Estadística Descriptiva y Probabilidad

(Teoría y problemas)

3ª Edición

Autores
I. Espejo Miranda
F. Fernández Palacín
M. A. López Sánchez
M. Muñoz Márquez
A. M. Rodríguez Chía
A. Sánchez Navas
C. Valero Franco

UCA Universidad de Cádiz

Servicio de Publicaciones

Edita: Servicio de Publicaciones de la Universidad de Cádiz
C/ Dr. Marañón, 3
11002 Cádiz

http://www.uca.es/publicaciones

ISBN: 978-84-9828-058-6

Depósito legal:

Estadística Descriptiva y Probabilidad. Teoría y Problemas (Revisión: Febrero 2006)
I. Espejo Miranda, F. Fernández Palacín, M. A. López Sánchez, M. Muñoz Márquez, A. M. Rodríguez Chía, A. Sánchez Navas, C Valero Franco

Capítulo 2

Análisis conjunto de variables

En el capítulo anterior se ha considerado un único carácter, sin embargo, es frecuente estudiar conjuntamente varios caracteres y preguntarse si existe o no algún tipo de relación entre ellos. Este capítulo se dedica al estudio de la relación entre dos caracteres, comenzando con la organización y sintetización de la información, siguiendo un esquema análogo al establecido en el capítulo anterior, para concluir con el estudio de la relación entre ambos. Cuando se analiza la relación entre dos caracteres se pueden presentar dos casos extremos: el primero de ellos será aquel en que conocido el valor de un carácter se pueda obtener el valor del otro, el segundo se presenta cuando la información sobre un carácter no arroja ninguna información sobre el otro. Entre estas situaciones extremas se dan una infinidad de casos intermedios, por ello, el objetivo del capítulo será analizar el nivel de influencia existente entre los caracteres. Hay que indicar, no obstante, que dicho análisis no establecerá cuál es la causa y cuál el efecto entre ambos, sino sólo la intensidad de la relación.

1. Distribución conjunta de dos caracteres

Cuando el investigador está interesado en el estudio de dos caracteres de una población, se obtienen dos observaciones para cada individuo, que se recogen en forma de pares de valores, que deben ser organizados

X, Y	y_1	$\cdots$	y_j	$\cdots$	y_s	
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	n

Tabla 2.1: Distribuciones conjuntas y marginales de (X, Y)

en función de la naturaleza de dichos caracteres.

Al igual que en el caso unidimensional es interesante organizar los datos en forma de tabla de frecuencias, sin embargo, al tener que especificar los valores que toman ambos caracteres, la tabla debe ser de doble entrada o bidimensional, véase la tabla 2.1. Supongamos que X toma r valores distintos $x_1, x_2, \ldots, x_r$, e Y toma s valores distintos $y_1, y_2, \ldots, y_s$. Se define la frecuencia absoluta del par (x_i, y_j), que se denota por n_{ij}, como el número de veces que se observa dicho par de valores. Esta distribución se denomina *distribución conjunta* de (X, Y).

Se verifica que $\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = n$, valor que aparece recogido en la parte inferior derecha de la tabla. Conservando la notación, f_{ij}, con $f_{ij} = \frac{n_{ij}}{n}$, es la frecuencia relativa del par (x_i, y_j) y por lo tanto $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = 1$.

Si la distribución es de atributos, la *tabla* se llama *de contingencia* y si es de variables se denomina *de correlación*. Inicialmente, se centra el estudio en el caso en el que los caracteres sean variables, para abordar el estudio de tablas de contingencia en posteriores apartados de este capítulo.

La situación de los valores no nulos en la tabla de doble entrada da una idea intuitiva de la posible relación entre ambos caracteres, así,

el que las mayores frecuencias se den alrededor de una diagonal viene a indicar la existencia de relación, mientras que el que no se dé esta circunstancia va a suponer, generalmente, la ausencia de la misma.

2. Distribuciones marginales

En la tabla 2.1, se han sumado las frecuencias que aparecen en cada una de las filas y columnas, colocándose los resultados en los márgenes, donde:

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \ , \quad n_{i\cdot} = \sum_{j=1}^{s} n_{ij} \quad y$$

$$n = n_{\cdot\cdot} = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij},$$

de tal forma que la primera y última columna de la tabla 2.1, constituyen la *distribución marginal de* X, y la primera y última fila la *distribución marginal de* Y. Lógicamente se verifica que:

$$\sum_{i=1}^{r} f_{i\cdot} = 1 \ , \quad \sum_{j=1}^{s} f_{\cdot j} = 1 \quad y$$

$$\sum_{i=1}^{r} \sum_{j=1}^{s} f_{ij} = 1,$$

lo que garantiza la condición de ambas distribuciones.

Se interpreta $f_{i\cdot}$ como la proporción de datos que toman el valor x_i de X, independientemente del valor que tome Y. Una notación análoga se maneja para la variable Y.

Obsérvese que considerar la distribución marginal de una variable equivale a considerar la distribución de ésta independientemente de la otra.

3. Distribuciones condicionadas

Cuando se posee información previa de una de las variables en estudio, ésta puede modificar la información disponible de la otra. En

particular, cuando se considera la distribución de una variable para un valor fijo de la otra se obtiene la *distribución condicionada*. Más concretamente, las frecuencias condicionadas son:

$$f_{i_{|j}} = \frac{n_{ij}}{n_{\cdot j}} \quad \forall i = 1, 2, \cdots, r$$

$$f_{_{|i}j} = \frac{n_{ij}}{n_{i\cdot}} \quad \forall j = 1, 2, \cdots, s.$$

Que son, respectivamente, la condicionada de X para el valor y_j de Y, para $j = 1, 2, \cdots, s$, y la condicionada de Y para el valor x_i de X, para $i = 1, 2, \cdots, r$.

Ejemplo 2.1 *Un alumno de Estadística está interesado en estudiar la estatura y el peso del grupo de 21 alumnos varones que pertenecen a su clase. A tal efecto y una vez provisto de los adecuados aparatos de medida, metro y balanza, se dispone a realizar las mediciones. Los resultados que obtuvo se ofrecen en la tabla 2.2. La estatura y el peso se dan con una precisión de 0'01 metros y 1 kilogramo, respectivamente.*

Estatura	Peso	Estatura	Peso	Estatura	Peso
1'78	80	1'83	78	1'73	70
1'72	68	1'80	76	1'66	66
1'88	87	1'84	94	1'80	77
1'81	85	1'77	74	1'75	70
1'73	78	1'77	77	1'69	72
1'89	84	1'82	82	1'71	77
1'80	82	1'71	67	1'68	66

Tabla 2.2: Tabla de datos

El alumno domina el análisis descriptivo univariante y no tiene dificultad en aplicarlo a cada una de las variables que ha considerado, obteniendo así información sobre las medias, dispersiones, simetrías, etc., de la estatura y el peso. Sin embargo, conside-

ra la posibilidad de que entre las dos variables exista algún tipo de relación. A la vista de los datos, y pensando que en otra situación parecida el número de individuos en estudio fuera mucho más grande, decide agrupar a éstos en clases uniformes. Después de analizar la situación establece intervalos de amplitud 0'05 metros y 5 kilogramos, para estatura y peso respectivamente, reorganizando la información obtenida en una tabla de doble entrada. El resultado que obtuvo se recoge en la tabla 2.3.

	Estatura				
Peso	1'65-1'70	1'71-1'75	1'76-1'80	1'81-1'85	1'86-1'90
65-69	2	2			
70-74	1	2	1		
75-79		2	3	1	
80-84			2	1	1
85-89				1	1
90-94				1	

Tabla 2.3: Distribución conjunta

A la vista del resultado obtenido, el investigador observa que ha perdido precisión respecto a los datos originales. Efectivamente, utilizando la tabla de doble entrada lo único que se puede decir, por ejemplo, es que hay dos individuos que midiendo entre 1'71 y 1'75 metros pesan entre 75 y 79 kilogramos, ignorándose las mediciones exactas de éstos. No obstante, entiende que aunque el volumen de datos fuera muy grande la tabla de doble entrada seguiría siendo válida con la adición, tal vez, de algunas clases extremas y, además, piensa que el error que se cometería no sería muy grande si, llegado el caso, se viera en la necesidad de asignar a cada intervalo su marca de clase. Por otra parte, y en el haber de la abstracción realizada, un me-

ro análisis visual hace entender que entre las dos variables existe cierta relación, pues los valores no nulos de la tabla se distribuyen alrededor de una diagonal, obteniéndose un resultado que como era de esperar, y a falta de algún tipo de cuantificación que se realice más adelante, hace corresponder, en general, a los individuos de estatura baja los de poco peso y a los de estatura alta los de mayor peso. No se puede decir que conocida la estatura de un individuo quede determinado su peso pero sí que se puede acotar éste e incluso hacer una previsión aproximada de su valor. Tampoco se puede decir cuál de las dos variables determina los valores de la otra.

A continuación, el alumno piensa que quizás sería interesante ofrecer los valores de la tabla como proporciones del número total de observaciones, para ello divide cada elemento de la tabla por el número de individuos en estudio, obteniendo la tabla 2.4.

	Estatura				
Peso	1'65-1'70	1'71-1'75	1'76-1'80	1'81-1'85	1'86-1'90
65-69	2/21	2/21	0	0	0
70-74	1/21	2/21	1/21	0	0
75-79	0	2/21	3/21	1/21	0
80-84	0	0	2/21	1/21	1/21
85-89	0	0	0	1/21	1/21
90-94	0	0	0	1/21	0

Tabla 2.4: Distribución relativa conjunta

Ahora su interés se centra en conocer la proporción de sus compañeros que pertenecen a una de las clases de estatura independientemente del peso que tengan. Para ello, se da cuenta de que sólo tiene que sumar cada columna, obteniendo, por ejemplo, que

	Estatura					
Peso	1'65-1'70	1'71-1'75	1'76-1'80	1'81-1'85	1'86-1'90	$f(P)$
65-69	2/21	2/21	0	0	0	4/21
70-74	1/21	2/21	1/21	0	0	4/21
75-79	0	2/21	3/21	1/21	0	6/21
80-84	0	0	2/21	1/21	1/21	4/21
85-89	0	0	0	1/21	1/21	2/21
90-94	0	0	0	1/21	0	1/21
$f(E)$	3/21	6/21	6/21	4/21	2/21	1

Tabla 2.5: Distribuciones marginales

hay tres individuos cuya estatura está comprendida entre 1'65 y 1'70 metros, seis entre 1'71 y 1'75 metros, y así sucesivamente. Realizando la misma operación con las filas obtiene los resultados para el peso. Al objeto de organizar esta información decide añadir una fila y una columna en la tabla donde almacena los resultados, véase la tabla 2.5.

Siguiendo con los datos del ejemplo, nuestro investigador se pregunta por la proporción de compañeros que poseen una cierta estatura dentro de los del grupo que pesan entre 75 y 79 Kilogramos, que sabemos constituyen $\frac{6}{21}$ *del total.*

Le resulta fácil comprobar que de entre los que tienen ese peso hay $\frac{2}{6}$ *que tienen una estatura entre 1'71 y 1'75 metros. Observe que podría haber llegado al mismo resultado de haber dividido* $\frac{2}{21}$ *entre* $\frac{6}{21}$*, es decir, la proporción de individuos con altura en la clase* $[1'71, 1'75]$ *y peso en* $[75, 79]$ *entre el correspondiente a la proporción marginal de la variable peso en la clase* $[75, 79]$.

4. Independencia

La independencia-dependencia viene a medir la información que arroja sobre una de las variables el conocimiento que se tiene de la otra variable. Así, una información total implica *dependencia funcional*, la nula información *independencia*, y una información parcial *dependencia estadística*.

Formalmente, se dice que X es independiente de Y si se verifica que:

$$f_{i_{|j}} = f_{i\cdot} \quad \forall i = 1, \cdots, r \quad j = 1, 2, \cdots, s.$$

Es decir, si la frecuencia condicionada coincide con la marginal. De la misma forma se define la independencia de Y respecto de X.

La definición de distribución condicionada da una expresión alternativa para la independencia, y así X e Y son independientes si:

$$f_{ij} = f_{i\cdot} f_{\cdot j} \quad \forall i, j,$$

que además pone de manifiesto que la independencia se establece en un doble sentido; es decir, X es independiente de Y si y sólo si Y lo es de X.

Ejemplo 2.2 *En el ejemplo que se arrastra, nuestro joven estadístico se pregunta por la posibilidad de que exista algún tipo de relación entre las variables, en el sentido de que conocido el valor de una de las variables se pueda decir algo sobre la otra. Él observa que si el peso está comprendido entre los 65 y los 70 kilos la estatura debe estar entre 1'65 y 1'75 metros, y que no hay individuos que en ese rango de pesos mida más de 1'75 metros. Es más, este ejemplo le hace ver que si uno de los cruces de las clases, por ejemplo (x_i, y_j), tiene frecuencia nula, el conocimiento de que una de las variables toma valores en la clase x_i imposibilita que la otra variable tome valores en la clase y_j, y viceversa. Pensando en su problema, llega a la conclusión de*

que existiría una dependencia total o funcional si el conocimiento del valor de una de las variables determina el valor que tomará la otra. Esto implica que si X depende funcionalmente de Y en cada fila hay una sola frecuencia distinta de cero, y si Y depende funcionalmente de X ocurre lo mismo con las columnas.

Por otra parte, le resulta evidente que las variables son independientes si fijado cualquier valor de una de las variables la otra variable mantiene sus porcentajes iguales a los de su distribución condicionada. Entre estas dos situaciones extremas, descubre que existen muchas posibilidades intermedias.

Por otra parte, se dice que X depende funcionalmente de Y, si conocido el valor que toma Y queda determinado el valor de X.

Para acabar esta sección se comprueba con un contraejemplo que la dependencia funcional no se establece en doble sentido.

Ejemplo 2.3 *En la siguiente distribución:*

X/Y	y_1	y_2	y_3
x_1	12	0	4
x_2	0	7	0

X depende funcionalmente de Y, puesto que conocido el valor de Y queda determinado el de X, pero el recíproco no se da, puesto que si X toma el valor x_1, Y puede tomar el valor y_1 o el y_3.

5. Medidas de dependencia. Coeficientes de relación

Los términos *asociación, correlación, contingencia, concordancia* y otros similares, se suelen utilizar como equivalentes muy a menudo. No obstante, haciendo un uso más correcto de la terminología estadística, aún con significado semejante, se puede considerar:

- correlación de variables propiamente dichas, o sea, medidas en escala de intervalo.
- concordancia de ordenaciones, entendiéndose como tales las denominadas variables ordinales, y
- asociación o contingencia de variables nominales o atributos.

Así, para clasificar los coeficientes que detectan y miden el grado de relación, o dependencia estadística, se ha tenido en cuenta el tipo y la naturaleza de las variables sometidas a estudio.

5.1. Variables continuas. Correlación

5.1.1. Covarianza

Para facilitar el estudio y la notación de la covarianza, se introduce previamente el concepto de *momentos bidimensionales*.

Se define el momento de orden (h, k) respecto al origen como:

$$a_{h,k} = \sum_{i=1}^{r} \sum_{j=1}^{s} x_i^h y_j^k f_{ij}.$$

Es fácil ver que $a_{1,0}$ es la media de X y que $a_{0,1}$ es la media de Y.

Por otro lado, el momento de orden (h, k) respecto a la media viene dado por:

$$m_{h,k} = \sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - \bar{x})^h (y_j - \bar{y})^k f_{ij}.$$

Constatándose que $m_{1,0}$ es cero, al igual que $m_{0,1}$, que $m_{2,0}$ y $m_{0,2}$ son las varianzas de X e Y, respectivamente, y que es posible expresar los momentos respecto a la media en función de los momentos respecto al origen. En particular se da la relación

$$m_{1,1} = a_{1,1} - a_{1,0} a_{0,1} \ .$$

A $m_{1,1}$ se le denomina *covarianza* de la distribución, denotándosele también por S_{xy}. Este coeficiente juega un importante papel en el estudio de la relación lineal entre las variables. Para analizar esta cuestión, se consideran las representaciones gráficas de la figura 2.1 que reflejan distintas situaciones, dichas representaciones reciben el nombre de *nube de puntos* o, también, *diagrama de dispersión.*

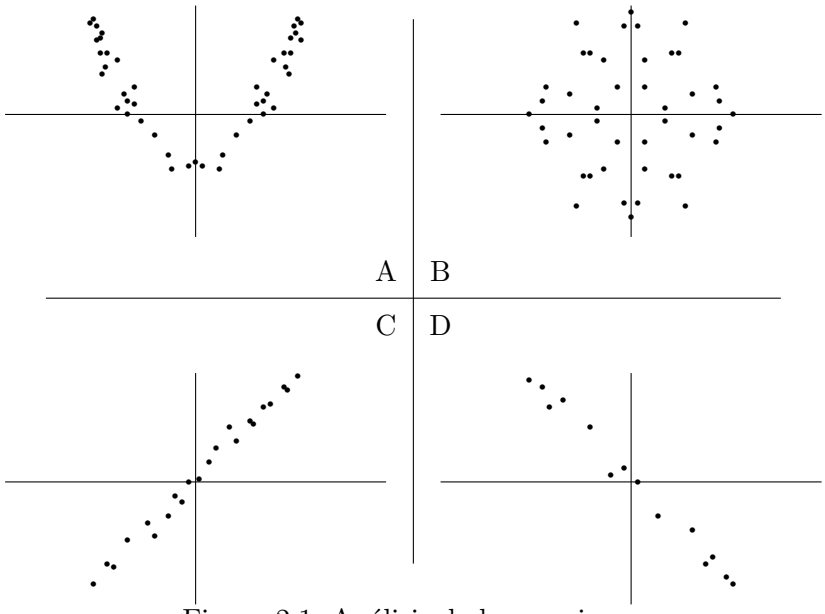

Figura 2.1: Análisis de la covarianza

El punto que viene determinado por la media de X y la media de Y constituye el *centro de gravedad* de las nubes de puntos en todos los casos.

Como se sabe, la covarianza viene dada por la expresión

$$S_{xy} = \sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \bar{x})(y_j - \bar{y})f_{ij} \ .$$

S_{xy} es una medida simétrica y se puede leer como la suma de los productos de las desviaciones de X por las desviaciones de Y con respecto a sus medias respectivas; de tal forma, que si el signo de la desviación

de X coincide con la de Y, como ocurre en el primer y tercer cuadrante, se genera un sumando positivo; y cuando el signo es distinto -segundo y cuarto cuadrante- la aportación a la covarianza es negativa. Por tanto, la concentración de valores en los distintos cuadrantes determina el signo y la cuantía de S_{xy}. Así, en los casos A y B de la figura 2.1, S_{xy} se aproxima a cero, en el caso C va a ser alta y positiva, y en el D alta y negativa. Por tanto, se está en condiciones de afirmar que la covarianza detecta la relación lineal entre las variables y el sentido de ésta, pero no distingue entre la no presencia de relación, caso B, y la existencia de alguna dependencia no lineal, caso A. De todas formas, aún para el estudio de relaciones lineales la covarianza adolece de ciertos problemas, como el de venir acompañada de las unidades de las variables y el de depender del número de observaciones.

5.1.2. Coeficiente de correlación de Pearson

Para obviar las carencias de la covarianza se introduce el *coeficiente de correlación lineal* o *coeficiente de correlación de Pearson*

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

que es una medida adimensional, ordinal, toma valores en el intervalo $[-1, 1]$ y tiene el signo de S_{xy}, por lo que cuando la relación lineal entre X e Y es exacta y directa, es decir, todos los puntos se encuentran sobre una recta con pendiente positiva, vale 1, cuando es exacta e inversa, es decir, todos los puntos se encuentran sobre una recta con pendiente negativa, vale -1 y cuando no hay relación lineal 0; con un análisis lógico para las posiciones intermedias. Cuando r vale cero, se dice que las *variables* están *incorreladas*.

En el caso lineal, al cuadrado de r se le llama *coeficiente de determinación* y se le denota por R^2, representando una medida cardinal o cuantitativa para medir la relación lineal entre las variables. Se estudia este coeficiente con más detalle en el capítulo siguiente.

Se concluye este apartado indicando que la independencia implica incorrelación, pero el recíproco no siempre es cierto. Este resultado es

consecuencia de que la independencia supone la descomposición de los momentos de orden (h, k) (respecto al origen o respecto a la media) en el producto de los momentos $(h, 0)$ y $(0, k)$; así, $a_{1,1} = a_{1,0}a_{0,1}$ y por tanto $S_{xy} = m_{1,1} = a_{1,0}a_{0,1} - a_{1,0}a_{0,1} = 0$, con lo que $r = 0$ y las variables están incorreladas. En sentido contrario, la incorrelación sólo implica esa descomposición para el momento $(1, 1)$. En cierta forma, se puede decir que la incorrelación es una independencia de primer orden o lineal.

Ejercicio 2.1 *Demuestre que las variables X e Y de la siguiente distribución:*

X	Y
2	8
1	5
0	4
−1	5
−2	8

están incorreladas, pero no son independientes; es más, existe una relación funcional entre ellas. Indíquela.

Por tanto, el coeficiente de correlación de Pearson mide el grado de relación lineal entre dos variables cuantitativas indicando el sentido directo o inverso de la relación. Es el más común de todos los coeficientes porque es la base de otras muchas medidas de relación entre variables de distinta naturaleza, de hecho, a menudo se tiende a interpretar cualquier coeficiente como si del de Pearson se tratase.

5.1.3. Coeficiente de correlación biserial

Se utiliza para establecer el grado de correlación entre dos variables cuantitativas cuando una de ellas ha sido dicotomizada previamente. Se trata de una modificación del coeficiente de correlación de Pearson entre una variable continua X y otra Y que se ha dicotomizado y que en origen responde a una estructura de distribución normal[1].

[1] La distribución normal se estudiará en el capítulo 5

El coeficiente de correlación biserial se denota por r_b y se puede calcular indistintamente por cualquiera de las siguientes expresiones:

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{S_x}\left(\frac{pq}{y}\right) = \frac{\bar{X}_p - \bar{X}}{S_x}\left(\frac{p}{y}\right) \ ,$$

donde:

X es la variable continua

Y es la variable dicotomizada

$\bar{X}_p$ es la media de X cuando Y vale 0

$\bar{X}_q$ es la media de X cuando Y vale 1

$\bar{X}$ es la media de la distribución marginal de X

S_x es la desviación típica de la marginal de X

p es la proporción de elementos con asignación 0 en la variable Y

q es la proporción de elementos con asignación 1 en la variable Y, $(q = 1 - p)$

y es el valor de la ordenada correspondiente a un valor de x que divide el área de la distribución normal tipificada en dos partes, una igual a p y otra igual a q.

Se interpreta de forma análoga al coeficiente de correlación de Pearson en lo referente a la intensidad de la relación, no a su sentido; además, cuando la correlación es alta y el requisito de normalidad de Y no se cumple de forma estricta, el coeficiente de correlación biserial puede valer más de 1 o menos de -1.

Como variante, aunque con idéntica interpretación y similar notación y expresión, se debe tener presente el *coeficiente de correlación biserial–puntual*, que se utiliza para medir la correlación entre una variable continua y otra dicotómica por naturaleza, definido por:

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{S_x}\sqrt{pq} = \frac{\bar{X}_p - \bar{X}}{S_x}\sqrt{\frac{p}{q}} \ .$$

Observación 2.1 *Desde el punto de vista práctico, el coeficiente de correlación biserial se usa sobre todo para hacer inferencias. Su cálculo necesita conocer la distribución normal, puesto que es necesario obtener el valor y.*

Ejemplo 2.4 *Con la finalidad de buscar el mayor rendimiento de la tierra, un agricultor, preocupado por su cosecha de naranjas, está interesado en estudiar el grado de relación entre la cantidad de fruta recogida y la lluvia caída en los últimos 10 años. Para ello parte de la siguiente información, obtenida por él mismo, en la que ha clasificado los años en secos (S) o lluviosos (L):*

Naranjas (Tm)	*Año*	*Naranjas (Tm)*	*Año*
10'01	*L*	*9'57*	*L*
8'2	*L*	*5'9*	*S*
7'23	*S*	*6'8*	*S*
11'45	*L*	*6'8*	*S*
8'50	*L*	*7'9*	*L*

Para estudiar a partir de estos datos la relación entre las variables, se recurre al coeficiente de correlación biserial-puntual[2]*, realizando la división de la cosecha en dos series, la obtenida en temporada de sequía, con valor asignado 1, y la obtenida en temporada de lluvia, con asignación el valor 0. Se denota por X la cantidad de naranjas y por Y si la temporada es de lluvia o de sequía.*

$$r_{bp} = \frac{9'27167 - 6'6825}{1'70077}\sqrt{0'6 \cdot 0'4} = 0'7457.$$

Lo que indica una relación de dependencia relativamente fuerte entre las variables.

[2]Se ha utilizado el coeficiente de correlación biserial-puntual y no el coeficiente de correlación biserial, debido a que aunque la variable "lluvia caída" es en principio continua y probablemente Normal, el uso del coeficiente de correlación biserial requiere conocimientos hasta ahora no adquiridos, como se indica en la observación 2.1.

Dada la inseguridad ante las medidas de la concentración de lluvia anual por metro cuadrado que obtuvo el agricultor, éste decide prescindir por completo de sus datos y recurrir a la información que sobre el tema proporciona anualmente el instituto meteorológico, el cuál le proporciona la cantidad de lluvia caída cada año. Se denota por X la cantidad de naranjas y por Y los m^3 de lluvia.

De esta forma los datos han sido transformados en:

Nar. (Tm)	*Lluvia (m^3)*	*Nar. (Tm)*	*Lluvia (m^3)*
10'01	*1'3*	*9'57*	*1'4*
8'2	*0'9*	*5'9*	*0'67*
7'23	*0'87*	*6'8*	*0'56*
11'45	*1'75*	*6'8*	*0'87*
8'50	*0'96*	*7'9*	*1'24*

Con esta información se analiza la relación de las variables con el coeficiente de correlación de Pearson, ya que ambas son continuas.

$$\begin{aligned} r &= \frac{S_{xy}}{S_x S_y} = 0'917511 \\ R^2 &= 0'841827. \end{aligned}$$

Con esto se concluye que existe una fuerte dependencia lineal y además directa entre ambas variables, es decir, la cosecha de naranjas es mayor cuando mayor es la cantidad de lluvia caída.

5.2. Variables ordinales. Concordancia

5.2.1. Coeficiente de correlación por rangos de Spearman

Este coeficiente se utiliza para medir la relación entre dos sucesiones de valores ordinales. Es el coeficiente de correlación de Pearson para las llamadas variables cuasi–cuantitativas, discretas, o bien, para aquellas cuantitativas que han sido transformadas en ordinales (n primeros

números naturales para cada variable) tiene la forma

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

donde:

r_s es el coeficiente de correlación por rangos de Spearman

d_i es la diferencia entre el valor ordinal de la variable X y el de la variable Y en el elemento i-ésimo

n es el tamaño de la muestra

Se verifica que $-1 \leq r_s \leq 1$.

Si hay un gran número de elementos con el mismo valor en alguna de las dos variables, es decir, si hay muchos empates, es conveniente recurrir a las correcciones de este coeficiente. Quedando el coeficiente como

$$r_s = \frac{x^2 + y^2 - \sum_{i=1}^{n} {d_i}^2}{2\sqrt{x^2 y^2}} \ ,$$

con:

$$x^2 = \frac{n^3-3}{12} - \sum_{i=1}^{n} T_{x_i} \ , \qquad T_{x_i} = \frac{t_{x_i}^3 - t_{x_i}}{12} \ ,$$

$$y^2 = \frac{n^3-3}{12} - \sum_{i=1}^{n} T_{y_i} \ , \qquad T_{y_i} = \frac{t_{y_i}^3 - t_{y_i}}{12} \ ,$$

donde:

t_{x_i} es el número de empates en el rango i de la variable X

t_{y_i} es el número de empates en el rango i de la variable Y

Sus características e interpretación son similares a las del coeficiente de correlación de Pearson.

5.2.2. Coeficiente τ de Kendall

De forma análoga al coeficiente de Spearman, el coeficiente τ considera el orden de los n objetos o elementos tanto de una variable como de la otra e intenta medir el grado de concordancia o correspondencia entre ellos. Dicho coeficiente viene dado por

$$\tau = \frac{P - Q}{P + Q} \ ,$$

donde:

τ es el coeficiente de Kendall

P el número de coincidencias o acuerdos

Q el número de no coincidencias o desacuerdos

Nuevamente, si hay gran número de empates, conviene aplicar una corrección, quedando el coeficiente como

$$\tau = \frac{P - Q}{\sqrt{\frac{1}{2}n(n-1) - T_x}\sqrt{\frac{1}{2}n(n-1) - T_y}} \ ,$$

con:

$$\begin{aligned} T_x &= \frac{1}{2}\sum_{i=1}^{n} t_{x_i}(t_{x_i} - 1) \ , \\ T_y &= \frac{1}{2}\sum_{i=1}^{n} t_{y_i}(t_{y_i} - 1) \ , \end{aligned}$$

donde t_{x_i} y t_{y_i} coinciden con los definidos para el coeficiente de correlación de Spearman.

Sus características e interpretación son similares a las del coeficiente de correlación de Pearson.

X	rg(X)	Y	rg(Y)	d_i	$d_i{}^2$
5	1	1	3	-2	4
6	2	3	$9'5$	$-7'5$	$56'25$
7	3	2	$6'5$	$-3'5$	$12'25$
8	4	1	3	1	1
9	5	1	3	2	4
10	6	0	1	5	25
11	7	2	$6'5$	$0'5$	$0'25$
12	8	2	$6'5$	$1'5$	$2'25$
13	9	3	$9'5$	$-0'5$	$0'25$
14	10	2	$6'5$	$3'5$	$12'25$
					$117'5$

Tabla 2.6: Cálculo del coeficiente de correlación de Spearman

5.2.3. Coeficiente γ de Goodman–Kruskal

Se utiliza para medir el grado de concordancia entre dos variables ordinales, estando especialmente indicado cuando hay muchas observaciones y pocos valores posibles, es decir, muchos empates.

Su expresión e interpretación es muy similar a la del coeficiente de Kendall, considerando la proporción de pares semejantes y la proporción de pares no semejantes entre los empatados, resultando

$$\gamma = \frac{n_s - n_d}{n_s + n_d}$$

donde:

γ es el coeficiente de Goodman-Kruskal

n_s es el números de pares semejantes o no invertidos

n_d es el número de no semejantes o invertidos

Ejemplo 2.5 *Se pretende estudiar la relación existente entre la edad (E) y el número de hermanos (H) de un grupo de 10 chicos, para ello se cuenta con los siguientes*

datos:

E	*6*	*12*	*8*	*11*	*10*	*7*	*9*	*14*	*13*	*5*
H	*3*	*2*	*1*	*2*	*0*	*2*	*1*	*2*	*3*	*1*

Se calculará el coeficiente de correlación por rangos de Spearman, dado que se están tratando variables cuantitativas. Obtendremos primero la versión original de dicho coeficiente.

A partir de los cálculos recogidos en la tabla 2.6, se obtiene

$$\begin{aligned} r_s &= 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 117'5}{10 \cdot 99} \\ &= 1 - \frac{705}{990} = 1 - 0'7121 = 0'2879. \end{aligned}$$

No obstante, debido al elevado número de empates debería emplearse el coeficiente modificado para dicho caso, o incluso al coeficiente de Goodman-Kruskal. Se calculará el coeficiente modificado de Spearman.

$$\begin{aligned} x^2 &= \frac{10^3-3}{12} - 0 = \frac{997}{12} \\ y^2 &= \frac{10^3-3}{12} - (\frac{24}{12} + \frac{62}{12} + \frac{6}{12}) = \frac{905}{12}. \end{aligned}$$

Por tanto el coeficiente modificado queda como

$$r_s = \frac{\frac{997+905}{12} - 117'5}{2\sqrt{\frac{997 \cdot 905}{12^2}}} = 0'2589$$

que es ligeramente inferior al original. Del resultado obtenido se concluye la escasa concordancia entre la edad y el número de hermanos.

5.3. Atributos. Contingencia

5.3.1. Coeficiente χ^2

El coeficiente χ^2 se utiliza para medir el grado de asociación entre dos variables cualitativas con h y k categorías respectivamente. Este estadístico está basado en la comparación de las frecuencias observadas con las esperadas bajo una cierta hipótesis, generalmente de independencia, respondiendo a la expresión

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \ ,$$

donde:

o_{ij} son las frecuencias observadas o empíricas

e_{ij} son las frecuencias esperadas o teóricas

Cuando h y k toman el valor 2, es decir, cuando se está trabajando con una tabla de contingencia 2×2, se aplica la denominada corrección de Yates, resultando el coeficiente:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(|o_{ij} - e_{ij}| - 0'5)^2}{e_{ij}}.$$

El coeficiente siempre toma valores no negativos, pero al tratarse de una medida no acotada, es de difícil interpretación por sí sola, si bien, cuanto más relacionadas estén las variables sometidas a estudio más se alejará el coeficiente del valor 0. Su valor depende del número de observaciones y de las categorías en que éstas se dividen, por tanto el coeficiente χ^2 y sus derivados no son comparables con cualquier otro coeficiente obtenido con distinto número de categorías.

Este coeficiente χ^2 es la base de otros obtenidos a partir de él y que solucionan el problema de su falta de acotación.

5.3.2. Coeficiente de contingencia

Es uno de los coeficientes derivados del χ^2, resultando útil bajo las mismas condiciones que aquel pero con mayores posibilidades de interpretación. Se denota por C y se define como

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

siendo n el tamaño muestral.

Se cumple que $0 \leq C \leq 1$ y mide la intensidad de la relación sin indicar su sentido.

5.3.3. Coeficiente de Cramer

Es otro de los coeficientes derivados del χ^2. Se caracteriza por V y su expresión es

$$V = \sqrt{\frac{\chi^2}{n(m-1)}}$$

siendo:

n el tamaño muestral

m el mínimo entre h y k

h el número de categorías de la variable X

k el número de categorías de la variable Y

Se verifica que $0 \leq V \leq 1$ y se interpreta igual que el coeficiente de contingencia, teniendo en cuenta que sólo proporciona información sobre la relación entre las variables y no sobre el sentido de la misma.

5.3.4. Coeficiente φ

Se trata de un coeficiente especialmente indicado para medir la asociación entre dos variables dicotómicas. Su expresión es

$$\varphi = \frac{n_{11}n_{22} - n_{21}n_{12}}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}$$

donde:

n_{11} es el número de veces que se da el par $(X = 0, Y = 0)$

n_{12} es el número de veces que se da el par $(X = 0, Y = 1)$

n_{21} es el número de veces que se da el par $(X = 1, Y = 0)$

n_{22} es el número de veces que se da el par $(X = 1, Y = 1)$

En cuanto a su interpretación, el coeficiente toma valores en el intervalo $[-1, 1]$, midiendo de forma similar al coeficiente de Pearson la intensidad de la asociación entre las dos variables; salvo que alguna de las frecuencias n_{ij} sea nula, en cuyo caso el coeficiente vale 1 ó -1.

En el caso en que se estudie el grado de correlación entre dos variables cuantitativas dicotomizadas, X e Y, siempre y cuando éstas respondan a variables continuas bajo una ley normal (que se estudiará más adelante), el coeficiente φ suele denominarse *coeficiente de correlación tetracórica.*

Ejemplo 2.6 *De cara a la planificación del próximo curso sería conveniente analizar la relación entre el nivel de estudios del padre y la orientación del alumno hacia las ciencias. Se cuenta para ello con la información obtenida en el centro*

	Estudios padre			
Orientación	*Nulo*	*Básico*	*Medio*	*Superior*
Orientado	*23*	*12*	*34*	*32*
No orientado	*18*	*42*	*16*	*27*

Como se trata de una tabla de contingencia, se calcula el coeficiente χ^2 y sus derivados para hacer posible la interpretación.

	Nulo	*Básico*	*Medio*	*Superior*	
Orientado	*23*	*12*	*34*	*32*	*101*
No orientado	*18*	*42*	*16*	*27*	*103*
	41	*54*	*50*	*59*	*204*

e_{ij}	1	2	3	4
1	$\frac{101\cdot 41}{204}$	$\frac{101\cdot 54}{204}$	$\frac{101\cdot 50}{204}$	$\frac{101\cdot 59}{204}$
2	$\frac{103\cdot 41}{204}$	$\frac{103\cdot 54}{204}$	$\frac{103\cdot 50}{204}$	$\frac{103\cdot 59}{204}$

e_{ij}	1	2	3	4
1	20′30	26′73	24′75	29′21
2	20′70	27′26	25′24	29′79

$$\begin{aligned}
\chi^2 &= \frac{(23-20'30)^2}{20'30} + \frac{(12-26'73)^2}{26'73} \\
&\quad + \frac{(34-24'75)^2}{24'75} + \frac{(32-29'21)^2}{29'21} \\
&\quad + \frac{(18-20'70)^2}{20'70} + \frac{(42-27'26)^2}{27'26} \\
&\quad + \frac{(16-25'24)^2}{25'24} + \frac{(27-29'79)^2}{29'79} \\
&= 0'36 + 8'12 + 3'46 + 0'26 \\
&\quad + 0'35 + 7'97 + 3'38 + 0'26 \\
&= 24'16
\end{aligned}$$

$$C = \sqrt{\frac{24'16}{24'16 + 204}} = 0,3254$$

$$V = \sqrt{\frac{24'16}{204 \cdot 1}} = 0,3441.$$

Luego podemos concluir que el grado de asociación entre las variables es pequeña.

Ejemplo 2.7 *En el conservatorio de música de una ciudad se pretende estudiar la relación existente entre el sexo del alumnado y su afición por los instrumentos de viento. Para ello, controlados los 482 estudiantes se tiene:*

	Hombre	*Mujer*
Aficionado	*150*	*97*
No aficionado	*123*	*112*

Dada la naturaleza dicotómica de las variables, se recurre al coeficiente φ

$$\varphi = \frac{150 \cdot 112 - 123 \cdot 97}{\sqrt{247 \cdot 235 \cdot 273 \cdot 209}} = \frac{4869}{57548'8} = 0'08.$$

Con esto se pone de manifiesto la inexistencia de relación entre el sexo y la preferencia por los instrumentos de viento.

Ejemplo 2.8 *Volviendo al ejemplo planteado en el estudio de variables continuas, véase ejemplo 2.4, y considerando un caso aún más general, se supone que la información que conservó el agricultor después de la cosecha de cada año es tan sólo el recuerdo de si fue buena o mala. Así los datos con los que se cuenta para el estudio de las variables son:*

	Mala	*Buena*
Seco	*0*	*4*
Lluvioso	*5*	*1*

Haciendo uso ahora, dado que las variables aparecen dicotomizadas, del coeficiente de correlación tetracórica

$$r_t = \frac{5 \cdot 4 - 0 \cdot 1}{\sqrt{6 \cdot 4 \cdot 5 \cdot 5}} = \frac{20}{24'4948} = 0'8165.$$

> *Poniendo nuevamente de manifiesto la relación entre la cantidad de naranjas y la lluvia. Hay que tener en cuenta que el signo que acompaña al coeficiente depende de la asignación de valores a la hora de dicotomizar las variables, por consiguiente, es interpretable la intensidad de la relación, no el sentido de la misma.*

Son varios los coeficientes de relación que a lo largo de esta sección se han ido enumerando, coincidiendo con los que por sus características, naturaleza y facilidad de cálculo son más utilizados y, por consiguiente, conocidos en los distintos campos donde su aplicación tiene cabida.

6. Ejercicios

6.1. Ejercicio resuelto

2.1 Se ha clasificado el peso de los huevos, Y, de un cierto tipo de pez en función del peso de la madre, X, obteniéndose los resultados de la tabla adjunta.

X \ Y	[25,27)	[27,29)	[29,31)	[31,33)
[500,550)	15	11	18	0
[550,600)	12	14	0	12
[600, 650)	0	3	7	18

Calcule:

a) La distribución del peso del huevo.

b) La distribución del peso de la madre cuando el huevo tiene su peso comprendido entre $[25, 27)$.

c) La media, la mediana y la moda del peso de los huevos.

d) El nivel de representatividad de la media del peso de la madre cuando el huevo está comprendido entre $[25, 27)$.

e) Estudiar si las variables son independientes.

f) El grado de dependencia lineal entre estas variables.

Solución:

a) En realidad el primer apartado lo que está pidiendo es la distribución marginal de la variable Y. Por tanto,

$$\begin{aligned} n_{[25,27)} &= n(y \in [25,27)) = n(x \in [500,550), y \in [25,27)) + \\ &\quad n(x \in [550,600), y \in [25,27)) + n(x \in [600,650), y \in [25,27)) \\ &= 15 + 12 + 0 = 27 \end{aligned}$$

procediendo de igual forma con el resto de intervalos donde Y toma valores, se obtiene que:

Y	n_i
[25,27)	27
[27,29)	28
[29,31)	25
[31,33)	30

b) Se pide la distribución de la variable X condicionada a que la variable Y tome valores en el intervalo $[25,27)$, es decir,

$$\begin{aligned} f_{_{|[25,27)}[500,550)} &= f(x \in [500,550)/y \in [25,27)) \\ &= \frac{f(x \in [500,550), y \in [25,27))}{f(y \in [25,27))} \\ &= \frac{\frac{15}{110}}{\frac{27}{110}} = \frac{15}{27} = \frac{5}{9} \end{aligned}$$

procediendo de igual forma, se tiene:

$X/Y \in [25,27)$	f_i
$[500,550)$	$\frac{5}{9}$
$[550,600)$	$\frac{4}{9}$
$[600,650)$	0

c) Se calcula la media de variable Y,

$$\bar{y} = \frac{26 \cdot 27 + 28 \cdot 28 + 30 \cdot 25 + 32 \cdot 30}{110} = 29'05.$$

Para calcular la mediana se tiene en cuenta el apartado a), donde se ve que el primer intervalo cuya frecuencia absoluta acumulada supera el 50 % de los datos, es decir, 55, es el intervalo [29,31). Por tanto la mediana viene dada por

$$Me = 29 + \frac{55 - 55}{25} 2 = 29.$$

Para calcular la moda, se observa que todos los intervalos tienen igual amplitud y que el intervalo con mayor frecuencia es el [31,33), por tanto la moda es

$$Mo = 31 + \frac{0}{25 + 0} 2 = 31.$$

d) Para calcular el nivel de representatividad de la media se utiliza el coeficiente de variación, para ello, se calcula previamente la media y la desviación típica de la variable requerida. La distribución de esta variable se ha calculado en el apartado b), por tanto

$$\bar{x}/y \in [25, 27) = 525\frac{5}{9} + 575\frac{4}{9} = \frac{4925}{9} \qquad y$$

$$S^2_{x/y\in[25,27)} = 525^2\frac{5}{9} + 575^2\frac{4}{9} - \left(\frac{4925}{9}\right)^2 = \frac{50000}{81}$$

Con lo que el coeficiente de variación es

$$CV = \frac{\sqrt{\frac{50000}{81}}}{\frac{4925}{9}} = 0'045.$$

lo que supone que la media es muy representativa debido a que es muy pequeño el coeficiente de variación.

e) Para tratar la independencia se considera un par $(x, y) \in [500, 550) \times [25, 27)$. Se sabe que

$$f(x \in [500, 550), y \in [25, 27)) = \frac{15}{110},$$

además, se tiene que

$$f(x \in [500, 550)) = \frac{44}{110}$$

y que

$$f(y \in [25, 27)) = \frac{27}{110}$$

con lo cual,

$$f(x \in [500, 550), y \in [25, 27)) = \frac{15}{110} \neq \frac{1188}{12100} =$$

$$= f(x \in [500, 550)) f(y \in [25, 27)).$$

Por tanto, se tiene que las variables no son independientes.

f) Para cuantificar el grado de dependencia lineal entre dos variables se calcula el coeficiente de determinación

$$R^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}.$$

Se necesita calcular S_X^2, S_Y^2 y S_{XY}:

$$S_X^2 = \frac{525^2 \cdot 44 + 575^2 \cdot 38 + 625^2 \cdot 28}{110} - \left(\frac{62450}{110}\right)^2 = 1583'47$$

$$S_Y^2 = \frac{26^2 \cdot 27 + 28^2 \cdot 28 + 30^2 \cdot 25 + 32^2 \cdot 30}{110} - \left(\frac{3196}{110}\right)^2 = 5'14$$

$$\begin{aligned} S_{XY} \;=\; & \frac{1}{110}\,(26 \cdot 525 \cdot 15 + 26 \cdot 575 \cdot 12 + 26 \cdot 625 \cdot 0 + 28 \cdot 525 \cdot 11 \\ & +28 \cdot 575 \cdot 14 + 28 \cdot 625 \cdot 3 + 30 \cdot 525 \cdot 18 + 30 \cdot 575 \cdot 0 \\ & +30 \cdot 625 \cdot 7 + 32 \cdot 525 \cdot 0 + 32 \cdot 575 \cdot 12 + 32 \cdot 625 \cdot 18) \\ & -\frac{199590200}{12100} = 44'03. \end{aligned}$$

Con lo cual

$$R^2 = \frac{44'03^2}{5'14 \cdot 1583'47} = 0'24,$$

de donde se deduce que el grado de dependencia lineal es bastante bajo.

6.2. Ejercicios propuestos

2.1. Durante el año 1993 se han observado la población y el número de viviendas de renta libre unifamiliares en 32 municipios de la provincia de Cádiz. Los datos obtenidos se han tabulado, obteniéndose:

Y \ X	0-10	10-30	30-70	70-150	150-250
[0-2)	3				
[2-5)	3				
[5-10)	1	3	1	1	
[10-30)	2	2	6	1	
[30-80)			2	1	2
[80-180)		1	1	1	1

donde:
X = Número de viviendas
Y = Población en miles de personas

a) Obtenga las distribuciones marginales de X e Y.

b) Indique qué distribución es más homogénea.

c) Obtenga la distribución de las viviendas unifamiliares para los municipios entre dos mil y treinta mil habitantes.

d) Calcule los momentos: a_{01}, a_{02}, a_{10}, a_{11}, m_{02}, m_{20}, m_{21}.

e) Entre las poblaciones de más de 10.000 habitantes, indique cuál es el número de viviendas libres construidas más frecuente.

f) Obtenga la covarianza y el coeficiente de correlación de las variables X e Y e interprételo.

2.2. De la variable bidimensional (X, Y) se conoce su coeficiente de correlación, $r = 0'83$, y sus varianzas, $S_x^2 = 5'32$ y $S_y^2 = 8'41$. Si se multiplican por 3 los valores de X y por 2 los valores de Y, ¿que repercusión tienen estas transformaciones en la covarianza y en el coeficiente de correlación?

2.3. La tabla 2.7 muestra una serie histórica sobre el Olivar Español que recoge la superficie, rendimiento y producción, durante el periodo 1965-1979. donde:

Año	X	Y	Z
1965	73'6	69'8	8'5
1966	98'1	62'5	6
1967	99'8	98'5	8'7
1968	107'7	102'5	6
1969	107'7	97'4	3'7
1970	122	113'8	8'9
1971	127	118	7'9
1972	138'1	128'1	10'1
1973	152'1	145'8	6'8
1974	144'8	139'8	5
1975	160'7	152'9	11'1
1976	150'2	143'4	9'8
1977	152'1	146	9'5
1978	167'3	162'1	10'8
1979	165	160'2	10

Tabla 2.7: Datos ejercicio 2.3

X = Superficie en miles de Ha.
Y = Rendimiento en Qm/Ha..
Z = Producción en miles de Tm..

Se pide:

a) El diagrama de dispersión de las variables X e Y.

b) Las medidas más representativas para cada una de las variables, indicando su representatividad

c) El estudio de la relación entre las variables XY, XZ e YZ.

2.4. La siguiente tabla muestra la relación existente entre la lluvia caída, en l/m^2, en el periodo octubre–mayo y la producción obtenida

en kilogramos por olivo.

X	300	400	500	600	700
Y	13	26	40	57	64
Y	24	21	31	45	69
Y	17	17	38	51	57
Y	11	26	34	58	76
Y	20	30	27	44	74

donde X representa la lluvia caída e Y la producción obtenida en kilogramos por olivo.

a) Represente el diagrama de dispersión.

b) Indique si existe alguna tendencia.

c) Cuantifique y comente la relación existente entre las dos variables.

2.5. Dada la siguiente tabla de doble entrada con valores porcentuales:

Y X	2	3	4
0	$0'22$	$0'13$	$0'04$
1	$0'16$	$0'11$	$0'05$
2	$0'08$	$0'16$	$0'05$

a) Obtenga la distribución marginal de X. Calcule su media, moda y mediana.

b) Calcule la media de Y cuando X toma el valor 3.

c) Estudie la dependencia de las variables X e Y.

2.6. Estudiar la coherencia de los siguientes resultados correspondientes a una variable bidimensional:

$$S_{xy} = -179'5,\ S_x^2 = 36'8,\ S_y^2 = 525,\ M_e(X) = -12'3,\ \bar{Y} = 0$$

2.7. De los modelos de una determinada marca de automóviles se considera el consumo medio y el tiempo de aceleración de 0 a 100

Km./h., obteniéndose los siguientes resultados:

	Cons. (lit.)				
Acel. (seg.)	[5, 6)	[6, 7)	[7, 8)	[8, 9)	[9,12)
[7, 9)			1	3	2
[9,11)		1	2	4	4
[11,14)	1	5	1	3	3
[14,18)	3	2	1		

a) Dibuje y comente el diagrama de dispersión.
b) Obtenga el consumo medio de carburante.
c) Obtenga el tiempo de aceleración medio.
d) Indique cuál de las dos medias es más representativa.
e) Estudie la relación existente entre las dos características.

2.8. A un grupo de estudiantes se les preguntó por el tiempo que tardan en llegar desde su hogar hasta la Facultad, X (minutos), el tiempo que le dedican diariamente al estudio, Y (horas), y las calificaciones obtenidas en la asignatura de Estadística, Z, obteniéndose las siguientes respuestas:

(40, 4, 4), (45, 3, 3), (30, 4, 5), (40, 4, 5), (80, 2, 5), (20, 3, 5)
(10, 1'5, 6), (10, 4, 6), (20, 4, 6), (45, 3, 3), (20, 4, 4), (30, 4, 7)
(30, 3, 7), (20, 4, 6), (30, 1, 6), (10, 5, 5), (15, 5, 5), (20, 6, 5)
(20, 3, 7), (20, 4, 5), (20, 5, 6), (60, 2, 3), (60, 5, 5)

a) Obtenga el diagrama de dispersión correspondiente al tiempo dedicado al estudio y las calificaciones obtenidas en Estadística.
b) ¿Se aprecia alguna tendencia?
c) Estudie las relaciones exitentes entre XY, XZ e YZ.

2.9. Al mismo grupo del ejercicio anterior se le ha pedido que escriba un dígito al azar entre 0 y 9 así como el número de hermanos que tiene, obteniéndose los siguientes pares de valores:

(7, 4), (0, 1), (2, 1), (2, 0), (9, 4), (7, 4), (6, 3), (8, 5)
(7, 3), (3, 2), (7, 3), (2, 1), (7, 4),(7, 3), (8, 4), (8, 5)
(5, 3), (3, 1), (4, 2), (4, 2), (5, 3), (2, 0), (4, 2)

¿Existe alguna relación entre las variables?, ¿de qué tipo?

2.10. Sea la variable bidimensional (X, Y) de la que se han obtenido 25 pares de valores, con los siguientes resultados:

$$r = 0'65, \qquad \sum_{i=1}^{25} x_i = 238, \qquad \sum_{i=1}^{25} y_i = 138$$

$$\sum_{i=1}^{25} x_i^2 = 12678, \qquad \sum_{i=1}^{25} y_i^2 = 2732$$

a) Calcule medias, varianzas y covarianza de X e Y.
b) Indique qué variable es más homogénea.

2.11. En cada uno de los estanques, A y B, se tienen 100 ejemplares de una variedad de dorada todas ellas afectadas por un parásito. La alimentación es idéntica en ambos estanques salvo en un producto encaminado a eliminar dichos parásitos, suministrado únicamente a los del estanque A. Posteriormente, se encuentra que en 71 ejemplares del A y en 58 del B han desaparecido los parásitos. Halle el coeficiente de contingencia y el coeficiente de Cramer e interprete los resultados.

2.12. Demuestre que el coeficiente de Cramer está comprendido entre 0 y 1.

2.13. Demuestre que el valor máximo del coeficiente de contingencia de una tabla $k \times k$ es $\sqrt{\frac{(k-1)}{k}}$.

2.14. Antes de un campeonato de fútbol las apuestas indican que las posiciones que ocuparán al finalizar éste cinco de los equipos participantes es $A > B > C > D > E$. Un jugador apuesta que el orden final será $A > D > B > E > C$. Mida el grado de similitud entre ambas ordenaciones.

2.15. Se mide el tiempo que 10 estudiantes tardan en realizar dos experimentos en los que predominan el cálculo mental y la capacidad espacial, respectivamente. Si los valores obtenidos son:

Estudiante	1	2	3	4	5	6	7	8	9	10
Tarea 1	32	37	45	50	48	56	78	69	77	79
Tarea 2	41	33	46	47	40	71	70	65	75	83

Estudie la relación entre los resultados obtenidos en ambas tareas.

2.16. Dos grupos de estudiantes deciden clasificar a 11 profesores. Los resultados se muestran a continuación:

Prf.	Es.	Og.	Mt.	Ig.	Fs.	Ge.	Cá.	FQ.	Oc.	Bi.	In.
Gr.I	7	4	2	8	9	10	11	6	1	3	5
Gr.II	8	2	1	5	3	11	9	10	7	4	6

Compare ambas clasificaciones.

2.17. En un grupo de 100 personas se estudian los atributos Color del Cabello (Moreno, Rubio, Castaño) y Color de los Ojos (Negro, Marrón, Azul y Verde), obteniéndose la siguiente tabla de contigencia:

Ojos \ Cabello	Moreno	Rubio	Castaño
Negro	20	8	4
Marrón	16	2	11
Azul	5	8	8
Verde	10	5	3

¿Están relacionados dichos atributos?