TEMA 7. ESTADÍSTICA BIDIMENSIONAL

En el tema anterior s'estudiava només una variable, però és força freqüent que sobre una mateixa població s'estudiïn dues característiques per tal de determinar si existeix alguna relació entre elles. Una **variable estadística bidimensional** resulta quan estudiem dues característiques diferents dels individus d'una població, i està formada per dues variables unidimensionals.
Si s'observa que hi ha alguna relació entre les variables observades intentarem calcular:

- el grau de relació;
- una fórmula o equació que descriguí el més fidelment possible aquesta relació.

La variable bidimensional (X, Y) queda determinada pels parells de dades (x_1, y_1), (x_2, y_2), …, (x_n, y_n).

Taules de freqüències

Per organitzar les dades es poden ordenar en **taules de doble entrada**.
Ex: S'ha fet un estudi sobre el nombre d'autobusos (X) que agafa una sèrie de persones i el temps que utilitzen en total per desplaçar-se (Y)

Y \ X	1	2	3	Total
[10,20)	14	4	1	**19**
[20,30)	10	16	3	**29**
[30,40)	14	4	6	**24**
[40,50)	7	6	5	**18**
Total	**45**	**30**	**15**	**90**

La taula ens mostra que s'han entrevistat a 90 persones de les quals 19 utilitzen entre 10 i 19 minuts per desplaçar-se; d'aquestes 19, 14 fan servir només un autobús, 4 fan servir dos i 1 utilitza 3 autobusos.
Quants usuaris utilitzen tres autobusos al dia? 15
Quants usuaris utilitzen dos autobusos i necessiten entre 30 i 40 minuts per arribar a la seva destinació? 4.

Per estudiar per separat cadascuna de les variables es fan les **taules de freqüències marginals**.
Per l'exemple anterior

X = nombre d'autobusos

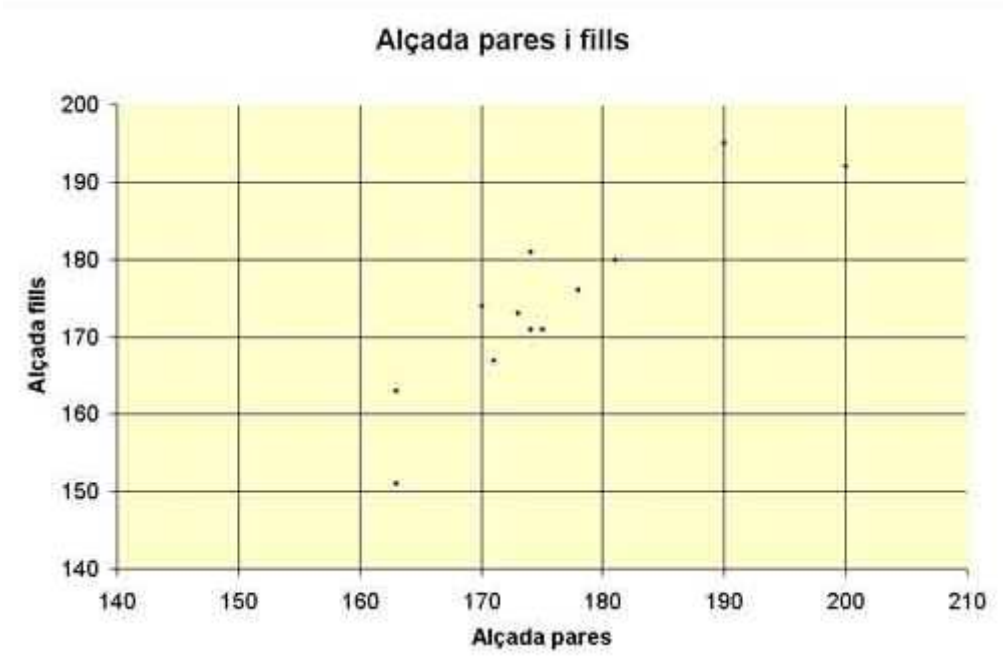
x_i	freqüència
1	45
2	30
3	15
Total	90

Y = temps de desplaçament (minuts)

y_i	freqüència
[10,20)	19
[20,30)	29
[30,40)	24
[40,50)	18
Total	90

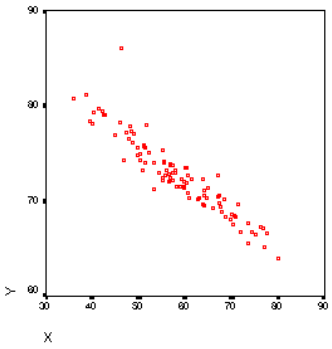
Diagrames de dispersió
Gràficament, les dades bivariants donen lloc a un diagrama de dispersió o núvol de punts. Cada individu és representat per un punt $\mathbf{P_i}$, les coordenades del qual corresponen als valors que hi adquireixen les dues variables.
Ex:

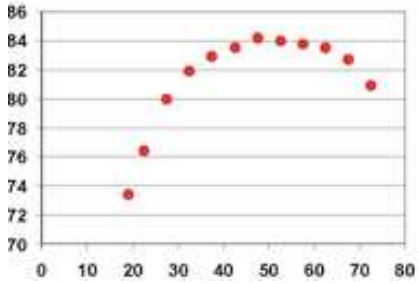
pares	178	170	163	181	174	190	175	171	200	174	173	163
fills	176	174	163	180	171	195	171	167	192	181	173	151



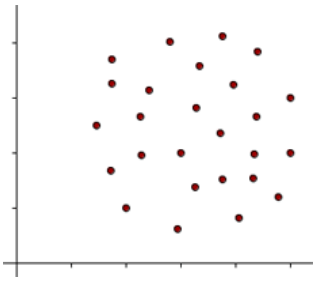
El gràfic resultant pot presentar diversos aspectes:

- Els punts es situen regularment damunt la trajectòria d'una línia de forma senzilla: recta, paràbola, funció exponencial, ...

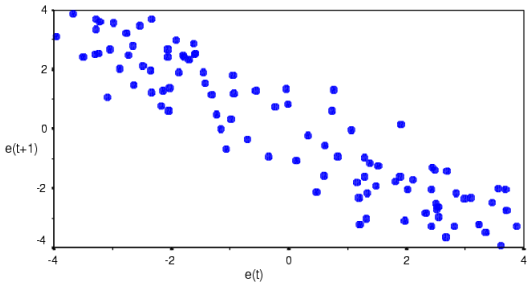




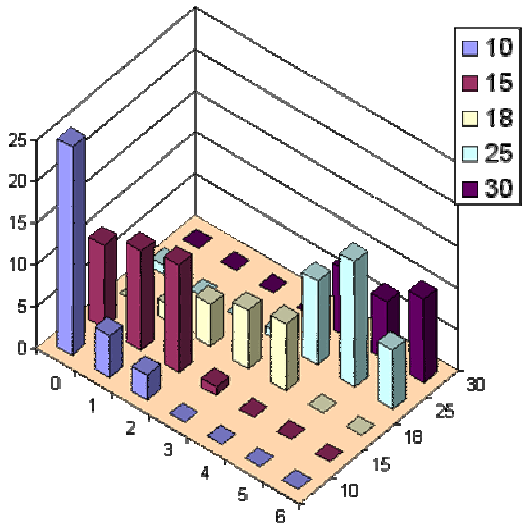
- Els punts es disposen en forma de *núvol*, de manera que no s'hi pot endevinar cap mena de regularitat



- I, entre aquests dos extrems, poden presentar-se tota mena de situacions intermèdies, en què els punts adoptin *més o menys* la forma d'una línia senzilla, sense coincidir-hi exactament



Com en un diagrama de dispersió no queda reflectida la freqüència amb la que es dona un determinat parell de valors (x_i , y_j), es fa una representació tridimensional, on dos dels eixos es fan servir per X i Y respectivament i el tercer per expressar f_{ij}.



Correlació. Relació entre variables

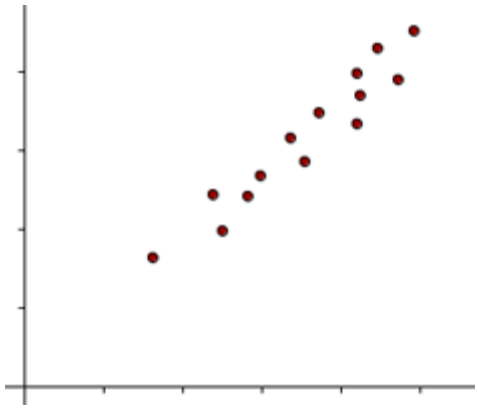
El diagrama de dispersió mostra la relació entre dues variables x i y, que pot ser:

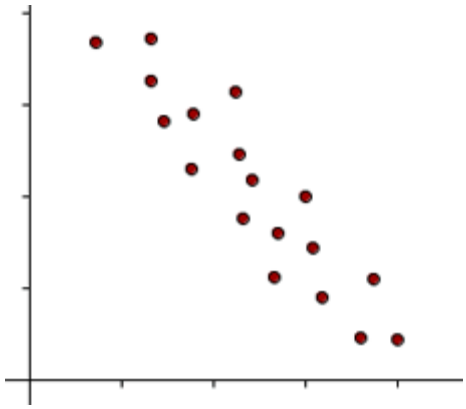
- funcional, si conegut el valor d'una d'elles es pot determinar de forma exacta el valor de l'altre;
- estadística, si conegut el valor d'una d'elles es pot estimar de forma aproximada el valor de l'altre.

La relació o dependència entre les dues variables que intervenen en una distribució bidimensional es determina amb la correlació. Dues variables estan correlacionades si els canvis en una de elles influeixen en els canvis de l'altre.

● Tipus de correlació:

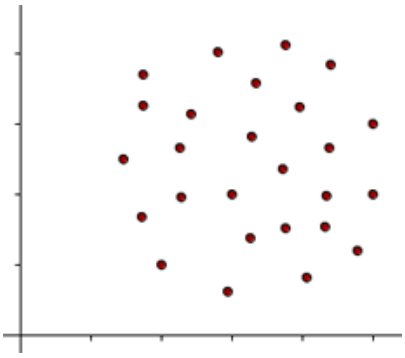
a) Directe / Inversa. En augmentar una d'elles l'altre augmenta/disminueix.





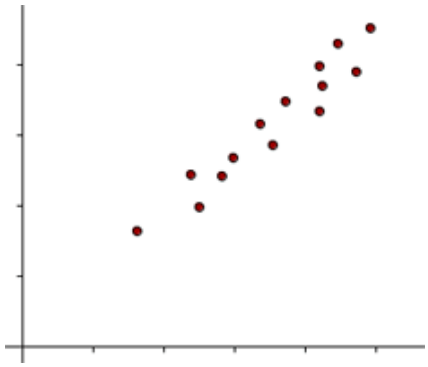
b) lineal o curvilínia, si el núvol de punts s'aproximen en la seva forma a una forma recta o corba.

La correlació és nul·la quan no hi ha dependència de cap tipus entre les variables. En aquest cas diem que les variables estan incorrelacionades i el diagrama de punts té una forma arrodonida.

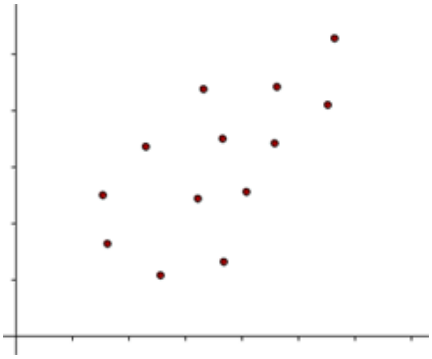


● Grau de correlació. Indica la proximitat que hi ha entre els punts del diagrama. Es pot diferenciar:

a) Correlació forta. La relació entre variables serà més forta quant més a prop es trobin els punts d'una recta, paràbola, ... o del gràfic d'una funció coneguda



b) Correlació feble. La relació serà més feble quant menys s'ajustin els punts a la forma d'un gràfic conegut: recta, paràbola, ...



Covariància

La covariància d'una variable bidimensional (X, Y) és una mesura estadística

$$\sigma_{xy} = \frac{\sum f_{ij} \cdot (x_i - \bar{x}) \cdot (y_j - \bar{y})}{N} = \frac{\sum f_{ij} \cdot x_i \cdot y_j}{N} - \bar{x} \cdot \bar{y}$$

La covariància és una mesura que permet saber la relació entre dues variables

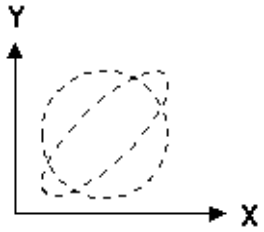
$\sigma_{xy} > 0 \rightarrow$ directa

$\sigma_{xy} < 0 \rightarrow$ inversa

$\sigma_{xy} = 0 \rightarrow$ variables independents, no hi ha cap tipus de relació

● La covariància té com inconvenient el fet que el seu valor depèn de la unitat en que s'expressin les variables. Es a dir, el valor serà diferent si l'altura s'expressa, per exemple, en metres o centímetres.

● Variància i covariància. Les variables X i Y tenen la mateixa variància en el cas de l'el·lipse i del cercle, però la covariància en el cercle és zero i la de l'el·lipse és més o menys gran i positiva.



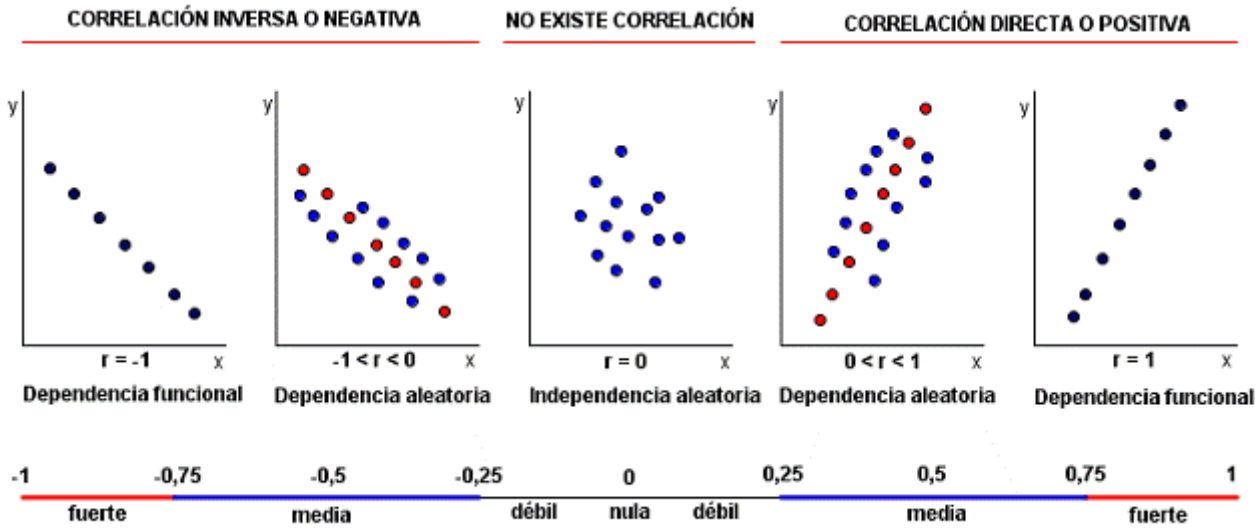
Correlació

El coeficient de correlació lineal o coeficient de Pearson **r** és una mesura de la variable bidimensional (X, Y) que determina el grau de **dependència lineal** entre les variables X i Y.

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

Observacions:

- r sempre està entre -1 i 1
- Si r és major a 0 la dependència serà positiva, si és menor a 0 la dependència serà negativa;
- Si r = 1 o r = - 1 els punts estan situats en línea recta i la correlació lineal perfecta;
- Quant més a prop de 1 o - 1 es troba r més forta és l'associació lineal entre les variables;
- Quant més a prop de 0 més dèbil serà la **dependència lineal**. Observem que això no implica que no existeixi relació entre les dues variables, potser que aquesta no sigui lineal;
- r no té dimensió ni unitats. No varia quan en les variables es canvia l'escala o es canvia d'origen;
- es tracta d'una mesura matemàtica que cal interpretar. Podem trobar un grau de correlació elevat entre dues variables però que aquestes no tinguin cap relació.



Ex: Per l'exemple inicial, en relació a un estudi sobre el nombre d'autobusos (X) que agafa una sèrie de persones i el temps que utilitzen en total per desplaçar-se (Y)

Y \ X	1	2	3	Total
[10,20)	14	4	1	**19**
[20,30)	10	16	3	**29**
[30,40)	14	4	6	**24**
[40,50)	7	6	5	**18**
Total	**45**	**30**	**15**	**90**

x_i	y_i	f_i	$x_i \cdot f_i$	$y_i \cdot f_i$	$x_i^2 \cdot f_i$	$y_i^2 \cdot f_i$	$f_i \cdot x_i \cdot y_i$
1	*15*	*14*	*14*	*210*	*14*	*3150*	*210*
1	*25*	*10*	*10*	*250*	*10*	*6250*	*250*
1	*35*	*14*	*14*	*490*	*14*	*17150*	*490*
1	*45*	*7*	*7*	*315*	*7*	*14175*	*315*
2	*15*	*4*	*8*	*60*	*16*	*900*	*120*
2	*25*	*16*	*32*	*400*	*64*	*10000*	*800*
2	*35*	*4*	*8*	*140*	*16*	*4900*	*280*
2	*45*	*6*	*12*	*270*	*24*	*12150*	*540*
3	*15*	*1*	*3*	*15*	*9*	*225*	*45*
3	*25*	*3*	*9*	*75*	*27*	*1875*	*225*
3	*35*	*6*	*18*	*210*	*54*	*7350*	*630*
3	*45*	*5*	*15*	*225*	*45*	*10125*	*675*
Suma		*90*	*150*	*2660*	*300*	*102425*	*4580*

$$\bar{x} = \frac{150}{90} = 1{,}67$$

$$\bar{y} = \frac{2660}{90} = 29{,}56$$

$$\sigma_{xy} = \frac{4580}{90} - 1{,}67 \cdot 29{,}56 = -44{,}17$$

S'observa una relació inversa

• El **coeficient de determinació lineal** o $\mathbf{r^2}$ ens indica el percentatge de dades que s'expliquen amb el coeficient de correlació r.
Ex: Si dues variables correlacionades linealment presenten un coeficient r = 0,9, el coeficient de determinació lineal r^2 indica que el 81% de les dades queden explicades per la correlació lineal.

<u>Rectes de regressió</u>
Les rectes de regressió són les que millor s'ajusten al diagrama de dispersió. Quan es sospita que hi ha una **dependència lineal** entre les variables, pot resultar útil calcular les equacions de la recta de regressió, que permeten fer prediccions de valors d'una variable en funció de l'altra.

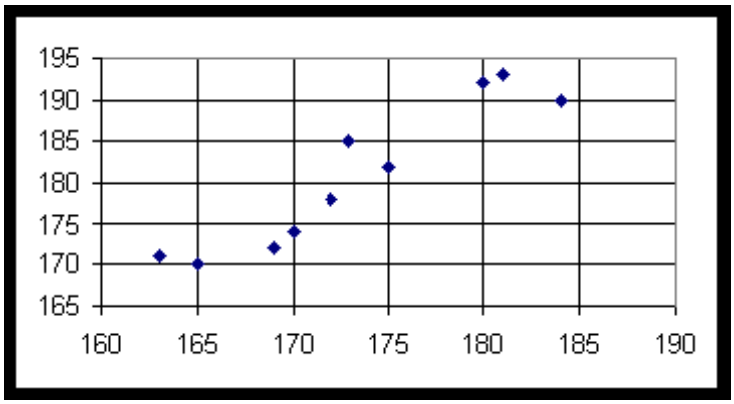
De y sobre x (per estimar valors de y des del valor de x): $y - \bar{y} = \frac{\sigma_{xy}}{\sigma_x^{\,2}}(x - \bar{x})$

De x sobre y (per estimar valors de x des de valors de y): $x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^{\,2}}(y - \bar{y})$

Ex: S'ha estudiat l'altura d'un grup de pares (X) i la dels seus fills (Y).

a) Hi ha una relació entre ambdues variables? De quin tipus és?

La representació dels resultats obtinguts sembla mostrar una certa relació lineal que haurem de confirmar amb el càlcul de r.



x	y	x^2	y^2	x.y
165	170	27225	28900	28050
170	174	28900	30276	29580
172	178	29584	31684	30616
184	190	33856	36100	34960
169	172	28561	29584	29068
163	171	26569	29241	27873
175	182	30625	33124	31850
181	193	32761	37249	34933
173	185	29929	34225	32005
180	192	32400	36864	34560
1732	1807	300410	327247	313495

$$s_x = 6{,}54 \quad s_y = 8{,}49 \quad s_{xy} = 52{,}26$$

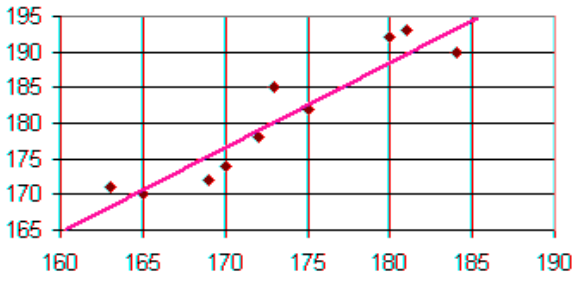
El coeficient de correlació lineal és r = 0,94. Indica l'existència d'una forta correlació lineal entre les variables.

b) Quina serà l'alçada del fill d'un pare que fa d'alçada 190 cm?

La recta de regressió és:

$$y - 180{,}7 = \frac{52{,}26}{42{,}77}(x - 173{,}2)$$

$$y = 1{,}22x - 31$$



Segons la recta de regressió, a un pare d'altura 190 cm li correspon una altura de 200,8 del fill.

$$y = 1{,}22 \cdot x - 31$$

$$y = 1{,}22 \cdot 190 - 31$$

$$y = 200{,}8\ c$$

La" bondat" de la predicció és de r^2, en aquest cas d'un 88%

● S'ha d'evitar utilitzar les rectes de regressió quan la representació gràfica de les variables o el càlcul de coeficient de correlació no ofereixi resultats fiables (propers a 1 o a -1). També en el cas de valors extrems de la distribució, molt grans o molt petits.

● El coeficient de correlació és molt sensible a valors que s'allunyen de la recta de regressió. Si en un conjunt de valors agrupats sobre una recta s'afegeix un nou valor separat de la mateixa, el nou coeficient de correlació canviarà de manera significativa indicant, en conjunt, una correlació menor. Aquesta propietat s'utilitza per determinar aquells valors "estranys" que, en suprimir-se, poden donar lloc a correlacions més significatives. No cal dir que és una tècnica que s'ha d'aplicar amb molta cura.