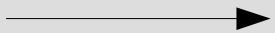
Tema 2: Distribucions estadístiques bidimensionals

– Organització de les dades:

- **Taules de contingencia**: mostren les freqüencies absolutes **conjuntes** de les dues variables
- A partir de la taula de contingència es poden calcular les **freqüencies** absolutes **marginals** (individuals) de cada variable

Dades brutes

X	Y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5
x_6	y_6
⋮	⋮
x_n	y_n



Taula de contingencia

$X \setminus Y$	y_1	y_2	$\cdots$	y_l	Suma
x_1	n_{11}	n_{12}	$\cdots$	n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2l}	$n_{2\bullet}$
⋮			⋮		⋮
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k\bullet}$
Suma	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet l}$	N

Distribucions estadístiques bidimensionals

– Taules de contingencia:

Exemple:

	Nota Mitjana Batxillerat	Nota Estadistica
Alumne 1	7,7	9
Alumne 2	7,1	7,5
Alumne 3	5,5	5
Alumne 4	6,2	5
Alumne 5	6,8	7
Alumne 6	5,8	5,5
Alumne 7	7,9	8,5
Alumne 8	6,7	6
Alumne 9	7,2	6,5
Alumne 10	5,4	4
Alumne 11	6,6	6
Alumne 12	8	7
Alumne 13	6,8	8
Alumne 14	7,1	7,5
Alumne 15	7	9
Alumne 16	5,8	4,5
Alumne 17	6,8	7

Distribució marginal nota batxillerat

Notes Batx. | Est.	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9	
5,4	1	0	0	0	0	0	0	0	0	0	0	1
5,5	0	0	1	0	0	0	0	0	0	0	0	1
5,8	0	1	0	1	0	0	0	0	0	0	0	2
6,2	0	0	1	0	0	0	0	0	0	0	0	1
6,6	0	0	0	0	1	0	0	0	0	0	0	1
6,7	0	0	0	0	1	0	0	0	0	0	0	1
6,8	0	0	0	0	0	0	2	0	1	0	0	3
7	0	0	0	0	0	0	0	0	0	0	1	1
7,1	0	0	0	0	0	0	0	2	0	0	0	2
7,2	0	0	0	0	0	1	0	0	0	0	0	1
7,7	0	0	0	0	0	0	0	0	0	0	1	1
7,9	0	0	0	0	0	0	0	0	0	1	0	1
8	0	0	0	0	0	0	1	0	0	0	0	1
	1	1	2	1	2	1	3	2	2	1	2	17

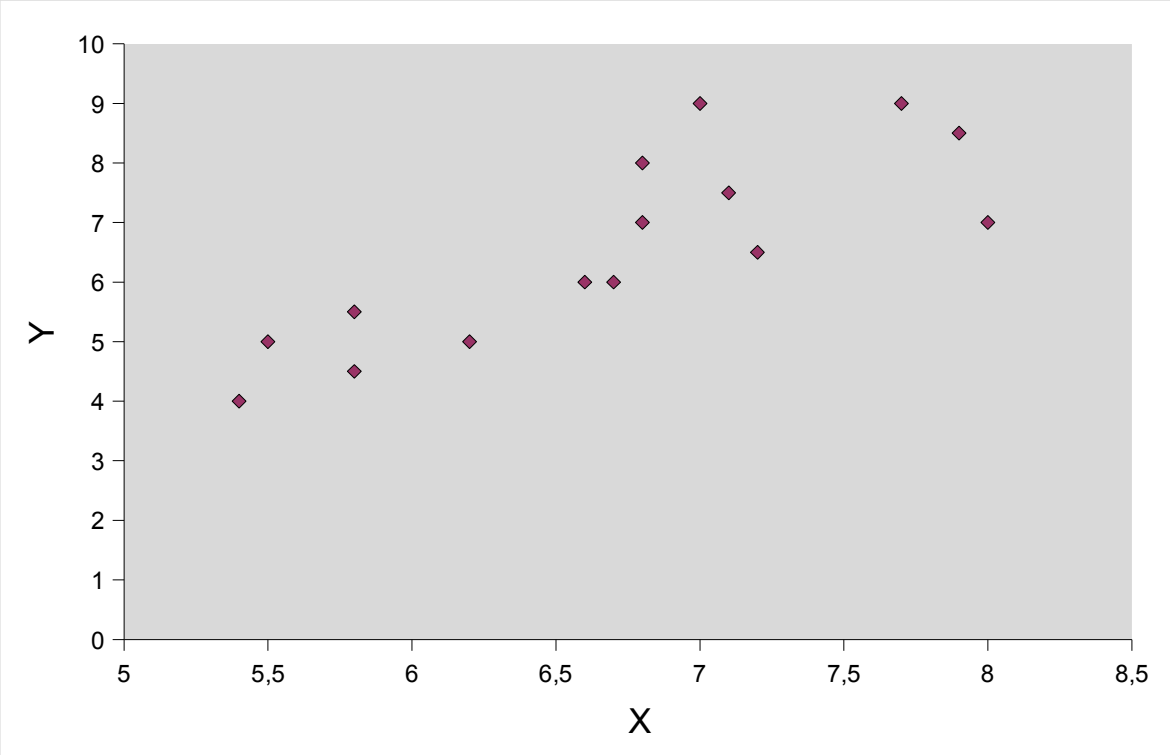
Distribució marginal nota estadística

Total

Distribucions estadístiques bidimensionals

- Representació gràfica conjunta de dades bidimensionals:

  Diagrama de dispersió: un **punt** per a cada parell de valors
  (coordenada x: primera variable, coordenada y: segona variable)



Distribucions estadístiques bidimensionals

– Mesura de la relació entre les variables d'una distribució bidimensional:

- **Coeficient de correlació**: mesura el grau de *relació lineal* entre les variables
- **Coeficient de contingència**: mesura el grau de *dependència* entre les variables

Distribucions estadístiques bidimensionals

- Coeficient de correlació:

$$r = \frac{Cov_{XY}}{s_X \cdot s_Y}$$

s_X, s_Y: desviacions típiques de X i Y

$\bar{x}$, $\bar{y}$: mitjanes de X i Y

(Covariància **poblacional**)

$$Cov_{XY} = \frac{(x_1-\bar{x})\cdot(y_1-\bar{y})+(x_2-\bar{x})\cdot(y_2-\bar{y})+\cdots}{N} = \frac{x_1\cdot y_1+x_2\cdot y_2+\cdots}{N} - \bar{x}\cdot\bar{y}$$

(dades brutes)

$$Cov_{XY} = \frac{(x_1-\bar{x})\cdot(y_1-\bar{y})\cdot n_{11}+(x_1-\bar{x})\cdot(y_2-\bar{y})\cdot n_{12}+\cdots+(x_2-\bar{x})\cdot(y_1-\bar{y})\cdot n_{21}+\cdots}{N}$$

$$Cov_{XY} = \frac{x_1\cdot y_1\cdot n_{11}+x_1\cdot y_2\cdot n_{12}+\cdots+x_2\cdot y_1\cdot n_{21}+\cdots}{N} - \bar{x}\cdot\bar{y}$$

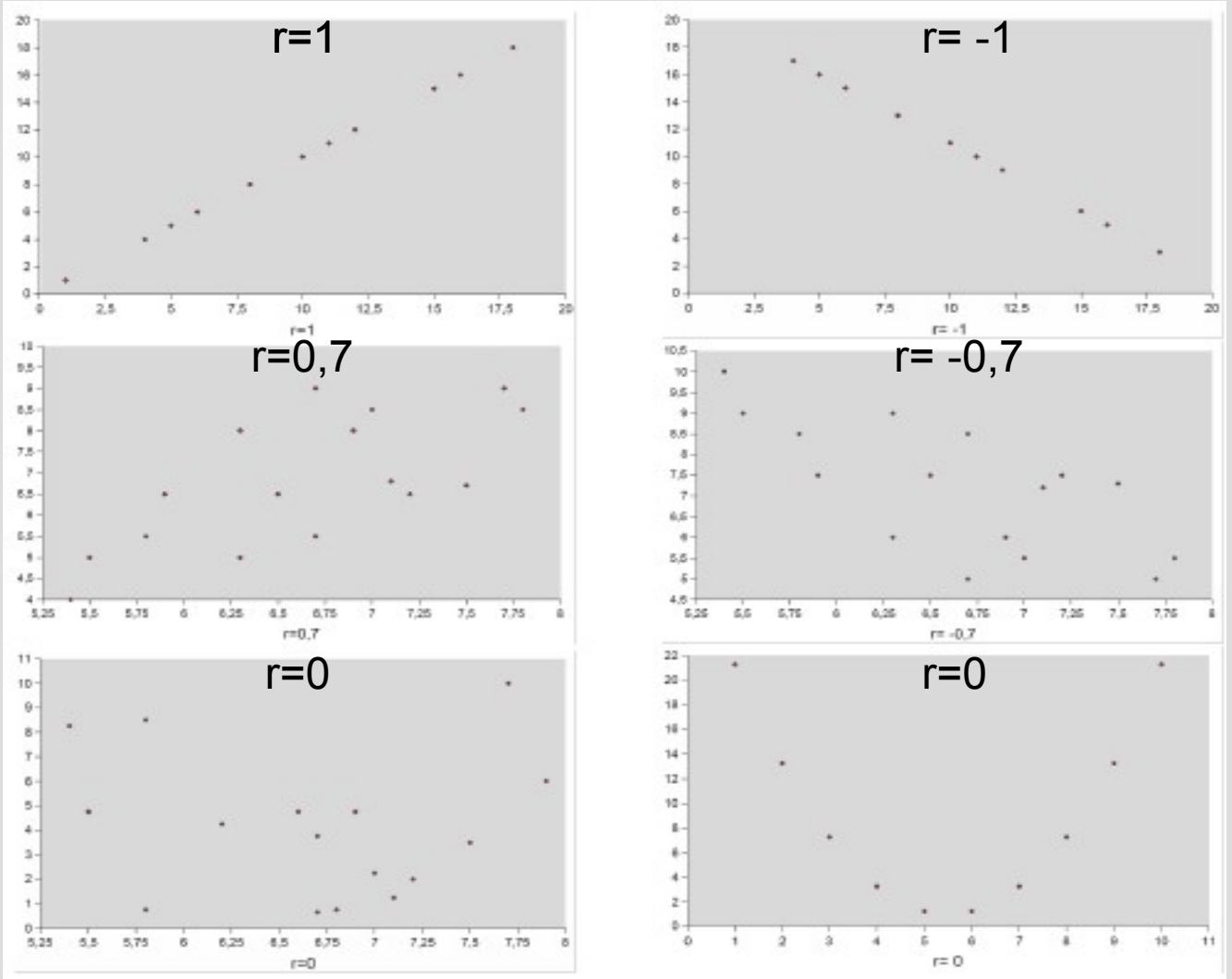
(dades en taula de contingència)

Distribucions estadístiques bidimensionals

- Coeficient de correlació: relació amb el diagrama de dispersió

Propietat:

$-1 \leq r \leq 1$



Distribucions estadístiques bidimensionals

- Coeficient de correlació: Exemple

Dades brutes

	Nota Mitjana Batxillerat **X**	Nota Estadistica **Y**
Alumne 1	7,7	9
Alumne 2	7,1	7,5
Alumne 3	5,5	5
Alumne 4	6,2	5
Alumne 5	6,8	7
Alumne 6	5,8	5,5
Alumne 7	7,9	8,5
Alumne 8	6,7	6
Alumne 9	7,2	6,5
Alumne 10	5,4	4
Alumne 11	6,6	6
Alumne 12	8	7
Alumne 13	6,8	8
Alumne 14	7,1	7,5
Alumne 15	7	9
Alumne 16	5,8	4,5
Alumne 17	6,8	7

$$\bar{x} = \frac{7,7 + 7,1 + \cdots + 6,8}{17} = 6,73$$

$$\bar{y} = \frac{9 + 7,5 + \cdots + 7}{17} = 6,65$$

$$Var_X = \frac{7,7^2 + 7,1^2 + \cdots + 6,8^2}{17} - 6,73^2 = 0,58 \longrightarrow s_X = \sqrt{0,58} = 0,76$$

$$Var_Y = \frac{9^2 + 7,5^2 + \cdots + 7^2}{17} - 6,65^2 = 2,2 \longrightarrow s_Y = \sqrt{2,2} = 1,48$$

$$Cov_{XY} = \frac{7,7 \cdot 9 + 7,1 \cdot 7,5 + \cdots + 6,8 \cdot 7}{17} - 6,73 \cdot 6,65 = 0,93$$

$$r = \frac{0,93}{0,76 \cdot 1,48} = 0,83 \longrightarrow$$ Forta correlació lineal entre les variables

Distribucions estadístiques bidimensionals

- Coeficient de correlació: Exemple

Taula de contingència

Notes Batx. **X** | Est. **Y**	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5	9	
5,4	1	0	0	0	0	0	0	0	0	0	0	1
5,5	0	0	1	0	0	0	0	0	0	0	0	1
5,8	0	1	0	1	0	0	0	0	0	0	0	2
6,2	0	0	1	0	0	0	0	0	0	0	0	1
6,6	0	0	0	0	1	0	0	0	0	0	0	1
6,7	0	0	0	0	1	0	0	0	0	0	0	1
6,8	0	0	0	0	0	0	2	0	1	0	0	3
7	0	0	0	0	0	0	0	0	0	0	1	1
7,1	0	0	0	0	0	0	0	2	0	0	0	2
7,2	0	0	0	0	0	1	0	0	0	0	0	1
7,7	0	0	0	0	0	0	0	0	0	0	1	1
7,9	0	0	0	0	0	0	0	0	0	1	0	1
8	0	0	0	0	0	0	1	0	0	0	0	1
	1	1	2	1	2	1	3	2	2	1	2	17

$$\bar{x} = \frac{5{,}4 \cdot 1 + 5{,}5 \cdot 1 + \cdots + 8 \cdot 1}{17} = 6{,}73$$

$$\bar{y} = \frac{4 \cdot 1 + 4{,}5 \cdot 1 + \cdots + 9 \cdot 2}{17} = 6{,}65$$

$$Var_X = \frac{5{,}4^2 \cdot 1 + \cdots + 8^2 \cdot 1}{17} - 6{,}73^2 = 0{,}58$$

$$s_X = \sqrt{0{,}58} = 0{,}76$$

$$Var_Y = \frac{4^2 \cdot 1 + \cdots + 9^2 \cdot 2}{17} - 6{,}65^2 = 2{,}2$$

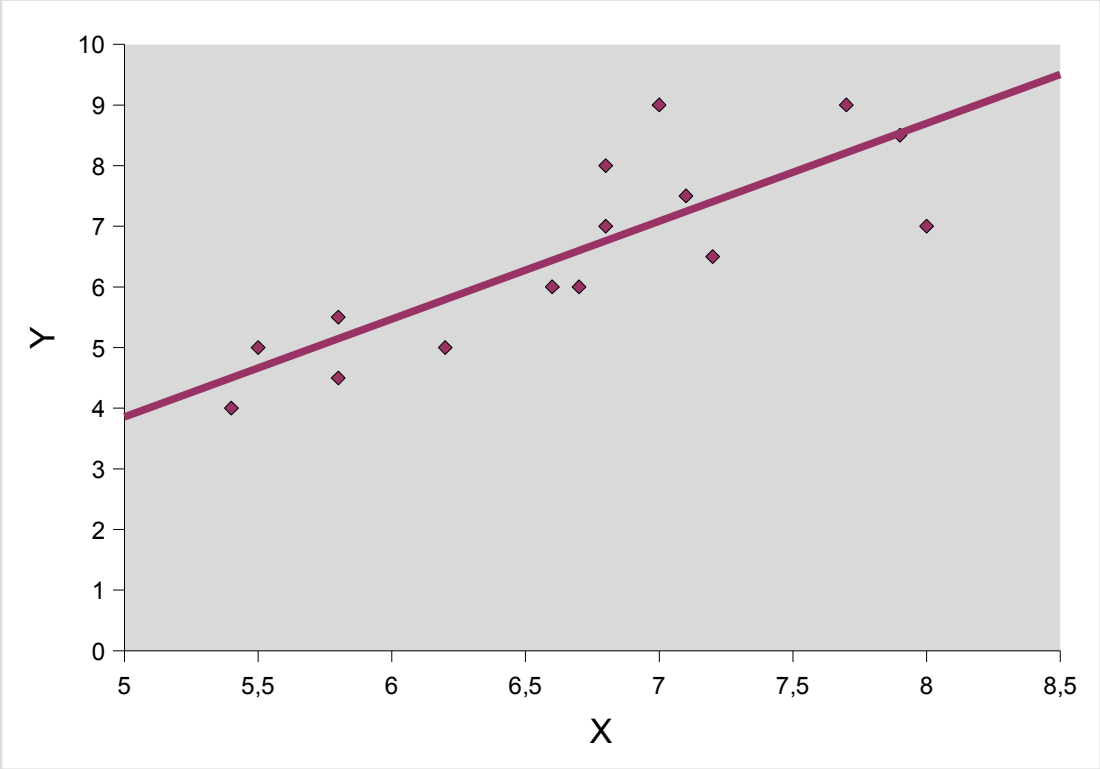
$$s_Y = \sqrt{2{,}2} = 1{,}48$$

$$Cov_{XY} = \frac{5{,}4 \cdot 4 \cdot 1 + \cdots + 8 \cdot 9 \cdot 0}{17} - 6{,}73 \cdot 6{,}65 = 0{,}93$$

$$r = \frac{0{,}93}{0{,}76 \cdot 1{,}48} = 0{,}83$$

Distribucions estadístiques bidimensionals

- Coeficient de correlació:
  - **Recta de regressió** : recta que millor aproxima el conjunt de punts del diagrama de dispersió



$$\hat{Y} = aX + b$$

$$a = \frac{Cov_{XY}}{Var_X}$$

$$b = \bar{y} - a\,\bar{x}$$

En el nostre exemple:

$$a = \frac{0,93}{0,58} = 1,61$$

$$b = 6,65 - 1,61 \cdot 6,73 = -4,2$$

$$\hat{Y} = 1,61\,X - 4,2$$

Si, p. ex., x=6,5 ⟶ ŷ=6,28 (predicció)

Distribucions estadístiques bidimensionals

- Coeficient de contingència:
  - Mesura el grau d'independència entre les variables
  - Concepte **d'independència estadística**:

    dues variables es consideren independents si la proporció dels seus valors, mesurada respecte al conjunt total de valors, és la mateixa que la proporció de valors mesurada respecte al subconjunt de valors que es té quan un dels valors de l'altra variable es manté fixat

  - Si dues variables són independents, llavors Cov_{XY}=0 (r=0)
    però, el fet que Cov_{XY}=0 no implica que X i Y siguin independents

Distribucions estadístiques bidimensionals

- Coeficient de contingència: càlcul

$$C=\sqrt{\frac{\chi^2}{N+\chi^2}}$$

(independència) $0 \leq C \leq \sqrt{1-\frac{1}{min(k,l)}}$ (dependència)

$$\chi^2=\sum_i \sum_j \frac{(n_{ij}-e_{ij})^2}{e_{ij}}$$

$$e_{ij}=\frac{n_{i*} \cdot n_{*j}}{N}$$

On:

k=nº de files (valors de X) de la taula de contingència

l=nº de columnes (valors de Y) de la taula de contingència

Distribucions estadístiques bidimensionals

- Coeficient de contingència: exemple

Sexe\Consum tabac	Fumador	No fumador	
Home	17	30	47
Dona	21	44	65
	38	74	112

n_{ij}

17	30
21	44

e_{ij}

38*47/112=15,95	74*47/112=31,05
38*65/112=22,05	74*65/112=42,95

$n_{ij}-e_{ij}$

1,05	-1,05
-1,05	1,05

$$\chi^2=\frac{1{,}05^2}{15{,}95}+\frac{(-1{,}05)^2}{31{,}05}+\frac{(-1{,}05)^2}{22{,}05}+\frac{1{,}05^2}{42{,}95}=0{,}1803$$

$$C=\sqrt{\frac{0{,}1803}{112+0{,}1803}}=\sqrt{0{,}0016}=0{,}04$$

$$C_{max}=\sqrt{1-\frac{1}{2}}=\sqrt{0{,}5}=0{,}7071$$

$$\frac{0{,}04}{0{,}7071}=0{,}056=5{,}6\,\%$$ → Alt nivell d'independència entre les variables