Variables estadísticas bidimensionales

BENITO J. GONZÁLEZ RODRÍGUEZ (bjglez@ull.es)
DOMINGO HERNÁNDEZ ABREU (dhabreu@ull.es)
MATEO M. JIMÉNEZ PAIZ (mjimenez@ull.es)
M. ISABEL MARRERO RODRÍGUEZ (imarrero@ull.es)
ALEJANDRO SANABRIA GARCÍA (asgarcia@ull.es)

Departamento de Análisis Matemático
Universidad de La Laguna

Índice

1. Introducción

En el análisis estadístico es conveniente a veces contrastar los datos procedentes de dos caracteres estudiados sobre un mismo individuo. En este sentido se plantea la consideración de variables estadísticas bidimensionales, así como la detección de posibles relaciones entre los dos caracteres investigados.

Definición 1.1. *Una* variable estadística bidimensional *es el conjunto* (X,Y) *de valores que pueden tomar dos caracteres diferentes X e Y medidos sobre cada uno de los individuos de una población o muestra. Los caracteres X e Y se denominan* caracteres *o* variables marginales *y pueden ser ambos cuantitativos, ambos cualitativos o uno de cada tipo; a su vez, los caracteres cuantitativos puede ser variables estadísticas tanto discretas como continuas.*

Ejemplo 1.2. *La siguiente tabla muestra algunos ejemplos de variables bidimensionales:*

(X,Y)	X	Y
(sexo, color del pelo)	*cualitativo*	*cualitativo*
(profesión, antigüedad en la empresa)	*cualitativo*	*cuantitativo*
(peso, estatura)	*cuantitativo (v.e. continua)*	*cuantitativo (v.e. continua)*
(número de hermanos, número de hijos)	*cuantitativo (v.e. discreta)*	*cuantitativo (v.e. discreta)*
(temperatura, pulsaciones)	*cuantitativo (v.e. continua)*	*cuantitativo (v.e. discreta)*

2. Ordenación de los datos

Centraremos nuestra atención en el estudio de variables bidimensionales cuyos caracteres marginales X e Y son ambos cuantitativos. Cada uno de los valores correspondientes a la variable bidimensional (X,Y) se representa mediante un par ordenado (x_i,y_j), donde x_i es el valor que mide el primer carácter e y_j es el valor que mide el segundo carácter. En consecuencia, las variables marginales X e Y toman los valores $\{x_1,x_2,...,x_m\}$, $\{y_1,y_2,...,y_r\}$ para ciertos m, r, respectivamente, y la variable bidimensional (X,Y) tomará los valores $\{(x_i,y_j)\}_{1\leq i\leq m,1\leq j\leq r}$.

Definición 2.1. *El número de elementos que tienen el valor* x_i *para el primer carácter y el valor* y_j *para el segundo se denomina* frecuencia absoluta *del par* (x_i,y_j)*, y se denota* n_{ij}*; es decir,* n_{ij} *es el número de veces*

que aparece repetido el par (x_i, y_j) *en las observaciones. La* frecuencia relativa *del par* (x_i, y_j) *es*

$$f_{ij} = \frac{n_{ij}}{N},$$

donde N denota el número total de pares observados.

Adviértase que:

- $\sum_{i=1}^{m} \sum_{j=1}^{r} n_{ij} = N.$
- $\sum_{i=1}^{m} \sum_{j=1}^{r} f_{ij} = 1.$

Definición 2.2. *Se define la* frecuencia (absoluta) marginal *del valor* x_i *como la suma de las frecuencias correspondientes a los pares* (x_i, y_j)*, para* $1 \leq j \leq r$*:*

$$n_{x_i} = \sum_{j=1}^{r} n_{ij}.$$

Análogamente se define la frecuencia (absoluta) marginal *del valor* y_j*:*

$$n_{y_j} = \sum_{i=1}^{m} n_{ij}.$$

Nótese que n_{x_i} (respectivamente, n_{y_j}) representa el número de veces que aparece el valor x_i (respectivamente, y_j) en el total de pares obtenidos. Se verifica:

- $\sum_{i=1}^{m} n_{x_i} = \sum_{i=1}^{m} \sum_{j=1}^{r} n_{ij} = N.$
- $\sum_{j=1}^{r} n_{y_j} = \sum_{j=1}^{r} \sum_{i=1}^{m} n_{ij} = N.$

A partir de las frecuencias absolutas marginales se obtienen las frecuencias relativas marginales.

Definición 2.3. *Las* frecuencias relativas marginales *son los cocientes*

$$f_{x_i} = \frac{n_{x_i}}{N}, \quad f_{y_j} = \frac{n_{y_j}}{N}.$$

Para ellas, se cumple:

- $\sum_{i=1}^{m} f_{x_i} = \frac{1}{N} \sum_{i=1}^{m} n_{x_i} = 1.$

- $\sum_{j=1}^{r} f_{y_j} = \frac{1}{N}\sum_{j=1}^{r} n_{y_j} = 1$.

Todos estos datos se disponen en una tabla de doble entrada, como se indica a continuación:

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_r	n_{x_i}	f_{x_i}
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1r}	n_{x_1}	f_{x_1}
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2r}	n_{x_2}	f_{x_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ir}	n_{x_i}	f_{x_i}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
x_m	n_{m1}	n_{m2}	$\cdots$	n_{mj}	$\cdots$	n_{mr}	n_{x_m}	f_{x_m}
n_{y_j}	n_{y_1}	n_{y_2}	$\cdots$	n_{y_j}	$\cdots$	n_{y_r}	N	
f_{y_j}	f_{y_1}	f_{y_2}	$\cdots$	f_{y_j}	$\cdots$	f_{y_r}		1

En el caso de que alguna de las variables marginales esté agrupada en intervalos de clase, serán éstos los que figuren en la cabecera de la fila o columna correspondiente; el recuento de frecuencias se hará por clases, y se incorporarán a la tabla las marcas de clase.

Ejemplo 2.4. *Se ha elegido al azar en un colegio a* 30 *niños a los que se les ha tomado la edad en años y el peso en kilogramos, resultando la siguiente tabla:*

peso \ *edad*	x_i	9	10	11	12	n_{x_i}	f_{x_i}
$[20,25)$	22.5	1				1	$1/30 = 0.03$
$[25,30)$	27.5	2	1	1		4	$4/30 = 0.13$
$[30,35)$	32.5	1	3	4	2	10	$10/30 = 0.30$
$[35,40)$	37.5		2	3	5	10	$10/30 = 0.30$
$[40,45)$	42.5			1	4	5	$5/30 = 0.17$
n_{y_j}		4	6	9	11	30	
f_{y_j}		$4/30 = 0.13$	$6/30 = 0.20$	$9/30 = 0.30$	$11/30 = 0.37$		1

Observación 2.5. *A veces tenemos una tabla de la forma*

x_i	x_1	x_2	$\cdots$	x_n
y_i	y_1	y_2	$\cdots$	y_n

con $x_1 < x_2 < \ldots < x_i < \ldots < x_n$, $y_1 < y_2 < \ldots < y_j < \ldots < y_n$. *En tal caso la ordenación de los datos en una tabla de doble entrada no es significativa, ya que en el cuerpo central de la tabla resulta una matriz diagonal unitaria indicativa de que tanto las frecuencias absolutas de cada par* (x_i, y_i) *como las marginales de* x_i *e* y_i $(1 \leq i \leq n)$ *valen* 1.

3. Medidas de centralización y dispersión marginales

Definición 3.1. *Se llaman* medidas de centralización y dispersión marginales *las correspondientes a las variables marginales X e Y:*

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i n_{x_i}}{N}, \qquad \bar{y} = \frac{\sum_{j=1}^{r} y_j n_{y_j}}{N};$$

$$\sigma_x^2 = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_{x_i}}{N} = \frac{\sum_{i=1}^{m} x_i^2 n_{x_i}}{N} - \bar{x}^2,$$

$$\sigma_y^2 = \frac{\sum_{j=1}^{r} (y_j - \bar{y})^2 n_{y_j}}{N} = \frac{\sum_{j=1}^{r} y_j^2 n_{y_j}}{N} - \bar{y}^2;$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 n_{x_i}}{N}} = \sqrt{\frac{\sum_{i=1}^{m} x_i^2 n_{x_i}}{N} - \bar{x}^2},$$

$$\sigma_y = \sqrt{\frac{\sum_{j=1}^{r} (y_j - \bar{y})^2 n_{y_j}}{N}} = \sqrt{\frac{\sum_{j=1}^{r} y_j^2 n_{y_j}}{N} - \bar{y}^2}.$$

Se toman como x_i (respectivamente, y_j) valores de la variable o marcas de clase, según proceda.

Ejemplo 3.2. *Obtener las medidas de centralización y dispersión marginales para los datos del Ejemplo 2.4.*

RESOLUCIÓN. Son las siguientes:

$$\bar{x} = \frac{(22.5 \cdot 1) + (27.5 \cdot 4) + (32.5 \cdot 10) + (37.5 \cdot 10) + (42.5 \cdot 5)}{30} \simeq 34.833,$$

$$\bar{y} = \frac{(9\cdot 4)+(10\cdot 6)+(11\cdot 9)+(12\cdot 11)}{30} = 10.900,$$

$$\sigma_x^2 = \frac{(22.5^2\cdot 1)+(27.5^2\cdot 4)+(32.5^2\cdot 10)+(37.5^2\cdot 10)+(42.5^2\cdot 5)}{30} - 34.833^2$$

$$\simeq 26.245,$$

$$\sigma_x \simeq 5.123,$$

$$\sigma_y^2 = \frac{(9^2\cdot 4)+(10^2\cdot 6)+(11^2\cdot 9)+(12^2\cdot 11)}{30} - 10.900^2$$

$$= 1.090,$$

$$\sigma_y \simeq 1.044.$$

□

4. Representación gráfica

Tienen especial interés los denominados *diagramas de dispersión* o *nubes de puntos*. Si las variables marginales no están agrupadas en intervalos, se representa cada par (x_i, y_j) en un diagrama cartesiano. Si sólo una de ellas está agrupada se trabaja con sus marcas de clase, representando los pares resultantes mediante puntos del plano, como en el caso anterior. Si ambas variables marginales están agrupadas dividimos el plano en casillas, dibujando dentro de cada una un número de puntos igual a la frecuencia absoluta correspondiente a sendos intervalos en la X y en la Y.

En un diagrama de dispersión no quedan reflejadas las veces que se repite un par o un intervalo; hemos de recurrir a un diagrama de barras en tres dimensiones, de las cuales dos son para la variable bidimensional y la tercera (altura) para las frecuencias. Precisamente denominamos *diagramas de frecuencias* a las gráficas de este tipo. En los diagramas de frecuencias en donde las dos variables están agrupadas en intervalos, la frecuencia es el volumen del paralelepípedo correspondiente.

En la sección de ejercicios resueltos pueden verse algunos ejemplos de diagramas de dispersión.

5. Regresión y correlación

5.1. Regresión

Al observar dos caracteres en un mismo individuo se plantea naturalmente la cuestión de determinar la existencia de algún tipo de dependencia entre ellos, y si es posible hallar una expresión matemática que las

relacione. El problema de la regresión consiste precisamente en intentar ajustar al diagrama de dispersión una curva de ecuación conocida (recta, exponencial, parábola, hipérbola, etc.), sugerida por el propio diagrama, con el fin de poder efectuar una predicción del valor de una de las variables a partir de la otra. Cuando la función que mejor se ajusta a la nube de puntos es una recta nos hallamos ante un problema de *regresión lineal*.

Definición 5.1. *La* recta de regresión de Y sobre X *proporciona los valores aproximados de Y conocidos los de X, y tiene por ecuación*

$$r_{y_x}:\ y-\overline{y}=\frac{\sigma_{xy}}{\sigma_x^2}(x-\overline{x});$$

La recta de regresión de X sobre Y *proporciona los valores aproximados de X conocidos los de Y, y tiene por ecuación*

$$r_{x_y}:\ x-\overline{x}=\frac{\sigma_{xy}}{\sigma_y^2}(y-\overline{y}).$$

Aquí,

$$\sigma_{xy}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{r}(x_i-\overline{x})(y_j-\overline{y})n_{ij}}{N}=\frac{\sum_{i=1}^{m}\sum_{j=1}^{r}x_iy_jn_{ij}}{N}-\overline{x}\,\overline{y}$$

es la covarianza *de X e Y, mientras que $\overline{x}$, $\overline{y}$ son las medias marginales y σ_x^2, σ_y^2 las varianzas marginales de X e Y, respectivamente. Nótese que ambas rectas de regresión se cruzan en el punto $(\overline{x},\overline{y})$, llamado* centro de gravedad *de la distribución. Las pendientes de dichas rectas,*

$$\beta_{y_x}=\frac{\sigma_{xy}}{\sigma_x^2},\quad \beta_{x_y}=\frac{\sigma_{xy}}{\sigma_y^2}$$

son los coeficientes de regresión lineal *de Y sobre X y de X sobre Y, respectivamente.*

Observación 5.2. *Hay que tener muy presente que:*

i) *No toda nube de puntos se ajusta apropiadamente a un modelo de regresión lineal. Los modelos lineales suponen una explicación simplificada y ágil de la realidad, y cuentan con un vasto respaldo teórico desde las matemáticas y la estadística; pero existe todo un abanico de técnicas de regresión (cuyo estudio excede el alcance de este curso), que emplean modelos basados en cualquier clase de función matemática, y que pueden ser más adecuados para el análisis del problema particular considerado.*

ii) *Un conjunto de datos proporciona una prueba de linealidad solamente sobre aquellos valores de las variables marginales cubiertos por el conjunto de datos; para valores fuera de éstos, no hay evidencia de linealidad. Es, por tanto, inadecuado utilizar una recta de regresión estimada para predecir los*

valores de una de las variables marginales correspondientes a valores de la otra variable que están fuera del rango cubierto por los datos.

Ejemplo 5.3. *Hallar las rectas de regresión correspondientes a las variables del Ejemplo 2.4.*

RESOLUCIÓN. Nos apoyaremos en los resultados del Ejemplo 3.2. En primer lugar, obtenemos la covarianza y los coeficientes de regresión lineal:

$$\begin{aligned}\sigma_{xy} =& \frac{1}{30}\{22.5\cdot(9\cdot 1)\\ &+27.5\cdot[(9\cdot 2)+(10\cdot 1)+(11\cdot 1)]\\ &+32.5\cdot[(9\cdot 1)+(10\cdot 3)+(11\cdot 4)+(12\cdot 2)]\\ &+37.5\cdot[(10\cdot 2)+(11\cdot 3)+(12\cdot 5)]\\ &+42.5\cdot[(11\cdot 1)+(12\cdot 4)]\}\\ &-(34.833\cdot 10.900)\\ \simeq& 3.570,\end{aligned}$$

$$\beta_{y_x} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{3.570}{26.245} \simeq 0.136,$$

$$\beta_{x_y} = \frac{\sigma_{xy}}{\sigma_y^2} = \frac{3.570}{1.090} \simeq 3.275.$$

Por tanto, las rectas de regresión son:

$$r_{y_x}:\ y = 10.900 + 0.136\,(x - 34.833), \qquad r_{x_y}:\ x = 34.833 + 3.275\,(y - 10.900).$$

□

5.2. Correlación

La correlación estudia el tipo de dependencia que existe entre las variables marginales de una variable bidimensional (X,Y), intentando cuantificarla mediante los llamados *coeficientes de correlación*.

Definición 5.4. *El* coeficiente de correlación lineal *o* de Pearson *es la media geométrica de los coeficientes de regresión lineal:*

$$\rho = \sqrt{\beta_{y_x}\cdot\beta_{x_y}} = \sqrt{\frac{\sigma_{xy}}{\sigma_x^2}\cdot\frac{\sigma_{xy}}{\sigma_y^2}} = \frac{\sigma_{xy}}{\sigma_x\cdot\sigma_y}.$$

Nótese que el signo de este coeficiente es el mismo que el de los coeficientes de regresión lineal, y se corresponde con el signo de la covarianza. Se demuestra que $|\rho| \leq 1$.

El coeficiente de correlación lineal proporciona la siguiente información sobre las rectas de regresión y el grado de dependencia entre ambas variables.

i) Si $\rho = 0$, entonces $\sigma_{xy} = 0$. Por tanto, las rectas de regresión son $y = \bar{y}$ y $x = \bar{x}$, perpendiculares ente sí. Las variables X e Y se dicen *linealmente incorreladas*, esto es, no están vinculadas por una dependencia lineal.

ii) Si $\rho = 1$, se comprueba sin dificultad que las dos rectas de regresión coinciden y tienen pendiente positiva, de modo que una de ellas crece si, y sólo si, crece la otra. En este caso decimos que X e Y presentan una *correlación positiva perfecta*.

iii) Si $\rho = -1$, entonces ambas rectas coinciden y tienen pendiente negativa, de modo que una de las variables crece si, y sólo si, la otra decrece. Se dice que X e Y presentan *correlación negativa perfecta*.

iv) Si $0 < |\rho| < 1$, las variables están tanto más correladas cuanto más próximo sea $|\rho|$ a 1. De forma orientativa, podemos adoptar la siguiente escala:

- $0 < |\rho| < 0.2$: correlación mala.
- $0.2 \leq |\rho| < 0.5$: correlación regular.
- $0.5 \leq |\rho| < 0.8$: correlación buena.
- $0.8 \leq |\rho| < 1$: correlación muy buena.

Ejemplo 5.5. *Calcular y discutir el coeficiente de correlación lineal para las variables del Ejemplo 2.4.*

RESOLUCIÓN. El coeficiente de correlación lineal es

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{3.570}{5.123 \cdot 1.044} \simeq 0.67,$$

lo que indica una correlación positiva buena ($0.5 < \rho \simeq 0.67 < 0.8$) entre ambas variables. □