Descripción conjunta de varias variables

Introducción

En lo estudiado anteriormente se ha visto cómo una serie de datos procedentes de una variable, X, se puede representar gráficamente de modo que resulte más intuitivo hacerse una idea de cómo se distribuyen las observaciones.

Otros conceptos que también ayudan en el análisis son los estadísticos de tendencia central, que nos indican hacia dónde tienden a agruparse los datos, y los estadísticos de dispersión, que nos indican si las diferentes modalidades que presenta una variable están muy agrupadas alrededor de cierto valor central o, por el contrario, las variaciones que presentan las modalidades con respecto al valor central son grandes. También se puede determinar si los datos se distribuyen de forma simétrica a un lado y a otro de un valor central.

En este tema se trata de estudiar una situación muy usual y por tanto de gran interés en la práctica:

Si Y es otra variable definida sobre la misma población que X, ¿será posible determinar si existe alguna relación entre las modalidades de X y de Y?

Un ejemplo trivial consiste en considerar una población formada por alumnas/os de un instituto y definir sobre ella las variables

$$\begin{aligned} X &\equiv \text{Altura medida en cm} \\ Y &\equiv \text{Altura medida en metros} \end{aligned}$$

ya que la relación es determinista y clara: $Y = X/100$.

Obsérvese que aunque la variable Y, como tal, puede tener cierta dispersión, vista *como función* de X, su dispersión es nula.

Un ejemplo que nos interesa realmente más lo tenemos cuando sobre la misma población

definimos dos variables distintas, por ejemplo, peso y altura de las personas. Intuitivamente esperamos que exista cierta relación entre ambas variables, por ejemplo, $Y = X - 110 \pm dispersión$ que nos expresa que (en media) a mayor altura se espera mayor peso. La relación no es exacta y por ello será necesario introducir algún término que exprese la dispersión de Y con respecto a la variable X.

Es fundamental de cara a realizar un trabajo de investigación experimental, conocer muy bien las técnicas de estudio de variables bidimensionales (y $n-$dimensionales en general). Baste para ello pensar que normalmente las relaciones entre las variables no son tan evidentes como se mencionó arriba. Por ejemplo: ¿se puede decir que en un grupo de personas existe alguna relación entre $X =$ tensión arterial e $Y =$ edad?

Tablas de doble entrada

Consideramos una población de n individuos, donde cada uno de ellos presenta dos caracteres que representamos mediante las variables X e Y. Representamos mediante

$$X \hookrightarrow x_1, x_2, \dots x_k$$

las k modalidades que presenta la variable X, y mediante

$$Y \hookrightarrow y_1, y_2, \dots y_p$$

las p modalidades de Y.

Con la intención de reunir en una sola estructura toda la información disponible, creamos una tabla formada por $k-p$ casillas, organizadas de forma que se tengan k filas y p columnas. La casilla denotada de forma general mediante el subíndice ij hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades x_i e y_j.

$X \setminus Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_p	
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\cdot}$
$\dots\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$n_{\cdot\cdot}$

De este modo, para $i = 1, \dots, k$, y para $j = 1, \dots, p$, se tiene que n_{ij} es el número de individuos o **frecuencia absoluta**, que presentan a la vez las modalidades x_i e y_j.

El número de individuos que presentan la modalidad x_i, es lo que llamamos **frecuencia absoluta marginal** de x_i y se representa como $n_{i\cdot}$. Es evidente la siguiente igualdad

$$n_{i\cdot} = n_{i1} + n_{i2} + \ldots + n_{ip} = \sum_{j=1}^{p} n_{ij}$$

Obsérvese que hemos escrito un símbolo $\bullet$ *en la parte de las jotas* que simboliza que estamos considerando los elementos que presentan la modalidad x_i, independientemente de las modalidades que presente la variable Y. De forma análoga, se define la frecuencia absoluta marginal de la modalidad y_j como

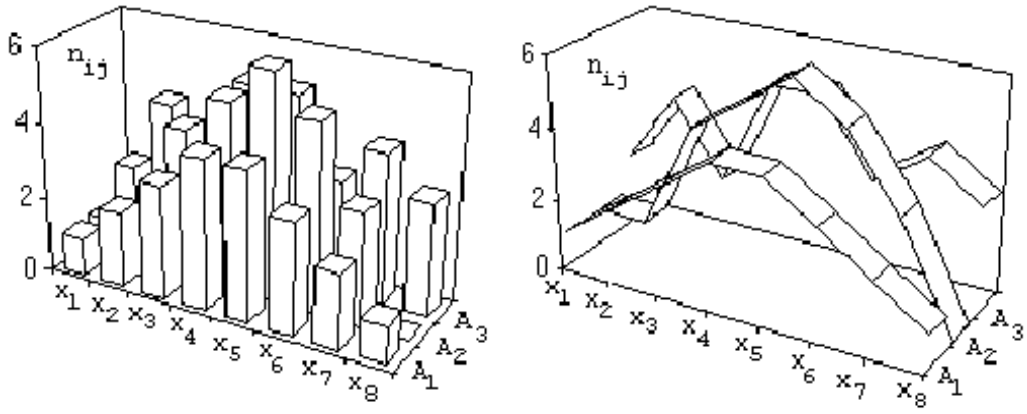
$$n_{\cdot j} = n_{1j} + n_{2j} + \ldots + n_{kj} = \sum_{i=1}^{k} n_{ij}$$

Estas dos distribuciones de frecuencias, $n_{i\cdot}$ para $i = 1, \ldots, k$, y $n_{\cdot j}$ para $j = 1, \ldots, p$ reciben el nombre de **distribuciones marginales** de X e Y respectivamente.

El número total de elementos de la población (o de la muestra), n lo obtenemos de cualquiera de las siguientes formas, que son equivalentes:

$$n = n_{\cdot\cdot} = \sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{p} n_{\cdot j} = \sum_{i=1}^{k} \sum_{j=1}^{p} n_{ij}$$

Las distribuciones de frecuencias de las variables bidimensionales también pueden ser representadas gráficamente. Al igual que en el caso unidimensional existen diferentes tipos de representaciones gráficas, aunque éstas resultan a ser más complicadas:



Distribuciones marginales

A la proporción de elementos (tanto por uno) que presentan simultáneamente las modalidades x_i e y_j la llamamos frecuencia relativa f_{ij}

$$f_{ij} = \frac{n_{ij}}{n}$$

siendo las frecuencias relativas marginales las siguientes cantidades:

$$f_{i\cdot} = \sum_{j=1}^{p} f_{ij} = \frac{n_{i\cdot}}{n}$$

$$f_{\cdot j} = \sum_{i=1}^{k} f_{ij} = \frac{n_{\cdot j}}{n}$$

Es obvio que

$$f_{\cdot\cdot} = \sum_{i=1}^{k} f_{i\cdot} = \sum_{j=1}^{p} f_{\cdot j} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij} = 1.$$

Observación

Es importante observar que las tablas bidimensionales aportan más información que las vistas anteriormente. De hecho, si quisiésemos estudiar la variable X y la Y por separado, hubiese bastado con utilizar:

Mod.X	**Marg. Abs.**	**Marg. Rel.**
x_1	$n_{1\cdot}$	$f_{1\cdot} = \frac{n_{1\cdot}}{n}$
...	...	...
x_i	$n_{i\cdot}$	$f_{i\cdot} = \frac{n_{i\cdot}}{n}$
...	...	...
x_k	$n_{k\cdot}$	$f_{k\cdot} = \frac{n_{k\cdot}}{n}$
	n	1

Mod.Y	**Marg. Abs.**	**Marg. Rel.**
y_1	$n_{\cdot 1}$	$f_{\cdot 1} = \frac{n_{\cdot 1}}{n}$
...	...	...
y_j	$n_{\cdot j}$	$f_{\cdot j} = \frac{n_{\cdot j}}{n}$
...	...	...
y_p	$n_{\cdot p}$	$f_{\cdot p} = \frac{n_{\cdot p}}{n}$
	n	1

Toda esa información se puede resumir en una sola tabla del siguiente modo:

$X \setminus Y$	y_1	y_2	$\dots$	y_j	$\dots$	y_p	
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2p}	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	n_{i1}	n_{i2}	$\dots$	n_{ij}	$\dots$	n_{ip}	$n_{i\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_k	n_{k1}	n_{k2}	$\dots$	n_{kj}	$\dots$	n_{kp}	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$n_{\cdot\cdot}$

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

Distribuciones condicionadas

De todos los elementos de la población, n, podemos estar interesados, en un momento dado, en un conjunto más pequeño que está formado por aquellos elementos que han presentado la modalidad y_j, para algún $j = 1, \dots, p$. El número de elementos de este conjunto sabemos que es $n_{\cdot j}$ La variable X definida sobre este conjunto se denomina **variable condicionada** y se suele denotar mediante $X|y_j$ o bien $X|Y = y_j$.

La distribución de frecuencias absolutas de esta nueva variable es exactamente la columna j de la tabla. Por tanto sus frecuencias relativas, que denominaremos frecuencias relativas condicionadas son

$$f_i^j = \frac{n_{ij}}{n_{\cdot j}}$$

$\forall i = 1, \dots, k$.

De la misma forma, es posible dividir la población inicial en k subconjuntos, cada uno de ellos caracterizados por la propiedad de que el $i-$ésimo conjunto todos los elementos verifican la propiedad de presentar la modalidad x_i. Sobre cada uno de estos conjuntos tenemos la variable condicionada $Y|x_i$ cuya distribución de frecuencias relativas condicionadas es:

$$f_j^i = \frac{n_{ij}}{n_{i\cdot}}$$

$\forall j = 1, \ldots, p$.

De este modo, la distribución de cada una de las variables condicionadas se puede representar en tablas como sigue:

Mod. $X\|y_j$	**Fr. Abs.**	**Fr. Rel.**
x_1	n_{1j}	f_1^j
$\ldots$	$\ldots$	$\ldots$
x_i	n_{ij}	$f_i^j = \frac{n_{ij}}{n_{\cdot j}}$
$\ldots$	$\ldots$	$\ldots$
x_k	n_{kj}	f_k^j
		1

Mod. $Y\|x_i$	**Fr. Abs.**	**Fr. Rel.**
y_1	n_{i1}	f_1^i
$\ldots$	$\ldots$	$\ldots$
y_j	n_{ij}	$f_j^i = \frac{n_{ij}}{n_{i\cdot}}$
$\ldots$	$\ldots$	$\ldots$
y_p	n_{ip}	f_p^i
		1

Observación

Estas relaciones se volverán a encontrar en términos de probabilidades teóricas, como definición de probabilidad condicionada.

Dependencia funcional e independencia

Dependencia funcional

La dependencia funcional, que nos refleja cualquier fórmula matemática o física, es a la que estamos normalmente más habituados. Al principio del tema consideramos un ejemplo en el que sobre una población de alumnos definíamos las variables

$$\begin{aligned} X &\equiv \text{Altura medida en cm} \\ Y &\equiv \text{Altura medida en metros,} \end{aligned}$$

al tomar uno de los alumnos, hasta que no se realice una medida sobre el mismo, no tendremos claro cual será su altura. Podemos tener cierta intuición sobre qué valor es

más probable que tome (alrededor de la media, con cierta dispersión). Sin embargo, si la medida X ha sido realizada, no es necesario practicar la de Y, pues la relación entre ambas es exacta (dependencia funcional): $Y = X/100$.

Ello puede describirse como que conocido el valor $X = x_i$, la distribución de $Y|x_i$ sólo toma un valor con una frecuencia del 100 %.

Esto se traduce en una tabla bidimensional de X e Y, del siguiente modo: La variable Y depende funcionalmente de la variable X si para cada fila $X = x_i$, existe un único j tal que $n_{ij} \neq 0$.

Análogamente, tenemos dependencia funcional de X con respecto a Y haciendo el razonamiento simétrico, pero por columnas, es decir, X depende funcionalmente de la variable Y si para cada columna $Y = y_j$, existe un único i tal que $n_{ij} \neq 0$.

Claramente, si la dependencia funcional es recíproca, la tabla es necesariamente cuadrada ($k = p$).

Ejemplo

Consideramos una población formada por 12 individuos, donde hay 3 franceses, 7 argentinos y 3 guineanos. Definimos las variables:

$$\begin{aligned} X &\equiv \text{Continente de nacimiento: Europa, América, Africa} \\ Y &\equiv \text{Nacionalidad: Francés, Guineano, Argentino} \\ Z &\equiv \text{Hablar castellano: Si, No.} \end{aligned}$$

Entonces, sobre esta población, podemos construir las siguientes tablas:

X \ Z	**Si**	**No**	
Europa	0	3	3
América	7	0	7
África	2	0	2
	9	3	12

X \ Y	**Francés**	**Guineano**	**Argentino**	
Europa	3	0	0	3
América	0	0	7	7
África	0	2	0	2
	3	2	7	12

y nos damos cuenta de que, según la definición

Z depende funcionalmente de X.

X no depende funcionalmente de Z.

X e Y dependen funcionalmente la una de la otra de modo recíproco.

Independencia

Hemos visto que la dependencia funcional implica una estructura muy particular de la tabla bidimensional, en la que en todas las filas (o en todas las columnas) existe un único elemento no nulo. Existe un concepto que de algún modo es el opuesto a la dependencia funcional, que es el de **independencia.**

Se puede expresar de muchas maneras el concepto de independencia, y va a implicar de nuevo una estructura muy particular de la tabla bidimensional, en el que todas las filas y todas las columnas van a ser proporcionales entre sí.

Para enunciar lo que es la independencia de dos variables vamos a basarnos en el siguiente razonamiento: Si la variable Y es independiente de X, lo lógico es que la distribución de frecuencias relativas condicionadas $Y|x_1$ sea la misma que la de $Y|x_2, \ldots Y|x_{k.}$

Esto se puede escribir diciendo que

$$f_j^1 = \cdots = f_j^i = \cdots = f_j^k = f_{\cdot j}$$

$\forall j = 1, \ldots, p.$

Pues bien, diremos que la variable Y es independiente de X si se verifica la relación anterior. Hay otras formas equivalentes de enunciar la independencia: Cada una de las siguientes relaciones expresa por sí sola la condición de independencia.

Proposición (Independencia en tablas de doble entrada)

Cada una de las siguientes relaciones expresa por sí sola la condición de independencia entre las variables X e Y

$$\begin{aligned} \frac{n_{ij}}{n_{i\cdot}} &= \frac{n_{\cdot j}}{n_{\cdot\cdot}} \\ f_{ij} &= f_{i\cdot} \cdot f_{\cdot j} \\ n_{ij} &= \frac{n_{i\cdot} \cdot n_{\cdot j}}{n_{\cdot\cdot}} \end{aligned}$$

Se puede observar que las relaciones anteriores implican que la independencia es siempre recíproca, es decir, si X es independiente de Y, entonces Y es independiente de X.

Ejemplo

Si tenemos dos variables que son

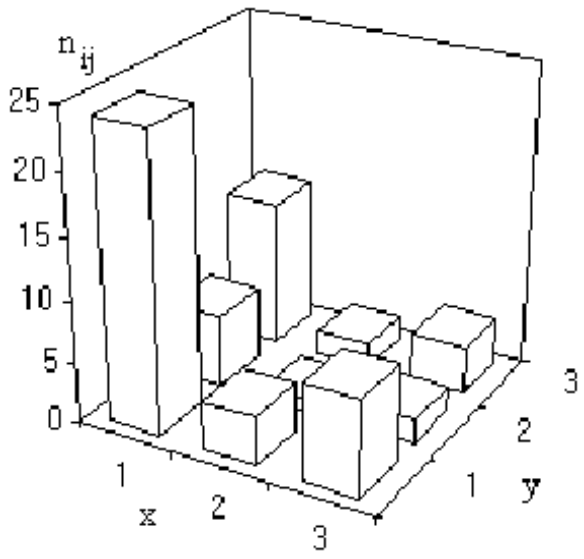
$$X \equiv \text{Número de cladifurdios ciclotómicos}$$
$$Y \equiv \text{coeficiente de saturación de ciclopondrinas}$$

y están distribuidas en una tabla del modo:

X \ Y	**1**	**3**	**5**	
0	24	4	8	36
1	6	1	2	9
2	12	2	4	18
	42	7	14	63

se puede decir que ambas variables son independientes. Obsérvese la proporcionalidad existente entre todas las filas de la tabla (incluidas la marginal). Lo mismo ocurre entre las columnas.

Cuando las variables son independientes, las diferencias entre las filas (o columnas) pueden entenderse como cambios de escala, como se observa en la siguiente gráfica.



Medias y varianzas marginales y condicionadas

Asociados a las distribuciones marginales y condicionadas, definidas en las secciones anteriores, podemos definir algunos estadísticos de tendencia central o dispersión, generalizando los que vimos en los temas dedicados al análisis de una variable. Las medias marginales de la variable X e Y se definen del siguiente modo:

$$\bar{x} = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} x_i = \sum_{i=1}^{k} f_{i.} x_i$$

$$\bar{y} = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} y_j = \sum_{j=1}^{p} f_{.j} y_j$$

Las varianzas marginales respectivas son

$$s_X^2 = \frac{1}{n_{..}} \sum_{i=1}^{k} n_{i.} \left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{k} f_{i.} \left(x_i - \bar{x}\right)^2$$

$$s_Y^2 = \frac{1}{n_{..}} \sum_{j=1}^{p} n_{.j} \left(y_j - \bar{y}\right)^2 = \sum_{j=1}^{p} f_{.j} \left(y_j - \bar{y}\right)^2$$

Para cada una de las p variables condicionadas $X|y_j$ definimos la media condicionada y la varianza condicionada mediante:

$$\begin{aligned} \bar{x}_j &= \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} f_i^j x_i \\ s_{X_j}^2 &= \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} \left(x_i - \bar{x}_j\right)^2 = \sum_{i=1}^{k} f_i^j \left(x_i - \bar{x}_j\right)^2 = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i^2 - \bar{x}_j^2 \end{aligned}$$

y lo mismo hacemos para las k condicionadas $Y|x_i$

$$\begin{aligned} \bar{y}_i &= \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j = \sum_{j=1}^{p} f_j^i y_j \\ s_{Y_i}^2 &= \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} \left(y_j - \bar{y}_i\right)^2 = \sum_{j=1}^{p} f_j^i \left(y_j - \bar{y}_i\right)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{p} n_{ij} y_j^2 - \bar{y}_i^2 \end{aligned}$$

Es interesante observar que podemos considerar que las $n_{..}$ observaciones de la variable X han sido agrupadas en p subgrupos, cada uno de ellos caracterizado por la propiedad de que $Y = y_j$ para algún $j = 1, \ldots, p$. Así se puede afirmar que las medias de las marginales es la media ponderada de las condicionadas, y la varianza de las marginales es la media ponderada de las varianzas condicionadas más la varianza ponderada de las medias condicionadas. De modo más preciso:

Proposición

Las medias y varianzas marginales de las variables X e Y se pueden escribir de modo equivalente como:

$$\begin{aligned}\overline{\overline{x}} &= \sum_{j=1}^{p} f_{\cdot j}\bar{x}_j = \frac{1}{n_{\cdot\cdot}}\sum_{j=1}^{p} n_{\cdot j}\bar{x}_j \\ s_X^2 &= \sum_{j=1}^{p} f_{\cdot j}s_{X_j}^2 + \sum_{j=1}^{p} f_{\cdot j}\left(\bar{x}_j - \overline{\overline{x}}\right)^2\end{aligned}$$

$$\begin{aligned}\overline{\overline{y}} &= \sum_{i=1}^{k} f_{i\cdot}\bar{y}_j = \frac{1}{n_{\cdot\cdot}}\sum_{i=1}^{k} n_{i\cdot}\bar{y}_j \\ s_Y^2 &= \sum_{i=1}^{k} f_{i\cdot}s_{Y_i}^2 + \sum_{i=1}^{k} f_{i\cdot}\left(\bar{y}_j - \overline{\overline{y}}\right)^2\end{aligned}$$

Covarianza y coeficiente de correlación

Cuando analizábamos las variables unidimensionales considerábamos, entre otras medidas importantes, la media y la varianza. Ahora, hemos visto que estas medidas también podemos considerarlas de forma individual para cada una de las componentes de la variable bidimensional.

Si observamos con atención los términos

$$\begin{aligned}s_X^2 &= \sum_{i=1}^{k} f_{i\cdot}\left(x_i - \overline{\overline{x}}\right)\left(x_i - \overline{\overline{x}}\right) \\ s_Y^2 &= \sum_{j=1}^{p} f_{\cdot j}\left(y_j - \overline{\overline{y}}\right)\left(y_j - \overline{\overline{y}}\right)\end{aligned}$$

vemos que las cantidades $\left(x_i - \overline{\overline{x}}\right)$ e $\left(y_j - \overline{\overline{y}}\right)$ van al cuadrado y por tanto no pueden ser negativas.

La covarianza S_{XY} es una manera de generalizar la varianza y se define como:

$$S_{XY} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}\left(x_i - \overline{\overline{x}}\right)\left(y_j - \overline{\overline{y}}\right).$$

Como se ve, la fórmula es muy parecida a las de las varianzas. Es sencillo comprobar que se verifica la siguiente expresión de S_{XY} más útil en la práctica:

$$S_{XY} = \sum_{i=1}^{k}\sum_{j=1}^{p} f_{ij}x_iy_j - \overline{\overline{x}}\cdot\overline{\overline{y}},$$

o si las observaciones no están ordenadas en una tabla de doble entrada, entonces se tiene que

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} (x_i - \bar{x})(y_j - \bar{y}).$$

o, lo que es lo mismo,

$$S_{XY} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{p} x_i y_j - \bar{x} \cdot \bar{y},$$

Ejemplo

Se han clasificado 100 familias según el número de hijos o de hijas, en la tabla siguiente:

H \ M	**0**	**1**	**2**	**3**	**4**
0	4	6	9	4	1
1	5	10	7	4	2
2	7	8	5	3	1
3	5	5	3	2	1
4	2	3	2	1	0

1. Hallar las medias, varianzas y desviaciones típicas marginales.
2. ¿Qué número medio de hijas hay en aquellas familias que tienen 2 hijos?
3. ¿Qué número medio de hijos hay en aquellas familias que no tienen hijas?
4. ¿Qué número medio de hijos tienen aquellas familias que a lo sumo tienen 2 hijas?
5. Hallar la covarianza.

Solución: En primer lugar, definimos las variables $X =$ número de hijos, e $Y =$ número de hijas y construimos la tabla con las frecuencias marginales y con otras cantidades que son útiles en el cálculo de medias y varianzas:

Y	$y_1 = 0$	$y_2 = 1$	$y_3 = 2$	$y_4 = 3$	$y_5 = 4$				
X	0	1	2	3	4	$n_{i\cdot}$	$n_{i\cdot}x_i$	$n_{i\cdot}x_i^2$	$x_i \sum_{j=0}^{4} n_{ij}y_j$
$x_1 = 0$	4	6	9	4	1	**24**	0	0	0
$x_2 = 1$	5	10	7	4	2	**28**	28	28	44
$x_3 = 2$	7	8	5	3	1	**24**	48	96	62
$x_4 = 3$	5	5	3	2	1	**16**	48	144	63
$x_5 = 4$	2	3	2	1	0	**8**	32	128	40
$n_{\cdot j}$	**23**	**32**	**26**	**14**	**5**	**100**	**156**	**396**	**209**
$n_{\cdot j}y_j$	0	32	52	42	20	**146**			
$n_{\cdot j}y_j^2$	0	32	104	126	80	**342**			

de este modo, las medias marginales son

$$\begin{aligned}\overline{\overline{x}} &= \frac{1}{n_{..}}\sum_{i=1}^{5} n_{i.}x_i = \frac{156}{100} = 1{,}56\\ \overline{\overline{y}} &= \frac{1}{n_{..}}\sum_{j=1}^{5} n_{.j}y_j = \frac{146}{100} = 1{,}46\end{aligned}$$

Calculamos, después, las varianzas marginales

$$\begin{aligned}s_X^2 &= \frac{1}{n_{..}}\sum_{i=1}^{k} n_{i.}x_i^2 - \overline{\overline{x}}^2 = \frac{396}{100} - 1{,}56^2 = 1{,}53\\ s_Y^2 &= \frac{1}{n_{..}}\sum_{j=1}^{p} n_{.j}y_j^2 - \overline{\overline{y}}^2 = \frac{342}{100} - 1{,}46^2 = 1{,}29\end{aligned}$$

que nos dan directamente las desviaciones típicas marginales,

$$\begin{aligned}s_X &= \sqrt{1{,}53} = 1{,}24\\ s_Y &= \sqrt{1{,}29} = 1{,}14\end{aligned}$$

El número medio de hijas en las familias con 2 hijos se obtiene calculando la distribución condicionada de $Y|_{X=2} = Y|x_3$

$Y\|_{X=2}$	n_{3j}	$n_{3j}y_j$
$y_1 = 0$	7	0
$y_2 = 1$	8	8
$y_3 = 2$	5	10
$y_4 = 3$	3	9
$y_5 = 4$	1	4
	24	31

Así

$$\bar{Y}|_{X=2} = \bar{Y}|x_3 = \frac{1}{n_{3.}}\sum_{j=1}^{5} n_{3j}y_j = \frac{31}{24} = 1{,}29$$

Del mismo modo, el número medio de hijos varones de las familias sin hijas, se calcula con la distribución condicionada $X|_{Y=0} = X|y_1$

$X\|_{Y=0}$	n_{i1}	$n_{i1}x_i$
$x_1 = 0$	4	0
$x_2 = 1$	5	5
$x_3 = 2$	7	14
$x_4 = 3$	5	15
$x_5 = 4$	2	8
	23	42

$$\bar{X}|_{Y=0} = \bar{X}|y_1 = \frac{1}{n_{\cdot 1}} \sum_{i=1}^{5} n_{i1}x_i = \frac{42}{23} = 1{,}83$$

El número medio de hijos en las familias que a lo sumo tienen dos hijas, se calcula usando las marginales de la tabla obtenida a partir de las columnas y_1, y_2 e y_3

$X\|_{Y\leq 2}$	n_{i1}	n_{i2}	n_{i3}	$n_{i1} + n_{i2} + n_{i3}$	$(n_{i1} + n_{i2} + n_{i3})\, x_i$
$x_1 = 0$	4	6	9	19	19
$x_2 = 1$	5	10	7	22	22
$x_3 = 2$	7	8	5	20	40
$x_4 = 3$	5	5	3	13	39
$x_5 = 4$	2	3	2	7	28
				81	129

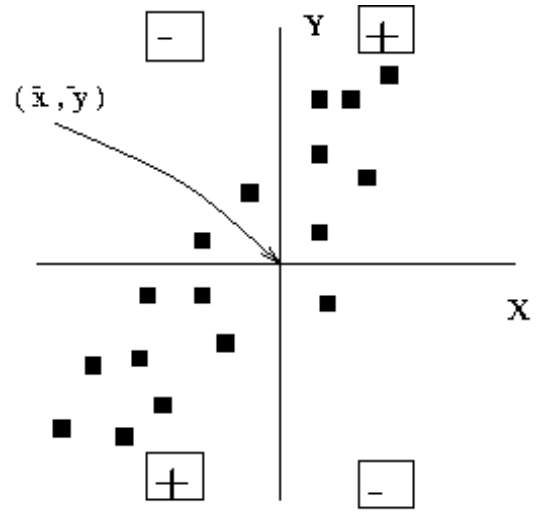
$$\bar{X}|_{Y\leq 2} = \frac{129}{81} = 1{,}59$$

La covarianza es:

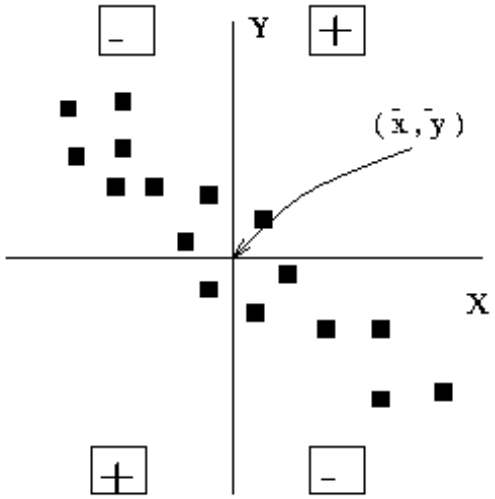
$$S_{XY} = \frac{1}{n_{\cdot\cdot}} \sum_{i=1}^{k} x_i \sum_{j=1}^{p} n_{ij}y_j - \overline{\overline{x}} \cdot \overline{\overline{y}} = \frac{209}{100} - 1{,}56 \cdot 1{,}46 = -0{,}188$$

Una interpretación geométrica de la covarianza

Consideremos la *nube de puntos* formadas por las n parejas de datos (x_i, y_i). El centro de gravedad de esta nube de puntos es $\left(\overline{\overline{x}}, \overline{\overline{y}}\right)$, o bien podemos escribir simplemente $(\overline{x}, \overline{y})$ si los datos no están ordenados en una tabla de doble entrada. Trasladamos los ejes XY al nuevo centro de coordenadas. Queda así dividida la nube de puntos en cuatro cuadrantes como se observa en la figura. Los puntos que se encuentran en el primer y tercer cuadrante contribuyen positivamente al valor de S_{XY}, y los que se encuentran en el segundo y el cuarto lo hacen negativamente.



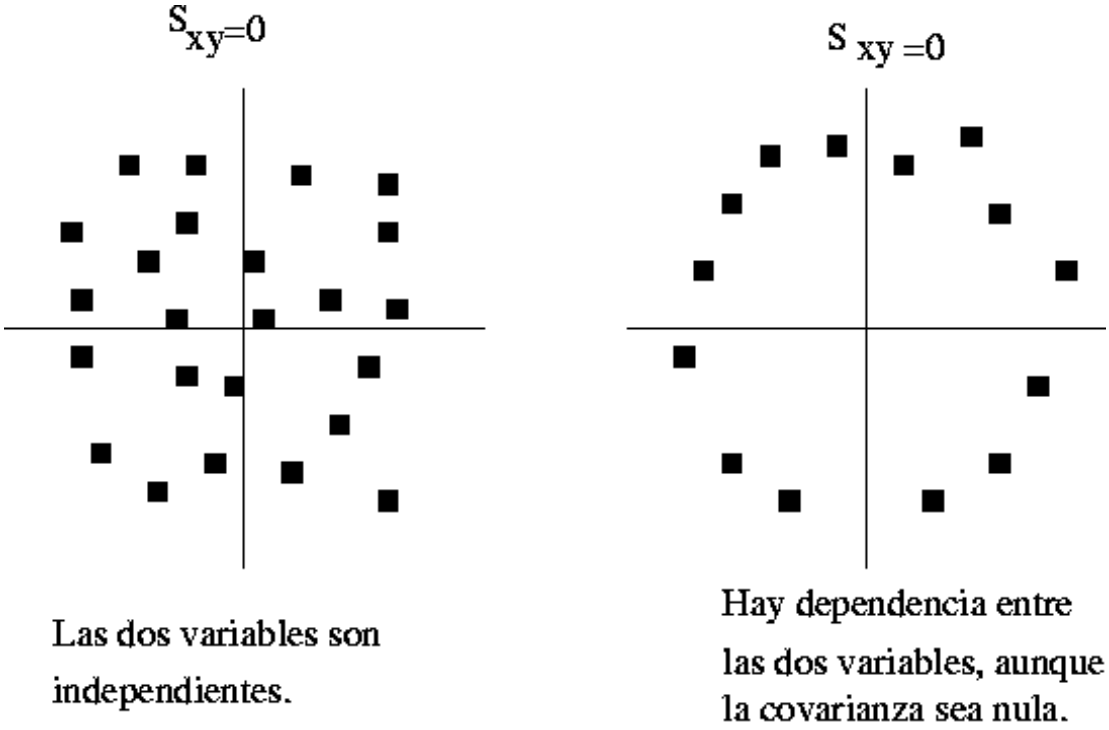
Cuando X crece, Y crece
Casi todos los puntos pertenecen
a los cuadrantes primero y tercero



Cuando X crece, Y decrece
Casi todos los puntos pertenecen
a los cuadrantes segundo y cuarto

De este modo,

- Si hay mayoría de puntos en el tercer y primer cuadrante, ocurrirá que $S_{XY} \geq 0$, lo que se puede interpretar como que la variable Y tiende a aumentar cuando lo hace X;
- Si la mayoría de puntos están repartidos entre el segundo y cuarto cuadrante entonces $S_{XY} \leq 0$, es decir, las observaciones Y tienen tendencia a disminuir cuando las de X aumentan;
- Si los puntos se reparten con igual intensidad alrededor de $\left(\overline{\overline{x}}, \overline{\overline{y}}\right)$, entonces se tendrá que S_{XY}=0.



Así, cuando los puntos se reparten de modo más o menos homogéneo entre los cuadrantes primero y tercero, y segundo y cuarto, se tiene que $S_{XY} \approx 0$. Eso no quiere decir de ningún modo que no pueda existir ninguna relación entre las dos variables, ya que ésta puede existir como se aprecia en la figura de la derecha.

La Covarianza

Si $S_{XY} > 0$ las dos variables crecen o decrecen a la vez (nube de puntos creciente).

Si $S_{XY} < 0$ cuando una variable crece, la otra tiene tendencia a decrecer (nube de puntos decreciente).

Si los puntos se reparten con igual intensidad alrededor de $(\overline{x}, \overline{y})$, $S_{XY} = 0$ (no hay relación lineal).

De este modo, podemos utilizar la covarianza para medir la variación conjunta (*covariación*) de las variables X e Y. Esta medida no debe ser utilizada de modo exclusivo para medir la relación entre las dos variables, ya que es sensible al cambio de unidad de medida, como se observa en el siguiente resultado:

Proposición

$$S_{X,a+bY} = bS_{XY}$$

Demostración: Para simplificar las notaciones, vamos a considerar que los datos no están agrupados en una tabla estadística, entonces,

$$S_{X,a+bY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\left[(a + by_i) - (a + b\bar{y})\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\, b\,(y_j - \bar{y}) = b\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_j - \bar{y}) = bS_{XY}$$

Así pues, es necesario definir una medida de la relación entre dos variables, que no esté afectada por los cambios de unidad de medida. Una forma posible de conseguir este objetivo es dividir la covarianza entre el producto de las desviaciones típicas de cada variable, ya que así se obtiene un coeficiente adimensional, r, que se denomina *coeficiente de correlación lineal de Pearson*

$$r = \frac{S_{XY}}{S_X S_Y}.$$

El coeficiente de correlación lineal posee las siguientes propiedades:

1. Carece de unidades de medida (adimensional).
2. Es invariante para transformaciones lineales (cambio de origen y escala) de las variables.
3. Sólo toma valores comprendidos entre -1 y 1: $-1 \leq r \leq 1$
4. Cuando $|r|$ esté próximo a uno, se tiene que existe una relación lineal muy fuerte entre las variables.
5. Cuando $r \approx 0$, puede afirmarse que no existe relación lineal entre ambas variables.

Interpretación geométrica de r

Si los datos son observaciones que no están ordenadas en una tabla bidimensional, tendremos parejas de valores para cada sujeto o elemento (x_i, y_i), para $i = 1, \ldots, n$. La fórmula de la covarianza, en este caso, es

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})\,.$$

Podemos a escribir las observaciones en forma de vectores de la siguiente manera:

$$\begin{aligned}\vec{X} &= (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n \\ \vec{Y} &= (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n \\ \overrightarrow{\bar{x}} &= (\bar{x}, \bar{x}, \ldots, \bar{x}) \in \mathbb{R}^n \\ \overrightarrow{\bar{y}} &= (\bar{y}, \bar{y}, \ldots, \bar{y}) \in \mathbb{R}^n\end{aligned}$$

Si denotamos $\vec{v} \cdot \vec{w}$ el producto escalar de los vectores $\vec{v}$ y $\vec{w}$, es inmediato comprobar que en realidad las definiciones de varianza y covarianza tienen una idea geométrica muy simple: son productos escalares en los que intervienen los vectores $\vec{X} - \overrightarrow{\bar{x}}$ e $\vec{Y} - \overrightarrow{\bar{y}}$,

$$\begin{aligned}S_{XY} &= \frac{1}{n}\left(\vec{X} - \overrightarrow{\bar{x}}\right)\left(\vec{Y} - \overrightarrow{\bar{y}}\right) \\ S_X^2 &= \frac{1}{n}\left(\vec{X} - \overrightarrow{\bar{x}}\right)\left(\vec{X} - \overrightarrow{\bar{x}}\right) = \frac{1}{n}\left|\vec{X} - \overrightarrow{\bar{x}}\right|^2 \\ S_Y^2 &= \frac{1}{n}\left(\vec{Y} - \overrightarrow{\bar{y}}\right)\left(\vec{Y} - \overrightarrow{\bar{y}}\right) = \frac{1}{n}\left|\vec{Y} - \overrightarrow{\bar{y}}\right|^2\end{aligned}$$

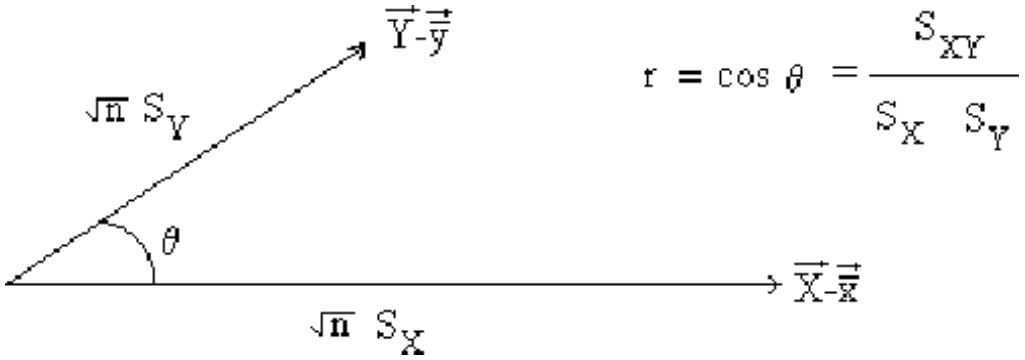
Con esta descripción geométrica de las varianzas y covarianzas, podemos poner de manifiesto la existencia de paralelismo entre las desviaciones de las variables X e Y, con respecto a sus centros de gravedad ya que

$$\left(\vec{X} - \overrightarrow{\bar{x}}\right) \cdot \left(\vec{Y} - \overrightarrow{\bar{y}}\right) = \left|\vec{X} - \overrightarrow{\bar{x}}\right| \cdot \left|\vec{Y} - \overrightarrow{\bar{y}}\right| \cdot \cos\theta,$$

donde θ es el ángulo entre los vectores $\vec{X} - \overrightarrow{\bar{x}}$ e $\vec{Y} - \overrightarrow{\bar{y}}$ como se puede ver en la figura.

Despejando:

$$\cos\theta = \frac{\left(\vec{X} - \overrightarrow{\bar{x}}\right) \cdot \left(\vec{Y} - \overrightarrow{\bar{y}}\right)}{\left|\vec{X} - \overrightarrow{\bar{x}}\right| \cdot \left|\vec{Y} - \overrightarrow{\bar{y}}\right|} = \frac{\frac{1}{n}\left(\vec{X} - \overrightarrow{\bar{x}}\right) \cdot \left(\vec{Y} - \overrightarrow{\bar{y}}\right)}{\sqrt{\frac{1}{n}}\left|\vec{X} - \overrightarrow{\bar{x}}\right| \cdot \sqrt{\frac{1}{n}}\left|\vec{Y} - \overrightarrow{\bar{y}}\right|} = \frac{S_{XY}}{S_X S_Y}$$

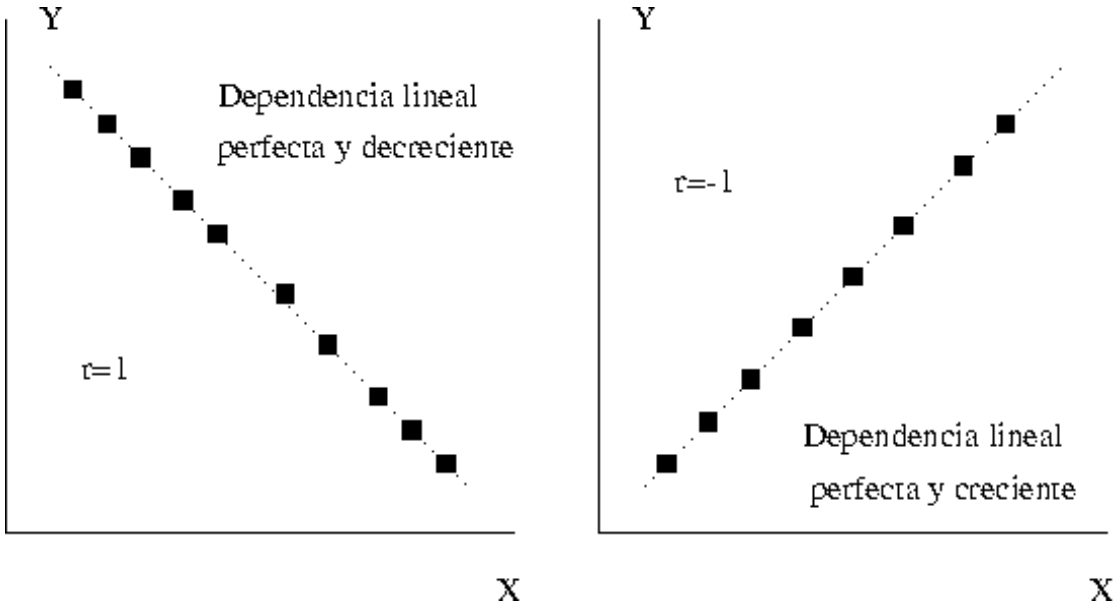


Interpretación geométrica de r como el coseno del ángulo que forman los vectores de las desviaciones con respecto a sus respectivas medias de X y de Y.

Si los vectores $\vec{X} - \overrightarrow{\overline{x}}$ e $\vec{Y} - \overrightarrow{\overline{y}}$ son totalmente paralelos entonces cos $= \pm 1$. En este caso existirá una constante de proporcionalidad m tal que:

$$\vec{Y} - \overrightarrow{\overline{y}} = m \cdot \left(\vec{X} - \overrightarrow{\overline{x}} \right).$$

Esta es la ecuación de una recta, es decir, $r = \pm 1$ si y sólo si las desviaciones con respecto a la media de ambas variables son proporcionales, es decir, si y sólo si las observaciones están perfectamente alineadas.



La magnitud que expresa el coseno del ángulo que forman los vectores $\vec{X} - \overrightarrow{\overline{x}}$ e $\vec{Y} - \overrightarrow{\overline{y}}$ tiene un papel muy destacado como veremos más adelante en regresión lineal. Lo hemos denominado anteriormente como coeficiente de correlación lineal de Pearson y se representa mediante la letra r:

$$r = \cos\theta = \frac{S_{XY}}{S_X S_Y}.$$

Son evidentes entonces las siguientes propiedades de r :

1. Cualesquiera que sean los valores (x_i, y_i), $i = 1, \ldots, n$, se tiene que $-1 \leq r \leq 1$, ya que r es el coseno del ángulo que forman las variaciones con respecto a sus valores medios.

2. Si las desviaciones con respecto al valor central de las observaciones x_i, son proporcionales a las desviaciones de y_i con respecto a su valor central $\bar{y}$,

$$\vec{Y} - \overrightarrow{\bar{y}} = m \cdot \left(\vec{X} - \overrightarrow{\bar{x}}\right) \Longleftrightarrow y_i = (\bar{y} - m\bar{x}) + m \cdot x_i \text{ donde } i = 1, \ldots, n$$

entonces los vectores $\vec{X} - \overrightarrow{\bar{x}}$ e $\vec{Y} - \overrightarrow{\bar{y}}$ son paralelos y por tanto $r = \pm 1$. En este caso se puede decir de modo exacto que conocido X lo es también Y, (y recíprocamente).

3. Por el contrario, si no existe dicha relación, el ángulo que formen $\vec{X} - \overrightarrow{\bar{x}}$ e $\vec{Y} - \overrightarrow{\bar{y}}$ será mayor, siendo el caso extremo en que ambos sean perpendiculares ($r = 0$). Cuando $r = 0$ decimos que las variables X e Y son incorreladas.

Otra propiedad interesante de r es la siguiente:

Proposición

El coeficiente de correlación entre dos variables no se ve afectada por los cambios de unidades.

Demostración:

Consideramos la variable bidimensional (X, Y) y sometemos a Y a un cambio de unidad $Z = a + bY$. Entonces

$$\frac{S_{XZ}}{S_X S_Z} = \frac{b \cdot S_{XY}}{b \cdot S_X S_Y} = \frac{S_{XY}}{S_X S_Y}$$

Por tanto ambas variables XZ y XY tienen el mismo coeficiente de correlación.