Distribuciones Bidimensionales

Tablas estadísticas bidimensionales

El objetivo de las Ciencias es establecer leyes que relacionen las propiedades de los objetos que se estudian, con ello se consigue que pueda explicarse y predecirse una característica, en función de los valores de otras características. Habitualmente las relaciones se refieren a propiedades que han sido cuantificadas, lo que permite expresarlas en términos de correspondencias numéricas.

La determinación, o la constatación, de una ley requerirá que se observen y registren ordenadamente, todas las variables que intervienen en la expresión de dicha ley. Por ello en este capítulo, situándonos en la relación más sencilla posible, la que liga los valores de dos variables, nos ocuparemos de la tabulación, síntesis y medida de la relación, de las distribuciones bidimensionales, es decir, de aquellas en las que simultáneamente se han registrado dos variables en cada individuo. Para ilustrar estos conceptos utilizaremos como ejemplo la situación experimental de Sternberg, donde se relaciona la longitud de listas de palabras sin sentido, presentadas a los sujetos, con el tiempo de respuesta necesario para decidir si una nueva palabra pertenece a la lista original.

Supongamos que en una serie de repeticiones de esta tarea, se han obtenido los siguientes datos:

(2,2) (2,3) (2,3) (2,4) (2,2) (3,3) (3,4) (3,2) (3,3) (3,2) (3,5) (3,4) (3,3) (4,5)
(4,4) (4,5) (4,3) (4,2) (4,4) (4,3) (5,4) (5,5) (5,5) (5,3) (5,5) (6,5) (6,6) (6,4)
(6,4) (6,3) (6,5) (7,5) (7,6) (7,5) (7,6) (7,4) (7,3)

donde el primer número representa la longitud de la lista y el segundo el tiempo de respuesta, expresado en décimas de segundo.

Podríamos, de forma análoga al caso de las distribuciones unidimensionales, tomar los posibles pares de valores y tabular el número de veces que ocurre cada uno de ellos, en la forma siguiente:

Valores	Frecuencias
(2,2)	2
(2,3)	2
(2,4)	1
(3,2)	2
(3,3)	3
(3,4)	2
...	...

Sin embargo, esta forma de organizar los resultados, no pone de manifiesto la relación entre ambas variables que es lo más importante en estas distribuciones.

Buscando resaltar esta relación, adoptaremos una tabulación en forma de tabla de doble entrada

	2	3	4	5	6
2	2	2	1	-	-
3	2	3	2	1	-
4	1	2	2	2	-
5	-	1	1	3	-
6	-	1	2	2	1
7	-	1	1	2	2

Donde los números que inician las filas representan los valores de la variable longitud de la lista. Las cifras que encabezan las columnas, indican el valor de la variable tiempo de respuesta y los números interiores son las frecuencias, es decir, el número de veces que se presenta cada par.

Distribuciones marginales

Si en la distribución anterior, contabilizamos cuantos ensayos se han hecho con listas de dos elementos, obtendremos que han sido 5. Los ensayos con listas de 3 elementos, han sido 8. Si continuamos haciendo este recuento obtendremos la siguiente distribución:

Nº de elementos	Nº de ensayos
2	5
3	8
4	7
5	5
6	6
7	6
Total	37

Esta distribución unidimensional, es la distribución de la variable número de elementos de la lista, independientemente de los valores que tome la variable tiempo de respuesta y se denomina distribución marginal de dicha variable, como puede verse esta distribución se obtiene totalizando por filas en la tabla de la distribución bidimensional:

	2	3	4	5	6	Totales
2	2	2	1	-	-	5
3	2	3	2	1	-	8
4	1	2	2	2	-	7
5	-	1	1	3	-	5
6	-	1	2	2	1	6
7	-	1	1	2	2	6

Si ahora, en esta distribución unidimensional, calculamos la media y la varianza obtendremos las denominadas media y varianza marginal, de la variable número de ensayos:

Nº de elementos	Nº de ensayos	$x_i \cdot n_i$	$x_i^2 \cdot n_i$
2	5	10	20
3	8	24	72
4	7	28	112
5	5	25	125
6	6	36	216
7	6	42	294
Total	37	165	839

La media será $165/37 = 4,46$

y la varianza $839/37 - 4,46^2 = 2,78$

De manera análoga, si contamos el número de ensayos en que el tiempo de respuesta fue de dos décimas, tendremos que han sido 5. El número de ensayos en que el tiempo de respuesta fue tres décimas, es 10. Haciendo este cálculo, para todos los valores de la variable tiempo de respuesta, obtendríamos su distribución marginal que será:

T. respuesta	Nº ensayos
2	5
3	10
4	9
5	10
6	3
Total	37

Esta distribución puede obtenerse en la tabla bidimensional totalizando por columnas:

	2	3	4	5	6
2	2	2	1	-	-
3	2	3	2	1	-
4	1	2	2	2	-
5	-	1	1	3	-
6	-	1	2	2	1
7	-	1	1	2	2
Totales	5	10	9	10	3

Los cálculos de media y varianza marginal del tiempo de respuesta serán:

T. de respuesta	Nº de ensayos	$y_j \cdot n_j$	$y_j^2 \cdot n_j$
2	5	10	20
3	10	30	90
4	9	36	144
5	10	50	250
6	3	18	108
Total	37	144	612

La media es $144/37 = 3,89$

y la varianza $612/37 - 3,89^2 = 1,41$

Distribuciones condicionadas

Las distribuciones condicionadas describen el comportamiento de una variable en la subpoblación, o submuestra, determinada por todos los elementos que tienen un mismo valor en la otra variable. Así si consideramos la distribución del tiempo de respuesta en los ensayos cuya lista tiene dos elementos, tendremos el siguiente subconjunto de observaciones:

(2,2) (2,3) (2,3) (2,4) (2,2)

tabulando estos datos obtenemos la siguiente tabla estadística:

T. respuesta	Nº ensayos
2	2
3	2
4	1
5	-
6	-
Total	5

Como puede observarse, esta distribución de frecuencias es idéntica a la que aparece en la primera fila de la tabla bidimensional, por ello para calcular las siete distribuciones condicionadas del tiempo de respuesta, una por cada valor de la variable longitud de la lista, no es necesario volver a tabular los datos, sino que basta con considerar las frecuencias que aparecen en la fila correspondiente al valor de la longitud de la lista que se desee.

Análogamente, si quisiéramos conocer la distribución de la variable longitud de la lista, en aquellas pruebas en que el tiempo de respuesta ha sido de 4 décimas, tendríamos que estudiar el siguiente subconjunto de ensayos:

(2,4) (3,4) (3,4) (4,4) (4,4) (5,4) (6,4) (6,4) (7,4)

Nº de elementos	Nº de ensayos
2	1
3	2
4	2
5	1
6	2
7	1
Total	9

Como vemos, esta distribución de frecuencias se corresponde con la columna encabezada por el valor 4, en la tabla bidimensional. Esto nos indica que las diferentes distribuciones condicionadas de esta variable vienen dadas por las correspondientes columnas de la tabla.

Las medias y varianzas de estas distribuciones reciben el nombre de medias y varianzas condicionadas. Por ejemplo la media y la varianza del tiempo de respuesta en los ensayos con listas de longitud dos serán:

T. de respuesta	Nº de ensayos	$y_j \cdot n_j$	$y_j^2 \cdot n_j$
2	2	4	8
3	2	6	18
4	1	4	16
5	-	-	-
6	-	-	-
Total	5	14	42

La media es $14/5 = 2,8$ y la varianza $42/5 - 2,8^2 = 0,56$

En general, las medias y varianzas condicionadas diferirán de una distribución a otra, por ejemplo las medias del tiempo de reacción condicionadas por las distintas longitudes de la lista son:

$$\bar{y}_2 = 2,8 \quad \bar{y}_3 = 3,25 \quad \bar{y}_4 = 3,71 \quad \bar{y}_5 = 4,4 \quad \bar{y}_6 = 4,5 \quad \bar{y}_7 = 4,83$$

Esto nos indica como cambia el tiempo medio de reacción en función de la longitud de la lista, sobre esta idea volveremos en el capítulo próximo cuando hablemos de regresión.

Relaciones Estadísticas

Si en una distribución bidimensional a cada valor, o modalidad, de una variable le corresponde un único valor, o modalidad, de la otra variable, se dice que la segunda variable depende funcionalmente de la primera. Este tipo de relación puede expresarse mediante una función:

$$y = f(x)$$

Por ejemplo: En un test de 100 preguntas de verdadero o falso, en que haya obligación de contestar a todas las preguntas, la puntuación obtenida depende

funcionalmente del número de aciertos, ya que a un cierto número de aciertos le corresponde una y solo una puntuación. Esta relación podría expresarse por medio de la siguiente función:

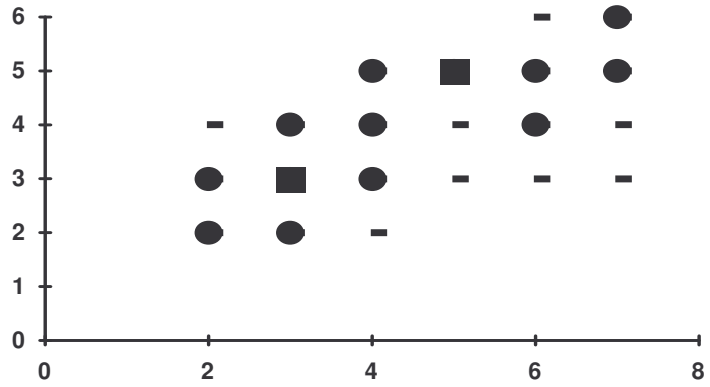
$$y = 2x - 100$$

Las distribuciones bidimensionales en las que existe dependencia funcional se reconocen fácilmente, ya que si la segunda variable depende funcionalmente de la primera, en cada fila aparecerá una sola frecuencia distinta de cero.

La situación opuesta es la independencia de las variables. En este caso los valores que toma una variable no guardan ninguna relación con los valores de la otra variable. Esta situación se refleja en que todas las distribuciones condicionadas de una variable, son iguales entre si e iguales a su distribución marginal.

En Psicología, ya sea por la falta de control o por una variabilidad intrínseca, las relaciones entre variables incluyen un término aleatorio que hace que para cada valor de la variable, puedan aparecer distintos valores de la otra variable, pero también es cierto que no todos los valores serán posibles, ni se darán con la misma frecuencia.

Por ejemplo, en la tabla de la distribución bidimensional de las variables, longitud de la lista y tiempo de respuesta, las mayores frecuencias tienden a presentarse alrededor de la diagonal principal, indicando con ello la tendencia a que aumente el tiempo de respuesta, cuando aumenta la longitud de la lista. Esta situación puede apreciarse claramente, si representamos gráficamente el conjunto de resultados de los ensayos:



donde, en abscisas se ha representado la longitud de la lista, en ordenadas los tiempos de respuesta y el grosor del trazo indica la frecuencia.

Este tipo de relaciones son las que denominamos relaciones estadísticas. En ellas una variable está determinada en parte por los valores de la otra variable, pero manteniendo un cierto grado de aleatoriedad. Nuestra primera tarea será medir la relación existente entre ambas variables, como expresión del grado en que una variable fija el valor de la otra.

Covarianza

Un indicador del grado de relación lineal que existe entre dos variables, es la covarianza, que denotaremos por S_{XY} y que se define por la expresión:

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

Si los datos están tabulados, cada par de valores (x_i, y_j) se presenta un cierto número de veces que indica su frecuencia n_{ij} y en consecuencia la expresión anterior se transforma en:

$$S_{XY} = \frac{\sum_{j=1}^{h}\sum_{i=1}^{k}(x_i - \bar{x})(y_j - \bar{y}) \cdot n_{ij}}{n}$$

Esta expresión es un indicador del grado de relación lineal de ambas variables y del sentido de la misma, porque si ambas variables tienen una relación directa, a valores altos de X corresponderán valores altos de Y y a valores bajos de X valores bajos de Y, en consecuencia las diferencias a la media serán grandes y del mismo signo, con lo cual la covarianza tomará un valor alto y positivo. Si la relación es inversa, a valores altos de X corresponderán valores bajos de Y y viceversa, por ello las diferencias a las respectivas medias serán grandes, pero de distinto signo y la covarianza será negativa, aunque su valor absoluto será alto. En el caso de que exista poca relación entre las variables, las diferencias serán aleatoriamente positivas y negativas y tenderán a compensarse con lo cual la covarianza tendrá un valor pequeño en términos absolutos.

Al igual que ocurría en el caso de la varianza, la formula anterior sirve para definir la covarianza, pero es incómoda para realizar los cálculos pues en general las diferencias a las medias tendrán fracciones decimales. Se prefiere por ello utilizar las siguientes expresiones:

$$S_{XY} = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y}$$

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

Coeficiente de correlación de Pearson

La covarianza no es un buen indicador de la correlación, o grado de relación, lineal entre dos variables pues no está normalizada, es decir no hay un valor que sea el máximo alcanzable, ni tampoco un mínimo, por lo que su interpretación es difícil. También sucede que la covarianza depende de las unidades empleadas en las variables, si cambiamos de escala una de las variables la covarianza cambiará de valor, sin que se haya alterado el grado de relación entre las variables. Por ello, se prefiere como indicador del grado de relación lineal entre dos variables, el coeficiente de correlación lineal, o coeficiente de correlación de Pearson, que es simplemente la covarianza entre dos variables, medidas en unidades de sus respectivas desviaciones típicas, o más brevemente, la covarianza de las puntuaciones típicas. La expresión será por tanto:

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

Si sustituimos la covarianza y las desviaciones típicas, por las expresiones empleadas para su cálculo y multiplicamos numerador y denominador por n, tendremos:

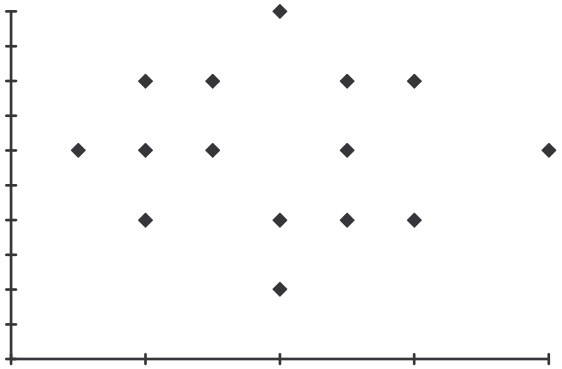
$$r_{XY} = \frac{\sum_{i=1}^{n} x_i \cdot y_i - n\overline{xy}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\ \sqrt{\sum_i y_i^2 - n\bar{y}^2}}$$

que es la fórmula habitualmente indicada en los textos, para su cálculo.

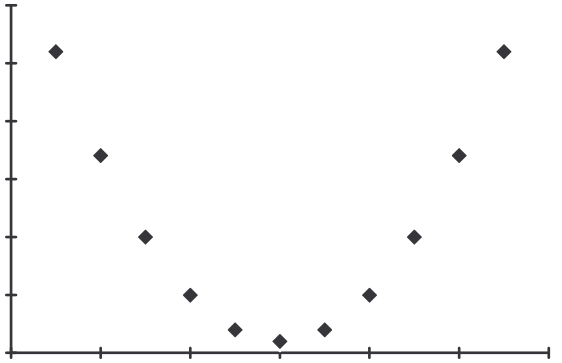
El coeficiente de correlación de Pearson es un número sin dimensiones, comprendido entre -1 y 1. En razón de su propia definición, goza de la propiedad de ser invariante frente a cambios de escala y origen, por ello el valor obtenido es el mismo si se calcula a partir de las puntuaciones iniciales, de las puntuaciones diferenciales o de las puntuaciones típicas, o de cualquier otra puntuación obtenida por transformación lineal de las originales.

El coeficiente de correlación nos proporciona dos tipos de información, su signo nos indica el sentido de la relación. Si el coeficiente es positivo indica una relación directa, al aumentar una variable aumenta la otra y viceversa al disminuir una variable disminuye la otra. Por el contrario, si el coeficiente es negativo estamos frente a una relación inversa, donde al aumentar una variable disminuye la otra. El valor absoluto del coeficiente de Pearson nos señala el grado de relación, cuando su valor es 1, estamos en un caso de relación perfecta o relación funcional, esto quiere decir que si representásemos los pares de valores X e Y obtendríamos una linea recta perfecta.

Cuando el valor es cero, diremos que las variables están incorreladas, esto nos indica la ausencia de relación lineal entre las dos variables, bien porque las variables son independientes, lo que daría origen a un grafico del tipo siguiente:



o bien porque la relación existente entre ellas es de otro tipo, cuadrática, etc. en cuyo caso la grafiva que obtendriamos sería:



Ejemplo: En una muestra de 15 estudiantes se ha pasado un test de inteligencia, cuyas puntuaciones se reflejan en la variable X, y se ha determinado su rendimiento en una prueba, las puntuaciones de esta última se recogen en la variable Y. La tabla siguiente nos proporciona los valores de estas variables, así como los cálculos básicos necesarios para determinar el valor del coeficiente de correlación de Pearson:

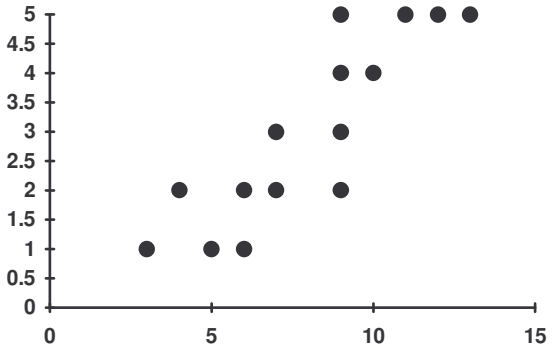
Sujeto	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
1	9	5	45	81	25
2	12	5	60	144	25
3	6	1	6	36	1
4	9	4	36	81	16
5	7	2	14	49	4
6	9	2	18	81	4
7	5	1	5	25	1
8	9	3	27	81	9
9	7	3	21	49	9
10	3	1	3	9	1
11	10	4	40	100	16
12	6	2	12	36	4
13	11	5	55	121	25
14	4	2	8	16	4
15	13	5	65	169	25
Totales	120	45	415	1078	169

La media de X es 120/15 = 8
La media de Y es 45/15 = 3
La covarianza es 415/15 - 24 = 3,67
La varianza de X es 1078/15 - 64 = 7,87 y la desviación típica 2,8
La varianza de Y es 169/15 - 9 = 2,27 y la desviación típica 1,5

Por consiguiente el coeficiente de correlación es:

$$r = \frac{3,67}{2,8 \cdot 1,5} = 0,87$$

Esto nos indica que la relación entre inteligencia y rendimiento, en esa prueba, es directa, como cabría esperar, y bastante alta. Podemos observar ambas cuestiones en la representación gráfica:



Hemos de señalar que la existencia de correlación, aunque esta sea alta, no implica necesariamente la existencia de una relación causal. En principio ambas cosas son de naturaleza diferente, la correlación es una medida estadística que indica el grado de variación conjunta, covariación, de dos variables, mientras que la idea de relación causal es de orden metafísico. Cierto es que si existe una relación causal entre dos variables, esto dará origen a una correlación alta entre ambas, pero la recíproca no es cierta, puede existir una correlación alta entre dos variables sin que exista relación causal entre ellas. La explicación, más habitual, para situaciones de este tipo es la existencia de una tercera variable, de la cual dependen las otras dos y que da origen a esa variación conjunta de los valores.