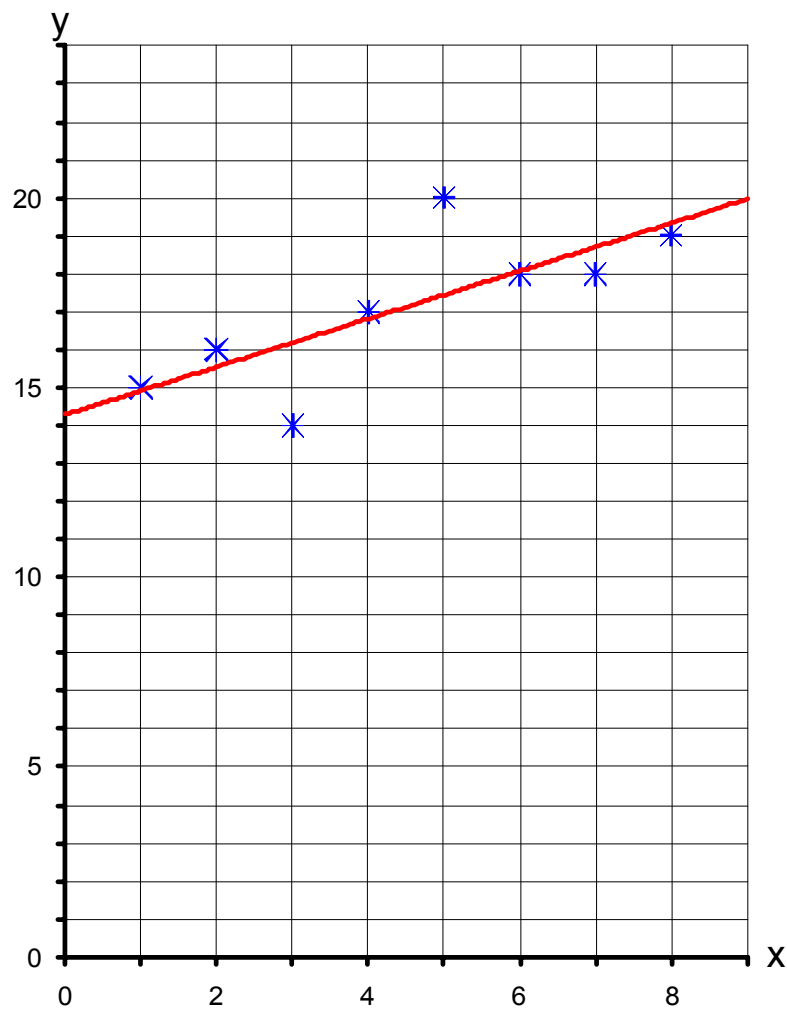


IES Arquitecte Manuel Raspall

Matemàtiques

Estadística

Distribucions Bidimensionals



Batxillerat

Distribucions bidimensionals

FULL DE TREBALL A

A.1 Un entrenador ha cronometrat els temps de pas per diferents distàncies d'un atleta que ha començat a córrer en una pista d'atletisme. Ha obtingut els següents resultats:

Temps (minuts)	0	2	3	3	4
Distància (m)	150	600	750	900	1.200

A partir d'aquestes dades es vol saber quina distància havia fet quan portava 1.5 minuts de cursa i predir quina distància haurà fet al cap de 5.5 minuts.

A.2 Una fàbrica de medicaments ha desenvolupat un nou analgèsic i vol determinar la dosi òptima per un adult. Per a fer-ho subministra diferents quantitats d'analgèsic a 7 pacients i observa la variació de temperatura corporal, en valor absolut, al cap d'un hora. Els resultats son els següents:

Mil·ligram	350	400	500	550	600	700	800
Variació temperatura	0.2	0.4	0.7	1.1	1.1	1.6	1.9

Si es vol aconseguir una variació de temperatura d'1° ¿Quina dosi caldrà subministrar?. Justifica la teva resposta.

UNA MICA DE VOCABULARI:

Algunes vegades ens pot interessar conèixer si dues variables estadístiques estan relacionades. Per a fer-ho, es pot començar considerant que les variables estadístiques son les variables X , Y d'uns eixos de coordenades i dibuixar tots els punts. Aquest dibuix s'anomena **diagrama de dispersió** o **núvol de punts** i ens permet fer una primera estimació de si la suposada relació existeix, de si és forta o fluixa i de quin tipus de relació hi ha.

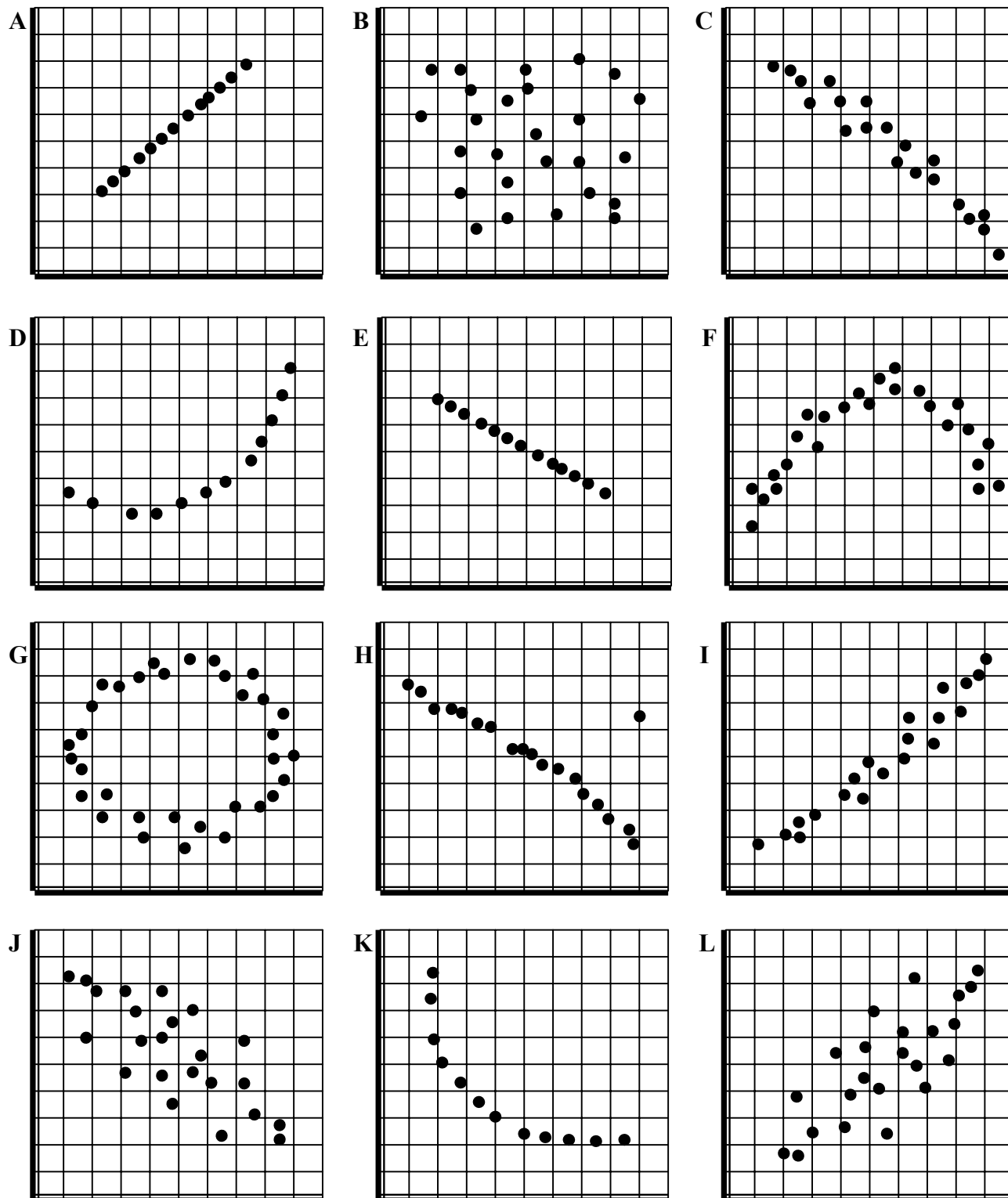
Si existeix relació entre dues variables, aquesta pot ser de molts tipus. Si es tracta, per exemple, de dos variables físiques relacionades per alguna fórmula matemàtica (com el primer exercici) direm hi ha una **relació funcional** doncs existeix una funció (fórmula) que les relaciona. En aquest cas, els punts del diagrama de dispersió estan perfectament ordenats seguint una línia recta o corba.

Quan dues variables estan relacionades però no existeix una relació funcional entre elles els punts del diagrama de dispersió envolten una recta o corba invisible. En aquests casos direm que existeix una **relació lineal o curvilínia** segons el cas.

En aquest tema estudiarem els casos en que hi hagi relació lineal. En aquests casos, la recta a la que s'aproximen tots els punts s'anomena **recta de regressió**. El problema matemàtic al que ens enfrontem consistirà, en primer lloc en determinar la *intensitat* de la relació i en segon lloc en trobar l'equació de la recta de regressió per a poder fer prediccions fiables.

A.3

Observa els diagrames de dispersió següents:



Respon a les preguntes següents per als diagrames anteriors:

- Existeix alguna relació entre les variables X i Y ?
- Si existeix, ¿És lineal, corbada,...?
- Es la relació directa (positiva) o inversa (negativa)?
- És la relació funcional?. Si és que no: ¿És forta o fluixa?

COEFICIENT DE CORRELACIÓ LINEAL

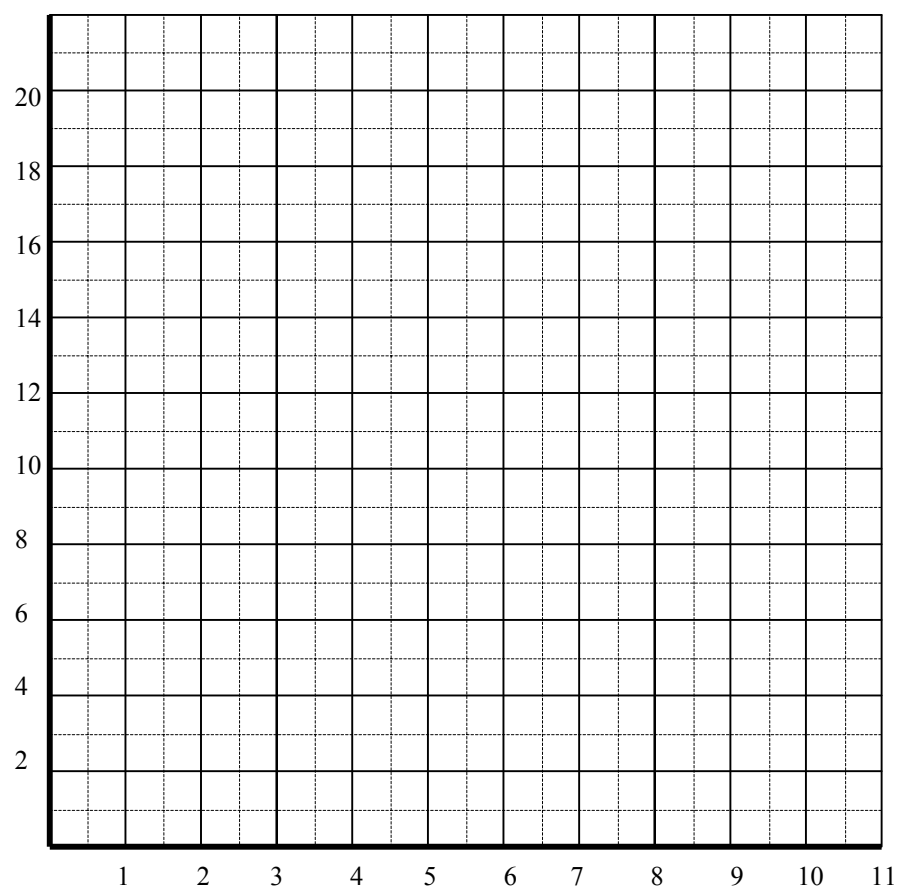
FULL DE TREBALL B

B.1 Per tal de quantificar la possible relació que hi ha entre dos variables estadístiques necessitem un valor numèric que determini d'una manera matemàticament rigorosa quina és la *intensitat* d'aquesta possible relació. Aquest valor s'anomena **coeficient de correlació lineal**. Per entendre bé d'on surt aquest coeficient ho veurem en un exemple numèric molt senzill.

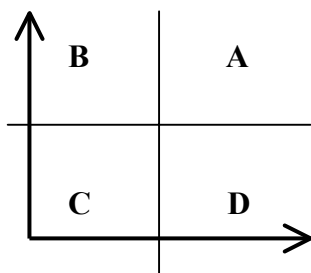
Treballarem tal com vam desenvolupar les fórmules de la desviació tipus: utilitzant un model en que entenem bé el per què del que fem trobarem unes formules que després caldrà simplificar. Les fórmules simplificades seran les que utilitzarem sempre.

Pots utilitzar la següent taula i el següent gràfic per a fer els exercicis:

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	x_i^2	y_i^2	$x_i \cdot y_i$
8	20						
5	9						
10	16						
5	5						
7	11						
2	6						
3	5						
11	21						
1	1						
8	16						



- a) Dibuixa els punts en el diagrama de dispersió
- b) Calcula la mitjana de les dues variables: $\bar{x}$ i $\bar{y}$
- c) Si considerem aquests valors $\bar{x}, \bar{y}$ com a coordenades podem dibuixar un nou punt $(\bar{x}, \bar{y})$ en el núvol de punts. Dibuixa'l. Aquest punt s'anomena **punt mitjà**. Observa que aquest punt constitueix una espècie de *centre de gravetat* del núvol de punts. Sembla, doncs, necessari que qualsevol recta que vulgui ajustar-se bé al núvol de punts tingui que passar, necessàriament, per aquest punt.
- d) Calcula la desviació tipus de les dues variables: σ_x i σ_y (utilitza les columnes x_i^2 i y_i^2 de la taula)
- e) Dibuixa, en color blau, uns altres eixos de coordenades amb origen en el **punt mitjà** $(\bar{x}, \bar{y})$
- f) Calcula, per a cada punt, les distàncies que hi ha entre el punt i aquests nous eixos blaus. (observa que pots utilitzar la 2^a i 3^a columna de la taula)
- g) Calcula el producte $(x_i - \bar{x})(y_i - \bar{y})$ per a cada punt (utilitza la columna corresponent de la taula)
- h) Considera els quadrants A, B, C i D dels nous eixos de coordenades:



- i) Omple la taula posant el signe dels valors segons la diferent posició del punt:

	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
En A			
En B			
En C			
En D			

- j) Observa que si sumem tots els valors $(x_i - \bar{x})(y_i - \bar{y})$, aquesta suma serà, en valor absolut, més alta contra més alineats estiguin els punts, doncs la majoria de valors $(x_i - \bar{x})(y_i - \bar{y})$ seran del mateix signe. Aquesta suma és, per tant un paràmetre matemàtic que ens pot indicar, d'alguna manera, el grau de relació que hi ha entre les dues variables.

Calcula, doncs, la suma $\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})$

- k) El paràmetre que acabes de calcular en l'exercici anterior no és gens significatiu, doncs depèn de la quantitat de punts que tenim. Es a dir, si en comptes de tenir 10 punts en tinguéssim 100 el valor de la suma seria molt més alta encara que la relació fos la mateixa o fins i tot més baixa. Per a que el paràmetre ens serveixi igual independentment de la quantitat de punts que tinguem caldrà dividir la suma pel nombre de punts que tenim. Aquest valor s'anomena **covariància** i és designa per σ_{xy} :

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

Calcula la covariància de l'exemple que estem fent.

- l) La covariància encara no és prou representativa doncs imagina que volem estudiar la relació que hi ha entre longitud i pes de les formigues i dels elefants. La covariància en el cas dels elefants seria molt més elevada que en les formigues, no per que la relació sigui més forta si no per què la magnitud de les mesures en els elefants és més gran. Aquest problema es resol fàcilment dividint la covariància entre el producte de les desviacions tipus de les dues variables. Així definim el **coeficient de correlació lineal** com:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Calcula el coeficient de correlació lineal de l'exemple que estem treballant.

INTERPRETACIÓ DEL COEFICIENT DE CORRELACIÓ LINEAL:

El coeficient de correlació lineal és un valor que oscil·la entre -1 i 1 .

- Si $r = -1$ es tracta d'una relació funcional inversa (la recta de regressió tindrà pendent negatiu)
- Si r és prop de -1 existeix una relació inversa no funcional entre les variables que serà més forta contra més prop de -1 estigui el valor.
- Si r és prop de 0 no existeix relació entre les variables.
- Si r és prop de 1 existeix una relació directa no funcional entre les variables que serà més forta contra més prop de 1 estigui el valor.
- Si $r = 1$ es tracta d'una relació funcional directa (la recta de regressió tindrà pendent positiu)

CÀLCUL DE LA COVARIÀNCIA (fórmula simplificada)

B.2 Igual com vam fer amb la desviació tipus cal que simplifiquem la fórmula de la covariància per tal que sigui més senzilla de calcular.

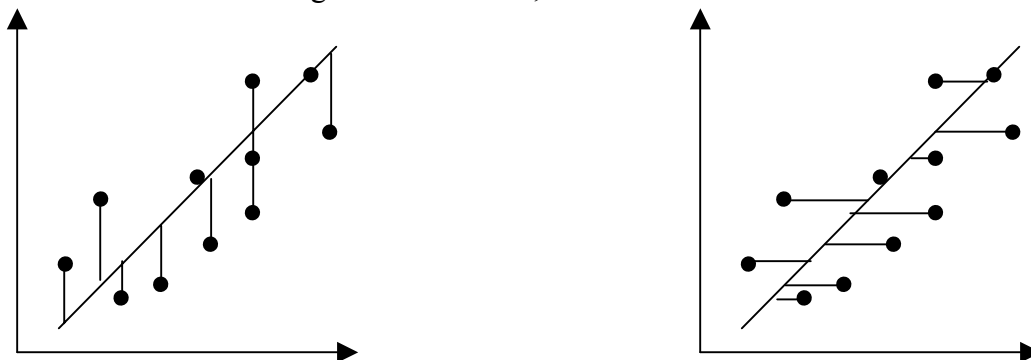
- a) Demostra que:

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

- b) Calcula novament la covariància de l'exemple utilitzant la nova fórmula (observa que et caldrà utilitzar les últimes columnes de la taula)

RECTA DE REGRESIÓ

B.3 Un dels objectius fonamentals d'aquest tema és fer prediccions. Es a dir, una vegada determinat que sí existeix una determinada correlació lineal entre dues variables ens plantejem la possibilitat de, coneixen una nova dada x , endevinar la nova dada y . Per a fer-ho, tan sols ens cal trobar l'equació matemàtica de la recta que més s'aproxima a la majoria de punts. La tècnica matemàtica que permet la recerca d'aquesta equació s'anomena mètode dels *mínims quadrats* i consisteix en trobar la recta que fa que el quadrat de la longitud dels segments que va dels punts a la recta sigui mínima. Aquest mètode es pot aplicar de dues maneres: Considerant el segment en vertical, o considerant-lo en horitzontal:



Aquesta diferència de mètodes produeix dues rectes **diferents** una serveix per aïllar la x en funció de la y i l'altra serveix per aïllar la y en funció de la x .

Recta de regressió de y en funció de la x : $y = Ax + B$

$$A = \frac{\sigma_{xy}}{\sigma_x^2} \quad B = \bar{y} - A\bar{x}$$

Recta de regressió de x en funció de la y : $x = A'y + B'$

$$A' = \frac{\sigma_{xy}}{\sigma_y^2} \quad B' = \bar{x} - A'\bar{y}$$

B.4 Utilitzant l'exercici B.1

- Calcula la recta de regressió de y en funció de x
- Dibuixa-la en el diagrama de dispersió
- Calcula la recta de regressió de x sobre y
- Dibuixa-la en el diagrama de dispersió
- ¿Coincideixen les dues rectes?

B.5 Utilitzant les dades de les noies de la classe

- Dibuixa el diagrama de dispersió
 - Calcula les mitjanes, variàncies, desviacions tipus, covariància, i coeficient de correlació.
 - Comenta raonadament si creus que existeix relació entre les dues dades.
 - Calcula la recta de regressió de y sobre x i utilitza-la per endevinar el pes que pot tenir una noia de la vostra edat que mesuri uns 168 cm. d'alçada.
- Podeu utilitzar la taula de la pàgina següent per a fer l'exercici:

Qualsevol calculadora científica (excepte l'antiga *Casio Fx 82*) calcula tots els paràmetres estadístics de la correlació lineal. Per a fer-ho cal posar la calculadora en **MODE LR** (*Linear Regression*) i introduir les dades separant-les per una coma (consulteu el vostre manual o pregunteu al professor com es fa).

C.3 El nombre de bacteris per unitat de volum, presents en un cultiu després d'un cert nombre d'hores, és expressat en la següent taula:

Hores	0	1	2	3	4	5
Bacteris per unitat de volum	12	19	23	34	56	62

Fes el diagrama de dispersió i calcula:

- Les mitjanes i desviacions tipus de les variables *nombre d'hores* i *nombre de bacteris*.
- La covariància i el coeficient de correlació lineal de la variable bidimensional.
- ¿És pot fer una predicció fiable del bacteris que hi haurà a les 6 hores? Justifica la resposta i fes, si es cau, la predicció
- ¿I a les 20 hores?. Justifica la resposta i fes, si es cau, la predicció.

C.4 Dibuixeu el núvol de punts de les dades següents:

x	1	1	2	3	5	8	10	10
y	9	8	7	5	3	3	10	1

Estimeu i calculeu el coeficient de correlació lineal i la recta de regressió. Hi ha una dada causant del baix valor que té r ; traieu-la i calculeu novament el coeficient de correlació i la recta de regressió. Dibuixeu les dues rectes i compareu els resultats.

C.5 L'encarregat d'un bar ha triat una mostra a l'atzar de 10 setmanes durant un any i n'ha mesurat la temperatura mitjana així com també, el nombre de caixes de begudes refrescants venudes, i ha obtingut la taula següent:

Temperatura	10	28	11	32	30	18	25	5	8	15
Nº de caixes	21	65	18	72	75	40	65	10	12	25

Utilitzant aquestes dades, ¿Es possible planificar la compra de begudes refrescants en funció de la temperatura?. ¿De quina manera?

C.6 S'ha observat que existeix una forta correlació entre el nombre d'hectàrees cremades en incendis a Espanya i el nombre d'hectàrees cremades a la resta d'Europa. ¿Significa això que Espanya te la culpa de que es cremin els boscos a Europa?.

Quan existeix correlació entre dues variables pot ser que:

- Que una sigui causa de l'altra
- Que hi hagi influència mútua
- Que les dues estiguin influïdes per terceres variables.
- Que la correlació sigui fruit de l'atzar

Suposem que entre les següents variables s'ha observat una forta correlació. ¿A quina de les causes anteriors creus que es deu la correlació:

- a) Gruix de la neu i nombre d'esquiadors
- b) Sou de funcionaris i consum d'aliments precuinats
- c) Preu del tabac i grau de contaminació industrial
- d) Pluja i collita agrícola
- e) Oferta de serveis i demanda de serveis.
- f) Nivell cultural i dieta alimentària
- g) Absentisme laboral i nivells de lectura.
- h) Consum i sou
- i) Hores de treball i unitats produïdes

C.7 Un individu vol obtenir un benefici de 15 000 € invertint en Borsa . Per això investiga alguns experts inversors i recull la següent informació d'ells: diners guanyats després de fer certa inversió.

Inversió (en milers d'€)	200	300	50	400	70	130	250	30
Beneficis (en milers d'€)	20	5	10	20	0	30	2	10

¿Quants diners cal invertir en borsa per tal d'obtenir els 15 mil euros de benefici desitjat?

C.8 Les qualificacions en Matemàtiques i física de 20 alumnes de segon de BUP han estat les següents:

Nota Matemàtiques (x_i)	2	4	5	6	6	7	7	8	10
Nota Física (y_i)	3	5	4	6	5	6	7	9	10
Nombre alumnes (f_i)	2	3	6	2	2	2	1	1	1

- a) Calculeu el coeficient de correlació lineal entre x i y .
- b) Si un alumne no ha pogut assistir a l'examen de Matemàtiques però sí al de Física, i té un 5 en aquesta darrera assignatura, ¿Quina nota és d'esperar que hagués tret de Matemàtiques?
- c) Si un alumne té un 4 de Matemàtiques ¿Quina nota és previsible que tregui a física?
- d) Seria just posar la nota d'un alumne utilitzant aquest mètode? ¿Per què?

C.9 Demostra que el producte dels coeficients de regressió (els pendents de les dues rectes de regressió) és igual al quadrat del coeficient de correlació lineal.

C.10 De dues variables x i y , sabem que la recta de regressió de y sobre x és: $y = 0.8x - 1$, i la de y sobre x és: $x = \frac{8}{19}y + \frac{71}{19}$. Trobeu:

- a) Les dues mitjanes
- b) El coeficient de correlació lineal.
- c) ¿Quin és el signe del coeficient de correlació? ¿per què?

C.11

Estimat Alumne d'estadística de 1r de Batxillerat:

Soc el president del tribunal electoral i he rebut una greu denúncia del partit progressista respecte a presumptes irregularitats en el vot per correu en les últimes eleccions. M'he pres la confiança de demanar-li un **informe tècnic** en el que exposi, **justificadament**, la seva opinió al respecte, doncs és de tot hom reconegut el seu alt prestigi com alumne del Institut.

La denúncia presentada pel partit progressista afirma que el partit conservador ha comès certes irregularitats en el vot per correu en les recents eleccions. Les proves presentades pels denunciants són ambigües i, per contra, el partit conservador afirma que el bon resultat en el vot per correu és fruit d'una campanya electoral exhaustiva dedicada, específicament, a aquest sector de la població.

En un intent d'esbrinar la veritat he recollit les estadístiques d'eleccions passades per tal de comparar els següents valors:

x = Diferència (positiva), en percentatge, entre els vots obtinguts pel partit progressista i els vots obtinguts pel partit conservador **en les urnes**

y = Diferència (positiva), en percentatge, entre els vots obtinguts pel partit progressista i els vots obtinguts pel partit conservador **en el vot per correu**

x	2.5	3.7	1.3	1.6	7.6	9.4	6.2	4.9	4.5	3.9
y	4.2	6	1.4	3.1	9.9	14.6	9.8	7.5	6.3	5.3

En les recents eleccions en que s'ha produït el conflicte les diferències en percentatge de vots han estat les següents: $x = 5.1$ $y = 15.9$

Li prego m'envii amb prestesa les conclusions a les seves investigacions el més justificades possibles per tal d'ajudar-me a dictar la sentència que anul·li o no el vot per correu.

Atentament i agraïnt-li el seu inestimable ajut:
Sr. Just Deltot i Massa

(Aquesta història està basada en uns fets reals succeïts amb els vots de Filadèlfia a les eleccions senatorials de l'estat de Pensilvània de l'any 95. En aquest cas real el president del tribunal va utilitzar l'informe d'un prestigiós matemàtic per dictar sentència)