ESTADÍSTICA DESCRIPTIVA BIDIMENSIONAL

UD000573_V(02)

ÍNDICE

MOTIVACIÓN

¿Qué ocurre si queremos estudiar a un individuo bajo dos variables simultáneas? Esto será posible gracias a la Estadística Descriptiva Bidimensional.

Nos permitirá conocer, además de cómo se comportan las variables por sí solas, cómo son las dos características de ese individuo de manera conjunta y si ambas variables están relacionadas entre ellas o no, es decir, si son dependientes unas de otras.

El estudio de la Estadística Descriptiva Bidimensional se hace necesario en mediciones de ingeniería, estudios de poblaciones, sociología, etc., ya que nos permite conocer la dependencia o no de una variable sobre otra.

PROPÓSITOS

Con el estudio de esta unidad didáctica, conseguirás:

- Crear tablas de variables estadísticas bidimensionales.
- Calcular la covarianza.
- Estudiar la correlación y calcular el coeficiente de correlación de Pearson.
- Realizar un estudio analítico sobre la regresión lineal.

PREPARACIÓN PARA LA UNIDAD

En esta unidad didáctica ampliaremos los conceptos de la Estadística Descriptiva unidimensional realizando, en este caso, un estudio bidimensional.

Introduciremos los conceptos de regresión y correlación, descubiertos recientemente por Dalton, que nos permitirán observar la relación existente entre dos variables estadísticas.

Podrás observar que la mecánica de los ejercicios es sencilla, pues se basa en la aplicación directa de las fórmulas respectivas, si bien es importante prestar atención a los cálculos iniciales que, de ser operados erróneamente, alterarán los resultados considerablemente.

1. VARIABLES ESTADÍSTICAS BIDIMENSIONALES

Hablamos de variables estadísticas bidimensionales cuando, de una serie de individuos de una población, se estudian dos caracteres simultáneamente, obteniéndose así dos series de datos.

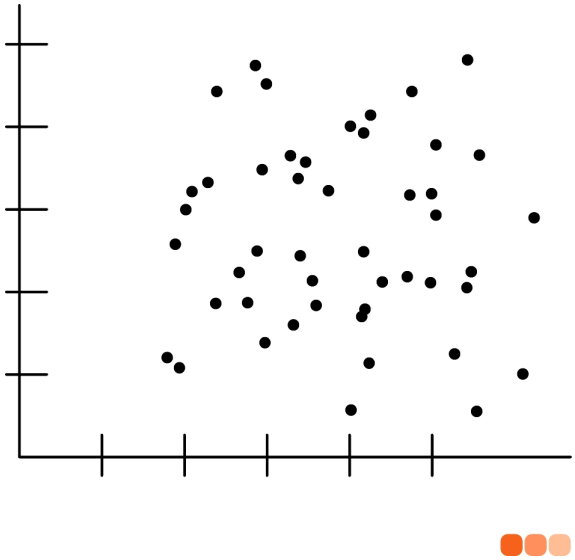
Así, una variable estadística bidimensional es un par de dos variables unidimensionales, y se representa (X, Y). En consecuencia, los valores que toma la variable bidimensional también son pares de valores de ambas variables; por ejemplo, si X toma los valores x_1, x_2..., x_n y la variable Y, los valores y_1, y_2..., y_n, la variable bidimensional (X, Y) tomará los valores (x_1, y_1), (x_2, y_2)..., (x_n, y_n).

Ejemplo

Son variables bidimensionales:

- La producción y las ventas de una empresa.
- La altura y el peso asociados a una determinada edad.
- Los ingresos y gastos familiares de cada trabajador de una empresa.

Así, podemos representar los valores de una variable bidimensional en un sistema de coordenadas tomando como abscisa el primer valor del par (x_i) y como ordenada el segundo valor del par (y_i). La representación de todos los valores de la variable se llama nube de puntos o diagrama de dispersión.



Nube de puntos

1.1. TABLAS BIDIMENSIONALES DE FRECUENCIAS

Veamos algunos ejemplos:

1. Las calificaciones de 40 alumnos de Matemáticas y Física han sido las siguientes:

X = calificación de Matemáticas	3	4	5	6	6	7	7	8	10
Y = calificación de Física	2	5	5	6	7	6	7	9	10
Número de alumnos	4	6	12	4	5	4	2	1	2

En esta distribución los valores de la variable bidimensional (X, Y) aparecen repetidos; en consecuencia, la frecuencia absoluta de cada par viene dada por el número de alumnos que han obtenido las correspondientes calificaciones.

A este tipo de tabla se le llama **tabla simple.**

2. Se han clasificado 50 familias dependiendo del número de hijos (X) e hijas (Y), obteniéndose los siguientes resultados:

X / Y	0	1	2	3	4	5	6	
0	2	-	4	3	1	-	-	10
1	3	-	9	-	-	3	-	15
2	-	6	-	6	-	-	1	13
3	1	4	-	-	2	1	-	8
4	-	-	2	-	1	-	-	3
5	-	-	-	1	-	-	-	1
	6	10	15	10	4	4	1	50

A este tipo de tabla se le llama **de doble entrada,** y se utiliza cuando existe una gran cantidad de datos o bien estos se encuentran agrupados en clases.

Reto

Realiza el siguiente ejercicio.

¿Qué es la nube de puntos?

Solución:

El conjunto de puntos que representa los pares de las variables bidimensionales (x, y) bajo el sistema de ejes cartesianos.

1.2. CÁLCULO DE PARÁMETROS

Si hay pares de datos que se repiten, se agrupan, siendo $\mathbf{n_{ij}}$ la frecuencia absoluta del par (x_i, x_j).

Recordemos las fórmulas para distribuciones de variables estadísticas unidimensionales:

- **Frecuencias marginales:** son las frecuencias de x_i e y_j.

$$\text{Frecuencia de } x_i = \sum_{j=1}^{n} n_{ij} = f_i$$

$$\text{Frecuencia de } y_j = \sum_{i=1}^{n} n_{ij} = f_j$$

- **Medias:**

$$\bar{x} = \frac{\sum x_i f_i}{N} \qquad \bar{y} = \frac{\sum y_j f_j}{N}$$

$$\text{Donde N} = \sum f_i = \sum f_j$$

- **Varianzas:**

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{N} = \frac{\sum x_i^2 f_i}{N} - \bar{x}^2$$

$$S_y^2 = \frac{\sum (y_j - \bar{y})^2 f_j}{N} = \frac{\sum y_j^2 f_j}{N} - \bar{y}^2$$

Recuerda

La raíz cuadrada positiva de las varianzas se llama desviación típica, y se representa por S_x y S_y.

Veamos un nuevo parámetro para variables bidimensionales:

- **Covarianza:**

  Se representa por S_{xy}:

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})n_{ij}}{N} = \frac{\sum x_i y_i n_{ij}}{N} - \bar{x}\bar{y}$$

  Más adelante veremos el significado de la covarianza, así como su interpretación según el signo.

Truco

Para calcular los parámetros anteriores resulta útil construir una tabla con cada uno de los productos que aparecen en los sumandos.

Calcula la covarianza del ejemplo 1 del punto 1.1. Tablas bidimensionales de frecuencias.

X = calificación de Matemáticas	3	4	5	6	6	7	7	8	10
Y = calificación de Física	2	5	5	6	7	6	7	9	10
Número de alumnos	4	6	12	4	5	4	2	1	2

Debemos aplicar la formula

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})n_{ij}}{N} = \frac{\sum x_i y_i n_{ij}}{N} - \bar{x}\bar{y}$$

Veamos qué significa cada letra:

x_i, y_i: cada uno de los valores que toman x e y en cada momento.

$\overline{x}, \overline{y}$: la media de x y de y, respectivamente ($\overline{x} = 6{,}22$, $\overline{y} = 6{,}33$).

n_{ij} : número de alumnos para el valor en x_i e y_j.

N: número total de alumnos (40).

Sustituimos y obtenemos:

$$S_{xy} = \frac{\sum x_i y_i n_{ij}}{N} - \overline{xy} =$$

$$= \frac{(3 \cdot 2 \cdot 4) + (4 \cdot 5 \cdot 6) + (5 \cdot 5 \cdot 12) + (6 \cdot 6 \cdot 4) + (6 \cdot 7 \cdot 5) + (7 \cdot 6 \cdot 4) + (7 \cdot 7 \cdot 2) + (8 \cdot 9 \cdot 1) + (10 \cdot 10 \cdot 2)}{40}$$

$$- 6{,}22 \cdot 6{,}33 = \frac{1336}{40} - 39{,}40 = 94{,}19$$

Recuerda

Recuerda que tienes a tu disposición un servicio de tutorías para resolver cualquier duda que te pueda surgir, bien mediante carta, Campus Virtual, fax o llamada telefónica.

2. REGRESIÓN LINEAL

Con la regresión lineal se trata de determinar si existe una curva que se ajuste lo más posible a la nube de puntos **(curva de regresión).**

Atención

El gráfico a través del cual se representa una variable estadística bidimensional es la nube de puntos o diagrama de dispersión, donde cada valor (x_i,y_i) viene representado por un punto.

Dada una variable estadística bidimensional (X, Y), hallar la ecuación de la recta que mejor represente a la variable; es decir, que más se ajuste a la nube de puntos (x_i, y_j).

El método utilizado se denomina método de los mínimos cuadrados. Se trata de determinar la recta para la cual la suma

$$\sum (y_j - y_j')^2 ,$$

es mínima, donde y_j' es la ordenada de la recta correspondiente a la abscisa x_j:

- x_i = valor observado de la variable independiente.
- y_j' = valor estimado mediante la recta de regresión.
- y_j = valor observado de la variable dependiente.

De la aplicación de este método se observa que las rectas de regresión pasan por el punto $(\bar{x}, \bar{y})$.

2.1. RECTA DE REGRESIÓN DE Y SOBRE X

La expresión de la recta de regresión de la variable Y sobre la variable X es la siguiente:

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} \cdot (x - \bar{x})$$

Donde

$$r_{yx} = \frac{S_{xy}}{S_x^2}$$

es el coeficiente de regresión de Y sobre X.

Sustituyendo en esta ecuación los valores de x podemos obtener, con cierta aproximación, los valores esperados para la variable y, que llamamos estimaciones o previsiones.

2.2. RECTA DE REGRESIÓN DE X SOBRE Y

La expresión de la recta de regresión de la variable X sobre la variable Y es la siguiente:

$$x - \bar{x} = \frac{\sigma_{xy}}{\sigma_y^2} \cdot (y - \bar{y})$$

Donde

$$r_{xy} = \frac{S_{xy}}{S_y^2}$$

es el coeficiente de regresión de X sobre Y.

Con esta ecuación estimamos valores de x, conocidos los de y.

La forma más utilizada para calcular una curva de regresión se realiza mediante **el método de mínimos cuadrados,** que se calcula así:

$$\sum (y_j - y_j')^2$$

3. CORRELACIÓN LINEAL

Cuando hablamos de correlación entre dos variables hacemos referencia a la dependencia que existe entre ellas.

Así, podemos distinguir los siguientes tipos de correlación:

1. **Correlación lineal o curvilínea:** si la nube de puntos se agrupa en torno a una línea recta o una curva.
2. **Correlación funcional:** si existe una función que se ajusta a los valores de la distribución.

Además, según sea la relación entre las variables diremos que la correlación es:

1. **Directa o positiva,** si cuando crece una de las variables la otra también crece.
2. **Inversa o negativa,** si cuando crece una de las variables la otra decrece.
3. **Nula,** si no existe ninguna relación entre las variables. En este caso los puntos en la nube de puntos están distribuidos al azar. Se dice que ambas variables están incorreladas.

Reto

Completa y realiza el siguiente ejercicio.

La correlación es de tipo _____ [1] si existe una función que satisfaga todos los valores de las distribuciones.

La correlación es _____[2] cuando no existe ninguna relación entre las variables.

Se denomina correlación positiva o _____[3].

Solución:

[1] Funcional.

[2] Nula.

[3]Directa.

3.1. COEFICIENTE DE CORRELACIÓN LINEAL

El coeficiente de correlación lineal o coeficiente de Pearson es el valor numérico que mide de forma cuantitativa la dependencia funcional entre las variables X e Y.

Se representa por r:

$$r = \frac{S_{xy}}{S_x \cdot S_y}$$

El signo del coeficiente de correlación lineal tiene gran importancia puesto que nos indica el tipo de correlación que existe entre las variables. Si r es positivo, la correlación es directa, si es negativo, la correlación será inversa, y si r es igual a cero, no existe correlación entre las variables.

Además, podemos observar que el signo del coeficiente de correlación lineal coincide con el signo de la covarianza, ya que, en el denominador de r, encontramos las desviaciones típicas, que siempre son valores positivos.

Se demuestran las siguientes propiedades:

1. El valor de r varía entre -1 y 1.

2. Si el coeficiente r > 0, al aumentar una variable también aumenta la otra. La correlación es positiva y será más fuerte a medida que se acerque a 1, y más débil a medida que se aproxime a 0. En este caso se dice que las variables están en dependencia aleatoria.
3. Si el coeficiente r < 0, al aumentar una variable disminuye la otra. La correlación es negativa y será más fuerte a medida que se acerque a –1, y más débil a medida que se aproxime a 0. En este caso se dice que las variables están en dependencia aleatoria.
4. Si r = ±1, tenemos correlación perfecta. Hay dependencia funcional entre las variables y las dos rectas de regresión coinciden. En este caso se dice que las dos variables están en dependencia funcional.
5. Si r = 0, las rectas de regresión son perpendiculares entre sí. La correlación es nula y se dice que las dos variables son aleatoriamente independientes.

3.1.1. Observación

La fiabilidad de los cálculos obtenidos a partir de la recta de regresión será tanto mejor cuanto mayor sea el coeficiente de correlación lineal en valor absoluto. Entonces:

- Si r es muy pequeño, no tiene sentido realizar ningún tipo de estimaciones o previsiones.
- Si r es próximo a –1 o a 1, probablemente los valores reales se aproximen a las estimaciones.
- Si r = -1 o 1, las estimaciones realizadas coinciden con los valores reales.

Atención

El signo de la covarianza nos va a dar el signo del coeficiente de correlación lineal, en el que:

- Si $S_{xy}>0$, entonces la correlación es directa.
- Si $S_{xy}<0$, entonces la correlación es inversa.
- Si $S_{xy}=0$, entonces la correlación es nula.

4. EJERCICIO RESUELTO

Se realiza un estudio acerca de la relación existente entre la mortalidad infantil en cada país y el número de camas de hospital por cada 1.000 habitantes. Para ello tenemos los siguientes datos sobre 10 países concretos que pueden considerarse representativos del resto, donde X representa el número de camas por cada 1.000 habitantes e Y, el porcentaje de mortalidad infantil. Calcular:

a) El coeficiente de correlación lineal.

b) Para un país que tiene 150 camas, el índice de mortalidad infantil que se espera.

TABLA DE DATOS:

X	50	100	70	60	120	180	200	250	30	90
Y	5	2	2,5	3,75	4	1	1,25	0,75	7	3

SOLUCIÓN:

Construiremos una tabla auxiliar que nos ayudará en los cálculos de los parámetros:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
50	5	2.500	25	250
100	2	10.000	4	200
70	2,5	4.900	6,25	175
60	3,75	3.600	14,0625	225

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
	120	4	14.400	16	480
	180	1	32.400	1	180
	200	1,25	40.000	1,5625	250
	250	0,75	62.500	0,5625	187,5
	30	7	900	49	210
	90	3	8.100	9	270
SUMAS	1.150	30,25	179.300	126,44	2.427,5

Podemos calcular fácilmente las medias, las desviaciones típicas y la covarianza:

$$\bar{x} = \frac{1.150}{10} = 115 \text{ camas por cada 1.000 habitantes.}$$

$$S_x = \sqrt{\frac{179.300}{10} - 115^2} = 68{,}59.$$

$$\bar{y} = \frac{30{,}25}{10} = 3{,}025\% \text{ de mortalidad infantil.}$$

$$S_y = \sqrt{\frac{126{,}44}{10} - 3{,}025^2} = 1{,}87.$$

$$S_{xy} = \frac{2.427{,}5}{10} - 115 \cdot 3.025 = -105.125.$$

a) Coeficiente de correlación lineal:

$$r = \frac{-105.125}{68{,}59 \cdot 1{,}87} = -0{,}82 \Rightarrow \qquad \boxed{r = -0{,}82.}$$

Es decir, una correlación negativa y alta.

b) Recta de regresión de y sobre x:

$$y - 3{,}025 = \frac{-105.125}{68{,}59^2} \cdot (x - 115) \Rightarrow y = -0{,}22 + 5{,}59$$

Veamos la estimación para x = 150 camas. Sustituyendo en la recta obtenida, resulta como índice de mortalidad infantil esperado:

$$y = 2{,}29\%$$

CONCLUSIONES

En numerosas ocasiones se hace necesario el estudio simultáneo de dos o más variables en un individuo determinado. Este estudio resulta posible gracias a la Estadística Descriptiva Multidimensional. En esta unidad didáctica, nos hemos centrado en el estudio de dos variables; es por lo que se ha denominado bidimensional.

El estudio del coeficiente de correlación nos permite conocer la dependencia de las dos variables analizadas.

RECAPITULACIÓN

- **Medias:**

$$\bar{x} = \frac{\sum x_i f_i}{N} \quad \bar{y} = \frac{\sum y_j f_j}{N}$$

$$N = \sum f_i = \sum f_j$$

- **Varianzas:**

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2 f_i}{N} = \frac{\sum x_i^2 f_i}{N} - \bar{x}^2$$

$$S_y^2 = \frac{\sum (y_j - \bar{y})^2 f_j}{N} = \frac{\sum y_j^2 f_j}{N} - \bar{y}^2$$

La raíz cuadrada positiva de las varianzas se llama desviación típica, y se representa por S_x y S_y.

- **Covarianza:**

Se representa por Sxy:

$$S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) n_{ij}}{N} = \frac{\sum x_i y_j n_{ij}}{N} - \bar{x} \cdot \bar{y}$$

- **Recta de regresión de Y sobre X:**

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} \cdot (x - \bar{x})$$

Donde

$$r_{yx} = \frac{S_{xy}}{S_x^2}$$

es el coeficiente de regresión de Y sobre X.

AUTOCOMPROBACIÓN

1. El coeficiente de correlación lineal es:

a) El cociente entre la varianza y el producto de las varianzas marginales.

b) El cociente entre la covarianza y el producto de las varianzas marginales.

c) El cociente entre el producto de las varianzas marginales y la varianza.

d) El cociente entre el producto de las varianzas marginales y la covarianza.

2. La covarianza de la siguiente distribución conjunta es:

X / Y	0	1	2	3
1	0,05	0,1	0,3	0,002
2	0,1	0,15	0,2	0,08

a) 0,7475.

b) 1,53.

c) 0,2491.

d) -0,0315.

3. Las rectas de regresión de Y sobre X y de X sobre Y no coinciden si:

a) r = 0.

b) R = 1.

c) R = -1.

d) Todas son correctas.

4. Consideremos las calificaciones en Matemáticas (X) y Física (Y) de 15 alumnos. Los resultados son:

X	8	4	5	9	2	3	8	4	5	8	9	10	5	4	3
Y	7	5	6	7	5	4	8	6	4	7	7	9	6	5	5

Las varianzas marginales son:

a) 5,8 y 2.

b) 6,23 y 6,06.

c) 2 y 6,293.

d) 6,06 y 5,8.

5. La recta de regresión de Y sobre X del ejercicio anterior es:

a) $y = 38,2x + 3,025$.

b) $y = 3,24x + 2,025$.

c) $y = 0,4849x + 3,2475$.

d) $y = 38,2x + 30,48$.

6. Consideremos las calificaciones en Matemáticas (X) y Física (Y) de 15 alumnos. Los resultados son:

X	8	4	5	9	2	3	8	4	5	8	9	10	5	4	3
Y	7	5	6	7	5	4	8	6	4	7	7	9	6	5	5

La covarianza es:

a) 38,2.

b) 3,052.

c) 6,293.

d) 2,004.

7. Consideremos las calificaciones en Matemáticas (X) y Física (Y) de 15 alumnos. Los resultados son:

X	8	4	5	9	2	3	8	4	5	8	9	10	5	4	3
Y	7	5	6	7	5	4	8	6	4	7	7	9	6	5	5

¿Cuál es la nota esperada en Física de un alumno que ha obtenido un 5,5 en Matemáticas?

a) 5,9144.

b) 6,385.

c) 4,063.

d) Ninguna es correcta.

8. Las rectas de regresión son perpendiculares entre sí si:

a) r = 0.

b) r = 1.

c) r = -1.

d) Ninguna es cierta.

9. Si la correlación es nula la dependencia es:

a) Fuerte.

b) Débil.

c) Nula.

d) Inversa.

10. Señala la afirmación verdadera:

a) La covarianza es siempre positiva.

b) Toda distribución bidimensional tiene una sola distribución marginal.

c) La línea de regresión es la línea que pasa por todos los puntos de la nube de puntos.

d) Las rectas de regresión de X sobre Y y de Y sobre X pasan siempre por el punto ($\overline{X}, \overline{Y}$).

SOLUCIONARIO

1.	b	**2.**	d	**3.**	a	**4.**	c	**5.**	c
6.	b	**7.**	a	**8.**	a	**9.**	c	**10.**	d

PROPUESTAS DE AMPLIACIÓN

En esta unidad didáctica hemos estudiado qué es una variable bidimensional y dos tipos de coeficientes, pero dentro del campo de la estadística existen diversas maneras de comprobar el comportamiento de estas variables a través de distribuciones marginales. Asimismo, existen diferentes formas de representar estas variables en sus correspondientes diagramas.

Si se desea profundizar más en este campo se pueden consultar páginas web o material especializado sobre este tipo de mediciones.

Te proponemos que compruebes si tienes materiales complementarios, o clases grabadas dentro de la unidad. Si es así, descárgalos para ampliar la información sobre el tema y **recuerda marcar he terminado.**

Te proponemos también que entres **en la sección de agenda** y compruebes qué clases en directo y/o talleres tienes disponibles, para complementar tus estudios, o tu preparación a la hora de afrontar los exámenes.

BIBLIOGRAFÍA

- VV.AA. EULER. *Matemáticas I.* Madrid: S.M , 2001.
- VV.AA. EULER. *Matemáticas II* aplicadas a las ciencias sociales. Madrid: S.M, 2000.
- COLERA, J y otros. *Matemáticas. En tus manos.* Madrid: Anaya, 2002.
- BESCÓS, E y PEÑA, Z. *Proyecto Exedra. Matemáticas I.* Madrid: Oxford, 2001.
- VV.AA. *Matemáticas I.* Madrid: Edelvives, 2003.
- VV.AA. *Matemáticas II.* Madrid: Edelvives, 2003.