

UNITAT 4: ESTADÍSTICA DESCRIPTIVA

L'estadística és la ciència que s'ocupa de l'ordenació i anàlisi de dades procedents de mostres, i de la realització d'inferències sobre les poblacions de les que hem obtingut les dades.

Les dues grans funcions de l'Estadística (descriptiva de dades i realització d'inferències) són el reflex de la pròpia història d'aquesta ciència. L'Estadística es el producte de la trobada entre l'estadística antiga i el càlcul de probabilitats del segle XIX. Etimològicament la paraula prové d'estat. Des dels temps dels egipcis i dels romans ja es feia servir per saber el nombre d'habitants i altres qüestions econòmiques relatives a les nacions. La recol·lecció de dades per a un posterior anàlisi del tema estudiat, això és el que entenem per estadística descriptiva; la missió d'aquesta és descriure situacions i processar informació, fent servir taules, gràfics,...

Tanmateix les conclusions extreïdes s'esgoten en el mateix conjunt de dades observades, doncs l'objectiu consistia en fer-se una idea clara del que hi havia, i el que hi havia es contava i es mesurava. No va ser fins el segle XIX que gràcies a **Karl C. Pearson** (1857-1936) es va desenvolupar l'estadística analítica o inferencial basada en la teoria de probabilitats, que tracta d'obtenir lleis generals a partir de l'observació de les dades.

4.1. Conceptes previs

En iniciar un estudi estadístic cal primer definir el conjunt d'elements (**població**) sobre el qual farem l'estudi i quina característica o propietat volem estudiar (**caràcter estadístic**). A continuació com recollim les dades i si la fem sobre la població o solament sobre una part d'ella (**mostra**). Aquests són els termes bàsics.

Població: conjunt d'elements dels quals volem estudiar alguna característica determinada.

Mostra: subconjunt de la població del qual extraurem les dades per a la realització de l'estudi. La principal característica d'una bona mostra és que sigui representativa de la població.

Caràcter estadístic: característica que volem estudiar d'una població. Els caràcters estadístics poden classificar-se en qualitius i quantitius. Si el caràcter no es pot mesurar numèricament és **qualitatiu**. Si el caràcter es pot mesurar numèricament és **quantitatiu** o variable estadística.

Les variables estadístiques es classifiquen en discretes i contínues.

Variables discretes quan només poden prendre valors aïllats, com per exemple les pàgines d'un llibre.

Variables contínues quan poden prendre qualsevol valor dins d'un interval, per exemple el pes d'un nouat.

A1 Volem saber a les properes eleccions a l'Ajuntament de Barcelona quina serà la distribució dels edils per partits. Com hem de fer la mostra? Dissenya com faries l'estudi fent servir els conceptes previs.

4.2 Freqüència absoluta i relativa

Freqüència absoluta d'un valor x és el nombre de vegades que es repeteix aquest. La representarem per f_i .

Freqüència relativa d'un valor x és el quocient entre la freqüència absoluta i el nombre de dades (N). La representarem h_i .

$$h_i = \frac{f_i}{N}$$

Freqüència absoluta acumulada d'un valor x és el nombre de valors que són més petits o iguals que ell. La representarem per F_i .

$$F_i = f_1 + f_2 + f_3 + \dots + f_i = \sum f_i$$

Freqüència relativa acumulada d'un valor x és el quocient entre la freqüència absoluta acumulada i el nombre de dades (N). La representarem H_i .

$$H_i = \frac{\sum f_i}{N} = \sum h_i$$

Aquesta informació es recull i s'organitza en una taula estadística. Quan el nombre de valors que pot prendre la variable estadística és molt gran, aquests valors s'agrupen en intervals anomenats **classes**. A l'hora de fer càlculs prendrem com a representant de cada interval el valor mig que anomenem **marca de classe**.

A2 Construeix la taula de freqüències. Les notes de l'últim examen de matemàtiques en un grup de 20 alumnes van ser:

2,8,9,8,7,1,7,6,7,10,4,4,4,1,8,9,8,4,2,6

A3 Construeix la taula de freqüències corresponent a l'estudi del nombre de germans dels alumnes d'una escola.

Nombre de germans	f_i
0	47
1	102
2	36
3	10
4	4
5	1

A4 Construeix la taula de freqüències corresponent a les alçades dels jugadors de bàsquet en una lliga.

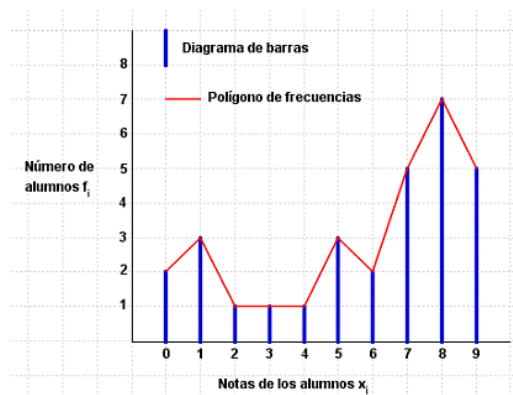
Classes alçada (cm)	f_i
[165,175)	9
[175,185)	34
[185,195)	33
[195,205)	19
[205,215)	5

4.3. Representacions gràfiques

Diagrames de barres i polígon de freqüències

S'utilitzen en variables de tipus discret. L'eix d'abscisses es situen els diferents valors de la variable. L'eix d'ordenades les freqüències relatives, absolutes acumulades o no. L'alçada de les barres és la freqüència.

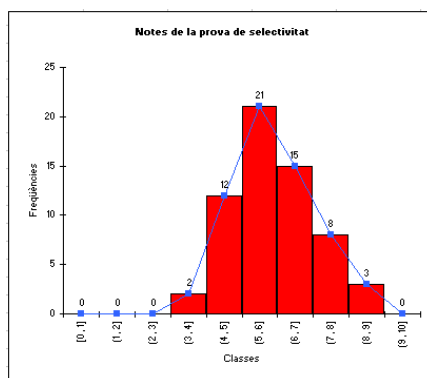
Si unim els punts superiors de les barres per segments obtenim un polígon de freqüències.



Histogrames

S'utilitzen en distribucions amb dades agrupades en intervals.

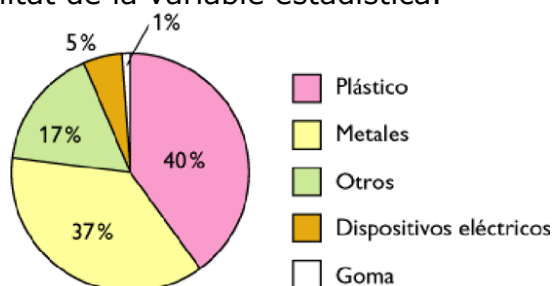
A l'eix d'abscisses es situen els diferents intervals i a l'eix d'ordenades les freqüències. Es dibuixa un rectangle per a cada classe, de base l'amplitud del interval. L'àrea dels rectangles ha de ser proporcional a la freqüència. Si tots els intervals són d'igual amplitud l'altura correspondrà a la freqüència de la classe.



Diagrames de sectors

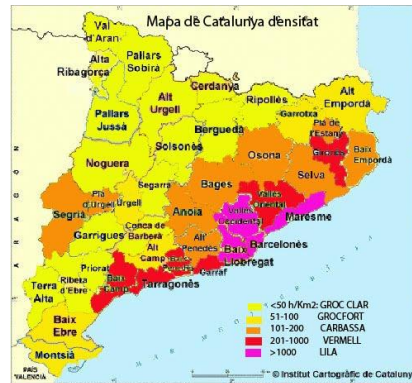
S'utilitzen bàsicament en distribucions de variable qualitativa.

Es construeixen dibuixant sobre un cercle un sector d'àrea proporcional a la freqüència de cada modalitat de la variable estadística.



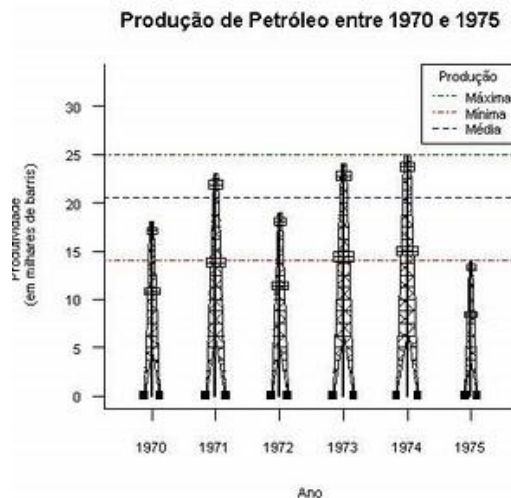
Cartogrames

Aquest gràfics es fan sobre mapes. Mitjançant colors, símbols o dibuixos s'indica la distribució d'un caràcter estadístic.



Pictogrames

Una altre forma de representar les dades estadístiques és el pictograma. És un gràfic en el que fem servir figures al·lusives al fenomen que estudiem. Si la variable és la producció d'ordinadors, el dibuix seria un ordinador.



4.4. Paràmetres de centralització i dispersió

Solament podem calcular aquestes mesures quan les variables són quantitatives. Els paràmetres són de dos tipus, centralització quan expliquen al voltant de quins valors se situen les dades en una distribució estadística i dispersió quan me indiquen si les dades estan agrupades o disperses.

4.4.1. Paràmetres de centralització

Mitjana aritmètica

S'anomena mitjana aritmètica el quocient entre la suma de tots els valors de la variable i el nombre total de dades. Es representa per $\bar{x}$.

$$\bar{x} = \frac{x_1 \cdot f_1 + x_2 \cdot f_2 + \dots + x_n \cdot f_n}{N} = \frac{\sum x_i \cdot f_i}{N}$$

Mediana

S'anomena mediana d'una distribució estadística el valor de la variable que, ordenades les dades de més petit a més gran, ocupa la posició central, deixant per sota i per sobre el mateix nombre de dades. Es representa per M .

Si el nombre de dades es senar, la mediana un cop ordenades les dades és el valor central, de 7 dades seria la quarta.

Si el nombre de dades es parell, no hi ha un únic valor central sinó dos; la mediana és la mitjana aritmètica entre els dos valors centrals.

Si les dades estan agrupades en classes aplicarem la següent fórmula:

$$M = Li + c \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$$

- Li : límit inferior de la classe de la mediana
- c : amplitud de la classe de la mediana.
- F_{i-1} : és la freqüència absoluta de la classe anterior de la mediana.
- f_i : es la freqüència absoluta de la classe de la mediana.

Moda

S'anomena moda d'una distribució estadística el valor de la variable que té freqüència absoluta més alta. Es representa per M_o .

Si aquest valor no és únic, tindrem dues modes o tres etc. Si tots els valors de la variable tenen la mateixa freqüència, la distribució no té moda.

Quantils

Són els valors que divideixen el conjunt de dades en parts que tenen la mateixa freqüència. Les dades han d'estar agrupades en classes. Els més típics són els **quantils**, les dades es distribueixen en quartes parts i els **percentils**, el conjunt de dades es reparteix en 100 parts.

Si les dades estan agrupades en classes aplicarem les següents fórmules:

Quantil 1	Q_1	$Q_1 = Li + c \cdot \frac{\frac{N}{4} - F_{i-1}}{f_i}$
Quantil 3	Q_3	$Q_3 = Li + c \cdot \frac{\frac{3N}{4} - F_{i-1}}{f_i}$
Percentil 65	P_{65}	$P_{65} = Li + c \cdot \frac{\frac{65N}{100} - F_{i-1}}{f_i}$

on:

- Li : límit inferior de la classe del quantil o percentil
- c : amplitud de la classe del quantil o percentil.
- F_{i-1} : és la freqüència absoluta de la classe anterior al quantil o percentil.
- f_i : es la freqüència absoluta de la classe del quantil o percentil.

A5 Calcula la mitjana aritmètica, la mediana i la moda de l'activitat 2.

A6 Calcula la mitjana aritmètica, la mediana i la moda de l'activitat 3.

A7 Calcula la mitjana aritmètica, la mediana, la moda, els quantils i el percentil 32 de l'activitat 4.

4.4.2. Paràmetres de dispersió

Si tenim dues classes de 5 alumnes cada una i les notes de la classe A són 5,5,5,5,5 i les de la classe B 0,1,5,9,10, veurem que són dos distribucions que tenen la mateixa mitjana aritmètica i moda 5, en canvi, són molt diferents.

Rang

S'anomena rang o recorregut d'una distribució estadística a la diferència entre el valor de la variable més gran i més petit. S'anomena rang interquartílic a la diferència entre el tercer i el primer quartil.

$$Q = Q_3 - Q_1$$

Variància

S'anomena variància d'una distribució estadística a la mitjana aritmètica dels quadrats de les desviacions de cada valor. Es representa per s^2 .

$$s_x^2 = \frac{f_1 \cdot (x_1 - \bar{x}) + \dots + f_n \cdot (x_n - \bar{x})}{N} = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N}$$

Desviació típica

S'anomena desviació típica o desviació estàndard a la arrel quadrada positiva de la variància. Es representa per s .

$$s_x = \sqrt{\frac{f_1 \cdot (x_1 - \bar{x}) + \dots + f_n \cdot (x_n - \bar{x})}{N}} = \sqrt{\frac{\sum f_i \cdot (x_i - \bar{x})^2}{N}}$$

A8 Calcula el rang i la desviació típica de l'activitat 2.

A9 Calcula el rang i la desviació típica de l'activitat 3.

A10 Calcula el rang i la desviació típica de l'activitat 4.

A11 a) Construeix el histograma de freqüències absolutes acumulades que correspon a aquesta taula:

Intervals	n
10 < x < 15	3
15 < x < 20	7
20 < x < 25	16
25 < x < 30	12
30 < x < 35	9
35 < x < 40	5
40 < x < 45	2

b) Calcula'n la mitjana, la mediana, la moda, la desviació típica i els quartils.

A12 En un total de 50 famílies es va estudiar la variable x=nombre de fills. En la taula següent es resumeixen els resultats obtinguts:

Número de fills	0	1	2	3	4	5
Freqüència absoluta		15		5	4	
Freqüència relativa	0,2					0,02

a) Completa la taula.

b) Calcula els valors dels paràmetres de centralització.

c) Calcula la desviació típica.

A13 Les edats, en anys , de tres grups de amics són:

Grup 1	20	18	18	19	17
Grup 2	31	29	30	32	26
Grup 3	42	26	35	39	46

Compara els tres grups d'edat, a partir de la seva mitjana i desviació típica.

A14 La primera prova d'unes oposicions constava d'un test puntuat de 0 a 80, s'han presentat 296 opositors, amb els resultats següents:

Puntuació	n	Puntuació	n
[0,10)	1	[40,50)	81
[10,20)	11	[50,60)	53
[20,30)	29	[60,70)	38
[30,40)	58	[70,80)	25

La normativa de la oposició diu que han de passa a la següent fase el 60% del candidats que tinguin millor puntuació.

a) Quina és la nota mínima que cal haver superat per passar a la fase següent?

b) Quin tant per cent dels opositors es troben al interval $(\bar{x}-s, \bar{x}+s)$? i al interval?

A15 Les notes obtingudes per dos grup de primer de Batxillerat en un control de matemàtiques han estat les següents:

Grup A	1, 9, 8, 2, 5, 8, 3, 5, 8, 5, 6, 0, 10, 10, 5, 3, 10, 2, 9, 1
Grup B	5, 6, 4, 7, 5, 7, 4, 3, 8, 6, 9, 0, 5, 5, 5, 6, 6, 5, 4, 10

a) Construeix les taules de freqüències dels dos grups.

b) Calcula la mitjana, la mediana i la moda del dos grups i digues que podem observar.

c) Calcula la desviació típica i amb els resultats que observes diu quines conclusions podem treure.

A16 Les notes de matemàtiques de la primera avaluació en un grup de 4t.d'ESO van ser:

Notes	n
Insuficient	8
Suficient	5
Bé	4
Notable	8
Excel·lent	5

a) Fes el histograma. Comenta-ho.

b) Calcular els paràmetres de centralització.

c) Calcula els paràmetres de dispersió.

d) Calcula els Quartils.

e) Calcula la nota que deixa per sota el 70% del alumnes? i la que deixa per sobre el 80%?

A17 Busca una estadística sobre un tema que t'interessi. Calcula tots els paràmetres i un cop tinguis aquest resultats comenta'ls.

4.5. Distribucions bidimensionals

És possible que quan fem un estudi estadístic en una població ens interessi observar una única variable estadística. No obstant això, en la majoria de casos, ens pot interessar veure la relació entre dues o més variables.

Quan observem dues variables estadístiques, cadascun dels individus tindrà associats un parell de valors. Es tracta d'estudiar la relació o dependència que pugui haver-hi entre elles, si és que aquesta existeix. L'objectiu en un estudi bidimensional és quantificar i formalitzar la relació entre les dues variables per poder fer-ne prediccions.

Alguns exemples de variables bidimensionals són:

- Extensió en km² i nombre d'habitants en els països europeus.
- El consum elèctric i el rebut que paguem a Barcelona.
- El temps que estudiem i la nota que obtenim en una matèria.
- Latitud i temperatura mitjana en les capitals del món.
- El temps jugat i el nombre de punts encistellat pels jugadors de l'ACB.

La manera més habitual d'expressar els valors obtinguts en una distribució estadística bivariant de la que tinguem moltes dades és mitjançant una taula de doble entrada o taula de contingència.

Exemple 1:

Dades de 30 famílies: la variable x és el nombre de fills, i la variable y el nombre de filles.

	0	1	2	3	4
0	2	3	1	1	2
1	4	2	2	-	1
2	3	3	1	1	-
3	1	-	2	1	-

Quan tenim poques dades normalment les donem com a parelles de valors o en una taula simple.

Exemple 2:

Dades de 12 ciutats europees on X és la temperatura mitjana (°C) i Y la latitud.

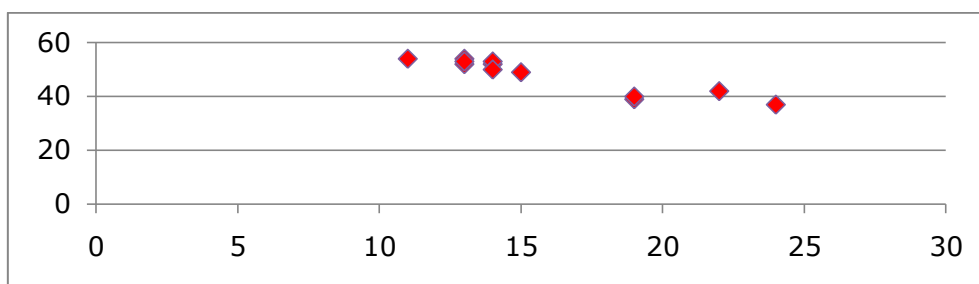
Capitals	x	y
Amsterdam	13	54
Atenes	24	37
Berlín	13	52
Brussel·les	14	52
Copenhaguen	11	54
Dublín	13	53
Lisboa	19	39
Londres	14	53
Luxemburg	14	50
Madrid	19	40
París	15	49
Roma	22	42

Exemple 3:

La informació recollida per una casa discogràfica sobre el nombre de concerts fets durant l'estiu per 15 grups musicals i les vendes de discos d'aquests grups (expressades en milers de discos).

Discos/ Concerts	[10,30)	[30,40)	[40,80)
[1,5)	3	-	-
[5,10)	1	4	1
[10,20)	-	1	5

Si representem cada parella de valors obtenim un gràfic que denominem núvol de punts. El núvol de punts que hem representat és el de l'exemple 2. L'eix de les X (temperatura mitjana en °C) i l'eix de les Y (la latitud).



4.5.1. Concepte de correlació

La **correlació** és el concepte que estudia la relació o dependència entre dues variables que intervenen en una distribució bidimensional.

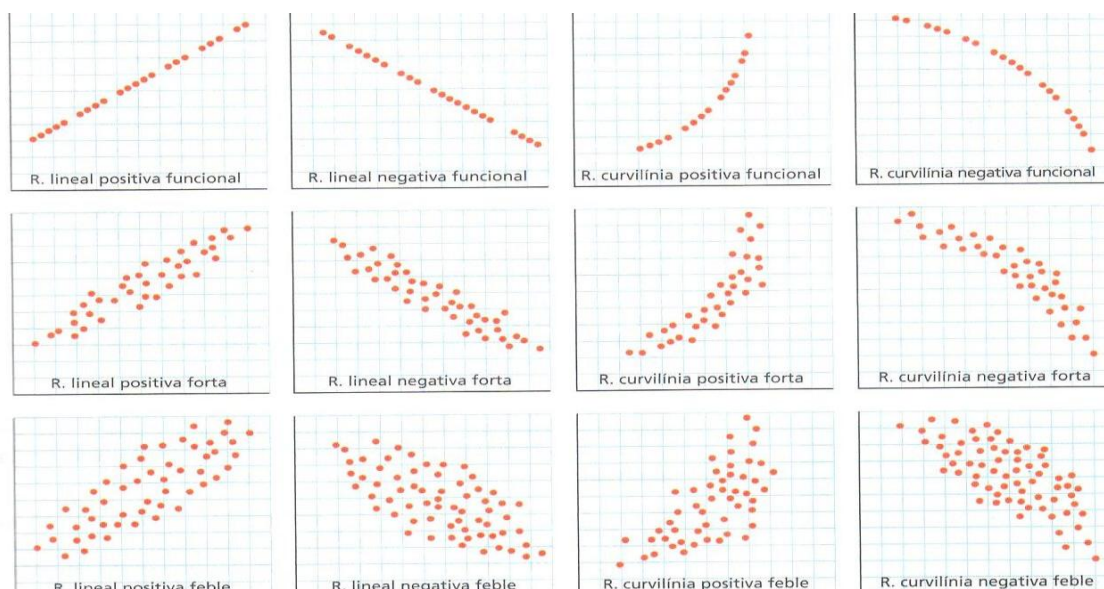
La correlació entre dues variables es pot classificar de varies formes.

1) La correlació pot ser **lineal** o **curvilínia**, si els punts del diagrama de punts es situen sobre una línia recta o una corba.

2) La correlació pot ser **positiva** (directa) o **negativa** (inversa); positiva és quan una variable creix i l'altra variable creix també. Negativa és quan una variable creix i en canvi l'altra decreix.

La correlació pot ser **nul·la** si no hi ha relació entre les dues variables. En el núvol de punts podrem observar els punts escampats a l'atzar.

3) La correlació és de tipus **funcional**, si hi ha alguna funció matemàtica que satisfà tots els punts de la distribució. En cas contrari, serà més **forta** o **feble** en funció de la dependència de les dues variables.



4.5.2. Càlcul de paràmetres

Per estudiar la correlació lineal entre dues variables calcularem una sèrie de paràmetres que ens permetran saber quin tipus de relació hi ha entre les dues variables. Aquests paràmetres són: les mitjanes aritmètiques i les desviacions típiques de les dues variables, la covariància i finalment el coeficient de correlació lineal de Pearson.

S'anomena **covariància** d'una variable bidimensional (x,y) la mitjana aritmètica dels productes de les desviacions de cada una de les variables respecte de les seves mitjanes respectives. La covariància es representa per s_{xy} .

$$s_{xy} = \frac{\sum f \cdot (x - \bar{x}) \cdot (y - \bar{y})}{N} \quad \text{o} \quad s_{xy} = \frac{\sum f \cdot (x) \cdot (y)}{N} - \bar{x} \cdot \bar{y}$$

El **coeficient de correlació lineal de Pearson** és el paràmetre que indica quin tipus de correlació lineal hi ha entre les dues variables. Es representa per r.

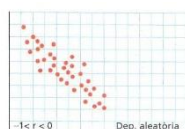
$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

El signe del coeficient r ve donat pel signe de la covariància, ja que les desviacions típiques són sempre positives. El signe de la covariància determina si la correlació és directa o inversa.

- Si la covariància és positiva la correlació és directa.
- Si la covariància és negativa, la correlació es inversa.



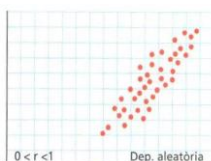
Si **r=-1** tots els valors de la variable bidimensional es troben sobre una recta de pendent negativa.
Relació funcional, inversa.



Si **-1 < r < 0** la correlació és negativa i serà més forta a mesura que s'acosti a -1 i més feble a mesura que s'acosti a 0.
Relació molt forta/forta/feble/molt feble, inversa.



Si **r=0** llavors no hi ha cap tipus de relació. Si el valor és proper a 0 es pot dir el mateix.
Relació nul·la (variables independents).



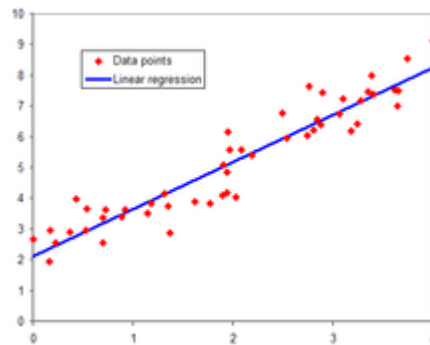
Si **0 < r < 1**, la correlació és positiva i serà més forta a mesura que s'acosti a 1 i més feble a mesura que s'acosti a 0.
Relació molt forta/forta/feble/molt feble, directa.



Si **r=1** tots els valors de la variable bidimensional es troben sobre una recta de pendent positiva.
Relació funcional, directa.

4.5.3. Regressió lineal

Si entre dues variables hi ha una correlació forta, el diagrama de punts es condensa al voltant d'una recta. Si x és la variable independent i y és la variable dependent, llavors el problema consisteix en trobar l'equació de la recta que s'ajusti millor al núvol de punts.



Aquesta equació de la recta em permet fer prediccions, però, cal ser prudent a l'hora de fer extrapolacions.

L'equació de la recta de regressió de y sobre x és:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

Quina fiabilitat podem concedir a aquests càlculs obtinguts mitjançant la recta de regressió? Com més gran sigui el coeficient de correlació lineal en valor absolut més fiabilitat. En el cas que $r=1$ o -1 , les estimacions coincidiran amb els valors reals.

A19 Les alçades en polzades de dotze pares i els seus fills són les següents:

Pares	65	63	67	64	68	62	70	66	68	67	69	71
Fills	68	66	68	65	69	66	68	65	71	67	68	70

Calcula'n el coeficient de correlació lineal i extrau les conclusions entre les dues variables estudiades.

A20 A les biblioteques de sis poblacions s'ha analitzat l'afluència de lectors (X en milers de persones) i el nombre de llibres prestats (Y).

X	0.5	1	1.3	1.7	2	2.5
Y	180	240	250	300	340	400

- Calcula i comenta el coeficient de Pearson.
- Calcula la recta de regressió. Si acudissin 1500 lectors a una biblioteca quants llibres es deixarien en préstec?

- A21** Un centre comercial sap els clients que el poden visitar en funció de la distància, en kilòmetres, a què se situï d'un nucli de població, segons les dades que figuren en la taula següent:

Nº de clients (centenars)	8	7	6	4	2	1
Distància (kilòmetres)	15	19	25	23	34	40

- Calcula'n la covariància i el coeficient de correlació.
- Dibuixa el núvol de punts.
- Si el centre se situa a 30 km, quants clients pot tenir?

- A22** A partir de la taula de valors següent:

x	2	4	5	6	8	11
y	18	12	10	8	7	5

- Calcula el coeficient de correlació.
- Multiplica cada valor de x per 2 i suma-hi 6; multiplica cada valor de y per 3 i resta-hi 15. Troba el nou coeficient de correlació. Compara i comenta els resultats.

- A23** La taula següent recull les qualificacions de quaranta alumnes en les matèries de matemàtiques i física:

Matemàtiques	3	4	5	6	6	7	7	8	10
Física	2	5	5	6	7	6	7	9	10
Nº alumnes	4	6	12	4	5	4	2	1	2

- Calcula la covariància.
- Calcula el coeficient de correlació i comenta'l.
- Quina nota de matemàtiques es pot esperar d'un alumne que treu un 7 a física?

- A24** La taula següent mostra les temperatures mitjanes en una ciutat al llarg del primer semestre:

	Gener	Febrer	Març	Abril	Maig	Juny
Tª màxima (x)	16	17	19	19	21	25
Tª mínima (y)	7	10	12	13	15	19

- Calcula el coeficient de correlació.
- Quin tipus de relació hi ha entre les dues variables?
- Determina la recta de regressió.
- Quina temperatura mínima hem d'esperar si la màxima és de 30º?