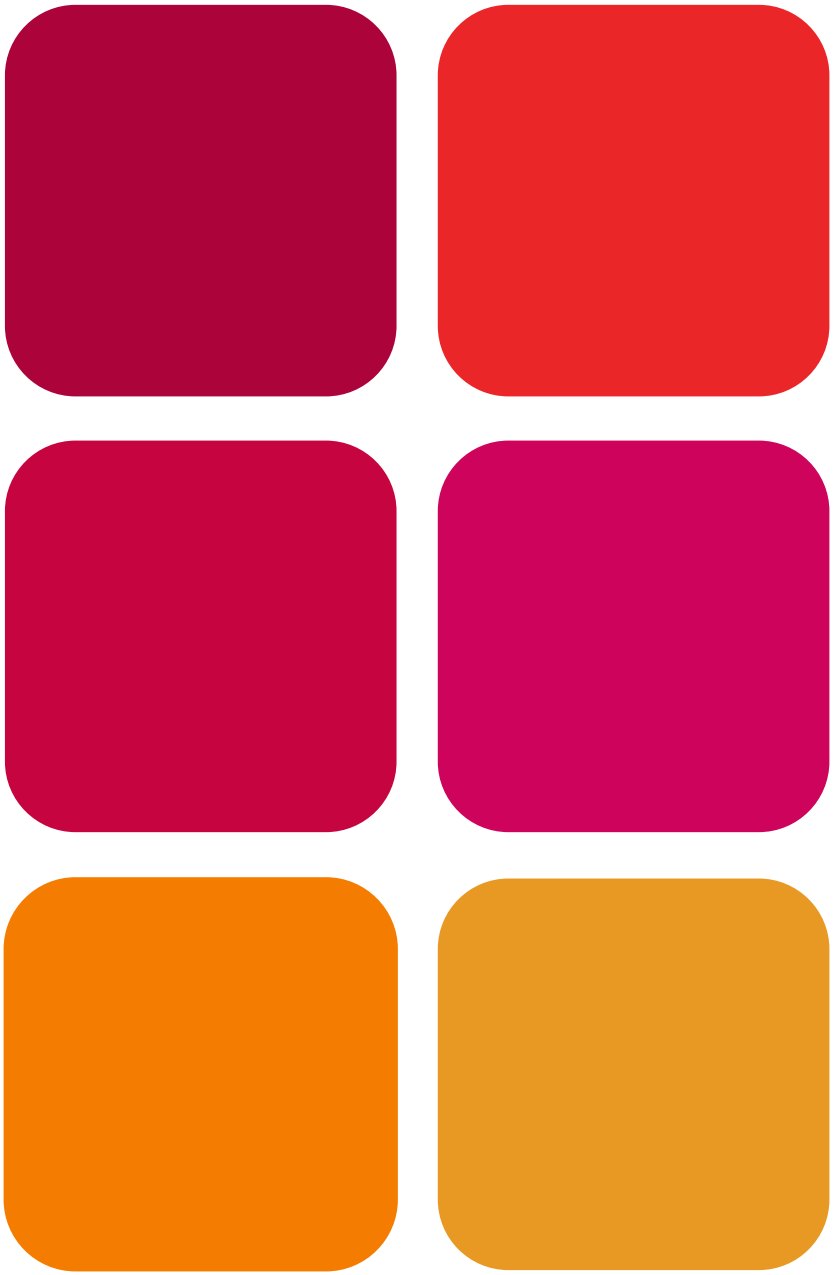


www.sapientia.uji.es | 21

UNIVERSITAT JAUME·I

Estadística. Volum I

Joaquín Castelló Benavent
María Victoria Ibáñez Gual
Vicente Martínez García
Amelia Simó Vidal

Estadística.
Volum I

Joaquín Castelló Benavent
María Victoria Ibáñez Gual
Vicente Martínez García
Amelia Simó Vidal

Departament de Matemàtiques

■ Codi d'assignatura C23

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
http://www.tenda.uji.es e-mail: publicacions@uji.es

Col·lecció Sapientia, 21
Primera edició, 2010
www.sapientia.uji.es

ISBN: 978-84-692-9048-4

ÍNDEX

TEMA 1

DESCRIPCIÓ D'UNA MOSTRA: DISTRIBUCIONS DE FREQÜÈNCIES I MÈTODES GRÀFICS

1.1. INTRODUCCIÓ

Atés l'aspecte aplicat que fonamentalment té l'estadística, començarem amb alguns exemples:

Exemple 1.1 La regidoria de Benestar Social d'una determinada ciutat desitja esbrinar si la mitjana de fills per família ha baixat respecte a la dècada anterior. Per a aquest fi, ha enquestat 50 famílies i n'ha obtingut les dades següents:

2	3	4	3	2	4	3	5	1	2	2	0	4	3	2	2	3	1	0	2	2	3	2	2	2
2	3	3	2	1	6	4	2	2	3	3	2	2	2	4	3	3	2	3	3	2	3	2	4	1

Exemple 1.2 Una cadena hotelera té la intenció d'obrir un nou hotel en una determinada ciutat. Abans de decidir el preu de les habitacions, el gerent de la cadena investiga els preus per habitació de 40 hotels de la mateixa categoria de la dita ciutat. Les dades obtingudes, en euros, varen ser:

39	49	37	56	43	49	50	61	51	45
53	39	43	50	60	47	51	42	44	58
33	43	41	58	44	38	61	43	53	45
40	54	39	47	33	45	47	42	45	48

L'**estadítica** és la ciència que s'encarrega de la recopilació, la representació i l'ús de dades sobre una o diverses característiques d'interés per prendre decisions o extraure conclusions generals a partir d'aquestes dades.

El **mètode estadístic** consta dels passos següents:

Pas 1. Plantejament del problema en termes precisos: àmbit d'aplicació (*població*) i característica (o característiques) objecte d'estudi (*variable(s)*).

Pas 2. Recollida de dades de la població d'interés (*mostreig*).

Pas 3. Organització, presentació i resum de les dades o de la mostra (*estadística descriptiva*).

Pas 4. Models matemàtics (*teoria de la probabilitat*).

Pas 5. Obtenció de conclusions generals o verificació d'hipòtesi (*inferència estadística*).

L'**estadística descriptiva** és la part de l'estadística que s'encarrega d'organitzar, resumir i donar una primera descripció (sense obtindre conclusions generals) de les dades obtingudes en el mostreig.

1.2. CONCEPTES BÀSICS

Anomenarem **població** el conjunt d'individus o ents subjectes a estudi (en l'exemple 1.1, el conjunt de totes les famílies de la ciutat; en l'exemple 1.2, el conjunt de tots els hotels d'aquesta categoria en la dita ciutat). Algunes poblacions són finites i poden conéixer-se (el conjunt de tots els hotels), altres són infinites o abstractes (el conjunt de totes les peces fabricades per una màquina).

Anomenarem **variable** la característica que volem estudiar en la població (en el primer exemple, el nombre de fills; en el segon, el preu per habitació). Les denotarem mitjançant lletres majúscules: X, $Y\ldots$ Podem classificar les variables en dos grans grups, les variables qualitatives i les variables quantitatives.

Les **variables qualitatives** són aquelles que no es poden mesurar, és a dir, aquelles que prenen valors als quals no es pot assignar cap número. Expressen qualitats o categories; per exemple: sexe, professió, color dels ulls, etc.

Les **variables quantitatives**, al contrari, són mesurables, és a dir, els valors que s'hi observen poden expressar-se de forma numèrica. Aquestes variables poden classificar-se en:

Discretes, quan prenen els seus valors en un conjunt finit o numerable. Per exemple, el nombre de fills, el nombre d'obrers en una fàbrica, les vegades que ix cara en llançar una moneda 10 vegades, etc.

Contínues, quan poden prendre qualsevol valor en un interval. Per exemple, el pes, l'estatura, etc.

Nota 1.1 La distinció entre variables discretes i variables contínues és més teòrica que pràctica, ja que les limitacions en els aparells de mesura fan que totes les variables quantitatives es comporten com a discretes quan es pretén observar-les. Aquesta distinció serà important en els models teòrics, quan estudiem la part de teoria de la probabilitat. De moment, farem més flexible el concepte de variable contínua considerant contínua aquella variable que pren un gran nombre de valors diferents. En aquest sentit, podem considerar la variable preu com a contínua.

Anomenarem **mostra** un subconjunt finit d'elements seleccionats entre els de la població. Per exemple, les 50 famílies del primer exemple o els 40 hotels del segon. El nombre d'observacions de la mostra l'anomenarem **grandària mostral**. Normalment el denotarem per n.

Anomenarem **dada** cada valor observat de la variable. Si la variable la representem per X, cada dada diferent de la mostra la representarem per x_i. El subíndex i indica el lloc que la dada ocupa en la mostra, quan totes les dades diferents s'han ordenat de més xicoteta a més gran. En l'exemple 1.1: $x_1 = 0$, $x_2 = 1$ ... En l'exemple 1.2: $x_1 = 33$, $x_2 = 37$ ...

1.3. DISTRIBUCIONS DE FREQÜÈNCIES

Si observem les dades dels exemples anteriors, és obvi que el primer pas en l'organització de les dades consistirà a agrupar aquelles que es repeteixen. Per a aquest propòsit establim les definicions següents:

Definició 1.1 La **freqüència absoluta** (n_i) d'un valor x_i de la variable és el nombre de vegades que aquest valor es repeteix en la mostra.

Propietat 1.1 *La suma de totes les freqüències absolutes és la grandària mostral:* $\sum n_i = n$.

Definició 1.2 La **freqüència relativa** (f_i) d'un valor x_i de la variable és el quocient entre la freqüència absoluta del valor i la grandària mostral: $f_i = \dfrac{n_i}{n}$.

Propietat 1.2 *La suma de totes les freqüències relatives és la unitat.*

Definició 1.3 La **freqüència absoluta acumulada** (N_i) d'un valor x_i de la variable és el nombre de dades en la mostra iguals o inferiors a x_i. Es calcula com $N_i = \sum_{k=1}^{i} n_k = N_{i-1} + n_i$.

Propietat 1.3 *L'última freqüència absoluta acumulada és la grandària mostral.*

Definició 1.4 La **freqüència relativa acumulada** (F_i) d'un valor x_i de la variable és el quocient entre la freqüència absoluta acumulada del valor i la grandària mostral. Es calcula com $F_i = \dfrac{N_i}{n} = \sum_{k=1}^{i} f_k$.

Propietat 1.4 *L'última freqüència relativa acumulada és la unitat.*

Definició 1.5 Una **distribució de freqüències** d'una variable és una taula que conté els diferents valors de la variable, sense repetir-los, ordenats de més baix a més alt amb les freqüències corresponents.

Exemple 1.3 Per a les dades de l'exemple 1.1 tenim:

x_i	n_i	f_i	N_i	F_i
0	2	0.04	2	0.04
1	4	0.08	6	0.12
2	21	0.42	27	0.54
3	15	0.30	42	0.84
4	6	0.12	48	0.96
5	1	0.02	49	0.98
6	1	0.02	50	1.00

Una vegada ordenades les dades, és molt fàcil obtindre informació de la mostra.

Exemple 1.4 Responeu les preguntes següents:

1. Quantes famílies tenen com a màxim dos fills?

   Podem mirar en la columna de les n_i: $2 + 4 + 21 = 27$, o en la de les $N_i = 27$.

2. Quantes famílies tenen més d'un fill o com a molt tres?

   Mirem en la columna de les n_i: $21 + 15 = 36$, o en la de les N_i: $42 - 6 = 36$.

3. Quin percentatge de famílies té més de tres fills?

   Si mirem en la columna de les f_i: $0{,}12 + 0{,}02 + 0{,}02 = 0{,}16$, concloem que el 16 % de les famílies té més de tres fills. Si mirem en la columna de les F_i: $1 - 0{,}84 = 0{,}16$, obtenim el mateix resultat.

Exemple 1.5 Si fem el mateix amb les dades de l'exemple 1.2, obtenim:

x_i	n_i	f_i	N_i	F_i
36	2	0.05	2	0.05
37	1	0.025	3	0.075
38	1	0.025	4	0.1
39	3	0.075	7	0.175
40	1	0.025	8	0.2
41	1	0.025	9	0.225
42	2	0.05	11	0.275
43	4	0.1	15	0.375
44	2	0.05	17	0.425
45	4	0.1	21	0.525
47	4	0.1	25	0.625
48	1	0.025	26	0.650
49	1	0.025	27	0.675
50	2	0.05	29	0.725
51	2	0.05	31	0.775
⋮	⋮	⋮	⋮	⋮

La taula és enorme!!!

Quan els valors diferents que pot prendre una variable són molts, s'obté una taula molt gran i, en conseqüència, és poc aclaridora. Això passarà sovint, quan la variable objecte d'estudi siga contínua. La solució és agrupar els diferents valors de la variable en intervals de classe, tenint sempre en compte que el que es guanya quant a l'organització i la facilitat per a manipular les dades, es perd en informació.

Agrupar en intervals de classe consisteix a agrupar les dades en un nombre xicotet d'intervals que verifiquen:

- Que no se superposen entre si, de forma que no existisca ambigüitat respecte a la classe a què pertany una dada particular.
- Que cobrisquen tot el rang de valors de la variable.

Anomenarem:

- **Límits superior i inferior** de la classe els extrems de l'interval. Els representarem per L_i i l_i, respectivament.

- **Marca de classe**, c_i, el punt mitjà de l'interval, és a dir, $c_i = \dfrac{L_i + l_i}{2}$.
- **Amplitud d'una classe**, a_i, la diferència entre els extrems superior i inferior de l'interval: $a_i = L_i - l_i$.
- **Freqüència de classe**, n_i, el nombre d'observacions de cada classe. Si dividim aquesta freqüència entre el nombre total d'observacions, tenim la **freqüència relativa de classe**, f_i. Anàlogament, definim N_i i F_i.

A continuació, donarem algunes indicacions per a respondre a la pregunta: com construïm una distribució de freqüències agrupada en intervals?

1. Començarem per determinar el **recorregut** de la variable, R_e, que es defineix com la diferència entre el valor observat més alt i el més baix.
2. El nombre de classes depén de la grandària de la mostra. Per a mostres no molt grans, $n < 50$, pot escollir-se un nombre de classes igual a $\sqrt{n}$. O bé s'usa la fórmula de Sturtges: $\dfrac{\log n}{\log 2} + 1$. A més, en general, el nombre de classes no ha de passar de 15 o 20, en casos de mostres molt grans.
3. Determinem l'amplitud dels intervals. És molt més còmode que tots els intervals tinguen la mateix amplitud (sempre que siga possible i excepte el primer i l'últim). Si és així,

$$a_i = a = \frac{R_e}{nombre\ d'intervals}.$$

Nota 1.2 Els passos 2 i 3 poden intercanviar-se.

Nota 1.3 Perquè no hi haja ambigüitat, prendrem els intervals tancats per l'esquerra i oberts per la dreta (excepte l'últim).

Exemple 1.6 Representarem ara la distribució de freqüències de l'exemple anterior, agrupant les dades.

El valor més baix és 33 i el més alt 61, per tant: $R_e = 61 - 33 = 28$. Com que $n = 40$, considerarem 6 classes, l'amplitud de les quals serà: $28/6 = 4.\hat{6}$. Així, l'amplitud és un nombre decimal periòdic i ens quedarien intervals un poc estranys. Podem fer el següent: prenem com a primer valor 32, en lloc de 33, i com a últim 62, en lloc de 61. D'aquesta manera, l'amplitud és 5 i la distribució de freqüències queda:

$[l_i, L_i[$	c_i	n_i	f_i	N_i	F_i
$[32, 37[$	34.5	2	0.05	2	0.05
$[37, 42[$	39.5	7	0.175	9	0.225
$[42, 47[$	44.5	12	0.3	21	0.525
$[47, 52[$	49.5	10	0.25	31	0.775
$[52, 57[$	54.5	4	0.1	35	0.875
$[57, 62]$	59.5	5	0.125	40	1.00

D'aquesta manera, podem respondre fàcilment a les preguntes:

1. Quants hotels tenen un preu entre 33 i 37 euros?
   2 hotels.
2. Quants hotels tenen un preu igual o superior a 47 euros?
   19 hotels.
3. Quin percentatge d'hotels costa, com a màxim, 42 euros?
   El 22.5 % dels hotels.

1.4. MÈTODES GRÀFICS

1.4.1. DIAGRAMA DE SECTORS

És un diagrama en forma circular en el qual, a cada valor de la variable, s'associa un sector circular proporcional a la seua freqüència. És adequat per a representar variables qualitatives.

Exemple 1.7 Una mostra de determinada població és enquestada abans de la convocatòria d'un referèndum, per poder efectuar una predicció sobre el resultat. El 50 % dels enquestats ha contestat que s'hi pronunciarà a favor, el 40 %, en contra i el 10 % restant ha dit que s'abstindrà.

El gràfic següent mostra el diagrama de sectors d'aquest exemple:

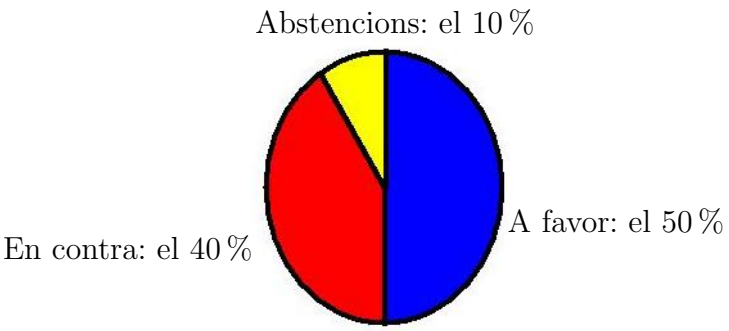


Figura 1.1: Diagrama de sectors de l'exemple 1.7

1.4.2. DIAGRAMA DE BARRES

Cada valor de la variable es representa mitjançant una barra d'alçada proporcional a la freqüència. És adequat tant per a representar variables qualitatives com variables quantitatives discretes. En la figura 1.2 se'n pot veure un exemple.

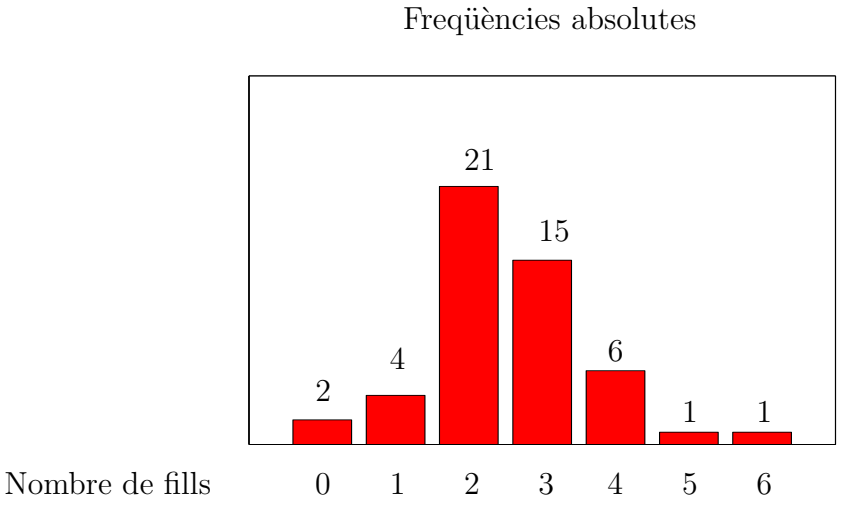


Figura 1.2: Diagrama de barres de l'exemple 1.1

1.4.3. POLÍGON DE FREQÜÈNCIES

Sobre cada valor de la variable (o interval) tracem una alçada igual a la seua freqüència (absoluta o acumulada). En el cas de dades discretes, unim mitjançant segments de recta l'extrem de cada ordenada amb la següent. En la figura 1.3 pot veure's el polígon de freqüències relatives acumulades (F_i) de l'exemple 1.6.

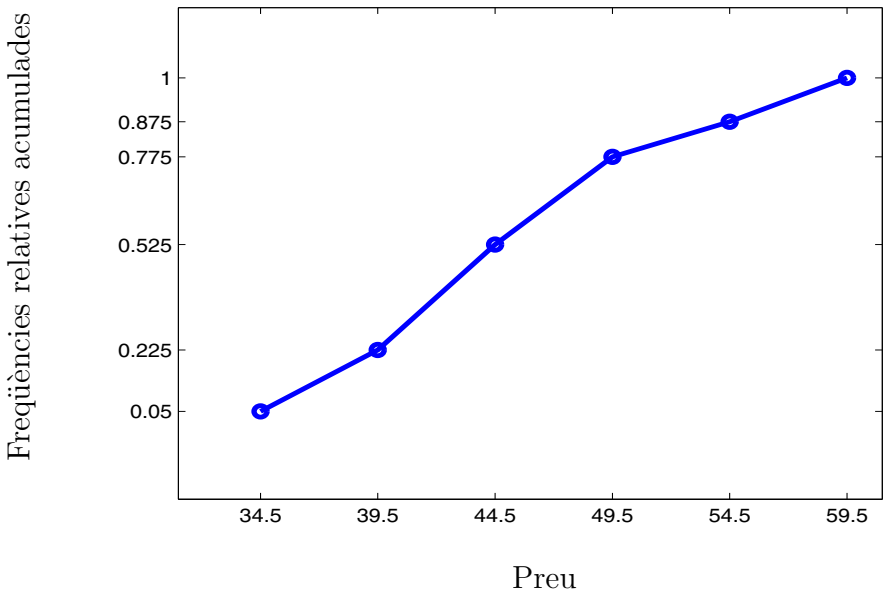


Figura 1.3: Polígon de freqüències relatives acumulades de l'exemple 1.6

1.4.4. HISTOGRAMES

És la representació equivalent al diagrama de barres, però per a dades agrupades per intervals. Sobre cada classe alcem un rectangle d'àrea proporcional a la freqüència de la classe. Caldrà, doncs, parar compte i veure si tots els intervals tenen la mateixa amplitud abans de fer el dibuix.

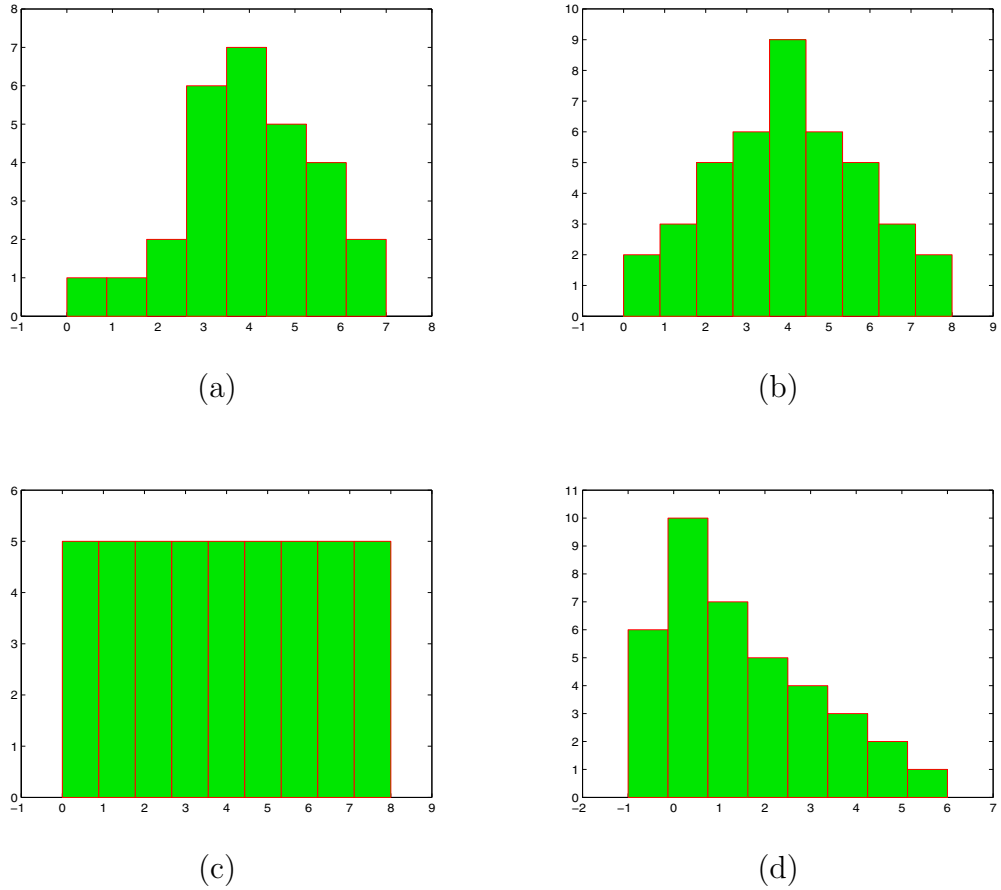


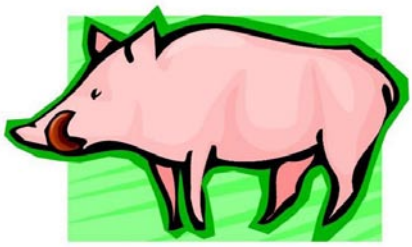
Figura 1.4: Diversos tipus d'histogrames

Els histogrames (i també els diagrames de barres) proporcionen molta informació respecte de l'estructura de les dades: el valor central de la distribució de les freqüències, la seua dispersió, simetria, etc. La figura 1.4 mostra diversos casos d'histogrames: en el primer (a), podem veure una distribució asimètrica que és típica de dades econòmiques, com pot ser la distribució de la renda, les dimensions d'una població, el consum d'electricitat en una ciutat, etc. El histograma (b) mostra una distribució simètrica i campaniforme que es presenta en la majoria de medicions físiques, en processos de fabricació, etc. Aquest tipus de distribució s'anomena **normal** perquè és el més habitual. L'histograma (c) presenta una distribució uniforme, que podria correspondre, per exemple, a l'última xifra del número premiat en una loteria. Finalment, la distribució (d) és molt asimètrica i apareix, per exemple, en estudiar temps entre avaries d'una determinada maquinària, en arribades a una finestra d'atenció al client, en temps transcorregut entre dos accidents de trànsit en una determinada carretera, etc.

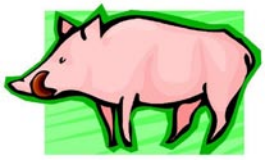
1.4.5. PICTOGRAMES

Expressen, amb dibuixos al·lusius al tema d'estudi, les freqüències de les modalitats de la variable. Els gràfics es fan de forma que queden representades les diferents escales del mateix dibuix en correspondència amb la grandària de la seua freqüència. L'escala dels dibuixos ha de ser de tal manera que l'àrea de cadascun siga proporcional a la freqüència de la modalitat que representa. Són molt utilitzats en variables qualitatives.

Exemple 1.8 Per a mostrar el consum de carn de porc en un mes en diferents ciutats, s'usaria la representació següent:



100 kg a la ciutat A



60 kg en B

40 kg en C

22 kg en D

1.5. PROBLEMES PROPOSATS

(1) Classifica les variables següents:

a) Color dels ulls.	*b*) Marques d'automòbil.
c) Alçada en cm.	*d*) Nivell d'estudis.
e) Anys d'estudis realitzats.	*f*) Nombre d'alumnes d'una classe.
g) Temperatura d'un malalt en ºC.	*h*) Professió.

(2) Els 100 estudiants d'una classe que es van presentar al primer examen parcial d'estadística en la convocatòria de febrer varen obtindre les qualificacions següents:

7	3	2	4	5	1	8	6	1	5	3	2	4	9	8	1	0	2	4	1
2	5	6	5	4	7	1	3	0	5	8	6	3	4	0	10	2	5	7	4
0	2	1	5	6	4	3	5	2	3	9	7	3	4	3	5	7	4	6	5
6	1	0	5	7	8	5	2	3	10	4	6	2	1	1	2	6	7	4	5
4	7	6	3	5	0	2	8	2	7	8	5	2	7	1	4	6	3	5	6

a) Obtín la distribució de freqüències de les qualificacions.

b) Quin percentatge d'estudiants va obtindre un 5?

c) Quants estudiants van obtindre un 6 o més?

d) Quin percentatge d'estudiants va aprovar?

e) Representa gràficament les freqüències no acumulades.

f) Representa gràficament les freqüències acumulades.

(3) En una determinada ciutat s'ha dut a terme un mostreig a partir del qual els establiments hotelers de la mostra s'han agrupat pel nombre de places que poseeixen, i s'ha obtingut la taula següent:

Places	Nre. d'hotels
$[0, 100[$	25
$[100, 200[$	37
$[200, 300[$	12
$[300, 400[$	22
$[400, 500[$	0
$[500, 600[$	21
$[600, 700[$	13
$[700, 800[$	5
$[800, 900[$	3
$[900, 1000]$	2

a) Construeix una taula de freqüències completa.

b) Determina el nombre d'establiments amb un nombre de places superior o igual a 400.

c) Quin percentatge d'establiments té menys de 700 places?

d) Representa gràficament les freqüències.

e) Representa gràficament les freqüències acumulades.

(4) A continuació apareixen els guanys, en euros, obtinguts en 25 quioscos per la venda diària d'un determinat diari:

55.31 81.47 64.90 70.89 86.02 77.25 76.73 84.51 56.02 84.92
90.23 78.01 88.05 73.37 87.09 55.31 81.47 64.90 70.89 86.02
77.25 76.73 84.51 56.02 84.92

Obtín la distribució de freqüències agrupades en 8 intervals amb les marques de classe 57, 62, 67, 72, 77, 82, 87, 92.

(5) Donades les següents qualificacions d'estadística d'un grup de 30 estudiants:

5.3 6.5 6 5 7.5 8 7 6.5 6 4.5
4.5 3.5 4 7 6.5 5 7 4.5 5 5.5
7.5 6.5 1 6 9.5 4 6 7.5 7 7.5

a) Obtín la distribució de freqüències.

b) Determina el percentatge de suspensos.

c) Calcula el percentatge d'estudiants amb qualificacions compreses entre 5 i 7.5, ambdues incloses.

(6) S'ha realitzat un estudi sobre el preu (en euros) per habitació de 50 hotels d'una determinada ciutat i s'han obtingut els resultats següents:

70	30	50	40	50	70	40	75	80	50
50	75	30	70	100	150	50	75	120	80
40	50	30	50	100	30	40	50	70	50
30	40	70	40	70	50	40	70	100	75
70	80	75	70	75	80	70	70	120	80

Determina:

a) La distribució de freqüències dels preus.

(a.1) Sense agrupar.

(a.2) Agrupant les dades en 5 intervals de la mateixa amplitud.

b) Representa gràficament ambdues distribucions.

c) Percentatge d'hotels amb un preu superior a 75 euros.

d) Quants hotels tenen un preu superior o igual que 50 euros però inferior o igual que 100?

(7) Completa la taula següent:

$[l_i, L_i[$	n_i	f_i	N_i
$[0, 10[$	60		60
$[10, 20[$		0,4	
$[20, 30[$	30		170
$[30, 40[$		0,1	
$[40, 50]$			200

(8) Les dades proporcionades a continuació corresponen al pes, en kg, de 80 persones:

60	66	77	70	66	68	57	70	66	52	75	65	69	71	58	66	67	74	61	63
69	80	59	66	70	67	78	75	64	71	81	62	64	69	68	72	83	56	65	74
67	54	65	65	69	61	67	73	57	62	67	68	63	67	71	68	76	61	62	63
76	61	67	67	64	72	64	73	79	58	67	71	68	59	69	70	66	62	63	66

a) Obtín la distribució de freqüències de manera que les dades estiguen agrupades en intervals d'amplitud 5.

b) Calcula el percentatge de persones que pesen menys de 65 kg.

c) Representa gràficament les freqüències absolutes acumulades.

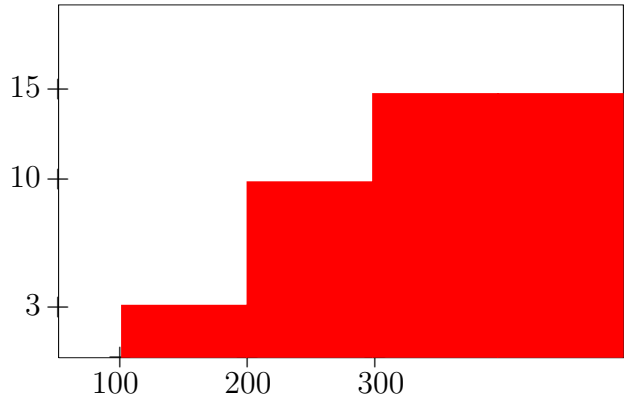
(9) A l'entrada d'un centre comercial, un enquestador recull informació sobre el nombre de desplaçaments que fan al mes les persones que hi acudeixen. Quan ha entrevistat 60 persones entrega la informació obtinguda, que resulta ser la següent:

2	8	5	6	1	3	2	8	5	3	2	4	1	3	4	4	3	5	2	6
1	7	6	2	5	3	8	4	6	2	8	7	6	4	3	2	6	1	1	1
2	2	4	7	6	2	1	3	4	5	8	2	2	6	5	3	2	3	4	3

a) Representa en una taula de freqüències, sense agrupar, les observacions anteriors. Quin percentatge de persones fa tres o menys visites al mes? Quantes persones en fan entre 4 i 7 (ambdós inclosos) al mes?

b) Representa les dades en una taula de freqüències, agrupant les dades en tres intervals. Quin percentatge de persones hi acudeix més de tres vegades al mes?

(10) El gràfic següent representa el diagrama de una distribució de freqüències absolutes acumulades. Troba la taula de freqüències completa.



SOLUCIONS

(1) *a*) És una variable qualitativa discreta: color A, color B, color C, etc.

b) És una variable qualitativa discreta: marca X, marca Y, marca Z, etc.

c) És una variable quantitativa contínua: 1.93, 1.935, 1.76, 1.67, etc.

d) És una variable qualitativa discreta: sense estudis, elementals, etc.

e) És una variable quantitativa discreta: 0, 1, 2, 3, etc.

f) És una variable quantitativa discreta: 0, 1, 12, 3033, 5004, etc.

g) És una variable quantitativa contínua: 36.1, 36.51, 36.512, 36.78, 37.1, 39.12, etc.

h) És una variable qualitativa discreta: metge, professor, pallasso, etc.

(2) *a*) Distribució de freqüències:

x_i	n_i	f_i	F_i	N_i
0	6	0.06	0.06	6
1	10	0.10	0.16	16
2	13	0.13	0.29	29
3	11	0.11	0.40	40
4	13	0.13	0.53	53
5	16	0.16	0.69	69
6	11	0.11	0.80	80
7	10	0.10	0.90	90
8	6	0.06	0.96	96
9	2	0.02	0.98	98
10	2	0.02	1.00	100

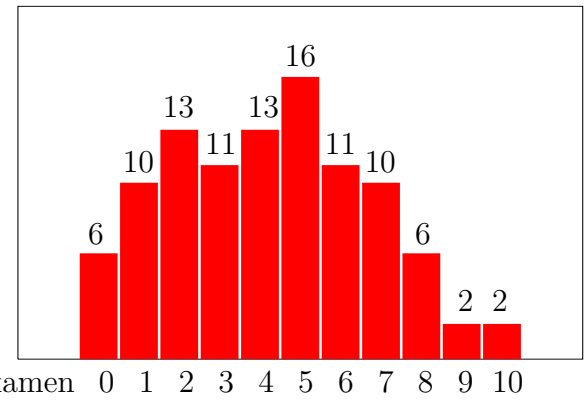
$\Rightarrow n = 100$

b) El 16 %.

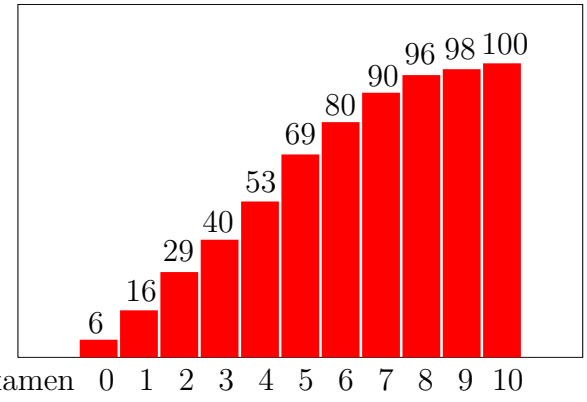
c) 31 estudiants.

d) El 47 %.

e)



f)



(3) *a*) Distribució de freqüències:

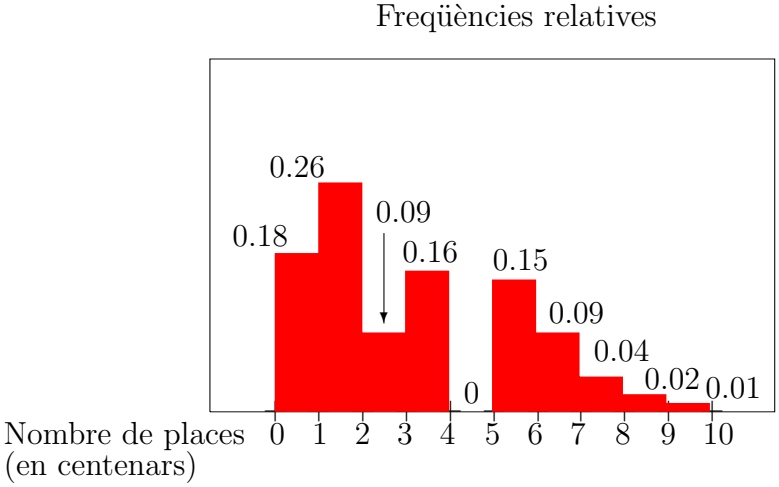
$[l_i, L_i[$	n_i	f_i	F_i	N_i
$[0, 100[$	25	0.18	0.18	25
$[100, 200[$	37	0.26	0.44	62
$[200, 300[$	12	0.09	0.53	74
$[300, 400[$	22	0.16	0.69	96
$[400, 500[$	0	0	0.69	96
$[500, 600[$	21	0.15	0.84	117
$[600, 700[$	13	0.09	0.93	130
$[700, 800[$	5	0.04	0.97	135
$[800, 900[$	3	0.02	0.99	138
$[900, 1000]$	2	0.01	1.00	140

$\Rightarrow n = 140$

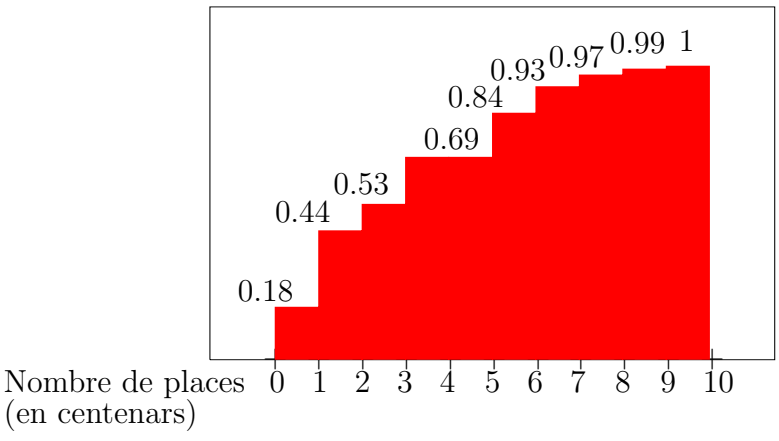
b) 44.

c) El 93 %.

d)



e)



(4) Distribució de freqüències:

$[l_i, L_i[$	c_i	n_i	f_i	F_i	N_i
[54.5, 59.5[	57	4	0.16	0.16	4
[59.5, 64.5[	62	0	0	0.16	4
[64.5, 69.5[	67	2	0.08	0.24	6
[69.5, 74.5[	72	3	0.12	0.36	9
[74.5, 79.5[	77	5	0.20	0.56	14
[79.5, 84.5[	82	2	0.08	0.64	16
[84.5, 89.5[	87	8	0.32	0.96	24
[89.5, 94.5[	92	1	0.04	1.00	25

$\Rightarrow n = 25$

(5) *a*) Distribució de freqüències:

x_i	n_i	f_i	F_i	N_i
1	1	0.035	0.035	1
3.5	1	0.035	0.07	2
4	2	0.07	0.14	4
4.5	3	0.1	0.24	7
5	3	0.1	0.34	10
5.3	1	0.035	0.375	11
5.5	1	0.035	0.41	12
6	4	0.13	0.54	16
6.5	4	0.13	0.67	20
7	4	0.13	0.80	24
7.5	4	0.13	0.93	28
8	1	0.035	0.965	29
9.5	1	0.035	1.00	30

$\Rightarrow n = 30$

b) El 24 %.

c) El 69 %.

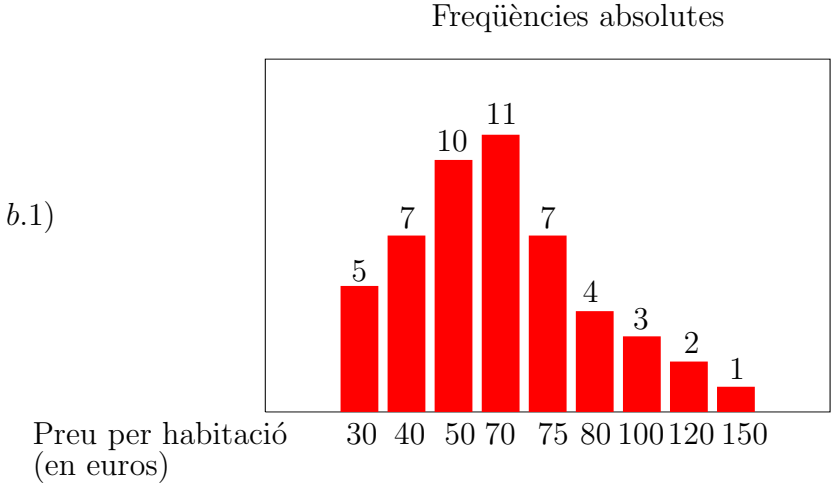
(6) *a*.1) Distribució de freqüències:

x_i	n_i	f_i	F_i	N_i
30	5	0,1	0,1	5
40	7	0,14	0,24	12
50	10	0,2	0,44	22
70	11	0,22	0,66	33
75	7	0,14	0,80	40
80	4	0,08	0,88	44
100	3	0,06	0,94	47
120	2	0,04	0,98	49
150	1	0,02	1	50

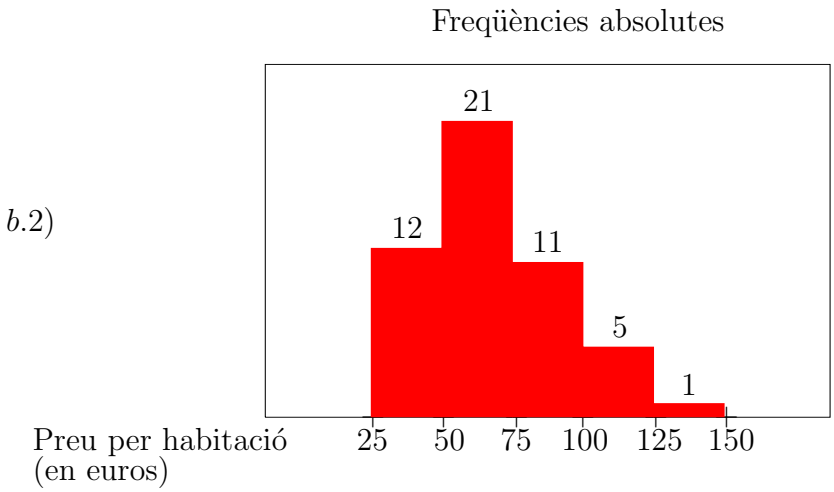
a.2) Distribució de freqüències (dades acumulades):

$[l_i, L_i[$	c_i	n_i	f_i	F_i	N_i
$[25, 50[$	37,5	12	0,24	0,24	12
$[50, 75[$	62,5	21	0,42	0,66	33
$[75, 100[$	87,5	11	0,22	0,88	44
$[100, 125[$	112,5	5	0,1	0,98	49
$[125, 150]$	137,5	1	0,02	1,00	50

b.1)



b.2)



c) El 34 %.

d) 35 hotels.

(7) Distribució de freqüències:

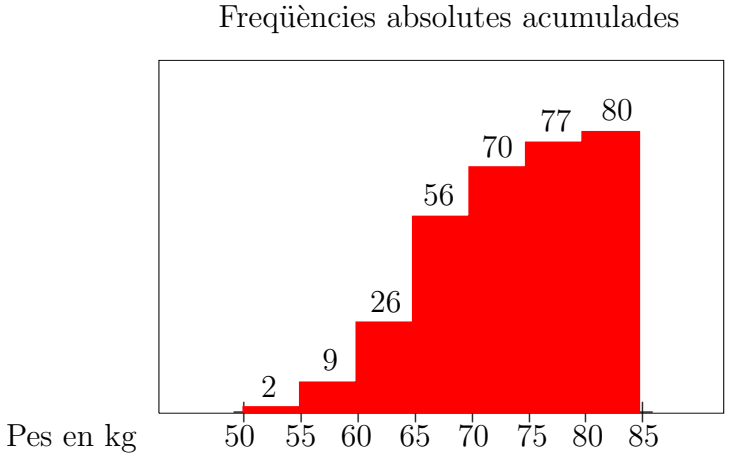
$[l_i, L_i[$	n_i	f_i	N_i
$[0, 10[$	60	**0.3**	60
$[10, 20[$	**80**	0,4	**140**
$[20, 30[$	30	**0.15**	170
$[30, 40[$	**20**	0,1	**190**
$[40, 50]$	**10**	**0.05**	200

(8) *a*) Distribució de freqüències:

$[l_i, L_i[$	c_i	n_i	f_i	F_i	N_i
[50, 55[	52.5	2	0.025	0.025	2
[55, 60[	57.5	7	0.0875	0.1125	9
[60, 65[	62.5	17	0.2125	0.325	26
[65, 70[	67.5	11	0.375	0.7	56
[70, 75[	72.5	11	0.175	0.875	70
[75, 80[	77.5	5	0.0875	0.9625	77
[80, 85[	82.5	1	0.0375	1	80

b) El 32.5 %.

c)



(9) a.1) Distribució de freqüències:

x_i	n_i	f_i	F_i	N_i
1	7	$0{,}116\hat{6}$	$0{,}116\hat{6}$	7
2	13	$0{,}26\hat{6}$	$0{,}3\hat{3}$	20
3	10	$0{,}16\hat{6}$	0,5	30
4	8	$0{,}13\hat{3}$	$0{,}63\hat{3}$	38
5	6	0,1	$0{,}73\hat{3}$	44
6	8	$0{,}13\hat{3}$	$0{,}86\hat{6}$	52
7	3	0,05	$0{,}916\hat{6}$	55
8	5	$0{,}0833\hat{3}$	1	60

$\Rightarrow n = 60$

a.2) El $66.\hat{6}\,\%$.

a.3) 25 persones.

b.1) Distribució de freqüències:

$[l_i, L_i[$	c_i	n_i	f_i	F_i	N_i
$[1, 4[$	2.5	30	0.5	0.5	30
$[4, 7[$	5.5	22	$0.36\hat{6}$	$0.83\hat{3}$	52
$[7, 9[$	8	8	$0.13\hat{3}$	1	60

b.2) El 50 %.

(10) Distribució de freqüències:

$[l_i, L_i[$	n_i	f_i	F_i	N_i
$[100, 200[$	3	0,2	0,2	3
$[200, 300[$	7	$0{,}46\hat{6}$	$0{,}6\hat{6}$	10
$[300, +\infty[$	5	$0{,}3\hat{3}$	1	15

$\Rightarrow n = 15$

TEMA 2

DESCRIPCIÓ D'UNA MOSTRA: MESURES DESCRIPTIVES

Per a dades qualitatives, la distribució de freqüències proporciona un resum concís i complet de la mostra, però per a variables quantitatives pot complementar-se utilitzant mesures descriptives numèriques tretes de les dades.

Les **mesures descriptives** són valors numèrics calculats a partir de la mostra i que ens resumeixen la informació que aquesta conté. En la part d'inferència estadística, les anomenarem **estadístics**.

2.1. MESURES DE POSICIÓ

Ens donen el valor que ocupa una determinada *posició* respecte de la resta de la mostra.

2.1.1. MESURES DE TENDÈNCIA CENTRAL

Ens donen un *centre* de la distribució de freqüències. Són valors que poden considerar-se com a *mesura resum* de totes les dades. Hi ha diferents formes de definir el *centre* de les observacions d'un conjunt de dades. Per ordre d'importància són:

1. **Mitjana aritmètica** (o simplement **mitjana**) ($\overline{x}$): és el quocient entre la suma de totes les dades i el nombre total d'aquestes (tenint en compte que si un valor es repeteix, cal considerar-ne totes les repeticions). Es calcula mitjançant:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i f_i.$$

   Si les dades estan agrupades en intervals, usarem la marca de classe, c_i, en lloc de x_i.

   És la mesura de centralització més important.

   Exemple 2.1 En l'exemple 1.1, del tema anterior, la mitjana de fills per família és:

$$\overline{x} = \frac{0\cdot 2 + 1\cdot 4 + 2\cdot 21 + 3\cdot 15 + 4\cdot 6 + 5\cdot 1 + 6\cdot 1}{50} = \frac{126}{50} = 2.52 \text{ fills.}$$

   Hauríem pogut calcular-la també com $\overline{x} = \sum_{i=1}^{k} x_i f_i$, és a dir:

$$\overline{x} = 0\times 0.04 + 1\times 0.08 + 2\times 0.42 + 3\times 0.3 + 4\times 0.12 + 5\times 0.02 + 6\times 0.02 = 2.52 \text{ fills.}$$

Exemple 2.2 El preu mitjà de les habitacions en l'exemple 1.2, del tema anterior, el calculem utilitzant les marques de classe, és a dir:

$$\overline{x} = \frac{34.5 \times 2 + 39.5 \times 7 + 44.5 \times 12 + 49.5 \times 10 + 54.5 \times 4 + 59.5 \times 5}{40} = 47.25 €.$$

O, equivalentment:

$$\overline{x} = 34.5 \times 0.05 + 39.5 \times 0.175 + 44.5 \times 0.3 + 49.5 \times 0.25 + 54.5 \times 0.1 + 59.5 \times 0.125 = 47.25 €.$$

Les propietats més importants de la mitjana són:

(a) Si a tots els valors d'una variable els sumem una constant C, la mitjana aritmètica queda augmentada en aquesta constant. És a dir, queda afectada pels canvis d'origen de la mateixa manera que les dades:

$$y_i = C + x_i \Rightarrow \overline{y} = C + \overline{x}.$$

Demostració:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} y_i n_i = \frac{1}{n}\sum_{i=1}^{k}(C + x_i)n_i = C\frac{1}{n}\sum_{i=1}^{k} n_i + \frac{1}{n}\sum_{i=1}^{k} x_i n_i = C + \overline{x},$$

ja que $\sum_{i=1}^{k} n_i = 1$

(b) Si tots els valors d'una variable els multipliquem per una constant C, la seua mitjana aritmètica queda multiplicada per la mateixa constant. És a dir, la mitjana aritmètica queda afectada pels canvis d'escala de la mateixa manera que les dades:

$$y_i = Cx_i \Rightarrow \overline{y} = C\overline{x}.$$

Demostració:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{k} y_i n_i = \frac{1}{n}\sum_{i=1}^{k} Cx_i n_i = C\frac{1}{n}\sum_{i=1}^{k} x_i n_i = C\overline{x}$$

(c) Com a corol·lari d'ambdues propietats anteriors, si considerem la transformació lineal $y_i = A + Cx_i$, on A i C són dues constants qualssevol, la mitjana arirmètica de la nova variable és: $\overline{y} = A + C\overline{x}$.

Demostració:
És evident a partir de les propietats anteriors.

(d) La suma de totes les diferències entre els valors de la variable i la mitjana és 0.

$$\sum_{i=1}^{k}(x_i - \overline{x})n_i = 0,$$

és a dir, la mitjana és el centre de gravetat de les observacions.

Demostració:

$$\sum_{i=1}^{k}(x_i - \overline{x})n_i = \sum_{i=1}^{k} x_i n_i - \sum_{i=1}^{k} \overline{x} n_i = n\overline{x} - n\overline{x} = 0$$

(e) La suma de les desviacions al quadrat dels valors de la variable respecte a una constant C qualsevol és mínima quan aquesta constant és la mitjana aritmètica. És a dir:

$$\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i \leq \sum_{i=1}^{k}(x_i - C)^2 n_i, \text{ per a qualsevol constant } C.$$

2. Mediana (Me): és el valor per al qual, quan totes les observacions s'ordenen de més baixa a més alta, la meitat d'aquestes és més petita que aquest valor i l'altra meitat, més gran. Si el nombre de dades és imparell, la mediana serà el valor central; si és parell, prendrem com a mediana la mitjana aritmètica dels dos valors centrals.

La forma més còmoda de calcular-la és usant les freqüències acumulades.

(a) Distribucions no agrupades

1) Calculem $\dfrac{n}{2}$.
2) Mirem en la distribució de freqüències la columna de les freqüències absolutes acumulades i hi busquem la freqüència N_i que faça complir que $N_{i-1} < \dfrac{n}{2} \leq N_i$:
   - Si $\dfrac{n}{2} < N_i$, aleshores la mediana és aquell valor la freqüència acumulada del qual és N_i, és a dir:

$$Me = x_i, \text{ de manera que } \frac{n}{2} < N_i.$$

- Si $\frac{n}{2} = N_i$ (noteu que això només pot passar quan n és parell), la mediana és la mitjana aritmètica d'aquells valors la freqüència acumulada dels quals és N_i i N_{i+1}, respectivament, és a dir:

$$Me = \frac{x_i + x_{i+1}}{2}, \text{ de manera que } \frac{n}{2} = N_i.$$

Exemple 2.3 Mediana del nombre de fills de l'exemple 1.1:

x_i	n_i	N_i
0	2	2
1	4	6 2
2	21	27
3	15	42
4	6	48
5	1	49
6	1	50

$n = 50$

$\frac{n}{2} = 25$

$N_2 = 6 < 25 \leq 27 = N_3$

Per tant, $Me = x_3 = 2$ fills.

(b) Distribucions agrupades per intervals

1) Calculem $\frac{n}{2}$.

2) Mirem en la distribució de freqüències la columna de les freqüències absolutes acumulades i hi busquem la freqüència N_i que faça complir que $N_{i-1} < \frac{n}{2} \leq N_i$. A aquesta freqüència, li correspon l'interval $[l_i, L_i[$ (que anomenarem **interval medià**); a continuació, per obtindre la mediana, aplicarem la fórmula següent:

$$Me = l_i + \frac{\left(\frac{n}{2} - N_{i-1}\right) a_i}{n_i}.$$

El raonament per a justificar la utilització d'aquesta fórmula és el següent: la freqüència absoluta acumulada fins a l'interval anterior al medià és N_{i-1}. Per a arribar a la meitat de les dades, és a dir, per a arribar fins a $\frac{n}{2}$, necessitem prendre $\left(\frac{n}{2} - N_{i-1}\right)$ dades de l'interval medià (el qual conté n_i) repartides en una amplitud a_i.

Com que a cada dada correspon una longitud $\frac{a_i}{n_i}$, a les $\left(\frac{n}{2} - N_{i-1}\right)$ dades, els corrrespondrà la longitud:

$$\frac{\left(\frac{n}{2} - N_{i-1}\right) a_i}{n_i}.$$

Exemple 2.4 Mediana del preu de les habitacions dels 40 hotels de l'exemple 1.2:

$[l_i, L_i[$	n_i	N_i
$[32, 37[$	2	2
$[37, 42[$	7	9
$[42, 47[$	12	21
$[47, 52[$	10	31
$[52, 57[$	4	35
$[57, 62]$	5	40

$n = 40$

$\frac{n}{2} = 20$

$N_2 = 9 < 20 \leq 21 = N_3$

Interval medià: $[37, 42[$

$Me = 37 + \frac{(\frac{40}{2} - 9) \cdot 5}{7} = 44{,}86$ €

Les propietats més importants de la mediana són:

(a) Si a tots els valors d'una variable els sumem una mateixa constant C, la mediana queda augmentada en aquesta constant. És a dir, la mediana queda afectada pels canvis d'origen de la mateixa manera que les dades:

$$y_i = C + x_i \Rightarrow Me_y = C + Me_x.$$

(b) Si tots els valors d'una variable els multipliquem per una constant C, la seua mediana queda multiplicada per la mateixa constant. És a dir, la mediana queda afectada pels canvis d'escala de la mateixa manera que les dades:

$$y_i = Cx_i \Rightarrow Me_y = CMe_x.$$

(c) Com a corol·lari de totes dues propietats anteriors, si considerem la transformació lineal $y_i = A + Cx_i$, on A i C són dues constants qualssevol, la mediana de la nova variable és: $Me_y = A + CMe_x$.

(d) La mediana fa mínima la suma de totes les desviacions absolutes dels valors de la variable respecte a una constant C qualsevol. És a dir:

$$\sum_{i=1}^{k} |x_i - Me|\, n_i \leq \sum_{i=1}^{k} |x_i - C|\, n_i, \text{ per a qualsevol constant } C.$$

3. Moda (Mo): és el valor de la variable que més es repeteix, és a dir, la freqüència (absoluta o relativa) del qual és més alta. No té per què ser única. Per a calcular-la distingirem:

(a) Distribucions no agrupades

Simplement observem en la columna de les freqüències absolutes i triem aquell valor o aquells valors de la variable que tenen més freqüència. Quan trobem dues modes, diem que la distribució és bimodal; quan en trobem tres, trimodal, etc.

Exemple 2.5 Moda del nombre de fills per família de l'exemple 1.1:

x_i	n_i
0	2
1	4
2	21
3	15
4	6
5	1
6	1

n_i més alt $= 21$

Correspon a n_3.

Per tant, $Mo = x_3 = 2$ fills.

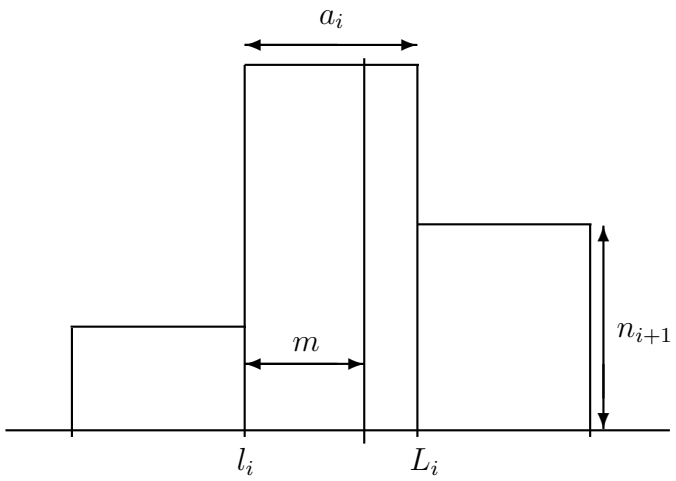


Figura 2.1: Representació de la moda en distribucions agrupades

(b) Distribucions agrupades per intervals

1) Intervals d'igual amplitud: observant les freqüències absolutes, determinem l'interval amb més freqüència $[l_i, L_i[$ (que anomenarem **interval modal**). A continuació, per a calcular la moda, aplicarem la fórmula següent:

$$Mo = l_i + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} \cdot a_i.$$

El raonament per a justificar la utilització d'aquesta fórmula (observeu la figura 2.1) és el següent: considerem els intervals anterior i posterior al modal, amb freqüències respectives n_{i-1} i n_{i+1}. Si aquestes freqüències són iguals, la moda és el centre de l'interval modal. En cas contrari, la moda estarà més prop d'aquell interval contigu, la freqüència del qual és més alta, és a dir, les distàncies de la moda als intervals contigus al modal són inversament proporcionals a les freqüències d'aquests intervals. Com a conseqüència d'açò, tindrem:

$$Mo = l_i + m, \text{ on } m \text{ permet que } \frac{m}{a_i - m} = \frac{n_{i+1}}{n_{i-1}}.$$

Si aïllem m i substituïm, obtenim la fórmula anterior.

Exemple 2.6 Moda del preu de les habitacions dels 40 hotels de l'exemple 1.2:

$[l_i, L_i[$	n_i
$[32, 37[$	2
$[37, 42[$	7
$[42, 47[$	12
$[47, 52[$	10
$[52, 57[$	4
$[57, 62]$	5

n_i més alt $= 12$, correspon a n_3.

Interval modal: $[42, 47[$

$$Mo = 42 + \frac{10}{7 + 10} \cdot 5 = 44{,}94 \text{ €}$$

2) Intervals de diferent amplitud: en primer lloc hem de calcular la **densitat de freqüència** de cada interval, que es defineix com:

$$d_i = \frac{n_i}{a_i}.$$

L'interval modal, $[l_i, L_i[$, serà ara el que tinga la densitat de freqüència més alta, i per a calcular la moda aplicarem la fórmula anterior, i substituirem les freqüències per les densitats de freqüència, és a dir:

$$Mo = l_i + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} \cdot a_i.$$

Nota 2.1 Comparació entre mitjana, mediana i moda:

Aquestes tres mesures de tendència central són les més importants i les més usuals, però quan utilitzem l'una o l'altra?

- La mitjana és la millor perquè fa ús de tota la informació, és a dir, pren en consideració tots els valors de la distribució. També té com a avantatge que és única. L'inconvenient principal és que és molt sensible a la presentació de dades anòmales o atípiques, que fan que el seu valor es desplace cap als dits valors. Així doncs, no és recomanable usar la mitjana en aquests casos. Un altre desavantatge és que pot no coincidir amb un dels valors de la distribució.

- Usarem la mediana quan falle la mitjana. Aquella usa menys informació que aquesta, ja que no depén dels valors de la variable, sinó del lloc que ocupen. Per aquest motiu, té l'avantatge de no estar afectada per observacions extremes. Un altre avantatge davant de la mitjana és que quasi sempre coincideix amb un valor de la variable.

- La moda és la que menys informació utilitza i és, per tant, la pitjor. A més, pot no ser única, la qual cosa és un altre inconvenient. L'avantatge més important és que podem obtindre-la, també, per a dades qualitatives.

- Si la distribució és simètrica i campaniforme, la mitjana, la mediana i la moda coincideixen.

- En el cas de distribucions campaniformes, la mediana està, amb freqüència, entre la mitjana i la moda (un poc més prop de la mitjana).

- La relació següent ens permet calcular aproximadament (en distribucions campaniformes) una d'aquestes mesures en funció de les altres:

$$Mo \simeq 3Me - 2\overline{x}.$$

Les mesures de centralització següents tenen un significat estadístic menys intuïtiu i s'utilitzen en situacions més específiques.

4. **Mitjana geomètrica** (G): es defineix com l'arrel enèsima del producte de les n dades:

$$G = \sqrt[n]{\prod_{i=1}^{k} x_i^{n_i}}.$$

   La mitjana geomètrica sol emprar-se per a fer mitjanes amb percentatges, taxes i nombres índexs.

   Propietat 2.1 *El logaritme de la mitjana geomètrica és igual a la mitjana aritmètica dels logaritmes dels valors de la variable.*

5. **Mitjana harmònica** (H): es defineix com el recíproc de la mitjana aritmètica dels recíprocs de les dades:

$$H = \frac{n}{\sum_{i=1}^{k} \frac{1}{x_i} \cdot n_i}.$$

   Sol utilitzar-se per a fer mitjanes amb velocitats, rendimients i, en general, magnituds expressades en termes relatius.

Nota 2.2 Si les dades estan agrupades, per a calcular totes dues mesures anteriors utilitzarem les marques de classe, és a dir, c_i en lloc de x_i.

Propietat 2.2 *Les tres mitjanes estan relacionades mitjançant:*

$$H \leq G \leq \overline{x}.$$

2.1.2. MESURES DE POSICIÓ NO CENTRALS: QUANTILS

Els **quantils** són valors de la distribució que la divideixen en parts iguals, és a dir, en intervals que contenen el mateix nombre de valors de la distribució. Els més usuals són:

1. **Percentils**: són 99 valors que divideixen la distribució en 100 parts iguals, després d'haver ordenat les dades. El **percentil d'ordre p** (P_p) és el menor valor superior al $p\,\%$ de les dades (ordenades les dades de més baixa a més alta, deixa el $p\,\%$ de les dades per davant). Els calculem a partir de les freqüències acumulades.

   (a) Dades no agrupades:
   - Calculem el $p\,\%$ de n, és a dir, $\dfrac{p \cdot n}{100}$.
   - Es busca en la taula el valor la freqüència acumulada del qual és la primera superior o igual al $p\,\%$ de n, és a dir:
   $$P_p = x_i \text{ que permeta que } N_{i-1} < \frac{p \cdot n}{100} \leq N_i.$$

   (b) Dades agrupades (utilitzem la mateixa idea que en el càlcul de la mediana):
   - Calculem el $p\,\%$ de n, és a dir, $\dfrac{p \cdot n}{100}$.
   - Busquem l'interval $[l_i, L_i[$ la freqüència acumulada del qual verifica $N_{i-1} < \dfrac{p \cdot n}{100} \leq N_i$.
   - A continuació, per a trobar el percentil, apliquem la fórmula següent:
   $$P_p = l_i + \frac{\left(\frac{p \cdot n}{100} - N_{i-1}\right) \cdot a_i}{n_i}.$$

2. **Quartils** (Q_i): són els tres valors que divideixen el conjunt de dades ordenades en quatre parts iguals. Són un cas particular dels percentils, de forma que:

$$Q_1 = P_{25}, Q_2 = P_{50} \text{ i } Q_3 = P_{75}.$$

Exemple 2.7 Calcula els tres quartils per a la distribució del nombre de fills de les 50 famílies de l'enquesta de l'exemple 1.1:

x_i	n_i	N_i
0	2	2
1	4	6
2	21	27
3	15	42
4	6	48
5	1	49
6	1	50

$$Q_1 = P_{25}; \quad \frac{25 \cdot 50}{100} = 12{,}5 \Rightarrow Q_1 = 2 \text{ fills}$$

$$Q_2 = P_{50}; \quad \frac{50 \cdot 50}{100} = 25 \Rightarrow Q_2 = 2 \text{ fills}$$

$$Q_3 = P_{75}; \quad \frac{75 \cdot 50}{100} = 37{,}5 \Rightarrow Q_3 = 3 \text{ fills}$$

Exemple 2.8 Calcula els tres quartils per a la distribució del preu per habitació dels 40 hotels de l'enquesta de l'exemple 1.2:

$[l_i, L_i[$	n_i	N_i
$[32, 37[$	2	2
$[37, 42[$	7	9
$[42, 47[$	12	21
$[47, 52[$	10	31
$[52, 57[$	4	35
$[57, 62]$	5	40

$$Q_1 = P_{25}; \quad \frac{25 \cdot 40}{100} = 10 \Rightarrow Q_1 \in [42, 47[$$

$$Q_1 = 42 + \frac{\frac{25 \cdot 40}{100} - 9}{12} \cdot 5 = 42{,}42 \text{ €}$$

$$q_2 = P_{50}; \quad \frac{50 \cdot 40}{100} = 20 \Rightarrow Q_2 \in [42, 47[$$

$$Q_2 = 42 + \frac{\frac{50 \cdot 40}{100} - 9}{12} \cdot 5 = 46{,}58 \text{ €}$$

$$Q_3 = P_{75}; \quad \frac{75 \cdot 40}{100} = 30 \Rightarrow Q_3 \in [47, 52[$$

$$Q_3 = 47 + \frac{\frac{75 \cdot 40}{100} - 12}{10} \cdot 5 = 56 \text{ €}$$

3. **Decils** (D_i): són els nou valors que divideixen la distribució, una vegada ordenades les dades de més baixa a més alta, en deu parts iguals. També són un cas particular dels percentils:

$$D_1 = P_{10}, \quad D_2 = P_{20}, \quad \ldots, \quad D_9 = P_{90}.$$

Propietat 2.3 *Evidentment, per a qualsevol distribució es verifica:*

$$Me = P_{50} = Q_2 = D_5.$$

2.2. MOMENTS

Els moments d'una distribució es defineixen com una generalització de la mitjana aritmètica i, com veurem més endavant, serveixen per a descriure algunes característiques importants de les distribucions de freqüències. La propietat més important és que dues distribucions són iguals quan tenen iguals tots els moments, i com més moments iguals tenen més paregudes són.

1. El **moment respecte a l'origen d'ordre r** (a_r) és la mitjana aritmètica de les potències r-èsimes de les dades. És a dir:

$$a_r = \frac{1}{n}\sum_{i=1}^{k} x_i^r n_i.$$

Propietat 2.4 *Òbviament es verifiquen:*

$$a_0 = \frac{1}{n}\sum_{i=1}^{k} x_i^0 n_i = \frac{n}{n} = 1 \quad \text{i} \quad a_1 = \frac{1}{n}\sum_{i=1}^{k} x_i n_i = \overline{x}.$$

2. El **moment d'ordre r respecte a la mitjana aritmètica** és:

$$m_r = \frac{1}{n}\sum_{i=1}^{k} (x_i - \overline{x})^r n_i.$$

Propietat 2.5 *Els moments d'ordre r respecte a la mitjana aritmètica més comuns verifiquen:*

$$m_0 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \overline{x})^0 n_i = \frac{n}{n} = 1 \quad \text{i} \quad m_1 = \frac{1}{n}\sum_{i=1}^{k} (x_i - \overline{x})^r n_i = \overline{x} - \overline{x} = 0.$$

2.3. MESURES DE DISPERSIÓ

Les mesures de tendència central tenen com a objectiu sintetitzar les dades en un valor representatiu. Les mesures de dispersió ens diuen fins a quin punt les de tendència central són representatives com a síntesi de la informació. Les mesures de dispersió quantifiquen la separació, la dispersió, i la variabilitat dels valors de la distribució respecte dels valors centrals.

Distingirem entre mesures de dispersió absolutes, que no són comparables entre diferents mostres, i les relatives, que ens permeten comparar-ne diverses.

2.3.1. MESURES DE DISPERSIÓ ABSOLUTES

Per ordre d'importància tenim:

1. **Variància** (s^2): és la mitjana dels quadrats de les distàncies entre cada observació i la mitjana aritmètica del conjunt de les observacions:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \sum_{i=1}^{k}(x_i - \overline{x})^2 f_i.$$

Si les dades estan agrupades per intervals, usarem les marques de classe per a calcular-la, és a dir, c_i en lloc de x_i.

En el cas extrem que totes les observacions siguen iguals, la mitjana coincideix amb aquest valor comú i, en conseqüència, la variància és 0. En general, com més disperses siguen les observacions, més grans seran les diferències dins dels quadrats i per tant, més alt és el valor de s^2.

Nota 2.3 La variància és el moment d'ordre 2 respecte de la mitjana, és a dir, $s^2 = m_2$.

Les propietats més importants de la variància són:

(a) La variància mai pot ser negativa: $s^2 \geq 0$.

(b) Una forma més senzilla de calcular la variància és:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} {x_i}^2 n_i - \overline{x} = \overline{x^2} - \overline{x} = a_2 - a_1^2$$

Demostració:

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)n_i = \\ &= \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - 2\overline{x}\frac{1}{n}\sum_{i=1}^{k} x_i + \overline{x}^2\frac{1}{n}\sum_{i=1}^{k} n_i = \\ &= \frac{1}{n}\sum_{i=1}^{k} x_i^2 n_i - 2\overline{x}^2 + \overline{x}^2 = \overline{x^2} - \overline{x}^2 = a_2 - a_1^2 \end{aligned}$$

(c) Si a tots els valors d'una variable, els sumem la mateixa constant C, la variància no canvia:

$$y_i = C + x_i \Rightarrow s_y^2 = s_x^2.$$

Demostració:

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \overline{y})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}(C + x_i - C - \overline{x})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = s_x^2$$

(d) Si tots els valors d'una variable, els multipliquem per una mateixa constant C, la seua variància queda multiplicada pel quadrat de la constant:

$$y_i = Cx_i \Rightarrow s_y^2 = C^2 s_x^2.$$

Demostració:

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k}(y_i - \overline{y})^2 n_i = \frac{1}{n}\sum_{i=1}^{k}(Cx_i - C\overline{x})^2 n_i = C^2 \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i = C^2 s_x^2$$

(e) Com a corol·lari de les propietats anteriors, si considerem la transformació lineal $y_i = A + Cx_i$, on A i C són dues constants qualssevol, la nova variància queda $s_y^2 = C^2 s_x^2$.

Exemple 2.9 Variància del nombre de fills per família de l'exemple 1.1:

x_i	n_i
0	2
1	4
2	21
3	15
4	6
5	1
6	1

$$\overline{x} = 2.52 \text{ fills}$$

$$s^2 = \frac{0^2 \cdot 2 + 1^2 \cdot 4 + 2^2 \cdot 21 + 3^2 \cdot 15 + 4^2 \cdot 6 + 5^2 \cdot 1 + 6^2 \cdot 1}{50} -$$

$$- (2.52)^2 = 1.25 \text{ (fills)}^2$$

Altres mesures de dispersió directament relacionades amb la variància són les dues següents:

2. Desviació típica (s): és l'arrel quadrada positiva de la variància. El motiu principal per a utilitzar-la és que la variància no està donada en les mateixes unitats que la variable, sinó en aquestes unitats al quadrat.

Les propietats més importants de la desviació típica, que es dedueixen fàcilment a partir de les corresponents propietats per a la variància, són:

(a) $s \geq 0$

(b) $y_i = C + x_i \Rightarrow s_y = s_x$

(c) $y_i = Cx_i \Rightarrow s_y = |C|\, s_x$

(d) $y_i = A + Cx_i \Rightarrow s_y = |C|\, s_x$

Exemple 2.10 En l'exemple 2.9, $s = \sqrt{1.25} = 1.12$ fills.

3. **Quasivariància** (s^{*2}): la definició és com la de la variància, però dividint entre $(n-1)$:

$$s^{*2} = \frac{1}{n-1} \cdot \sum_{i=1}^{k} (x_i - \overline{x})^2 \cdot n_i = \frac{n}{n-1} \cdot s^2.$$

Exemple 2.11 En l'exemple 2.9, $s^{*2} = \dfrac{50}{49} \times 1.25 = 1.27$ (fills)^2.

4. **Desviació mitjana respecte de la mitjana aritmètica** ($D_{\overline{x}}$): es defineix com la mitjana aritmètica de les desviacions, en valor absolut, respecte de la mitjana aritmètica:

$$D_{\overline{x}} = \frac{1}{n} \sum_{i=1}^{k} |x_i - \overline{x}|\, n_i.$$

Si pren valors grans, significa que els valors de la variable es distribueixen en valors allunyats de la mitjana.

Exemple 2.12 Per al nombre de fills per família de l'exemple 1.1:

x_i	n_i	$\lvert x_i - x\rvert$	$\lvert x_i - x\rvert\, n_i$
0	2	2.52	5.04
1	4	1.52	6.08
2	21	0.52	10.92
3	15	0.48	7.2
4	6	1.48	8.88
5	1	2.48	2.48
6	1	3.48	3.48
	50		44.08

$$D_{\overline{x}} = \frac{1}{n} \sum_{i=1}^{k} |x_i - \overline{x}|\, n_i = \frac{44.08}{50} = 0.88 \text{ fills.}$$

5. **Desviació mitjana respecte de la mediana** (D_{Me}): es defineix com la mitjana aritmètica de les desviacions, en valor absolut, respecte de la mediana:

$$D_{Me} = \frac{1}{n} \sum_{i=1}^{k} |x_i - Me|\, n_i.$$

Si pren valors grans, significa que els valors de la variable estan dispersos respecte de la mediana.

Exemple 2.13 Per al cas del nombre de fills:

x_i	n_i	$\lvert x_i - Me\rvert$	$\lvert x_i - Me\rvert\, n_i$
0	2	2	4
1	4	1	4
2	21	0	0
3	15	1	15
4	6	2	12
5	1	3	3
6	1	4	4
	50		42

$$D_{Me} = \frac{1}{n}\sum_{i=1}^{k} |x_i - Me|\, n_i = \frac{42}{50} = 0.84 \text{ fills.}$$

6. **Recorregut o rang mostral** (R_e): és la diferència entre els valors més alt i més baix de les observacions:

$$Re = x_{\text{máx}} - x_{\text{mín}}.$$

Com més recorregut, més dispersió.

Exemple 2.14 Per al cas del nombre de fills: $Re = 6 - 0 = 6$ fills.

7. **Recorregut interquartílic** (RQ): és la diferència entre el tercer i el primer quartil.

$$RQ = C_3 - C_1.$$

Com més RQ, més dispersió.

Exemple 2.15 Per al cas del nombre de fills: $RQ = 3 - 2 = 1$ fill.

2.3.2. MESURES DE DISPERSIÓ RELATIVES

Només considerarem el **coeficient de variació de Pearson**, que es defineix com el quocient entre la desviació típica i el valor absolut de la mitjana aritmètica:

$$CV = \frac{s}{|\overline{x}|}.$$

És adimensional i val per a comparar dues distribucions que no vénen en les mateixes unitats. Representa quantes vegades la mitjana aritmètica està continguda en la desviació típica. Com més alt és CV, més gran és la dispersió i menor la representativitat de la mitjana aritmètica.

Exemple 2.16 El coeficient de variació del nombre de fills és:

$$CV = \frac{1.12}{2.52} = 0.44$$

2.4. TIPIFICACIÓ D'UNA DISTRIBUCIÓ DE FREQÜÈNCIES

Es diu que una **variable** estadística està **tipificada** quan la seua mitjana arimètica és 0 i la seua variància (o la seua desviació típica) és 1.

Suposem que apliquem a les dades la transformació següent:

$$z_i = \frac{x_i - \overline{x}}{s_x},$$

és a dir, a cada valor de la variable, li restem la mitjana i després dividim entre la desviació típica. Es tracta, doncs, d'una transformació lineal $z_i = A + Cx_i$, on $A = -\frac{\overline{x}}{s_x}$ i $C = \frac{1}{s_x}$. Usant la propietat **c** de la mitjana i la propietat **d** de la desviació típica, és fàcil demostrar que la nova distribució de freqüències té mitjana 0 i desviació típica 1.

Aleshores direm que la mostra o la **distribució de freqüències** està **tipificada** i la transformació realitzada, l'anomenarem **tipificació**.

2.5. MESURES DE FORMA

Comparen la forma que té la representació gràfica de la distribució, bé siga l'histograma o el diagrama de barres, amb la distribució normal.

2.5.1. MESURES D'ASIMETRIA

Mesuren la simetria de la distribució. Suposem que hem representat gràficament una distribució de freqüències; tracem una perpendicular a l'eix d'abscisses per l'abscissa corresponent a $\overline{x}$. Direm que la **distribució** és **simètrica** si a ambdós costats de la perpendicular traçada existeix el mateix nombre de valors, equidistants dos a dos, i cada parell de punts equidistants amb la mateixa freqüència.

1. El **coeficient d'asimetria de Fisher** (g_1):

$$g_1 = \frac{1}{ns^3}\sum_{i=1}^{k}(x_i - \overline{x})^3 n_i = \frac{m_3}{s^3}.$$

   Si la distribució és simètrica, en el numerador tindrem tantes desviacions positives com negatives i per tant, $g_1 = 0$.

   Si $g_1 > 0$, la distribució és asimètrica positiva o asimètrica per la dreta.

   Si $g_1 < 0$, la distribució és asimètrica negativa o asimètrica per l'esquerra.

Exemple 2.17 Coeficient d'asimetria de Fisher per al nombre de fills:

x_i	n_i	$x_i - \overline{x}$	$(x_i - \overline{x})^3$	$(x_i - \overline{x})^3 n_i$
0	2	-2.52	-16.003	-32.006
1	4	-1.52	-3.512	-14.047
2	21	-0.52	-0.141	-2.953
3	15	0.48	0.11	1.658
4	6	1.48	3.242	19.451
5	1	2.48	15.253	15.253
6	1	3.48	42.144	42.144
	50			29.5

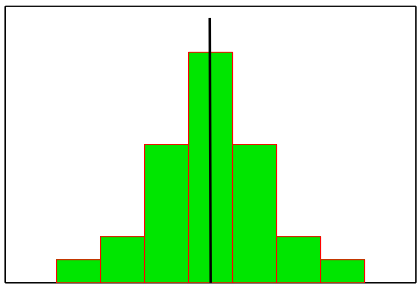
$\overline{x} = 2.52$ fills

$s_x = 1.12$ fills

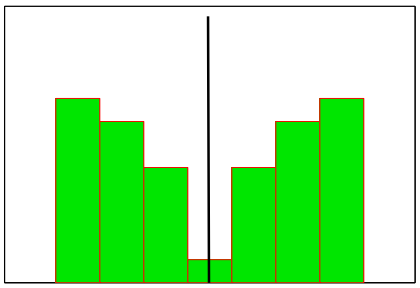
$$g_1 = \frac{29.5}{50 \times (1.12)^3} = 0.42 > 0$$

Distribució asimètrica per la d

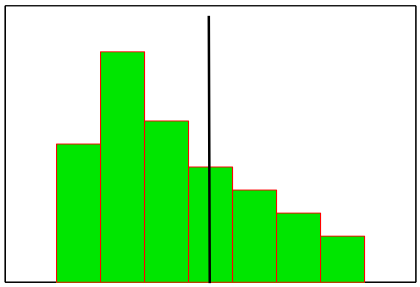
En el dibuix següent poden observar-se els diferents tipus d'asimetries:



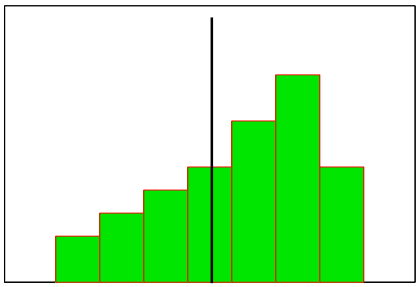
Simètrica



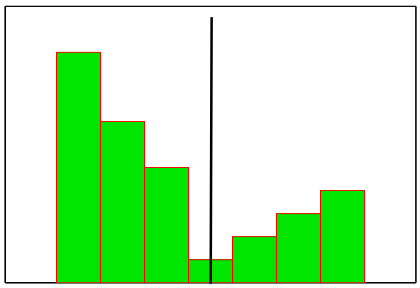
Simètrica



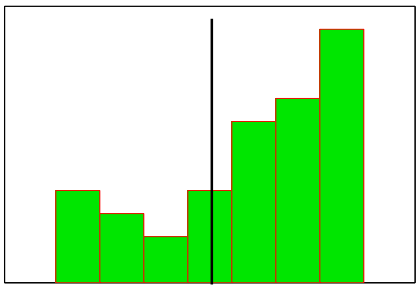
Asimètrica per la dreta



Asimètrica per l'esquerra



Asimètrica per la dreta



Asimètrica per l'esquerra

2. El **coeficient d'asimetria de Pearson** (A_s): és molt més fàcil de calcular que l'anterior, però sols és aplicable a les distribucions que només tenen una moda i que tenen forma de campana. Es defineix com:

$$A_s = \frac{\overline{x} - Mo}{s}.$$

Si la distribució és simètrica, $\overline{x} = Mo$ i per tant, $A_s = 0$. Si $A_s > 0$, la distribució és asimètrica positiva. Si $A_s < 0$, la distribució és asimètrica negativa.

Exemple 2.18 Per al cas de l'exemple 1.1: $Mo = 2$ fills, $\overline{x} = 2.52$ fills i $s = 1.12$ fills. Per tant:

$$A_s = \frac{\overline{x} - Mo}{s} = \frac{2.52 - 2}{1.12} = 0.46 > 0 \Rightarrow \text{ Distribució asimètrica positiva.}$$

2.5.2. MESURES D'APUNTAMENT O CURTOSI

Mesuren la quantitat de dades que s'agrupen en torn a la mitjana. Sols tenen sentit en les distribucions campaniformes, és a dir, unimodals simètriques o lleugerament asimètriques.

Si per a valors pròxims a la mitjana les freqüències són més altes que en la distribució normal, la gràfica serà molt apuntada en aquesta zona, i es diu que la **distribució** és **leptocúrtica**. Quan són més baixes que en la normal, direm que la **distribució** és **platicúrtica**. Finalment, quan la distribució de freqüències és igual d'apuntada que la normal, direm que és una **distribució mesocúrtica**.

El **coeficient d'apuntament** o **curtosi** (g_2) es defineix com:

$$g_2 = \frac{1}{ns^4}\sum_{i=1}^{k}(x_i - \overline{x})^4 n_i - 3 = \frac{m_4}{s^4} - 3.$$

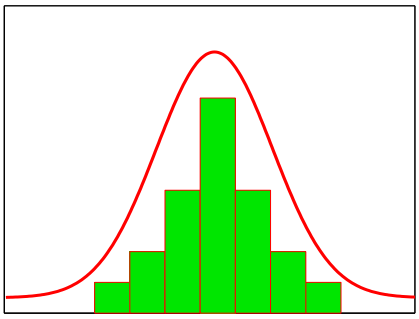
Si $g_2 > 0$, és leptocúrtica; si $g_2 < 0$, platicúrtica i si $g_2 = 0$, mesocúrtica. En la figura 2.2, pot observar-se una representació gràfica de la curtosi.

Exemple 2.19 Curtosi per al nombre de fills de l'exemple 1.1:

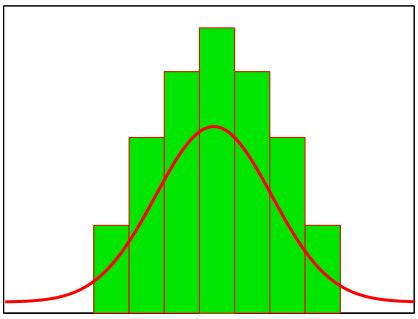
x_i	n_i	$x_i - \overline{x}$	$(x_i - \overline{x})^4$	$(x_i - \overline{x})^3 n_i$
0	2	-2.52	40.327	80.655
1	4	-1.52	5.338	21.352
2	21	-0.52	0.073	1.533
3	15	0.48	0.053	0.795
4	6	1.48	4.798	28.788
5	1	2.48	37.827	37.827
6	1	3.48	146.662	146.662
	50			317.612

$$g_2 = \frac{317.612}{50 \times (1.12)^4} - 3 = 1.037 > 0$$

Distribució leptocúrtica



Platicúrtica



Leptocúrtica

Figura 2.2: Representació de diversos tipus d'apuntament

2.6. MESURES DE CONCENTRACIÓ

Les mesures de concentració tracten de posar de manifiest el grau d'igualtat en el repartiment total dels valors de la variable. Són, per tant, indicadors del grau d'equidistribució de la variable. Aquestes mesures tenen especial aplicació a variables de tipus econòmic: rendes, salaris, etc.

Suposem que tenim n subjectes i els valors de la variable (rendes, salaris, etc.) són:

$$x_1 \leq x_2 \leq \cdots \leq x_n$$

i ens interessa estudiar fins a quin punt la suma total de valors (renda total, suma dels salaris, etc.) està equitativament repartida. Les dues situacions extremes són:

- Concentració màxima: dels n subjectes, només un rep el total i la resta, res:

$$x_1 = x_2 = \cdots = x_{n-1} = 0, \qquad x_n \neq 0.$$

- Concentració mínima: tots tenen el mateix valor:

$$x_1 = x_2 = \cdots = x_{n-1} = x_n.$$

Nota 2.4 Cal considerar que, des d'un punt de vista estadístic, els termes dispersió i concentració no són oposats. Recordem que el primer fa referència a la variabilitat de les dades respecte de la mitjana; mentre que el segon, com acabem d'assenyalar, a la no-equitat en el repartiment de la suma total de la variable.

1. **Índex de concentració de Gini** (I_{co}). Es construeix a partir de les quantitats següents:

   (a) Calculem, en primer lloc, els productes $x_i n_i$, que ens indiquen el total percebut (renda total, guanys totals, etc.) pels n_i subjectes amb valor x_i (renda, guany, etc.). Aquest producte, és anomenat **riquesa de l'i-èsim grup**.

   (b) Calculem les riqueses acumulades de la variable, que denotarem per u_i:

$$\begin{aligned} u_1 &= x_1 n_1 \\ u_2 &= x_1 n_1 + x_2 n_2 \\ u_3 &= x_1 n_1 + x_2 n_2 + x_3 n_3 \\ &\vdots \\ u_k &= x_1 n_1 + x_2 n_2 + \cdots + x_k n_k \end{aligned}$$

   (c) Les riqueses acumulades (u_i), les representem en tant per cent del total (u_k). Denotem aquests percentatges per q_i:

$$q_i = \frac{u_i}{u_k} \cdot 100.$$

   (d) Expressem les freqüències relatives acumulades en tant per cent. Denotem aquests percentatges per p_i:

$$p_i = \frac{N_i}{n} \cdot 100 = F_i \cdot 100.$$

   Una vegada efectuats aquests càlculs, es defineix l'índex de concentració de Gini a partir de la fórmula:

$$I_{co} = \frac{\sum_{i=1}^{k-1} (p_i - q_i)}{\sum_{i=1}^{k-1} p_i}.$$

   S'hi obté que $0 \leq I_{co} \leq 1$.

Podem observar que:

- Si $q_i = 0$, per a $i = 1, 2, \ldots, (k-1)$ i $q_k \neq 0$, aleshores $I_{co} = \dfrac{\sum_{i=1}^{k-1} p_i}{\sum_{i=1}^{k-1} p_i} = 1$ i la concentració és màxima.
- Si per a cada i, $q_i = p_i$, aleshores $I_{co} = 0$ i el repartiment és equitatiu, ja que cada percentatge d'individus poseeix el mateix percentatge de riquesa.

2. **Corba de Lorenz:** una forma d'estudiar gràficament la concentració és mitjançant la corba de Lorenz. Es construeix representant en l'eix d'abscisses el percentatge de freqüències acumulades (p_i) i en el d'ordenades, els percentatges acumulats del total de la variable (q_i). En unir aquests punts obtenim la corba de Lorenz. Per a una millor interpretació, se sol dibuixar un quadrat de costat 100 (en la figura 2.3 OABC) i la seua diagonal (OB).

   Noteu que:

   - Com que per a $p_i = 0$, és $q_i = 0$, la gràfica sempre passa pel punt $(0, 0)$.
   - Com que per a $p_i = 100$, és $q_i = 100$, la gràfica sempre passa pel punt $(100, 100)$.
   - Com que $p_i \geq q_i$, la gràfica sempre està situada per davall de la diagonal del quadrat (OB) o hi coincideix.
   - En el cas d'existir repartiment equitatiu, és a dir, concentració mínima ($p_i = q_i$), la corba coincideix amb la diagonal.
   - Si la concentració és màxima, la corba de Lorenz està formada pels costats del quadrat: OA i OB (observeu la figura 2.4).

   Es demostra que, aproximadament:

$$I_{co} = \frac{\text{Àrea entre la corba i la diagonal OB}}{\text{Àrea del triangle OAB}}.$$

Nota 2.5 L'índex de Gini té l'avantatge, sobre la corba de Lorenz, de resumir la informació en una sola xifra, però quan realitzem comparacions entre dues distribucions, aquest avantatge té com a contrapartida negativa que dues distribucions amb aspectes molt diferents poden tindre el mateix índex de Gini.

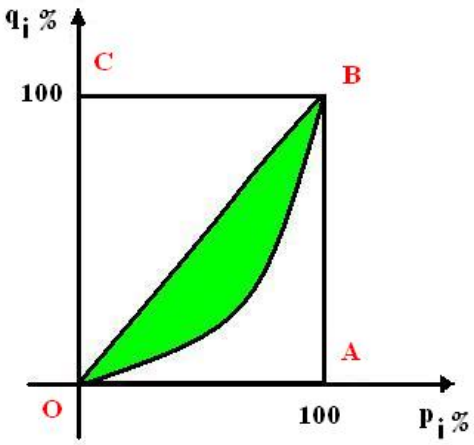


Figura 2.3: Corba de Lorenz

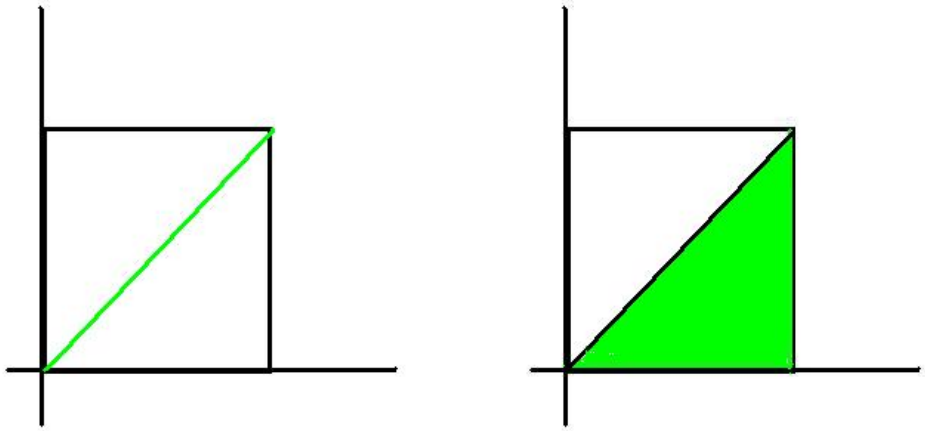


Figura 2.4: Comparació de dues corbes de Lorenz

Exemple 2.20 Està el nombre de fills molt concentrat en unes poques famílies en l'exemple 1.1?

x_i	n_i	$x_i n_i$	u_i	q_i	F_i	p_i	$p_i - q_i$
0	2	0	0	0	0.04	4	4
1	4	4	4	3.17	0.12	12	8.83
2	21	42	46	36.51	0.54	54	17.49
3	15	45	91	72.22	0.84	84	11.78
4	6	24	115	91.27	0.96	96	4.73
5	1	5	120	95.24	0.98	98	2.76
6	1	6	126	100	1	100	–
							49.59

$$I_{co} = \frac{49.59}{348} = 0.142$$

Poca concentració

2.7. PROBLEMES PROPOSATS

(1) Donada la distribució $x_i = \{45924, 45926, 45928, 45930, 45932\}$ amb freqüències respectives $n_i = \{5, 7, 5, 2, 9\}$, calcula'n la mitjana artimètica. Pots proposar-ne un mètode senzill de càlcul?

(2) La taula següent mostra la distribució dels salaris (en euros) en 2005 en la indústria turística d'un determinat país:

$l_i - L_i$	n_i
$0 - 600$	2140
$600 - 900$	1525
$900 - 1200$	845
$1200 - 1500$	950
$1500 - 1800$	1105
$1800 - 2100$	2347
$2100 - 2700$	615
$2700 - 4000$	323
> 4000	150

Calcula:

a) El salari mitjà per treballador (pren com a marca de classe de l'últim interval 5000).

b) El salari més freqüent.

c) El salari que permet ser superior a la meitat dels restants.

(3) Demostra que la mitjana aritmètica de la variable Z, obtinguda com a suma de les dades d'altres dues variables X i Y, és la suma de les mitjanes aritmètiques d'aquestes.

(4) Calcula la mediana, la moda, el primer i el tercer quartil, el quart decil i el nonagèsim centil de la distribució:

x_i	5	10	15	20	25
n_i	3	7	5	3	2

(5) Donada la distribució de freqüències següent:

x_i	0	10	20	30	40
n_i	2	4	7	5	2

Calcula:

a) Mitjana, moda, mediana, primer i tercer quartil, i quaranta-cinqué centil.

b) Variància, desviació típica, coeficient de variació, i desviació mitjana respecte de la mitjana i respecte de la mediana, i recorregut i recorregut interquartílic.

c) El coeficient d'asimetria i la curtosi.

d) Comenta els resultats.

(6) Dues empreses A i B tenen 100 treballadors cadascuna. Els salaris per dia i treballador són:

- A l'empresa A, 20 persones perceben 15 euros i 80 perceben 120 euros.
- A l'empresa B, 20 persones perceben 120 euros i 80 perceben 15 euros.

a) Calcula mitjana, variància, desviació típica i coeficient de variació en cada cas. Compara els resultats.

b) Obtín la corba de Lorenz i l'índex de concentració de Gini en cada cas. Analitza-ho i compara els resultats.

SOLUCIONS

(1) $\overline{x} = 45928.2143$

(2) *a*) $\overline{x} = 1366.35$ euros

b) $Mo = 1865.31$ euros

c) $Me = 1354.74$ euros

(4) $Me = 12.5$, $Mo = 10$, $C_1 = 10$, $C_3 = 15$, $D_4 = 10$, $P_{90} = 20$

(5) *a*) $\overline{x} = 20.5$, $Mo = 20$, $Me = 20$, $C_1 = 10$, $C_3 = 30$, $P_{45} = 20$

b) $s_x^2 = 124.75$, $s_x = 11.17$, $CV = 0.55$, $D_{\overline{x}} = 8.65$, $D_{Me} = 8.5$, $Re = 40$, $RQ = 20$

c) $A_s = 0.045$, $g_1 = -0.09848$, $g_2 = -0.65$

(6) *a*) Empresa A: $\overline{x} = 99$, $s_x^2 = 1764$, $s_x = 42$, $CV_x = 0.42$
Empresa B: $\overline{y} = 36$, $s_y^2 = 1764$, $s_y = 42$, $CV_y = 1.17$

b) Empresa A: $I_{co} = 0.84$, Empresa B: $I_{co} = 0.58$

TEMA 3
DISTRIBUCIONS BIDIMENSIONALS

3.1. INTRODUCCIÓ

Es vol fer un estudi d'acceptació de dos models d'impressores. Per a aquest fi, es consideren les vendes en una tenda durant un període de 25 dies, durant els quals les vendes foren:

Model A	0	2	2	2	1	3	3	3	3	4	4	2	3
Model B	2	1	2	2	3	1	1	1	2	0	1	1	1

Model A	3	3	3	2	3	2	4	2	2	3	3	3
Model B	1	1	2	2	1	1	1	2	2	2	2	1

En molts processos de la vida ordinària és necessari estudiar simultàniament dues característiques en una determinada població, és a dir, dues variables. L'estudi conjunt permet determinar les relacions que guarden. Suposarem, inicialment, que estem observant dues variables encara que el tractament que presentarem es podria generalitzar sense dificultat per a qualsevol nombre de variables.

Al llarg del tema usarem la notació següent:

- Representarem les variables per X i Y. En l'exemple anterior, X = nombre d'impressores del model A que es venen en un dia. Y = nombre d'impressores del model B que es venen en un dia.
- n: nombre de parells d'observacions. En l'exemple, $n = 25$.
- x_i: cada dada diferent observada en la mostra de X.
- k: nombre de valors diferents de X. En l'exemple, $k = 5$.
- y_j: cada dada diferent observada en la mostra de Y.
- h: nombre de valors diferents de Y. En l'exemple, $h = 4$.

3.2. DISTRIBUCIONS DE FREQÜÈNCIES BIVARIANTS

3.2.1. DISTRIBUCIÓ CONJUNTA

Com en el cas d'una variable, quan volem descriure conjuntament dues variables, el primer que farem serà representar les dades en una taula de freqüències.

Definició 3.1 La **freqüència absoluta conjunta** (n_{ij}) d'un parell (x_i, y_j) és el nombre de vegades que aquest apareix en la mostra.

Exemple 3.1 Per al parell $(x_1, y_3) = (0, 2)$ de l'exemple de la introducció es té que $n_{13} = 1$.

Propietat 3.1 *La suma de totes les freqüències absolutes conjuntes és el nombre total de parells d'observacions, és a dir,* $\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = n$.

Definició 3.2 La **freqüència relativa conjunta** (f_{ij}) d'un parell (x_i, y_j) és:

$$f_{ij} = \frac{n_{ij}}{n}.$$

Exemple 3.2 Per al cas de l'exemple anterior, $f_{13} = 0.04$.

Propietat 3.2 *La suma de totes les freqüències relatives conjuntes és la unitat, és a dir,* $\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = 1$.

Definició 3.3 Una **distribució de freqüències conjunta** és una taula de doble entrada on, en la primera columna, representarem, ordenats de més baix a més alt, els valors observats de la variable X i en la primera fila, els de la variable Y. Al centre, les corresponents n_{ij}, f_{ij} o ambdues.

Exemple 3.3 Per a l'exemple introductori de les impressores es té:

x_i \ y_j	0	1	2	3	$n_{i\cdot}$
0	0	0	1	0	1
1	0	0	0	1	1
2	0	3	5	0	8
3	0	8	4	0	12
4	1	2	0	0	3
$n_{\cdot j}$	1	13	10	1	25

3.2.2. DISTRIBUCIONS MARGINALS

Definició 3.4 Les **distribucions marginals** són les dues distribucions unidimensionals que podem obtindre considerant separadament les dades de cadascuna de les variables X i Y.

Definició 3.5 Les **freqüències marginals** són les que s'obtenen en les distribucions marginals. Les obtindrem a partir de les conjuntes.

(1) La **freqüència absoluta marginal per a X** ($n_{i\cdot}$) és el nombre de vegades que es repeteix el valor x_i sense tindre en compte els valors de Y (observeu la taula de l'exemple 3.3), és a dir:

$$n_{i\cdot} = \sum_{j=1}^{h} n_{ij} \quad \text{(noteu que sumem la fila } i\text{)}.$$

(2) La **freqüència absoluta marginal per a Y** ($n_{\cdot j}$) és el nombre de vegades que es repeteix el valor y_j sense tindre en compte els valors de X (observeu la taula de l'exemple 3.3), és a dir:

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij} \quad \text{(noteu que sumem la columna } j\text{)}.$$

Exemple 3.4 En l'exemple de la introducció: $x_3 = 2$; $n_{3\cdot} = 3 + 5 = 8$. D'altra banda, $y_2 = 1$; $n_{\cdot 2} = 3 + 8 + 2 = 13$.

De la mateixa forma podem definir les freqüències relatives marginals: $f_{i\cdot}$ i $f_{\cdot j}$, que es calculen a partir de les freqüències absolutes marginals.

Les distribucions d'aquestes freqüències marginals poden tabular-se de forma separada, com ho hem fet al tema anterior, o en la taula conjunta, col·locant les $n_{i\cdot}$ i les $f_{i\cdot}$ en les dues últimes columnes, i les $n_{\cdot j}$ i les $f_{\cdot j}$ en les dues últimes files.

3.2.3. DISTRIBUCIONS CONDICIONADES

A partir de la distribució de freqüències conjunta, podem definir un altre tipus de distribucions unidimensionals, tant per a X com per a Y, que definim a continuació.

Definició 3.6 Les **distribucions condicionades** són les que s'obtenen a partir de les conjuntes fixant el valor d'una de les variables.

Exemple 3.5 Nombre d'impressores venudes del model A, atés que sabem que se n'ha venut una del model B.

Definició 3.7 Les **freqüències condicionades** són les que s'obtenen en les distribucions condicionades. Les obtindrem a partir de les conjuntes.

(1) La **freqüència absoluta condicionada per a $\mathbf{X = x_i}$ donada $\mathbf{Y = y_j}$** ($n_{i(j)}$) és el nombre de vegades que es repeteix x_i quan només considerem els casos en què $Y = y_j$. És a dir, $n_{i(j)} = n_{ij}$; $(1 \leq i \leq k)$.

(2) La **freqüència absoluta condicionada per a $\mathbf{Y = y_j}$ donada $\mathbf{X = x_i}$** ($n_{(i)j}$) és el nombre de vegades que es repeteix y_j quan només considerem els casos en què $X = x_i$. És a dir, $n_{(i)j} = n_{ij}$; $(1 \leq j \leq h)$.

Propietat 3.3 $\sum_{i=1}^{k} n_{i(j)} = n_{\cdot j}$ $\quad i \quad$ $\sum_{i=1}^{h} n_{(i)j} = n_{i\cdot}$.

En les distribucions condicionades no se solen usar les freqüències absolutes, perquè, com ja sabem, depenen del nombre de dades i el nombre de dades és diferent per a cada distribució, ja que dependrà de la freqüència del valor que fixem de l'altra variable. Són molt més útils les **freqüències relatives condicionades**, que definim a continuació:

Definició 3.8 Freqüència relativa condicionada per a $X = x_i$ donada $Y = y_j$ ($f_{i(j)}$):

$$f_{i(j)} = \frac{n_{ij}}{n_{\cdot j}}.$$

Definició 3.9 Freqüència relativa condicionada per a $Y = y_j$ donada $X = x_i$ ($f_{(i)j}$):

$$f_{(i)j} = \frac{n_{ij}}{n_{i\cdot}}.$$

Exemple 3.6 Calcula la distribució de freqüències del nombre d'impressores venudes del model A, quan sabem que del model B s'ha venut una impressora. És a dir, calcula la disribució de freqüències de X condicionada que $Y = 1$, o siga, condicionada a y_2.

x_i	$n_{i(2)}$	$f_{i(2)}$
0	0	0
1	0	0
2	3	0.23
3	8	0.62
4	2	0.15
	13	1

Nota 3.1 Si la taula resulta molt gran caldrà agrupar una o ambdues variables en intervals de classe, de la mateixa manera que hem vist al tema 1. En aquest cas, totes les definicions que hem vist en aquest tema, es generalitzen como ho vàrem fer al tema 1.

3.2.4. INDEPENDÈNCIA ESTADÍSTICA

Des d'un punt de vista exclusivament intuïtiu, podem dir que dues variables són independents quan en fixar el valor d'una no canvia la distribució de freqüències de l'altra. Més precisament:

Definició 3.10 Direm que X i Y són **variables independents** estadísticament quan totes les freqüències relatives condicionades són iguals a les corresponents freqüències marginals. És a dir:

$$f_{i(j)} = f_{i\cdot}; \quad \forall j = 1, \ldots, h \quad \text{i} \quad f_{(i)j} = f_{\cdot j}; \quad \forall i = 1, \ldots, k.$$

Definició 3.11 Direm que X i Y són **variables independents** estadísticament quan la freqüència relativa conjunta és igual al producte de les freqüències relatives marginals. És a dir:

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}; \quad \forall i = 1, \ldots, k \quad \text{i} \quad \forall j = 1, \ldots, h$$

o, equivalentment:

$$\frac{n_{ij}}{n} = \frac{n_{i\cdot}}{n} \cdot \frac{n_{\cdot j}}{n}; \quad \forall i = 1, \ldots, k \quad \text{i} \quad \forall j = 1, \ldots, h.$$

3.3. REPRESENTACIÓ GRÀFICA: DIAGRAMA DE DISPERSIÓ

Com en el cas univariant, la forma de la distribució conjunta s'aprecia a primera vista, i es reté més fàcilment en la memòria, amb una adequada representació gràfica.

El **diagrama de dispersió** (també anomenat **núvol de punts**) s'obté representant cada parell observat (x_i, y_j) com un punt en el plan cartesià. Sol utilitzar-se amb les dades sense agrupar. Si les dades estan agrupades per intervals, prenem les marques de classe.

És el tipus de gràfic més útil, ja que ens permet visualitzar la relació entre ambdues variables.

Exemple 3.7 Com podem observar en el diagrama de dispersió per a l'exemple de les impressores, en augmentar X disminueix Y.

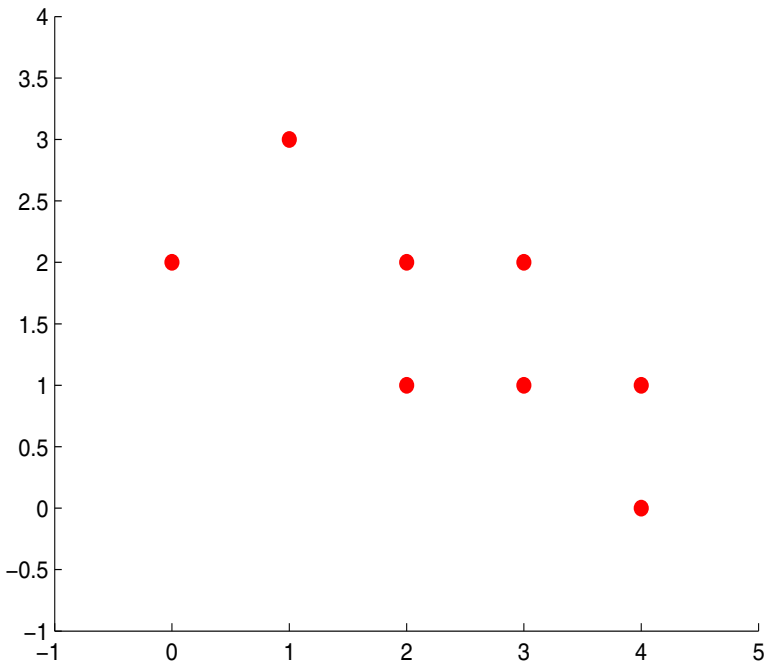
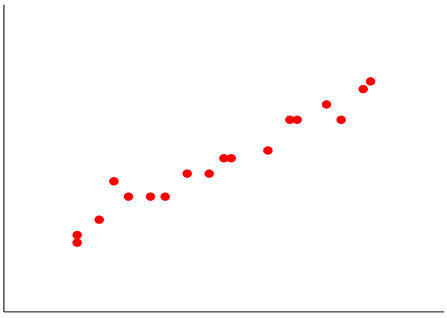


Figura 3.1: Diagrama de dispersió de l'exemple 3.3

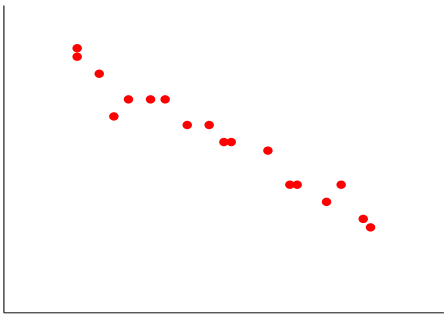
Alguns casos que se solen presentar en la pràctica són els següents:

1. Els punts s'agrupen al voltant d'una recta $y \simeq a + bx$.

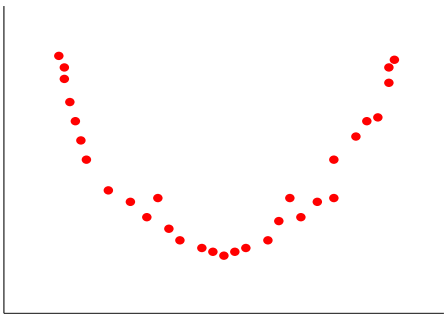
   a) Pendent positiu ($b > 0$): relació lineal directa.



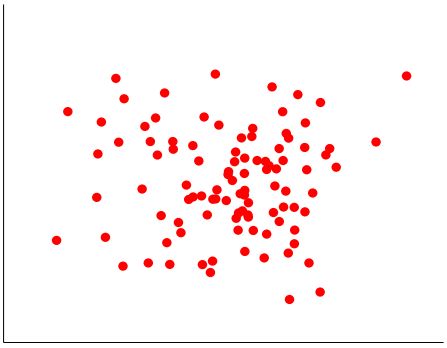
   b) Pendent negatiu ($b < 0$): relació lineal inversa.



2. Els punts s'agrupen al voltant d'una paràbola $y \simeq ax^2+bx+c$: relació quadràtica.



3. No s'aprecia cap relació.



3.4. MESURES DESCRIPTIVES D'UNA DISTRIBUCIÓ BIDIMENSIONAL

3.4.1. MOMENTS

De la mateixa forma que es defineixen els moments en les distribucions unidimensionals, també es poden definir en les distribucions bidimensionals. De nou, algun cas particular ens donarà certa informació sobre la distribució de freqüències i, en general, podrem afirmar que els moments la caracteritzen.

1. Moments respecte a l'origen: es defineix el **moment d'ordre r, s respecte a l'origen** (a_{rs}) $(r = 0, 1, \ldots; s = 0, 1, \ldots)$ com:

$$a_{rs} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{h} x_i^r y_j^s n_{ij}.$$

Alguns casos particulars interessants són:

- $a_{00} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{h} x_i^0 y_j^0 n_{ij} = \frac{n}{n} = 1.$
- $a_{10} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{h} x_i^1 y_j^0 n_{ij} = \frac{1}{n} \sum_{i=1}^{k} x_i n_{i\cdot} = \overline{x}$, que és la mitjana marginal de la variable X.
- $a_{01} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{h} x_i^0 y_j^1 n_{ij} = \frac{1}{n} \sum_{j=1}^{h} y_j n_{\cdot j} = \overline{y}$, que és la mitjana marginal de la variable Y.

2. Moments centrals o respecte a la mitjana: es defineix el **moment d'ordre r, s respecte a la mitjana** (m_{rs}) $(r = 0, 1, \ldots; s = 0, 1, \ldots)$ com:

$$m_{rs} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^r(y_j - \overline{y})^s n_{ij}.$$

Alguns casos particulars interessants són:

- $m_{00} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^0(y_j - \overline{y})^0 n_{ij} = \frac{n}{n} = 1.$
- $m_{10} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^1(y_j - \overline{y})^0 n_{ij} = \frac{1}{n}\sum_{i=1}^{k} x_i n_{i\cdot} - \frac{1}{n}\sum_{i=1}^{k}\overline{x} n_{i\cdot} = \overline{x} - \overline{x} = 0.$
- $m_{01} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^0(y_j - \overline{y})^1 n_{ij} = \frac{1}{n}\sum_{j=1}^{h} y_j n_{\cdot j} - \frac{1}{n}\sum_{j=1}^{h}\overline{y} n_{\cdot j} = \overline{y} - \overline{y} = 0.$
- $m_{20} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^2(y_j - \overline{y})^0 n_{ij} = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_{i\cdot} = s_x^2.$
- $m_{02} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})^0(y_j - \overline{y})^2 n_{ij} = \frac{1}{n}\sum_{j=1}^{h}(y_j - \overline{y})^2 n \cdot j = s_y^2.$

3.4.2. MESURES DE DEPÈNDENCIA LINEAL

En l'estudi conjunt de dues variables, el que ens interessa principalment és saber si existeix algun tipus de relació entre aquestes variables. En l'apartat anterior, amb la representació gràfica del diagrama de dispersió, hem pogut fer-nos una primera idea de si hi existeix algun tipus de relació. En aquesta secció, presentem mesures descriptives que ens permetran analitzar si hi existeix alguna relació de tipus lineal, és a dir, de la forma $Y = a + bX$.

Definició 3.12 Covariància (s_{xy}). Es defineix com:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h}(x_i - \overline{x})(y_j - \overline{y})n_{ij}.$$

És a dir, $s_{xy} = m_{11}$.

- Si hi ha relació lineal directa (a valors grans de X corresponen valors grans de Y), aleshores $s_{xy} > 0$ i és gran en valor absolut.
- Si hi ha relació lineal inversa (a valors grans de X corresponen valors xicotets de Y), aleshores $s_{xy} < 0$ i és gran en valor absolut.
- Si no hi ha relació lineal, aleshores $s_{xy} \simeq 0$.

- La fórmula següent ens permet calcular la covariància d'una forma més senzilla:

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j n_{ij} - \overline{xy} = a_{11} - a_{10}a_{01}.$$

Les propietats més importants de la covariància són:

(1) Si a tots els valors de la variable X sumem una constant C, i a tots els valors de la variable Y sumem una constant C', la covariància no varia. És a dir:

$$z_i = C + x_i, \quad t_j = C' + y_j \Rightarrow s_{zt} = s_{xy}.$$

(2) Si tots els valors de la variable X els multipliquem per una constant C, i tots els valors de la variable Y els multipliquem per una constant C', la covariància queda multiplicada pel producte de les constants. És a dir:

$$z_i = C \cdot x_i, \quad t_j = C' \cdot y_j \Rightarrow s_{zt} = CC' s_{xy}.$$

(3) Com a corol·lari de totes dues propietats anteriors, si considerem les transformacions lineals $z_i = a + bx_i$ i $t_j = a' + b'y_j$, on a, b, a', b' són constants qualssevol, aleshores $s_{zt} = bb's_{xy}$.

Exemple 3.8 Per al cas dels models d'impressores, es té: $\overline{x} = 2.6$ impressores, $\overline{y} = 1.44$ impressores i, en conseqüència:

$$s_{xy} = \frac{0 \times 0 \times 0 + 0 \times 1 \times 0 + \cdots}{25} - (2.6 \times 1.44) = -0.344.$$

Suposem ara que cada impressora del model A val 120 euros i que el preu d'una impressora del model B és de 150 euros. Aleshores, la quantitat invertida en la compra d'impressores model A i impressores model B la podem obtindre posant:

$$Z = 120\,X, \quad T = 150\,Y.$$

D'aquesta manera, Z representa els diners invertits en la compra d'impressores del model A, i T els diners invertits en la compra d'impressores del model B, i s'obté així:

$$s_{zt} = 120 \cdot 150 \cdot s_{xy} = 6192.$$

L'inconvenient principal de la covariància com a mesura de la relació lineal entre dues variables és la depèndencia respecte de les unitats i, en conseqüència, respecte dels canvis d'escala. Definim a continuació una mesura que no està afectada per les unitats i, en conseqüència, tampoc pels canvis d'unitats de mesura.

Definició 3.13 Coeficient de correlació (r_{xy}). Es defineix com:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Les propietats principals són:

(1) És adimensional.

(2) $-1 \leq r_{xy} \leq 1$.

(3) Si hi ha relació lineal directa, aleshores $r_{xy} > 0$ i pròxim a 1.

(4) Si hi ha relació lineal inversa, aleshores $r_{xy} < 0$ i pròxim a -1.

(5) Si no hi ha relació lineal, aleshores $r_{xy} \simeq 0$.

Exemple 3.9 Per al cas dels dos models d'impressores:

$$s_x = \sqrt{0.8} = 0.89, \quad s_y = \sqrt{0.41} = 0.64 \;\Rightarrow\; r_{xy} = \frac{-0.344}{0.89 \times 0.41} = -0.9427.$$

Finalitzarem el tema donant una condició necessària per a la independència estadística.

Teorema 3.1 *Si X i Y són independents, aleshores $s_{xy} = 0$.*

Demostració: Aplicant la primera propietat de la covariància, i tenint en compte que $f_{ij} = \dfrac{n_{ij}}{n}$, podem escriure:

$$s_{xy} = \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j f_{ij} - \overline{xy}.$$

Ara bé, com que les variables són independents, la freqüència relativa conjunta és el producte de les freqüències relatives marginals. Per tant:

$$\begin{aligned} s_{xy} &= \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j f_{ij} - \overline{xy} = \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j f_{i\cdot} f_{\cdot j} - \overline{xy} = \\ &= \sum_{i=1}^{k} x_i f_{i\cdot} \sum_{j=1}^{h} y_j f_{\cdot j} - \overline{xy} = \overline{xy} - \overline{xy} = 0. \end{aligned}$$

Corol·lari 3.1 *Si X i Y són independents, aleshores $r_{xy} = 0$.*

Demostració: És evident, tenint en compte el resultat anterior, ja que:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

Nota 3.2 Existeixen casos en què la covariància entre dues variables pot ser zero sense que aquestes siguen independents.

3.5. PROBLEMES PROPOSATS

(1) Les següents són les qualificacions obtingudes per 25 alumnes de 2n curs de la Diplomatura en Ciències Empresarials en les assignatures Matemàtiques i Comptabilitat:

Matemàtiques	4	5	5	5	6	6	7	7	7	7	7	7	7
Comptabilitat	3	5	5	6	7	7	7	7	7	7	8	8	8

Matemàtiques	8	8	8	8	8	8	9	9	9	9	9	10
Comptabilitat	7	7	8	8	8	8	8	8	8	10	10	10

a) Obtín la taula de freqüències conjunta.

b) Quina proporció d'alumnes obté més d'un 5 en ambdues assignatures? Quina proporció d'alumnes obté més d'un 5 en Matemàtiques? Quina proporció d'alumnes obté més d'un 5 en Comptabilitat?

c) Són independents les qualificacions en Matemàtiques i Comptabilitat?

d) Representa el diagrama de dispersió i comenta'l.

e) Calcula el coeficient de correlació i interpreta'n el resultat.

(2) Es pretén fer un estudi sobre la utilització d'un escàner en una determinada oficina. Per a aquest fi es van mesurar, durant un dia, els minuts transcorreguts entre les successives utilitzacions (X) i el nombre de pàgines escanejades (Y), i es van obtindre els resultats següents:

X	8	8	3	5	7	8	6	5	8	8	8	7	7	8	8
Y	2	7	2	7	2	7	7	7	2	7	11	11	7	7	11

X	8	8	9	8	14	9	11	11	9	9	11	9	9
Y	11	19	7	19	7	7	19	7	7	11	7	19	19

a) Escriu la distribució de freqüències conjunta. Quin és el percentatge de vegades que transcorren més de 8 minuts des de la utilització anterior de l'escàner i s'escanegen menys d'11 pàgines?

b) Escriu les distribucions de freqüències marginals. Quantes vegades s'escanegen com a màxim 11 pàgines? Quantes pàgines s'escanegen com a màxim el 80 % de les ocasions?

c) Troba la distribució de freqüències del nombre de pàgines escanejades condicionada que hagen transcorregut 8 minuts entre utilitzacions successives.

d) Dibuixa el diagrama de dispersió.

(3) De la distribució (x_i, y_j, n_{ij}), per a 100 observacions, es té que:

$$\sum_i x_i n_{i\cdot} = 500, \quad \sum_j y_j n_{\cdot j} = 1000, \quad \sum_i \sum_j x_i y_j n_{ij} = 6000.$$

a) Quant val la covariància entre X i Y?

b) I la covariància entre U i Z, si $X = \dfrac{3U+4}{2}$ i $Y = \dfrac{2Z+3}{2}$?

(4) L'assignatura Comptabilitat Financera consta de dues parts, una de teòrica i una altra de pràctica. A l'examen final es varen presentar 10 alumnes, que van obtindre les qualificacions següents:

Teoria	5	7	6	9	3	1	2	4	6	8
Pràctica	6	5	8	6	4	2	1	3	7	8

Calcula la covariància i el coeficient de correlació lineal. Dibuixa el núvol de punts. Comenta'n els resultats.

(5) Entre els empleats d'una empresa es disposa d'informació sobre els seus salaris (en milers d'euros) i el nombre de vehicles de motor que s'han adquirit en els últims 5 anys:

Vehicles / Salaris	0	1	2	3
$[18, 27[$	2	3	1	0
$[27, 45]$	0	0	2	2

Calcula:

a) El percentatge d'empleats que cobra menys de 27000 euros i que té més d'un vehicle.

b) La covariància.

c) L'ajuda mitjana per empleat si l'empresa dóna una ajuda de 100 euros per a l'adquisició dels vehicles a tots els empleats (adquirisquen o no vehicle) més 300 euros per cada vehicle adquirit.

d) La covariància entre l'ajuda i el salari.

SOLUCIONS

(1) *a*)

x_i \ y_j	3	5	6	7	8	9	10	$n_{i\cdot}$
4	1							1
5		2	1					3
6				2				2
7				4	3			7
8				2	4			6
9					3		2	5
10							1	1
$n_{\cdot j}$	1	2	1	8	10	0	3	25

b) 84 %, 84 %, 88 %

c) No són independents.

e) $r_{xy} = 0.8782$

(2) a.1)

x_i \ y_j	2	7	11	19	$n_{i\cdot}$
3	1				1
5		2			2
6		1			1
7	1	1	1		3
8	2	4	3	2	11
9		3	1	2	6
11		2		1	3
14		1			1
$n_{\cdot j}$	4	14	5	5	28

a.2) 21.43 %

b.1) Mireu la taula anterior.

b.2) 23

b.3) 23

c)

$y_{(5)j}$	$n_{(5)j}$
2	2
7	4
11	3
19	2

(3) a) $s_{xy} = 10$

b) $s_{uz} = \dfrac{20}{3}$

(4) $s_{xy} = 4.6$, $r_{xy} = 0.8$

(5) a) 10 %

b) $s_{xy} = 5.4$

c) 550 €

d) 1620 €

TEMA 4

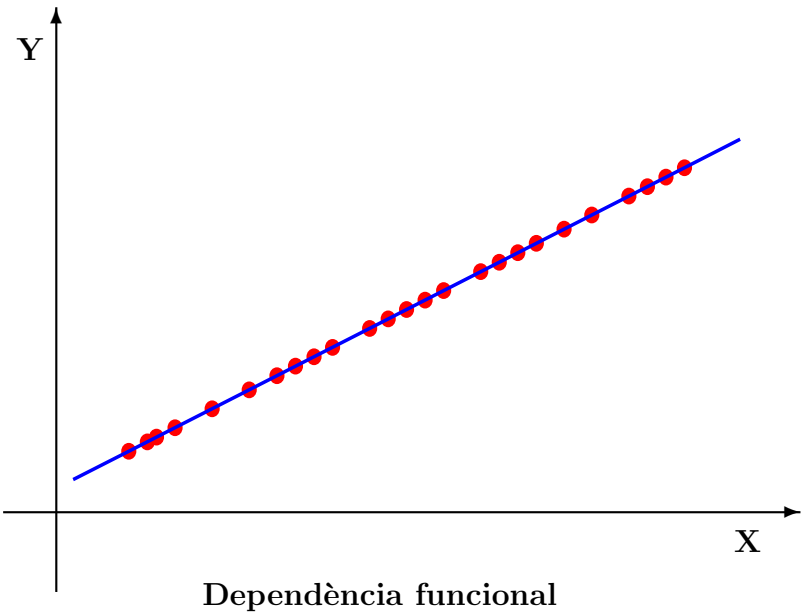
REGRESSIÓ I CORRELACIÓ LINEAL

4.1. INTRODUCCIÓ. MÈTODE DELS MÍNIMS QUADRATS

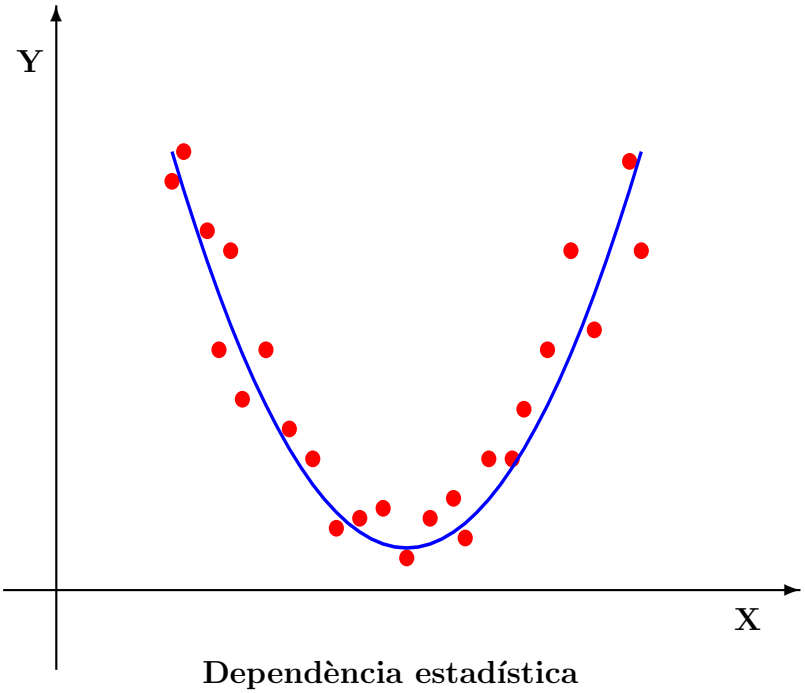
En el tema anterior vàrem veure que el diagrama de dispersió o núvol de punts ens permet visualitzar la relació entre dues variables X i Y. En representar el diagrama de dispersió podem trobar les situacions següents:

- Distribucions estadístiques per a les quals el núvol de punts es disposa de tal forma que existeix una funció matemàtica els punts de la qual són una part de la seua representació gràfica.
- Distribucions estadístiques per a les quals el núvol de punts, sense coincidir exactament amb la gràfica d'una funció matemàtica, s'hi aproxima encara que siga poc.
- Distribucions estadístiques per a les quals el núvol de punts presenta un aspecte de tal manera que no existeix concentració de punts pròxima a cap gràfica d'una funció matemàtica, i es distribueix d'una forma uniforme en una regió del pla.

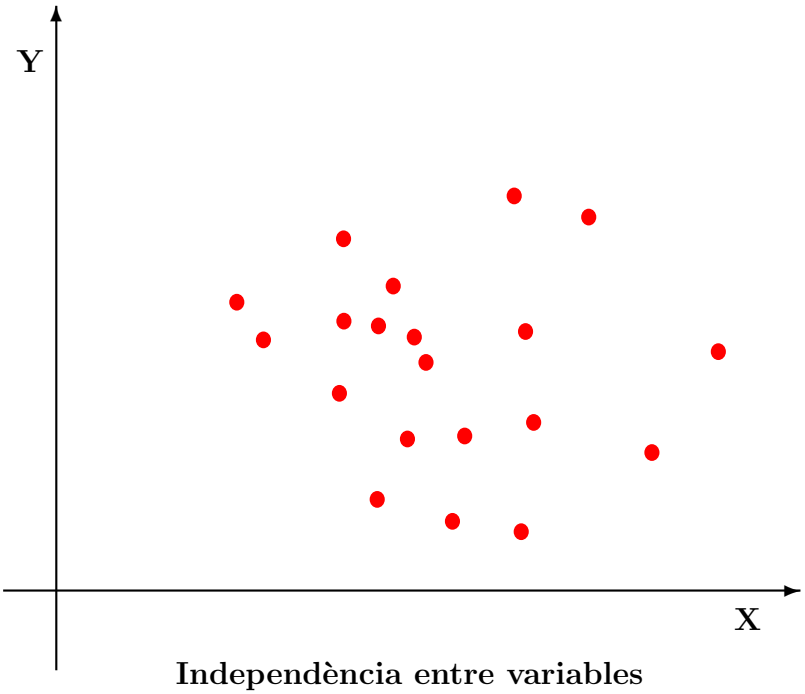
En el primer cas es diu que hi ha una **dependència funcional** o exacta entre les variables X i Y, és a dir, existeix una funció matemàtica de manera que $Y = f(X)$. En el segon cas es diu que hi ha una **dependència estadística** o aproximada entre ambdues variables: $Y \simeq f(X)$. I en l'últim cas diem que les **variables** són **independents**.



Dependència funcional



Dependència estadística



Independència entre variables

Les **tècniques de regressió** s'ocupen del segon cas que hem citat anteriorment i tenen per objecte modelitzar, és a dir, trobar una funció que aproxime el màxim possible la relació de dependència estadística entre variables i predir-ne els valors d'una (Y) a partir dels valors de l'altra (o les altres): (X o $X_1, X_2, \ldots, X_n$). La variable (o variables) coneguda, l'anomenarem **variable(s) independent(s) o explicativa(ves)**, i la variable que volem predir, **variable dependent o explicada**.

Anomenarem **regressió de Y sobre X** la funció que explica la variable Y (dependent) per a cada valor de la variable X (independent):

$$Y \simeq f(X).$$

Diem que la regressió és:

- **Lineal**, quan el model o funció de regressió seleccionada és una recta. En qualsevol altre cas l'anomenarem **regressió no lineal**.
- **Simple**, quan sols tenim una variable independent. **Múltiple**, quan tenim dues o més variables independents.

El procediment que seguirem per a efectuar la regressió serà el següent:

1) Elegir un tipus de funció o corba que creguem que millor relaciona ambdues variables. Açò, ho podrem fer observant el núvol de punts.

2) Obtindre l'equació de la corba entre les infinites d'aquest tipus que hi ha en el pla, que millor s'adapte al conjunt de punts. L'objectiu d'obtindre aquesta equació és predir el valor de la variable Y per a un valor concret, x_0, de la variable X.

3) Obtindre una mesura del grau d'aquesta associació o correlació. Açò ens dóna la fiabilitat de les prediccions que farem amb aquesta equació.

Els dos primers passos s'engloben dins del que es coneix com a **teoria de la regressió**, mentre que el tercer és el que es coneix com a **teoria de la correlació**.

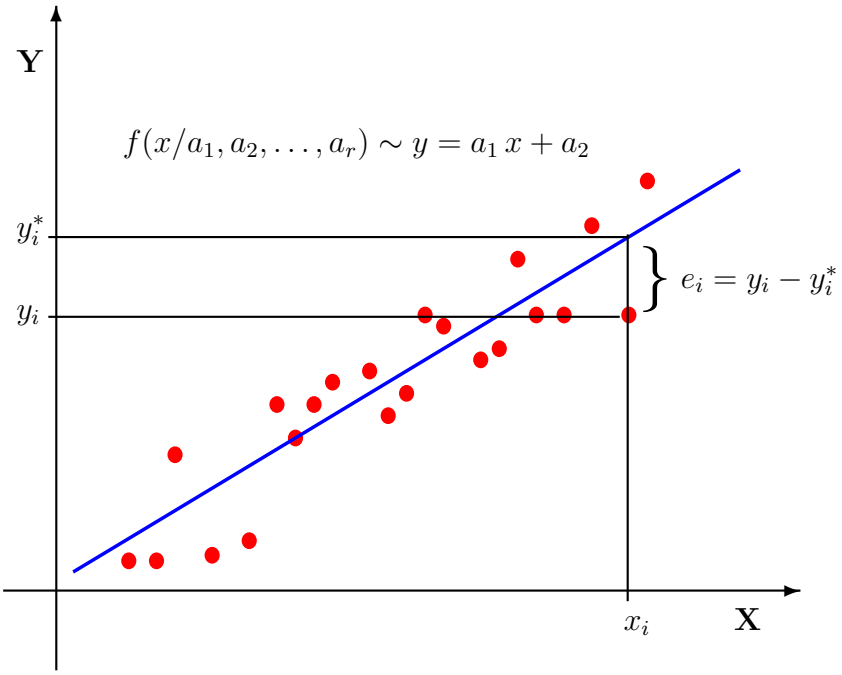
El problema que planteja el segon pas, l'obtenció de la funció, es coneix com a **problema de l'ajustament**, i es poden usar diferents mètodes matemàtics per tal de resoldre'l, com per exemple: el dels mínims quadrats, el dels polinomis ortogonals, el dels moments, el de la corba logística, etc. Nosaltres sols desenvoluparem el primer.

Nota 4.1 En aquest tema, només considerarem la mostra original, sense ordenar ni agrupar en una taula de freqüències, és a dir:

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

4.1.1. EL MÈTODE DELS MÍNIMS QUADRATS

Donats els punts $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, suposem que hem elegit una funció $y = f(x/a_1, a_2, \ldots, a_r)$ que volem ajustar a aquest conjunt de punts i en la qual intervenen r paràmetres $(a_1, a_2, \ldots, a_r)$. Considerem el núvol de punts corresponent:



Per a cada valor x_i de X, tenim dos valors de Y:

- El **valor observat**, y_i, en la mostra (o en el núvol de punts).
- El **valor teòric**, y_i^* (en general distint de l'anterior), que s'obté en substituir x_i per x en la funció, és a dir, $y_i^* = f(x_i/a_1, a_2, \ldots, a_r)$ ($y_i^* = a_1\, x_i + a_2$, en el cas lineal).

Així, per a cada x_i tenim una diferència entre tots dos valors de Y. Aquesta diferència s'anomena **residu**(e_i):

$$e_i = y_i - y_i^*.$$

El **mètode dels mínims quadrats** consisteix a determinar els paràmetres $(a_1, a_2, \ldots, a_r)$ de tal forma que la suma dels residus al quadrat siga mínima. És a dir, busquem minimitzar l'expressió:

$$\Psi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} \left(y_i - f(x_i/a_1, a_2, \ldots, a_r)\right)^2.$$

Nota 4.2 Observeu que minimitzem la suma de les distàncies verticals dels punts a la funció que pretenem aproximar, és a dir, les desviacions, al quadrat, dels valors y_i que realment té la variable respecte dels valors y_i^* que ens subministra el model que volem aproximar. Es considera el quadrat d'aquesta diferència perquè les desviacions, realment, sumen i no es compensen les que es produeixen per defecte amb les que es produeixen per excés.

La teoria de l'anàlisi matemàtica ens diu que la condició necessària per a obtindre el mínim és que les primeres derivades parcials respecte a cada un dels paràmetres s'anul·len, és a dir:

$$\begin{cases} \dfrac{\partial \Psi(a_1, a_2, \ldots, a_r)}{\partial a_1} = 0 \\ \dfrac{\partial \Psi(a_1, a_2, \ldots, a_r)}{\partial a_2} = 0 \\ \vdots \\ \dfrac{\partial \Psi(a_1, a_2, \ldots, a_r)}{\partial a_r} = 0 \end{cases}$$

Resolent aquest sistema, denominat **sistema d'equacions normals**, queden determinats els paràmetres $(a_1, a_2, \ldots, a_r)$, així com la funció corresponent.

4.2. MODEL DE REGRESSIÓ LINEAL SIMPLE

4.2.1. RECTA DE REGRESSIÓ

En el model de regressió lineal simple la funció elegida per a aproximar la relació entre les variables és una recta, és a dir, una funció de la forma $y = a + b\,x$, on a i b són els paràmetres que hem de determinar. Aquesta recta s'anomena **recta de regressió de Y sobre X**.

A continuació, en deduirem l'equació usant el mètode dels mínims quadrats: donat un valor x_i de X, tenim els corresponents valors de Y, l'observat y_i, i el teòric $y_i^* = a + b\,x_i$. Així doncs, hem de minimitzar:

$$\Psi = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^*)^2 = \sum_{i=1}^{n} \left(y_i - (a + b x_i)\right)^2 .$$

Derivant respecte als paràmetres a i b i igualant a zero:

$$\begin{cases} \dfrac{\partial \Psi(a,b)}{\partial a} = -2 \displaystyle\sum_{i=1}^{n} (y_i - a - b\,x_i) = 0, \\ \dfrac{\partial \Psi(a,b)}{\partial b} = -2 \displaystyle\sum_{i=1}^{n} (y_i - a - b\,x_i)\, x_i = 0, \end{cases}$$

obtenim un sistema de dues equacions normals i dues incògnites a i b. Aquest sistema pot escriure's com:

$$\begin{cases} \sum_{i=1}^{n} a + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \,, \\ a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \,. \end{cases} \tag{4.1}$$

Aïllant a de la primera equació i operant, s'obté que:

$$n\,a = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i \Rightarrow a = \frac{1}{n} \sum_{i=1}^{n} y_i - b\,\frac{1}{n} \sum_{i=1}^{n} x_i \Rightarrow a = \overline{y} - b\,\overline{x}.$$

Substituint el valor obtingut per a a en la segona equació del sistema (4.1) i tenint en compte que $\sum_{i=1}^{n} x_i = n\,\overline{x}$, s'obté que:

$$(\overline{y} - b\,\overline{x}) \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

$$\overline{y} \sum_{i=1}^{n} x_i - b\,\overline{x} \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

$$\overline{y}\,n\,\overline{x} - b\,\overline{x}\,n\,\overline{x} + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2,$$

$$b \left(\sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2 \right) = \sum_{i=1}^{n} x_i^2 - n\,\overline{x}\,\overline{y}\,. \tag{4.2}$$

Dividint per n ambdós membres de (4.2) i operant:

$$b \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{x}^2 \right) = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \overline{xy}$$

$$b\,s_x^2 = s_{xy}$$

$$b = \frac{s_{xy}}{s_x^2}\,.$$

Per tant, la recta de regressió de Y sobre X és:

$$\boxed{y = a + b\,x}$$

on:

$$\boxed{b = \frac{s_{xy}}{s_x^2} \quad \text{i} \quad a = \overline{y} - b\,\overline{x}}$$

El pendent de la recta de regressió de Y sobre X (paràmetre b) es denomina **coeficient de regressió de Y sobre X**.

Aplicant un raonament anàleg a l'anterior podem obtindre l'expressió de la **recta de regressió de X sobre Y**. En aquest cas, l'equació quedaria establerta per:

$$\boxed{x = a' + b'\,y}$$

on:

$$\boxed{b' = \frac{s_{xy}}{s_y^2} \quad \text{i} \quad a' = \overline{x} - b'\,\overline{y}}$$

El pendent de la recta de regressió de X sobre Y (paràmetre b') es denomina **coeficient de regressió de X sobre Y**.

Nota 4.3 La recta de regressió de X sobre Y no s'obté aïllant la X de la recta de regressió de Y sobre X.

Exemple 4.1 La despesa dels consumidors d'un país en béns i serveis (Y) i la renda corresponent (X) (ambdues en milions d'euros), en deu anys, han sigut:

x_i	5.4	6	7.2	8.4	9	10.2	11.4	12.6	15	16.2
y_i	3.6	3.6	4.2	4.8	5.4	6	6.6	7.2	9	9.6

Per a calcular la recta de regressió de la despesa dels consumidors (Y) en funció de la seua renda (X), construïm la taula que apareix a continuació, a partir de la qual ens resultarà més senzill calcular els paràmetres dels quals depén la recta de regressió:

	x_i	y_i	x_i^2	$x_i y_i$
	5.4	3.6	29.16	19.44
	6	3.6	36	21.6
	7.2	4.2	51.84	30.24
	8.4	4.8	70.56	40.32
	9	5.4	81	48.6
	10.2	6	104.4	61.2
	11.4	6.6	129.96	75.24
	12.6	7.2	158.76	90.72
	15	9	225	135
	16.2	9.6	262.44	155.52
Sumes:	101.4	60	1149.08	677.88

$$\overline{x} = \frac{101.4}{10} = 10.14, \quad \overline{y} = \frac{60}{10} = 6$$

$$s_x^2 = \frac{1149.08}{10} - (10.14)^2 = 12.0884$$

$$s_{xy} = \frac{677.88}{10} - 10.14 \times 6 = 6.948$$

$$b = \frac{6.948}{12.0884} = 0.5748$$

$$a = 6 - 0.5748 \times 10.14 = 0.1715$$

Així, la recta de regressió és:

$$\boxed{y = 0.1715 + 0.5748\, x}$$

Algunes propietats de les rectes de regressió:

- Les dues rectes de regressió es tallen en el punt $(\overline{x}, \overline{y})$, el qual es denomina **centre de gravetat** de la distribució conjunta.
- Les seues equacions en la forma punt-pendent queden establertes per:

Y sobre X: $\boxed{y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})}$ $\quad$ X sobre Y: $\boxed{x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})}$

- Tant el signe de b com el de b' coincideixen amb el signe de la covariància (ja que les variàncies són sempre positives). Una covariància positiva ens conduirà a dos coeficients de regressió positius i així els pendents de les rectes de regressió seran positius i donaran lloc a rectes creixents. Tanmateix, una covariància negativa ens conduirà a dos pendents negatius i així les rectes de regressió seran decreixents. En cas que la covariància siga zero, aquestes seran paral·leles als eixos coordenats i perpendiculars entre si.

4.2.2. MESURES DE LA BONDAT D'AJUSTAMENT. CORRELACIÓ

Recordem que per a cada valor x_i de X podem calcular la diferència (el residu) entre el valor observat de Y, y_i, en el núvol de punts i el corresponent valor teòric, y_i^*, obtingut en la recta de regressió. Si tots els punts del núvol estan sobre la recta, els residus valen zero i, en conseqüència, la dependència és funcional i, per tant, el grau de dependència és el màxim possible. A mesura que s'allunyen els punts

observats de la funció (és a dir, a mesura que els residus augmenten) anem perdent intensitat en la dependència.

En aquesta secció definirem alguns paràmetres que ens donaran una mesura d'aquest grau d'intensitat en la dependència.

Definició 4.1 Es defineix la **variància residual** com la mitjana aritmètica de tots els residus elevats al quadrat:

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Si la variància residual és gran, els residus són grans i la dependència és xicoteta. Per tant, l'ajustament és roín. Si la variància residual és xicoteta (prop de zero), la dependència és gran i aleshores l'ajustament és bo.

Nota 4.4 És fàcil demostrar que la mitjana aritmètica dels residus en la regressió lineal de Y sobre X és zero, és a dir, $\overline{e}^2 = 0$. Per tant, la variància residual rep aquest nom per ser la variància dels residus.

Definició 4.2 Anomenarem **variància deguda a la regressió** la variància dels valors teòrics, és a dir, dels y_i^*. Tenint en compte que la mitjana aritmètica d'aquests és la mateixa que la dels valors observats, és a dir: $\overline{y^*} = \overline{y}$, la variància deguda a la regressió és:

$$s_{y^*}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i^* - \overline{y})^2$$

i pot provar-se que:

$$s_y^2 = s_e^2 + s_{y^*}^2,$$

és a dir, la variància total de la variable Y és la suma de dues variàncies: la variància de Y^*, que representa la part de la dispersió o variabilitat de la variable Y explicada per la regressió (és a dir, per la relació lineal amb la variable X), i la variància residual, que representa la part de la varibilitat no explicada per la regressió.

Així doncs, quan augmenta la variància deguda a la regressió, disminueix la variància residual i l'ajustament és bo. I al contrari, quan disminueix la variància deguda a la regressió, augmenta la variància residual i l'ajustament és roín.

La variància deguda a la regressió serveix per a veure en quina mesura millora la descripció d'una variable a través de l'altra.

El problema d'utilitzar la variància residual és que queda afectada per les unitats de mesura i això impossibilita la comparació de la dependència entre grups de variables. Tenint en compte la relació entre els difrents tipus de variàncies, podem obtindre una mesura relativa (és a dir, que no depenga de les unitats) que estiga entre 0 i 1, per a la bondat d'ajustament dividint la variància deguda a la regressió entre la variància total de Y. En la definició següent, precisem amb més detall aquest concepte:

Definició 4.3 Es defineix el **coeficient de determinació** (R^2) com:

$$R^2 = \frac{s^2_{y^*}}{s^2_y} \quad \text{o bé,} \quad R^2 = 1 - \frac{s^2_e}{s^2_y}.$$

El coeficient de determinació (multiplicat per cent) representa el percentatge de la variabilitat de Y explicada per la recta de regressió, és a dir, per la relació amb la variable X.

Algunes propietats del coeficient de determinació:

- $0 \leq R^2 \leq 1$
- Si $R^2 = 1$, aleshores tots els residus valen zero i l'ajustament és perfecte; si $R^2 = 0$, l'ajustament és inadequat.
- El coeficient de determinació de la recta de regressió de Y sobre X és el mateix que el de la recta de regressió de X sobre Y, i es verifica que:

  $$R^2 = b \cdot b',$$

  és a dir, el coeficient de determinació és una mesura del grau de relació lineal entre les variables.

  Demostració: per definició, $y^*_i = a + b\,x_i$. Aplicant les propietats de la variància:

  $$s^2_{y^*} = b^2 s^2_x$$

  d'on:

  $$R^2 = \frac{s^2_{y^*}}{s^2_y} = \frac{b^2 s^2_x}{s^2_y} = b\frac{\frac{s_{xy}}{s^2_x} s^2_x}{s^2_y} = b\frac{s_{xy}}{s^2_y} = b \cdot b'.$$

- El coeficient de determinació és el quadrat del coeficient de correlació lineal, és a dir:

  $$R^2 = r^2_{xy}.$$

  Demostració:

  $$R^2 = b \cdot b' = \frac{s_{xy}}{s^2_x} \cdot \frac{s_{xy}}{s^2_y} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = r^2_{xy}.$$

4.2.3. PREDICCIÓ

L'objectiu últim de la regressió és la predicció d'una variable per a un valor determinat de l'altra.

La **predicció** de Y per a $X = x_0$ és, simplement, el valor obtingut en la recta de regressió de Y sobre X en substituir el valor de x per x_0, és a dir: $y_0 = a + bx_0$. Evidentment, la fiabilitat d'aquesta predicció augmentarà quan la correlació entre les variables ho faça (és a dir, quan R^2 augmente).

Exemple 4.2 En l'exemple anterior, determina la despesa per a enguany si la renda és de 45.3 milions d'euros. Dóna una mesura de la bondat de la predicció. Quin és el percentatge de variabilitat en la despesa atribuïble a la renda dels consumidors?

Solució Ja hem calculat la recta de regressió de la despesa en funció de la renda:

$$y = 0.1715 + 0.5748\, x.$$

Com que la renda es mesura en milions d'euros, la predicció de la despesa serà:

$$y(45.3) = 0.1715 + 0.5748 \times 45.3 = 26.21 \text{ milions d'euros.}$$

Una mesura de la bondat de la predicció, ens la proporciona el coeficient de correlació lineal entre ambdues variables:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{6.948}{\sqrt{12.0884} \times \sqrt{4.320}} = 0.9952\,.$$

Així doncs, com que està pròxim a la unitat, la predicció és molt fiable.

El percentatge de variabilitat en la despesa atribuïble a la renda dels consumidors, ens el dóna el coeficient de determinació:

$$R^2 = r_{xy}^2 = 0.9904.$$

La bondat de la predicció, també sol expressar-se en termes de percentatge, en aquest cas seria un valor prop del 99 %.

4.3. REGRESSIÓ LINEAL MÚLTIPLE

En aquesta secció considerarem que estem estudiant $p + 1$ variables i que el nostre objectiu és obtindre una funció que modelitze la relació de dependència d'una d'aquestes variables (Y), que anomenarem **variable dependent o explicada** com a funció de les p restants $(X_1, \ldots, X_p)$, que anomenarem **variables independents o explicatives**. Els valors de la mostra, ara estaran ordenats de la manera següent:

Y	y_1	y_2	$\cdots$	y_n
X_1	x_{11}	x_{12}	$\cdots$	x_{1n}
X_2	x_{21}	x_{22}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
X_p	x_{p1}	x_{p2}	$\cdots$	x_{pn}

Només estudiarem el cas en què la funció de regressió considerada siga una funció de tipus lineal, és a dir:

$$y = b_0 + b_1\, x_1 + \cdots + b_p\, x_p,$$

on $b_0, b_1, \ldots, b_p$ són els paràmetres. Aquesta equació és l'equació d'un hiperplà en l'espai $\mathbb{R}^p$, anomenat **hiperplà de regressió de Y sobre $\mathbf{X}_1, \ldots, \mathbf{X}_p$**.

Per a la deducció de l'hiperplà de regressió usarem àlgebra matricial, denotarem per:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

el vector de valors observats de la variable Y. Per:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

el vector dels paràmetres. Per:

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \vdots & \vdots & & v \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}$$

la matriu en la qual tots els elements de la primera columna són iguals a 1 i la resta de columnes contenen els valors observats de les variables explicatives $X_1, \ldots, X_p$. Finalment, per:

$$y^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix}$$

el vector de valors teòrics, on:

$$y_i^* = b_0 + b_1\, x_{1i} + \cdots + b_p\, x_{pi}.$$

Matricialment aquest vector pot escriure's com:

$$y^* = X \cdot b.$$

Per a trobar el valor de b apliquem el mètode dels mínims quadrats. En aquest cas, haurem de minimitzar la funció:

$$\Psi = (y - y^*) \cdot (y - y^*)^t = (y - X \cdot b) \cdot (y - X \cdot b)^t,$$

on usem el superíndex t per a representar la matriu transposada.

Derivant en l'expressió anterior respecte a b i igualant a 0, obtenim el vector dels paràmetres:

$$\boxed{b = (X^t \cdot X)^{-1} \cdot X^t \cdot y}$$

Exemple 4.3 Les dades següents representen la formació fixa del capital brut (Y), el PIB a preus de mercat (X_1) i la formación fixa del capital brut d'un any anterior (X_2), per al període 81-85. Ajusteu un model lineal que explique la formació fixa del capital brut d'un any en funció del PIB i de la formació fixa del capital brut de l'any anterior.

Y	X_1	X_2
3.26	15.2	3.37
3.27	15.4	3.26
3.19	15.6	3.27
3.03	15.9	3.19
3.15	16.3	3.03

Solució Utilitzant les dades de la taula construïm les matrius següents:

$$y = \begin{bmatrix} 3.26 \\ 3.27 \\ 3.19 \\ 3.03 \\ 3.15 \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & 15.2 & 3.37 \\ 1 & 15.4 & 3.26 \\ 1 & 15.6 & 3.27 \\ 1 & 15.9 & 3.19 \\ 1 & 16.3 & 3.03 \end{bmatrix},$$

a partir de les quals calculem:

$$X^tX = \begin{bmatrix} 5 & 78.4 & 16.12 \\ 78.4 & 1230.06 & 252.55 \\ 16.12 & 252.55 & 52.0344 \end{bmatrix}, \quad (X^tX)^{-1} = \begin{bmatrix} 16354.92 & \text{-612.84} & \text{-2092.26} \\ \text{-612.84} & 23.20 & 77.27 \\ \text{-2092.26} & 77.27 & 273.15 \end{bmatrix},$$

$$X^t y = \begin{bmatrix} 15.9 \\ 249.196 \\ 51.2879 \end{bmatrix}, \quad b = (X^t X)^{-1} X^t y = \begin{bmatrix} 19.242664 \\ \text{-}0.6580132 \\ \text{-}1.7795794 \end{bmatrix}.$$

I així, l'equació del model és:

$$\boxed{y = 19.242664 - 0.6580132\, x_1 - 1.7795794\, x_2}$$

4.3.1. VARIÀNCIA RESIDUAL. COEFICIENT DE DETERMINACIÓ MÚLTIPLE

Una mesura de la bondat d'ajustament, també en el cas de la regressió lineal múltiple, és la **variància residual** (s_e^2), que en aquest cas serà:

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{1i} - \cdots - b_p x_{ip})^2,$$

la qual podem expressar matricialment com:

$$s_e^2 = \frac{1}{n}\left(y^t y - b^t X^t y\right)$$

i té el mateix significat que en el cas de la regressió lineal simple.

Perquè la mesura de la bondat d'ajustament no depenga de les unitats de mesura, definim, de la mateixa forma que en el cas de la regressió simple, el **coeficient de determinació múltiple**:

$$R^2 = 1 - \frac{s_e^2}{s_y^2},$$

que mesura el grau d'associació lineal simultània entre les p variables.

Exemple 4.4 Considerem les dades de l'exemple anterior, calculant:

$$b^t X^t y = 50.59157235, \quad y^t y = 50.6 \Rightarrow s_e^2 = 0.00168552944, \quad s_y^2 = 0.0076.$$

Per tant,

$$R^2 = 1 - \frac{0.00168552944}{0.0076} = 0.77822.$$

Així, la fiabilitat de l'ajustament serà del 77.82 %.

4.3.2. UN CAS PARTICULAR: EL PLA DE REGRESSIÓ

En cas que hàgem de trobar el pla de regressió de Y sobre X_1 i X_2, podem obtindre els coeficients en funció de les mitjanes aritmètiques, variàncies i covariàncies de les variables que intervenen en el problema:

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2\,,$$

$$b_1 = \frac{s_{x_1y}s^2_{x_2} - s_{x_2y}s_{x_1x_2}}{s^2_{x_1}s^2_{x_2} - s^2_{x_1x_2}},$$

$$b_1 = \frac{s_{x_2y}s^2_{x_1} - s_{x1y}s_{x_1x_2}}{s^2_{x_1}s^2_{x_2} - s^2_{x_1x_2}},$$

on el coeficient de determinació múltiple, en aquest cas és:

$$R^2 = b_1\frac{s_{x_1y}}{s^2_y} + b_2\frac{s_{x_2y}}{s^2_y}\,.$$

4.4. REGRESSIÓ NO LINEAL. COEFICIENT DE CORRELACIÓ GENERAL

4.4.1. MODELS DE REGRESSIÓ NO LINEAL SIMPLE

El model lineal de regressió és el més senzill, però en ocasions el núvol de punts ens pot indicar que no és adequat. Per tant, haurem de recórrer a altres models, és a dir, a buscar altres funcions que ajusten millor les dades que tenim.

- **Model potencial:** busquem una funció de regressió de la forma:

  $$y = k\,x^b,$$

  on k i b són els paràmetres que cal determinar. Aquest model, podem reduir-lo al cas lineal prenent logaritmes:

  $$\ln y = \ln k + b\ln x.$$

  Si ara fem el canvi de variables $z = \ln y$ i $t = \ln x$ i posem $a = \ln k$, només hem de calcular la recta de regressió de Z sobre T:

  $$z = a + b\,t$$

  i després, una vegada calculats els paràmetres a i b d'aquesta recta, obtindre els paràmetres buscats k i b, tenint en compte que:

  $$k = e^a \quad \text{i} \quad b = b.$$

- **Model exponencial:** busquem una funció de regressió de la forma:

  $$y = c\,k^x,$$

on c i k són els paràmetres que cal determinar. Aquest model, també podem reduir-lo al cas lineal prenent logaritmes:

$$\ln y = \ln c + (\ln k)\, x.$$

Si ara fem el canvi de variable $z = \ln y$ i posem $a = \ln c$ i $b = \ln k$, només hem de calcular la recta de regressió de Z sobre X

$$z = a + bx,$$

i després, una vegada calculats els paràmetres a i b d'aquesta recta, obtindrem els paràmetres buscats c i k, tenint en compte que:

$$c = e^a \quad \text{i} \quad k = e^b.$$

- **Model parabòlic:** busquem una funció de regressió de la forma:

$$y = a + b\,x + c\,x^2,$$

on a, b i c són els paràmetres que cal determinar.

En aquest cas, utilitzarem el mètode dels mínims quadrats per a la determinació dels paràmetres. Considerem:

$$e_i = y_i - y_i^* = y_i - a - b\,x_i - c\,x_i^2$$

i minimitzem:

$$\Psi(a,b,c) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a - b\,x_i - c\,x_i^2\right)^2.$$

Derivant parcialment respecte de cadascuna de les variables d'aquesta equació (que són els paràmetres que busquem) i igualant a zero:

$$\left.\begin{aligned} \frac{\partial \Psi(a,b,c)}{\partial a} &= -2\sum_{i=1}^{n}\left(y_i - a - b\,x_i - c\,x_i^2\right) = 0 \\ \frac{\partial \Psi(a,b,c)}{\partial b} &= -2\sum_{i=1}^{n}\left(y_i - a - b\,x_i - c\,x_i^2\right) x_i = 0 \\ \frac{\partial \Psi(a,b,c)}{\partial c} &= -2\sum_{i=1}^{n}\left(y_i - a - b\,x_i - c\,x_i^2\right) x_i^2 = 0 \end{aligned}\right\}$$

i simplificant, obtenim el sistema d'equacions normals:

$$\left.\begin{aligned} a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i \end{aligned}\right\}$$

que, una vegada resolt, ens proporciona la paràbola de regressió de Y sobre X.

4.4.2. MESURES DE LA BONDAT D'AJUSTAMENT

De nou, la bondat d'ajustament de cada un dels models analitzats, la mesurarem mitjançant la **variància residual**:

$$s_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*)^2$$

o mitjançant el coeficient de **determinació general**:

$$R^2 = 1 - \frac{s_e^2}{s_y^2}$$

l'arrel quadrada positiva del qual

$$R = \sqrt{1 - \frac{s_e^2}{s_y^2}}$$

rep el nom de **coeficient de correlació general de Pearson**.

Exemple 4.5 Donada la distribució

x_i	2	5	8	11	14
y_i	4	7	12	21	25

estima un model de regressió parabòlic i calcula el coeficient de correlació general.

Solució Per tal d'ajustar una paràbola a aquestes dades, plantegem el sistema d'equacions normals:

$$\left.\begin{aligned} a\,n + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3 &= \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 &= \sum_{i=1}^{n} x_i^2 y_i \end{aligned}\right\}$$

i per a aquest fi formem la taula següent:

	y_i	x_i	x_i^2	x_i^3	x_i^4	$y_i x_i$	$y_i x_i^2$
	4	2	4	8	16	8	16
	7	5	25	125	625	35	175
	12	8	64	512	4096	96	768
	21	11	121	1331	14641	231	2541
	25	14	196	2744	38416	350	4900
Sumes:	69	40	410	4720	57764	720	8400

d'on obtenim el sistema:

$$\left.\begin{array}{l} 5\,a + 40\,b + 410\,c = 69 \\ 40\,a + 410\,b + 4720\,c = 720 \\ 410\,a + 4720\,b + 57764\,c = 8400 \end{array}\right\}$$

que té per solució:

$$a = 1.057, \quad b = 1.1105, \quad c = 0.048,$$

d'on la paràbola de regressió de Y sobre X és:

$$\boxed{y = 1.057 + 1.105\,x + 0.048\,x^2}$$

Per a calcular el coeficient de correlació general, calcularem prèviament la variància residual. Per a això formem una altra taula:

y_i	x_i	$y_i^* = 1.057 + 1.105\,x_i + 0.048\,x_i^2$	$e_i = y_i - y_i^*$	e_i^2	y_i^2
4	2	3.459	0.541	0.292681	16
7	5	7.782	-0.782	0.611524	49
12	8	12.969	-0.969	0.938961	144
21	11	19.020	1.980	3.920400	441
25	14	25.935	-0.935	0.874225	625
69				6.637791	1275

a partir de la qual calculem:

$$\overline{y} = \frac{69}{5} = 13.8, \quad s_y^2 = \frac{1275}{5} - 13.8^2 = 64.56, \quad s_e^2 = \frac{6.637791}{5} = 1.327,$$

$$R = \sqrt{1 - \frac{1.327}{64.56}} = 0.9897 \;\Rightarrow\; \text{fiabilitat del } 98.97\,\% \ .$$

4.5. PROBLEMES PROPOSATS

(1) El Ministeri de Turisme d'un determinat país ha observat que el nombre de places hoteleres ocupades és diferent segons el preu de l'habitació. En la taula següent es detallen el total de places ocupades en un any amb els preus corresponents:

Preu (en euros)	Nombre d'habitacions ocupades
40	4720
80	2615
120	1870
200	945
350	450

a) Representa gràficament les dades i comprovar que existeix una relació lineal entre el nombre d'habitacions ocupades i el preu per habitació.

b) Troba l'equació de la recta de regressió. Quantes habitacions s'ocuparien a 275 €?

c) En quina mesura podem considerar que el nivell d'ocupació depén de l'estructura dels preus?

(2) Donades les següents dades:

x_i	-2	-1	0	1	2
y_i	4	1	0	1	4

a) Analitza-les gràficament. Raona si escau o no fer-hi un ajustament lineal.

b) Ajusta el model no lineal més adequat tenint en compte la representació gràfica feta en l'apartat anterior.

c) Interpreta la bondat d'ajustament.

(3) Donada la distribució

x_i	y_i
3	1
7	6
13	11
16	24
21	36

a) Ajusta un model exponencial.

b) Calcula el coeficient de correlació general de Pearson.

(4) Un determinat partit polític es planteja el problema de fins a quin punt li poden compensar les despeses en publicitat de la pròxima campanya electoral segons el resultat obtingut. En els últims processos electorals les despeses en publicitat (en milers d'euros) i el nombre de diputats elegits van ser:

Despeses en publicitat	Diputats elegits
12	3
18	4
27	4
33	6
51	8

La comissió electoral del partit està estudiant la possibilitat de realitzar una despesa de 60000 € en publicitat per a la pròxima campanya electoral.

a) Quin serà el nombre de diputats elegits d'aquest partit d'acord amb el pressupost esmentat si la imatge del partit no varia respecte a les eleccions anteriors?

b) Amb quina confiança pot esperar-se aquest resultat?

c) Quin serà el percentatge de causes diferents de la publicitat que influiran en les eleccions?

(5) Coneixent les dades següents: $\overline{x} = 3$, $\overline{y} = 2$, $s_x^2 = 6$, $s_y^2 = 8$ i que la recta de regressió de Y sobre X és $y = 4 - 0.667\,x$, obteniu la recta de regressió de X sobre Y.

SOLUCIONS

(1) *a*) Es representa el núvol de punts

b) $y = 3976.5 - 11.75\,x$, $y(275) \approx 745$ habitacions

c) $r_{xy} = -0.86$. Així tenim un 86 % de fiabilitat

(2) *a*) $r_{xy} = 0$. Per tant, no existeix relació lineal entre les variables

b) $y = x^2$

c) Ajustament perfecte

(3) *a*) $y = 0.93 \times (1.21)^x$

b) $R = 0.83$

(4) *a*) 9 diputats

b) La predicció és molt fiable, ja que $r_{xy} = 0.96$

c) 7.3 %

(5) $x = 4 - 0.5\,y$

TEMA 5
NOMBRES ÍNDEXS

5.1. INTRODUCCIÓ

En moltes ocasions les variables socioeconòmiques com el volum d'importacions, el nombre de vendes d'una empresa, o el valor del PIB, varien amb el temps i pot aparéixer la necessitat de fer comparacions en funció de les dites variables per a diferents temps, tant per separat, com en grups o conjunts de les variables. En aquest tema tractarem el problema de la comparació d'una sèrie d'observacions respecte a una situació inicial fixada arbitràriament. Les mesures estadístiques que descriuen aquests canvis són els **nombres índexs**.

Els exemples de nombres índexs són molt abundants: a més dels més coneguts, com poden ser l'índex de preus de consum (consulteu la pàgina web de l'Institut Nacional d'Estadística: http://www.ine.es/) o els indicadors de la borsa, n'existeixen d'altres, menys coneguts popularment, però que tenen una gran influència en l'economia mundial. En citarem dos, i indicarem algunes pàgines web on se'n pot ampliar la informació. Però és necessari advertir que per a consultes posteriors a la publicació d'aquest text és convenient actualitzar les dates.

- Índex de Confiança de Mercats Emergents

  Poden consultar-se, per exemple, les pàgines web:

  - http://www.iberglobal.com/Newsletter/alerta_geo_abril_2006.htm
  - http://www.iberglobal.com/Newsletter/alerta_geo_abril_2007.htm
  - http://www.iberglobal.com/ (busqueu notícies sobre el dit índex)

- Índex de Confiança dels Consumidors

  Consulteu:

  - http://www.finanzas.com/id.4892123/noticias/noticia.htm
  - http://economy.blogs.ie.edu/archives/2007/01/indice_de_confi.php
  - http://www.cincodias.com/ (buscar notícies sobre dit índex)
  - http://www.agendadeprensa.com/informes/ico_enero08.pdf

A continuació, formalitzarem el concepte de nombre índex:

Definició 5.1 Nombre índex és aquella mesura estadística que ens permet estudiar els canvis que es produeixen en una magnitud simple o complexa respecte al temps.

Anomenarem **període base** o **període de referència** el període inicial, i la situació que volem comparar, l'anomenarem **període actua**l o **període corrent**.

Els nombres índexs poden ser: **simples**, si només comparem una variable, o **complexos**, si comparem un grup de variables. Aquests últims poden ser: **ponderats** o **sense ponderar**.

5.2. ÍNDEXS SIMPLES I COMPLEXOS

5.2.1. ÍNDEXS SIMPLES

Definició 5.2 Siga X una variable i siguen x_0 i x_t els valors de la dita variable mesurats en els períodes base i actual, respectivament. El **nombre índex simple** I per a la magnitud citada es defineix com el quocient entre ambdós valors:

$$I = I_0^t = \frac{x_t}{x_0}\,.$$

És a dir:

$$I = I_{t_{base}}^{t_{actual}} = \frac{x_{t_{actual}}}{x_{t_{base}}}\,.$$

El nombre índex simple I mesura en tant per u la variació que ha experimentat la magnitud X entre els períodes considerats. De vegades es multipliquen per 100 i expressen percentatges.

$$I > 1 \text{ (o } 100) \rightarrow \text{augment}, \qquad I < 1 \text{ (o } 100) \rightarrow \text{disminució}.$$

Exemple 5.1 Donats els preus de dos articles A i B per la taula següent:

Anys	Preus	
	Article A	Article B
1994	10	20
1995	12	25
1996	15	28

els índexs simples amb base en 1994 són:

Anys	Índexs simples	
	Article A	Article B
1994	1	1
1995	1.2	1.25
1996	1.5	1.4

on els valors dels índexs s'han obtingut aplicant la definició. Per exemple, per a l'article A:

$$I_0^{95} = \frac{12}{10} = 1.2, \qquad I_0^{96} = \frac{15}{10} = 1.5.$$

Els índexs simples més usuals són:

- El **preu relatiu**: és la raó entre el preu d'un bé en el període actual (p_t) i el preu del mateix bé en el període base (p_0):

$$p_0^t = \frac{p_t}{p_0}\,.$$

(Si seguim la notació anterior: $X = P =$ preu d'un bé, aleshores $x_t = p_t$ i $x_0 = p_0$).

- La **quantitat relativa**: és el quocient entre la quantitat produïda o venuda d'un bé en el període actual (q_t) i la quantita produïda o venuda del mateix bé en el període base (q_0).

$$q_0^t = \frac{q_t}{q_0}\,.$$

(Si posem: $X = Q =$ quantitat produïda o venuda d'un bé, aleshores $x_t = q_t$ i $x_0 = q_0$).

- El **valor relatiu**: anomenarem **valor d'un bé** en un període arbitrari el producte del preu d'aquest bé per la quantitat produïda (o venuda). Per tant, el valor relatiu (V_0^t) és:

$$V_0^t = \frac{p_t\, q_t}{p_0\, q_0}\,.$$

Així, el valor relatiu d'un bé és igual al producte del seu preu relatiu per la seua quantitat relativa, és a dir:

$$V_0^t = p_0^t\, q_0^t\,,$$

ja que:

$$V_0^t = \frac{p_t\, q_t}{p_0\, q_0} = \frac{p_t}{p_0} \cdot \frac{q_t}{q_0} = p_0^t\, q_0^t\,.$$

5.2.2. ÍNDEXS SIMPLES EN CADENA

Els índexs en cadena són un conjunt d'índexs per als quals la base és sempre el període precedent. D'aquesta manera, cadascun representa una comparació percentual respecte al període anterior.

Exemple 5.2 Per a les mateixes dades de l'exemple anterior es té que:

Anys	Índexs en cadena	
	Article A	Article B
1994		
1995	1.2	1.25
1996	1.25	1.12

on, per exemple, per a l'article A:

$$I_{96}^{95} = \frac{15}{12} = 1.25.$$

5.2.3. ÍNDEXS COMPLEXOS: NO PONDERATS I PONDERATS

En la realitat succeeix que, generalment, no estem interessats a comparar preus, quantitats o valors de béns individuals, sinó que les dites magnituds es comparen per a grups. És a dir, habitualment serà necessari estudiar les variacions d'un conjunt de N variables. Per exemple, analitzar l'evolució de preus dels quatre cereals bàsics a Espanya. Per a aquest fi, la informació subministrada pels índexs simples de cadascuna de les variables s'ha de resumir en un únic índex que anomenarem **índex complex**.

Existeixen dos tipus d'índexs complexos: ponderats i no ponderats.

NOMBRES ÍNDEXS COMPLEXOS NO PONDERATS

Per a resumir la informació obtinguda a través dels índexs simples, el més lògic és calcula d'alguna forma la mitjana d'aquests. Segons el tipus de mitjana que s'utilitze, apareixen els diferents nombres índexs complexos.

Considerem les variables $X_1, X_2, \ldots, X_N$ que fan prendre els valors:

Variable	Període base	Període actual	Índexs simples
X_1	x_{10}	x_{1t}	$I_1 = \frac{x_{1t}}{x_{10}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_i	x_{i0}	x_{it}	$I_i = \frac{x_{it}}{x_{i0}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_N	x_{N0}	x_{Nt}	$I_N = \frac{x_{Nt}}{x_{N0}}$

A partir de la taula i prenent els índexs simples, podem definir els següents índexs complexos no ponderats. Per ordre d'importància, considerarem:

- **Índex mitjana aritmètica d'índexs simples**: consisteix a calcular la mitjana aritmètica simple dels índexs de totes les variables:

$$\overline{I} = \frac{1}{N} \sum_{i=1}^{N} I_i \, .$$

- **Índex mitjana agregativa**: quan les mitjanes en què estan expressades les variables siguen homogènies, es poden comparar les mitjanes dels valors de les variables en cada període (base i actual):

$$I_A = \frac{\sum_{i=1}^{N} x_{it}}{\sum_{i=1}^{N} x_{i0}} .$$

- **Índex mitjana geomètrica d'índexs simples**:

$$I_G = \sqrt[N]{\prod_{i=1}^{N} I_i} \quad .$$

- **Índex mitjana harmònica d'índexs simples**:

$$I_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{I_i}} .$$

Exemple 5.3 Donada la producció de tres tipus de cítrics expressades en milions de quilograms, calculeu-ne els índexs complexos, mitjana aritmètica i mitjana agregativa amb base en 1994.

Anys	Taronges	Mandarines	Pomelos
1994	450	200	120
1995	400	180	98
1996	425	220	150

Solució:

Índexs simples			$\overline{I}$	$\sum x_{it}$	I_A
Taronges	Mandarines	Pomelos	I. C.	Total cítrics	I. C.
100	100	100	100	770	100
88.89	90	81.67	86.85	678	88.05
94.44	110	125	109.81	795	103.25

Els índexs complexos no ponderats més usuals són:

- **Índex de preus de Sauerbeck**: és l'índex mitjana aritmètica dels preus relatius:

$$S_p = \frac{1}{N}\sum_{i=1}^{N} I_i\,,$$

on $I_i = \dfrac{p_{it}}{p_{i0}}$.

- **Índex de preus Bradstreet-Dûtot**: és l'índex mitjana agregativa dels preus relatius:

$$BD_p = \frac{\sum_{i=1}^{N} p_{it}}{\sum_{i=1}^{N} p_{i0}}\,.$$

- **Índex de quantitats de Sauerbeck**: és l'índex mitjana aritmètica de les quantitats relatives:

$$S_q = \frac{1}{N}\sum_{i=1}^{N} I_i\,.$$

on $I_i = \dfrac{q_{it}}{q_{i0}}$.

- **Índex de quantitats Bradstreet-Dûtot**: és l'índex mitjana agregativa de les quantitats relatives:

$$BD_q = \frac{\sum_{i=1}^{N} q_{it}}{\sum_{i=1}^{N} q_{i0}}\,.$$

NOMBRES ÍNDEXS COMPLEXOS PONDERATS

En els índexs anteriors no hem tingut en compte la diferent importància relativa que pot tindre cadascuna de les variables simples dins del conjunt. És necessari de vegades que, als índexs simples, se'ls assignen pesos o ponderacions (ω_i) que en consideren la importància relativa.

D'aquesta forma obtindríem els següents **índexs complexos ponderats**:

- **Índex mitjana aritmètica ponderat**:

$$\overline{I}^{*} = \frac{\sum_{i=1}^{N} \omega_i I_i}{\sum_{i=1}^{N} \omega_i}\,.$$

- **Índex mitjana agregativa ponderat**:

$$I_A^* = \frac{\sum_{i=1}^{N} x_{it}\omega_i}{\sum_{i=1}^{N} x_{i0}\omega_i}.$$

- **Índex mitjana geomètrica ponderat**:

$$I_G^* = \sqrt[\Sigma\,\omega_i]{\prod_{i=1}^{N} I_i^{\omega_i}}.$$

- **Índex mitjana harmònica ponderat**:

$$I_H^* = \frac{\sum_{i=1}^{N} \omega_i}{\sum_{i=1}^{N} \frac{\omega_i}{I_i}}.$$

Els índexs complexos ponderats més usuals són:

- **Índex de preus de Laspeyres**: és la mitjana aritmètica ponderada dels índexs simples de preus usant els pesos $\omega_i = p_{i0} \cdot q_{i0}$ (valor de la quantitat consumida del bé i-èsim en el període base, a preus del dit període):

$$L_p = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}} = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}.$$

- **Índex de preus de Paasche**: és la mitjana aritmètica ponderada dels índexs simples de preus usant els pesos $\omega_i = p_{i0} \cdot q_{it}$ (valor de la quantitat consumida del bé i-èsim en el període actual, a preus del període base):

$$P_p = \frac{\sum_{i=1}^{N} \frac{p_{it}}{p_{i0}} \cdot p_{i0} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}} = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}}.$$

  L'índex de Paasche requereix calcular les ponderacions per a cada període corrent, per la qual cosa la seua elaboració és més costosa; per això s'utilitza menys que el de Laspeyres.

- **Índex de preus d'Edgeworth**: és la mitjana agregativa ponderada dels preus, amb pesos $\omega_i = q_{i0} + q_{it}$ (quantitat total consumida del bé i-èsim en el període base i en l'actual):

$$E_p = \frac{\sum_{i=1}^{N} p_{it}\,(q_{i0} + q_{it})}{\sum_{i=1}^{N} p_{i0}\,(q_{i0} + q_{it})}\,.$$

- **Índex de preus de Fisher**: és la mitjana geomètrica dels índexs de preus de Laspeyres i Paasche, és a dir:

$$F_p = \sqrt{L_p \cdot P_p}\,.$$

- **Índex de quantitats de Laspeyres**: és la mitjana aritmètica ponderada dels índexs simples de quantitats usant els pesos $\omega_i = p_{i0} \cdot q_{i0}$, i té l'expressió següent:

$$L_q = \frac{\sum_{i=1}^{N} p_{i0} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{i0}}\,.$$

- **Índex de quantitats de Paasche**: és la mitjana aritmètica ponderada dels índexs simples de quantitats usant els pesos $\omega_i = p_{it} \cdot q_{i0}$ (valor de la quantitat consumida del bé i-èsim en el període base, a preus actuals):

$$P_q = \frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{it} \cdot q_{i0}}\,.$$

- **Índex de quantitats d'Edgeworth**: és la mitjana agregativa ponderada amb pesos $\omega_i = p_{i0} + p_{it}$, i té l'expressió següent:

$$E_q = \frac{\sum_{i=1}^{N} q_{it}\,(p_{i0} + p_{it})}{\sum_{i=1}^{N} q_{i0}\,(p_{i0} + p_{it})}\,.$$

- **Índex de quantitats de Fisher**: és la mitjana geomètrica dels índexs de quantitat de Laspeyres i Paasche, és a dir:

$$F_q = \sqrt{L_q \cdot P_q}\,.$$

Exemple 5.4 Donats els preus i les quantitats de tres articles de consum des de 1990 fins a 1994, calculeu els índexs complexos ponderats de preus de Laspeyres i Paasche prenent com a base 1990.

Solució:

Anys	Article A		Article B		Article C		Índexs	
	P.	C.	P.	C.	P.	C.	L_p	P_p
1990	2	8	3	5	1	3	100	100
1991	3	7	4	6	2	3	147.06	145.71
1992	2	10	5	6	2	5	138.23	139.53
1993	4	12	7	7	4	8	232.35	243.40
1994	5	11	8	8	5	10	279.41	302.78

on, per exemple:

$$L_{90}^{91} = \frac{3\cdot 8 + 4\cdot 5 + 2\cdot 3}{2\cdot 8 + 3\cdot 5 + 1\cdot 3} = \frac{50}{34} = 1.1405, \quad L_{90}^{92} = \frac{2\cdot 8 + 5\cdot 5 + 2\cdot 3}{2\cdot 8 + 3\cdot 5 + 1\cdot 3} = \frac{47}{34} = 1.3823,$$

$$P_{90}^{91} = \frac{3\cdot 7 + 4\cdot 6 + 2\cdot 3}{2\cdot 7 + 3\cdot 6 + 1\cdot 3} = \frac{51}{35} = 1.4571, \quad P_{90}^{92} = \frac{2\cdot 10 + 5\cdot 6 + 2\cdot 5}{2\cdot 10 + 3\cdot 6 + 1\cdot 5} = \frac{60}{43} = 1.3953\,.$$

5.3. PROPIETATS DELS NOMBRES ÍNDEXS

Existeixen una sèrie de propietats que han de ser verificades pels nombres índexs; aquestes són:

- **Existència**: tot nombre índex ha d'existir i ha de tindre un valor finit distint de zero, és a dir:
  - $-\infty < I < +\infty$.
  - $I \neq 0$.

  Aquesta propietat, no la compleixen els índexs complexos mitjana geomètrica i mitjana harmònica.

- **Identitat**: si es fan coincidir el període base i el període actual, el nombre índex ha de ser igual a la unitat, és a dir, $I_0^0 = 1$. Tots els índexs més usuals ho verifiquen.

- **Inversió**: si denotem per I_0^t un índex amb base 0 i període actual t, en intercanviar els períodes entre si (I_t^0) el nou índex és: $I_t^0 = \dfrac{1}{I_0^t}$. Això implica que $I_t^0 \cdot I_0^t = 1$. Aquesta propietat, la compleixen els índexs simples i entre els complexos, els de Bradstreet-Dûtot, Edgeworth i Fisher.

- **Proporcionalitat**: si en el període actual totes les magnituds experimenten una variació proporcional, el nombre índex ha de quedar afectat per aquesta variació.

  Aquesta propietat, la compleixen tots els índexs simples: suposem que els valors x_t experimenten una variació proporcional d'ordre k, i en el període t' són $x'_t = k\,x_t$, aleshores el nou índex simple serà:

$$I' = \frac{x'_t}{x_0} = \frac{k\,x_t}{x_0} = k\,I.$$

  Els índexs complexos més usuals compleixen aquesta propietat.

  Exemple 5.5 Vejam que l'índex de preus de Paasche compleix aquesta propietat. Siguen $p_{it'} = k\,p_{it}$, aleshores:

$$P'_p = \frac{\sum_{i=1}^{N} p_{it'} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}} = \frac{\sum_{i=1}^{N} k\,p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}} = k\,\frac{\sum_{i=1}^{N} p_{it} \cdot q_{it}}{\sum_{i=1}^{N} p_{i0} \cdot q_{it}} = k\,P_p.$$

  Però en el cas dels índexs de Paasche, Edgeworth i Fisher es pot plantejar una objecció de tipus econòmic: en variar els preus en qualsevol proporció és difícil mantindre el supòsit que les quantitats romanguen constants, o en el cas variar les quantitats, que els preus romanguen constants.

  Així doncs, entre els índexs complexos ponderats, l'índex de Laspeyres és l'únic que compleix adequadament la propietat de proporcionalitat i és, per tant, el que més s'utilitza.

- **Homogeneïtat**: un nombre índex no ha de veure's afectat per les unitats de mesura.

- **Circular**: si considerem els períodes 0, t, t', t'', s'ha de complir que:

$$I_0^t I_t^{t'} = I_0^{t'}, \qquad I_{t'}^t I_t^{t''} = I_{t'}^{t''} \quad \text{i} \quad I_0^t I_t^{t'} I_{t'}^{t''} = I_0^{t''}.$$

  Açò, ho compleixen tots els índexs simples i el de Bradstreet-Dûtot.

5.4. ALGUNS PROBLEMES EN LA CONSTRUCCIÓ I LA UTILITZACIÓ DELS NOMBRES ÍNDEXS

- **Variables**: en índexs simples les variables seran les donades; però si s'elaboren índexs complexos, de primer cal plantejar-se quines seran les variables que se seleccionaran. El més habitual és prendre una subpoblació integrada pels productes que es consideren més rellevants. En l'exemple que hem vist dels cítrics: taronges, mandarines i pomelos.

- **Període de temps pres com a base**: se'n sol elegir un de no allunyat excessivament del període corrent, o l'índex perdrà representativitat, es quedarà obsolet. Per això cal renovar periòdicament la informació relativa a l'any base.

- **Renovació de l'índex: canvi de base i enllaç**:

  - En els índexs simples el canvi del període pres com a base es fa per la propietat circular, igualant a 1 (o 100) el valor, preu o quantitat del nou any base. Si representem per h el nou període base:

$$I_h^t = \frac{1}{I_0^h} \cdot I_0^t, \quad \forall t.$$

  - El problema que es planteja té més dificultat en els índexs complexos, sobretot en els ponderats, ja que s'han de canviar les ponderacions, encara més quan per a ponderar magnituds actuals s'utilitzen pesos relatius referits al període base. En aquest cas la nova sèrie ha de ser recalculada. D'altra banda, per a poder relacionar sèries d'índexs referits a diferents períodes base, cal enllaçar ambdues sèries, l'antiga i la nova. L'operació d'enllaç és molt senzilla matemàticament; de nou, n'hi ha prou d'assignar al nou any base l'índex 1 (o 100, en percentatges), i aplicar la propietat circular a la sèrie d'índexs antiga. Si representem per h el nou període base:

$$I_h^t = \frac{1}{I_0^h} \cdot I_0^t, \quad \forall t < h.$$

    D'aquesta forma ambdues sèries s'enllacen numèricament, encara que sempre cal tindre en compte, en fer les comparacions, les ponderacions que veritablement es van utilitzar en la construcció de l'índex.

Exemple 5.6 Suposem que per a un conjunt de béns tenim les dades següents:

Anys	$\sum_i p_{it} q_{i0}$	$\sum_i p_{it} q'_{i0}$
1980	5	
1981	5.5	
1982	6	
1983	6.5	8
1984		9
1985		10
1986		10.5

Calcula els índexs de preus de Laspeyres corresponents sobre la base dels anys 1980 i 1983. Calcula també els índexs de preus dels períodes 80, 81 i 82 sobre la base de l'any 1983.

Solució:

Els índexs de preus de Laspeyres són:

$$L_{80}^{80} = \frac{5}{5} = 100\,\%, \quad L_{80}^{81} = \frac{5.5}{5} = 110\,\%, \quad L_{80}^{82} = \frac{6}{5} = 120\,\%, \quad L_{80}^{83} = \frac{6.5}{5} = 130\,\%,$$

$$L_{82}^{83} = \frac{8}{8} = 100\,\%, \quad L_{83}^{84} = \frac{9}{8} = 112.5\,\%, \quad L_{83}^{85} = \frac{10}{8} = 125\,\%, \quad L_{83}^{86} = \frac{10.5}{8} = 131.25\,\%,$$

i els índexs de preus dels períodes 80, 81 i 82 sobre la base de l'any 1983:

$$L_{83}^{80} = \frac{L_{80}^{80}}{L_{80}^{83}} = \frac{100\,\%}{130\,\%} = 76.9\,\%,$$

$$L_{83}^{81} = L_{80}^{81} \cdot L_{83}^{80} = 110\,\% \cdot 76.9\,\% = 84.6\,\%,$$

$$L_{83}^{82} = L_{80}^{82} \cdot L_{83}^{80} = 120\,\% \cdot 76.9\,\% = 92.3\,\%.$$

5.5. DEFLACIÓ

Una de les aplicacions més importants dels nombres índexs és la possibilitat de provocar deflació en les sèries (de preus, de valors, de rendes, de sous, etc.). És ben conegut de tots que el poder adquisitiu dels diners varia amb el temps. El fenomen es coneix com a **inflació**.

Anomenarem **preus constants** els preus que regeixen un determinat període fix, i **preus corrents** els preus que regeixen al llarg de diversos períodes.

Si es té una variable en moneda corrent de cada any (euros, dòlars, etc.) difícilment se'n pot analitzar el creixement o el decreixement real. El mateix ocorreria si es desitja establir comparacions amb altres variables expressades en unitats monetàries distintes. Açò és degut que l'activitat econòmica té un fort component monetari, per la qual cosa les variacions que reflecteixen les sèries, a més de tindre increments o decrements reals, estan influïdes per efectes monetaris molt importants que cal eliminar si es pretén estudiar l'evolució en termes reals d'una economia.

L'operació de convertir les sèries monetàries en valors reals (constants) s'anomena **deflació**. Per a expressar una sèrie donada en diners corrents, en diners constants d'un any T, cal dividir la sèrie primitiva entre els índexs de preus adequats (eliminem la influència dels preus), prenent com a base l'any T, és a dir: $\dfrac{x_t}{I_T^t}$.

Nota 5.1 L'índex ha d'estar expressat en tant per u, i si ho necessitem, podem utilitzar la relació $I_T^t = \dfrac{1}{I_t^T}$.

Si volem obtindre una fórmula de fàcil aplicació, en aquests casos, podem definir les variables següents:

x_t = quantitat amb valor en diners de l'any t,
x_T = quantitat x_t amb valor en diners corrents de l'any T.

Així obtenim:

$$x_T = \frac{x_t}{I_T^t} \qquad \text{i} \qquad x_T = x_t \cdot I_t^T.$$

L'índex utilitzat per a aquesta operació es denomina **deflactor**. Podem usar com a deflactors els índexs de preus més usuals, el de Laspeyres i el de Paasche. En general, el que se sol utilitzar és l'índex de preus de consum.

Exemple 5.7 La mitjana dels sous que una empresa ha pagat mensualment als empleats, durant els anys que s'indiquen en la taula, ha sigut:

Any	Sou mitjà	Sou mitjà en PTA constant de 1982
1982	98735	98735
1983	113940	101641.39
1984	131373	105266.83
1985	147663	108735.64
1986	162282	109872.71
1987	178834	114931.88

Sabent que l'Índex de Preus de Consum corresponent a aquest període està donat per la taula següent:

Any	IPC (base 1972)
1982	463.3
1983	519.4
1984	578.1
1985	629.0
1986	684.4
1987	720.7

determina l'evolució dels sous a preus constants de 1982.

Solució:

Any	IPC (base 1972)$\rightarrow I_0^t$	IPC (base 1982)$\rightarrow I_{h=T}^t$
1982	463.3	463.3:463.3=1.00
1983	519.4	519.4:463.3=1.121
1984	578.1	578.1:463.3=1.248
1985	629.0	629.0:463.3=1.358
1986	684.4	684.4:463.3=1.477
1987	720.7	720.7:463.3=1.556

i usant aquesta taula per a provocar deflació en els sous resulta que:

Any	Sou (en PTA constant de 1982)
1982	98735:1.00=98735
1983	113940:1.121=101641
1984	131373:1.248=105267
1985	147663:1.358=108736
1986	162282:1.477=109873
1987	178834:1.556=114932

5.6. ÍNDEX DE PREUS DE CONSUM I ALTRES ÍNDEXS ELABORATS A ESPANYA

5.6.1. ÍNDEX DE PREUS DE CONSUM

L'**Índex de Preus de Consum**, IPC, és un dels índexs de més importància en l'actualitat. S'hi pretén analitzar l'evolució en el temps de la despesa en consum privat a preus constants (d'un any pres com a base) per a un determinat estrat de població. A Espanya l'elabora l'Institut Nacional d'Estadística (INE). Al nostre país va començar a publicar-se en 1939 sobre la base de 1936 i ha experimentat diverses renovacions sobre les bases de 1958, 1964, 1976, 1983, 1992, 2001 i 2006.

Per a elaborar-lo se selecciona una sèrie de béns. Aquests béns, una selecció de 491 articles, formen el que s'anomena **cistell de consum**. Els componentes del cistell de consum es determinen a través de l'**Enquesta de Pressupostos Familiars** i són el conjunt de béns i serveis que les famílies adquireixen normalment; canvien amb el temps en funció dels usos de consum.

Es publica cada mes i es pren com a base la mitjana aritmètica simple dels índexs mensuals de l'any 2006.

Una vegada determinat el cistell de consum, es valoren les quantitats corresponents consumides a preus del període base i de l'actual. L'índex de preus utilitzat

en la majoria dels països, i en particular a Espanya, és l'índex de Laspeyres encadenat amb actualització de ponderacions anuals. Aquesta actualització anual té els avantatges següents:

- L'IPC s'adapta als canvis del mercat i dels hàbits de consum en un termini molt breu de temps.
- En l'IPC es poden incloure nous béns o serveis quan apareixen en el mercat, així com eliminar els que es consideren poc significatius.

Es calculen dotze índexs independents, per a dotze grups de béns i serveis de consum en què s'estructura el cistell de consum: aliments i begudes no alcohòliques, begudes alcohòliques i tabac, vestit i calçat, habitatge, parament, medicina, transport, comunicacions, cultura i oci, ensenyament, hotels, cafés i resturants, altres béns i serveis.

A l'índex poden desagregar-se tantes variables com es vulga. L'INE elabora aquest índex a escala general o global, per a comunitats autònomes, per a capitals de província, per a nuclis urbans i per a àrees rurals.

Com que és obvi que en l'estudi no poden incloure's totes les famílies, s'hi pren un conjunt de la forma més àmplia possible i representativa, anomenat **estrat de refèrencia**.

5.6.2. ALTRES ÍNDEXS

- **Índex de preus de consum harmonitzat** (IPCA): és un indicador estadístic que proporciona una mesura comuna de la inflació entre els països de la Unió Europea. En la pàgina web de l'INE en podem trobar la metodologia i els resultats detallats.
- **Índexs implícits de preus**: mesuren l'evolució dels preus i es deriven de la Comptabilitat Nacional (valors del producte nacional, despeses de consum i inversió, estalvi, etc.). Aquests valors contenen, implícitament, les variacions en els preus de les magnituds macroeconòmiques. Els índexs que s'hi calculen són índexs de preus de Paasche. S'usen també per a la deflació de sèries de valors.
- **Índexs de producció industrial**: hi ha dues sèries d'índexs de producció industrial de periodicitat mensual: l'una recull les variacions de l'oferta industrial dins de la majoria de les branques de l'activitat industrial i l'altra especifica les variacions en la producció de béns d'equipament.
- **Índexs de preus industrials**: mesuren l'evolució dels preus dels béns d'equipament. S'usen per a provocar deflació en les sèries de valors industrials.
- **Índexs de preus agrícoles**: s'elaboren dos índexs, l'**índex de preus pagats** (pels béns i serveis que es necessiten) i l'**índex de preus percebuts**. La sèrie formada pel quocient d'aquests índexs s'anomena **relació de paritat** i mostra les variacions del poder adquisitiu del sector agrícola.

- **Índexs de l'activitat comercial**: s'elaboren índexs que reflecteixen l'evolució del comerç interior del país, com ara els **índexs de vendes de preus al detall** i **de l'engròs**. El comportament del comerç exterior s'estudia amb els **índexs de preus** i **de quantitats d'exportacions** i **importacions**. El quocient entre l'índex de preus d'importacions i exportacions rep el nom de **relació real d'intercanvi** i permet conéixer l'evolució del poder de compra d'un país davant de l'estranger.

- **Índexs d'activitat financera**: s'elabora una gran quantitat d'índexs: **índexs de cotitzacions de borsa**, **índexs de fons d'inversió**, etc. Generalment s'utilitza l'índex mitjana aritmètica ponderat, on les ponderacions són el volum de contractació negociat de cada títol en l'any base.

5.7. PROBLEMES PROPOSATS

(1) Donada l'estadística sobre la contractació efectiva de les borses espanyoles, en milions de pessetes:

Anys	Madrid	Barcelona	Bilbao	València
1972	67993	28878	12179	2817
1973	100049	43360	19782	3865
1974	113385	40685	21198	6892
1975	102500	31116	23582	6837
1976	131180	35426	14350	4775
1977	74279	17253	16724	7839

a) Calcula els índexs simples sobre la base de 1972.

b) Calcula els índexs de Sauerbeck i de Bradstreet-Dûdot.

(2) Les quantitats emprades en jocs d'atzar a Espanya, en milions de pessetes, durant el període 1982-1987, han sigut:

Any	1982	1983	1984	1985	1986	1987
Quantitat	1700470	1829785	2011267	2147043	2238579	2518765

Expressa aquesta sèrie en pessetes constants de 1982 tenint en compte que l'IPC sobre la base de 1980 ha sigut:

Any	1982	1983	1984	1985	1986	1987
IPC (base 1980)	131.1	147.0	163.6	178.0	193.7	204.0

(3) Si tenim la informació següent sobre un conjunt de nombres índexs simples $I_{83}^{81} = 0.95$, $I_{85}^{83} = 0.8$, quant valdran I_{81}^{83} i I_{81}^{85}?

(4) Les relacions entre dos països, Libertònia i Eslàvia, queden reflectides en les taules següents:

Libertònia va exportar a Eslàvia

	2000		2006	
Producte	Preu	Quantitat	Preu	Quantitat
E_1	20	800	32	1400
E_2	7	1500	11	600
E_3	12	200	14	500

Libertònia va importar d'Eslàvia

	2000		2006	
Producte	Preu	Quantitat	Preu	Quantitat
C_1	4	200	5	410
C_2	10	100	9	300
C_3	11	50	15	100
C_4	8	320	10	150

Calcula:

a) Els índexs de preus de Laspeyres i Paasche per a l'exportació i per a la importació sobre la base de l'any 2000.

b) Els corresponents índexs de quantitats.

c) La raó real d'intercanvi.

(5) El propietari d'un apartament té pactat, en 2002, un lloguer amb el seu inquilí de 300 € mensuals. Es vol revisar el lloguer sobre la base de l'IPC grup habitatge. Quant caldrà que pague en els anys 2003, 2004 i 2005?

Anys	2002	2003	2004	2005
IPC grup habitatge (base 2001)	102.257	105.215	108.895	114.689

(*Font: INE, Espanya*)

SOLUCIONS

(1) *a*)

Anys	Madrid	Barcelona	Bilbao	València
1972	1	1	1	1
1973	1.47	1.50	1.62	1.37
1974	1.67	1.41	1.74	2.45
1975	1.51	1.08	1.94	2.43
1976	1.93	1.23	1.18	1.70
1977	1.09	0.60	1.37	2.78

b) $\mathrm{S}_{p\,72}^{\,73} = 1.49$, $\mathrm{S}_{p\,72}^{\,74} = 1.82$, $\mathrm{S}_{p\,74}^{\,77} = 0.75$, ...
$\mathrm{BD}_{73} = 1.493$, $\mathrm{BD}_{74} = 1.628$ ....

(2)

Anys	Quantitat en milions de pessetes de l'any 1982
1982	1700470
1983	1631843
1984	1611721
1985	1581376
1986	1515113
1977	1618640

(3) $I_{81}^{83} = 1.0526$, $I_{81}^{85} = 1.315$

(4) *a*) Per a l'exportació: $L_p = 155.4\,\%$, $P_p = 152.9\,\%$
Per a la importació: $L_p = 119.1\,\%$, $P_p = 111.7\,\%$

b) Per a l'exportació: $L_q = 132.2\,\%$, $P_q = 130.1\,\%$
Per a la importació: $L_q = 141.3\,\%$, $P_q = 132.5\,\%$

c) $R_{2000}^{2006} = 1.37$

(5) Caldrà que pague 308.68 € en 2003, 319.45 € en 2004, i 336.44 € en 2005.

TEMA 6

SÈRIES TEMPORALS

6.1. INTRODUCCIÓ

Una **sèrie temporal** consisteix, típicament, en un conjunt d'observacions d'una variable Y, preses al llarg del temps en intervals regulars (cada dia, cada mes, etc.), i és, per tant, un conjunt de dades de la forma:

$$\{y_t : t = 1, 2, \cdots, n\}$$

en el qual el subíndex t indica el temps en què la dada y_t va ser observada.

El seu estudi permet analitzar l'evolució que en el transcurs del temps ha experimentat la variable, tant per a descriure'n les propietats com per a caracteritzarne els trets principals i poder predir-ne els valors futurs. Aquesta descripció pot consistir en mesures descriptives i representacions gràfiques.

Normalment, en problemes d'**estadística bàsica**, les observacions són mútuament independents, però en estudiar variables mesurades en el temps, les observacions són clarament **no independents**. Cadascuna tendeix a un valor que està més prop al de les observacions més pròximes que al de les més allunyades. Aquest tipus de comportament s'anomena **correlació serial**.

Exemple 6.1 Xifres oficials de població espanyola des de 1997 fins a 2007, segons la revisió anual del padró municipal de l'1 de gener de cada any.

Any	Població	Any	Població
1997	39669394	2003	42717064
1998	39852651	2004	43197684
1999	40202160	2005	44108530
2000	40499791	2006	44708964
2001	41116842	2007	45200737
2002	41837894		

Exemple 6.2 Consum d'electricitat: en la taula següent tenim el consum mensual d'electricitat a Espanya en el període 2002-2006. Cal destacar que no hi estan incloses les energies renovables ja que no n'existeixen dades mensuals. Les dades estan expressades en milers de TEP (tona equivalent de petroli, és una unitat d'energia equivalent a l'energia que hi ha en un tona de petroli).

(Fonts: INE, Ministeri d'Indústria, Turisme i Comerç)

Any / Mes	2002	2003	2004	2005	2006
Gener	1.616	1.691	1.711	1.880	1.964
Febrer	1.424	1.570	1.640	1.762	1.782
Març	1.483	1.554	1.732	1.782	1.842
Abril	1.433	1.438	1.544	1.614	1.579
Maig	1.450	1.499	1.566	1.631	1.720
Juny	1.463	1.596	1.634	1.740	1.759
Juliol	1.555	1.695	1.746	1.829	1.942
Agost	1.406	1.589	1.619	1.676	1.750
Setembre	1.434	1.542	1.640	1.664	1.765
Octubre	1.492	1.585	1.617	1.645	1.731
Novembre	1.497	1.587	1.694	1.743	1.730
Desembre	1.552	1.691	1.770	1.899	1.920

6.2. REPRESENTACIÓ GRÀFICA

Tota anàlisi d'una sèrie temporal ha d'iniciar-se amb una representació gràfica d'aquesta; en l'eix d'abscisses cal posar el temps i en el d'ordenades, els valors de la sèrie. Açò ens permet detectar les característiques més importants del fenomen, com ara el moviment a llarg termini, l'amplitud de les oscil·lacions, la possible existència de cicles, les ruptures, els valors anòmals, etc.

Mirem els gràfics de les sèries temporals que hem vist en els exemples.

Exemple 6.3 Xifres de població. Vegeu la figura 6.1.

Exemple 6.4 Consum d'electricitat. Vegeu la figura 6.2.

6.3. CARACTERÍSTIQUES D'UNA SÈRIE TEMPORAL

Una de les formes més senzilles d'analitzar una sèrie temporal és descompondre-la en una suma de quatre sumands:

$$y_t = m_t + s_t + c_t + u_t$$

on m_t rep el nom de **tendència** i recull el component de la sèrie que representa l'evolució a llarg termini de la sèrie; s_t representa un **component estacional fix**, per exemple aquelles oscil·lacions d'una sèrie temporal que es completen dins d'un any (o un període inferior a un any); c_t representa el **component cíclic fix**, per exemple les oscil·lacions que es produeixen en un període superior a un any i que es deuen principalment a l'alternança d'etapes de prosperitat i de depressió en l'activitat econòmica. Per a acabar, u_t recull la **variació residual** i representaria la part aleatòria, la deguda a l'atzar. El component cíclic és molt difícil d'obtindre i es necessita una sèrie temporal molt llarga per a poder separar-lo de la resta.

Nota 6.1 En l'exemple 6.1 observem que només hi ha tendència; en l'exemple 6.2 podem veure tendència i estacionalitat.

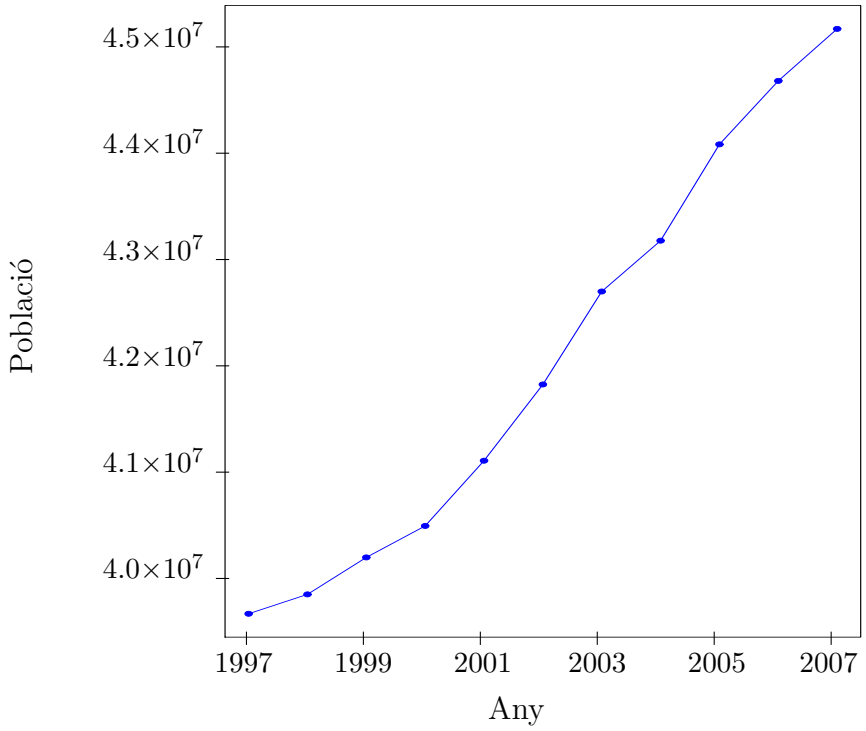


Figura 6.1: Xifres de població a Espanya des de 1997 fins 2007

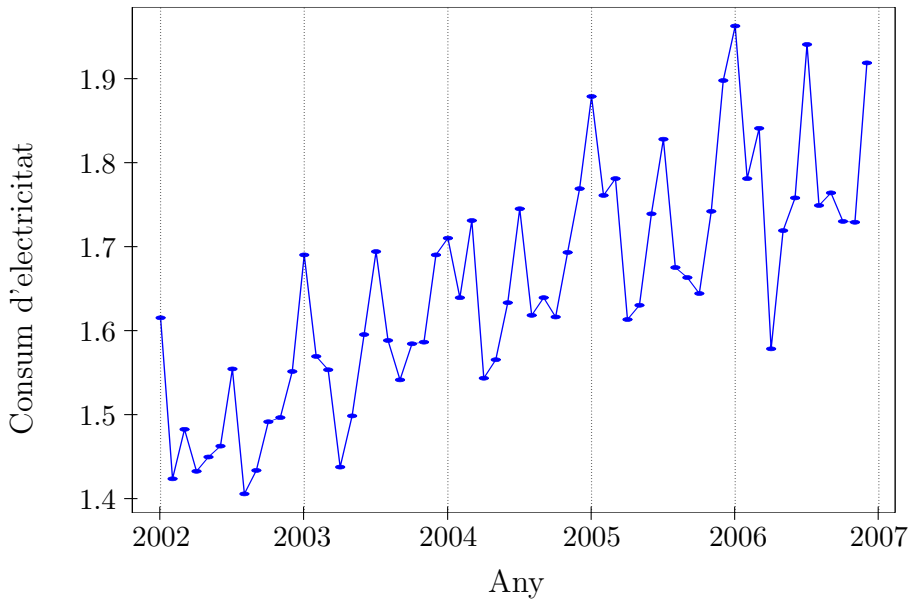


Figura 6.2: Dades de consum d'electricitat en milers de TEP per mesos

6.4. ANÀLISI DE LA TENDÈNCIA

En aquest apartat estudiem procediments per a aïllar la tendència i les variacions estacionals. Es pot fer amb dos objectius diferents: estimació de la tendència amb objecte de conéixer quines són les pautes de comportament al llarg del temps de la variable objecte d'estudi, o per a la predicció de valors futurs. Existeixen molts mètodes, entre els quals n'estudiarem únicament dos: un de global i un altre de local.

6.4.1. ANÀLISI SENSE COMPONENT ESTACIONAL

Suposem que tenim una sèrie temporal que podem descompondre com:

$$y_t = m_t + u_t$$

No tenim ni component estacional, ni component cíclic. A continuació, veurem com calcular la tendència en aquest cas.

MÈTODE DE REGRESSIÓ POLINÒMICA

Consisteix a ajustar un polinomi a les dades usant el mètode dels mínims quadrats, és a dir, tractar y_t com a variable resposta i t com a variable independent, com vam veure en el tema 4 (només hem vist $p = 1$ o $p = 2$).

En aquest cas, podrem expressar la tendència com $m_t = \sum_{j=0}^{p} b_j t^j$, on b_j s'estima a partir de les dades, minimitzant: $\sum_{t=1}^{n} (y_t - m_t)^2$.

Nota 6.2 Recordem que en el tema 4 també vam veure que prenent logaritmes en la sèrie podem usar una regressió lineal per a estimar una tendència exponencial.

El gran avantatge que representa aquest mètode és que podem donar-hi una mesura de la bondat calculant el coeficient de determinació i interpretant-lo de la manera ja coneguda.

Exemple 6.5 Ajustament de la tendència de les dades de la població espanyola en el període 1996-2007:

$$m_t = 1.25824 \times 10^{11} - 1.26241 \times 10^{8}\, t + 31675\, t^2\,.$$

MÈTODE DE MITJANES MÒBILS

Es basa en la suavització de la sèrie a partir del càlcul reiterat de valors mitjans. Una **mitjana mòbil** d'una sèrie temporal y_t és una sèrie temporal definida per:

$$m_t = \frac{1}{2p+1} \sum_{j=-p}^{p} x_{t+j}$$

on p és un enter positiu. $2p + 1$ s'anomena **ordre** de la mesura mòbil.

Nota 6.3 Observeu que el valor de m_t està indefinit prop del principi i del final de la sèrie. Una forma de completar aquesta definició, per als valors extrems, és deixar que la suma vaja des de màx$(-p, 1-t)$ fins a mín$(p, n-t)$ i dividir entre el nombre dels sumands corresponents.

Exemple 6.6 Calculem les mitjanes mòbils d'ordre 5 ($p = 2$) amb les dades de l'exemple 6.1.

Any	Tendència (m_t)	Any	Tendència (m_t)
1997	39908068	2003	42595603
1998	40055999	2004	43314027
1999	40268168	2005	43986596
2000	40701868	2006	44303979
2001	41274750	2007	44672744
2002	41873855		

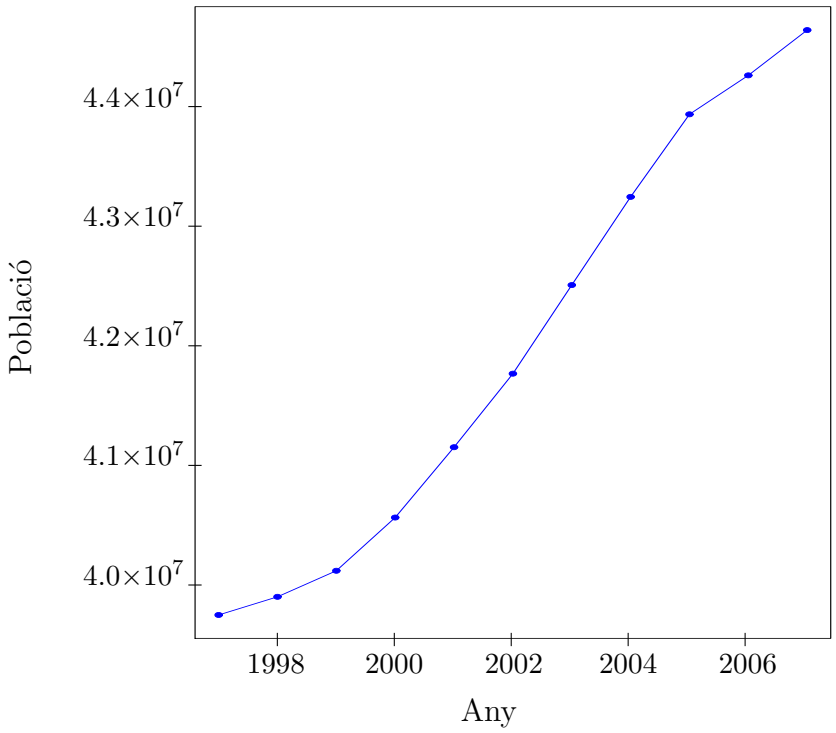


Figura 6.3: Sèrie de població després d'haver-hi aplicat el mètode de les mitjanes mòbils

6.4.2. ANÀLISI AMB COMPONENT ESTACIONAL

Els mètodes utilitzats per a eliminar la tendència poden adaptar-se d'una forma natural quan necessitem eliminar tant la tendència com l'estacionalitat, és a dir, quan tenim:

$$y_t = m_t + s_t + u_t\,.$$

Nota 6.4 Observeu que, per la definició de component estacional, existeix un d (període que tarda a completar-se una oscil·lació) de tal manera que $s_t = s_{t+d}$ i $\sum_{j=1}^{d} s_j = 0$. (Per exemple, si l'estacionalitat és anual, aleshores $d = 12$.)

MÈTODE DE LES MESURES MENSUALS (TENDÈNCIA LINEAL)

Suposem que la tendència és lineal, és a dir: $m_t = a + b\,t$. Aleshores:

(1) Estimem aquesta tendència usant les mesures anuals de les dades observades:

$$\bar{y}_j^{anual} = \frac{\sum_{k=1}^{d} y_{j-1)d+k}}{d}; \qquad j = 1, \ldots, N$$

on $N = \dfrac{n}{d}$ és el nombre de períodes o anys, i hi ajustem una recta

$$\bar{y}_j^{anual} = a + bj$$

pel mètode dels mínims quadrats (tema 4).

(2) Calculem les mitjanes mensuals:

$$\bar{y}_k = \frac{1}{N}\sum_{j=1}^{N} y_{(j-1)d+k} \qquad k = 1, \ldots, d\,.$$

(3) Per a aïllar el component estacional de la variació deguda exclusivament al pas del temps restem a cada mitjana mensual la proporció que hi correspon de l'increment anual:

$$\bar{y}'_k = \bar{y}_k - \frac{b(k-1)}{d} \qquad k = 1, \ldots, d\,.$$

(4) Finalment, per a estimar-ne el component estacional, restem a cada mitjana mensual corregida la mitjana global corregida:

$$\hat{s}_k = \bar{y}'_k - \frac{1}{d}\sum_{k=1}^{d} \bar{y}'_k\,.$$

Exemple 6.7 Apliquem aquest mètode per a estimar el component estacional de les dades de l'exemple 6.2.

(1) Mitjanes anuals:

Any	2002 (1)	2003 (2)	2004 (3)	2005 (4)	2006 (5)
Mitjana	1.48	1.59	1.66	1.74	1.79

Ajustant una recta a aquestes dades obtenim $b = 0.08$.

(2) Mitjanes mensuals:

Mes	Gener	Febrer	Març	Abril	Maig	Juny	Juliol
Mitjana	1.77	1.64	1.68	1.52	1.57	1.64	1.75

Mes	Agost	Setembre	Octubre	Novembre	Desembre
Mitjana	1.61	1.61	1.61	1.65	1.77

(3) Mitjanes mensuals corregides:

Mes	Gener	Febrer	Març	Abril	Maig	Juny	Juliol
Mitjana	1.77	1.63	1.66	1.50	1.54	1.60	1.71

Mes	Agost	Setembre	Octubre	Novembre	Desembre
Mitjana	1.56	1.56	1.55	1.58	1.69

(4) Component estacional:

Mes	Gener	Febrer	Març	Abril	Maig	Juny	Juliol
Component	0.15	0.01	0.05	-0.11	-0.07	-0.01	0.09

Mes	Agost	Setembre	Octubre	Novembre	Desembre
Component	-0.05	-0.06	-0.06	-0.03	0.08

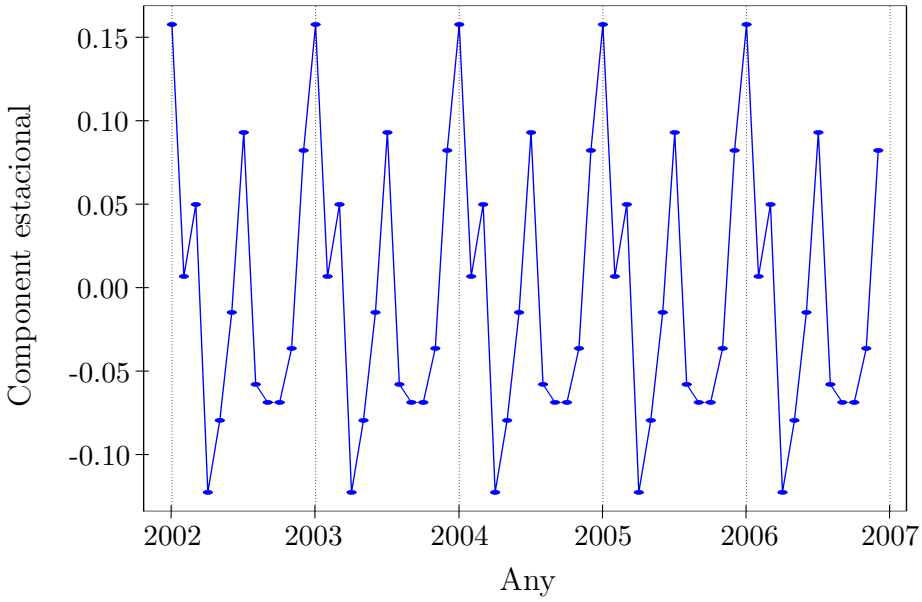


Figura 6.4: Component estacional de l'exemple 2 amb el mètode de les mitjanes mensuals

MÈTODE DE LES MITJANES MÒBILS

(1) Es tracta, en primer lloc, d'aplicar una mitjana mòbil per a suavitzar la tendència (prenem tots els valors i mitjanem), distingint si el període és parell o imparell. Així doncs:

a) Si $d = 2q + 1 \longrightarrow \hat{m}_t = \dfrac{1}{d} \displaystyle\sum_{j=-q}^{q} y_{t+j}\,.$

b) Si $d = 2q \longrightarrow \hat{m}_t = \dfrac{1}{d}(0.5\, y_{t-q} + \displaystyle\sum_{j=-q+1}^{q-1} y_{t+j} + 0.5\, y_{t+q})\ \forall q < t \leq n - q\,.$

En el cas més habitual de $d = 12$ i:

$$\hat{m}_t = \frac{1}{12}(0.5\, x_{t-6} + x_{t-5} + \cdots + x_{t+5} + 0.5\, x_{t+6}).$$

(2) El segon pas consisteix a estimar el component estacional.

a) Per a cada $k = 1, \ldots, d$ calculem la mitjana w_k de les desviacions $\{y_{k+jd} - \hat{m}_{k+jd} : q < k + jd \leq n - q\}$. És a dir:

$$w_k = \frac{1}{N-1} \sum_{j=1}^{N-1} (y_{k+jd} - \hat{m}_{k+jd}).$$

Com que aquestes mitjanes no sumen zero, estimem el component estacional com:

$$\hat{s}_k = w_k - \frac{1}{d}\sum_{i=1}^{d} w_i, \qquad k = 1, \ldots, d$$

que ja sumen 0.

(3) Per a acabar, reestimem la tendència de $\{d_t\}$, amb:

$$d_t = y_t - \hat{s}_t\,,$$

utilitzant un filtre de mitjanes mòbils per a dades sense estacionalitat (vist anteriorment), o ajustant un polinomi a les $\{d_t\}$.

Exemple 6.8 Apliquem aquest mètode per a estimar el component estacional de les dades de l'exemple 2:

Mes	Gener	Febrer	Març	Abril	Maig	Juny	Juliol
Component	0.15	0.02	0.05	-0.13	-0.08	-0.01	0.07

Mes	Agost	Setembre	Octubre	Novembre	Desembre
Component	-0.03	-0.05	-0.05	-0.01	0.07

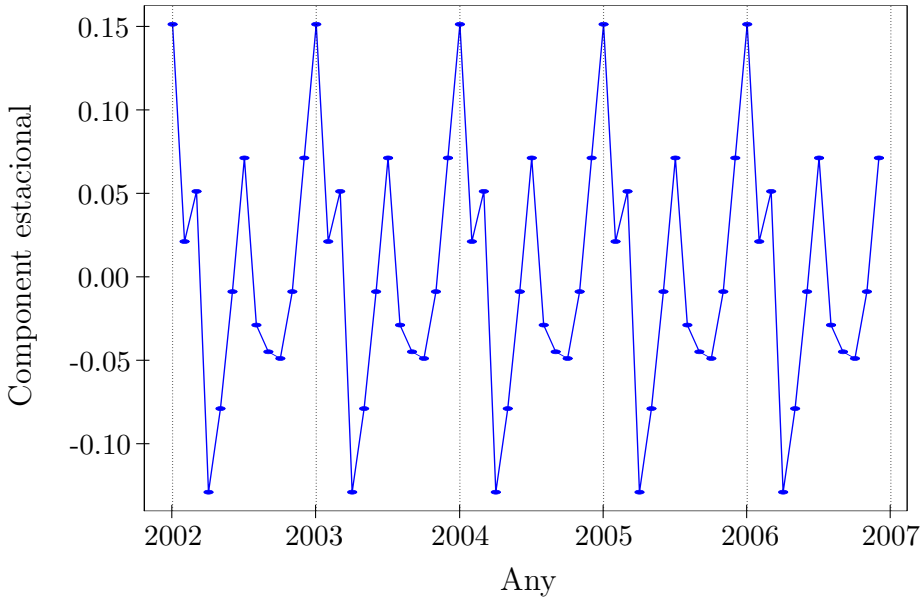


Figura 6.5: Component estacional de l'exemple 2 amb el mètode de les mitjanes mòbils

6.5. PROBLEMES PROPOSATS

(1) Amb quin component d'una sèrie temporal associaríeu cadascun dels fets següents:

a) Una vaga de treballadors.

b) Un increment de la producció del blat a causa de la incorporació de noves tècniques de conreu.

c) Un augment de les vendes d'automòbils durant el mes de maig.

d) Una recessió en el volum de construcció d'habitatges durant tres anys.

(2) En la figura 6.6 veiem la representació gràfica de cinc sèries temporals recollides en estudis independents. Identifica els components de cadascuna sabent que es corresponen amb els estudis següents:

a) Sèrie temporal d'exportacions anuals de taulells a Itàlia (en milers de metres quadrats) des de 1990 fins a 2003.

b) Sèrie del nombre mensual d'automòbils matriculats a Espanya en el període de 1998-2003.

c) Dades mensuals proporcionades per l'INE sobre el consum de gasolina a Espanya, des de gener de l'any 2000 fins gener de 2007. (Les dades estan en milers de tones.)

d) La Conselleria de Medi Ambient de la Generalitat Valenciana, desenvolupa una campanya de vigilància dels nivells de contaminació per ozó en l'atmosfera. Per a aquest fi disposa de diverses estacions de mesurament repartides per tota la comunitat. En la figura 6.6 es consideren les dades recollides diàriament en l'estació meteorològica de Penyeta Roja (Castelló), durant els anys 2006 i 2007.

e) Dades de vendes mensuals d'una empresa durant els últims anys.

(3) El volum de facturació (en milers d'euros) d'un hipermercat durant els 15 anys que està obert, ha seguit l'evolució següent:

Anys	1	2	3	4	5	6	7	8
Facturació	2500	3400	3800	4200	4700	5200	5500	6000
Anys	9	10	11	12	13	14	15	
Facturació	6500	6200	7500	8200	9000	9300	9000	

a) Estima quin serà el volum de facturació d'aquest hipermercat d'ací a 3 anys a través de la recta de tendència.

b) Calcula el coeficient que mesura el grau de bondat d'ajustament i comenta el resultat obtingut.

(4) S'han recollit dades de l'evolució de les despeses en vestit i en calçat per persona i dia durant els anys 2005, 2006 i 2007:

Any / Trimestre	2005	2006	2007
1r	8	9	11
2n	11	14	16
3r	6	8	9
4t	16	18	19

a) Identifica si aquesta sèrie temporal presenta tendència i component estacional.

b) Calcula els components que hages identificat en l'apartat anterior.

(5) L'Institut Nacional d'Estadística, en l'apartat "Estadística de Transport de Viatgers," publica les dades de milers de viatgers que han utilitzat el ferrocarril com a mitjà de transport interurbà. Les dades estan presentades amb mesures mensuals, des de l'any 1996 fins a final de 2007. En la figura 6.7 veiem la representació d'aquestes dades. A continuació es detallen les dades dels quatre últims anys:

Anys	Gener	Febrer	Març	Abril	Maig	Juny
2004	47424	46820	50148	45446	50141	49013
2005	49220	47225	49305	52144	52769	50978
2006	50808	50012	52291	48557	53890	51260
2007	51067	49175	54683	49288	53762	50565

Anys	Juliol	Agost	Set.	Oct.	Nov.	Des.
2004	45578	35083	47275	51468	50495	45820
2005	46700	37162	48535	52693	51913	48215
2006	49034	38196	48593	55246	52961	48000
2007	48489	35856	46582	53909	51133	45980

a) Quins components pots identificar en aquesta sèrie temporal? Utilitza el mètode de les mitjanes mòbils per a calcular el component estacional d'aquesta sèrie. Dibuixa-la.

b) Resta a cada dada el seu component estacional. Dibuixa la sèrie resultant.

c) Ajusta un model lineal a la sèrie obtinguda en l'apartat anterior.

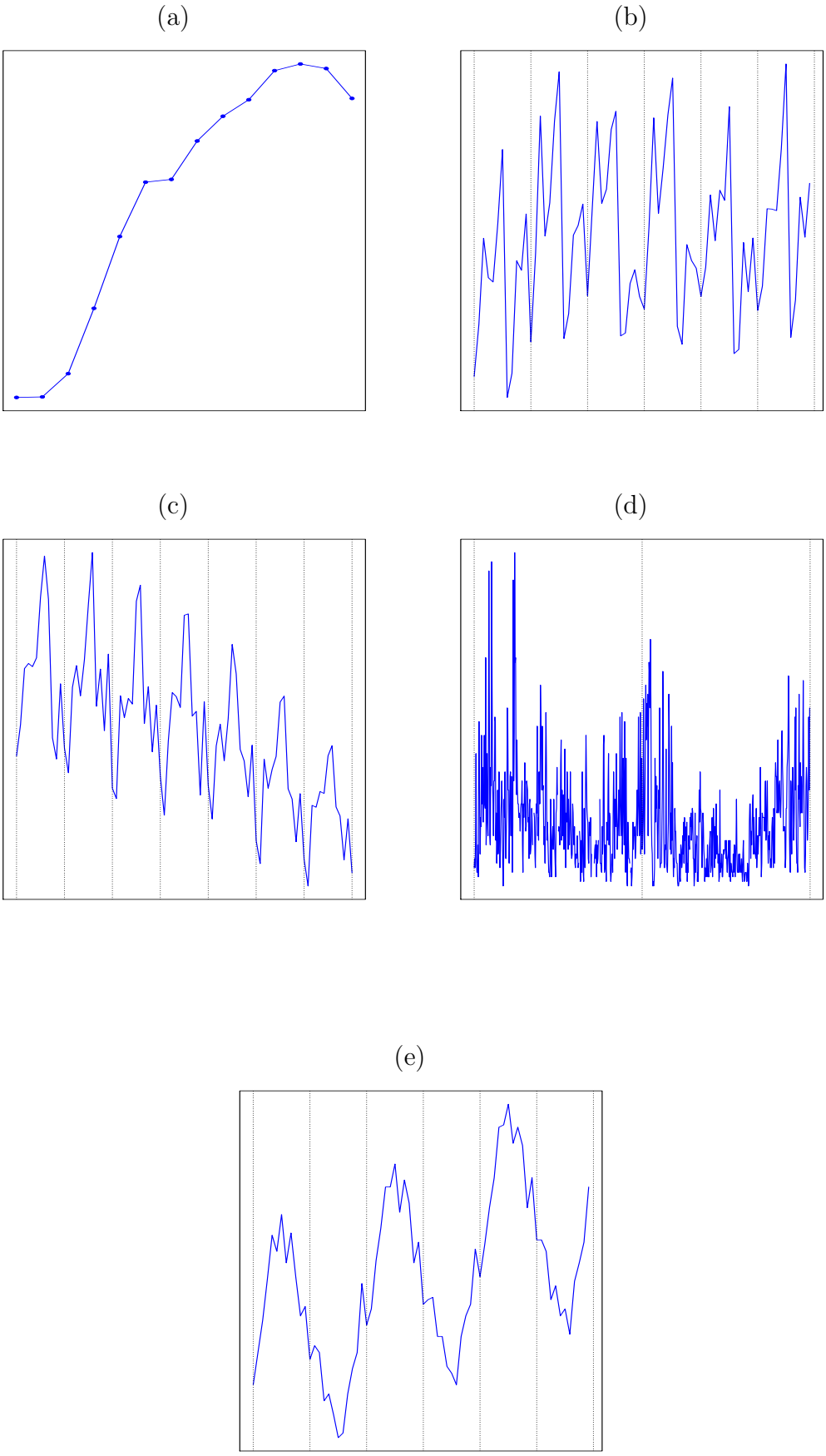


Figura 6.6: Sèries temporals del problema 2

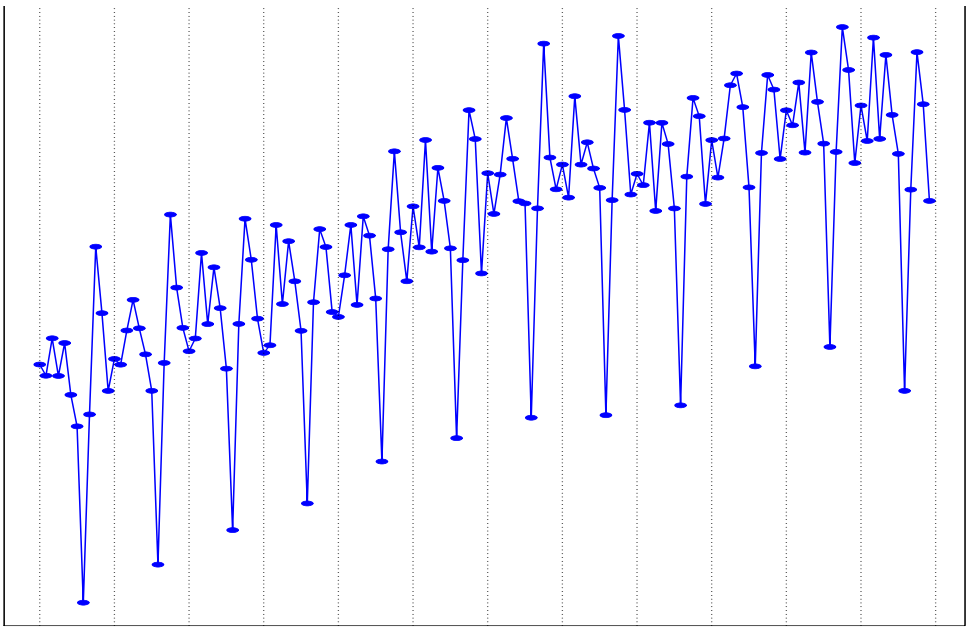


Figura 6.7: Sèrie temporal del problema 5. Milers de viatgers (per mesos) que utilitzen el tren per als trajectes interurbans

SOLUCIONS

(1) *a*) Variació residual
b) Tendència
c) Component estacional
d) Component cíclic

(2) *a*) Tendència
b) Component estacional
c) Tendència i estacionalitat
d) Variació residual
e) Tendència, estacionalitat i un componente cíclic de 2 anys de duració

(3) *a*) $m_t = 2229.52 + 479.64\,t$. Predicció: 10863.10 milers d'euros
b) 0.98

(4) *a*) Tendència component estacional
b) Pel mètode de les mitjanes mensuals (en aquest cas trimestrals)
Tendència: $y_t = 8.58 + 1.75\,t$
Component estacional: $-1.6875,\ 0.9375,\ -4.4375,\ 5.1875$

(5) *a*) Tendència i component estacional

b)

Mes	Component estacional
Gener	1087.63
Febrer	−524.54
Març	2763.35
Abril	642.41
Maig	4076.98
Juny	1526.56
Juliol	−1640.54
Agost	−12014.18
Setembre	−789.20
Octubre	4095.78
Novembre	2646.13
Desembre	−1870.38

c) $y_t = 47292.81 + 60.25\,t$

BIBLIOGRAFIA

[1] Calot, G., *Curso de Estadística Descriptiva*, Paraninfo, 1988.

[2] Canavos, G. C., *Probabilidad y Estadística*, McGraw-Hill, 1988.

[3] Durá, J. M. y López, J. M., *Fundamentos de Estadística*, Ariel, 1992.

[4] Escuder Vallés, R., *Métodos Estadísticos Aplicados a la Economía*, Ariel, 1987.

[5] García Barbancho, A., *Estadística Elemental Moderna*, Ariel, 1992.

[6] López de la Manzanera, J., *Problemas de Estadística*, Pirámide, 1989.

[7] Martín Guzmán, P. y Martín-Pliego J., *Curso Básico de Estadística Económica*, A.C., 1993.

[8] Martín Pliego, F. J., *Curso práctico de estadística económica*, A.C., 1987.

[9] Martín Pliego, F. J., *Introducción a la Estadística Económica y Empresarial*, A.C., 1994.

[10] Mendenhall, W. y Reinmuth, J., *Estadística para Administración y Economía*, Grupo Editorial Iberoamérica, 1981.

[11] Montiel, A. M., Rius, F. y Barón, F. J., *Elementos Básicos de Estadística Económica y Empresarial*, Prentice Hall, 1997.

[12] Murgui, J. S., Aybar, C., Casino, A., Colom, C., Cruz, M. y Yagüe, R., *Estadística para Economía y Administración de Empresas: Aplicaciones y Ejercicios*, Puchardes, 1992.

[13] Newbold, P., *Estadística para los negocios y la Economía*, Prentice Hall, 1997.

[14] Peña, D., *Estadística: modelos y métodos, Vol. 1 (Fundamentos)*, Alianza Universidad, 1991.

[15] Spiegel, M. R., *Estadística*, McGraw Hill, 1997.

[16] Tomeo Perucha, V. y Uña Juárez, I., *Lecciones de Estadística Descriptiva. Curso teórico-práctico*, Thomson, 2003.