E i V.

- Quin tipus de relació us suggereix aquesta gràfica?
- Comproveu aquesta relació mitjançant el càlcul.

b) Feu una transformació logarítmica (logaritme neperià) d'aquests valors. Quina és la relació que hi ha entre ln(V) i ln(E)?

24. Audiència de les pel·lícules (mediametria)

El 1999, es va analitzar el públic de les sales de cinema segons si les pel·lícules estaven produïdes a França o als Estats Units. Els resultats de les dues distribucions d'ús es representen en la taula següent:

Taula 46. Audiències de les pel·lícules franceses i americanes el 1999

Nombre d'entrades *Ui*	**Nombre de pel·lícules *Ni*(Fr)**	**Nombre de pel·lícules *Ni*(US)**
Menys de 10.000	47	32
Entre 10.000 i 25.000	20	27
Entre 25.000 i 50.000	16	21
Entre 50.000 i 100.000	19	16
Entre 100.000 i 250.000	23	32
Entre 250.000 i 500.000	16	18
Entre 500.000 i 1 milió	13	22
Entre 1 milió i 2 milions	7	16
Més de 2 milions	3	8
Total	**164**	**192**

Podem ajustar aquestes distribucions d'ús amb una distribució coneguda? Interpreteu els resultats.

Exercicis sobre les funcions

25. Augment de la producció en ciències (cienciometria)

El nombre de desenvolupaments científics segueix des del seu inici una llei de creixement exponencial. Per exemple, si a és la dimensió inicial d'un conjunt d'autors d'un conjunt d'articles en el moment $t = 0$ i b la taxa de creixement, la dimensió o la mida del conjunt en el moment t serà:

$$C(t) = ae^{bt}$$

Aquests augments no poden prosseguir sense fi, i en algun moment es produeix una disminució de la velocitat de creixement fins que s'anul·la i s'arriba a la saturació.

En aquest moment definim una altra llei de creixement, representada per l'equació següent:

$$C(t) = \frac{k}{1 + ae^{-bt}}$$

Aquesta corba es coneix amb el nom de *corba logística* o *corba en S*.

El punt d'inflexió d'aquesta corba indica el moment en què comença aquesta disminució. En el cas d'un desenvolupament científic d'una durada de trenta anys i sabent que les constants a, b i k són, respectivament, $a = 100$, $b = 0{,}8$ i $k = 1.000$, determineu aquest punt d'inflexió de dues maneres diferents: per un mètode numèric i per un mètode analític.

El punt d'inflexió

És el punt en què canvia el radi de curvatura de la corba; de còncava passa a convexa o a la inversa.

a) Per al mètode numèric, dibuixeu la corba amb 50 punts.
b) Per al mètode analític, calculeu les derivades primeres i segones i el límit quan t tendeix a l'infinit.

26. Volum de respostes d'un motor de cerca (webmetria)

Volem mesurar el rendiment d'un motor de cerca. Farem noranta-vuit consultes i anotarem la resposta per a cadascuna, és a dir, el nombre de documents obtinguts. Aquests valors es presenten en la taula següent, que s'ha de llegir de la manera següent: en plantejar la primera pregunta s'han obtingut 5 documents, en plantejar la segona pregunta s'han obtingut 5 documents, en plantejar la tercera pregunta s'han obtingut 4 documents, etc.

Volem saber si les preguntes són molt específiques (s'obtenen pocs documents) o si, per contra, són molt generals (s'obtenen molts documents).

5	19	132	75	61
5	6	28	52	115
4	18	38	70	118
65	33	1	73	50
19	16	35	15	54
31	17	36	67	55
26	15	7	59	58
29	10	2	3	2
29	12	6	7	3
12	13	39	2	3
156	14	30	148	123
105	20	42	79	160
142	11	27	51	103
1	40	4	1	42
135	12	29	53	108
138	7	22	44	87
126	8	23	45	90
101	9	24	47	94
2	9	25	48	97
	21	1	83	

a) Després de classificar els valors per ordre creixent, dibuixeu la corba que representa el volum del nombre de documents obtinguts en funció de les preguntes plantejades. Quina forma té?

b) Què es pot veure per als valors grans? Expliqueu-ho.

27. Gestió de les adquisicions d'una biblioteca (bibliometria)

Si l'objectiu d'una biblioteca és presentar una oferta documental adequada a la demanda dels lectors, haurà de tenir en compte dos elements: el nombre d'obres de cada categoria del fons F i l'ús U que se'n fa cada any.

Per al fons en conjunt:
F és el nombre total d'obres.
U és el nombre d'obres que s'han deixat en préstec l'any anterior.

Per a cada categoria d'obres i (i varia d'1 a N):
F_i és el nombre d'obres de la categoria i.
U_i és el nombre d'obres d'aquesta categoria i que s'han deixat en préstec l'any anterior.

Per a distribuir al llarg de l'any següent les seves compres d'obres, un bibliotecari decideix usar la fórmula empírica següent:

$$F_i = \frac{F}{U} \cdot U_i$$

a) Expliqueu el significat d'aquesta fórmula i demostreu que pot ajudar a fer la distribució de les compres.

b) Sabent que l'any 2000 els volums de préstecs es van repartir segons les set categories següents, proposeu una distribució de les compres per a l'any 2001:

Categories	Obres	Préstecs 2000
A	241	230
B	216	114
C	100	99
D	226	273
E	129	109
F	43	37
G	239	422

Comenteu els resultats.

c) El bibliotecari vol fer una altra projecció modificant la seva fórmula de la manera següent:

$$F_i = \frac{F}{\sum_{i=1}^{N} U_i^x} U_i$$

Feu una projecció amb $x = \frac{2}{3}$. Comenteu els resultats.

28. Ús de la llei de Lotka (cienciometria)

Al llarg de tres anys s'ha fet un estudi sobre la distribució de la producció d'articles per autors que treballen en el camp de l'apoptosi (mort cel·lular programada). Els valors obtinguts s'han extret del banc bibliogràfic Medline. Només s'ha tingut en compte el primer autor.

Nombre d'articles publicats	Nombre d'autors
1	339
2	85
3	53
4	22
5	18
6	10
7	3
8	3
9	3
10	6
13	2

Demostreu que la distribució anterior es pot descriure amb l'ajuda d'una funció de tipus potència. Doneu-ne l'expressió matemàtica.

Exercicis sobre les equacions

29. Temps de reacció per a escollir una operació en un terminal interactiu (mediametria)

El temps de reacció (en segons) per a escollir una operació en un menú d'una pantalla tàctil d'un terminal interactiu depèn del nombre d'alternatives (b) i de dues constants, c i k, que caracteritzen un usuari. L'equació que permet calcular aquest temps és $t = c + k \cdot \log_2 (b)$, en què $\log_2$ és el logaritme de base 2.

Hem cronometrat un usuari que en un primer terminal que ofereix 4 opcions ha reaccionat en 10 segons, mentre que en un segon terminal que ofereix 6 opcions ha reaccionat en 20 segons.

A partir d'aquestes mesures, calculeu les constants c i k i escriviu l'equació que caracteritza l'usuari.

30. Taxa d'ocupació d'una línia de transmissió per un terminal (mediametria)

La taxa d'ocupació E (en erlangs) d'una línia de transmissió per un terminal es defineix amb l'equació $E = \frac{C \cdot t}{3.600}$, en què C és el nombre de transmissions per hora i t la durada mitjana en segons d'una transmissió.

a) A què correspon una taxa d'ocupació E propera a 1 erlang?
b) Calculeu les taxes d'ocupació de les línies de transmissió per a una companyia de serveis informatius (C = 0,8 i t = 2.700), per a un servei de gestió d'estocs documentals informatitzats (C = 0,45 i t = 1.200) i per a un banc d'informacions (C = 15 i t = 120). Traieu-ne conclusions.

31. Obsolescència dels articles (cienciometria)

Tenim un conjunt de N articles científics i $U(t)$ és el nombre acumulat d'articles citats com a mínim una vegada durant un període de temps t. Suposem que $U(t)$ varia de manera proporcional al nombre d'articles no citats segons un factor $\alpha(t)$, que decreix de manera exponencial amb el temps. Així, tenim $\alpha(t) = A \cdot e^{-at}$, en què A i α són dues constants positives. La proporció d'articles citats en la data t és, doncs, $R(t) = \frac{U(t)}{N}$.

a) Demostreu que $R(t)$ verifica l'equació exponencial:

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = A \cdot e^{-at} \cdot \left(1 - R(t)\right)$$

b) Demostreu que $R(t) = 1 - \frac{\left(1 - R(0)\right)}{b} \cdot b^{-e^{at}}$ amb $b = e^{\frac{A}{a}}$.

c) Què ocorre quan t tendeix a l'infinit?

32. Exponent de la llei de Lotka (cienciometria)

Tenim l'equació diferencial $\frac{\mathrm{d}G}{\mathrm{d}u} + a \cdot \frac{G}{u} = 0$, en què G representa el nombre d'autors que han publicat u articles i a és una constant positiva que cal determinar. Calculeu la funció G que soluciona aquesta equació i interpreteu-ne el resultat.

Exercicis sobre els conjunts

33. Escriptura d'una equació booleana (infometria)

a) En fer una investigació documental informatitzada amb un sistema que usa un programa informàtic booleà, els usuaris formulen preguntes de la forma A EXCEPTE B, A I NO B, $A \circ B$.

Les taules de veritats són unes eines que permeten escriure tots els casos de xifres possibles juntament amb les seves combinacions. Per exemple, la primera fila de la taula de veritat de l'operador EXCEPTE següent es llegeix: si A és veritat (= 1) i B és fals (= 0), llavors A EXCEPTE B és veritat (= 1):

Operador EXCEPTE:

A	*B*	*A* EXCEPTE *B*
1	0	1
0	1	0
1	1	0
0	0	0

Les taules dels operadors Y, O i NO són:

Operador Y

A	*B*	*A* Y *B*
0	1	0
0	0	0
1	0	0
1	1	1

Operador O

A	*B*	*A* O *B*
1	1	1
1	0	1
0	1	1
0	0	0

Operador NO

A	NO *A*
1	0
0	1

Amb aquestes taules, demostreu que les dues consultes (*A* EXCEPTE *B*) i (*A* Y NO *B*) són equivalents.

b) La mateixa pregunta amb les dues consultes:

- (*A* Y *B*) O (*A* Y *C*).
- *A* Y (*B* O *C*).

34. Avaluació d'un servei d'informació (infometria)

Per a una pregunta concreta, un banc d'informacions *G* conté 162 documents pertinents (G_0) i 849 documents no pertinents (G_1). *U* és la resposta del sistema a aquesta pregunta, i conté 28 documents pertinents i 35 documents no pertinents.

a) Calculeu la similitud entre els conjunts *U* i G_0 amb l'ajuda dels coeficients de similitud.

b) Calculeu la relació i la precisió.

c) Demostreu que $\frac{1}{D(U,G_0)} = \frac{1}{2R} + \frac{1}{2P}$, és a dir, que l'invers de Dice és igual a la mitjana harmònica de *R* i *P* (la mitjana harmònica d'un conjunt de nombres és la inversa de la mitjana aritmètica de les inverses dels nombres). Demostreu numèricament aquesta relació.

35. Cerca d'informació a Internet (webmetria)

Un internauta planteja la pregunta "Les lleis sobre la repressió del frau" a un motor de cerca i la resposta inclou 5 llocs. Per a cadascun, el motor comptabilitza el nombre d'ocurrències de les paraules *llei*, *repressió* i *frau*. Els resultats es donen en la taula següent:

	Llei	**Repressió**	**Frau**
Lloc A	4	2	4
Lloc B	4	0	4

	Llei	Repressió	Frau
Lloc C	5	4	1
Lloc D	1	1	6
Lloc E	4	5	1

a) Representeu la pregunta i els diferents llocs usant la modelització de Salton.
b) Quin és el lloc que respon millor la pregunta?

36. Comparació de les produccions científiques de diferents països (cienciometria)

Volem definir un índex de no-semblança que caracteritza amb un valor positiu, dins de les tres disciplines científiques de la física, la biologia i la química, la distància entre dos països segons el nombre d'articles referenciats en els bancs d'informacions internacionals. Disposem dels valors següents per a tres països: França, el Regne Unit i Alemanya.

	Articles de física	Articles de biologia	Articles de química
França	98	439	506
Regne Unit	220	259	311
Alemanya	508	498	380

a) Podem escollir la diferència absoluta entre el nombre total d'articles referenciats com a índex de no-semblança entre dos països?
b) Si usem la distància euclidiana com a índex de no-semblança, calculeu els diferents valors de distància entre els tres països. Agrupeu els països més propers.
c) Feu el mateix usant la distància de χ^2. Aquesta última es calcula sobre els percentatges de les files ponderant cada coordenada amb la inversa del nombre d'articles referenciats.
Quina és la funció de la ponderació?
d) Compareu els dos resultats.

Solucionari

Solucions als exercicis sobre les sèries

20.

a) Apliquem la llei de Bradford. El nucli r és igual a 10,4 (ja que $\frac{10.132}{127} = 10{,}4$) i la raó q és igual a 3,4. Així, doncs, el nombre de publicacions de cada grup es calcula amb la fórmula:

$$G_i = 10{,}4 \cdot 3{,}4^i \qquad i = 1, 2\ldots$$

Tenim:

$$G_1 = 10{,}4,\ G_2 = 10{,}4 \cdot 3{,}4 = 35{,}36 \text{ i } G_3 = 10{,}4 \cdot 3{,}4^2 = 120{,}22$$

com els efectius calculats. Aquests valors són molt propers als valors observats pel bibliotecari. Així, doncs, ell suposa que els articles del camp dels mitjans de comunicació es desglossen d'acord amb aquesta llei.

b) Amb 164 publicacions, tenim 396 articles. Volem conèixer el nombre de revistes necessàries per a obtenir el doble, és a dir, 792 articles. Mantenint aquest mateix model (grups amb aproximadament 132 articles), és necessari constituir 6 grups (en efecte, $\frac{792}{132} = 6$).

Aplicant la fórmula de Bradford, calculem el nombre total de revistes que cal analitzar:

$$\sum_{i=0}^{5} 10{,}4 \cdot 3{,}4^i = 10{,}4 \cdot \frac{1-3{,}4^6}{1-3{,}4} = 6.690 \text{ revistes}$$

Això no té sentit, ja que el nombre de revistes que cobreixen aquest camp no supera les 1.000. Així, doncs, mai no arribarem a comptar 792 articles en el camp dels mitjans de comunicació.

c) Si analitzem les 1.000 publicacions, el nombre de grups constituïts és:

$$10{,}4 \cdot \frac{1-3{,}4^n}{1-3{,}4} = 1.000, \text{ és a dir, } -3{,}4^n = \frac{1.000}{10{,}4} \cdot (-2{,}4) - 1\ .$$

És a dir,

$$3{,}4^n = 232 \text{ és a dir, } n = \frac{\mathrm{Ln}(232)}{\mathrm{Ln}(3{,}4)} = 4{,}45\ .$$

Si considerem $n = 4$, podem esperar trobar $4 \cdot 132 = 528$ articles.
Si considerem $n = 5$, podem esperar trobar $5 \cdot 132 = 660$ articles.
Quedem molt lluny dels 792 articles previstos en la pregunta anterior; no aconseguirem doblar el nombre d'articles comptats.

21.

a) Els resultats del recompte es presenten en la taula 1. El nombre total d'articles és 159.

Taula 1

Nre. d'articles	Nre. de publicacions	Total articles
26	1	26
20	1	20
19	1	19
10	1	10

Nre. d'articles	Nre. de publicacions	Total articles
6	2	12
5	2	10
4	3	12
3	4	12
2	7	14
1	24	24
		159

Proposem constituir quatre classes de revistes, cada una aproximadament amb 40 articles. Aquesta opció empírica sembla el millor equilibri possible entre una homogeneïtat de quantitats d'articles per classe i un gran nombre de classes. En efecte, si haguéssim creat grups més grans, per exemple amb unes classes de 65 documents, hauríem tingut una primera classe composta pels 26 + 20 + 19 = 65 primers articles, una segona classe composta pels 10 + 12 + 10 + 12 + 12 = 56 articles següents, i una tercera classe composta pels 14 + 24 = 38 últims articles. Aquest repartiment no és gens homogeni. Amb classes d'uns 80 articles, hauríem tingut un repartiment homogeni però format només per dues classes.

Així, doncs, el repartiment és el següent:

Taula 2

	Classes			
k	**1**	**2**	**3**	**4**
Revistes	Núm. 1 - núm. 2	Núm. 3 - núm. 6	Núm. 7 - núm. 16	Núm. 17 - núm. 46
Nombre d'articles	46	41	36	36
Nombre de revistes: G^k	2	4	10	30
Mitjana: (r_k)	23	10,25	3,6	1,2
$G^k \cdot r_k$	46	41	36	36

b) Sabent que, per a cada classe *k*, G^k és el nombre de revistes que conté i r_k el nombre mitjà d'articles per revista, podem dir que $G^k \cdot r_k$ representa el nombre d'articles de la classe *k*. Les classes s'han constituït de manera que tinguem un repartiment equitatiu dels articles. Així, doncs, podem dir que $G^1 r_1 = G^2 r_2 = G^3 r_3 = G^4 r_4$.

En conseqüència

$$G^2 = \frac{r_1}{r_2} G^1 = q_1 G^1$$

$$G^3 = \frac{r_2}{r_3} G^2 = q_2 G^2$$

$$G^4 = \frac{r_3}{r_4} G^3 = q_3 G^3$$

Així, doncs, tenim $q_1 = 2$, $q_2 = 2{,}5$ i $q_3 = 3$. Per a aplicar la llei de Bradford hem d'escollir un únic paràmetre q per a substituir els tres últims. Agafarem la mitjana de q_1, q_2 i q_3, és a dir, $q = 2{,}5$.

c) Totes les classes contenen el mateix nombre d'articles *m*. Llavors, si $R(k)$ és la suma acumulada del nombre d'articles per classe, tindrem $R(k) = km$. Triem el valor $m = 40$.

d) J_k és la suma acumulada de les revistes, per la qual cosa tenim:

$$J_k = G^1 + G^2 + \ldots + G^k$$

Per tant, $J_k = G^1 + qG^1 + \ldots + q^{k-1}G^1 = (1 + q + \ldots + q^{k-1})\, G^1$.

Trobem una sèrie geomètrica de raó q el primer terme de la qual és G^1.

Així, doncs, $J_k = \left(\frac{q^k - 1}{q - 1}\right) \cdot G^1$.

e) Hem demostrat que:

$$J_k = \left(\frac{q^k - 1}{q - 1}\right) \cdot G^1, \text{ és a dir, } q^k = \frac{J_k (q - 1)}{G^1} + 1 .$$

Si $t = \frac{G^1}{q - 1}$ (notacions de l'enunciat), llavors podem escriure:

$$q^k = \frac{J_k}{t} + 1$$

Passem als logaritmes. Tenim $k \cdot \log q = \log\left(\frac{J_k}{t} + 1\right)$.

És a dir, $R(k) \cdot \frac{\log q}{m} = \log\left(\frac{J_k}{t} + 1\right)$.

O bé, $j = \frac{m}{\log q}$.

Així, doncs, $R(k) = j \cdot \log\left(\frac{J_k}{t} + 1\right)$.

Usant la propietat de la funció logarítmica i substituint t pel seu valor, obtenim:

$$R(k) = j \cdot \log\left(\frac{J_k + t}{t}\right) = j \cdot \log(J_k + t) - j \cdot \log t = j \cdot \log\left(J_k + \frac{G^1}{q - 1}\right) - j \cdot \log\left(\frac{G^1}{q - 1}\right)$$

O bé, segons el resultat de la pregunta 5, tenim $\frac{J_k}{q^k - 1} = \frac{G^1}{q - 1}$.

És a dir, $R(k) = j \cdot \log\left(J_k + \frac{J_k}{q^k - 1}\right) - j \cdot \log\left(\frac{G^1}{q - 1}\right)$

o bé $\log\left(J_k + \frac{J_k}{q^k - 1}\right) = \log(J_k) + \log\left(1 + \frac{1}{q^k - 1}\right)$.

Així, doncs,

$$R(k) = j \cdot \log(J_k) + j \cdot \log\left(1 + \frac{1}{q^k - 1}\right) - j \cdot \log\left(\frac{G^1}{q - 1}\right) = j \cdot \log(J_k) + j \cdot \log\left(\frac{q - 1}{G^1} \times \frac{q^k}{q^k - 1}\right) .$$

Per tant, $R(k) = j \cdot \log(J_k) + B_k$.

f) El $R(k)$ observat correspon al nombre acumulat dels articles per classe; per exemple, per a k = 2, $R(k)$ és igual a 46 + 41 = 87. Els resultats de les altres classes es presenten en la taula 3.

Per calcular els $R(k)$ teòrics, apliquem la fórmula

$$R(k) = j \cdot \log(J_k) + B_k$$

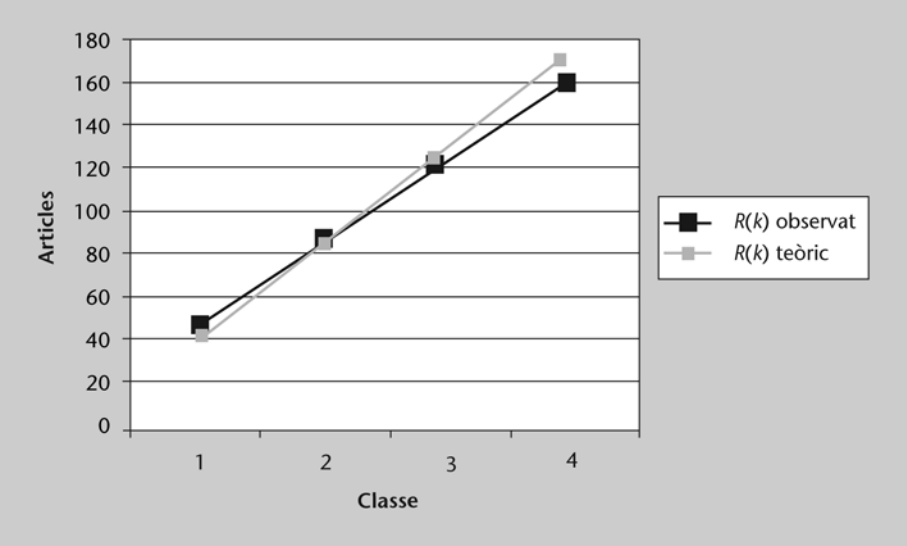
amb q = 2,5 i m = 40.

Taula 3

	Classes			
k	1	2	3	4
G^k observat	2	4	10	30
J_k observat	2	6	16	46
log (J_k) observat	0,30	0,78	1,20	1,66
Nre. d'articles observat	46	41	36	36
$R(k)$ observat	46	87	123	159
J_k teòric	2	7,4	21,98	61,35
log (J_k) teòric	0,30	0,87	1,34	1,78
B_k teòric	9,74	–4,95	–9,67	–11,43
$R(k)$ teòric	40	82,42	125,23	168,28

Les xifres s'han arrodonit a dos decimals.

Les corbes teòriques i experimentals es representen en la gràfica següent:



Constatem una bona alineació. La llei de Bradford amb els paràmetres $q = 2,5$ i $m = 40$ és, doncs, un bon model teòric per a aquest conjunt d'articles.

22.

a) Creem tres classes de revistes, cada classe amb un centenar d'articles. Aquesta opció és empírica, ja que és la que permet tenir el nombre més gran de classes per a un repartiment el més homogeni possible dels articles publicats (vegeu la taula 1).

Taula 1

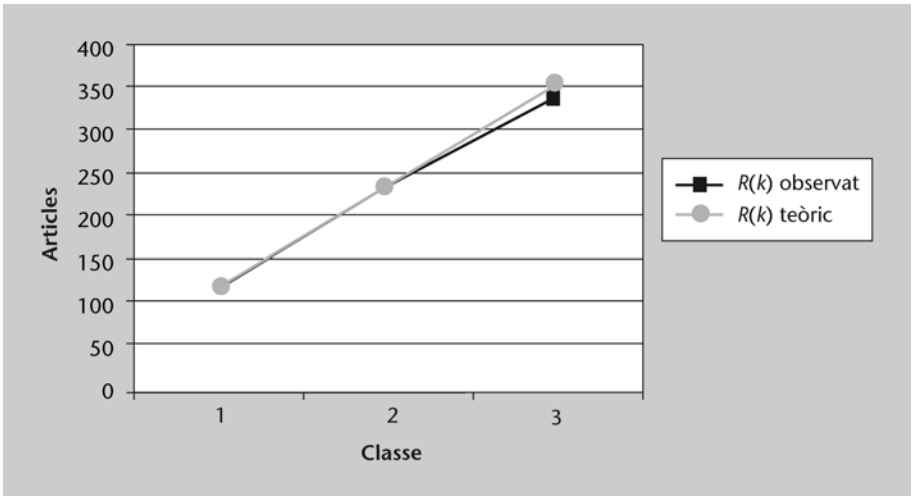
	Classes		
k	1	2	3
Números de les revistes	Núm. 1 - Núm. 2	Núm. 3 - Núm. 11	Núm. 12 - Núm. 46
G^k = nombre de revistes	2	9	35
Mitjana m_k	58	13,11	2,83
Nre. d'articles	116	118	99

Seguint els mateixos passos que en l'exercici 21, obtenim $q = 4{,}52$ i $m = 116$. Així, doncs, podem calcular els efectius teòrics. El conjunt dels resultats es representa en la taula següent:

Taula 2

	Classes		
k	**1**	**2**	**3**
G^k observat	2	9	35
Nre. d'articles observat	116	118	99
***R*(*k*) observat**	116	234	333
J_k observat	2	11	46
log (J_k) observat	0,30	1,04	1,66
J_k teòric	2	11,04	51,90
log (J_k) teòric	0,30	1,04	1,71
B_k teòric	62,70	47,33	44,31
***R*(*k*) teòric**	116	232	348

b) Les corbes teòrica i experimental de la relació entre $R(k)$ i $\log(J_k)$ es representen en la gràfica següent:

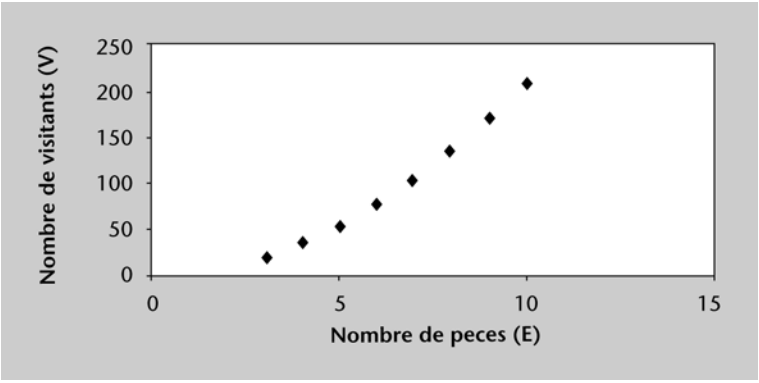


La dispersió de la literatura de biologia teòrica segueix una llei de Bradford.

23.

a) Representem en una gràfica la variació del nombre de visitants en funció del nombre d'obres visitades.

Relació entre V i E



La representació gràfica següent suggereix que les variables estan vinculades per una relació funcional de tipus potència (equació d'una paràbola).

$$V = 2 \cdot E^2$$

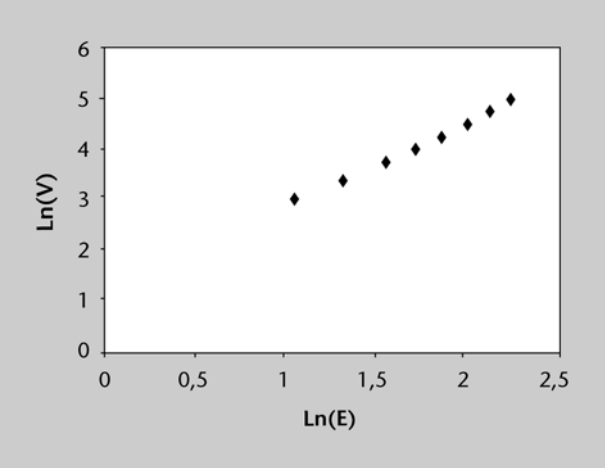
Verifiquem aquesta relació per mitjà del càlcul:

E	3	4	5	6	7	8	9	10
$2 \cdot E^2$	18	32	50	72	98	128	162	200
V	18	32	50	72	98	128	162	200

b) En la taula anterior es representen les transformacions logarítmiques (logaritme neperià) de les variables V i E.

ln(E)	1,10	1,39	1,61	1,79	1,95	2,08	2,20	2,30
ln(V)	2,89	3,47	3,91	4,28	4,58	4,85	5,09	5,30

Relació entre ln(V) i ln(E)



Constatem que els punts estan alineats. Per tant, podem escriure que hi ha una relació funcional lineal entre ln(V) i ln(E):

Busquem una relació de la forma següent: $\ln(V) = B + A \cdot \ln(E)$

Un càlcul simple de regressió lineal (fet sobre els valors arrodonits de la taula anterior) dóna, amb quatre xifres significatives:

$$A = 2{,}0028 \qquad B = 0{,}6862$$

Tornem a trobar la relació prèvia.

24.

En general, les distribucions d'usos són de tipus hiperbòlic, és a dir, que hi ha una relació del tipus $N_i = \frac{k}{U_i^a}$ $k > 0$, $a > 0$.

Per trobar els paràmetres a i k, calculem els logaritmes neperians:

$$\ln(N_i) = \ln(k) - a \cdot \ln(U_i).$$

Una regressió lineal permet calcular els paràmetres $\ln(k)$ i a.

U_i	$\ln(U_i)$	$\ln(N_i\text{Fr})$	$\ln(N_i\text{Us})$
5.000	8,52	3,85	3,47
17.500	9,77	3,00	3,30

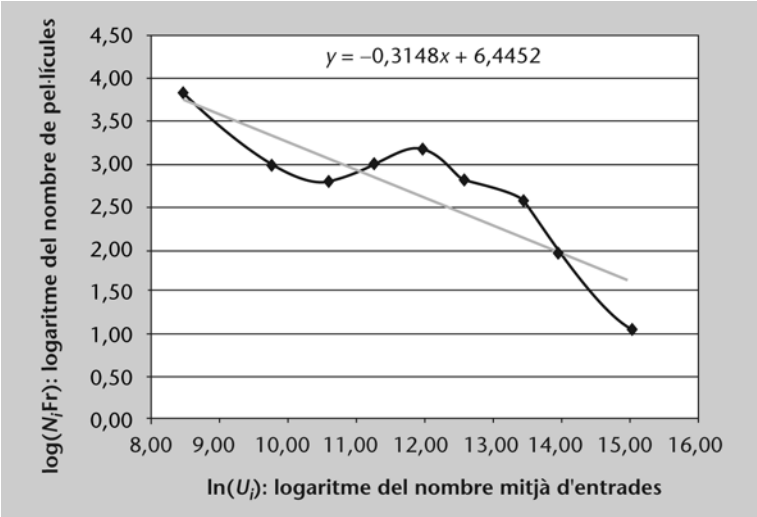
U_i	$\ln(U_i)$	$\ln(N_i\text{Fr})$	$\ln(N_i\text{Us})$
37.500	10,56	2,77	3,04
75.000	11,23	2,94	2,77
175.000	12,07	3,14	3,47
375.000	12,69	2,77	2,89
750.000	13,53	2,56	3,09
1.500.000	14,22	1,95	2,77
4.000.000	15,20	1,10	2,08

Per a les pel·lícules franceses, la recta de regressió corresponent té com a equació:
$y = -0{,}3148 \cdot x + 6{,}44$

Així, doncs, tenim $k = e^{6{,}44} = 626{,}40$ y $a = 0{,}31$.

I gràficament:

Figura 1. Audiència de les pel·lícules franceses



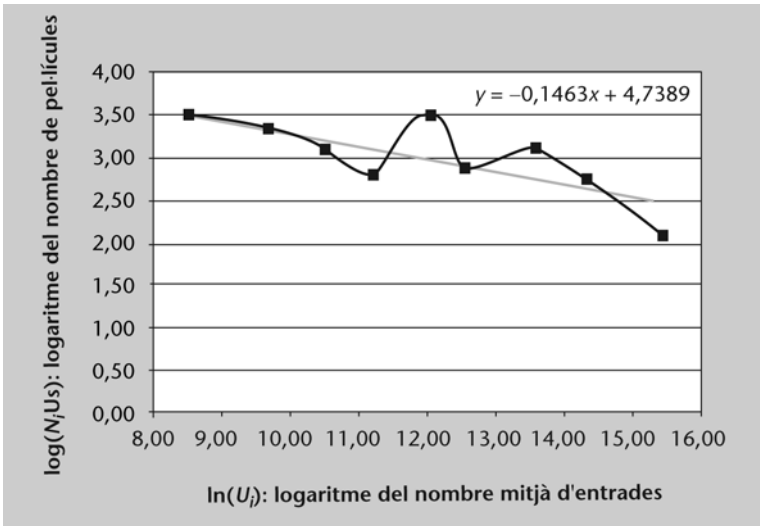
Per a les pel·lícules americanes, la recta de regressió té com a equació:

$$y = -0{,}1463 \cdot x + 4{,}73$$

És a dir, $k = e^{4{,}73} = 113{,}2$ i $a = 0{,}14$.

La representació gràfica és:

Figura 2. Audiència de les pel·lícules americanes



En les figures 1 i 2, si comparem les corbes de les audiències observades, designades com a N_i(Fr) i N_i(US), i les que resulten de l'ajust anterior, designades com a Teòrica (Fr) i Teòrica (US), observem que en el cas francès l'ajust per una sèrie hiperbòlica és relativament coherent, mentre que en el cas americà, vist l'escàs pendent (–0,14), una sèrie lineal és el més adequat. La raó és la diferència real de les produccions cinematogràfiques dels dos països. En efecte, a França hi ha unes poques pel·lícules que venen moltes entrades i moltes pel·lícules que en venen molt poques; ens trobem, doncs, prop dels resultats observats en bibliometria. Als Estats Units la situació és molt més homogènia. Igual que en el cas anterior, hi ha unes poques pel·lícules que venen moltes entrades, però la diferència és que hi ha moltes menys pel·lícules que venen poques entrades. La raó és el nombre més alt d'espectadors, però també els criteris de producció, que beneficien les grans productores en detriment dels realitzadors independents.

Figura 3. Audiència de les pel·lícules franceses, ajust per una sèrie hiperbòlica

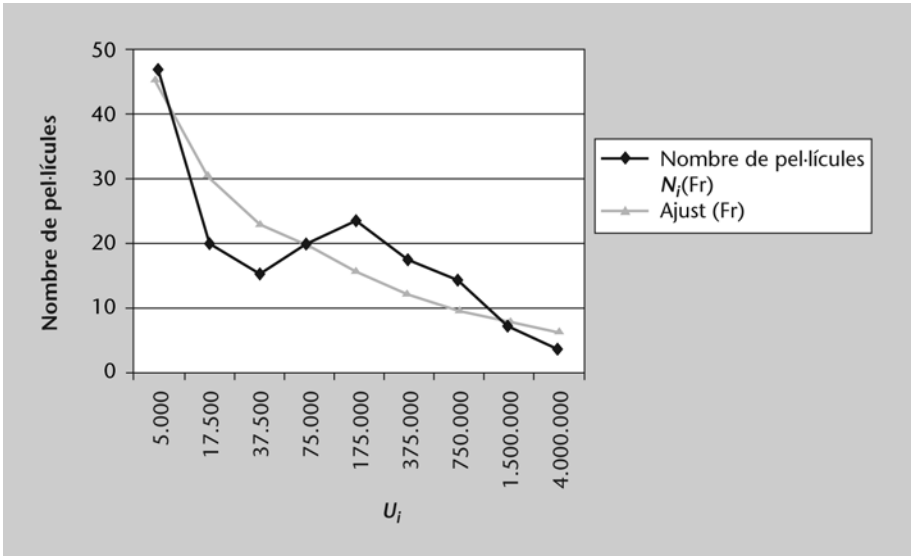
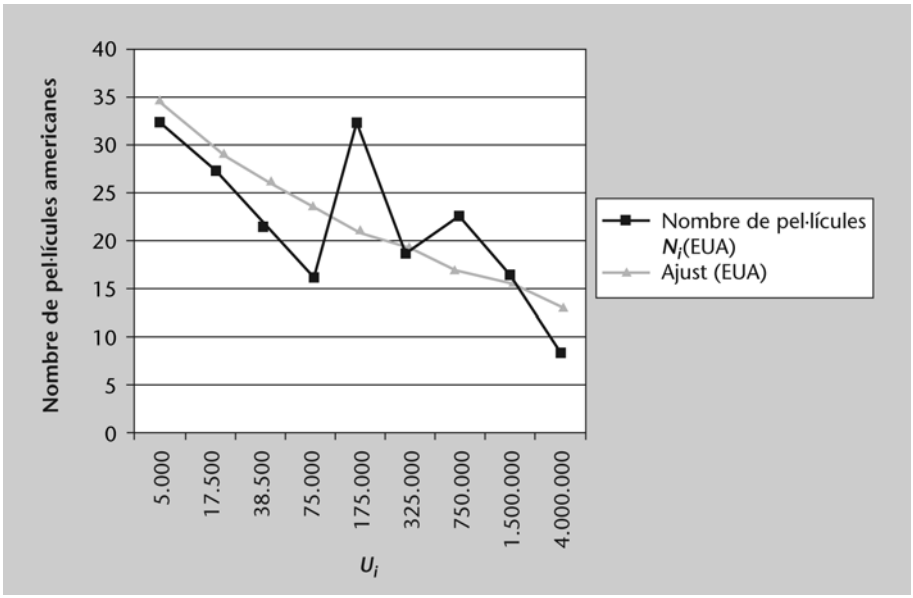


Figura 4. Audiència de les pel·lícules americanes, ajust per una sèrie hiperbòlica



Solucions als exercicis sobre les funcions

25.

a) Mètode numèric

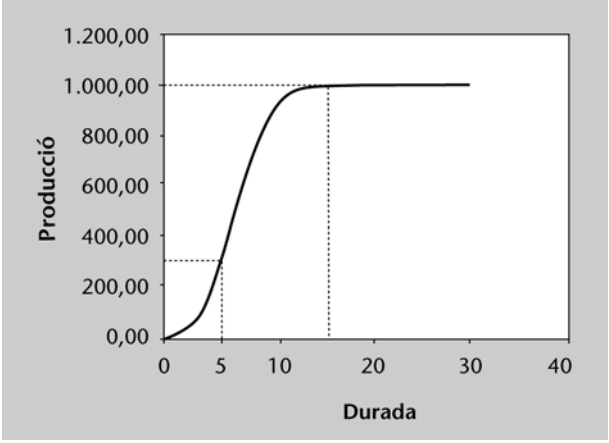
Substituint les constants pels seus valors, la funció C s'escriu:

$$C(t) = \frac{1.000}{\left(1 + 100e^{-0,8t}\right)}$$

Calculem el valor d'aquesta funció per a 50 punts escollits dins de l'interval [0, 30] amb l'ajuda d'una eina clàssica com un full de càlcul o algun programa de càlcul numèric. Els resultats són els que apareixen en la taula següent:

Durada (anys)	Producció	Durada (anys)	Producció
0	9,90	**6,1**	568,28
1	21,77	**6,2**	587,79
2	47,19	**6,3**	607,03
3	99,29	**6,4**	625,94
4	197,00	**6,5**	644,47
4,1	209,96	**6,6**	662,58
4,2	223,54	**6,7**	680,23
4,3	237,73	**6,8**	697,38
4,4	252,53	**6,9**	713,99
4,5	267,93	**7**	730,04
4,6	283,91	**8**	857,52
4,7	300,45	**9**	930,53
4,8	317,52	**10**	967,54
4,9	335,11	**11**	985,15
5	353,16	**12**	993,27
5,1	371,64	**13**	996,97
5,2	390,51	**14**	998,63
5,3	409,71	**15**	999,39
5,4	429,19	**16**	999,72
5,5	448,89	**17**	999,88
5,6	468,75	**18**	999,94
5,7	488,71	**19**	999,97
5,8	508,71	**20**	999,99
5,9	528,68	**22**	1.000,00
6	548,55	**23**	1.000,00

Hem escollit molts punts entre 4 i 6 per distingir millor en la gràfica la inflexió de la corba, que es dóna prop del punt $t = 5$ anys.



En el moment $t = 0$, $f(0) = \frac{1.000}{1+100} = 9{,}90$.

En el moment $t = 5$, $f(5) = 353{,}16$.

Quan t assoleix un valor molt alt, el numerador s'aproxima a 1 i la producció tendeix al valor 1.000.

En la gràfica i en la taula distingim clarament l'inici de la saturació, que té lloc prop del punt $t = 15$.

b) Mètode analític

Usem el càlcul diferencial per a trobar analíticament el punt d'inflexió de la corba i interpretar-lo.

Podeu consultar informació sobre el punt d'inflexió de la corba en l'annex 4 "Derivada i integral, exemple de resolució d'equacions diferencials".

- Càlcul de les derivades:

El càlcul de la derivada és $\frac{dC}{dt} = \frac{k \cdot a \cdot b \cdot e^{-bt}}{\left(1 + a \cdot e^{-bt}\right)^2}$; aquesta derivada és positiva, la qual cosa implica que C és una funció creixent. El càlcul de la derivada segona (derivada de la derivada) és:

$$\frac{d^2C}{d^2t} = \frac{k \cdot a \cdot b^2}{\left(1 + a \cdot b \cdot e^{-bt}\right)^3} e^{-bt}\left(ae^{-bt} - 1\right)$$

Aquests dos càlculs es poden fer manualment usant les regles que es donen en un manual de matemàtiques o amb un programa informàtic de matemàtiques.

- Càlcul del punt d'inflexió:

La corba té un punt d'inflexió quan la derivada segona s'anul·la. Per a conèixer aquest punt n'hi ha prou de resoldre l'equació següent:

$$ae^{-bt} - 1 = 0$$

$ae^{-bt} - 1 = 0$ únicament si $e^{-bt} = \frac{1}{a}$

$$e^{-bt} = \frac{1}{a}$$

únicament si $t = \frac{\ln(a)}{b}$.

Substituint les constants pel seu valor, obtenim $t = 5{,}75$ anys.

La derivada segona és positiva per a $t < \frac{\ln(a)}{b} = 5{,}75$ i negativa per a $t > \frac{\ln(a)}{b} = 5{,}75$, la qual cosa implica que la derivada és creixent per a $t < 5{,}75$ i decreixent per a $t > 5{,}75$; això significa que per a $t > 5{,}75$, la velocitat de creixement de la producció començarà a disminuir.

- Càlcul del límit:

Quan estudiem el límit d'aquesta funció, obtenim:

Tenim $\lim_{t\to\infty}\left(e^{-bt}\right) = 0$, i per això $\lim_{t\to\infty}\left(\frac{k}{1 + a \cdot e^{-bt}}\right) = k$.

Substituint k pel seu valor, podem dir que el límit és 1.000.

26.

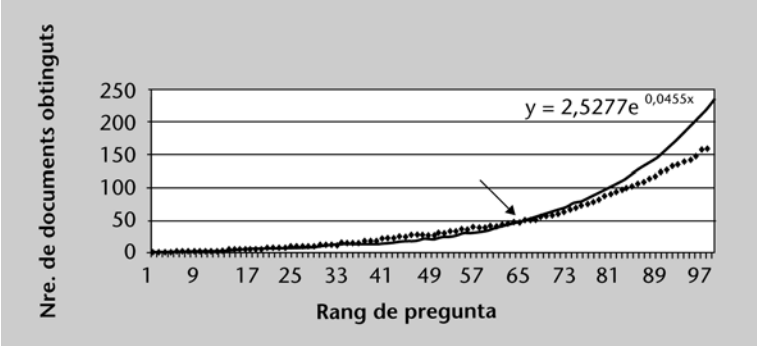
a) Per a observar una regularitat és necessari classificar les respostes obtingudes segons el nombre creixent de documents i dibuixar la corba de variació:

1	7	21	42	83
1	7	22	44	87
1	8	23	45	90

1	9	24	47	94
2	9	25	48	97
2	10	26	50	101
2	11	27	51	103
2	12	28	52	105
3	12	29	53	108
3	12	29	54	115
3	13	29	55	118
4	14	30	58	123
4	15	31	59	126
5	15	33	61	132
5	16	35	65	135
6	17	36	67	138
6	18	38	70	142
7	19	39	73	148
	19	40	75	156
	20	42	79	160

La corba té la forma d'una funció exponencial d'equació $y = be^{ax}$. Per a trobar els paràmetres a i b de la funció exponencial que millor l'ajusta, fem una regressió lineal després de la transformació logarítmica. Obtenim $a = 0{,}0455$ i $b = 2{,}5277$ (gràfica següent).

Figura 1. Variació del nombre de documents obtinguts



b) A partir de 50 documents (vegeu la fletxa), els valors obtinguts són inferiors als valors calculats. Això s'explica perquè el banc consultat té una mida finita; per més que plantegem unes preguntes cada cop més generals, les respostes no poden superar mai el nombre de documents que es troben al banc.

27.

a) La inversa de la taxa de rotació d'una obra correspon a la relació:

$$k = \frac{F}{U} = \frac{\sum_{i=1}^{N} F_i}{\sum_{i=1}^{N} U_i}$$

Això informa del grau general d'activitat d'una biblioteca.

Com que la col·lecció inclou N categories d'obres, és possible calcular per a cadascuna la inversa de la taxa de rotació particular característica del seu grau d'activitat propi. Però també podem aplicar a cada categoria el grau d'activitat general. Així, doncs, per a una categoria:

$$F_i = k \cdot U_i, \text{ si } i \text{ varia d'1 a } N.$$

Ens permet calcular el nombre d'obres necessàries dins d'una categoria concreta si coneixem el nombre de préstecs i el grau d'activitat general de la biblioteca.

Podem destacar que si hi ha una única categoria, tenim:

$$F_i = \frac{F}{U} \cdot U_i = F$$

Aquesta fórmula correspon a un desglossament de les compres, ja que el nombre d'obres previstes continua essent constant:

$$\sum_{i=1}^{N} F_i = \sum_{i=1}^{N} \frac{F}{U} U_i = \frac{F}{U} \cdot \sum_{i=1}^{N} U_i = F$$

b) La taxa de rotació de la col·lecció és $\frac{U}{F} = \frac{1.314}{1.194} = 1{,}10$. Així, doncs, aplicant la fórmula anterior, podem calcular els objectius per al 2001.

Destacarem que només hem de comprar obres per a les categories D i G, la taxa de rotació de les quals és superior a 1. Quan la taxa és inferior a 1 és necessari preveure una reducció de les compres (categories A, B, G, E i F).

c) L'exponent *x* és inferior a 1, per la qual cosa el nombre d'obres deixades en préstec disminueix. La taxa de rotació ponderada és, doncs, $\frac{218}{1.194} = 0{,}18$.

Categoria	Obres (F_i)	Préstecs (U_i)	Taxa de rotació	Objectiu 2001 (F_i)	Desviacions
A	241	230	0,95	209	–32
B	216	144	0,67	131	–85
C	100	99	0,99	90	–10
D	226	273	1,21	248	+22
E	129	109	0,84	99	–30
F	43	37	0,86	34	–9
G	239	422	1,77	383	+144

Categoria	Obres (F_i)	Préstecs (U_i)	Préstecs ponderats ($U_i^{2/3}$)	Objectiu 1993 (F_i)	Desviacions
A	241	230	37,54	205	–36
B	216	144	27,47	150	–66
C	100	99	21,40	117	+17
D	226	273	42,08	230	–4
E	129	109	22,82	125	+4
F	43	37	11,10	61	–18
G	239	422	56,26	307	+68

Els resultats mostren que l'exponent *x* té un paper de ponderació. Les desviacions són més homogènies. No hi ha cap increment sistemàtic per a les categories en les quals la rotació és superior a 1. Només la categoria G s'incrementa, mentre que la categoria D ja no ho fa. Per contra, les categories G i E augmenten quan tenen una taxa de rotació inferior a 1.

28. Els valors de les variables varien en sentit invers, per la qual cosa pensem en una funció de potència inversa del tipus:

$$G = \frac{k}{U^a}$$

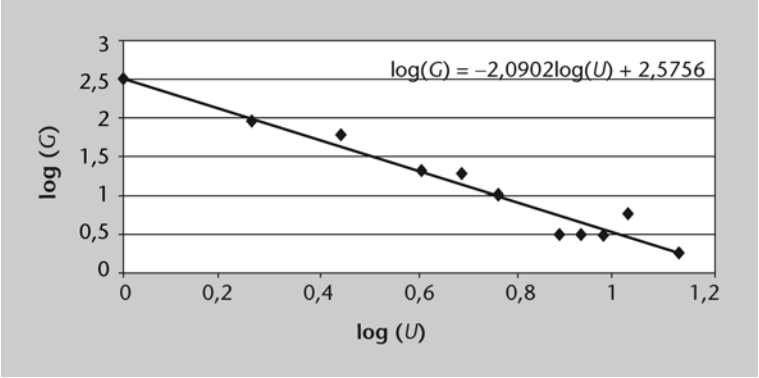
Un ajust lineal per pas a logaritmes permetrà calcular *k* i *a*.

$$\log(G) = \log\left(\frac{k}{U^a}\right) = \log(k) - \log\left(U^a\right) = \log(k) - a \cdot \log(U)$$

Els valors dels logaritmes en base 10 de la variable del nombre d'articles (representada per U) i de la variable del nombre d'autors (representada per G) són:

log(U)	log(G)
0	2,5302
0,3010	1,9294
0,4771	1,7243
0,6020	1,3424
0,6990	1,2553
0,7781	1
0,845	0,4771
0,903	0,4771
0,9542	0,4771
1	0,7782
1,1139	0,301

La corba corresponent és, doncs:



Les variacions es representen mitjançant una recta d'equació:

$$\log(G) = -2{,}09 \cdot \log(U) + 2{,}57$$

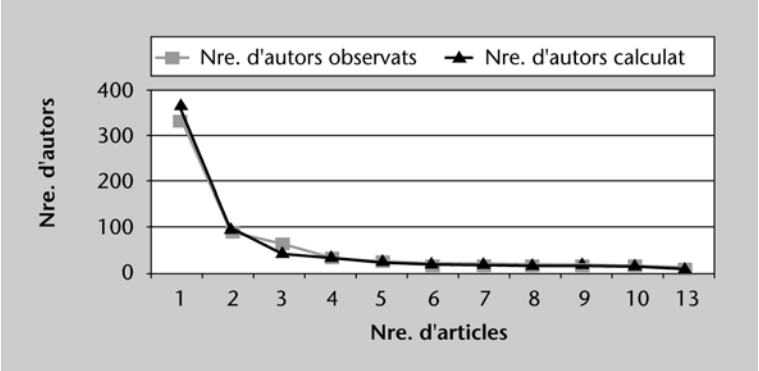
en què $a = 2{,}09$ i $\log(k) = 2{,}57$. Llavors, $k = 371{,}54$.

La distribució observada la descriu la funció de potència:

$$G = \frac{371{,}54}{U^{2{,}09}}$$

Això segueix d'una manera molt estreta una llei de Lotka en què $a = 2$.

Les distribucions dels valors observats i calculats es representen mitjançant les corbes següents:



Solucions als exercicis sobre les equacions

29. Substituint les variables pels seus valors en l'equació proposada, obtenim un sistema de dues equacions amb dues incògnites:

$$\text{Tenim } \begin{cases} 10 = c + k \cdot \log_2(4) \\ 20 = c + k \cdot \log_2(6) \end{cases}.$$

Resolució:

$$\begin{cases} c = 10 - k \cdot \log_2(4) \\ 20 = 10 - k \cdot \log_2(4) + k \cdot \log_2(6) \end{cases}$$

$$\begin{cases} c = 10 - k \cdot \log_2(4) \\ 10 = k \cdot (\log_2(6) - \log_2(4)) \end{cases}$$

$$\begin{cases} c = 10 - k \cdot \log_2(4) \\ k = \dfrac{10}{\log_2(6) - \log_2(4)} \end{cases}.$$

Així, doncs, $\begin{cases} k = 17{,}09 \\ c = -24{,}19 \end{cases}$.

Per a aquest usuari, l'equació del seu temps de reacció serà:

$$t = -24{,}19 + 17{,}09 \cdot \log_2(b)$$

30.

a) Una taxa d'ocupació d'1 erlang correspon a una línia totalment assignada. En conseqüència, una taxa d'ocupació $E > 1$ indica una línia que no està totalment ocupada.
b) Per a la companyia de serveis informatius, $E = 0{,}6$ erlang.
Per a un servei de gestió d'estocs documentals informatitzats, $E = 0{,}15$ erlang.
Per a un banc d'informacions, $E = 0{,}5$ erlang.
La línia més ocupada és la de la companyia de serveis.

31.

a) El nombre d'articles nous citats és proporcional a l'interval de temps que es té en compte Δt, al nombre d'articles no citats, $N - U(t)$ i un factor $\alpha(t)$.

Així, doncs, podem escriure:

$$U(t + \Delta t) - U(t) = \Delta t \cdot (N - U(t)) \cdot \alpha(t)$$

Després de la transformació, tenim:

$$\frac{U(t + \Delta t) - U(t)}{N \cdot \Delta t} = \left(1 - \frac{U(t)}{N}\right) \cdot \alpha(t)$$

És a dir:

$$\frac{R(t + \Delta t) - R(t)}{\Delta t} = (1 - R(t)) \cdot \alpha(t) = (1 - R(t)) \cdot Ae^{-at}$$

Si fem tendir Δt a 0 i apliquem la definició de la derivada, tenim:

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = A \cdot e^{-at} \cdot (1 - R(t))$$

Després de separar les variables R i t, obtenim:

$$\frac{\mathrm{d}R}{1 - R} = A \cdot e^{-at}\mathrm{d}t$$

Integrant els dos membres de l'equació:

$$\int \frac{\mathrm{d}R}{1-R} = \int A \cdot e^{-at}\mathrm{d}t$$

La integral de $\frac{\mathrm{d}R}{1-R}$ és $-\ln(1 - R(t))$ i la de $A \cdot e^{-at}\,\mathrm{d}t$ és $-\frac{A}{a} \cdot e^{-at} + C$, en què C és una constant, per la qual cosa podem escriure:

$$\ln\left(1 - R(t)\right) = \frac{A}{a} \cdot e^{-at} + C$$

Llavors, $R(t) = 1 - K \cdot e^{\frac{A}{a} \cdot e^{-at}}$, en què K és una constant tal que $K = e^C$.

Si plantegem $b = e^{\frac{A}{a}}$ (b > 1, ja que $\frac{A}{a} > 0$), llavors tenim la solució:

$$R(t) = 1 - K \cdot \ b^{e^{-at}}$$

Per a $t = 0$ $R(0) = 1 - K \cdot b$, $K = \frac{1 - R(0)}{b}$.

Així, doncs, $R(t) = 1 - \left(\frac{1 - R(0)}{b}\right) b^{e^{-at}}$ és la solució de l'equació $\frac{\mathrm{d}R}{1-R} = A \cdot e^{-at}\mathrm{d}t$.

b) Quan t tendeix a l'infinit, tenim $\lim_{t \to \infty} e^{-at} = 0$, ja que $a > 0$.

Així, doncs:

$$\lim_{t \to \infty} b^{e^{-at}} = 1 \quad \text{i} \quad \lim_{t \to \infty} R(t) = 1 - \frac{1 - R(0)}{b}$$

Com que b és més gran que 1, tenim $\frac{1 - R(0)}{b} < 1$.

Això significa que sempre, fins i tot al final d'un període de temps llarg, hi haurà una proporció d'articles no citats.

32. Després d'haver separat les variables u i G, tenim:

$$\frac{\mathrm{d}G}{G} = -a\frac{\mathrm{d}u}{u}$$

Així, doncs, integrant:

$$\ln(G) + C = -a\ln(u)$$

En què C és una constant, és a dir:

$$G = \frac{K}{u^a}$$

En què K és una constant.

Reconeixem la llei de Lotka amb un exponent que és igual a $a \cdot K$ és una constant que caracteritza el camp.

Solucions als exercicis sobre els conjunts

33.

a) Avaluem la consulta *A* Y NO *B:*

A	*B*	No *B*	*A* Y NO *B*
1	0	1	1
0	1	0	0
1	1	0	0
0	0	1	0

Veiem que l'última columna (*A* Y NO *B*) d'aquesta taula és idèntica a l'operació *A* EXCEPTE *B*, la qual cosa demostra l'equivalència de totes dues consultes.

b) Demostrem que, per a qualsevol dels valors donats (1 o 0) a *A*, *B* o C, es compleix la igualtat següent:

$$(A \text{ Y } B) \text{ O } (A \text{ Y } C) = A \text{ Y } (B \text{ O } C)$$

Amb l'ajuda de les taules de veritat, estudiem tots els casos possibles:

A	*B*	*C*	*A* Y *B*	*A* Y *C*	*B* O *C*	(*A* Y *B*) O (*A* Y *C*)	*A* Y (*B* O *C*)
1	**1**	**1**	1	1	1	1	1
1	**1**	**0**	1	0	1	1	1
1	**0**	**1**	0	1	1	1	1
0	**1**	**1**	0	0	1	0	0
1	**0**	**0**	0	0	0	0	0
0	**1**	**0**	0	0	1	0	0
0	**0**	**1**	0	0	1	0	0
0	**0**	**0**	0	0	0	0	0

Veiem que les dues últimes columnes són idèntiques, la qual cosa demostra que les dues consultes són equivalents.

34.

a) La consulta ens permet construir la taula de contingència següent:

	G_0 = pertinent	G_1 = no pertinent
U **= extret**	28	35
No extret	134 (162 – 28)	814 (849 – 35)

Per a mesurar la similitud entre *U* i G_0 escollim l'índex de Dice.

$$D(U,G_0) = 2 \cdot \frac{|U \cap G_0|}{|G_0| + |U|} = 2 \cdot \frac{28}{162 + (28 + 35)} = 0{,}25$$

b) La relació és el nombre de documents pertinents extrets sobre el nombre de documents pertinents, per la qual cosa es definirà amb:

$$R = \frac{|G_0 \cap U|}{|G_0|}$$

La precisió és el nombre de documents pertinents sobre el nombre total de documents extrets, per la qual cosa es definirà amb:

$$PR = \frac{|G_0 \cap U|}{|U|}$$

Numèricament, la relació i la precisió són:

$$R = \frac{28}{162} = 0,17 \quad PR = \frac{28}{63} = 0,44$$

c)

$$\frac{1}{D(U,G_0)} = \frac{|U| + |G_0|}{2|U \cap G_0|} = \frac{|U|}{2|U \cap G_0|} + \frac{|G_0|}{2|U \cap G_0|} = \frac{1}{2R} + \frac{1}{2P} = 2,94 + 1,14 = 4,07 \simeq \frac{1}{0,25}$$

35.

a) L'espai vectorial considerat té tres dimensions, cada una relativa a una de les paraules informatives. Els llocs es representen amb punts (figura 1) o bé amb vectors (figura 2). Les seves coordenades són les ocurrències de cada paraula informativa representada pels eixos.

La pregunta també es representa amb punts (figura 1) o bé amb el vector unitari $U(1,1,1)$ (figura 2). En efecte, considerem que cadascuna de les paraules *llei*, *repressió* i *frau* apareix el mateix nombre de vegades en la pregunta.

Figura 1. Llocs i pregunta representada per punts

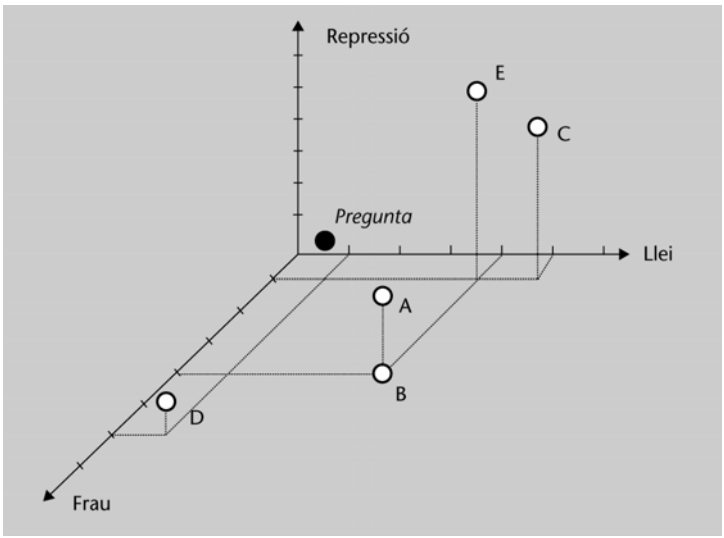
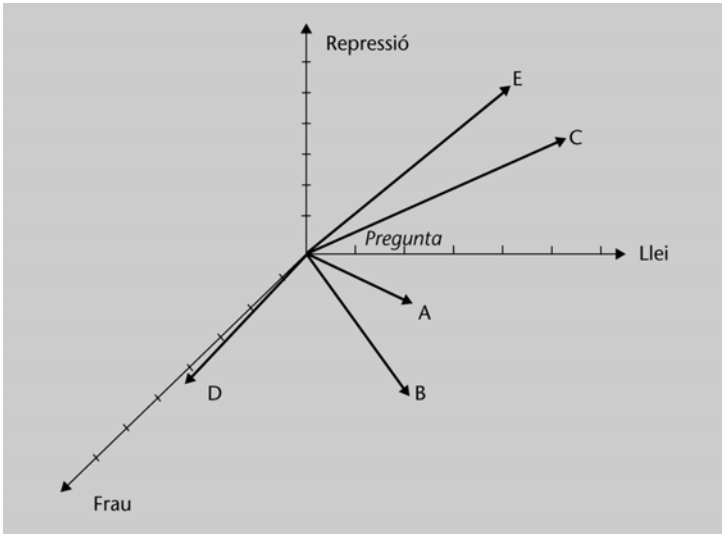


Figura 2. Llocs i pregunta representada per vectors



b) És molt difícil expressar gràficament quin és el lloc més proper a la pregunta. Per a determinar-ho, Salton usa l'angle d'inclinació dels vectors *A*, *B*, *C*, *D* i *E* amb el vector unitari *U*. Com més gran és el cosinus, més petit és l'angle entre els vectors.

$$\cos(U,A) = \frac{\sum_{k=1}^{t} a_k u_k}{\sqrt{\sum_{k=1}^{t} (a_k)^2}\sqrt{\sum_{k=1}^{t} (u_k)^2}}$$

O bé, $(u_1,u_2,u_3) = (1,1,1)$

Així, doncs, $\cos(U,A) = \frac{\sum_{k=1}^{t} a_k}{\sqrt{3 \cdot \sum_{k=1}^{t} (a_k)^2}}$, és a dir, cos(*U*,*A*) = 0,96.

De la mateixa manera, també calculem:

cos (*U*,*B*) = 0,82
cos (*U*,*C*) = 0,89
cos (*U*,*D*) = 0,75
cos (*U*,*E*) = 0,89

El cosinus entre *U* i *A* és el més gran, per la qual cosa el lloc *A* és el que millor respon a la pregunta.

36.

a) Aquest índex no és prou precís. En efecte, un valor nul significa que els dos països han publicat globalment el mateix nombre d'articles. No obstant això, no és necessari que hagin publicat el mateix nombre d'articles dins de cada disciplina. Així, doncs, no podem dir que hi ha una no-semblança nul·la (o una semblança completa).
b) L'espai vectorial considerat és de tres dimensions, i cada una correspon a les tres disciplines.

Tenim:

	Distància
d (França, Regne Unit)	292,08
d (Regne Unit, Alemanya)	380,56
d (França, Alemanya)	432,96

d (França, Regne Unit) < d (Regne Unit, Alemanya) < d (França, Alemanya)

Aquí, la classificació prevista és {França, Regne Unit}, Alemanya

c) Per començar, calculem els percentatges de les files de la taula. Per exemple, els 98 articles de física publicats a França corresponen al 9,40% del total dels 1.043 articles publicats a França. En l'espai considerat es defineix una dimensió per a cada tipus d'articles. Les entitats representades són els països, per la qual cosa les seves coordenades seran els percentatges de cada tipus d'articles.

	Articles de física %	**Articles de biologia %**	**Articles de química %**
França	9,40	42,09	48,51
Regne Unit	27,85	32,78	39,37
Alemanya	36,65	35,93	27,42

En cada eix fem una ponderació que és igual a l'invers de la suma d'articles referenciats en una disciplina. Com més gran sigui el nombre d'articles, més baix serà el coeficient de ponderació en l'eix.

Tenim:

$$\text{Física: } \frac{1}{826} = 0{,}0012$$

$$\text{Biologia: } \frac{1}{1.196} = 0{,}0008$$

$$\text{Química: } \frac{1}{1.197} = 0{,}0008$$

Això vol dir que concedirem més importància a la diferència d'articles referenciats en biologia que en física, ja que el total d'articles referenciats és més baix. La taula següent descompon els resultats i mostra la contribució de cada eix al càlcul de la distància.

	Contribució física	Contribució biologia	Contribució química	Distància
França / Regne Unit	0,4122	0,0724	0,0699	0,7446
Regne Unit / Alemanya	0,0938	0,0083	0,1193	0,4706
França/Alemanya	0,8994	0,0317	0,3718	1,1415

distància (Regne Unit/Alemanya) < distància (França/Regne Unit) < distància (França/Alemanya)

Aquí, la classificació prevista és {Alemanya, Regne Unit}, França.

El Regne Unit i Alemanya són els que estan més propers, ja que produeixen de mitjana el mateix nombre d'articles dins de cada disciplina; tenen el mateix "perfil científic". Per contra, França i Alemanya són els dos països més allunyats, ja que França és molt feble en física en comparació d'Alemanya, mentre que en química França és molt forta en comparació d'Alemanya.

Aquesta distància mesura els perfils, mentre que la distància anterior mesurava les diferències de producció.