Estadística Aplicada

Seminari 2

Seminari 1

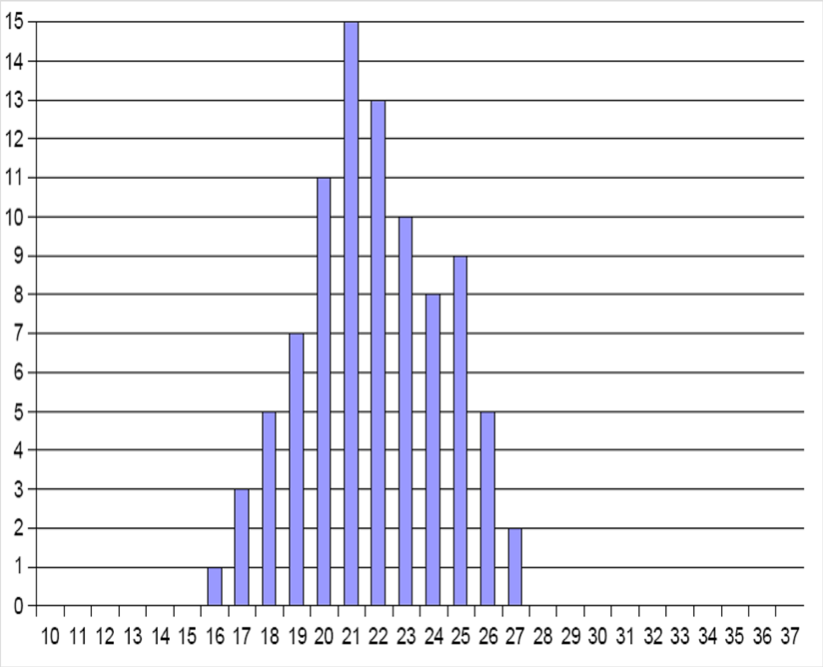
- Resum teoria (temes 1, 2 i 3): tipus d'estadístiques, tipus de variables, taules de freqüència, representacions gràfiques, mesures de tendència central (moda, mediana, mitjana)
- Exemples pràctics: classificació RCD Mallorca temporada 2006-07

Seminari 2

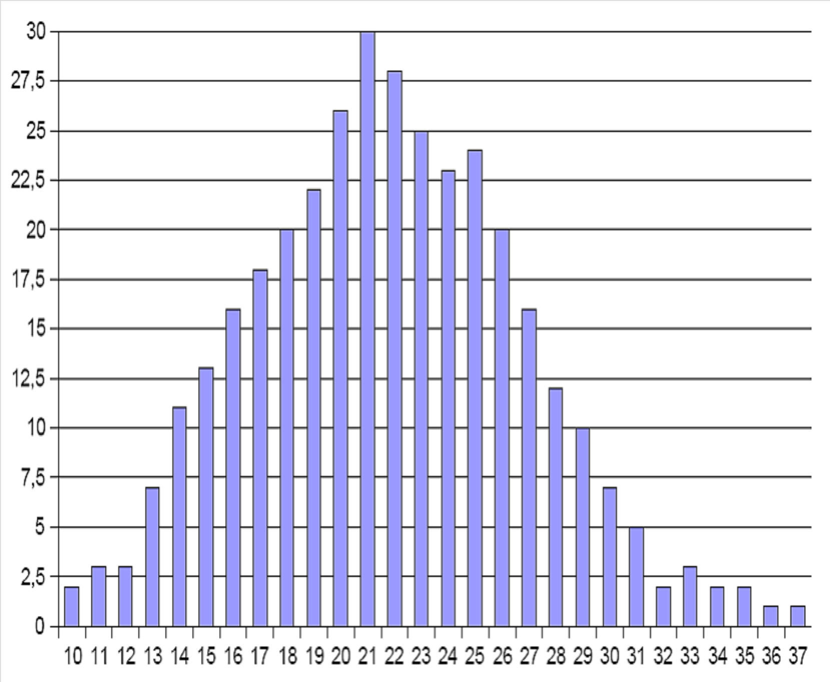
- Resum teoria (temes 4 i 5)
- Exemples pràctics

Resum teoria

– Mesures de dispersió



Moda= 21
Mediana= 22
Mitjana= 21,82



Moda= 21
Mediana= 22
Mitjana= 21,82

Les mesures de tendència central no són suficients per a descriure una distribució de valors

Resum teoria

– Mesures de dispersió

- ***Ratio* de variació**: medeix la concentració dels valors respecte a la moda. Única mesura de dispersió possible per a variables nominals.

$$RV = 1 - \frac{n_{moda}}{N}$$

(concentració) 0 ≤ *RV* ≤ 1 (dispersió)

Exemple:

Immigració per nacionalitats

Nacionalitat	Freq absoluta
Colòmbia	350
Ecuador	250
Perú	120
Argentina	100
Rumania	80
Marruecos	70
Senegal	30
	1000 =N

moda: Colòmbia

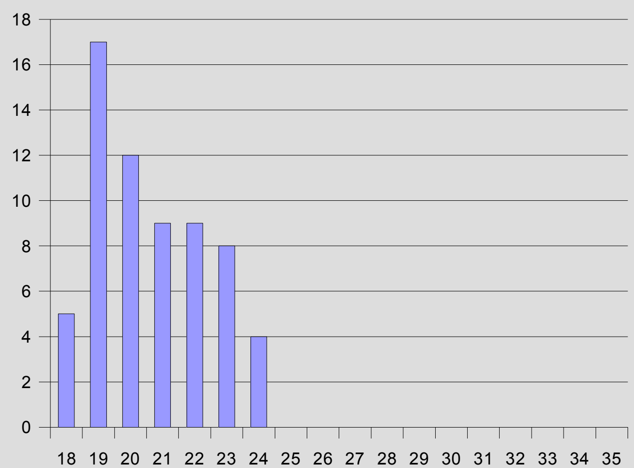
$$RV = 1 - \frac{350}{1000} = 0,65$$

Resum teoria

– Mesures de dispersió

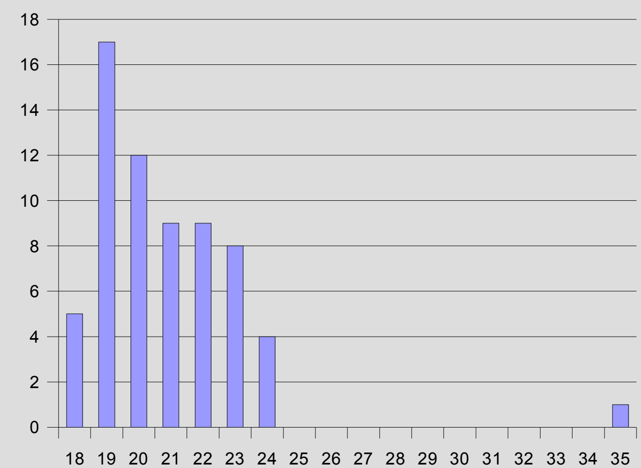
- **Rang**: diferència entre els valors màxim i mínim
- **Rang interquartílic (RIC)**: diferència entre el tercer i el primer quartils. Menos sensible a valors extrems que el rang.

Exemple:



Rang=6

RIC=3



Rang=17

RIC=3

Resum teoria

– Mesures de dispersió

• **Desviació típica**: mesura la diferència entre les dades i el seu valor mitjà

$$s = \sqrt{Var}$$

Var és la **Variància** de les dades

Es compleix que:

- al menys un 75% dels valors estan entre $\bar{x} - 2 \cdot s$ i $\bar{x} + 2 \cdot s$
- al menys un 89% dels valors estan entre $\bar{x} - 3 \cdot s$ i $\bar{x} + 3 \cdot s$
- al menys un 93% dels valors estan entre $\bar{x} - 4 \cdot s$ i $\bar{x} + 4 \cdot s$

Resum teoria

– Mesures de dispersió

- Exemple desviació típica i variància: nota d'estadística de 8 persones

Dades brutes

7
5
9
7
5
6
7
6
4

9 valors

$$\bar{x}=\frac{7+5+\cdots+4}{9}=6{,}22$$

$$Var\,x=\frac{7^2+5^2+\cdots+4^2}{9}-6{,}22^2=1{,}98$$

$$s=\sqrt{1{,}98}=1{,}4$$

Taula de freqüències

x_i	n_i
4	1
5	2
6	2
7	3
9	1

$n=9$

$$\bar{x}=\frac{4\cdot1+5\cdot2+\cdots+9\cdot1}{9}=6{,}22$$

$$Var\,x=\frac{4^2\cdot1+5^2\cdot2+\cdots+9^2\cdot1}{9}-6{,}22^2=1{,}98$$

Resum teoria

– Mesures de dispersió

- Variància mostral i variància poblacional

$$Var\,x=\frac{x_1^2+x_2^2+\cdots+x_n^2}{N}-\bar{x}^2$$

poblacional

Exemple: estudi de la satisfacció de tots els clients d'un hotel

$$Var\,x=\frac{x_1^2+x_2^2+\cdots+x_n^2}{N-1}-\bar{x}^2\cdot\frac{N}{N-1}$$

mostral

Exemple: estudi de la satisfacció d'alguns clients d'un hotel

La distinció entre els dos tipus de variàncies es necessària en estadística inferencial.

Resum teoria

– *z-scores* (tipificació de variables)

permeten comparar dades procedents de distints estudis estadístics

$$z=\frac{x-\bar{x}}{s}$$

Exemple:

Dues persones opten a una beca deportiva. La primera té una marca de 7.60m en salt de longitud i la segona una de 65.4s en 100 metres lliures de natació. Quina persona mereix més la beca?

Per saber quin dels candidats destaca més en el seu respectiu esport hem d'analitzar les marques de la resta de deportistes.
Si tenim que:

Salt de longitud: mitjana 7.40m, desviació típica 0.4m
100m lliures natació: mitjana 68.30s, desviació típica 1.6s

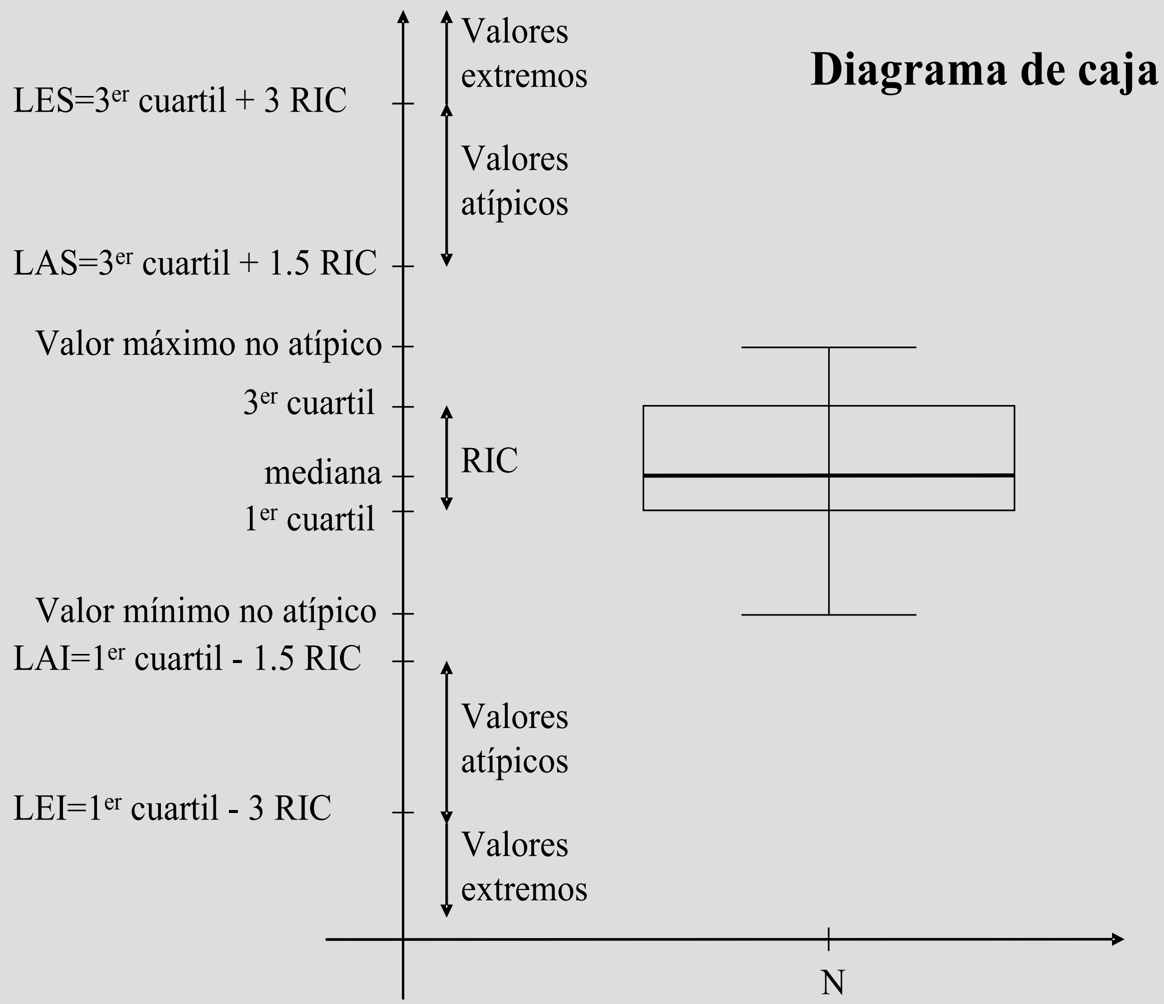
Llavors:

$$z_1=\frac{7.6-7.4}{0.4}=0.5 \qquad z_2=\frac{65.4-68.3}{1.6}=-1 \longrightarrow 1$$

Donariem la beca al nadador perquè destaca més en la seva especialitat.

Resum teoria

– Diagrames de capsa

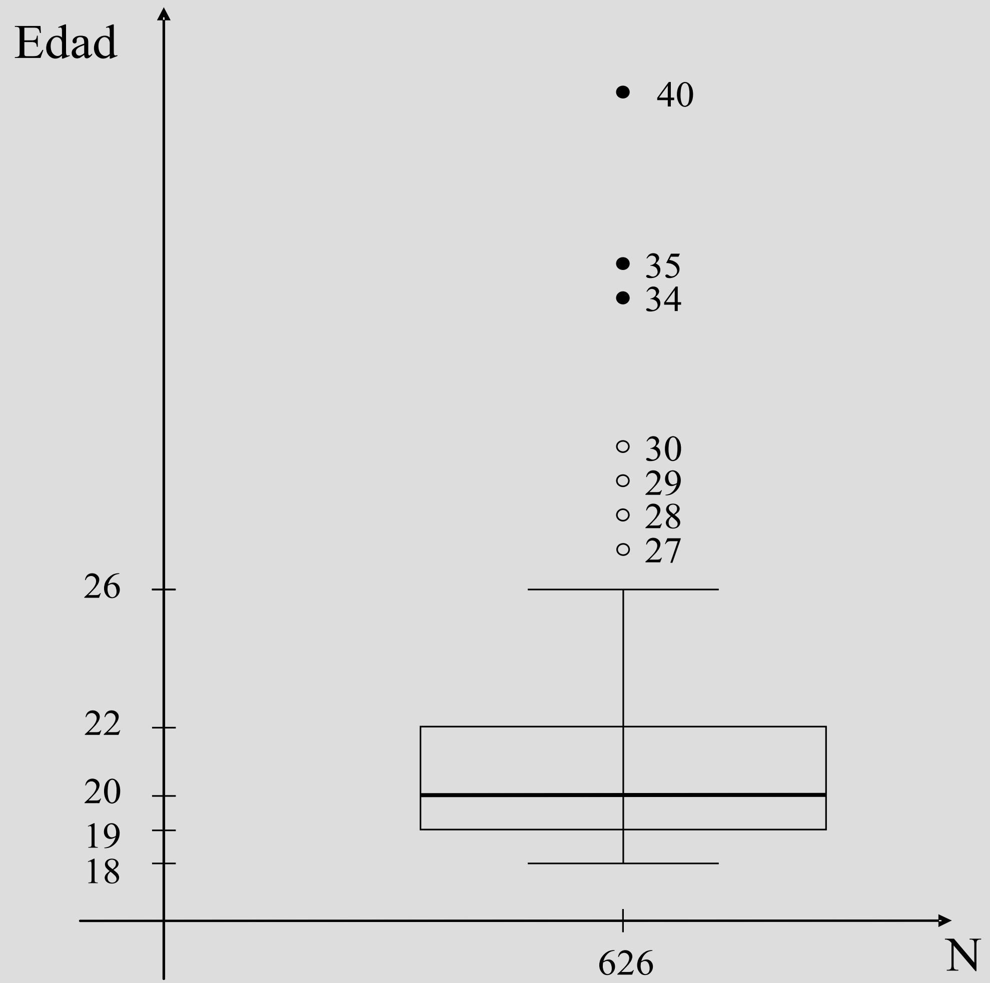


Resum teoria

– Diagrames de capsa (Exemple)

	Edat	Freq. Absoluta
	18	120
1er quartil	19	150
mediana	20	90
	21	70
3er quartil	22	65
	23	50
	24	30
	25	20
	26	10
	27	7
	28	8
	29	2
	30	1
	34	1
	35	1
	40	1
		626

RIC=22-19=3
1er quartil - 1,5 RIC=19-4,5=14,5
3er quartil + 1,5 RIC=22+4,5=26,5



Resum teoria

– Anàlisi estadístic bivariant: relacions entre 2 variables estadístiques

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

Resum teoria

– Anàlisi estadístic bivariant:

Organització de les dades: taules de contingencia

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
$\vdots$	$\vdots$
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}	$n_{2\bullet}$
$\vdots$			$\vdots$		$\vdots$
x_k	n_{k1}	n_{k2}	$\cdots$	n_{kl}	$n_{k\bullet}$
Suma	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet l}$	N

Resum teoria

– Anàlisi estadístic bivariant:

Organització de les dades: taules de contingencia

Exemple:

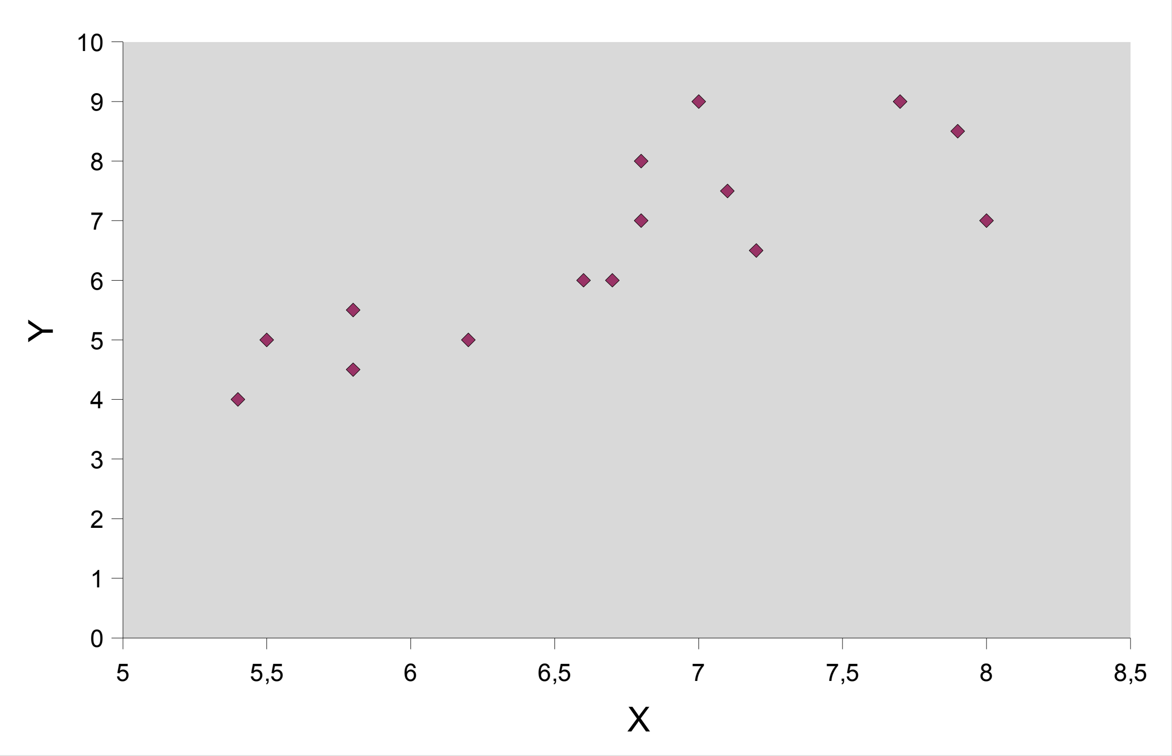
	Nota Mitjana Batxillerat	Nota Estadistica
Alumne 1	7,7	9
Alumne 2	7,1	7,5
Alumne 3	5,5	5
Alumne 4	6,2	5
Alumne 5	6,8	7
Alumne 6	5,8	5,5
Alumne 7	7,9	8,5
Alumne 8	6,7	6
Alumne 9	7,2	6,5
Alumne 10	5,4	4
Alumne 11	6,6	6
Alumne 12	8	7
Alumne 13	6,8	8
Alumne 14	7,1	7,5
Alumne 15	7	9
Alumne 16	5,8	4,5
Alumne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

Nota Batx. \ Nota 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

Resum teoria

– Anàlisi estadístic bivariant:

Representació gràfica de les dades: diagrama de dispersió



Resum teoria

– Anàlisi estadístic bivariant:

Mesura de la relació entre les dades:

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

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

Resum teoria

– Anàlisi estadístic bivariant:

Coeficient de contingència:

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

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

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

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

Resum teoria

– Anàlisi estadístic bivariant:

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

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

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

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

$$\frac{0{,}04}{0{,}7071} = 0{,}056 = 5{,}6\,\% \longrightarrow$$

Alt nivell d'independència entre les variables

Resum teoria

– Anàlisi estadístic bivariant:

Coeficient de correlació:

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

s_X, s_Y: desviacions típiques de X i Y

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

(covariància poblacional)

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

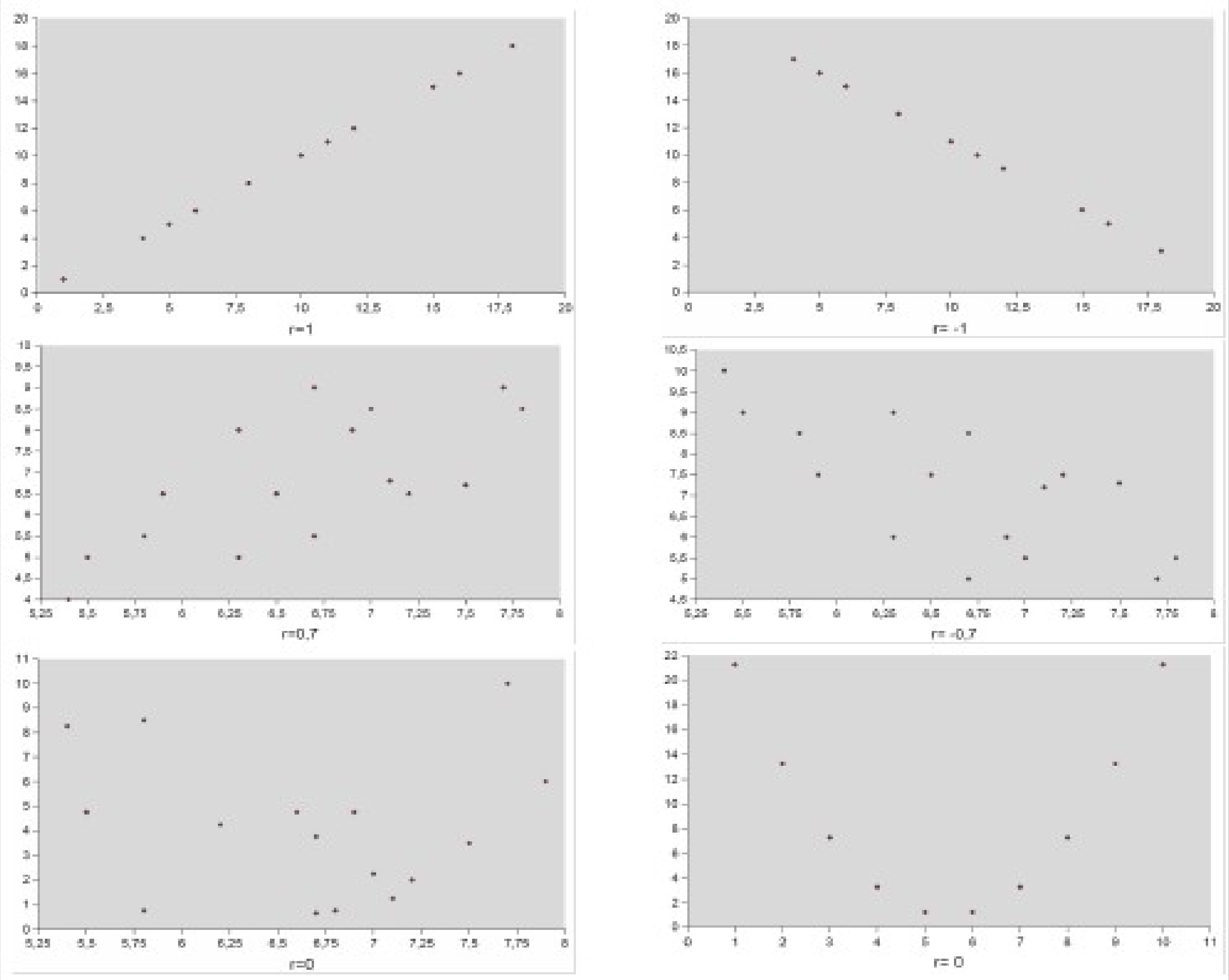
$$Cov_{XY}=\frac{(x_1-\bar{x})\cdot(y_1-\bar{y})+(x_2-\bar{x})\cdot(y_2-\bar{y})+\cdots}{N-1}$$

(covariància mostral)

Resum teoria

– Anàlisi estadístic bivariant:

Coeficient de correlació:



Resum teoria

– Anàlisi estadístic bivariant:

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

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

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

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

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

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

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

Resum teoria

– Anàlisi estadístic bivariant:

Coeficient de correlació (Exemple):

Taula de contingència

Nota Batx. **X** \ Nota 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

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

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

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

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

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

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

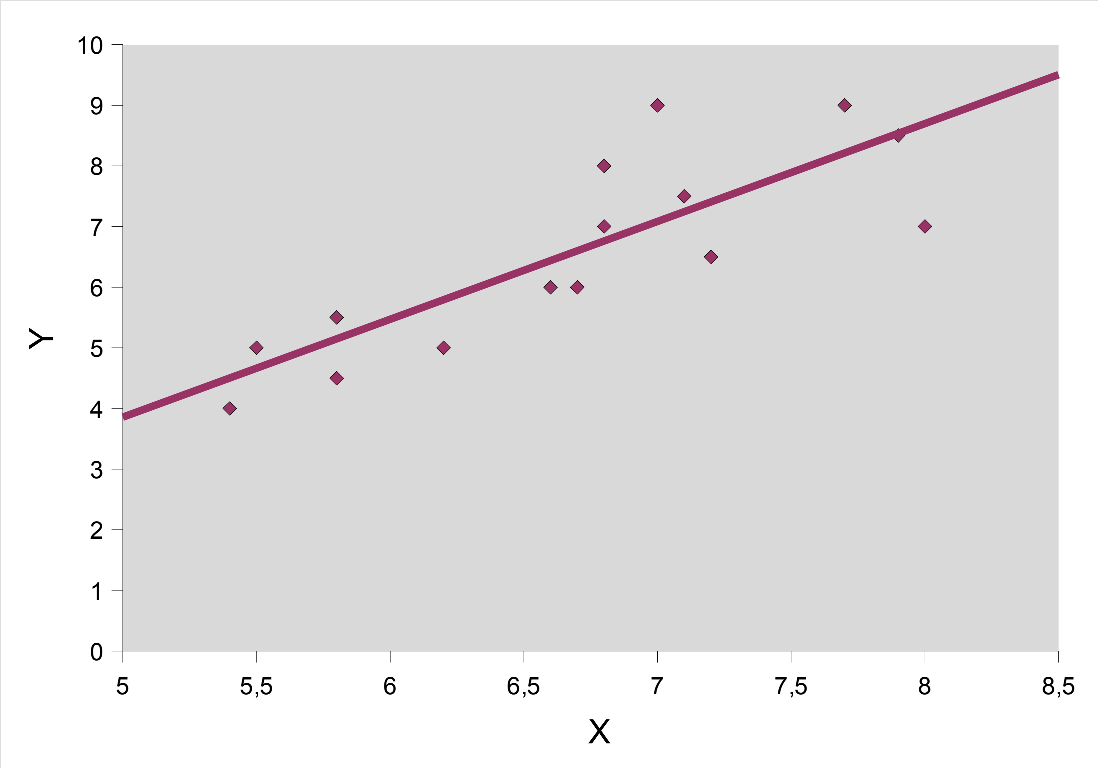
$$Cov=\frac{5{,}4\cdot 4\cdot 1+\cdots+8\cdot 9\cdot 0}{17}-6{,}73\cdot 6{,}65=0{,}93$$

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

Resum teoria

– Anàlisi estadístic bivariant:

Recta de regressió:



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

$$a = \frac{Cov}{Var_X}$$

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

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)

Exemples pràctics

- Activitats dirigides 2 i 3