UNIDAD 11 DISTRIBUCIONES BIDIMENSIONALES

1. INTRODUCCIÓN

En la unidad anterior se ha estudiado un único carácter de cierta población. Sin embargo, a menudo es necesario realizar el estudio de dos o más caracteres diferentes simultáneamente de todos los individuos de una población o muestra considerada. En el caso de que se estudien dos caracteres a la vez se estará generando una distribución bidimensional.

Se va a generalizar el estudio unidimensional para el caso de dos caracteres. No obstante, van a aparecer una serie de conceptos nuevos como el grado de relación entre los caracteres estudiados o bien la necesidad de medir la intensidad con la que éstos puedan estar relacionados. Así, se podrá detectar si existe cierto tipo de dependencia o variación conjunta (covariación).

Aunque los caracteres pueden ser ambos cualitativos, cuantitativos o uno cualitativo y otro cuantitativo, en esta unidad nos vamos a ocupar de los de tipo cuantitativo.

Ejemplo: Se puede estudiar el número de libros que leen al año 20 alumnos de nuestro instituto tomados al azar y preguntarnos:

- ¿Hay alguna relación entre la edad del alumnado y el número de libros leídos durante el año?
- ¿Se puede establecer alguna relación funcional que relacione ambas variables estadísticas y que permita, por ejemplo, estimar el número de libros leídos por un estudiante de 17 años?
- En caso afirmativo, ¿qué confianza se puede tener en el resultado obtenido?

2. DISTRIBUCIONES DE FRECUENCIAS BIDIMENSIONALES

En adelante, se va a considerar una población de tamaño N en la cual se han estudiado **simultáneamente** dos caracteres cuantitativos que vienen representados por las variables estadísticas X e Y.

La variable estadística X presentará n valores $x_1, x_2, .., x_n$.

La variable estadística Y presentará m valores $y_1, y_2, .., y_m$.

Al par (X, Y) formado por el conjunto de pares de valores (x_i, y_j) correspondientes a esas dos variables estadísticas unidimensionales X e Y, se le llama **variable estadística bidimensional**.

En el caso de que alguna o ambas variables estadísticas estén agrupadas en intervalos o clases, x_i y/o y_j representarán la marca de clase correspondiente.

Distribución de frecuencias bidimensional: está formada por los pares de valores (x_i, y_j) de la variable estadística (X, Y) junto con sus correspondientes frecuencias absolutas (o relativas). Se organizan en la tabla estadística de frecuencias.

Tabla estadística de frecuencias para distribuciones bidimensionales.
La forma habitual mediante la cual se va a presentar una distribución bidimensional es una **tabla de doble entrada** conocida con el nombre de tabla estadística bidimensional o tabla de correlación.

Distribución conjunta de X e Y

X \ Y	y_1	y_2	…	y_j	…	y_m
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}
…	…	…	…	…	…	…
x_n	n_{n1}	n_{n2}	…	n_{nj}	…	n_{nm}

Frecuencia absoluta conjunta n_{ij}: número de veces que se presenta el par (x_i, y_j).

La suma de las frecuencias absolutas es igual a N:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij} = N$$

Frecuencia relativa conjunta f_{ij}: $f_{ij} = \dfrac{n_{ij}}{N}$

La suma de las frecuencias relativas es igual a 1:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = 1$$

Ejemplo: En una encuesta se pregunta a las familias por el número de personas que las componen (X) y el números de veces que suelen ir al cine mensualmente (Y). Se ha confeccionado la siguiente tabla de doble entrada con las respuestas obtenidas:

X \ Y	**1**	**2**	**3**	**5**
2	3	4	2	1
3	5	3	1	0
4	1	2	0	0

Así, por ejemplo:

$n_{12}=4$ indica el número de familias encuestadas con dos componentes que suelen ir dos veces al cine mensualmente.

$n_{21}=5$ indica el número de familias con tres componentes que van una vez.

$n_{34}=0$ indica que no hay ninguna familia con cuatro miembros que acostumbre a ir cinco veces al cine.

Por otro lado: $N=\sum_{i=1}^{n}\sum_{j=1}^{m}n_{ij}=22$ familias encuestadas.

$f_{21}=\frac{n_{21}}{N}=\frac{5}{22}\approx 0.227\Rightarrow$ El 22.7% de las familias encuestadas las forman tres miembros y van una vez al cine mensualmente.

Tabla de entrada simple o de datos apareados: se usa cuando el número de observaciones de la distribución bidimensional es pequeño. En estos casos es más cómodo usar este tipo de tablas.

Datos apareados

X	Y	n_i
x_1	y_1	n_1
x_2	y_2	n_2
…	…	…
x_i	y_i	n_i
…	…	…
x_k	y_k	n_k
		N

Ejemplo: Construimos la tabla de entrada simple del ejemplo anterior:

X	Y	n_i
2	1	3
2	2	4
2	3	2
2	5	1
3	1	5
3	2	3
3	3	1
4	1	1
4	2	2
		$N=22$

A partir de la tabla simple también se puede construir la tabla de doble entrada.

$n_4=1$ nos indica que hay una familia con dos miembros y que suele ir al cine cinco veces al mes.

$f_2=\frac{n_2}{N}=\frac{4}{22}\approx 0.1818\Rightarrow$ El 18.18% de las familias encuestadas las forman dos miembros y van dos veces al cine mensualmente.

Observación: Si las variables son de tipo cualitativo la tabla de doble entrada recibe el nombre de **tabla de contingencia**.

Ejemplo: A la salida de unos grandes almacenes se ha preguntado a 80 personas su estado civil y la información obtenida, diferenciada por sexos, se ha resumido en la siguiente tabla de contingencia:

Sexo / Estado Civil	**Hombre**	**Mujer**	**Total**
Soltero	12	15	**27**
Casado	23	20	**43**
Divorciado	6	4	**10**
Total	**41**	**39**	**80**

3. DISTRIBUCIONES MARGINALES Y CONDICIONADAS

Las primeras se van a generar al considerar aisladamente cada una de las variables que forman parte de la distribución bidimensional.

Las segundas son distribuciones unidimensionales que se generan al considerar una de las variables de la distribución bidimensional, pero suponiendo alguna condición sobre la otra variable.

3.1. DISTRIBUCIONES MARGINALES

Se estudia aisladamente cada una de las variables sin hacer ninguna referencia a los valores de la otra. Así, distinguimos la distribución marginal de X y la distribución marginal de Y.

Distribución conjunta de X e Y

X \ Y	y_1	y_2	…	y_j	…	y_m	$n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{1\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_n	n_{n1}	n_{n2}	…	n_{nj}	…	n_{nm}	$n_{n\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	N

Distr. Marginal de X

X	$n_{i\bullet}$
x_1	$n_{1\bullet}$
x_2	$n_{2\bullet}$
…	…
x_i	$n_{i\bullet}$
…	…
x_n	$n_{n\bullet}$
	N

Distr. Marginal de Y

Y	$n_{\bullet j}$
y_1	$n_{\bullet 1}$
y_2	$n_{\bullet 2}$
…	…
y_j	$n_{\bullet j}$
…	…
y_m	$n_{\bullet m}$
	N

Frecuencia absoluta marginal $n_{i\bullet}$: número de individuos de la población que han presentado la modalidad x_i de la variable X con independencia de las modalidades de la variable Y con las que se da conjuntamente.

$$n_{i\bullet} = \sum_{j=1}^{m} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{ij} + \ldots + n_{im} \qquad \text{Además se cumple que} \quad \sum_{i=1}^{n} n_{i\bullet} = N$$

Frecuencia absoluta marginal $n_{\bullet j}$: número de individuos de la población que han presentado la modalidad y_j de la variable Y con independencia de las modalidades de la variable X con las que se da conjuntamente.

$$n_{\bullet j} = \sum_{i=1}^{n} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{nj} \qquad \text{Además se cumple que} \quad \sum_{j=1}^{m} n_{\bullet j} = N$$

Frecuencia relativa marginal $f_{i\bullet}$: proporción de individuos de la población que han presentado la modalidad x_i de la variable X con independencia de las modalidades de la variable Y con las que se da conjuntamente.

$$f_{i\bullet} = \frac{n_{i\bullet}}{N} \qquad \text{Además se cumple que} \quad \sum_{i=1}^{n} f_{i\bullet} = 1$$

Frecuencia relativa marginal $f_{\bullet j}$: proporción de individuos de la población que han presentado la modalidad y_j de la variable Y con independencia de las modalidades de la variable X con las que se da conjuntamente.

$$f_{\bullet j} = \frac{n_{\bullet j}}{N} \qquad \text{Además se cumple que} \quad \sum_{j=1}^{m} f_{\bullet j} = 1$$

Se pueden definir y calcular todo tipo de características unidimensionales como parámetros de centralización (media marginal de X o de Y, medianas marginales, modas, cuartiles,…) de dispersión (varianzas marginales, desviaciones típicas,…) o cualquier otro aspecto estudiado en la unidad anterior.

Ejemplo: *En una encuesta de familias sobre el número de individuos que la componen (X) y el número de personas activas en ellas (Y) se han obtenido los resultados resumidos en la tabla de doble entrada. A partir de ella se han obtenido las distribuciones marginales de ambas variables.*

Distribución conjunta de *X* e *Y*

X \ Y	1	2	3	$n_{i\bullet}$
1	7	0	0	7
2	10	2	0	12
3	11	5	1	17
4	10	6	6	22
5	8	6	4	18
$n_{\bullet j}$	46	19	11	***N*=76**

Distr. Marginal de *X*

X	$n_{i\bullet}$
1	7
2	12
3	17
4	22
5	18
	***N*=76**

Distr. Marginal de *Y*

Y	$n_{\bullet j}$
1	46
2	19
3	11
	***N*=76**

a) Calcular el número medio de miembros de las familias encuestadas y el más frecuente, así como su mediana y desviación típica.

X	$n_{i\bullet}$	$N_{i\bullet}$	$x_i \cdot n_{i\bullet}$	$x_i^2 \cdot n_{i\bullet}$
1	7	7	7	7
2	12	19	24	48
3	17	36	51	153
4	22	58	88	352
5	18	76	90	450
	***N*=76**		**260**	**1010**

$$\bar{x} = \frac{\sum x_i \cdot n_{i\bullet}}{N} = \frac{260}{76} \approx 3.42 \qquad Mo_X = 4$$

$$Me_X = 4 \text{ puesto que } \frac{N}{2} = \frac{76}{2} = 38$$

$$\sigma_X^2 = \frac{\sum x_i^2 \cdot n_{i\bullet}}{N} - \bar{x}^2 = \frac{1010}{76} - 3.42^2 \approx 1.59$$

$$\sigma_X = \sqrt{1.59} \approx 1.26$$

b) Si consideramos el número de personas activas por familia, calcular Q_{1_Y}, Q_{3_Y} y su coeficiente de variación.

Y	$n_{\bullet j}$	$N_{\bullet j}$	$y_j \cdot n_{\bullet j}$	$y_j^2 \cdot n_{\bullet j}$
1	46	46	46	46
2	19	65	38	76
3	11	76	33	99
	***N*=76**		**117**	**221**

Cálculo de Q_{1_Y}, Q_{3_Y} :

$$\frac{N}{4} = \frac{76}{4} = 19 \Rightarrow Q_{1_Y} = 1$$

$$\frac{3N}{4} = \frac{3 \cdot 76}{4} = 57 \Rightarrow Q_{3_Y} = 2$$

Cálculo del coeficiente de variación:

$$C_{v_Y} = \frac{\sigma_Y}{\bar{y}} = \frac{0.73}{1.54} \approx 0.47$$

$$\bar{y} = \frac{\sum y_j \cdot n_{\bullet j}}{N} = \frac{117}{76} \approx 1.54$$

$$\sigma_Y^2 = \frac{\sum y_j^2 \cdot n_{\bullet j}}{N} - \bar{y}^2 = \frac{221}{76} - 1.54^2 \approx 0.54 \Rightarrow \sigma_Y = \sqrt{0.54} \approx 0.73$$

Observación: Los cálculos de las distribuciones marginales pueden realizarse en la tabla de doble entrada sin necesidad de obtener dos tablas independientes para las variables estadísticas *X* e *Y*.

3.2. DISTRIBUCIONES CONDICIONADAS

Distribución de X condicionada a que $Y = y_j$ ($X/Y = y_j$): se considera la variable X pero sus valores están condicionados a que la variable Y tome el valor y_j.

Distribución de Y condicionada a que $X = x_i$ ($Y/X = x_i$): se considera la variable Y pero sus valores están condicionados a que la variable X tome el valor x_i.

Distribución conjunta de X e Y

X \ Y	y_1	y_2	…	y_j	…	y_m	$n_{i\bullet}$
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	$n_{1\bullet}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	$n_{i\bullet}$
…	…	…	…	…	…	…	…
x_n	n_{n1}	n_{n2}	…	n_{nj}	…	n_{nm}	$n_{n\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	…	$n_{\bullet j}$	…	$n_{\bullet m}$	N

Distr. de X condicionada a $Y = y_j$

$X/Y = y_j$	n_{ij}
x_1	n_{1j}
x_2	n_{2j}
…	…
x_i	n_{ij}
…	…
x_n	n_{nj}
	$n_{\bullet j}$

Distr. de Y condicionada a $X = x_i$

$Y/X = x_i$	n_{ij}
y_1	n_{i1}
y_2	n_{i2}
…	…
y_j	n_{ij}
…	…
y_m	n_{im}
	$n_{i\bullet}$

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}} \Rightarrow f_{ij} = f_{i/j} \cdot f_{\bullet j} \qquad f_{j/i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}} \Rightarrow f_{ij} = f_{j/i} \cdot f_{i\bullet}$$

De la misma manera que con las distribuciones marginales, se pueden definir y calcular parámetros de centralización o dispersión condicionados tales como medias, medianas, modas, cuartiles, varianzas, desviaciones típicas,… o cualquier otro aspecto estudiado en la unidad anterior.

Ejemplo: *Si consideramos el ejemplo anterior del punto 3.1. en el que se consideraba el número de personas que componen cada familia (X) y el número de personas activas en ellas (Y).*

a) Obtener las siguientes distribuciones condicionadas:

- *Distribución del número de familias con dos de sus miembros en activo (X/Y=2).*
- *Distribución del número de miembros activos en las familias de tres miembros (Y/X=3).*

Distribución conjunta de X e Y

X \ Y	**1**	**2**	**3**	$n_{i\bullet}$
1	7	0	0	7
2	10	2	0	12
3	11	5	1	17
4	10	6	6	22
5	8	6	4	18
$n_{\bullet j}$	46	19	11	**N=76**

Distribución de X condicionada a $Y = 2$

$X/Y = 2$	n_{i2}
1	0
2	2
3	5
4	6
5	6
	$n_{\bullet 2}$=19

Distribución de Y condicionada a $X = 3$

$Y/X = 3$	n_{3j}
1	11
2	5
3	1
	$n_{3\bullet}$=17

b) En el grupo de familias donde las personas activas son 3, calcular el número medio y más frecuente de miembros que las forman. Calcular también su coeficiente de variación.

$X/Y = 3$	n_{i3}	N_{i3}	$x_i \cdot n_{i3}$	$x_i^2 \cdot n_{i3}$
1	0	0	0	0
2	0	0	0	0
3	1	1	3	9
4	6	7	24	96
5	4	11	20	100
	$n_{\bullet 3}$=11		**47**	**205**

$$\bar{x}_{/Y=3} = \frac{\sum x_i \cdot n_{i3}}{n_{\bullet 3}} = \frac{47}{11} \approx 4.273 \qquad Mo_{X/Y=3} = 4$$

$$\sigma^2_{X/Y=3} = \frac{\sum x_i^2 \cdot n_{i3}}{n_{\bullet 3}} - \bar{x}^2_{/Y=3} = \frac{205}{11} - 4.273^2 \approx 0.38$$

$$\sigma_{X/Y=3} = \sqrt{0.38} \approx 0.617$$

$$C_{v_{X/Y=3}} = \frac{\sigma_{X/Y=3}}{\bar{x}_{/Y=3}} = \frac{0.617}{4.273} \approx 0.144$$

c) Calcular el número medio de miembros activos y el más frecuente en las familias formadas por 5 personas. Calcular también su mediana y su varianza.

$Y/X=5$	n_{5j}	N_{5j}	$y_j \cdot n_{5j}$	$y_j^2 \cdot n_{5j}$
1	8	8	8	8
2	6	14	12	24
3	4	18	12	36
	$n_{5\bullet}=18$		**32**	**68**

$$\bar{y}_{/X=5}=\frac{\sum y_j \cdot n_{5j}}{n_{5\bullet}}=\frac{32}{18}\approx 1.778 \quad Mo_{Y/X=5}=1$$

$$Me_{Y/X=5}=2 \text{ puesto que } \frac{n_{5\bullet}}{2}=\frac{18}{2}=9$$

$$\sigma^2_{Y/X=5}=\frac{\sum y_j^2 \cdot n_{5j}}{n_{5\bullet}}-\bar{y}^2_{/X=5}=\frac{68}{18}-1.778^2\approx 0.62$$

3.3. COVARIANZA

Dada una distribución bidimensional (X,Y), se define su **covarianza** o **varianza conjunta** de las variables X e Y como:

$$\sigma_{XY}=Cov(X,Y)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i-\bar{x})\cdot(y_j-\bar{y})\cdot n_{ij}}{N}$$

La covarianza nos proporciona información acerca del grado de dependencia lineal existente ente las variables. Su signo nos proporciona el sentido de esa variación conjunta.

- Si es positivo las dos variables varían en el mismo sentido, es decir, si una aumenta la otra también.
- Si es negativo lo harán en sentido opuesto, es decir, si una aumenta la otra disminuirá.

Es fácil demostrar que esta definición es equivalente a esta otra más operativa:

$$\sigma_{XY}=Cov(X,Y)=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}x_i \cdot y_j \cdot n_{ij}}{N}-\bar{x}\cdot\bar{y}$$

Para tablas de datos apareados: $\sigma_{XY}=\dfrac{\sum x_i \cdot y_i \cdot n_i}{N}-\bar{x}\cdot\bar{y}$

Si además la frecuencia absoluta conjunta de las observaciones es 1: $\sigma_{XY}=\dfrac{\sum x_i \cdot y_i}{N}-\bar{x}\cdot\bar{y}$

Ejemplo 1: *Tabla de datos apareados o tabla simple.*
Para calcular $\bar{x}$ e $\bar{y}$ se pueden obtener las distribuciones marginales o se puede operar directamente en esta tabla:

X	Y	n_i	$x_i \cdot n_i$	$y_i \cdot n_i$	$x_i \cdot y_i \cdot n_i$
2	1	3	6	3	6
2	2	4	8	8	16
2	3	2	4	6	12
2	5	1	2	5	10
3	1	5	15	5	15
3	2	3	9	6	18
3	3	1	3	3	9
4	1	1	4	1	4
4	2	2	8	4	16
		$N=22$	59	41	106

$$\bar{x}=\frac{\sum x_i \cdot n_i}{N}=\frac{59}{22}\approx 2.68$$

$$\bar{y}=\frac{\sum y_i \cdot n_i}{N}=\frac{41}{22}\approx 1.86$$

$$\sigma_{XY}=\frac{\sum x_i \cdot y_i \cdot n_i}{N}-\bar{x}\cdot\bar{y}$$

$$\Rightarrow \sigma_{XY}=\frac{106}{22}-2.68\cdot 1.86\approx -0.17$$

Ejemplo 2: *Tabla de datos apareados con frecuencia absoluta 1.*

X	Y	$x_i \cdot y_i$
2	4	8
3	9	27
4	16	64
5	24	120
6	30	180
7	34	238
8	38	304
9	42	378
44	197	1319

En este caso es claro que $N = 8$.

$$\bar{x} = \frac{44}{8} = 5.5$$

$$\bar{y} = \frac{197}{8} = 24.625$$

$$\sigma_{XY} = \frac{\sum x_i \cdot y_i}{N} - \bar{x} \cdot \bar{y} = \frac{1319}{8} - 5.5 \cdot 24.625 = 29.4375$$

Ejemplo 3: *Tabla de datos de doble entrada.*

Si es factible se puede hacer una tabla simple y proceder como en el ejemplo 2, obtener las distribuciones marginales para el cálculo de $\bar{x}$ e $\bar{y}$ o bien actuar del siguiente modo:

X \ Y	**1**	**2**	**3**	$n_{i\bullet}$	$x_i \cdot n_{i\bullet}$
1	3	0	2	5	5
2	5	3	0	8	16
3	2	3	4	9	27
4	0	1	5	6	24
$n_{\bullet j}$	10	7	11	**N=28**	**72**
$y_j \cdot n_{\bullet j}$	10	14	33	**57**	

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i \cdot n_{i\bullet}}{N} = \frac{72}{28} \approx 2.57$$

$$\bar{y} = \frac{\sum_{j=1}^{3} y_j \cdot n_{\bullet j}}{N} = \frac{57}{28} \approx 2.04$$

$$\sigma_{XY} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{3} x_i \cdot y_j \cdot n_{ij}}{N} - \bar{x} \cdot \bar{y} = \frac{1\cdot1\cdot3+1\cdot3\cdot2+2\cdot1\cdot5+2\cdot2\cdot3+3\cdot1\cdot2+3\cdot2\cdot3+3\cdot3\cdot4+4\cdot2\cdot1+4\cdot3\cdot5}{28} - 2.57\cdot2.04 =$$

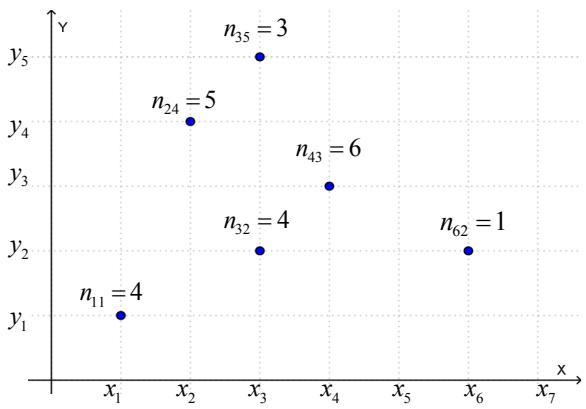
$$= \frac{159}{28} - 5.2428 \approx 0.44$$

4. *REPRESENTACIONES GRÁFICAS*

Son distintas según la naturaleza de los caracteres considerados. En esta unidad se van a estudiar el diagrama de dispersión también llamado nube de puntos y el estereograma. Ambos corresponden a caracteres cuantitativos.

4.1. DIAGRAMA DE DISPERSIÓN O NUBE DE PUNTOS

Dada una distribución bidimensional (X, Y), en un sistema de coordenadas cartesiano se representan en el eje de abscisas los valores de la variable X y en el eje de ordenadas los de la variable Y. Al representar por medio de un punto cada par observado (x_i, y_j) aparecerán tantos puntos como parejas de valores hayan sido observadas. A la gráfica obtenida se le llama **diagrama de dispersión** o **nube de puntos**.



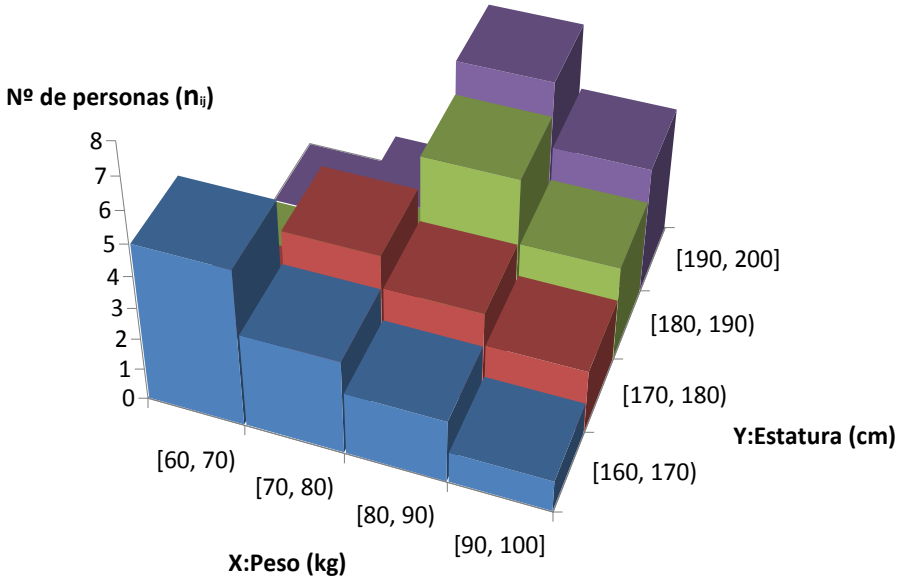
Esta representación gráfica es muy útil pues nos da una idea de la relación existente entre las variables como veremos en los puntos posteriores de esta unidad.

La frecuencia de cada pareja se puede poner de relieve escribiendo al lado de cada punto el número de veces que la pareja se presenta en la población. También se puede expresar mediante un círculo de centro el punto y superficie proporcional a n_{ij}.

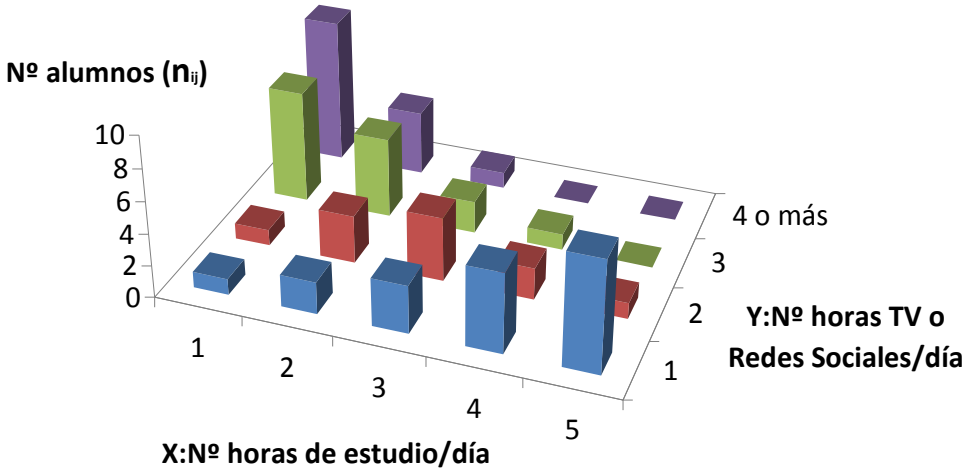
Si los datos están agrupados se toman las marcas de clase.

4.2. ESTEREOGRAMA

Es la generalización del histograma de frecuencias a una distribución estadística bidimensional. Se usa con caracteres de tipo cuantitativo y variables de tipo continuo como el de este gráfico:



Si las variables son de tipo discreto o en el caso de caracteres de tipo cualitativo, se obtendrá un **diagrama de barras tridimensional** como el siguiente:



5. *INDEPENDENCIA Y DEPENDENCIA ESTADÍSTICA*

Al estudiar dos variables estadísticas simultáneamente puede que exista algún tipo de relación entre ellas o, por el contrario, que no exista ningún tipo de relación. En caso de dependencia, esta puede ser perfecta o bien pueden existir ciertos grados de relación entre las variables consideradas.

Por tanto, se distingue:

Dependencia funcional: existe una función matemática que transforma los valores de una variable estadística en los valores de la otra. Así, la dependencia será perfecta y permitirá predecir los valores de una variable conocidos los de la otra.

Ejemplos: Distancia entre dos ciudades en un mapa y distancia en la realidad.
Número de agricultores y tiempo empleado en recoger la cosecha.

Dependencia estadística: existe cierto grado de relación entre los valores que toman las variables estadísticas pero no se puede formalizar matemáticamente mediante una función.

Ejemplos: Extensión en km^2 de una comunidad autónoma y número de habitantes de ésta.
Número de horas de estudio y número de asignaturas aprobadas.

Independencia estadística: ausencia de relación entre variables.

Ejemplos: Altura de una persona y el número de televisores que tiene en casa.
Número de hermanos de una persona y su altura.

Independencia	**Dependencia Estadística**	**Dependencia Funcional**
↓	↓	↓
Ausencia de relación	Hay relación pero no funcional	Relación perfecta Función matemática

Propiedades:

a) X e Y independientes $\Leftrightarrow$ $f_{ij} = f_{i\bullet} \cdot f_{\bullet j}$

b) X e Y independientes $\Rightarrow \begin{cases} f_{i/j=1} = f_{i/j=2} = f_{i/j=3} = \ldots = f_{i\bullet} \\ f_{j/i=1} = f_{j/i=2} = f_{j/i=3} = \ldots = f_{\bullet j} \end{cases}$

Es decir, las frecuencias relativas condicionadas coinciden con su correspondiente frecuencia relativa marginal.

6. *REGRESIÓN*

Objetivo: sustituir la dependencia de tipo estadístico existente entre dos variables estadísticas por una dependencia de tipo funcional que permita predecir, de forma aproximada, los valores de una de las variables a partir de los valores de la otra que se supone conocida. El proceso que permite determinar la función que mejor pueda explicar el comportamiento de una variable estadística conocidos los valores de la otra recibe el nombre de **regresión**.

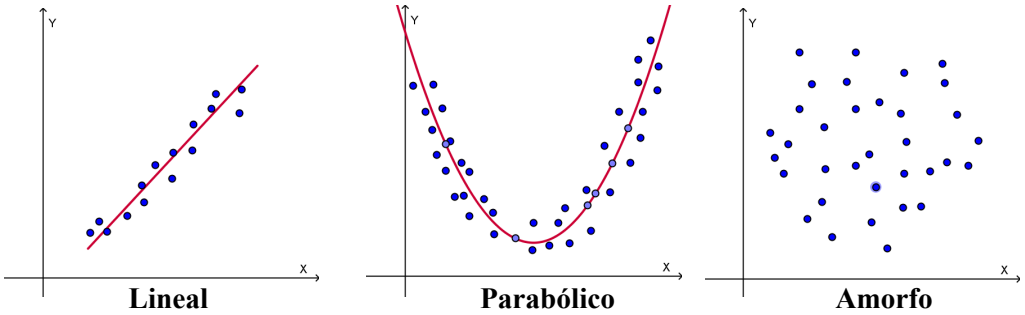
El proceso se llevará a cabo en dos fases:

1º) *Especificación del modelo elegido:* será necesario representar el diagrama de dispersión, observar el grado de dependencia estadística que existe entre las variables y elegir una línea curva que se adapte lo mejor posible a la nube de puntos.

Sir Francis Galton y la regresión

Naturalista británico (1822-1911). Encontró una fuerte correlación entre las variables:
- Estatura media de un matrimonio.
- Estatura media alcanzada por sus hijos.

Sin embargo, los hijos de padres de gran estatura tendían a igualarse al valor medio y lo mismo ocurría con los de muy baja estatura. Dicho de otro modo, estos hijos tendían a "regresar" a la altura media de la población. A partir de entonces se usa el término "**regresión**" para indicar la relación existente entre dos variables estadísticas.



Puede ocurrir que la nube de puntos sea amorfa como en el último diagrama de dispersión y por tanto no se pueda encontrar la relación funcional buscada. Es el caso de variables estadísticas independientes y diremos que están incorreladas.

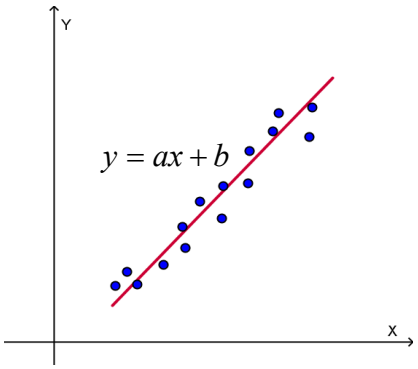
2º) *Determinación de los parámetros del modelo elegido:* una vez elegida la función analítica que mejor se adapte a la nube de puntos, habrá que calcular los parámetros desconocidos que la caracterizan. Para ello se usa, en general, un método de ajuste llamado ajuste por mínimos cuadrados que minimiza el error cuadrático medio, es decir, la media de los residuos al cuadrado. El residuo o error se define como la diferencia entre los valores observados y los ajustados por la función de regresión.

En esta unidad didáctica nos vamos a ocupar únicamente del modelo de regresión lineal.

6.1. REGRESIÓN LINEAL

Las ecuaciones lineales son ajustadas con mucha frecuencia pues presentan buenas propiedades matemáticas y suelen explicar de modo adecuado muchos fenómenos.

Rectas de regresión: Dada una variable aleatoria bidimensional (X, Y) y representada su nube de puntos, observamos que la mejor función que explica el comportamiento de la variable Y respecto a X es una recta como observamos en la siguiente figura:



Obteniendo el valor de los parámetros a y b por el método de ajuste mínimo cuadrático se obtiene:

$$r_{Y/X}:\ y-\bar{y}=\frac{\sigma_{XY}}{\sigma_X^2}(x-\bar{x})$$

↑

Recta de regresión de *Y* sobre *X*

De igual forma, si el mejor modelo que explica el comportamiento de la variable X a partir de Y es una función lineal $y = a'x + b'$, se obtiene:

$$r_{X/Y}:\ x-\bar{x}=\frac{\sigma_{XY}}{\sigma_y^2}(y-\bar{y}) \leftarrow \textbf{Recta de regresión de } \boldsymbol{X} \textbf{ sobre } \boldsymbol{Y}$$

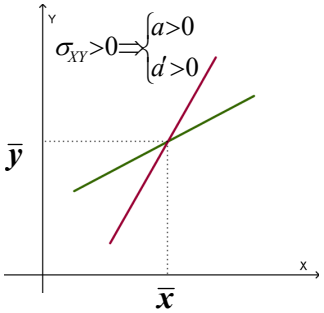
Observación: La recta de regresión de X sobre Y NO se obtiene despejando x de la recta de regresión de Y sobre X.

Coeficientes de regresión lineal: son las pendientes de las rectas de regresión, es decir:

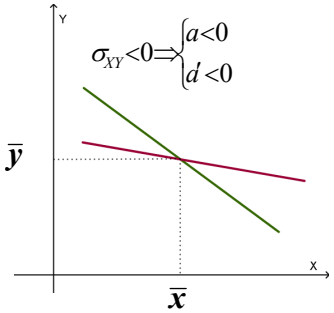
$$a=\frac{\sigma_{XY}}{\sigma_X^2} \qquad a'=\frac{\sigma_{XY}}{\sigma_Y^2}$$

Los coeficientes de regresión tienen el mismo signo y coincide con el signo de la covarianza σ_{XY}, con lo cual ambas van a ser crecientes o decrecientes.

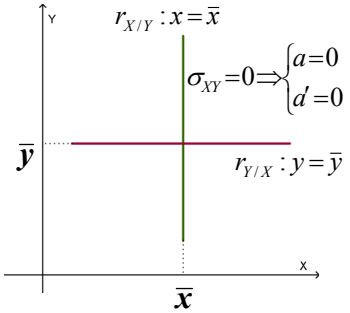
Si $\sigma_{XY}=0 \Rightarrow a=a'=0$ (pendiente nula) las dos rectas de regresión serán perpendiculares, cada una de ellas paralela a un eje de coordenadas, $r_{Y/X}: y=\bar{y}$; $\quad r_{X/Y}: x=\bar{x}$.



Crecientes



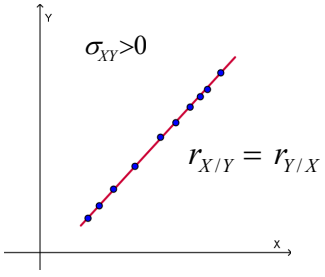
Decrecientes

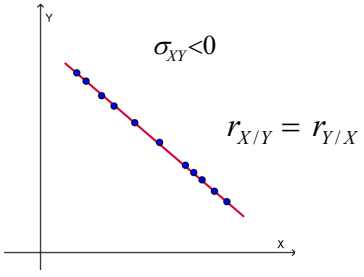


Perpendiculares

Centro de gravedad: ambas rectas pasan por el punto $(\bar{x}, \bar{y})$ llamado centro de gravedad de la distribución. Podemos observar este hecho en las figuras anteriores.

Caso de dependencia funcional: en este caso ambas rectas de regresión coinciden y se ajustan perfectamente a la nube de puntos. Es el caso de dependencia funcional perfecta:



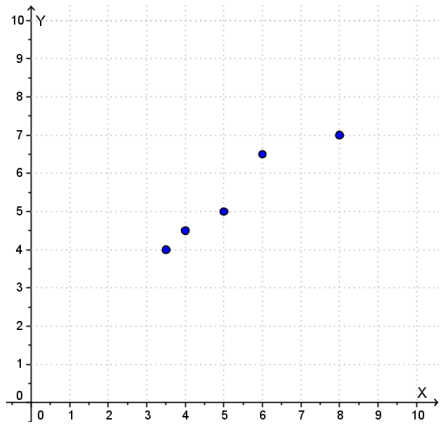


Ejemplo: *Las notas obtenidas por cinco alumnos en Matemáticas y Física son:*

Matemáticas (X)	6	4	8	5	3.5
Física (Y)	6.5	4.5	7	5	4

a) Representar la nube de puntos asociada a la tabla.
b) Obtener ambas rectas de regresión.
c) Calcular la nota esperada en Física para un alumno que tiene 7.5 en matemáticas.

a) La nube de puntos de la distribución es:



Al aumentar la nota en matemáticas aumenta la nota en Física.

Por tanto, la correlación es positiva.

b) Se calculan previamente los parámetros necesarios:

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
6	6.5	36	42.25	39
4	4.5	16	20.25	18
8	7	64	49	56
5	5	25	25	25
3.5	4	12.25	16	14
26.5	27	153.25	152.5	152

$N = 5$

$\bar{x} = \frac{26.5}{5} = 5.3 \qquad \bar{y} = \frac{27}{5} = 5.4$

$\sigma_X^2 = \frac{153.25}{5} - 5.3^2 = 2.56 \qquad \sigma_Y^2 = \frac{152.5}{5} - 5.4^2 = 1.34$

$\sigma_{XY} = \frac{152}{5} - 5.3 \cdot 5.4 = 1.78$

Recta de regresión de Y sobre X:

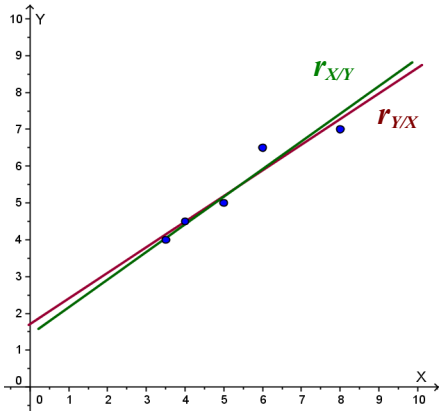
$$r_{Y/X} : y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x}) \Rightarrow y - 5.4 = \frac{1.78}{2.56}(x - 5.3) \Rightarrow \boldsymbol{r_{Y/X} : y = 0.695x + 1.715}$$

Recta de regresión de X sobre Y:

$$r_{X/Y} : \; x - \bar{x} = \frac{\sigma_{XY}}{\sigma_y^2}(y - \bar{y}) \Rightarrow x - 5.3 = \frac{1.78}{1.34}(y - 5.4) \Rightarrow \boldsymbol{r_{X/Y} : x = 1.328y - 1.873}$$

c) Como $x = 7.5 \Rightarrow y = 0.695 \cdot 7.5 + 1.715 \Rightarrow y = 6.9275 \Rightarrow$ La nota esperada en Física es **6.9.**

Observa la representación conjunta de la nube de puntos y ambas rectas de regresión.



Fíjate: La correlación lineal entre ambas variables es casi perfecta ya que las rectas están muy próximas.

Casi coinciden y el ángulo que forman se acerca mucho a cero.

Para determinar la fiabilidad de la predicción se van a definir en el siguiente punto parámetros que midan cuantitativamente esa fiabilidad.

7. CORRELACIÓN

Correlación: Grado de dependencia mutua existente entre las variables que intervienen en una distribución bidimensional.

Objetivo: Determinar una serie de medidas que cuantifiquen la intensidad con la que las variables puedan estar relacionadas según una determinada función de regresión. Por tanto, se estará midiendo la bondad del ajuste de la función de regresión elegida, es decir, en qué grado la elección de la función de regresión escogida es adecuada.

Nuestro estudio se va a centrar en un modelo de regresión lineal y para ello vamos a definir dos medidas:

1ª) Coeficiente de correlación lineal de Pearson r :

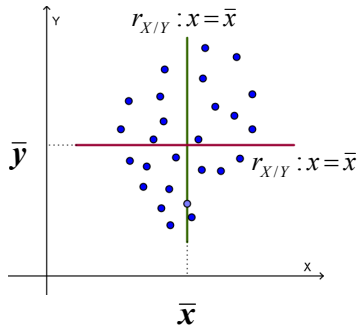
$$\boxed{r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}} \qquad -1 \le r \le 1$$

El signo de r depende del que tenga la covarianza.

Observación: r coincide para las dos rectas de regresión, es decir, cuantifica sin distinción la regresión lineal de *Y/X* así como la de *X/Y*.

Interpretación del valor de r :

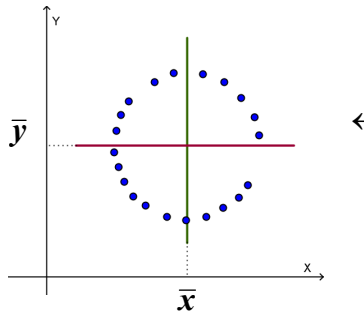
- Si $r = 0 \Rightarrow$ Correlación lineal nula. Son independientes según una recta.



Fíjate: Si $r = 0 \Rightarrow \sigma_{XY} = 0 \Rightarrow \left.\begin{matrix} y - \bar{y} = 0 \\ x - \bar{x} = 0 \end{matrix}\right\} \Rightarrow \left.\begin{matrix} r_{Y/X} : y = \bar{y} \\ r_{X/Y} : x = \bar{x} \end{matrix}\right\}$

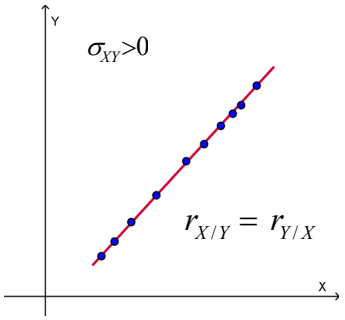
Ambas rectas de regresión forman un ángulo de 90º.

Observa: Que $r = 0$ no implica que las variables sean independientes ya que pueden estar relacionadas mediante otro tipo de curva.



← X e Y no son independientes. Están incorreladas linealmente, pero existe una función circular que las relaciona.

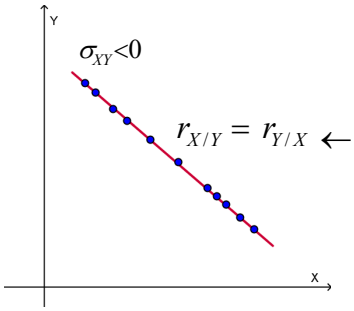
- Si $r = 1 \Rightarrow$ Correlación lineal perfecta positiva o directa (dependencia funcional).



← Las rectas de regresión coinciden, el ajuste con la nube de puntos es perfecto y como $\sigma_{XY} > 0$ son crecientes.

Ambas rectas de regresión forman un ángulo de 0º.

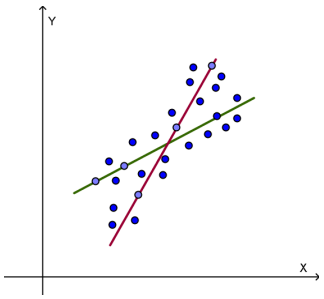
- Si $r = -1 \Rightarrow$ Correlación lineal perfecta negativa o inversa (dependencia funcional).



Las rectas de regresión coinciden, el ajuste con la nube de puntos es perfecto y como $\sigma_{XY} < 0$ son decrecientes.

Ambas rectas de regresión forman un ángulo de 0º.

- Si $0 < r < 1 \Rightarrow$ Cierto grado de correlación lineal positiva (las variables crecen en el mismo sentido), mayor cuanto más se acerque a 1 y más débil cuanto más se acerque a cero.

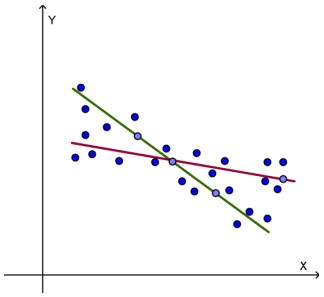


Si $r = 0.90 \Rightarrow$ Correlación positiva muy fuerte.
Si $r = 0.75 \Rightarrow$ Correlación positiva considerable.
Si $r = 0.50 \Rightarrow$ Correlación positiva media.
Si $r = 0.10 \Rightarrow$ Correlación positiva débil.

$\boldsymbol{r}$ informa del ángulo que determinan las rectas de regresión.

Será tanto mayor cuanto más próximo a cero esté $\boldsymbol{r}$ y menor cuanto más próximo a 1.

- Si $-1 < r < 0 \Rightarrow$ Cierto grado de correlación lineal negativa (las variables crecen en sentido opuesto), mayor cuanto más se acerque a -1 y más débil cuanto más se acerque a cero.



Si $r = -0.90 \Rightarrow$ Correlación negativa muy fuerte.
Si $r = -0.75 \Rightarrow$ Correlación negativa considerable.
Si $r = -0.50 \Rightarrow$ Correlación negativa media.
Si $r = -0.10 \Rightarrow$ Correlación negativa débil.

En este caso, el ángulo será de nuevo tanto mayor cuanto más se aproxime $\boldsymbol{r}$ a cero y menor cuanto más lo haga a -1.

2ª) Coeficiente de determinación (razón de correlación lineal) $\boldsymbol{r^2}$:

$$r^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2} \qquad 0 \le r^2 \le 1$$

Observaciones:

$\boldsymbol{r^2}$ coincide para las dos rectas de regresión, es decir, cuantifica sin distinción la regresión lineal de *Y/X* así como la de *X/Y*.

Expresión de en función de los coeficientes de regresión:

$$r^2 = a \cdot a' \text{ siendo } a = \frac{\sigma_{XY}}{\sigma_X^2} \text{ y } a' = \frac{\sigma_{XY}}{\sigma_Y^2} \text{ los coeficientes de regresión.}$$

Interpretación del valor de $\boldsymbol{r^2}$:

- Si $r^2 = 0 \Rightarrow$ Correlación lineal nula. Son independientes según una recta.
- Si $r^2 = 1 \Rightarrow$ Correlación lineal perfecta (ajuste perfecto). Dependencia funcional.
- Si $0 < r < 1 \Rightarrow$ Existe cierto grado de correlación lineal entre las variables sin distinguir si la variable explicada ha sido *X* o *Y*.

Observaciones:

- Para medir la bondad del ajuste entre dos variables por un modelo de regresión lineal suele usarse con mayor frecuencia el coeficiente de determinación lineal r^2, puesto que aporta la proporción (si se multiplica por 100 indica el porcentaje) de variación de una variable debido a la variación de la otra. Dicho de otro modo, indica la proporción de Y explicada por X y viceversa.
- Se considerará buena la aproximación si al calcular $r^2 \cdot 100$, el porcentaje obtenido es mayor al 75%, muy buena a partir de un 90% y un ajuste poco fiable si es inferior al 60%.
- Se usará el coeficiente de correlación lineal r cuando se quiera obtener el sentido de variación entre ambas variables estadísticas.
- A más datos utilizados la fiabilidad aumenta. Si se usan pocos datos (pocos puntos en la nube) el riesgo será grande a pesar de que los valores de r o r^2 nos indiquen lo contrario .
- La fiabilidad de un valor estimado y_0 de Y a partir de otro x_0 de X será mayor si el valor x_0 está próximo a $\overline{x}$. Si x_0 se aleja de $\overline{x}$ el riesgo en la estimación aumenta.

Ejemplo 1: (Ejemplo del punto anterior)
Las notas obtenidas por cinco alumnos en Matemáticas y Física son:

Matemáticas (X)	6	4	8	5	3.5
Física (Y)	6.5	4.5	7	5	4

Obtener el valor de r y r^2. Interpretar su significado.

Ya se obtuvieron los valores de los parámetros:

$$\overline{x} = \frac{26.5}{5} = 5.3 \quad \overline{y} = \frac{27}{5} = 5.4$$

$$\sigma_X^2 = \frac{153.25}{5} - 5.3^2 = 2.56 \quad \sigma_Y^2 = \frac{152.5}{5} - 5.4^2 = 1.34 \quad \sigma_{XY} = \frac{152}{5} - 5.3 \cdot 5.4 = 1.78$$

Por tanto:

$$\sigma_X = \sqrt{2.56} = 1.6 \quad \sigma_Y = \sqrt{1.34} = 1.158 \Rightarrow r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{1.78}{1.6 \cdot 1.158} = 0.961 \Rightarrow \mathbf{r = 0.961}$$

Es decir, el coeficiente de correlación lineal nos dice que existe una correlación positiva muy fuerte entre la calificación de ambas materias. Hay alto grado de fiabilidad.

Calculamos el coeficiente de determinación:

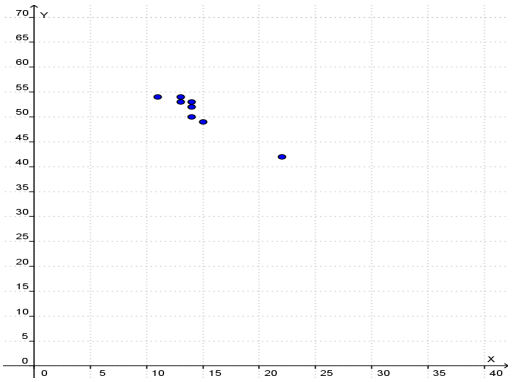
$$r^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \cdot \sigma_Y^2} = \frac{1.78^2}{2.56 \cdot 1.34} = 0.9236 \Rightarrow \mathbf{r^2 = 0.9236} \Rightarrow$$ El 92.36% de la nota en Física es explicada por la nota en Matemáticas (y viceversa).

Ejemplo 2: *Una variable bidimensional viene dada por la siguiente tabla:*

X	13	14	11	13	14	14	15	22
Y	54	52	54	53	53	50	49	42

a) Dibujar la nube de puntos asociada a la tabla, su centro de gravedad, las varianzas de X e Y y su covarianza.
b) Obtener ambas rectas de regresión.
c) Calcular su coeficiente de correlación lineal y el coeficiente de determinación.
d) Si x=12, ¿qué valor se espera que tome y?
e) ¿Es fiable esta predicción? Justifícalo

a) La nube de puntos de la distribución es: ⟶



Al aumentar la variable estadística X disminuye la variable Y. Por tanto la correlación es negativa.

Se realizan los cálculos necesarios para obtener los parámetros pedidos:

x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
13	54	169	2916	702
14	52	196	2704	728
11	54	121	2916	594
13	53	169	2809	689
14	53	196	2809	742
14	50	196	2500	700
15	49	225	2401	735
22	42	484	1764	924
116	407	1756	20819	5814

$N = 8 \quad \bar{x} = \frac{116}{8} = 14.5 \qquad \bar{y} = \frac{407}{8} = 50.875$

Centro de gravedad: $(\bar{x}, \bar{y}) = (\mathbf{14.5, 50.875})$

$\sigma_X^2 = \frac{1756}{8} - 14.5^2 = 9.25$

$\sigma_Y^2 = \frac{20819}{8} - 50.875^2 \approx 14.109$

$\sigma_{XY} = \frac{5814}{8} - 14.5 \cdot 50.875 = -10.9375 \Rightarrow \boldsymbol{\sigma_{XY} = -10.9375}$

b) Recta de regresión de Y sobre X:

$$r_{Y/X} : y - \bar{y} = \frac{\sigma_{XY}}{\sigma_X^2}(x - \bar{x}) \Rightarrow y - 50.875 = \frac{-10.9375}{9.25}(x - 14.5) \Rightarrow \boldsymbol{r_{Y/X} : y = -1.182x + 68.020}$$

Recta de regresión de X sobre Y:

$$r_{X/Y} : \; x - \bar{x} = \frac{\sigma_{XY}}{\sigma_y^2}(y - \bar{y}) \Rightarrow x - 14.5 = \frac{-10.9375}{14.109}(y - 50.875) \Rightarrow \boldsymbol{r_{X/Y} : x = -7.752y + 53.939}$$

c) $r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} = \frac{-10.9375}{3.041 \cdot 3.756} = -0.9576 \Rightarrow \boldsymbol{r = -0.9576}$.

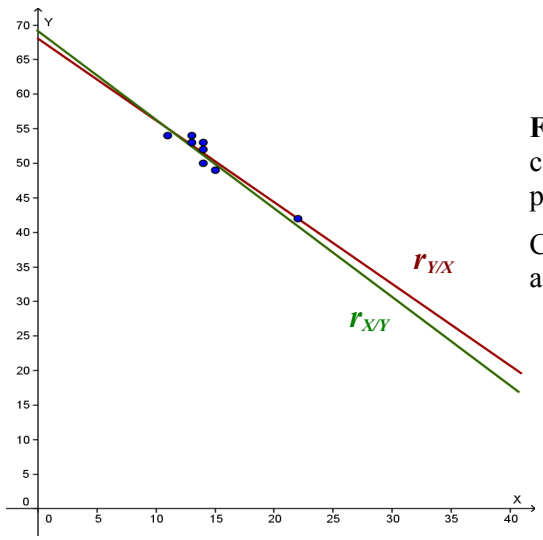
$r^2 = (-0.9576)^2 \Rightarrow \boldsymbol{r^2} = 0.917 \Rightarrow$ El 91.7% de la variable Y es explicada por la variable X (y viceversa).

$\sigma_X = \sqrt{9.25} = 3.041 \qquad \sigma_Y = \sqrt{14.109} = 3.756$

Existe una correlación fuerte e inversa. Al aumentar los valores de una de las variables, disminuyen los de la otra.

d) $x = 12 \Rightarrow y = -1.182 \cdot 12 + 68.020 \Rightarrow \boldsymbol{y = 53.836}$

e) La predicción es fiable ya que el coeficiente de correlación está muy próximo a -1.



Fíjate: La correlación lineal entre ambas variables es casi perfecta y, por tanto, las rectas están muy próximas.

Casi coinciden y el ángulo que determinan se acerca a cero.