UNIDAD

11 Estadística bidimensional

ÍNDICE DE CONTENIDOS

La Estadística sirve para describir datos. En cursos anteriores el alumno ha podido empezar a estudiar los rudimentos de la Estadística, en particular de la Estadística unidimensional. En esta unidad, después de un breve repaso de los conceptos estudiados anteriormente, nos vamos a ocupar de la Estadística bidimensional. En la Estadística unidimensional se estudia un grupo de datos, por ejemplo, el gasto mensual en libros de un cierto número de familias. Con la Estadística bidimensional podríamos estudiar la relación entre el gasto mensual en libros de un grupo de familias con sus ingresos mensuales, es decir, podemos estudiar la relación de dos características distintas de un determinado conjunto de individuos.

1. Estadística unidimensional

1.1. Población y muestra

Se ha llevado a cabo una encuesta a un grupo de 15 estudiantes de primero de bachillerato sobre el número de horas semanales que dedican al estudio de la asignatura de Matemáticas. Los encuestados han respondido que el número de horas que dedican al estudio de esta asignatura son:

10	8	5	3	1
8	4	4	2	1
5	6	7	8	8

Este es un ejemplo de **variable estadística unidimensional**, se dice unidimensional porque sólo estamos estudiando una característica, numérica en este caso, del grupo, el número de horas. El grupo al que se ha dirigido la encuesta, estudiantes de primero de bachillerato, se denomina **población**. La población son todos los estudiantes de primero de bachillerato de España, por ejemplo. Los datos de los 15 estudiantes que hemos recogido, son una **muestra** extraída de la población.

Una vez que tenemos una muestra de datos, el estudio de los mismos se puede llevar a cabo con dos finalidades distintas. Podemos estar interesados únicamente en sacar conclusiones sobre el número de horas que dedican al estudio de las Matemáticas los 15 estudiantes a los que hemos preguntado; esto es hacer **Estadística descriptiva**. Por otra parte, podríamos intentar sacar conclusiones sobre los hábitos de estudio de toda la población, en este caso, los estudiantes españoles de primero de bachillerato, mediante el estudio de la muestra de 15 que tenemos; esto sería hacer **Estadística inferencial** o inferencia estadística. Para este último punto de vista habría que utilizar modelos matemáticos de probabilidad, y es lo que se estudiará en el curso siguiente. En este curso y en esta unidad en particular, estudiaremos la Estadística desde el punto de vista descriptivo. En otras palabras, no pretendemos descubrir los hábitos de estudio de todos los estudiantes de primero de bachillerato, sino simplemente analizar lo que les ocurre a los 15 que hemos elegido.

Como se ha dicho en la introducción, aquí vamos a recordar cómo se calculan algunos de los parámetros asociados a una variable estadística unidimensional, que ya se estudiaron en cursos anteriores. Sólo aquellos que vamos a necesitar en el resto de la undiad didáctica.

1.2. Parámetros estadísticos

Los parámetros estadísticos asociados a una variable son ciertos números que se calculan a partir de los datos, que proporcionan información sobre el comportamiento conjunto de la variable.

Para ir recordándolos, vamos a utilizar el ejemplo de las horas de estudio de los 15 estudiantes de primero de bachillerato. Comenzamos ordenándolos de menor a mayor,

1	1	2	3	4	4	5	5	6	7	8	8	8	8	10

Si sumamos todos los datos y dividimos la suma por la cantidad de ellos que hay, obtenemos la **media** o *media aritmética*, que representaremos por $\overline{x}$. En nuestro caso,

$$\overline{x} = \frac{1+1+2+3+4+4+5+5+6+7+8+8+8+8+10}{15} = \frac{80}{15} = 5'33$$

lo que nos indica que, por término medio, estos quince estudiantes dedican 5'33 horas semanales al estudio de las matemáticas.

En general, si disponemos de los datos $x_1, x_2, \ldots, x_n$, la *media* es

$$\boldsymbol{\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}}$$

Donde n es el número de datos de que disponemos.

Para abreviar la expresión anterior, en particular la suma del numerador se utiliza el símbolo sumatorio $\sum$. Aunque lo correcto sería escribir $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$, para indicar que la suma es desde el primer dato x_1, hasta el enésimo x_n, escribiremos simplemente $\sum x_i$ para referirnos a la suma de todos los x_i.

Por tanto, la fórmula de la media se puede escribir de la forma siguiente:

$$\boldsymbol{\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x_i}{n}}$$

La media es un parámetro que mide la tendencia central de la variable estadística. Otros parámetros de tendencia central son, aunque no los usaremos en la estadística bidimensional; la **moda**, que es el dato que aparece un mayor número de veces, y la **mediana**, que es el dato que ocupa el lugar central, si previamente se han ordenado de menor a mayor.

En nuestro ejemplo, la *moda* es 8, que es el dato que aparece mayor número de veces, es decir, con mayor *frecuencia*. Y la *mediana*, dado que hay 15 datos, es el dato que ocupa el lugar octavo, $x_8 = 5$, una vez que están ordenados de menor a mayor:

MEDIANA ↓

1	1	2	3	4	4	5	**5**	6	7	8	8	8	8	10

Si el número de datos fuese par, entonces la mediana es la media aritmética de los dos datos que se encuentran en el centro.

Además de los parámetros que miden la tendencia central; media, moda y mediana, hay otros que sirven para medir cuál es la dispersión de los mismos, es decir, si están más o menos concentrados alrededor de la media. Los que vamos a recordar aquí son **varianza** y la **desviación típica**.

Varianza y desviación típica se utilizan para medir el promedio de las desviaciones de los datos con respecto de la media. Para medir esta desviación, podríamos calcular todas las diferencias entre dato y media, $(x_i - \overline{x})$, y sumarlas, para después dividir entre el número de datos. Sin embargo, este resultado sería nulo. Por esta razón, lo que se hace es utilizar los cuadrados de las diferencias, así se trata de una suma de números positivos o nulos, y no siempre dará cero.

La *varianza* es la suma de los cuadrados de las diferencias entre cada dato y la media, dividida por el número de datos, se denota s^2,

$$\boldsymbol{s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \frac{\sum (x_i - \overline{x})^2}{n}}$$

Entonces, para calcular la varianza, hay que restar la media a cada dato, después se eleva al cuadrado cada número y se suman todos, por último, se divide entre n. El proceso es algo largo, pero se puede abreviar, porque la fórmula admite una expresión más sencilla de utilizar. Se puede comprobar que la siguiente expresión proporciona el resultado, y es la que vamos a utilizar en la práctica:

$$\boldsymbol{s^2 = \frac{\sum x_i^2}{n} - \overline{x}^2}$$

A pesar de que la varianza mide el promedio de las desviaciones de la media, como estas desviaciones se han elevado al cuadrado, no están en las mismas unidades que los datos. Por esta razón, resulta más intuitivo a la hora de interpretar la variabilidad o dispersión de los datos, utilizar otro parámetro, la *desviación típica*, que es simplemente la raíz cuadrada de la varianza, y se denota s,

$$\boldsymbol{s = \sqrt{\frac{\sum x_i^2}{n} - \overline{x}^2}}$$

Calculemos la varianza y desviación típica en nuestro ejemplo del número de horas de estudio:

La varianza la calculamos con la fórmula

$$s^2 = \frac{\sum x_i^2}{n} - \overline{x}^2$$

Calculamos en primer lugar la suma de los cuadrados de los x_i,

$$\begin{aligned}
\sum x_i^2 &= x_1^2 + x_2^2 + \cdots + x_n^2 \\
&= 1^2 + 1^2 + 2^2 + 3^2 + 4^2 + 4^2 + 5^2 + 5^2 + 6^2 + 7^2 + 8^2 + 8^2 + 8^2 + 8^2 + 10^2 \\
&= 1 + 1 + 4 + 9 + 16 + 25 + 36 + 49 + 64 + 64 + 64 + 64 + 100 \\
&= 538
\end{aligned}$$

Ahora sustituimos en la fórmula de la varianza,

$$s^2 = \frac{538}{15} - 5'33^2 = 7'46$$

Por tanto, la desviación típica,

$$s = \sqrt{s^2} = \sqrt{7'46} = 2'73$$

En resumen, la media de tiempo dedicado al estudio de la asignatura de matemáticas de los 15 alumnos analizados es $\overline{x} = 5'33$ horas; y la desviación típica es $s = 2'73$, lo que quiere decir que la mayoría de los 15 estudiantes dedica al estudio de las

matemáticas un tiempo comprendido entre las $5'33 - 2'73 = 2'6$ horas y $5'33 + 2'73 = 8'06$ horas.

ACTIVIDADES

1. El número de libros leídos en el último mes por diez personas ha sido:

0	1	2	2	2	3	3	4	5	5

Calcular la media, moda, mediana, varianza y desviación típica.

Recuerda

- ✓ Una *variable estadística unidimensional* es una característica numérica de un grupo de individuos.
- ✓ Se llama *población* al conjunto de individuos del que se hace un estudio estadístico. Se llama *muestra* a un subconjunto de la población.
- ✓ Conocidos los datos $x_1, x_2, \ldots, x_n$ de una variable estadística. Los principales parámetros estadísticos se calculan de la forma siguiente:
  - Media. Es la suma de los datos dividido entre la cantidad de ellos $\overline{x} = \dfrac{\sum x_i}{n}$.
  - Moda. Es el dato que aparece con mayor frecuencia.
  - Mediana. Es el dato central, si el número de datos es impar. Si el número es par, es la media de los dos datos centrales.
  - Varianza. Es el promedio de los cuadrados de las diferencias entre cada dato y su media, se puede calcular de dos formas equivalentes, aunque la segunda es más conveniente, $s^2 = \dfrac{\sum (x_i - \overline{x})^2}{n} = \dfrac{\sum x_i^2}{n} - \overline{x}^2$.
  - Desviación típica. Es la raíz cuadrada de la varianza, mide la desviación de la media de los datos, $s = \sqrt{s^2}$.

2. Estadística bidimensional

Hemos recordado en la sección anterior algunas nociones de estadística unidimensional. Allí estudiábamos una característica numérica de un grupo de individuos. En la estadística bidimensional se trata de estudiar dos características numéricas de cada individuo.

2.1. Variables estadísticas bidimensionales

Tenemos un grupo de 8 estudiantes de primero de bachillerato que han hecho un examen de la asignatura de Matemáticas y otro de la asignatura de Física y Química,

las notas, de 0 a 10, de cada alumno en cada asignatura han sido las siguientes:

x_i	2	4	5	6	7	8	8	9
y_i	3	5	4	$5'5$	5	$7'5$	8	10

donde para cada par (x_i, y_i), la nota de Matemáticas es x_i, y la de Física y Química, y_i de un alumno concreto.

Este es un ejemplo de **variable estadística bidimensional** (X, Y), donde X = *nota de Matemáticas*, Y = *nota de Física y Química*. El objeto de estudiar las dos variables de manera conjunta es el de ver qué tipo de relación hay entre ambas, si es que hay alguna.

2.2. Nube de puntos

Una primera aproximación al estudio de la relación existente entre las dos variables que constituyen una variable bidimensional es su representación gráfica. Si cada par (x_i, y_i) lo representamos como un punto en unos ejes coordenados, se obtiene una gráfica que llamaremos **nube de puntos** o *diagrama de dispersión*. Para el ejemplo anterior, la nube de puntos se ha representado en la figura 11.1.

La forma de la nube de puntos de la figura 11.1 nos sugiere que existe algún tipo de dependencia entre la nota de Matemáticas y la de Física y Química.

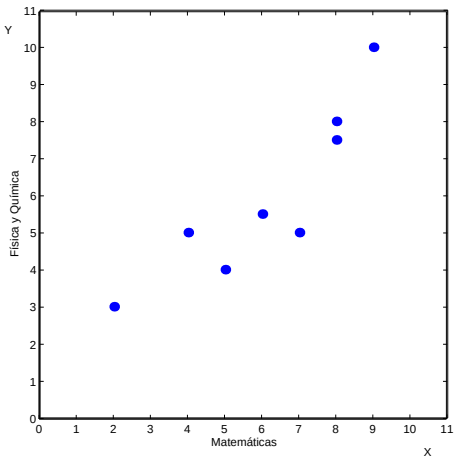


Figura 11.1: Nube de puntos

En efecto, según vemos en la figura, parece que cuando un alumno tiene una nota alta en la asignatura de Matemáticas, también la tiene en la asignatura de Física y Química. Lo mismo ocurre con alumnos que han obtenido nota baja en una de las asignaturas, también ha obtenido una nota baja en la otra. Se dice en este caso que parece existir una *dependencia linea positiva* entre las dos variables (debido a que la nube de puntos se asemeja a una recta con pendiente positiva). Mediante los parámetros bidimensionales que estudiaremos después, veremos cómo se puede medir cuantitativamente el grado de dependencia.

La nube de puntos puede presentar muchas formas distintas, en la figuras siguientes se muestran algunas de las diferentes posibilidades.

- En la figura 11.2, los puntos parecen acumularse alrededor de una recta con pendiente positiva, como en nuestro ejemplo anterior. Diremos entonces que entre las variables, hay *dependencia lineal positiva*.
- En la figura 11.3, los puntos parecen acumularse en torno a una recta con pendiente negativa. En este caso, diremos que entre las variables, hay *dependencia lineal negativa*.

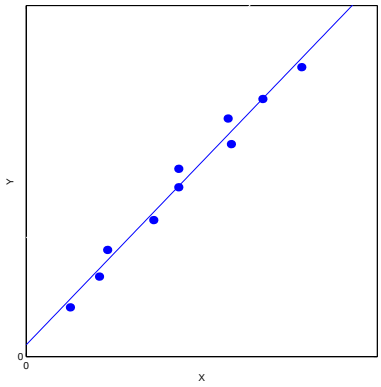


Figura 11.2: Dependencia lineal positiva

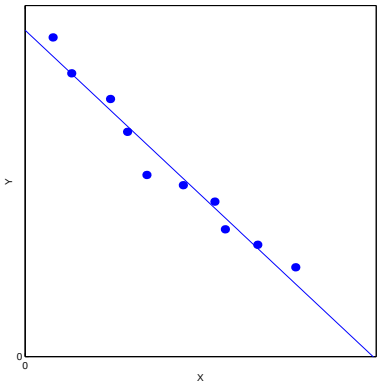


Figura 11.3: Dependencia lineal negativa

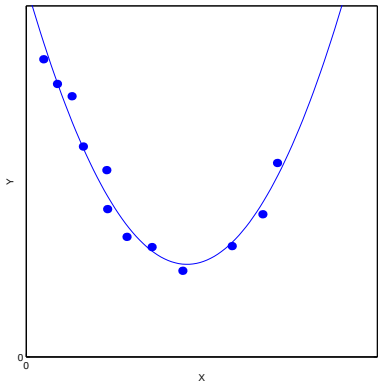


Figura 11.4: Dependencia no lineal

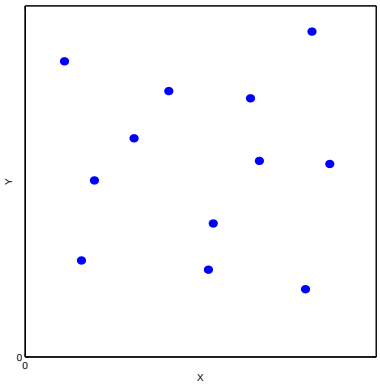


Figura 11.5: No hay dependencia

- En la figura 11.4, los puntos parecen seguir una curva, en este caso parecida a una parábola. Es decir, entre las variables parece haber algún tipo de dependencia, aunque ahora es *no lineal*.
- Por último, en la figura 11.5, los puntos están situados en el plano sin seguir ninguna pauta concreta, sin aproximarse a ninguna curva o recta que sugiera algún tipo de función. Diremos que entre las variables, *no hay dependencia*.

Pero, incluso dentro de cada uno de los casos anteriores se pueden establecer diferencias. Observemos las nubes de puntos de las figuras 11.6 y 11.7. En ambas tenemos, según acabamos de explicar, una dependencia lineal negativa. Sin embargo, son claramente distintas. Mientras que los puntos del diagrama de la figura 11.6 están prácticamente sobre la recta, en la figura 11.7, los puntos están mucho más dispersos. Para distinguir entre estas posibilidades, se dice que hay dependencia lineal *fuerte* y dependencia lineal *débil*, respectivamente.

No obstante, esta denominación de fuerte o débil resultaría ambigua si no hubiese otra posibilidad que la simple observación de la nube de puntos. Por esta razón, veremos más adelante un parámetro, el coeficiente de correlación, que establece numéricamente una medida para el grado de dependencia de las variables.

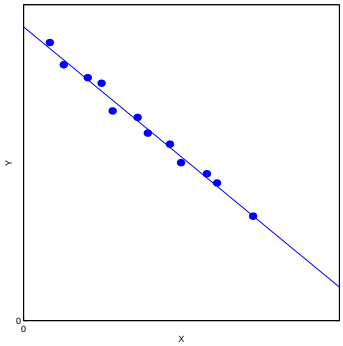


Figura 11.6: Dependencia *fuerte*

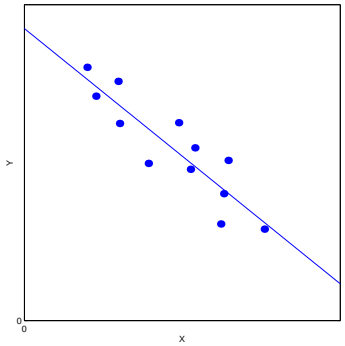


Figura 11.7: Dependencia *débil*

ACTIVIDADES

2. Dibujar las nubes de puntos correspondientes a las siguientes variables estadísticas bidimensionales. Indicar además, el tipo de dependencia entre las variables que se observa en el diagrama. Será útil para ello intentar dibujar la recta o la curva que mejor se aproxime a la nube de puntos:

a)

x_i	2	2	3	4	6	7	8	5
y_i	9	8	8	5	3	1	1	4

b)

x_i	1	1	2	4	5	6	7	8
y_i	2	3	6	7	7	6	4	2

c)

x_i	1	2	4	4	6	6	9	9
y_i	2	6	8	2	4	8	2	7

d)

x_i	1	2	2	4	5	6	7	7
y_i	1	2	3	4	6	7	8	9

Recuerda

- ✓ Una *variable estadística bidimensional* es un par de características numéricas (X, Y) de un cierto conjunto de individuos. Por ejemplo, el peso y la estatura de un grupo de personas.
- ✓ Una *nube de puntos* o *diagrama de dispersión* es la representación gráfica de los datos de una variable estadística bidimensional (X, Y). Se trata de representar los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, como puntos en un sistema de ejes coordenados.
- ✓ El aspecto de la nube de puntos sugiere la dependencia entre las variables X e Y de una variable bidimensional (X, Y):

  - Si la los puntos se acumulan en torno a una recta de pendiente positiva, se dice que hay *dependencia lineal positiva*.

  - Si se acumulan alrededor de una recta de pendiente negativa, se dice que hay *dependencia linea negativa*.

  - También puede haber dependencia *no lineal*, cuando los puntos se acumulan en torno a una curva, y puede no haber dependencia, si los puntos no siguen ninguna pauta reconocible.

3. Parámetros estadísticos bidimensionales

Como ya hemos comentado antes, el objeto de estudiar de manera conjunta dos variables estadísticas en lo que llamamos una variable bidimensional, es el de estudiar la posible dependencia de las dos variables. Para ello, además de la nube de puntos, se pueden calcular ciertos parámetros estadísticos que proporcionan información sobre el conjunto de datos. Hay unos parámetros que se refieren sólo a cada variable por separado, son las **medias** y **desviaciones típicas marginales**. Otros, que estudiaremos después, involucran a los datos de las dos variables.

3.1. Medias y desviaciones típicas marginales

Consideremos los siguientes datos correspondientes a una variable bidimensional (X, Y).

x_i	1	3	3	4	5	6	7	8	8	8
y_i	2	3	4	4	6	7	7	7	8	9

Si pensamos en los datos de la variable X de forma independiente, podemos calcular, tanto su media como su desviación típica con las fórmulas de la estadística unidimensional. Así, las medias se calculan mediante las fórmulas:

$$\overline{x} = \frac{\sum x_i}{n} \qquad \overline{y} = \frac{\sum y_i}{n}$$

donde $\sum x_i$, $\sum y_i$ son las sumas de x_i e y_i, respectivamente, y n es el número de (pares de) datos.

Para nuestro ejemplo, estas medias son:

$$\overline{x} = \frac{53}{10} = 5'3 \qquad \overline{y} = \frac{57}{10} = 5'7.$$

Las **medias marginales** admiten una interpretación gráfica importante. Si dibujamos la nube de puntos (en la figura 11.8 hemos dibujado la nube de los datos anteriores), el punto de coordenadas $(\overline{x}, \overline{y})$ se encuentra siempre en el **centro de gravedad** de la nube, esto es, el punto donde se puede suponer concentrada toda la masa.

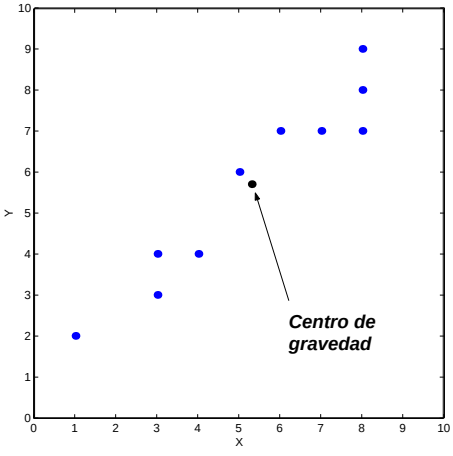


Figura 11.8: $(\overline{x}, \overline{y})$ es el centro de gravedad

ACTIVIDADES

3. Para los siguientes datos, dibujar la nube de puntos correspondiente y su *centro de gravedad*, es decir, el punto de coordenadas $(\overline{x}, \overline{y})$:

a)
x_i	2	3	4	5
y_i	6	5	3	2

b)
x_i	1	2	3	4
y_i	1	3	3	1

c)
x_i	1	2	3	4
y_i	1	3	5	4

.

También se pueden calcular de forma separada las varianzas y las desviaciones típicas marginales, mediante las fórmulas siguientes:

- Las **varianzas marginales**:

$$s_x^2 = \frac{\sum x_i^2}{n} - \overline{x}^2 \qquad s_y^2 = \frac{\sum y_i^2}{n} - \overline{y}^2$$

- Las **desviaciones típicas marginales**:

$$s_x = \sqrt{\frac{\sum x_i^2}{n} - \overline{x}^2} \qquad s_y = \sqrt{\frac{\sum y_i^2}{n} - \overline{y}^2}$$

Vamos a calcular las varianzas y desviaciones típicas del ejemplo con el que empezábamos este apartado. Para llevar a cabo estos cálculos, y los que vendrán después, conviene organizar la información en una tabla que nos permita realizar la tarea de una manera más cómoda. Una posibilidad es organizarlo como en la tabla que se pone a continuación, en la que aparecen (de momento) cuatro columnas; x_i, y_i, que son los datos originales, y x_i^2, y_i^2, que son los datos al cuadrado. En la última fila de la tabla se ponen las sumas de todas las columnas:

x_i	y_i	x_i^2	y_i^2
1	2	1	4
3	3	9	9
3	4	9	16
4	4	16	16
5	6	25	36
6	7	36	49
7	7	49	49
8	7	64	49
8	8	64	64
8	9	64	81
53	57	337	373

← SUMAS

Y ahora sólo queda sustituir en las fórmulas.

Varianzas:

$$s_x^2 = \frac{337}{10} - 5'3^2 = 5'61$$

$$s_y^2 = \frac{373}{10} - 5'7^2 = 4'81.$$

Desviaciones típicas:

$$s_x = \sqrt{5'61} = 2'37$$

$$s_y = \sqrt{4'81} = 2'19.$$

3.2. Covarianza

La **covarianza** es el primer parámetro conjunto que vamos a estudiar, en el sentido de que involucra los datos de las dos variables. Da una idea sobre la forma en que se distribuyen los puntos alrededor del centro de gravedad $(\overline{x}, \overline{y})$. Se representa por s_{xy} y es la media de los productos de las diferencias de las coordenadas de cada punto de

la nube (x_i, y_i) y las coordenadas de $(\overline{x}, \overline{y})$, es decir,

$$\boldsymbol{s_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n}}$$

Al igual que ocurría en el caso de las varianzas, la covarianza se puede calcular mediante una fórmula que es equivalente a la anterior, pero más sencilla de utilizar desde un punto de vista práctico, esta fórmula equivalente es la siguiente:

$$\boldsymbol{s_{xy} = \frac{\sum x_i.y_i}{n} - \overline{x}.\overline{y}}$$

Para utilizar esta fórmula, primero hay que calcular la suma de todos los productos $x_i.y_i$, dividir esta suma entre n y restarle el producto de las medias. A efectos prácticos, en la tabla de la que hemos hablado antes, añadimos una nueva columna, la de los productos $x_i.y_i$, cuyos componentes serán los productos de los de las columnas x_i, y_i. Para el ejemplo del apartado anterior, completando la tabla que habíamos empezado antes,

x_i	y_i	x_i^2	y_i^2	$x_i.y_i$
1	2	1	4	2
3	3	9	9	9
3	4	9	16	12
4	4	16	16	16
5	6	25	36	30
6	7	36	49	42
7	7	49	49	49
8	7	64	49	56
8	8	64	64	64
8	9	64	81	72
53	57	337	373	352

Entonces, la covarianza en este caso es

$$s_{xy} = \frac{\sum x_i.y_i}{n} - \overline{x}.\overline{y} = \frac{352}{10} - 5'3 \cdot 5'7 = 4'99$$

Aparte de la forma en la que se calcula la covarianza que, según acabamos de ver, no reviste una especial dificultad, nos interesa saber cuál es el significado de este parámetro. Como dijimos al principio, la covarianza da una idea de cómo están distribuidos los puntos alrededor del centro de gravedad $(\overline{x}, \overline{y})$. Y esto se aprecia mediante el signo.

- Si la covarianza es un número positivo, los puntos estarán dispuestos de manera que haya dependencia lineal positiva.

- Si la covarianza es un número negativo, los puntos estarán dispuestos de manera que haya dependencia lineal negativa.

- Si la covarianza es un número próximo a cero, esto indicará que, o bien no hay dependencia entre las variables, o esta dependencia no es lineal.

Hay una explicación geométrica para lo anterior, y aunque no vamos a entrar mucho en los detalles, indicaremos que realmente la covarianza no es otra cosa que la

media de los productos de las coordenadas de los puntos una vez trasladados los ejes al punto $(\overline{x}, \overline{y})$. Entonces, si cada vez que tenemos una nube de puntos, imaginamos los ejes centrados en $(\overline{x}, \overline{y})$, dependiendo de en qué cuadrante estén los puntos, los signos de los productos serán positivos o negativos. Por ejemplo, si la mayoría de los puntos se encuentran en el primer y tercer cuadrante, la media será positiva. Esto es lo que ocurre cuando hay dependencia lineal positiva.

ACTIVIDADES

4. A partir de los datos siguientes, dibujar la nube de puntos y calcular medias y la covarianza:

x_i	1	2	3	4
y_i	9	8	6	3

¿Qué tipo de dependencia hay entre las variables?

Recuerda

✓ Sea (X, Y) una variable estadística bidimensional. A partir de los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ de la varible, se pueden calcular los siguientes paráme-tros:

✓ *Medias marginales*:

$$\overline{x} = \frac{\sum x_i}{n} \qquad \overline{y} = \frac{\sum y_i}{n}$$

El punto $(\overline{x}, \overline{y})$ es el *centro de gravedad* de la nube de puntos.

✓ *Varianzas marginales:*

$$s_x^2 = \frac{\sum x_i^2}{n} - \overline{x}^2 \qquad s_y^2 = \frac{\sum y_i^2}{n} - \overline{y}^2$$

✓ *Desviaciones típicas marginales:*

$$s_x = \sqrt{\frac{\sum x_i^2}{n} - \overline{x}^2} \qquad s_y = \sqrt{\frac{\sum y_i^2}{n} - \overline{y}^2}$$

✓ *Covarianza:*

$$s_{xy} = \frac{\sum x_i.y_i}{n} - \overline{x}.\overline{y}$$

✓ El signo de la covarianza determina el tipo de dependencia entre las variables:

- Si $s_{xy} > 0$ hay dependencia lineal positiva.
- Si $s_{xy} < 0$ hay dependencia lineal negativa.

4. Correlación

A pesar de que el signo de la covarianza indica el tipo de dependencia que hay entre las variables, no nos sirve para determinar si esta dependencia es más o menos fuerte. La razón es que su valor depende de las unidades en las que estén dados los datos. Es decir, que si una de las variables está en centímetros, por ejemplo, y pasamos estas medidas a metros, el valor de la covarianza cambia (aunque no el signo). Lo cual impide que se pueda establecer una comparación acertada entre dos dependencias lineales positivas o dos dependencias negativas.

Para evitar este problema, se utiliza otro parámetro llamado **coeficiente de correlación lineal** o *coeficiente de Pearson* (nosotros lo llamaremos simplemente coeficiente de correlación), que se define como el cociente entre la covarianza y el producto de las desviaciones típicas:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

Veamos en primer lugar un ejemplo sencillo de cómo se calcula y, a continuación, comentaremos su significado.

Por ejemplo, queremos calcular el *coeficiente de correlación* para los datos siguientes:

x_i	1	3	4	6
y_i	1	2	5	5

En primer lugar, organizamos la información en la tabla que hemos aprendido a utilizar antes,

x_i	y_i	x_i^2	y_i^2	$x_i.y_i$
1	1	1	1	1
3	2	9	4	6
4	5	16	25	20
6	5	36	25	30
14	13	62	55	57

Y empezamos a calcular los parámetros estadísticos.

Medias:

$$\overline{x} = \frac{\sum x_i}{n} = \frac{14}{4} = 3'5 \qquad \overline{y} = \frac{\sum y_i}{n} = \frac{13}{4} = 3'25$$

Varianzas:

$$s_x^2 = \frac{\sum x_i^2}{n} - \overline{x}^2 = \frac{62}{4} - 3'5^2 = 3'25 \qquad s_y^2 = \frac{\sum y_i^2}{n} - \overline{y}^2 = \frac{55}{4} - 3'25^2 = 3'19$$

Desviaciones típicas:

$$s_x = \sqrt{3'25} = 1'80 \qquad s_y = \sqrt{3'19} = 1'78$$

Covarianza:

$$s_{xy} = \frac{\sum x_i.y_i}{n} - \overline{x}\cdot\overline{y} = \frac{57}{4} - 3'5 \cdot 3'25 = 2'87$$

Por fin, el coeficiente de correlación es

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{2'87}{1'80 \cdot 1'78} = 0'89$$

ACTIVIDADES

5. Multiplicar los todos los datos anteriores por 10 y calcular el nuevo coeficiente de correlación. Es decir, ahora consideramos los datos:

x_i	10	30	40	60
y_i	10	20	50	50

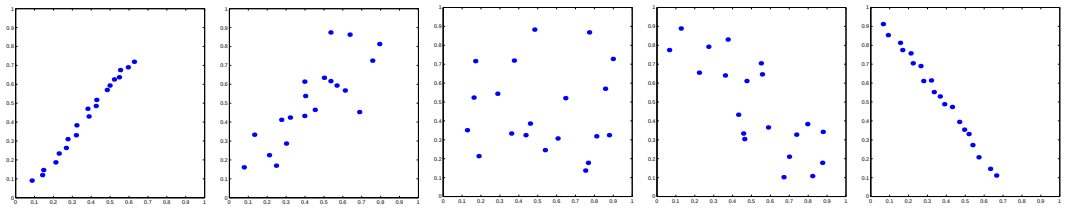
Ya sabemos calcular el coeficiente de correlación. Veamos ahora cómo lo podemos interpretar.

En primer lugar, es evidente que el signo del coeficiente de correlación y el de la covarianza coinciden, ya que las desviaciones típicas que aparecen en el denominador son siempre positivas (recordemos que las desviaciones típicas son raíces cuadradas y, por tanto, su resultado siempre es positivo.) Pero además, se puede demostrar que $r^2 \leq 1$, lo cual implica que r siempre es un número comprendido entre -1 y 1,

$$\boldsymbol{-1 \leq r \leq 1}$$

El hecho de que r esté más próximo a -1, a 1, o a 0 está directamente relacionado con la forma de la nube de puntos y, por tanto, con el tipo de dependencia entre las variables estadísticas.

En las gráficas de la figura 11.9 hemos dibujado diferentes nubes de puntos junto a los valores de sus correspondientes coeficientes de correlación.



$r = 0'98$ $r = 0'81$ $r = -0'02$ $r = -0'79$ $r = -0'99$

Figura 11.9: Coeficiente de correlación y nube de puntos

Según vemos en las figuras, cuando los valores de r son positivos, tenemos dependencia lineal positiva, tanto más fuerte, cuanto más próximo a 1 esté el valor de r. Cuando los valores de r son negativos, hay dependencia lineal negativa, tanto más fuerte, cuanto más próximo a -1 esté el valor de r. Por último, en el diagrama central tenemos un ejemplo de una nube de puntos en la que no se aprecia ninguna tendencia positiva o negativa concreta, en este caso, el valor de r está cercano a 0.

Por último, si $r = 1$ los puntos estará situados sobre una recta de pendiente positiva, es decir, no es que se parezcan a una recta, sino que están alineados. Si $r = -1$ los puntos estarán alineados en una recta con pendiente negativa. Y cuando $r = 0$,

estaríamos en el caso en el que no hay absolutamente ninguna dependencia o correlación, se dice a veces que las variables son *incorreladas.*

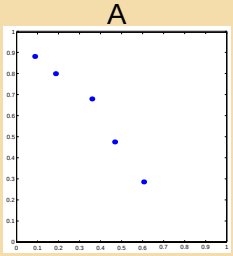
A la hora de calcular el coeficiente de correlación hay que tener una pequeña precaución. Lo habitual es que aparezcan decimales casi desde el primer cálculo, de los cuales tomamos los 2 o los 3 primeros, bien eliminando los demás (esto se llama truncar), o bien redondeando al segundo, al tercero, etc. Sea cual sea la manera en la que lo hagamos, cuando llegamos al valor de r se han acumulado pequeños errores que, finalmente nos pueden dar un valor de r poco mayor que 1, por ejemplo, 1'001. Pero esto, teóricamente es imposible, en estos casos lo mejor es rehacer los cálculos tomando un mayor número de decimales, con el objeto de aumentar la exactitud.

ACTIVIDADES

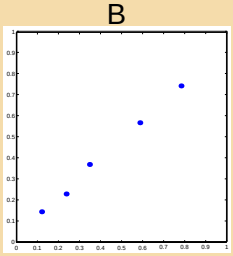
6. Asignar a cada una de las nubes de puntos siguientes su valor del coeficiente de correlación de entre los siguientes:

$$r = 0'60; \qquad r = 0'99; \qquad r = -0'98; \qquad r = -0'26.$$

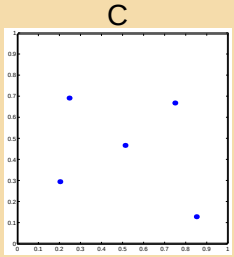
A



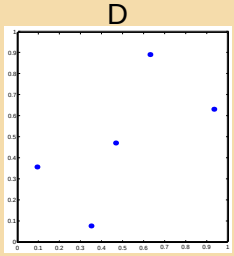
B



C



D



Recuerda

- ✓ El *coeficiente de correlación lineal* mide el grado de asociación lineal entre dos variables estadísticas. Se calcula mediante la fórmula

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

- ✓ Verifica $-1 \leq r \leq 1$.

  Si $r > 0$, hay dependencia lineal positiva entre las variables.

  Si $r < 0$, hay dependencia lineal negativa entre las variables.

  Si $r = 0$, no hay dependencia, las variables son incorreladas.

- ✓ La dependencia es mayor cuanto mayor sea el valor absoluto de r.

5. Regresión lineal

Si calculásemos el coeficiente de correlación de la nube de puntos de la figura 11.10, obtendríamos un resultado positivo. En efecto, se observa que hay una dependencia lineal positiva, ya que los puntos parecen acumularse alrededor de una recta de pendiente positiva, recta esta que también se ha dibujado en la figura. Ahora bien, ¿cuál es la recta que mejor se aproxima a los puntos? Esta recta cuya ecuación vamos a aprender a calcular es la llamada **recta de regresión**.

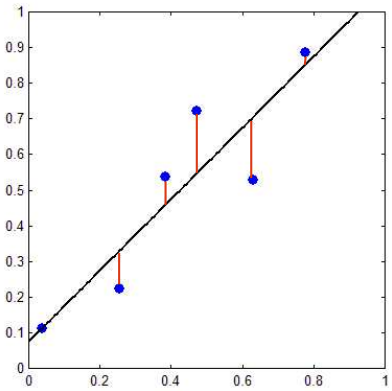


Figura 11.10: Recta de regresión

Cuando queremos estudiar la recta que mejor se aproxima a la nube de puntos, estamos haciendo *regresión lineal*. Sin embargo, hemos visto a lo largo de esta unidad otras nubes de puntos que más que acercarse a una recta, parecían aproximarse a alguna curva (una parábola por ejemplo). En este caso hablaríamos de *regresión no lineal*. En este curso sólo nos ocuparemos de la *regresión lineal*.

5.1. Rectas de regresión

Hemos dicho antes que la *recta de regresión* es la recta que mejor se aproxima a la nube de puntos. Vamos a intentar aclarar qué queremos decir con la expresión “mejor se aproxima".

Empezaremos diciendo que hay dos formas de precisar esta idea, que dan lugar a dos rectas de regresión, que se van a llamar, respectivamente, **recta de regresión de Y sobre X** y **recta de regresión X sobre Y**.

La recta de regresión de Y sobre X aparece como aquella recta que hace que las diferencias entre las ordenadas de la recta y las ordenadas de los puntos (estas diferencias son los segmentos verticales que se han dibujado en la figura 11.10, es decir, son las diferencias entre la altura del punto y la de la recta) sea, en promedio, mínima. Debido a que estas diferencias son positivas y negativas, dependiendo de que el punto esté por debajo o por encima de la recta, realmente lo que se hace es elevarlas al cuadrado, sumarlas y dividir entre el número de puntos, para calcular el promedio. Bien, pues imponiendo la condición de que este promedio sea mínimo, se puede demostrar que la ecuación que la verifica es

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

La recta de regresión de X sobre Y se obtiene de manera análoga, imponiendo la condición de que sean mínimas las diferencias entre las abscisas. De esta forma, se llega a la ecuación

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$

Es fácil darse cuenta de que las dos rectas de regresión pasan por el centro de masas de la nube de puntos $(\overline{x}, \overline{y})$ y además, ambas están inclinadas hacia el mismo lado, es decir, ambas tienen pendiente positiva o pendiente negativa, precisamente dependiendo de que la haya dependencia lineal positiva o negativa.

Para calcularlas es necesario calcular las medias, las varianzas y la covarianza. En el apartado siguiente vamos a ver un ejemplo.

5.2. Estimaciones con las rectas de regresión

¿Para qué sirven las rectas de regresión? Pues sirven para hacer estimaciones de una de las variables sobre la otra. Vamos a ver qué significa esto mediante un ejemplo que ya apareció cuando empezamos a hablar de variables estadísticas bidimensionales, las notas de la asignatura de Matemáticas y de la asignatura de Física y Química de un grupo de alumnos.

x_i	2	4	5	6	7	8	8	9
y_i	3	5	4	5′5	5	7′5	8	10

donde los x_i son las notas de Matemáticas y los y_i las de Física y Química.

La nube de puntos de estos datos, que ya habíamos dibujado, es la de la figura 11.11 y sugiere que cuando la nota de Matemáticas es alta, también lo es la de Física y Química; y lo mismo cuando la nota es baja.

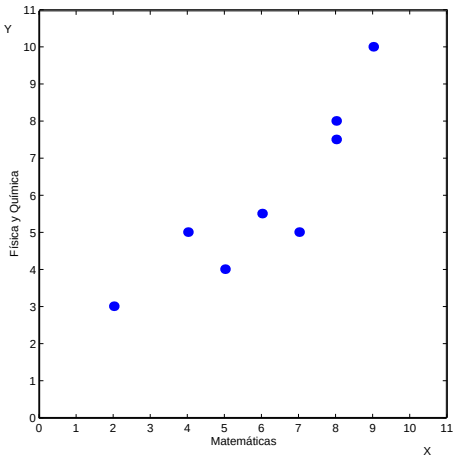


Figura 11.11: Dependencia positiva

Este grado de dependencia o de correlación la podemos cuantificar mediante el cálculo del coeficiente de correlación. Omitiremos la tabla y algunos de los cálculos que hemos repetido ya en varios ejemplos. (Se sugiere como ejercicio verificar los resultados.)

Medias:

$$\overline{x} = 6'125; \qquad \overline{y} = 6.$$

Varianzas:

$$s_x^2 = 4'8594 \qquad s_y^2 = 4'6875.$$

Desviaciones típicas:

$$s_x = 2'2044 \qquad s_y = 2'1651.$$

Covarianza:

$$s_{xy} = 4'25.$$

Coeficiente de correlación:

$$r = \frac{s_{xy}}{s_x \cdot s_y} = 0'8905.$$

Lo que confirma nuestra observación sobre la nube de puntos. Existe una correlación positiva fuerte entre las dos variables.

Ahora bien, supongamos que un estudiante ha hecho los dos exámenes y sólo conoce la nota de Matemáticas que es un 6'5. ¿Qué nota puede esperar en la asignatura de Física y Química? Para hacer esta estimación es para lo que sirve la recta de regresión de Y sobre X. Calculamos su ecuación sustituyendo en la expresión

$$y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$$

$$y - 6 = \frac{4'25}{4'8594}(x - 6'125)$$

que, simplificando y despejando y se convierte en

$$y = 0'8746x + 0'6431$$

Ahora, para estimar la nota de Física y Química, sustituimos en la ecuación de la recta el valor $x = 6'5$, entonces, $y = 6'328$, que será aproximadamente la nota que puede obtener.

Otro estudiante sabe que ha obtenido 9 en Física y Química. ¿Qué nota puede esperar en Matemáticas? Ahora sabemos que $y = 9$ y queremos calcular x. Para este caso, es más adecuada la recta de regresión de X sobre Y, ya que queremos calcular x conocido y.

Calculamos la recta de regresión de X sobre Y sustituyendo en la expresión

$$x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$$

$$x - 6'125 = \frac{4'25}{4'6875}(y - 6)$$

Simplificamos y despejamos x y obtenemos

$$x = 0'9067y + 0'685.$$

Para estimar la nota de Matemáticas, sustituimos en la ecuación de esta recta el valor $y = 9$, con lo que se obtiene $x = 8'8453$.

¿Qué fiabilidad tienen estas estimaciones? Desde luego, no se puede esperar que aporten el resultado exacto. Su exactitud dependerá precisamente del valor del coeficiente de correlación. Cuanto más próximo esté el coeficiente de correlación a -1 o a 1, mejor será la estimación. Por otra parte, hay que tener en cuenta que la estimación sólo tendrá sentido si el valor que se sustituye está en el rango de los datos que se tienen. Por ejemplo, si tenemos datos de peso y estatura de un grupo de personas y los pesos oscilan entre 60 y 70 kilogramos, no tendría sentido, con estos datos, intentar hacer una estimación de la estatura de una persona que pese 120 kilogramos.

ACTIVIDADES

7. Se ha preguntado a un grupo de personas cuántas horas semanales dedican a hacer deporte (X) y cuántas horas semanales dedican a ver la televisión. Las respuestas han sido las siguientes:

x_i	0	4	7	8	10
y_i	20	15	5	4	1

Calcular las ecuaciones de las dos rectas de regresión. Utilizando la recta adecuada, estimar el tiempo que dedicaría a ver la televisión una persona que dedica semanalmente 5 horas a hacer deporte.

Recuerda

✓ Una recta de regresión es la que mejor se ajusta a una nube de puntos. Hay dos:

Recta de regresión de Y sobre X: $y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x})$

Recta de regresión de X sobre Y: $x - \overline{x} = \frac{s_{xy}}{s_y^2}(y - \overline{y})$

✓ Las rectas de regresión sirven para hacer estimaciones de una variable sobre la otra.
Si conocemos x, utilizamos la recta de regresión de Y sobre X para calcular y.
Si conocemos y, utilizamos la recta de regresión de X sobre Y para calcular x.