UNIDAD 9:
DISTRIBUCIONES ESTADÍSTICAS BIDIMENSIONALES

1. VARIABLES ESTADÍSTICAS BIDIMENSIONALES

1.1. Introducción

Consideremos una población de N individuos descrita según dos caracteres cuantitativos X e Y, es decir, según dos variables estadísticas: el par (X,Y) se denomina **variable estadística bidimensional**.

Supongamos que X e Y son discretas, en el caso de que fueran continuas x_i e y_j designarían las marcas de clase (puntos medios de los intervalos):

$X : x_1, x_2, ..., x_k$

$Y : y_1, y_2, ..., y_p$

La tabla estadística que describe la población, da la frecuencia absoluta f_{ij} de individuos que presentan a la vez el valor x_i de la variable estadística X y el valor y_j de la variable estadística Y.

$X \setminus Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_p
x_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1p}
x_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
x_i	f_{i1}	f_{i2}	$\cdots$	f_{ij}	$\cdots$	f_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_k	f_{k1}	f_{k2}	$\cdots$	f_{kj}	$\cdots$	f_{kp}

1.2. Distribuciones marginales

Son distribuciones unidimensionales que estudian una variable independientemente de la otra. Sus tablas son:

X	Frecuencias absolutas (f_i)
x_1	f_1
x_2	f_2
$\vdots$	$\vdots$
x_i	f_i
$\vdots$	$\vdots$
x_k	f_k
	$\sum f_i = N$

Y	Frecuencias absolutas (f_j)
y_1	f_1
y_2	f_2
$\vdots$	$\vdots$
y_j	f_j
$\vdots$	$\vdots$
y_p	f_p
	$\sum f_j = N$

1.3. Características marginales

Consideremos la columna marginal de la tabla: da las frecuencias absolutas f_i de individuos que presentan el valor x_i de X, es decir, define la variable marginal X. Las características marginales de X se designan por:

Medida de tendencia central

Media (aritmética): $\boxed{\bar{x} = \frac{\sum f_i x_i}{N}}$

Medidas de dispersión

Varianza: $\boxed{\sigma_X{}^2 = \frac{\sum f_i \left(x_i - \bar{x}\right)^2}{N} = \frac{\sum f_i x_i{}^2}{N} - \bar{x}^2}$

Desviación típica: $\boxed{\sigma_X = +\sqrt{\sigma_X{}^2}}$

Coeficiente de variación: $\boxed{CV_X = \frac{\sigma_X}{\bar{x}}}$ (se usa para comparar distribuciones)

Análogamente, la distribución marginal de Y tiene las siguientes características:

Medida de tendencia central

Media (aritmética): $\boxed{\bar{y} = \frac{\sum f_{j\,y_j}}{N}}$

Medidas de dispersión

Varianza: $\boxed{\sigma_Y{}^2 = \frac{\sum f_j \left(y_j - \bar{y}\right)^2}{N} = \frac{\sum f_j y_j{}^2}{N} - \bar{y}^2}$

Desviación típica: $\boxed{\sigma_Y = +\sqrt{\sigma_Y{}^2}}$

Coeficiente de variación: $\boxed{CV_Y = \frac{\sigma_Y}{\bar{y}}}$ (se usa para comparar distribuciones)

1.4. Una característica conjunta: la covarianza

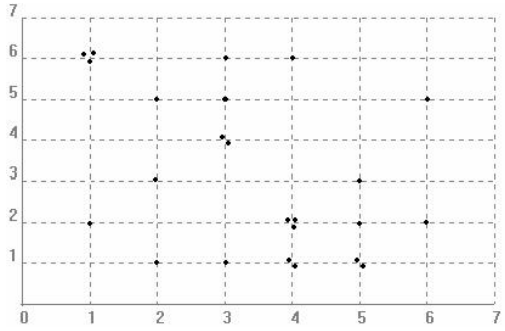
Se llama **covarianza** de una variable estadística bidimensional (X,Y) a la media aritmética de los productos de las desviaciones de cada una de las variables respecto de sus medias respectivas, es decir:

$$\boxed{\sigma_{(X,Y)} = \frac{\sum f_{ij}\left(x_i - \bar{x}\right)\left(y_j - \bar{y}\right)}{N} = \frac{\sum f_{ij} x_i y_j}{N} - \bar{x}\cdot\bar{y}}$$

1.5. Representación gráfica

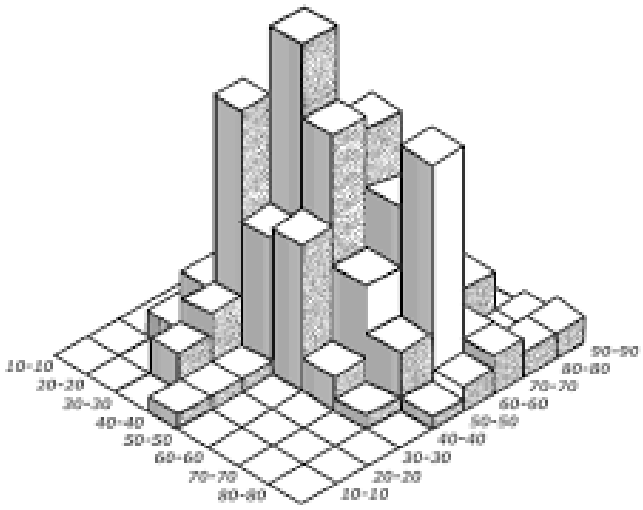
1.5.1. Diagrama de dispersión o nube de puntos

Consiste en unos ejes cartesianos, en los cuales van las variables. Supongamos, por comodidad, que las variables son discretas. En los puntos de confluencia de cada recta correspondiente a cada valor de la variable, se dibujan tantos puntos como frecuencia absoluta tenga el par.



1.5.2. Estereogramas

Consisten en una serie de barras o prismas rectangulares que tienen por altura f_{ij} y cuya base es un punto (variables estadísticas discretas) o un rectángulo de confluencia de los intervalos (variables estadísticas continuas).



2. REGRESIÓN Y CORRELACIÓN

2.1. Introducción

A la hora de establecer relaciones entre sucesos en un determinado campo de investigación, el investigador intenta traducirlas en estructuras manejables, haciendo uso fundamentalmente del lenguaje estadístico-matemático. Para ello, establece un conjunto de relaciones funcionales en donde un número finito de magnitudes (variables, atributos...) $X_1,...,X_n$ se suponen relacionadas con una variable Y a través de una determinada expresión:

$$Y = f\left(X_1,...,X_n\right)$$

Desde esta perspectiva el problema se puede abordar con dos enfoques:

(1) Regresión: determinación de la estructura de dependencia que mejor expresa el tipo de relación de la variable Y con las demás.

(2) Correlación: estudia el grado de dependencia existente entre las variables.

2.2. Regresión

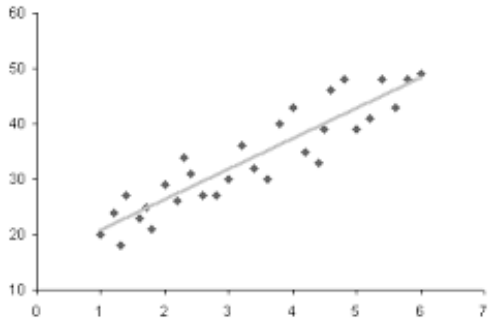
2.2.1. Regresión lineal

Sea $\left(x_i, y_j, f_{ij}\right)$ la nube de puntos a la que queremos ajustar la recta $y = ax + b$. Aplicando el método de mínimos cuadrados (Gauss y Legendre, principios del siglo XIX)[1] la ecuación que se obtiene es la siguiente:

$$y = \frac{\sigma_{(X,Y)}}{{\sigma_X}^2} x + \overline{y} - \frac{\sigma_{(X,Y)}}{{\sigma_X}^2} \overline{x}$$

es decir,

$$y - \overline{y} = \frac{\sigma_{(X,Y)}}{{\sigma_X}^2}\left(x - \overline{x}\right) \qquad \text{Recta de regresión de } Y / X$$



De forma análoga se obtiene la recta de regresión de X / Y:

$$x - \overline{x} = \frac{\sigma_{(X,Y)}}{{\sigma_Y}^2}\left(y - \overline{y}\right)$$

Las dos rectas de regresión se cortan en un punto que recibe el nombre de **centro de gravedad de la distribución**.

Vamos a estudiar los coeficientes de la recta de regresión:
Tenemos que

$$y = \frac{\sigma_{(X,Y)}}{{\sigma_X}^2} x + \overline{y} - \frac{\sigma_{(X,Y)}}{{\sigma_X}^2} \overline{x}$$

luego

$$a = \frac{\sigma_{(X,Y)}}{{\sigma_X}^2}$$

es la pendiente de la recta de regresión de Y / X, y

$$b = \overline{y} - \frac{\sigma_{(X,Y)}}{{\sigma_X}^2} \overline{x}$$

es la ordenada en el origen de dicha recta.

2.3. Análisis de la correlación

2.3.1. Introducción

Llamamos **correlación** al grado de dependencia mutua que puede existir entre las variables según una determinada función de regresión. El objetivo de la Teoría de la Correlación es la determinación

[1] En 1829, Gauss fue capaz de establecer la razón del éxito de este procedimiento: simplemente, el método de mínimos cuadrados es óptimo en muchos aspectos. El argumento concreto se conoce como teorema de Gauss-Markov.

de medidas que cuantifiquen la intensidad con que las variables se relacionan según distintas funciones de regresión, por lo tanto, lo que se va a pretender es "medir" la bondad de las funciones ajustadas a los datos dados.

2.3.2. Correlación y regresión lineal

Supongamos que dada una distribución bidimensional el modelo más adecuado que explica el comportamiento de una variable a partir de los valores de la otra es un modelo lineal, es decir, una ecuación del tipo $y = ax + b$ .Vamos a determinar el grado de asociación existente entre X e Y según esta función, esto es, vamos a cuantificar el grado en que las variables están correladas según una recta. Por lo tanto, lo que estamos midiendo es el grado de acierto al ajustar una recta a la distribución dada.

Llamamos **razón de correlación lineal** a la proporción de variabilidad de Y explicada por la recta de regresión, es decir:

$$r^2 = \frac{\sigma^2_{(X,Y)}}{\sigma_X{}^2 \cdot \sigma_Y{}^2}$$

Como consecuencia de la definición se tiene:

$$0 \le r^2 \le 1$$

Veamos la **interpretación** de r^2:

Si $r^2 = 1$ entonces se dice que existe correlación lineal perfecta entre X e Y, es decir, la correlación lineal es máxima. Por tanto, la recta ajustada **explica** perfectamente el comportamiento de una variable por la otra. En esta situación la recta ajustada va a pasar por todos los puntos de la nube.

Si $r^2 = 0$ entonces se dice que existe correlación lineal nula entre las variables X e Y, es decir, que no existe asociación lineal entre las variables. Por lo tanto, la recta ajustada **no explica** en absoluto el comportamiento de una variable en relación con la otra. Ahora bien, esto no quiere decir que las variables sean independientes, sino que no están relacionadas mediante una recta.

Si $0 < r^2 < 1$ entonces existe cierto grado de correlación lineal, que será mayor cuanto más próximo esté a 1 y ello nos informará de la validez del ajuste, y cuanto más próximo esté a cero menor será la intensidad de la relación lineal de las variables, por lo tanto, peor será el ajuste. Por tanto, cuanto más se acerque dicho valor a la unidad, **mayor poder explicativo** tendrá el modelo de regresión.

> ***Aclaración: poder explicativo vs poder predictivo***
> *Un modelo de regresión con un alto porcentaje de variaciones explicado, puede no ser bueno para predecir, ya que el que la mayoría de los puntos se encuentren cercanos a la recta de regresión, no implica que todos lo estén, y puede ocurrir, que justamente para aquel rango de valores en el que el investigador está interesado, se alejen de la recta y, por tanto, el valor predictivo puede alejarse mucho de la realidad.*
> *La forma de poder evaluar el poder predictivo del modelo queda fuera de lo que se pretende en esta materia.*
>
> ***Otra aclaración más: correlación vs causalidad***

Es muy importante resaltar el hecho, de que el que un modelo sea capaz de explicar de manera adecuada las variaciones de la variable dependiente en función de la independiente (estén correlacionadas), no implica que la primera sea causa de la segunda.

Definimos ahora el **coeficiente de correlación lineal (de Pearson)**.
Este coeficiente va a ser una medida de correlación que además de medir el grado de asociación lineal entre las variables nos va a informar sobre el carácter de esta relación, es decir, si es positiva o negativa, entendiendo que si es positiva las variables van a crecer en el mismo sentido y si es negativa las variables crecerán en sentidos opuestos. En la primera situación tendremos rectas de regresión crecientes y en la segunda rectas de regresión decrecientes.

$$r = \sqrt{r^2} = \frac{\sigma_{(X,Y)}}{\sigma_X \cdot \sigma_Y}$$

Como consecuencia de si definición se verifican:

(1) El signo de este coeficiente es el signo de la covarianza.

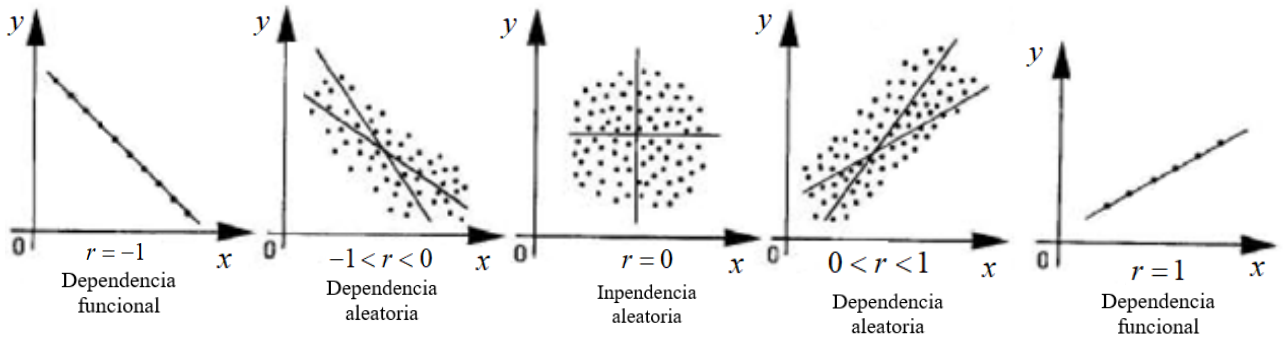
(2) $-1 \le r \le 1$

Interpretamos este coeficiente:
Si $r = 1$ se dice que existe correlación lineal perfecta positiva, es decir, la recta de regresión explica perfectamente el comportamiento de las variables (a medida que aumenta una aumenta la otra). Por tanto, la recta ajustada pasa por todos los puntos y las dos rectas coinciden.

Si $r = -1$ se dice que existe correlación lineal perfecta negativa, es decir, la recta de regresión explica perfectamente el comportamiento de las variables (a medida que aumenta una disminuye la otra). Por tanto, la recta ajustada pasa por todos los puntos, y aunque la intensidad de la asociación es máxima las variables crecen en sentidos opuestos. En este caso también coinciden las dos rectas.
Si $r = 0$ se dice que la correlación lineal es nula, es decir, X e Y no están relacionadas mediante una recta y, por lo tanto, la recta ajustada no explica en absoluto el comportamiento de las variables. Las rectas de regresión son perpendiculares entre sí y paralelas a los ejes de coordenadas.

Si $0 < r < 1$ se dice que existe cierto grado de correlación lineal positiva, es decir, que las variables crecen en el mismo sentido y por lo tanto las rectas de regresión son crecientes y la intensidad de correlación será mayor cuando $r \to 1$ y más débil cuando $r \to 0$.

Si $-1 < r < 0$ se dice que existe cierto grado de correlación lineal negativa, es decir, las variables crecen en sentido opuesto, por lo tanto, las rectas de regresión son decrecientes y la intensidad de la dependencia lineal será mayor cuando $r \to -1$ y menor cuando $r \to 0$.



3. EJERCICIOS

1. Se han lanzado dos dados varias veces. Designando por X el resultado del primer dado y por Y el del segundo, la información obtenida se dispone en la siguiente tabla:

X	1	2	2	3	5	4	1	3	3	4	1	2	5	4	3	4	4	5	3	1	6	5	4	6
Y	2	3	1	4	3	2	6	4	1	6	6	5	1	2	5	1	1	2	6	6	2	1	2	5

a) Calcular la tabla bidimensional.
b) Calcular las distribuciones marginales de cada dado.
c) Analizar dichos datos.
d) Calcular la covarianza de (X,Y).
e) Representar gráficamente dichos datos.

2. Se han obtenido los pesos (kg) y las tallas (cm) correspondientes a un grupo de individuos, obteniendo la siguiente información (las variables se han considerado discretas por simplificar):

$X \setminus Y$	160	162	164	166	168	170
48	3	2	2	1	0	0
51	2	3	4	2	2	1
54	1	3	6	8	5	1
57	0	0	1	2	8	3
60	0	0	0	2	4	4

a) Calcular el peso y la talla media.
b) Obtener las distribuciones marginales.
c) Medidas de dispersión marginales.
d) Covarianza de (X,Y).

3. 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 siguientes resultados:

$X \setminus Y$	1	2	3	4
1	7	0	0	0
2	10	2	0	0
3	11	5	1	0
4	10	6	6	0
5	8	6	4	0
6	1	2	3	1
7	1	0	0	1
8	0	0	1	1

a) Calcular el número medio de miembros de las familias encuestadas.
b) Cuantificar la dispersión que presenta la distribución del número de personas activas entorno a su media.
c) Calcular la covarianza.
d) Representar gráficamente dichos datos.

4. De una muestra de 24 puestos de venta en un mercado de abastos se recogía información acerca del número de balanzas (X) y el número de dependientes (Y): los resultados aparecen en la siguiente tabla:

$X \setminus Y$	1	2	3	4
1	1	2	0	0
2	1	2	3	1
3	0	1	2	6
4	0	0	2	3

a) Determinar las rectas de regresión.

b) ¿Es apropiado suponer que existe una relación lineal entre las variables?

c) Predecir el número de balanzas que pueden esperarse si son 6 los dependientes de un puesto. ¿Es aceptable esta predicción?

5. Sobre el gasto en espectáculos (Y en %) y la renta disponible mensual (X en euros) se dispone de la información referente a 6 familias:

y_j	0,3	0,5	0,6	0,9	1,0	1,4
x_i	360	420	442	600	900	1 260

Explicar el comportamiento de Y por X mediante una recta. ¿Es adecuado este ajuste?

6. Se realizó un experimento para estudiar el efecto de un determinado medicamento en la disminución de los ataques de corazón. La variable independiente fue la dosis de droga en miligramos, X , y la variable dependiente la reducción en los ataques al corazón, al compararlos con un grupo control, Y . Los datos recogidos fueron los siguientes:

x_i	0,50	0,75	1,00	1,25	1,50	1,75	2,00	2,25	2,50	2,75	3,00	3,25	3,50
y_j	10	8	12	12	14	12	16	18	17	20	18	20	21

a) Determinar la recta de regresión de Y / X .

b) ¿Qué valor se espera obtener si la dosis suministrada es de 4 miligramos?

c) ¿Es adecuado dicho ajuste?

7. El director de un restaurante tiene la siguiente distribución de frecuencias sobre el coste (en euros) de las comidas solicitadas por sus clientes en un fin de semana:

Número de comidas servidas	Coste por comida
X	Y
30	6,90
35	6,60
40	5,80
45	6,00
50	5,80
55	5,40
60	5,30

70	5,10
75	4,70
80	4,20
65	4,80

Se pide:

a) Determinar la recta de regresión de Y / X .

b) ¿Cuál es el coste por comida, si se han servido 65 comidas?

c) Decidir si dicho ajuste es adecuado.

8. Un estudio psicológico proporcionó los siguientes datos sobre el número de hijos en la familia, X , y el cociente intelectual promedio de esa familia, Y :

Número de hijos en la familia	Cociente intelectual promedio en la familia
X	Y
1	105
2	102
3	104
4	100
5	97
6	101
7	95
8	93
9	97
10	88

Se pide:

a) Determinar la recta de regresión de Y / X .

b) Decidir si dicho ajuste es adecuado.

9. La siguiente tabla proporciona los datos de un hospital sobre el tiempo que un enfermo pasa en él, Y , y el nivel de ingresos mensuales (en miles de euros) de la familia, X :

x_i	1,00	1,25	1,50	1,75	2,00	2,25
y_j	11	12	9	8	9	10

2.50	2,75	3,00	3,25	3,50	3,75	4,00
7	8	4	7	5	6	3

Se pide:

a) Determinar la recta de regresión de Y / X .

b) Ver si dicha recta modela adecuadamente el estudio.

c) ¿Qué número de días pasa ingresado un enfermo con unos ingresos de 5 000 euros? ¿Es adecuada dicha predicción?

10. Dos métodos de medida de la respuesta cardiaca fueron comparados en 10 animales experimentales, obteniéndose los siguientes datos:

Método I	Método II
X	Y
0,8	0,5
1,0	1,2
1,3	1,1
1,4	1,3
1,5	1,1
1,4	1,8
2,0	1,6
2,4	2,0
2,7	2,4
3,0	2,8

Se pide:

a) Determinar la recta de regresión de Y / X .

b) Si con el método I la respuesta cardiaca es 3,5, ¿qué valor se espera obtener con el método II?

c) Ver si dicha recta modela adecuadamente el estudio, y decidir si el valor obtenido en el apartado b es adecuado o no.

11. El consumo de energía per cápita en miles de kWh y la renta per cápita en miles de dólares en seis países de la UE son los siguientes:

	Consumo	Renta
Alemania	5,7	11,1
Bélgica	5,0	8,5
Dinamarca	5,1	11,3
España	2,7	4,5
Francia	4,6	9,9
Italia	3,1	6,5

a) Calcular el coeficiente de correlación lineal e interpretar el resultado.

b) ¿Qué predicción podemos hacer sobre el consumo de energía per cápita de Grecia si sabemos que su renta per cápita es de 4,4 miles de dólares? ¿Es aceptable la predicción realizada?

12. En las bibliotecas de seis poblaciones se ha analizado la afluencia de lectores X (en miles de personas) y el número de libros prestados Y . Los datos recogidos se muestran en la tabla siguiente:

X	0,5	1,0	1,3	1,7	2,0	2,5
Y	180	240	250	300	340	400

a) ¿Cuál es el número medio de libros prestados en el conjunto de todas las bibliotecas?

b) Escribe la recta de regresión que expresa el número de libros prestados en función de la afluencia de lectores.

c) Si acudiesen 1500 lectores a una biblioteca, ¿cuántos libros se prestarían?

13. La torre inclinada de Pisa es una maravilla arquitectónica. Su creciente inclinación ha generado numerosos estudios y obras sobre su futura estabilidad. En la siguiente tabla se presentan las medidas de su inclinación entre los años 1978 y 1987. Los datos de inclinación se han codificado como décimas de milímetros por exceso de 2,9 m, de manera que la inclinación en el año 1978, que fue de 2,9667 aparece en la tabla como 667.

Años	Inclinación	Años	Inclinación
1978	667	1983	713
1979	673	1984	717
1980	688	1985	725
1981	696	1986	742
1982	698	1987	757

a) ¿Crees que la inclinación de la torre tiene una tendencia lineal que crece con el tiempo? Justifica tu respuesta.
b) Calcula la recta de regresión de la inclinación en el tiempo.