ESTADÍSTICA

GRAU TURISME

TEMA 3: ANÀLISI DE DADES TURÍSTIQUES BIDIMENSIONALS

Prof. Rosario Martínez Verdú

TEMA 3: ANÀLISI DE DADES TURÍSTIQUES BIDIMENSIONALS

1. Distribucions bidimensionals de freqüències i diagrama de dispersió
2. Covariació i correlació
3. Regressió lineal
4. Anàlisi de la Bondat de l'Ajust i predicció

1.- Distribucions bidimensionals de freqüències i diagrama de dispersió

TIPUS DE DISTRIBUCIONS BIDIMENSIONALS CONJUNTES:

- **Distribucions amb freqüències conjuntes no unitàries**

Objectiu: analitzar dues variables simultàniament o conjunta a partir de l'ordenació de les dades en taules de doble entrada o de contingència.

Família	X nombre de membres	Y nombre de cotxes
1	1	0
2	3	1
3	1	1
4	5	2
5	5	2
6	3	2
7	1	0
8	3	0
9	5	1
10	1	1

a) Obtenir la distribució conjunta de freqüències de (X,Y).

b) Obtenir les distribucions marginals.

c) Són independents X i Y?

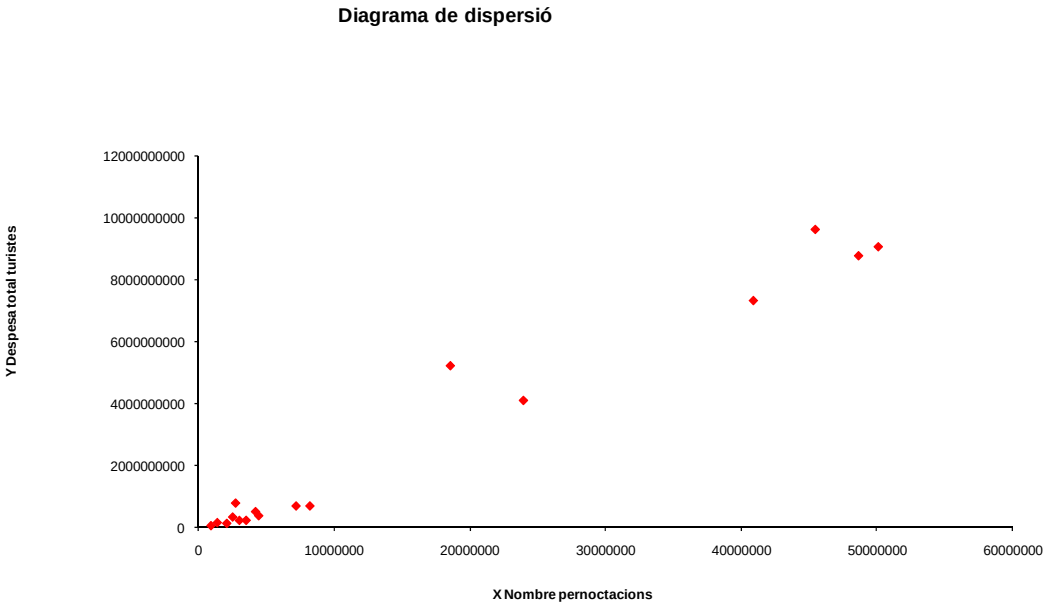
d) Obtenir la distribució de freqüències del nombre de membres de les famílies sense automòbil.

e) Obtenir la distribució de freqüències del nombre de cotxes de les famílies de tres membres.

•Distribucions amb freqüències conjuntes unitàries

Es disposa d'informació de 2009 sobre les N = 17 comunitats autònomes sobre:
•X: nombre de pernoctacions segons comunitat autònoma de destinació principal.
•Y: despesa total dels turistes segons comunitat autònoma de destinació principal, en euros.

CA	X Nombre de pernoctacions	Y Despesa total turistes
Andalusia	40915967	7337693516
Aragó	4417225	365913871
Astúries	2996498	212639740
Balears	48675674	8790673954
Canàries	50131606	9082284827
Cantàbria	2499509	323239643
Castella-La Manxa	3494811	216888633
Castella i Lleó	7178093	680693316
Catalunya	45484289	9643021029
Comunitat Valenciana	23949998	4101830190
Extremadura	2064935	118479497
Galícia	8196397	682792104
Madrid	18560724	5226855314
Múrcia	2715355	775103842
Navarra	1362676	142984959
País Basc	4184217	498036300
Rioja	899254	43334129



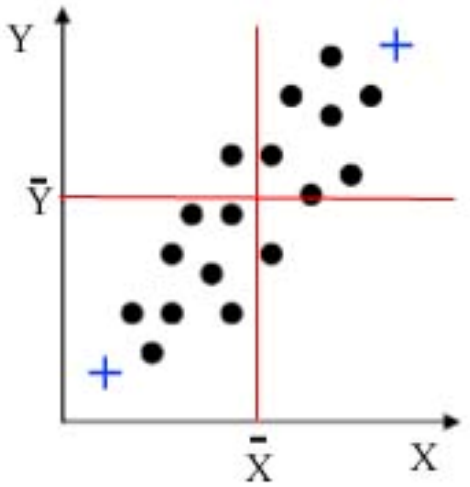
Font: enquesta d'ocupació hotelera 2009, INE i enquesta de despesa turística (Edetur) 2009, IET.

2.- COVARIACIÓ I CORRELACIÓ

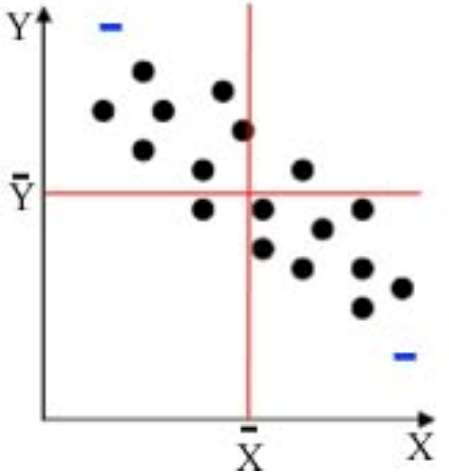
Objectiu: definir unes mesures estadístiques (**covariància i coeficient de correlació lineal**) que posen de manifest l'existència o no de relació de tipus lineal entre dues variables. Per a això, ens basem en dues característiques importants de la distribució conjunta de (X,Y):

Vector de Mitjanes: $\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix}$ Matriu de Variàncies-Covariànces: $\begin{pmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{pmatrix}$

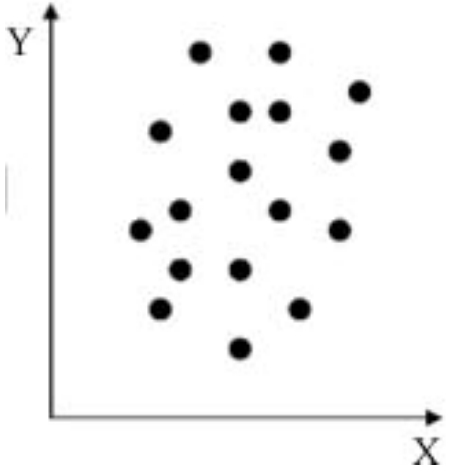
Covariància: $S_{XY} = \frac{1}{N}\sum (X_i - \bar{X})(Y_i - \bar{Y})$ → S_{XY} serveix per a mesurar la variació conjunta entre X i Y. Més que el valor, interessa analitzar-ne el signe.



$S_{XY}>0$ les variables varien en el mateix sentit



$S_{XY}<0$ les variables varien en sentit contrari



$S_{XY}=0$ no hi ha variació conjunta (incorrelació)

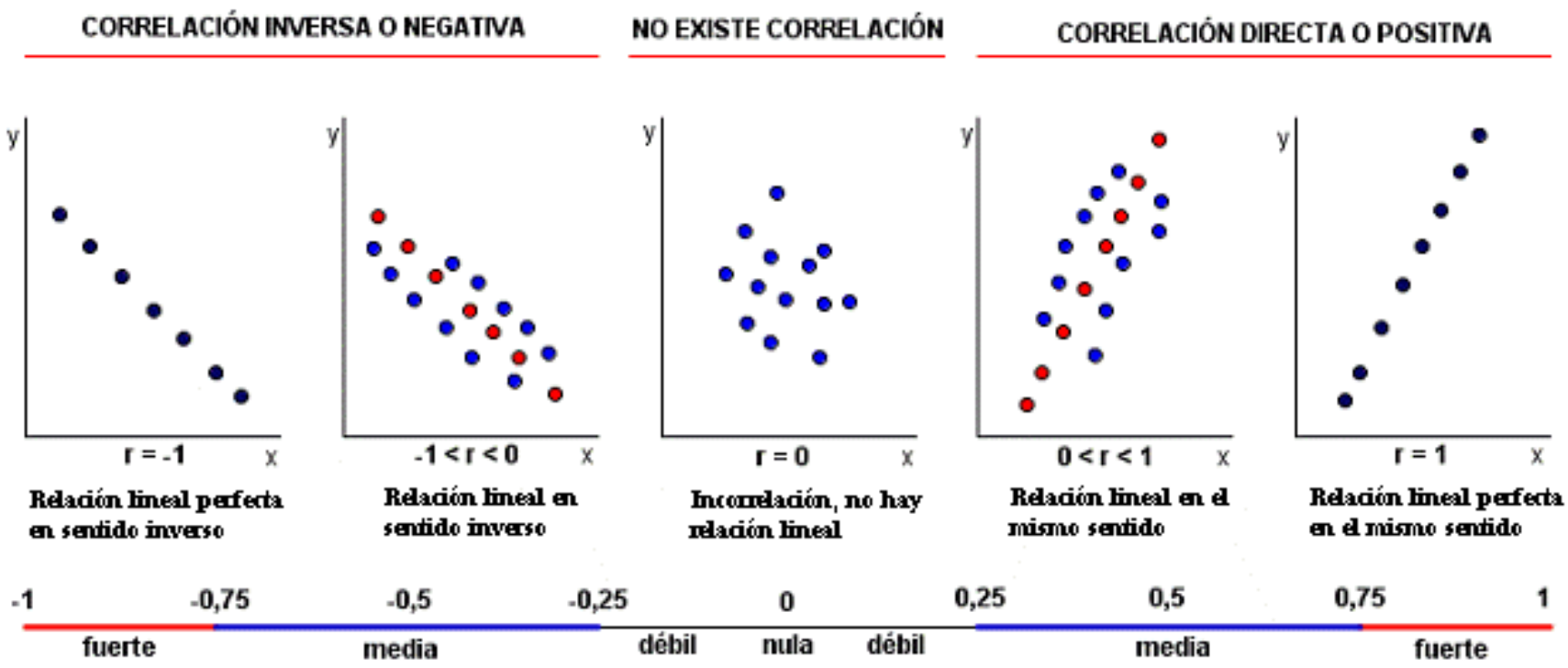
Coeficient de correlació lineal r_{XY}

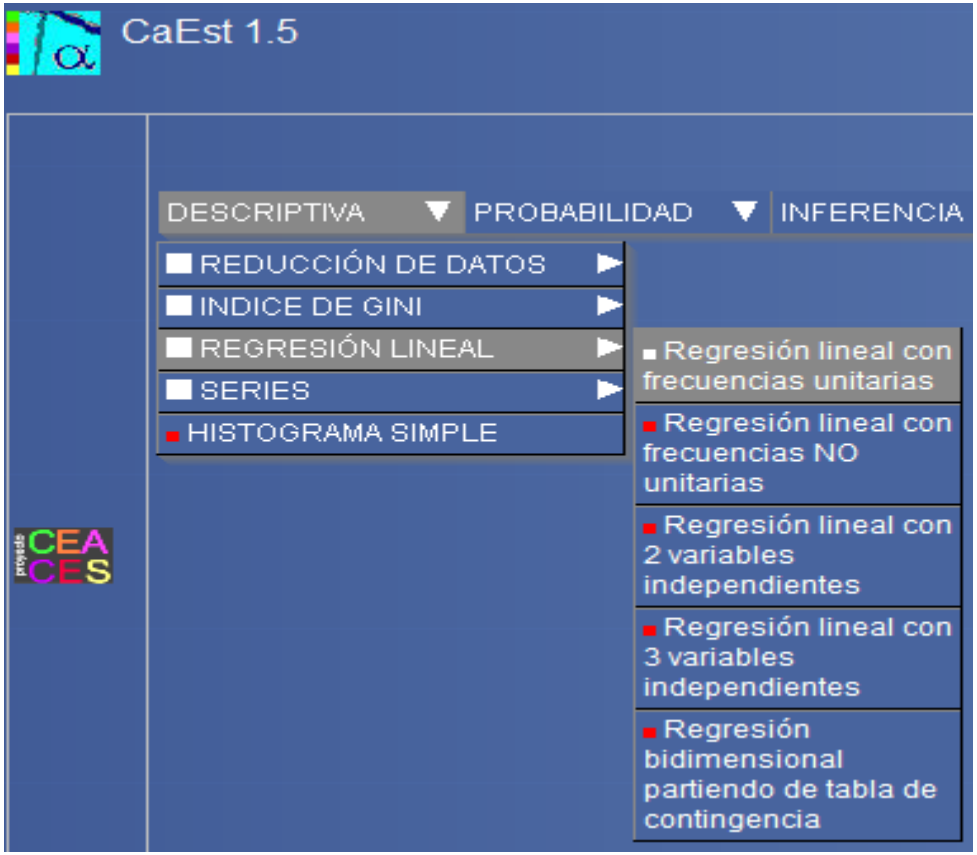
Està basat en la covariància. Mesura el grau o intensitat de la relació lineal entre dues variables i determina el sentit d'aquesta relació. Interessa interpretar-ne tant el valor com el signe. Es defineix com a:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} \qquad -1 \leq r_{XY} \leq 1$$

Signe r_{XY} = signe de S_{XY}

Interpretació del valor i del signe de r_{XY}





Amb la CaEst es poden calcular totes aquestes mesures:

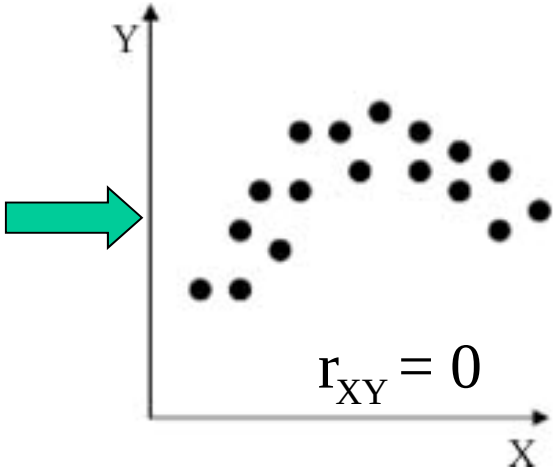
Exemple anterior:
X: nombre de pernoctacions
Y: despesa total dels turistes

Resultats de càlcul de les mesures amb Caest

Indicadors	Y	X	
Mitjana	2837792050,824	15748660,471	
Variàncies i covariància	12704809334987837000	324980124816179.93	63323770899631180 ← S_{XY}
Desv.típica	3564380638,342	18027205,13	
C. correlació	0,985		

r_{XY}

Si $\mathbf{r_{XY}=0}$**, són independents les variables?**
No necessàriament, l'única cosa que es pot concloure és que no hi ha relació lineal entre les variables, però les variables poden tenir un altre tipus de relació.

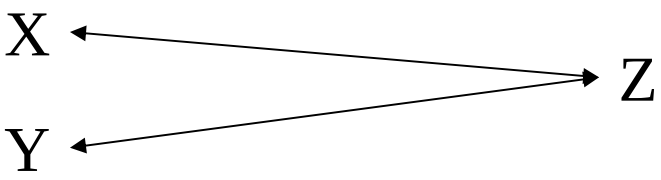


Correlacions espúries o sense sentit

De vegades és possible trobar un coeficient de correlació alt entre dues variables que no tenen relació justificada per cap teoria. D'això, se'n diu correlació espúria o sense sentit. Un exemple: el 1952 Neyman va analitzar la relació entre la taxa de naixements de xiquets i xiquetes i la població de cigonyes a diverses regions, i va trobar un alt coeficient de correlació entre aquestes variables.

Correlació indirecta

De vegades dues variables, X i Y, presenten un coeficient de correlació lineal alt entre elles, però aquesta relació és aparent o indirecta, ja que ambdues variables estan en realitat relacionades amb una tercera variable, Z. Per mesurar la vertadera relació entre X i Y es pot calcular el **COEFICIENT DE CORRELACIÓ PARCIAL**.



COEFICIENT DE CORRELACIÓ PARCIAL

És un coeficient de correlació lineal entre X i Y en el qual s'elimina la influència que exerceix una tercera variable Z sobre ambdues variables.

EXEMPLE

CA	X nombre de reclusos	Y nombre de biblioteques	Z població 2009
Andalusia	17495	869	8302923
Aragó	2644	374	1345473
Astúries	1547	159	1085289
Balears	1937	184	1095426
Canàries	3198	208	2103992
Cantàbria	724	71	589235
Castella-La Manxa	7021	453	2081313
Castella i Lleó	2227	609	2563521
Catalunya	10531	830	7475420
Comunitat Valenciana	8240	624	5094675
Extremadura	1408	501	1102410
Galícia	4904	550	2796089
Madrid	10515	513	6386932
Múrcia	967	129	1446520
Navarra	250	131	630578
País Basc	1472	323	2172175
Rioja	405	51	321702

Font: INE i Ministeri de l'Interior.

$r_{XY} = 0{,}816$ $\quad r_{XZ} = 0{,}945$ $\quad r_{YZ} = 0{,}849$

És real aquesta alta correlació positiva entre X i Y o hi ha una tercera variable Z (població 2009) que n'és la responsable? Calculem el coeficient de correlació parcial entre X i Y:

$$r_{XY}^{p} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{\left(1-r_{XZ}^2\right)\left(1-r_{YZ}^2\right)}}$$

$$= \frac{0{,}816 - 0{,}945 \times 0{,}849}{\sqrt{\left(1-0{,}945^2\right)\left(1-0{,}849^2\right)}} = 0{,}079$$

Si s'elimina la influència de la variable població (Z), gairebé no hi ha relació lineal entre el nombre de reclusos (X) i el nombre de biblioteques (Y).

3.- REGRESSIÓ LINEAL

Suposem que entre les variables X i Y hi ha **una relació de causa-efecte**. És a dir, una variable (la X) és la causa i l'altra (la Y) és l'efecte. Les variacions en X (la causa) provocaran variacions en Y (l'efecte).
Exemple: per a un conjunt de famílies, de les variables ingressos i despesa en turisme, quina serà X (la causa) i quina serà Y (l'efecte)?

Regressió Y/X (de Y respecte a X): és una funció matemàtica que ens explica els valors de la Y a partir dels valors de la X: Y= f(X).

- X serà la **variable independent** o explicativa.
- Y serà la **variable dependent** o explicada.

Utilitats de la regressió:

- Mesurar l'efecte que una variació (augment o disminució) en X provoca en Y.
- Fer prediccions per a la variable Y a partir de valors de X.

Model de regressió Y/X (de Y respecte a X): funció matemàtica que ens explica els valors de la Y a partir dels valors de la X: Y= f(X)

PROBLEMES DEL MODEL DE REGRESSIÓ:

1) Elegir una funció matemàtica que relacione ambdues variables.

Elegim una funció lineal (una recta) per → senzillesa / més s'usa

2) Quina és la recta que millor s'ajusta als punts del diagrama de dispersió?

Equació d'una recta: $Y^* = a + b\,X$

En definitiva, determinar els valors dels **coeficients *a* i *b*** de la recta de regressió.

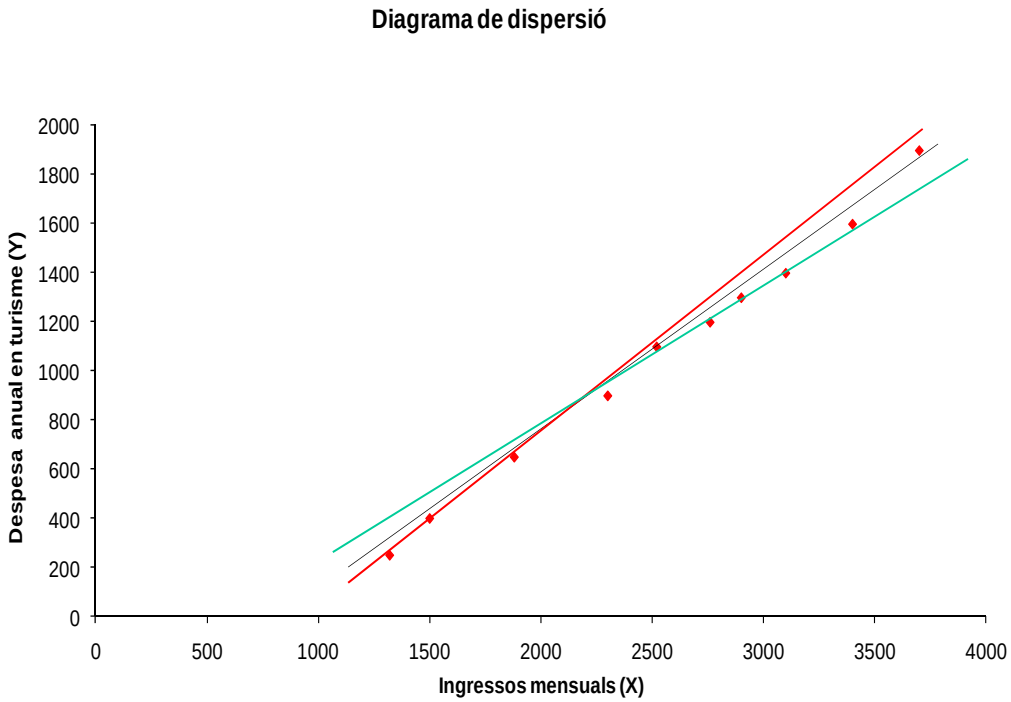
El **mètode minimoquadràtic** permet de determinar els valors dels coeficients *a* i *b* de la recta de regressió:

$$b = \frac{S_{XY}}{S_X^2} \qquad a = \bar{Y} - b\,\bar{X}$$

Interpretació dels coeficients *a* i *b* de la recta de regressió. Ho veurem amb un exemple.

•EXEMPLE:

Família	Ingresos mensuals en € (X)	Despesa anual en turisme en € (Y)
1	1880	650
2	2300	900
3	3700	1900
4	2760	1200
5	3400	1600
6	2900	1300
7	1320	250
8	1500	400
9	2520	1100
10	3100	1400



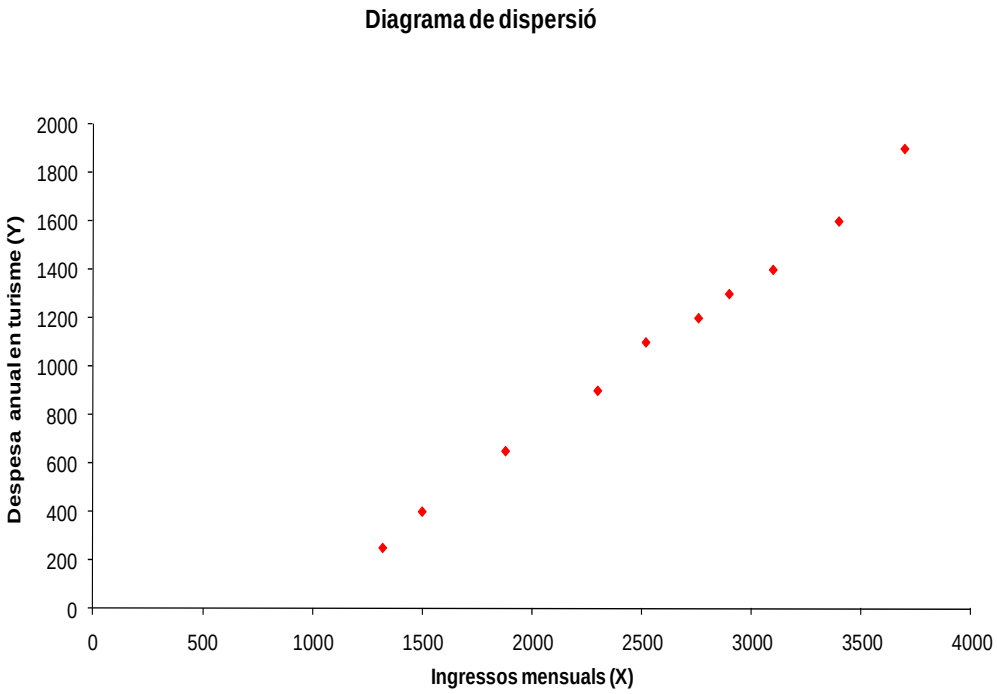
•EXEMPLE:

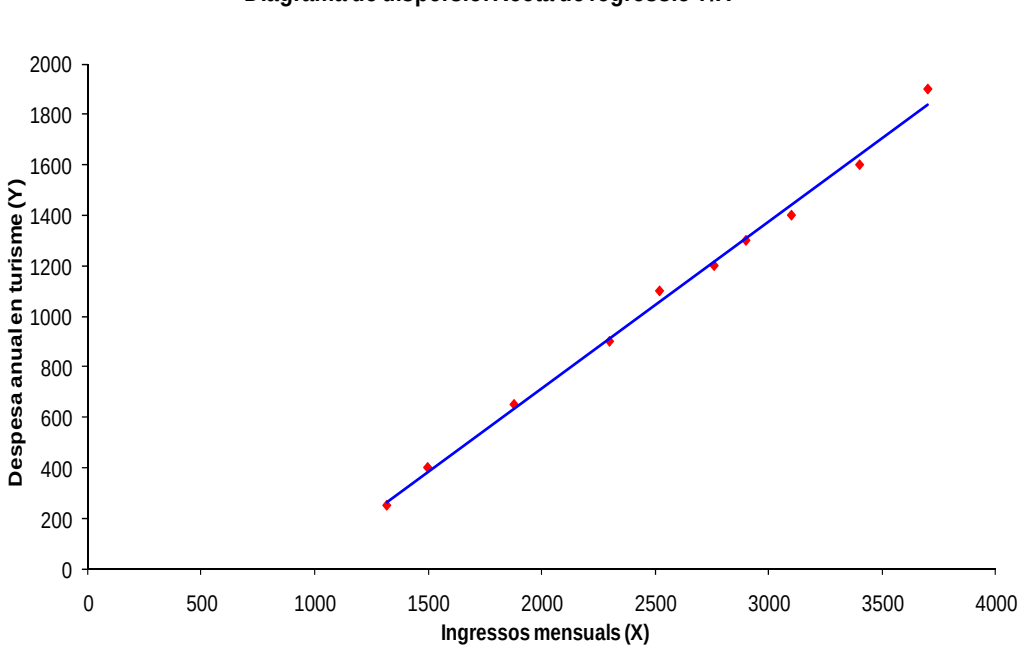
Família	Ingressos mensuals en € (X)	Despesa anual en turisme en € (Y)
1	1880	650
2	2300	900
3	3700	1900
4	2760	1200
5	3400	1600
6	2900	1300
7	1320	250
8	1500	400
9	2520	1100
10	3100	1400

Vector de Mitjanes: $\begin{pmatrix} \overline{X} = 2538 \\ \overline{Y} = 1070 \end{pmatrix}$

Matriu de Variàncies-Covariànces:

$$\begin{pmatrix} S_X^2 = 564036 & S_{XY} = 372940 \\ S_{XY} = 372940 & S_Y^2 = 247600 \end{pmatrix}$$





Matriu de Variàncies-Covariànces:

Vector de Mitjanes: $\begin{pmatrix} \bar{X} = 2538 \\ \bar{Y} = 1070 \end{pmatrix}$

$$\begin{pmatrix} S_X^2 = 564036 & S_{XY} = 372940 \\ S_{XY} = 372940 & S_Y^2 = 247600 \end{pmatrix}$$

$$b = \frac{S_{XY}}{S_X^2} = \frac{372940}{564036} = 0{,}661$$

$$a = \bar{Y} - b\,\bar{X} = 1070 - 0{,}661 \times 2538 = -607{,}618$$

El model de regressió lineal de Y/X és:

$$Y^* = -607{,}618 + 0{,}661\,X$$

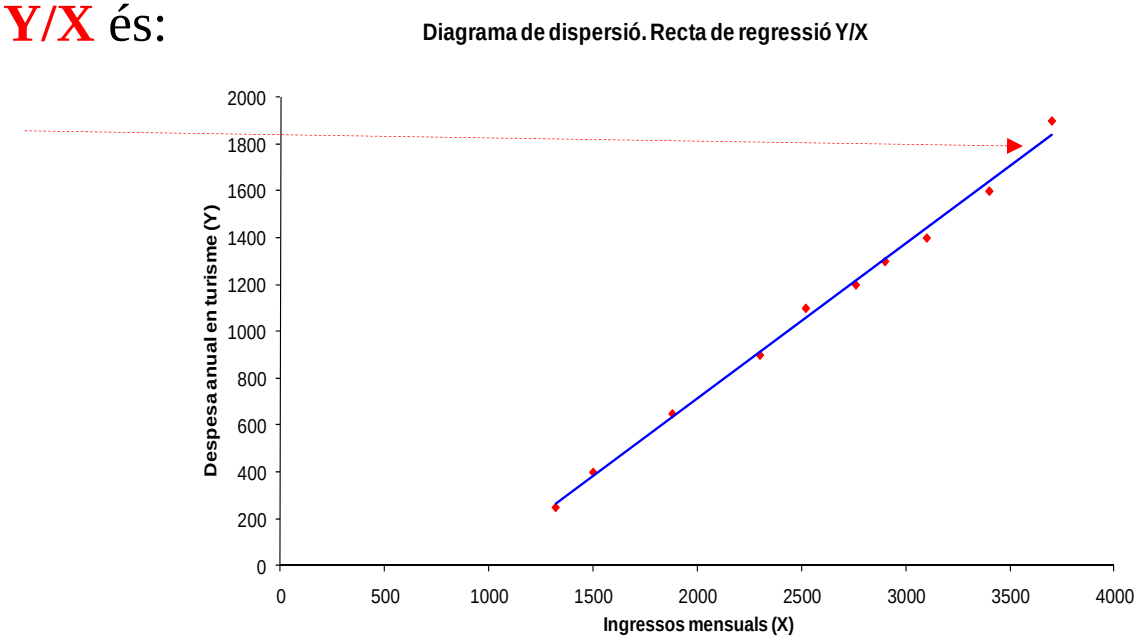
- Què és el coeficient a?

Si $X=0 \Rightarrow Y^* = -607{,}618$

Té sentit econòmic?

- Què és el coeficient b?

El pendent de la recta, què significa?



A banda de X i de Y, es creen dues noves variables:

• **Y***: la Y teòrica o explicada. Són els valors estimats de Y que ens proporciona el model de regressió lineal. La part dels valors de Y que podem explicar a partir dels valors de X.

$$Y^* = a + b\,X = -607{,}618 + 0{,}661\,X$$

• **e** : la variable error o residu. Són els errors que es cometen quan s'ajusta la recta de regressió. Allò que no explica el model de regressió.

$$e = Y - Y^*$$

Família	Ingressos mensuals en € (X)	Despesa anual en turisme en € (Y)	Y* teòrica Y*=a+bX	error e=Y-Y*
1	1880	650	635,1	14,9
2	2300	900	912,7	-12,7
3	3700	1900	1838,1	61,9
4	2760	1200	1216,7	-16,7
5	3400	1600	1639,8	-39,8
6	2900	1300	1309,3	-9,3
7	1320	250	264,9	-14,9
8	1500	400	383,9	16,1
9	2520	1100	1058,1	41,9
10	3100	1400	1441,5	-41,5
	∑ suma	10700	10700	0

4.- ANÀLISI DE LA BONDAT DE L'AJUST I PREDICCIÓ

	Y observada	Y* teòrica	e error
Mitjana	$\bar{Y}$	$\bar{Y}^* = \bar{Y}$	$\bar{e} = 0$
Variància	S_Y^2	$S_{Y^*}^2$	S_e^2

Variància explicada

Variància residual

Relació entre les tres variàncies: $S_Y^2 = S_{Y^*}^2 + S_e^2$

Coeficient de determinació: $R^2 = \dfrac{S_{Y^*}^2}{S_Y^2}$ $\qquad 0 \le R^2 \le 1$

- R^2 és la part de la variància de Y que explica el model de regressió.
- 1- R^2 és la part de la variància de Y que no explica el model, que es deu als errors que es cometen.

Propietat de la regressió lineal: $R^2 = \dfrac{S_{XY}^2}{S_X^2 \, S_Y^2} = r_{XY}^2$

Resultats de la CaEst 1.5:

Indicadors	Y	X	
Mitjana	1070	2538	
Variàncies i covariància	**247600**	564036	372940
Desv. típica	497,594	751,023	

REGRESSIÓ

C. correlació	0,998
C. determinació	0,996
Variància explicada	**246609,6**
Variància residual	**990,4**
Coeficient a	-607,618
Coeficient b	0,661
RECTA	Y*= -607,618 + 0,661X

Més informació sobre aquest tema en:

- PARRA, E; CALERO, F. J.: *Estadística para turismo*, Ed. McGraw-Hill, Madrid, 2007. Capítol 7.
- ESTEBAN, J. i altres: "Estadística descriptiva y nociones de probabilidad", Ed. Thomson, segona impressió 2006. Capítols 3 i 4.
- MONTIEL, A. M.; RIUS, F.; BARÓN F. J.: *Elementos básicos de estadística económica y empresarial*, Ed. Prentice Hall, Madrid, 1997. Capítols 5 i 6.
- RONQUILLO, A: *Estadística aplicada al sector turístico*, Ed Ramón Areces, Madrid, 1997. Capítol 6.
- http://www.uv.es/ceaces/descriptiva/simplem.htm
  http://www.uv.es/ceaces/base/regresion/simple.htm
- http://webpersonal.uma.es/de/J_SANCHEZ/Capitulo3.PDF