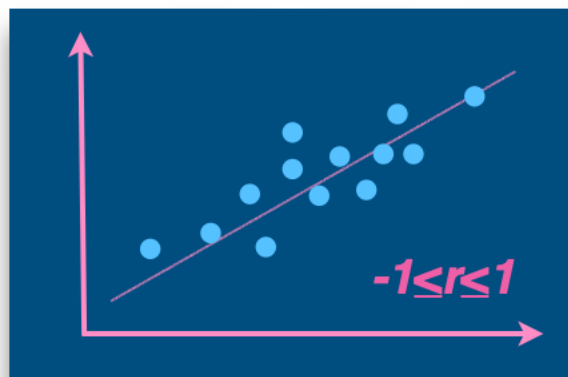


Matemàtiques I

Lliurament 4: Estadística bidimensional



Aquesta obra està subjecta a les condicions de llicència CREATIVE COMMONS no comercial i compartir igual.

Edició \LaTeX : © Josep Mulet

Versió: 19-11-2020

[Reconeixement-NoComercial-CompartirIgual 4.0 Internacional](#)



Índex

1	Introducció	3
2	Estadística descriptiva univariant	4
3	Estadística bidimensional	9
3.1	Núvol de punts	11
3.2	Covariància. Coeficient de correlació	13
3.3	Repàs de la funció lineal	18
3.4	Recta de regressió lineal	20

1. Introducció

■ Conceptes

En les unitats d'estadística de l'ESO s'han estudiat les distribucions d'una variable. Alguns exemples de variables estadístiques, les quals generalment s'indiquen com x_i , són:

- El nom del partit polític votat en unes eleccions autonòmiques. x_i = Nom del partit
- El nombre de fills en cadascuna de les famílies d'un bloc de pisos. x_i = Número de fills
- El nombre de mòbils que han tingut cadascun dels alumnes d'aquesta classe. x_i = Número de mòbils
- El pes dels alumnes de 3r d'ESO d'un institut en concret. x_i = Pes (kg)
- Salaris d'un grup de treballadors d'una fàbrica. x_i = Salari €

Les variables estadístiques es classifiquen segons el tipus:

- **Qualitativa**: indica una preferència: partit polític, color preferit, marca de mòbil, etc.
- **Quantitativa**: és una quantitat que es pot expressar numèricament. Aquesta la classificam en:
 - **Quantitativa discreta**: Es descriu amb un nombre enter com ara el nombre de fills, el nombre de mòbils...
 - **Quantitativa contínua**: Es descriu amb un nombre real (amb decimals). Per exemple: El pes, estatura, salari, temps...

Així mateix, per a cada estudi podem identificar la **població** (individus als quals afecta) i la **mostra** (la part de la població de la qual prendrem dades). A tall d'exemple:

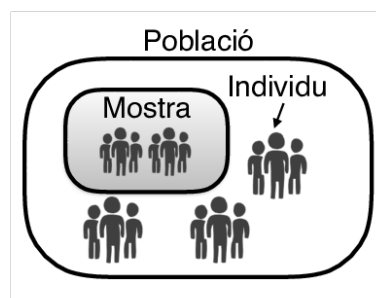


Figura 1: Concepte de individu, població i mostra.

- "Votació de partits polítics en unes eleccions autonòmiques".
 - Població: Tots els habitants de la comunitat majors d'edat inscrits al cens electoral.
 - Mostra: Cridarem telefònicament dos individus a l'atzar de cada població de la comunitat.
- "El nombre de fills en un de les famílies d'un bloc de pisos".
 - Població: Cadascuna de les famílies del bloc
 - Mostra: En aquest cas, coincideix amb la població.
- "El pes dels alumnes de 3r d'ESO d'un Institut en concret"
 - Població: Tots els alumnes matriculats a 3r d'ESO de l'Institut EscolaIB
 - Mostra: Seleccionam a l'atzar 5 alumnes de cada grup de 3r d'ESO.

Com veim, és imprescindible considerar una mostra quan la dimensió de la població és massa gran com per poder recopilar les dades de tots els seus individus.

EXERCICI RESOLT 1

Classifica les següents variables com a qualitatives o quantitatives, i aquestes últimes com a contínues o discretes.

1. Intenció de vot d'un partit.
2. Nombre de correus electrònics que reps en un mes.
3. Número de calçat.
4. Nombre de quilòmetres recorreguts en cap de setmana.
5. Marques de cervesa.
6. Nombre d'empleats d'una empresa.
7. Altura d'una persona.
8. Temperatura d'un malalt.

1. Qualitativa, 2. Quantitativa discreta, 3. Quantitativa discreta, 4. Quantitativa contínua, 5. Qualitativa, 6. Quantitativa discreta, 7. Quantitativa contínua, 8. Quantitativa contínua

2. Estadística descriptiva univariant

L'estadística parteix d'una gran quantitat de dades i intenta extreure i resumir-ne la informació més rellevant. Per exemple, al butlletí de notes a final de curs no apareixen totes les notes de les tasques del curs; tan sols un paràmetre (la mitjana) que resumeix l'evolució de l'alumne.



Parlam, doncs, dels **paràmetres estadístics** com aquells valors que extreuen informació rellevant de la distribució. Coneixem molt bé la mitjana aritmètica, però, n'existeixen molts altres, cadascun amb el seu significat i utilitat. Podem agrupar els paràmetres estadístics en dues categories:

- Paràmetres de **posició**: De forma semblant a la mitjana, ens indiquen quina posició ocupen els individus dins del rang de valors de la distribució.
- Paràmetres de **dispersió**: Ens informa de com diferents (variabilitat) són les dades amb respecte la mesura de posició.

Mesures de posició

Donada una variable estadística x_i amb cada valor repetit f_i vegades (freqüència absoluta), es defineixen els següents **paràmetres estadístics de posició**

- **Nombre de dades**: $N = \sum_i f_i$
- **Mitjana aritmètica**: $\bar{x} = \frac{\sum_i f_i x_i}{N}$
- **Moda**: El valor de x més freqüent.
- **Mediana**: Valor de x pel qual la freqüència acumulada assoleix el 50%.

El símbol \sum_i significa la suma de tots els elements enumerats per l'índex $i = 1, 2, 3, \dots$

Anem a veure un exemple que posa de manifest que la mitjana no dona tota la informació. Considereu dos alumnes: l'Andreu que ha obtingut les notes 5, 4 i 6; i na Maria ha tret 2, 9 i 4. Fixeu-vos que en ambdós casos la mitjana és 5, però està clar que les notes de l'Andreu varien poc mentre que les de na Maria tenen una gran dispersió. En el primer cas el rang=6-4=2 i, en el

segon cas, rang=9-2=7. Tot i que el rang és una mesura de dispersió fàcil d'obtenir, generalment s'utilitza la desviació típica que passam a definir.

Mesures de dispersió

- **Rang:** La diferència entre els valors major i menor de x .

- **Variància:** $Var = \frac{\sum_i f_i x_i^2}{N} - \bar{x}^2$

- **Desviació típica:** $\sigma = \sqrt{Var}$

- **Coeficient de variació:** $CV = \frac{\sigma_x}{\bar{x}}$

La **mitjana** és el **centre de gravetat** de la distribució; si el diagrama de barres estigues construït amb barres de fusta, la mitjana és el valor de x pel qual les barres queden en equilibri. En canvi, la **desviació típica** ens informa de la **dispersió**, és a dir, ens diu com d'allunyades estan les dades respecte de la mitjana. Podem pensar que una dada serà *normal* si es troba dins l'interval de x ($\bar{x} - \sigma, \bar{x} + \sigma$).

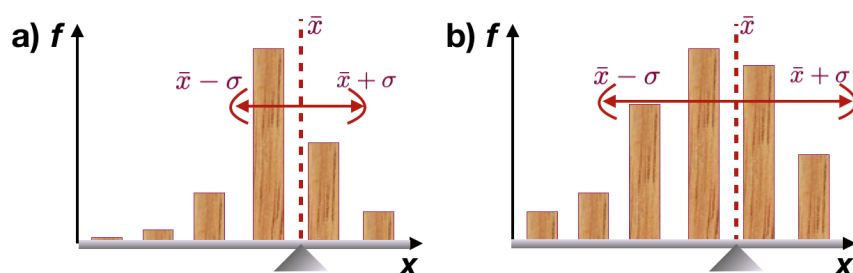


Figura 2: Significat de la mitjana i la desviació típica. Les dues distribucions tenen igual mitjana però a) té menor dispersió que b).

Tot seguit farem dos exemples de com es realitzen els càlculs d'aquests paràmetres en dos estudis estadístics: el primer d'una variable quantitativa discreta i, el segon, d'una agrupada en intervals.

EXERCICI RESOLT 2

Hem demanat pel nombre d'assignatures suspeses durant la primera avaluació a un grup de 15 alumnes i aquestes han estat les respostes:

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

- a) Fes un recompte i dibuixa un diagrama de barres.
b) Calcula la mitjana i la desviació típica.

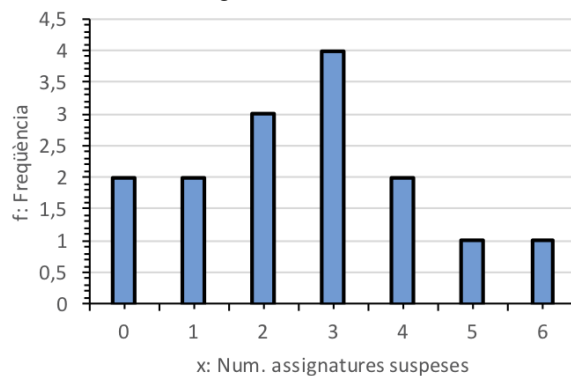
La variable estadística és $x_i = N$. d'assignatures suspeses i es tracta d'una variable quantitativa discreta. El rang de la variable és de 0 a 6.

Començam fent un recompte quan freqüent (f_i) és cada valor (x_i) de la variable:

x_i	0	1	2	3	4	5	6
f_i	2	2	3	4	2	1	1

(1)

Amb aquesta taula podem dibuixar un diagrama de barres



Per calcular els paràmetres estadístics necessitam calcular dues columnes més $f \cdot x$ i $f \cdot x^2$:

x_i	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
0	2	0	0
1	2	2	2
2	3	6	12
3	4	12	36
4	2	8	32
5	1	5	25
6	1	6	36
SUMES	15	39	143

(2)

Número total de dades $N = 15$

La mitjana s'obté de $\bar{x} = \frac{39}{15} = 2.6$

La desviació típica $\sigma = \sqrt{\frac{143}{15} - 2.6^2} = 1.67$

El coeficient de variació és $CV = \frac{1.67}{2.6} = 0.64$, aproximadament un 64 %.

(c) Josep Mulet (2020)

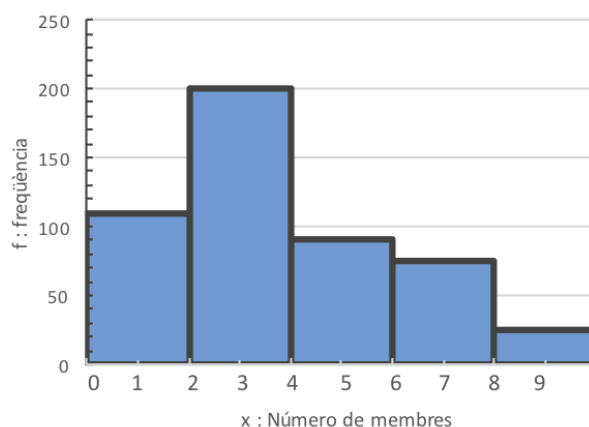
EXERCICI RESOLT 3

En un barri s'ha trobat que les famílies residents s'han distribuït, segons el número de membres, de la forma següent:

membres	n. famílies
0 – 2	110
2 – 4	200
4 – 6	90
6 – 8	75
8 – 10	25

- a) Representa un histograma.
b) Calcula la mitjana i la desviació típica.

a) Es tracta d'una variable discreta (número de membres) que s'ha agrupat en intervals. El número de famílies és la freqüència. Aleshores, el gràfic més adequat és fer un histograma. Amb la taula podem dibuixar un histograma



b) Per calcular els paràmetres estadístics necessitam calcular **la marca de classe** que és el punt mitjà de cada interval. Aquesta serà la nostra x_i per realitzar els càlculs.

x_i : marca	f_i	$f_i \cdot x_i$	$f_i \cdot x_i^2$
1	110	110	110
3	200	600	1800
5	90	450	2250
7	75	525	3675
9	25	225	2025
SUMES	500	1910	9860

(3)

El número de dades és $N = 500$

La mitjana s'obté de $\bar{x} = \frac{1910}{500} = 3.82$

La desviació típica $\sigma = \sqrt{\frac{9860}{500} - 3.82^2} = 2.26$

El coeficient de variació és $CV = \frac{2.26}{3.82} = 0.57$, aproximadament un 60%.

EXERCICIS PROPOSATS

- Una entitat bancària disposa de 50 sucursals en les quals hi treballen un nombre determinat d'empleats. Anomenem x_i el nombre d'empleats i f_i el nombre d'entitats en el quals hi treballen aquest nombre d'empleats. Hem obtingut les següents sumes: $\sum_i f_i = 50$, $\sum_i f_i \cdot x_i = 654$, $\sum_i f_i \cdot x_i^2 = 8862$. Amb aquestes dades calcula la mitjana i la desviació típica del nombre d'empleats.
- S'ha realitzat un estudi a 100 dones majors de 15 anys i se'ls hi ha demanat el nombre de fills que tenien. S'han presentat els resultats en forma de taula de freqüències:

N. fills (x_i)	0	1	2	3	4	5	6
N. do- nes (f_i)	13	20	25	20	11	7	4

- Identifica la variable estadística i indica el seu tipus.
- Representa un diagrama de barres.
- Calcula la mitjana del nombre de fills.
- Calcula la desviació típica del nombre de fills i el coeficient de variació.

3. Estadística bidimensional

■ Dependència funcional vs. correlació

En nombroses ocasions ens interessa estudiar simultàniament dos o més caràcters de la població. En particular, volem saber quina relació existeix entre aquestes variables. Aquest és l'objecte d'estudi de l'**estadística bidimensional**.

Considerem dos exemples:

- x = kg de pomes, y = preu en €

x (kg)	0	1	2	3	4	5	6	7	8	9
y (€)	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5

- x = nota matemàtiques, y = nota física d'un grup de 10 alumnes.

x (ma- tes)	3	4	5	6	7	7	7	8	9	10
y (física)	2	5	5	6	7	6	7	9	8	10

Existeixen dos tipus de relació entre variables $\{x_i, y_i\}$:

- **Dependència funcional:** Existeix una llei (una fórmula) que relaciona les dues variables $y_i = f(x_i)$. *Quan més quilos de pomes, major preu $y = 0.5 \cdot x$.* En aquest cas, el preu per quilogram és 0,5 €/kg.
- **Correlació estadística:** No existeix cap llei exacta sinó que vindrà donada per una tendència. Per exemple, *si un alumne és bo en matemàtiques, cal esperar que també ho sigui en física*, però, no existeix cap fórmula a partir de la qual es pugui predir nota que obtindrà l'alumne.

Així mateix, la correlació estadística la podem classificar segons el tipus (positiva-negativa) o segons la intensitat (forta-feble):

- **Correlació positiva:** Quan x_i augmenta, també ho fa y_i . Per exemple, *les notes de l'examen de física i les notes de l'examen de matemàtiques*.
- **Correlació negativa:** Quan x_i augmenta, y_i disminueix. Per exemple, *a major distància de la cistella de basquet, menor el nombre d'encerts*.

Cadascuna d'elles pot ésser correlació **forta** o correlació **feble**. Començarem l'estudi mesurant de forma qualitativa aquesta correlació. A mesura que avancem en el tema, però, aprendrem com descriure-la de forma quantitativa.

EXERCICI RESOLT 4

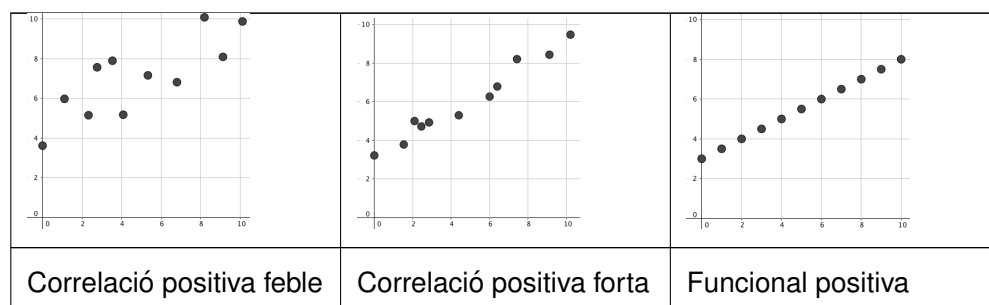
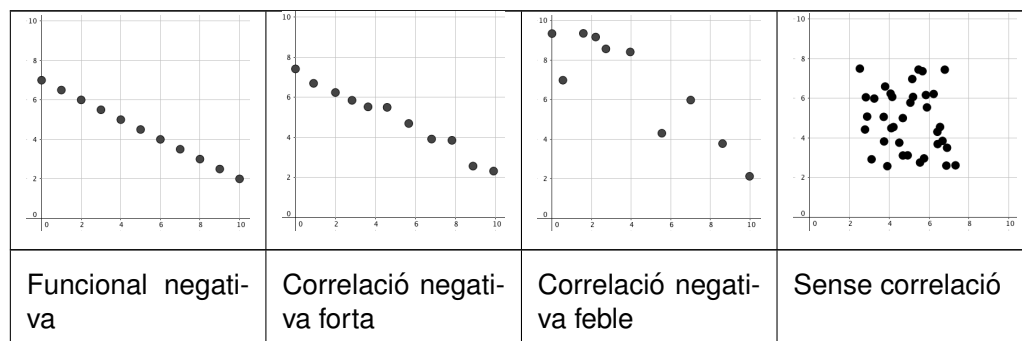
Per a cadascun dels casos següents analitza quin és el tipus de relació entre les variables (funcional o correlació). En cas de correlació indica si és positiva o negativa.

1. El radi d'una esfera – el costat d'aquesta
2. El nombre d'encerts d'un jugador de basquet – La distància a la cistella.
3. Les notes de l'examen de matemàtiques – Notes de l'examen de física.
4. La distància d'un trajecte en tren – El preu el bitllet.
5. El pes dels alumnes de 1r de batxillerat – La seva altura.
6. El nombre de membres de la família – El preu del rebut d'aigua mensual.

1. Funcional $D = 2R$, 2. Correlació negativa, 3. Correlació positiva, 4. Correlació positiva, 5. Correlació positiva, 6. Correlació positiva

3.1 Núvol de punts

Un núvol de punts s'obté de dibuixar els punts corresponents als parells (x_i, y_i) . Aquest núvol ens permet identificar visualment el tipus de relació entre les variables:



Com més concentrats estiguin els punts al voltant d'una línia recta diem que major és la correlació. Aquesta línia s'anomena **recta de regressió**. Si aquesta recta és creixent es diu que la correlació és positiva, mentre que si és decreixent la correlació és negativa.

El **centre de gravetat** del núvol és a les mitjanes de les variables $G(\bar{x}, \bar{y})$. Es compleix que la recta de regressió passa sempre pel centre de gravetat.

EXERCICI RESOLT 5

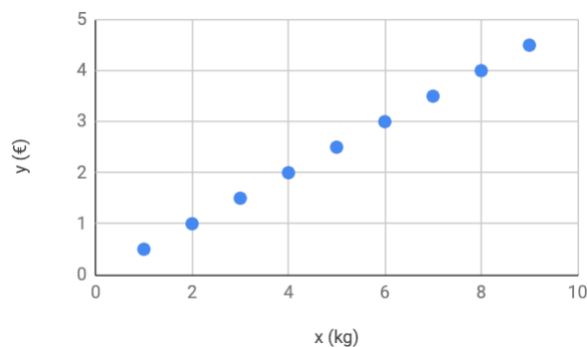
Representeu els núvols de punts per a les dades dels dos exemples de l'apartat anterior.

a) x =kg de pomes, y =preu en €

b) x =nota matemàtiques, y =nota física d'un grup de 10 alumnes.

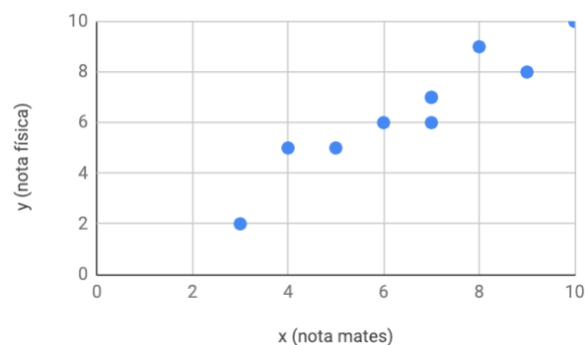
A partir del gràfic, interpretau quin tipus de relació es dona.

a) x =kg de pomes, y =preu en €



Els punts estan exactament sobre una línia recta creixent, aleshores es tracta d'una relació funcional positiva.

b) x =nota matemàtiques, y =nota física d'un grup de 10 alumnes.



Els punts estan bastant concentrats al voltant d'una recta creixent, aleshores existeix correlació positiva forta.

EXERCICIS PROPOSATS

3. S'ha fet un estudi sobre com afecta el nombre de cigarrets consumits en un dia a l'índex de mortalitat. S'han obtingut aquests resultats:

Núm. cigarrets	3	6	8	20	25
Índex de mortalitat	0,2	0,4	0,5	1,2	1,7

- a) Representau un núvol de punts.
b) Indicaue el tipus de correlació.

4. En l'estudi estadístic bidimensional s'han recollit les següents dades:

x_i	0	4	6	2	4	8	0	2	6	8
y_i	1	1	1	3	3	3	5	5	5	5

- a) Representau un núvol de punts.
b) Indicaue el tipus de correlació.

3.2 Covariància. Coeficient de correlació

A l'apartat anterior, el núvol de punts ens proporciona una forma de classificar la correlació entre les variables $\{x_i, y_i\}$ de forma qualitativa.

Anomenam **covariància** al paràmetre estadístic que relaciona **quantitativament** la relació entre les variables:

$$\sigma_{xy} = \frac{\sum_i x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y} \quad (4)$$

on N és el nombre de punts (x_i, y_i) i \bar{x} , \bar{y} les mitjanes de cada variable.

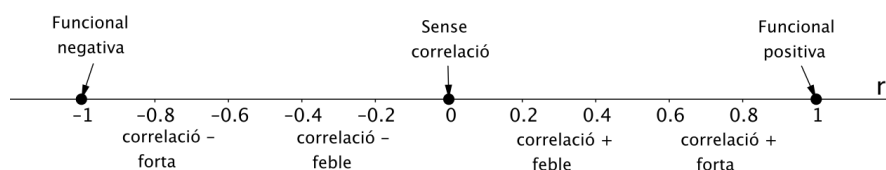
La covariància pot ésser negativa, zero o positiva segons els tipus de correlació que existeixi. La covariància té l'inconvenient, però, que depèn de l'escala de les dades que utilitzem, per la qual cosa es fa difícil tenir una idea de quan forta o feble és la relació.

Per poder quantificar millor la intensitat de la correlació empram el **coeficient de correlació lineal de Pearson** definit mitjançant

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \quad (5)$$

σ_{xy} : covariància, σ_x : desviació típica de la variable x_i , σ_y : desviació típica de la variable y_i

Aquest coeficient té el mateix signe que la covariància i, per tant, ens indicarà si es tracta d'una correlació positiva o negativa. A més a més, compleix que els seus valors sempre es troben entre $-1 \leq r \leq 1$.

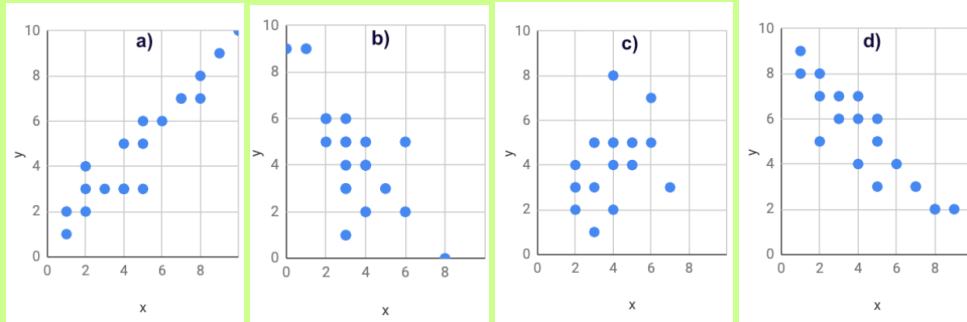


Vegem quin tipus de dependència existeix entre les variables x_i i y_i segons els valors de r .

- Si $r = -1$, tots els valors de la variable bidimensional (x_i, y_i) es troben situats sobre una recta decreixent; consegüentment, satisfan l'equació d'una recta. Llavors x_i, y_i presenten una **dependència funcional negativa**.
- Si $-1 < r < 0$, la **correlació és negativa** i serà més forta com més proper a -1 sigui el valor de r i més feble com més s'aproximi a zero.
- Si $r = 0$, llavors no existeix cap tipus de relació entre les variables, diem que estan **descorrelacionades**.
- Si $0 < r < 1$, la **correlació és positiva** i serà més forta com més proper a 1 sigui el valor de r i més feble com més s'aproximi a zero.
- Si $r = 1$, tots els valors de la variable bidimensional (x_i, y_i) es troben situats sobre una recta creixent; consegüentment, satisfan l'equació d'una recta. Llavors x_i, y_i presenten una **dependència funcional positiva**.

EXERCICI RESOLT 6

Ordenau de menor a major segons el valor del coeficient de correlació lineal de les següents distribucions bidimensionals. Assigna el valor correcte de r -0.92, -0.75, 0.34, 0.94 a cada apartat.



a) $r = 0.94$, b) $r = -0.75$, c) $r = 0.34$, d) $r = -0.92$.

De menor a major valor de r

d) < b) < c) < a)

EXERCICI RESOLT 7

Calculeu la covariància i el coeficient de correlació lineal dels dos exemples de l'apartat anterior.

a) x =kg de pomes, y =preu en €

b) x =nota matemàtiques, y =nota física d'un grup de 10 alumnes.

a) x =kg de pomes, y =preu en €

	x	y	x^2	y^2	$x \cdot y$
	0	0	0	0	0
	1	0.5	1	0.25	0.5
	2	1	4	1	2
	3	1.5	9	2.25	4.5
	4	2	16	4	8
	5	2.5	25	6.25	12.5
	6	3	36	9	18
	7	3.5	49	12.5	24.5
	8	4	64	16	32
	9	4.5	81	20.25	40.5
<i>Sumes</i>	45	22.5	285	71.25	142.5

(6)

Número de parelles $N = 10$

Mitjana x : $\bar{x} = \frac{45}{10} = 4.5$

Variància x : $Var_x = \frac{285}{10} - 4.5^2 = 8.25$

Desviació típica x : $\sigma_x = \sqrt{Var_x} = 2.87$

Mitjana y : $\bar{y} = \frac{22.5}{10} = 2.25$

Variància y : $Var_y = \frac{71.25}{10} - 2.25^2 = 2.06$

Desviació típica y : $\sigma_y = \sqrt{Var_y} = 1.436$

Covariància: $\sigma_{xy} = \frac{142.5}{10} - 4.5 \cdot 2.25 = 4.125$

Coeficient de correlació lineal: $r = \frac{4.125}{2.87 \cdot 1.436} = 1$

Donat que $r = 1$, tenim una relació funcional positiva.

b) x =nota matemàtiques, y =nota física d'un grup de 10 alumnes.

	x	y	x^2	y^2	$x \cdot y$
	3	2	9	4	6
	4	5	16	25	20
	5	5	25	25	25
	6	6	36	36	36
	7	7	49	49	49
	7	6	49	36	42
	7	7	49	49	49
	8	9	64	81	72
	9	8	81	64	72
	10	10	100	100	100
<i>Sumes</i>	66	65	478	469	471

(7)

Número de parelles $N = 10$

Mitjana x : $\bar{x} = \frac{66}{10} = 6.6$

Variància x : $Var_x = \frac{478}{10} - 6.6^2 = 4.24$

Desviació típica x : $\sigma_x = \sqrt{Var_x} = 2.06$

Mitjana y : $\bar{y} = \frac{65}{10} = 6.5$

Variància y : $Var_y = \frac{469}{10} - 6.5^2 = 4.65$

Desviació típica y : $\sigma_y = \sqrt{Var_y} = 2.156$

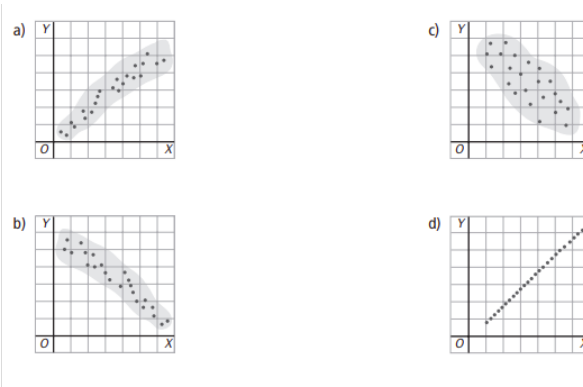
Covariància: $\sigma_{xy} = \frac{471}{10} - 6.6 \cdot 6.5 = 4.2$

Coefficient de correlació lineal: $r = \frac{4.2}{2.06 \cdot 2.156} = 0.946$

Donat que $r = 0.946$ és positiu i molt proper a 1, tenim una correlació positiva forta.

EXERCICIS PROPOSATS

5. Associa cada núvol de punts amb el seu coeficient de correlació: $r = 1$; $r = 0,92$; $r = -0,25$; $r = -0,78$;



6. En la següent variable bidimensional:

X	1	2	3	4	3	7	6	3	4	5
Y	45	30	30	25	25	10	20	15	10	15

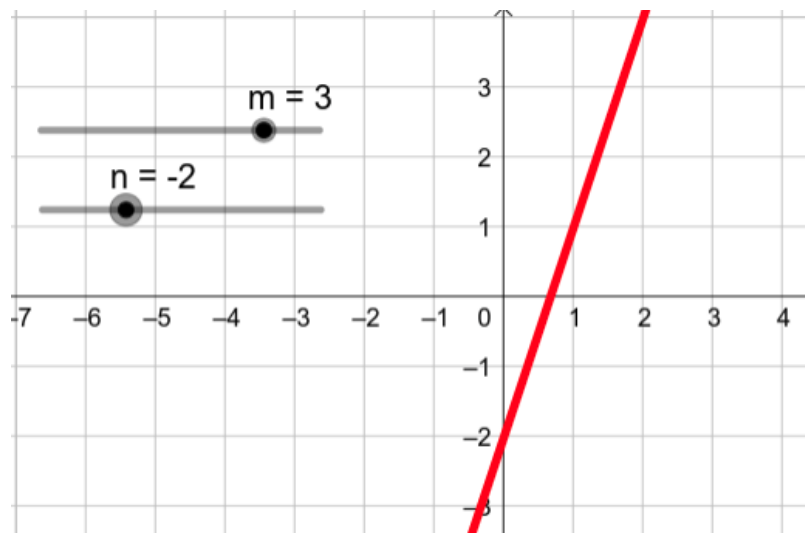
- Representa un núvol de punts. Indica el tipus de correlació.
- Calcula el centre de masses (\bar{X} , \bar{Y}) i la covariància.
- Determina el coeficient de correlació lineal.


3.3 Repàs de la funció lineal

En aquest apartat farem un petit repàs de l'equació de la recta que haurem d'utilitzar per obtenir la recta de regressió.

L'expressió d'una funció lineal és $y = mx + n$, essent m el pendent i n l'ordenada a l'origen (el punt de tall de la recta amb l'eix Y).

- Si $m < 0$: La recta és decreixent
- Si $m = 0$: es diu que la funció és constant i la gràfica és una recta horitzontal
- Si $m > 0$: La recta és creixent



 Simulació 1: <https://www.geogebra.org/m/GnGvayd4> : La funció lineal $y=mx+n$. Canvia els paràmetres m i n i observa com canvia la recta.

Alternativament, si ens donen un punt (x_0, y_0) i el pendent m podem escriure directament l'equació punt-pendent:

$$y - y_0 = m(x - x_0) \quad (8)$$

EXERCICI RESOLT 8

Calcula l'equació de la recta que passa pel punt $(5,3)$ i té pendent $m = -\frac{2}{3}$. Representa-la gràficament.

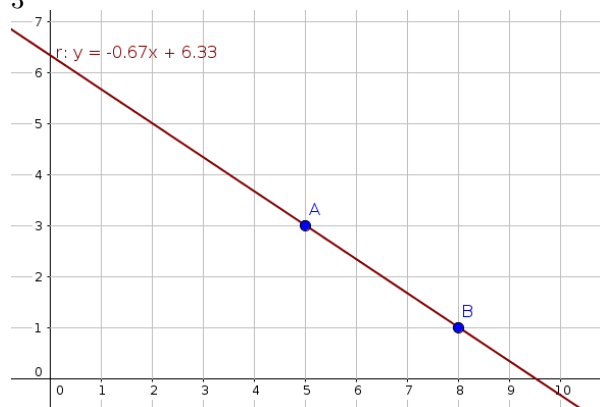
Escrivim l'equació punt-pendent $y - 3 = -\frac{2}{3}(x - 5)$

Eliminam els parèntesi i passam tot al membre de la dreta: $y = 3 - \frac{2}{3}x + \frac{2}{3}5$

Obtenim l'equació de la recta: $y = -\frac{2}{3}x + \frac{19}{3}$

La recta talla a l'eix Y en el punt $n = \frac{19}{3}$

Per representar una recta basta amb dos punts, un dels quals ja el tenim $(5,3)$. Si prenem $x = 8$, obtenim $y = -\frac{2}{3} \cdot 8 + \frac{19}{3} = 1$; llavors l'altre punt és $(8,1)$

**EXERCICIS PROPOSATS**

7. a) Calculeu la recta que passa pels punt $x = 3, y = -1$ i que té pendent $m = 4$.
b) Trobeu dos punts més d'aquesta recta.

3.4 Recta de regressió lineal

Suposem que tenim un diagrama núvol de punts com el que mostra la figura; ara intentem construir una línia recta que s'aproximi tant com es pugui al núvol de punts.

De totes les rectes possibles, sembla que n'existeix una que compleix la condició. Aquesta recta

l'anomenem **recta de regressió lineal**.

Ara bé, el mètode per aconseguir aquesta recta no és fàcil; com a primera aproximació es pot traçar a ull. Però, si necessitam més precisió, caldrà utilitzar un mètode analític.

■ **Mètode dels mínims quadrats**

Volem trobar la recta $y = mx + n$ que millor ajusta al núvol de punts. Existeixen molts de mètodes per determinar els valors de m i n . El més utilitzat es coneix com el mètode del mínims quadrats. Consisteix en fer mínima la suma dels quadrats de les diferències entre els valors experimentals i teòrics obtinguts amb l'equació. Els càlculs per l'obtenció de la recta queden fora de l'abast d'aquests apunts i simplement ens limitarem a donar el resultat.

Si s'aplica aquest mètode, s'obté que la recta de regressió passa pel **centre de gravetat** del núvol de punts (\bar{x}, \bar{y}) . Llavors, la seva equació ha d'ésser de la forma:

$$y - \bar{y} = m(x - \bar{x}) \quad (9)$$

El pendent de la recta és $m = \frac{\sigma_{xy}}{\sigma_x^2}$.

Aquesta es coneix com la recta de regressió de y sobre x . Amb aquesta recta podem fer prediccions del valor de y conegut el valor de x . Si, en canvi, ens donen una y com a dada, es pot estimar la x aïllant-la de l'equació. Tot seguit farem exemples de com es fan les estimacions basant-nos en la recta de regressió.

■ **La recta de regressió per fer estimacions**

La recta de regressió s'utilitza per fer prediccions d'una variable coneguda l'altra. Ara bé, quina fiabilitat podem concedir als càlculs obtinguts a través de les rectes de regressió? La predicció serà més bona o fiable com més s'apropi a ± 1 el coeficient de correlació r . Així doncs:

- Si r és proper a 0, no té sentit fer cap mena d'estimacions o prediccions.
- Si r és pròxim a -1 o $+1$, probablement els valors reals seran pròxims a les estimacions que feim.
- Si r és proper 1 o -1 , les estimacions fetes coincidiran amb els valors reals.

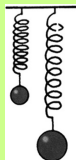
Existeixen dos tipus de prediccions: **interpolacions** i **extrapolacions**. En les interpolacions, es vol determinar un valor que es troba dins del rang de les dades. En les extrapolacions, en canvi, cau fora del rang de dades. Independentment del grau de correlació, cal tenir en compte que la validesa d'una extrapolació només serà certa si no ens allunyam massa del rang de dades.

EXERCICI RESOLT 9

La llei de Hooke relaciona l'allargament d'una molla amb la força que hi aplicam $F = kx$ on k és la constant d'elàstica. Hem anat penjat a la molla pesos de diferent valors i hem anotat l'allargament

Allargament x (m)	0	0.06	0.2	0.3	0.6
Pes F (N)	0	0.45	1.1	1.47	2.95

(10)



- a) Calcula la recta de regressió i estima el valor de la constant elàstica de la molla.
b) Estima el pes necessari per produir un allargament de 0.5 m.

Començarem calculant els paràmetres de cada distribució per separat així com la covariància i el coeficient de correlació lineal.

	x	y	x^2	y^2	$x \cdot y$
	0	0	0	0	0
	0.06	0.45	0.0036	0.202	0.027
	0.2	1.1	0.04	1.21	0.22
	0.3	1.47	0.09	2.16	0.441
	0.6	2.95	0.36	8.7	1.77
Sumes	1.16	5.97	0.494	12.28	2.46

(11)

Número de parelles $N = 5$

Mitjana x : $\bar{x} = \frac{1.16}{5} = 0.232$

Variància x : $Var_x = \frac{0.4936}{5} - 0.232^2 = 0.0449$

Desviació típica x : $\sigma_x = \sqrt{Var_x} = 0.212$

Mitjana y : $\bar{y} = \frac{5.97}{5} = 1.194$

Variància y : $Var_y = \frac{12.28}{5} - 1.194^2 = 1.103$

Desviació típica y : $\sigma_y = \sqrt{Var_y} = 1.015$

Covariància: $\sigma_{xy} = \frac{2.46}{5} - 0.232 \cdot 1.194 = 0.2146$

Coeficient de correlació lineal: $r = \frac{0.2146}{0.212 \cdot 1.015} = 0.998$

Donat que $r \approx 1$, tenim una correlació forta positiva.

El pendent de la recta y sobre x és

$$m = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{0.2146}{0.212^2} = 4.78 \quad (12)$$

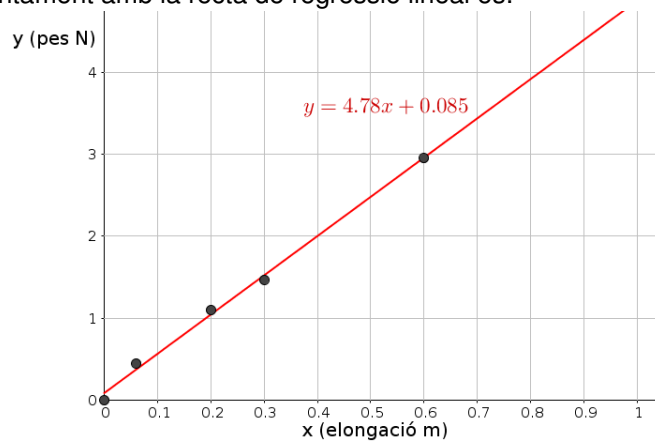
Podem aproximar la constant de la molla a $k = 4.78 \text{ N/m}$.
L'equació de la recta de regressió és

$$y - 1.194 = 4.78(x - 0.232) \quad (13)$$

que podem simplificar com $y = 4.78x + 0.085$.

b) L'estimació s'obté a partir de la recta de regressió $y = 4.78 \cdot 0.5 + 0.085 = 2.48 \text{ N}$. Aquesta predicció és bastant fiable ja que $r \approx 1$.

El núvol de punts juntament amb la recta de regressió lineal és:



del qual també es desprèn una correlació forta positiva.

EXERCICI RESOLT 10

Es determina la pèrdua d'efectivitat d'un medicament al llarg del temps i s'obtenen els resultats següents

Temps x (mesos)	1	2	3	4	5
% d'efectivitat restant	90	78	42	35	21

(14)

- a) Quin tant per cent d'efectivitat quedarà al cap de 6 mesos?
 b) Quan de temps haurà de passar perquè quedi el 50% d'efectivitat restant?

Començarem calculant els paràmetres de cada distribució per separat així com la covariància i el coeficient de correlació lineal.

	x	y	x^2	y^2	$x \cdot y$
	1	90	1	8100	90
	2	78	4	6084	156
	3	42	9	1764	126
	4	35	16	1225	140
	5	21	25	441	105
<i>Sumes</i>	15	266	55	17614	617

(15)

Número de parelles $N = 5$

Mitjana x : $\bar{x} = \frac{15}{5} = 3$

Variància x : $Var_x = \frac{55}{5} - 3^2 = 2$

Desviació típica x : $\sigma_x = \sqrt{Var_x} = 1.41$

Mitjana y : $\bar{y} = \frac{266}{5} = 53.2$

Variància y : $Var_y = \frac{17614}{5} - 53.2^2 = 692.56$

Desviació típica y : $\sigma_y = \sqrt{Var_y} = 26.32$

Covariància: $\sigma_{xy} = \frac{617}{5} - 3 \cdot 53.2 = -36.2$

Coeficient de correlació lineal: $r = \frac{-36.2}{1.41 \cdot 26.32} = -0.972$

Donat que $r \approx -1$, tenim una correlació forta negativa.

El pendent de la recta y sobre x és

$$m = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-36.2}{1.41^2} = -18.1 \quad (16)$$

L'equació de la recta de regressió és $y - 53.2 = -18.1(x - 3)$ que podem simplificar com $y = -18.1x + 107.5$.

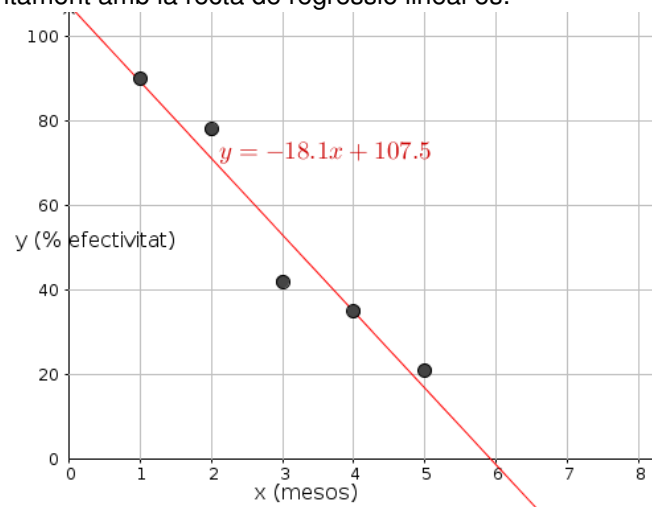
Les prediccions que ens demanen seran

a) $y = -18.1 \cdot 6 + 107.5 = -1.1 \%$. En aquest cas l'extrapolació fa que obtinguem un percentatge negatiu cosa que no té sentit. Això indica que el medicament haurà deixat de fer efecte passats 6 mesos.

b) Per fer l'altra predicció, substituïm el valor de y dins l'equació de la recta $50 = -18.1 \cdot x + 107.5$ i aïllem $x = \frac{50 - 107.5}{-18.1} = 3.17$ mesos.

Les dues prediccions són bastant fiables ja que $r \approx -1$.

El núvol de punts juntament amb la recta de regressió lineal és:



EXERCICIS PROPOSATS

8. Considereu les dades de l'exercici proposat 6. Es demana que:
- Calculeu la recta de regressió.
 - Estimeu el valor de y quan $x = 5$
 - Estimeu el valor de x quan $y = 100$
 - Justifiqueu la validesa de les prediccions b) i c).