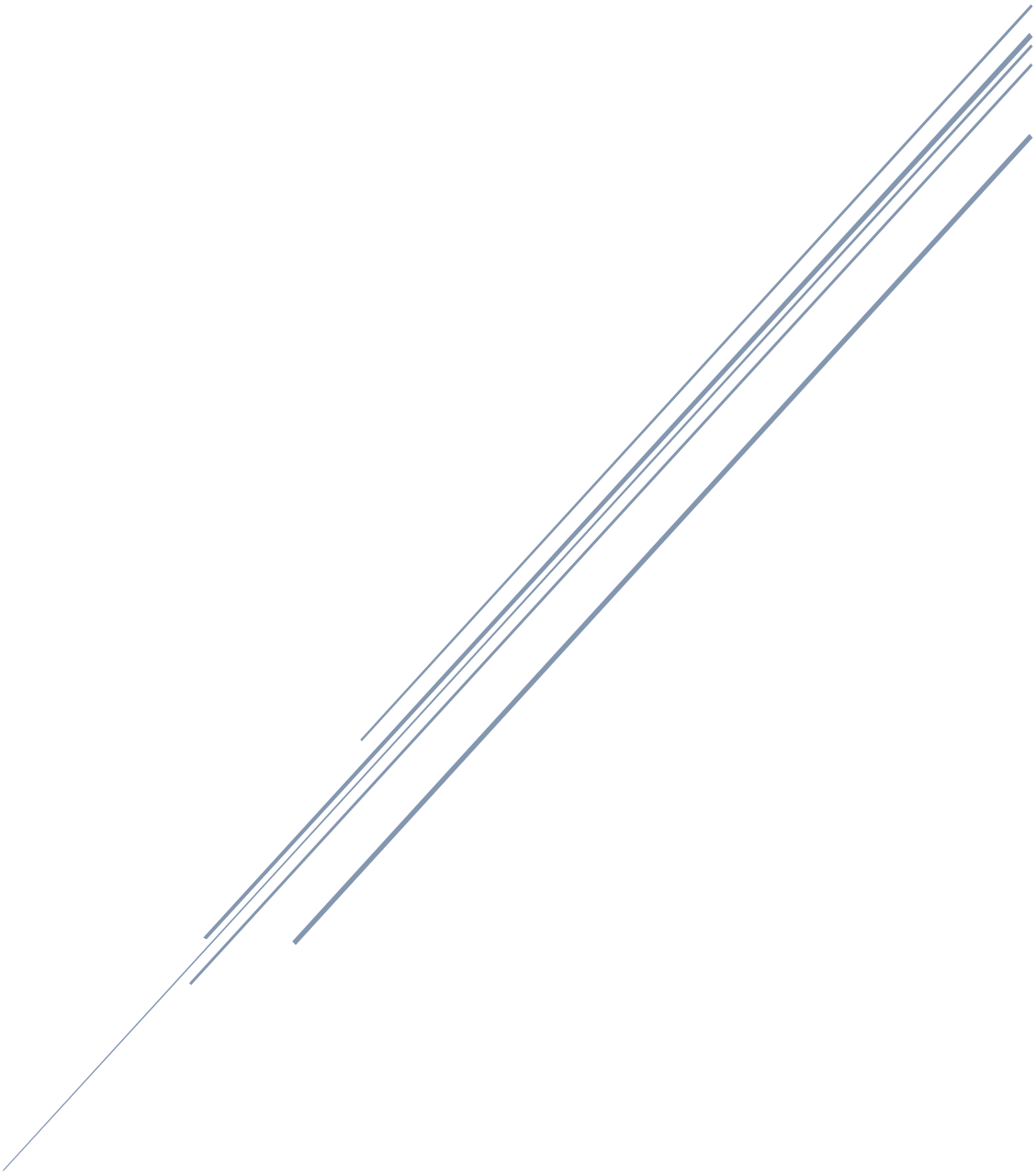
TEMA 6. TABULACIÓN DE DATOS BIVARIANTES

Curso OCW de “Estadística Descriptiva con Excel para Grados de Ciencias Sociales”



Grupo Decisión Multicriterio Zaragoza (GDMZ)

1. INTRODUCCIÓN

En los temas anteriores hemos abordado el tratamiento estadístico de una única variable. Pero cuando afrontamos el estudio de una población (por ejemplo, la situación socioeconómica de las familias de una ciudad), lo habitual es distintas características de sus individuos (tamaño familiar, gastos e ingresos, nivel de instrucción, etc.). Ello no sólo proporciona una visión más enriquecedora de la realidad, sino que permite estudiar si las distintas variables se influyen mutuamente ("¿afectan los ingresos de la familia a la estructura del gasto de la misma?"). Comenzamos presentando la notación necesaria para el análisis conjunto de dos variables.

2. DISTRIBUCIÓN DE FRECUENCIAS CONJUNTA

Cuando observamos conjuntamente dos variables *X* e *Y* sobre una misma población de tamaño *N*, estamos ante una variable estadística bidimensional que representamos por *(X, Y)*. Supongamos que la variable *X* toma *k* valores distintos x_i (*i*=1,..,*k*) y que la variable *Y* toma *h* valores distintos y_j (*j* =1,..,*h*). Se denomina **frecuencia absoluta conjunta**, y se denota por n_{ij}, al número de veces que se observa el par (x_i, y_j) y se verifica que:

$$\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} = N$$

Se denomina **frecuencia relativa conjunta**, y se denota por f_{ij}, a la proporción de veces sobre *N* que se presenta conjuntamente el par (x_i, y_j), es decir:

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

verificándose que $\sum_{i=1}^{k}\sum_{j=1}^{h} f_{ij} = 1$

Se denomina **distribución de frecuencias conjunta** a la terna:

$$\left(x_i, y_j, n_{ij}\right) \quad i = 1 \ldots k; j = 1 \ldots h$$

La representación numérica de los datos de una distribución de frecuencias conjunta se realiza mediante una tabla de doble entrada como la siguiente:

X \ Y	y_1	y_2	…	y_j	…	y_h
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1h}
…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ih}
…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kh}

En la primera columna aparecen los valores de la variable *Y* y en la primera fila los de la variable *X*. La intersección de la fila *i*-ésima con la columna *j*-ésima contiene la frecuencia absoluta conjunta n_{ij} del par (x_i, y_j). De forma análoga construimos la **distribución de frecuencias relativas conjunta**:

X \ Y	y_1	y_2	…	y_j	…	y_h
x_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1h}
…	…	…	…	…	…	…
x_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ih}
…	…	…	…	…	…	…
x_k	f_{k1}	f_{k2}	…	f_{kj}	…	f_{kh}

Si las variables son cualitativas, estas tablas se denominan tablas de contingencia.

Ejemplo:

La siguiente tabla recoge la distribución de frecuencias conjunta de las variables *X*: "*Saldos de caja (en €)*" e *Y*: "*Saldos de bancos (en €)*" de una empresa al final de la semana durante 66 semanas:

X \ Y	20.000 – 80.000	80.000 – 220.000
0 – 1.000	6	20
1.000 – 5.000	10	10
5.000 – 10.000	14	6

Y la tabla que muestra la distribución de frecuencias relativas conjunta es:

X \ Y	20.000 – 80.000	80.000 – 220.000
0 – 1.000	9,09%	30,30%
1.000 – 5.000	15,15%	15,15%
5.000 – 10.000	21,21%	9,09%

3. DISTRIBUCIONES MARGINALES

A partir de la distribución de frecuencias conjunta de dos variables se pueden obtener las distribuciones unidimensionales de cada una de las variables por separado. Estas distribuciones reciben el nombre de **marginales** y se obtienen considerando los valores que toma una de las variables con sus respectivas frecuencias, independientemente de los valores de la otra variable. Por tanto, las

distribuciones marginales se pueden caracterizar por todas las medidas de posición, dispersión y forma estudiadas en los temas anteriores.

X \ Y	y_1	y_2	...	y_j	...	y_h	$n_{i.}$
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1h}	$n_{1.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ih}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kh}	$n_{k.}$
$n_{.j}$	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.h}$	N

La distribución marginal de *X* se obtiene considerando sus valores, así como sus respectivas frecuencias, independientemente de los valores de la variable *Y*. Es decir, la distribución marginal de *X* viene dada por $(x_i, n_{i.})\ i = 1 \dots k$ donde $n_{i.}$ es la frecuencia absoluta marginal de *X* y se obtiene como suma de la *i*-ésima fila de frecuencias absolutas conjuntas:

$$n_{i.} = \sum_{j=1}^{h} n_{ij}$$

verificándose que $\sum_{i=1}^{k} n_{i.} = N$

En términos de frecuencias relativas la distribución marginal de X viene dada por $(x_i, f_{i.}), i = 1 \dots k$, donde $f_{i.}$es la **frecuencia relativa marginal de *X***:

$$f_{i.} = \frac{n_{i.}}{N}$$

y se verifica que $\sum_{i=1}^{k} f_{i.} = 1$

Análogamente, la distribución marginal de *Y* se obtiene considerando sus valores, así como sus respectivas frecuencias independientemente de los valores de la variable *X*. Es decir, la distribución marginal de *Y* viene dada por $\left(y_j, n_{j.}\right) j = 1 \dots h$, donde $n_{.j}$ es la frecuencia absoluta marginal de *Y* y se calcula como la suma de la *j*-ésima columna de frecuencias absolutas conjuntas:

$$n_{.j} = \sum_{i=1}^{k} n_{ij}$$

y se verifica que $\sum_{j=1}^{h} n_{.j} = N$

En términos de frecuencias relativas la distribución marginal de Y viene dada por $(y_j, f_{j.})\ j = 1 \ldots h$, donde $f_{.j}$ es la **frecuencia relativa marginal de Y**:

$$f_{.j} = \frac{n_{.j}}{N}$$

y se verifica que $\sum_{i=1}^{h} f_{.j} = 1$

Observación: A partir de una distribución conjunta se pueden obtener las distribuciones marginales; sin embargo, a partir de las marginales no se puede deducir la conjunta.

Ejemplo:

A partir de la distribución de frecuencias conjunta del ejemplo de los saldos de cajas y bancos se calculan las dos distribuciones marginales:

X \ Y	**20.000 - 80.000**	**80.000 – 220.000**	$n_{i.}$
0 – 1.000	6	20	26
1.000 – 5.000	10	10	20
5.000 – 10.000	14	6	20
$n_{.j}$	30	36	66

X	$n_{i.}$	$f_{i.}$
0 – 1.000	26	0,394
1.000 – 5.000	20	0,303
5.000 – 10.000	20	0,303
Total	66	1

Y	$n_{.j}$	$f_{.j}$
20.000 - 80.000	30	0,4545
80.000 – 220.000	36	0,5455
Total	66	1

4. DISTRIBUCIONES CONDICIONADAS

Otra distribución unidimensional que se pueden definir a partir de la distribución de frecuencias conjunta es aquellas que se construye definiendo previamente una condición. En este sentido, presentamos la distribución de una de las variables condicionada a que la otra variable tome un determinado valor y se denomina **distribución condicionada**. Al igual que las marginales, son distribuciones unidimensionales a las que se aplican las técnicas estadísticas estudiadas en los temas anteriores.

Para una distribución bidimensional $(x_i, y_j, n_{ij})\quad i = 1 \dots k; j = 1 \dots h,$ la **distribución de X condicionada a y_j** viene dada por:

$$(x_i, n_{ij})\quad i = 1 \dots k$$

o bien, en términos de frecuencias relativas condicionadas, por

$$\left(x_i, f_{X=x_i|Y=y_j}\right)\quad i = 1 \dots k$$

donde la **frecuencia relativa condicionada del valor x_i** es el cociente entre el número de veces que se presenta el par (x_i, y_j), es decir, n_{ij}, y el número de veces que aparece el valor y_j independientemente de los valores de X con los que aparece, es decir, $n_{.j}$. Por tanto:

$$f_{i|j} = f_{X=x_i|Y=y_j} = \frac{n_{ij}}{n_{.j}} \quad \forall\, i = 1, \dots, k$$

y se verifica que:

$$\sum_{i=1}^{k} f_{X=x_i|Y=y_j} = \sum_{i=1}^{k} \frac{n_{ij}}{n_{.j}} = \frac{\sum_{i=1}^{k} n_{ij}}{n_{.j}} = \frac{n_{.j}}{n_{.j}} = 1$$

Análogamente, la **distribución de Y condicionada a x_i** viene dada por:

$$(y_j, n_{ij})\quad \text{j} = 1 \dots h$$

o bien en términos de frecuencias relativas condicionadas por

$$\left(y_j, f_{Y=y_j|X=x_i}\right) \quad j = 1 \ldots h$$

donde la **frecuencia relativa condicionada del valor y_j** es el cociente entre el número de veces que se presenta el par (x_i, y_j), es decir, n_{ij}, y el número de veces que aparece el valor x_i independientemente de los valores de *Y*, es decir, $n_{i.}$ Por tanto:

$$f_{j|i} = f_{Y=y_j|X=x_i} = \frac{n_{ij}}{n_{i.}} \quad \forall\, j = 1, \ldots, h$$

y se verifica que:

$$\sum_{j=1}^{h} f_{Y=y_j|X=x_i} = \sum_{j=1}^{h} \frac{n_{ij}}{n_{i.}} = \frac{\sum_{j=1}^{h} n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1$$

$$\sum_{j=1}^{h} f_{Y=y_j|X=x_i} = \sum_{j=1}^{h} \frac{n_{ij}}{n_{i.}} = \frac{\sum_{j=1}^{h} n_{ij}}{n_{i.}} = \frac{n_{i.}}{n_{i.}} = 1$$

Ejemplo:

En el ejemplo de los saldos de cajas y bancos, las distribuciones de X: "*Saldos de caja (en €)*" condicionadas por los valores de Y: "*Saldos de bancos (en €)*" son:

X/YÎ[20.000,80.000]	n_{i1}	$f_{X=x_i\|Y=y_1}$
0 – 1.000	6	0,200
1.000 – 5.000	10	0,333
5.000 – 10.000	14	0,467
Total	20	1

X/YÎ[80.000,220.000]	n_{i2}	$f_{X=x_i\|Y=y_2}$
0 – 1.000	20	0,556
1.000 – 5.000	10	0,278
5.000 – 10.000	6	0,167
Total	36	1

Análogamente, las distribuciones de Y: "*Saldos de bancos (en €)*" condicionadas por los valores de X: "*Saldos de caja (en €)*" son:

Y/XÎ[0,1.000]	n_{1j}	$f_{Y=y_j\|X=x_1}$
20.000 - 80.000	6	0,231
80.000 – 220.000	20	0,769
Total	26	1

Y/XÎ[1.000,5.000]	n_{2j}	$f_{Y=y_j\|X=x_2}$
20.000 - 80.000	10	0,500
80.000 – 220.000	10	0,500
Total	20	1

Y/XÎ[5.000,10.000]	n_{3j}	$f_{Y=y_j\|X=x_3}$
20.000 - 80.000	14	0,700
80.000 – 220.000	6	0,300
Total	20	1

5. INDEPENDENCIA

Dos variables son independientes entre sí, si los valores que toma una de ellas no están afectados por los valores que toma la otra, lo que supone que las distribuciones condicionadas son idénticas a la distribución marginal correspondiente. En términos de frecuencias relativas se expresa del siguiente modo:

$$f_{i|j} = f_{i.} \quad \forall\ j = 1 \dots h$$

$$f_{j|i} = f_{.j} \quad \forall\ i = 1 \dots k$$

Como consecuencia, se puede enunciar el criterio de Independencia Estadística diciendo que dos variables son estadísticamente independientes si la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales para todos los valores de ambas variables, es decir, $f_{ij} = f_{i.} \times f_{.j} \quad \forall i = 1 \dots k,\ \forall\ j = 1 \dots h.$

Ejemplo:

Se han recogido los datos acerca de las variables *X*= '*Estatura (en cm)*' e *Y* = '*Sueldo mensual (en euros)*' para 15 personas. Los valores obtenidos se muestran en la siguiente tabla de doble entrada:

X \ Y	[800, 1100]	[1100, 1800]
[150, 170]	2	3
[170, 185]	4	6

Las variables X e Y son estadísticamente independientes porque se cumple que $f_{ij} = f_{i.} \times f_{.j} \quad \forall i = 1 \dots k,\ \forall\ j = 1 \dots h.$ Veámoslo:

En primer lugar, calculamos las frecuencias marginales:

X \ Y	[800, 1100]	[1100, 1800]	$n_{i.}$
[150, 170]	2	3	5
[170, 185]	4	6	10
$n_{.j}$	6	9	15

A continuación, comprobamos la condición para cada posible combinación de valores de las dos variables.

$$i = 1, j = 1 \quad f_{11} = \frac{2}{15} \quad f_{1.} \times f_{.1} = \frac{5}{15} \times \frac{6}{15} = \frac{2}{15}$$

$$i = 1, j = 2 \quad f_{12} = \frac{3}{15} \quad f_{1.} \times f_{.2} = \frac{5}{15} \times \frac{9}{15} = \frac{3}{15}$$

$$i = 2, j = 1 \quad f_{21} = \frac{4}{15} \quad f_{2.} \times f_{.1} = \frac{10}{15} \times \frac{6}{15} = \frac{4}{15}$$

$$i = 2, j = 2 \quad f_{22} = \frac{6}{15} \quad f_{2.} \times f_{.2} = \frac{10}{15} \times \frac{9}{15} = \frac{6}{15}$$

Como se ve, queda comprobado que las variables X e Y son estadísticamente independientes.

Tipos de dependencia

Existen diversos grados posibles de dependencia entre dos variables X e Y. Por un lado, existe la llamada **dependencia funcional** en la que existe una relación determinística entre las variables, es decir, se puede expresar una de las variables como una función matemática exacta de la otra y, por tanto, se pueden determinar unívocamente los valores de una conocidos los de la otra.

En el otro extremo se sitúa la **independencia** en el que no existe ningún tipo de relación entre las variables X e Y. Finalmente, entre estos dos casos extremos nos podemos encontrar con muchas situaciones intermedias, caracterizadas porque existe una cierta relación entre las variables que no puede ser expresada de forma exacta mediante una función matemática pero tampoco se puede decir que no existe relación alguna entre ellas. En este caso se dice que existe una **dependencia estadística** entre ellas. Este tipo de dependencia admite diversos grados de intensidad reflejando la existencia de una asociación más o menos fuerte entre las variables. Estas cuestiones serán estudiadas en el tema siguiente.