T. 7 – Organización y representación gráfica de datos multivariados

• Tras abordar en temas previos el tratamiento individualizado (univariado) de las variables, en éste y temas sucesivos se describen una serie de procedimientos asociados al tratamiento conjunto de dos o más variables, los cuales van a permitir extraer diversas facetas de la información compartida por esas variables. En bastantes momentos se va a ceñir esta exposición al caso bivariado (dos variables) por ser más sencillo en su presentación y por tratarse, con frecuencia, del caso particular más simple del modo general de abordar el problema a nivel multivariado (dos o más variables).

1. La distribución conjunta multivariada

• De modo análogo al caso univariado, un resumen básico de la información de un grupo de 2 o más variables consiste en la distribución conjunta de frecuencias de las mismas, la cual se basa en el conteo del número de casos (frecuencias) que presentan las distintas combinaciones de valores que a nivel empírico se hayan dado para esas variables. Las modalidades (X_i ,Y_i ,Z_i ...) de una distribución conjunta representarán, no a los valores de una variable concreta, sino a todas las posibles combinaciones de los valores de las variables que se consideren –excepto aquellas

combinaciones que no se hayan presentado a nivel empírico y que por tanto no tiene sentido incluir en la distribución de frecuencias.

Ejemplo: La siguiente tabla de datos procede de un estudio sobre las relaciones de pareja en que se obtuvo información en una muestra de 71 sujetos de las 3 variables siguientes: sexo (1: hombre; 2: mujer); nº de parejas estables a lo largo de los últimos 5 años; y situación emocional actual (1: satisfactoria; 2: ni satisfactoria ni insatisfactoria; 3: insatisfactoria).

ID	Sexo	Num_parejas	Sit_actual
1	1	1	3
2	1	4	2
3	2	1	1
4	2	2	1
5	2	1	3
6	1	0	1
7	2	3	2
...	...	...	...
71	1	1	1

La organización de los datos de la anterior tabla en forma de distribución conjunta de frecuencias (absolutas) quedaría como sigue, donde X = Sexo; Y = Num_parejas y Z = Sit_actual:

X_i, Y_i, Z_i	n_i
1,0,1	4
1,0,2	3
1,0,3	2
1,1,1	12
1,1,2	8
1,1,3	6
1,2,1	5
1,2,2	1
1,2,3	2
1,4,2	1
2,0,1	6
...	...
2,3,2	1
	71

• La distribución conjunta de frecuencias relativas o proporciones (p_i) y la de porcentajes ($\%_i$) pueden obtenerse a partir de las frecuencias absolutas dividiendo cada frecuencia absoluta entre el nº de casos (n) y multiplicando las frecuencias relativas por cien, respectivamente.

• El ordenamiento de las modalidades en una distribución conjunta de frecuencias carece de sentido, si bien, se suelen situar en orden alfabético/numérico creciente a fin de poder localizar más fácilmente cualquier combinación de valores de las variables.

• La obtención de las frecuencias acumuladas, ya sean absolutas, relativas o porcentajes, carece también aquí de sentido dado que las modalidades de la distribución no representan un continuo -al igual que ocurría con las distribuciones de frecuencias de las variables categóricas. Por ello mismo, los índices de posición y de dispersión descritos para las variables categóricas podrían ser aplicados también en una distribución conjunta de frecuencias.

• Inconvenientes: Si el número de variables es amplio o si alguna de las variables tiene muchos valores, el número de combinaciones de valores posibles puede llegar a ser muy numeroso, tanto que la visualización de la distribución de frecuencias resulte poco ventajosa en su propósito de resumir la información de los datos. Existen algunas alternativas que pueden ayudar a resolver este problema en algunas situaciones:

(1) En el caso de una variable (o más) con muchos posibles valores (como es lo más habitual con variables cuantitativas), una opción es colapsar esos valores en intervalos. De este modo se pierde en precisión de la información, si bien, se hace factible la reducción drástica el número de combinaciones de valores posibles.

Por **ejemplo**, supongamos que tenemos dos variables, cada una con los tiempos (seg.) empleados en ejecutar dos tareas de aptitudes mecánicas por un mismo grupo de sujetos. Suponiendo un rango de valores en ambas variables de entre 0 y 20 seg., una posible agrupación de los mismos a la hora de crear una distribución conjunta de frecuencias podría quedar como sigue:
(¿Cuántas filas tendría la tabla si no se hubiese hecho esta agrupación de valores en intervalos?)

X_i (seg.), Y_i (seg.)	n_i
0-5 ∩ 0-5	...
0-5 ∩ 5-10	...
0-5 ∩ 10-15	...
0-5 ∩ 15-20	...
5-10 ∩ 0-5	...
5-10 ∩ 5-10	...
...	...
15-20 ∩ 15-20	...

(2) En el caso de muchas variables, una alternativa consiste en aplicar alguno de los métodos estadísticos que se suelen englobar bajo el calificativo de "técnicas de reducción de datos" (por ejemplo, el análisis factorial, el escalamiento multidimensional o el análisis de correspondencias), métodos que escapan a los contenidos de la presente asignatura.

1.1. La tabla de contingencia

• En el caso de dos variables, una forma muy conveniente de visualizar la distribución conjunta de ambas es en forma de tabla de contingencia, esto es, una tabla de doble entrada en que cada lado de la tabla contiene las modalidades de una variable. En las casillas interiores de la tabla aparecen la frecuencias conjuntas (ya sean absolutas, relativas o porcentajes) de la combinación de los valores fila y columna correspondientes.

• Ejemplo: se llevó a cabo un estudio para evaluar si el estado de ánimo de los mayores de 65 años podía verse influido por el hecho de vivir en una residencia geriátrica o no. Se recogieron datos de una muestra de 500 personas de las variables "Estado de ánimo": negativo (−); neutro (±); positivo (+) y "Vivir en residencia" (Sí; No). La distribución conjunta de frecuencias es la siguiente:

	−	±	+
Sí	48	42	60
No	70	105	175

¿Cómo se ha construido esa tabla de contingencia? Realizando, a partir de la matriz de datos original, un recuento del nº de casos que presentan cada combinación de par de valores.

Caso	Residencia	Estado ánimo
1	S	−
2	N	±
3	S	−
4	S	+
...	...	...
500	N	±

• También a partir de esa tabla de datos, como ya se ha visto, es posible obtener:

- La distribución de cada variable por separado (= distribuciones marginales):

Residencia (X)

X_i	n_i	p_i
Sí	150	0.30
No	350	0.70
	500	1

Estado ánimo (Y)

Y_i	n_i	p_i
−	118	0.236
±	147	0.294
+	235	0.470
	500	1

- La distribución conjunta de ambas variables:

X_i, Y_i	n_i	p_i
Sí ∩ −	48	0.096
Sí ∩ ±	42	0.084
Sí ∩ +	60	0.120
No ∩ −	70	0.140
No ∩ ±	105	0.210
No ∩ +	175	0.350
	500	1

• En las tablas de contingencia es habitual incluir en los laterales las sumas de las celdas de filas y columnas => distribuciones marginales (= distribución de cada variable por separado)

	−	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	500

Ejemplo de la tabla de contingencia de las dos variables anteriores tal y como es obtenida con el programa SPSS:

*Tabla de contingencia Vivir residencia * Estado ánimo*

		Estado ánimo			Total
		Negativo	Neutro	Positivo	
Vivir residencia	Sí	48	42	60	150
	No	70	105	175	350
Total		118	147	235	500

• En las tablas de contingencia se pueden presentar también las frecuencias relativas o porcentajes:

p_{ij}	−	±	+	Total
Sí	0,096	0,084	0,120	0,300
No	0,140	0,210	0,350	0,700
Total	0,236	0,294	0,470	1

$\%_{ij}$	−	±	+	Total
Sí	9,6	8,4	12	30
No	14	21	35	70
Total	23,6	29,4	47	100

El siguiente "output" muestra cómo queda la tabla de contingencia anterior cuando es obtenida con SPSS en el caso de solicitar que en las casillas de la tabla aparezcan los porcentajes (las frecuencias relativas no es posible con SPSS):

Tabla de contingencia Vivir residencia * Estado ánimo

			Estado ánimo			Total
			Negativo	Neutro	Positivo	
Vivir residencia	Sí	Recuento	48	42	60	150
		% del total	9,6%	8,4%	12,0%	30,0%
	No	Recuento	70	105	175	350
		% del total	14,0%	21,0%	35,0%	70,0%
Total		Recuento	118	147	235	500
		% del total	23,6%	29,4%	47,0%	100,0%

• Las filas y columnas interiores (sin la columna y la fila de las distribuciones marginales) de una tabla de contingencia son referidas como distribuciones condicionales. Por ejemplo, la primera fila de nuestra tabla de ejemplo (48, 42, 60) es la distribución condicional de la variable "Estado de ánimo" para aquellos sujetos que viven en una residencia. La segunda fila (70, 105, 175) es la distribución condicional de la variable "Estado de ánimo" para aquellos sujetos que no viven en una residencia.

• La comparación de las distribuciones de una variable condicionales a los valores de la otra variable es fundamental a la hora de valorar si hay o no relación entre las 2 variables y este aspecto será tratado en el siguiente tema.

• A fin de designar los elementos de una tabla de contingencia a nivel simbólico:

- Las casillas interiores se representan como n_{ij} (frecuencias absolutas), p_{ij} (frecuencias relativas) o $\%_{ij}$ (porcentajes), donde i y j representan el nº de fila y el nº de columna.
- Las casillas de los márgenes derecho e inferior (distribuciones marginales) se representan como n_{i+} y n_{+j}, respectivamente. Si la tabla es de frecuencias relativas, p_{i+} y p_{+j}, y si de porcentajes, $\%_{i+}$ y $\%_{+j}$.

Ejemplo para una tabla de contingencia de 2x4 (nº de filas x nº de columnas) de frecuencias absolutas:

	Y_1	Y_2	Y_3	Y_4	n_{i+}
X_1	n_{11}	n_{12}	n_{13}	n_{14}	n_{1+}
X_2	n_{21}	n_{22}	n_{23}	n_{24}	n_{2+}
n_{+j}	n_{+1}	n_{+2}	n_{+3}	n_{+4}	N

Ejercicio 1: Asignar los valores correspondientes a cada una de las siguientes expresiones:

	−	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	500

n_{21} = n_{13} = p_{23} = n_{++} (n) =

n_{2+} = n_{+3} = p_{1+} = p_{+2} =

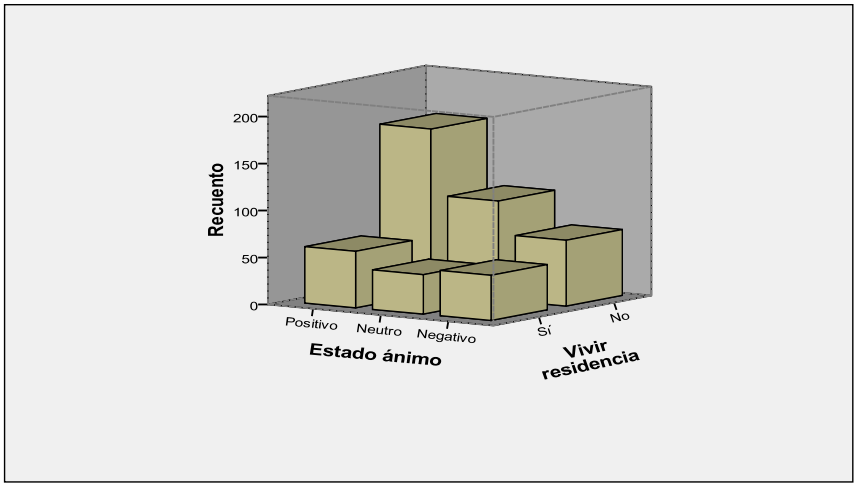
2. Representaciones gráficas

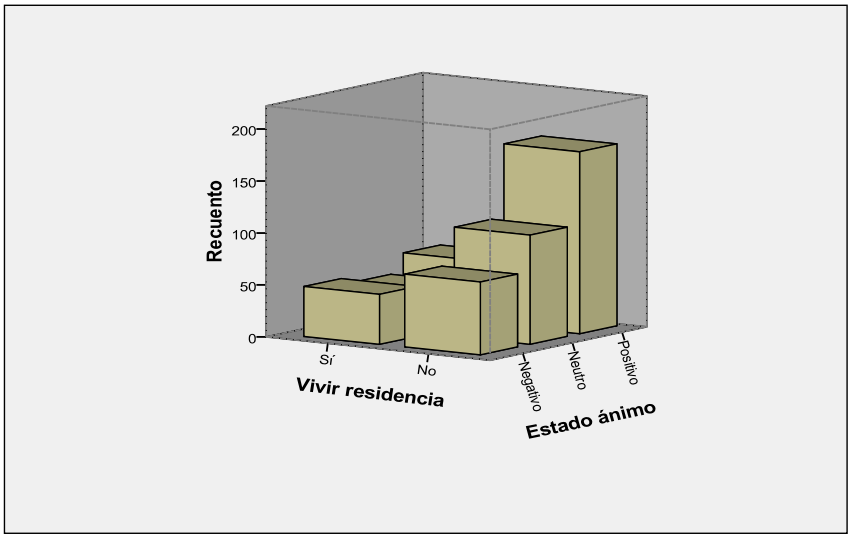
• Se presentan a continuación una serie de gráficos orientados a mostrar datos multivariados, si bien, la mayoría se ciñen al caso de 2 variables. Se diferencia su presentación en función del tipo de variables al que van dirigidos. No se van a presentar gráficos específicos para las variables ordinales, si bien, puede ser utilizado cualquiera de los orientados a variables categóricas o, si se asume naturaleza cuantitativa para las mismas, los orientados a este tipo de variables.

2.1. El caso de dos variables categóricas

• El diagrama de barras tridimensional o 3-D

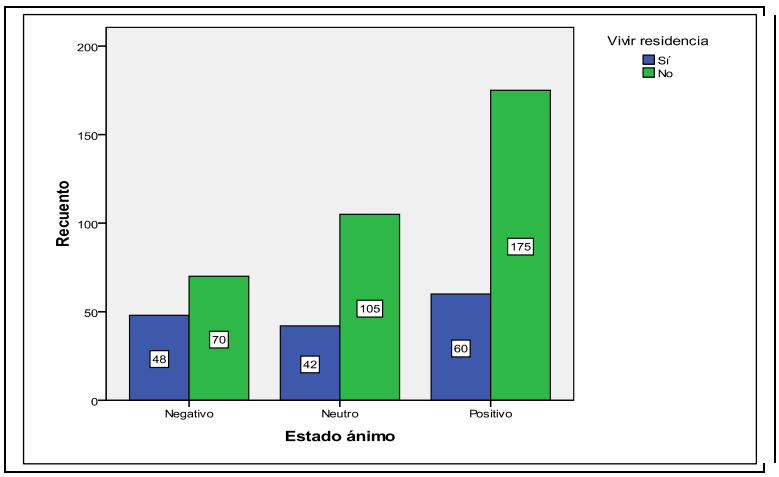
Ejemplos de diagrama de barras 3-D con la distribución conjunta de frecuencias absolutas de "Estado de ánimo" y "Vivir residencia", intercambiando la posición de ambas variables.

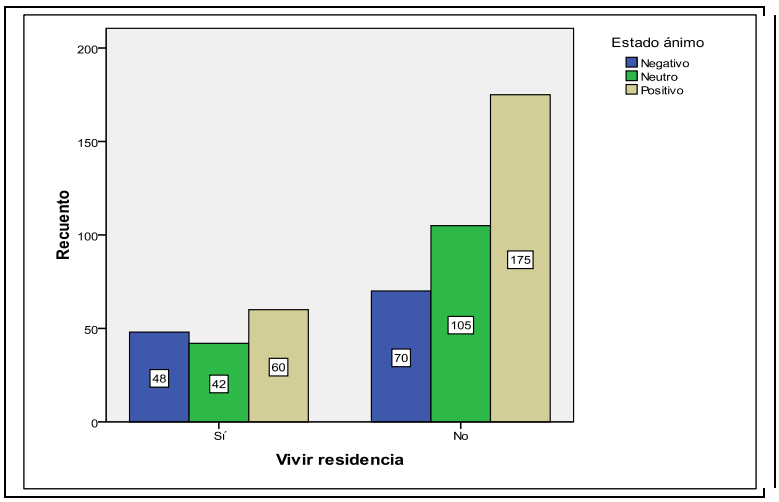




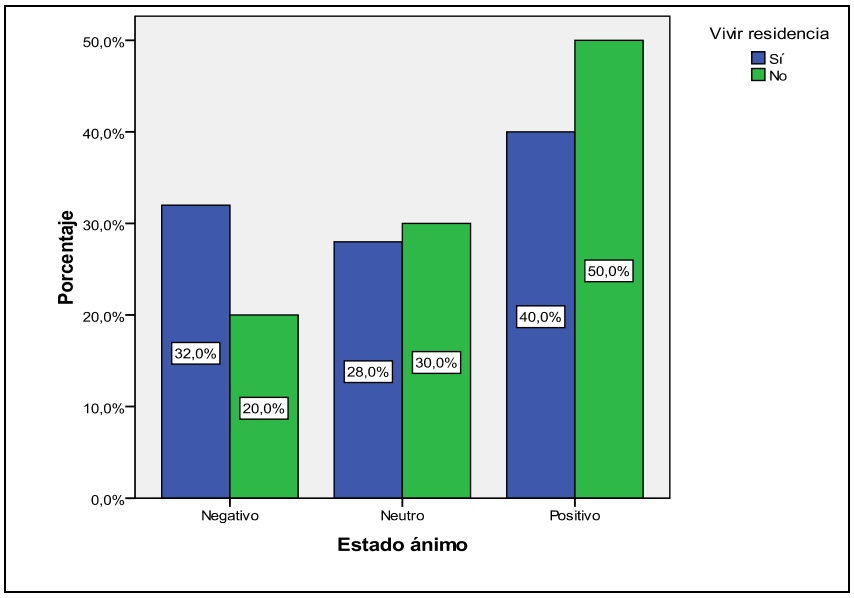
• El diagrama de barras

Ejemplos de diagrama de barras de la distribución conjunta de frecuencias absolutas de "Estado de ánimo" y "Vivir residencia" intercambiando la posición de ambas variables. Para diferenciar verbalmente ambos, haremos referencia al primero como diagrama de barras de frecuencias absolutas de la variable "Estado de ánimo" agrupada en función de "Vivir residencia", mientras que al segundo como diagrama de barras de la variable "Vivir residencia" agrupada en función de "Estado de ánimo". En ambos gráficos se representan frecuencias absolutas, por lo que las barras en ambos deben sumar el total del tamaño de la muestra (n = 500).





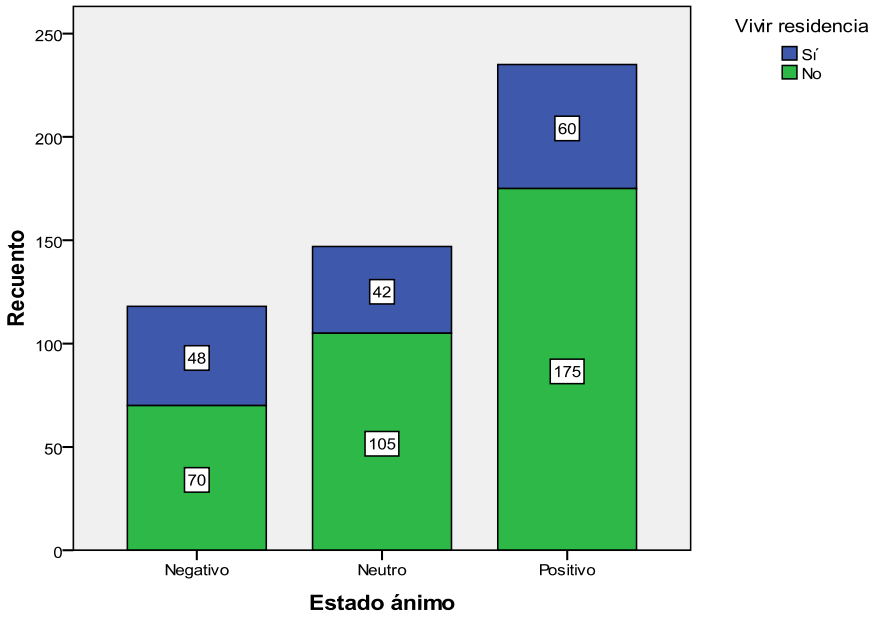
Este tipo de diagrama de barras de la distribución conjunta de dos variables sólo puede ser representado en SPSS para frecuencias absolutas, pero no para porcentajes. Cuando se solicita que se representen los porcentajes (la opción de frecuencias relativas no se ofrece en SPSS), lo que se representa no son los porcentajes en sí (ver gráfico de ejemplo a continuación), sino un tipo de porcentaje condicional que resulta útil para un determinado objetivo sobre el que se tratará en el próximo capítulo. Puede comprobarse como las barras de este gráfico no suman 100.

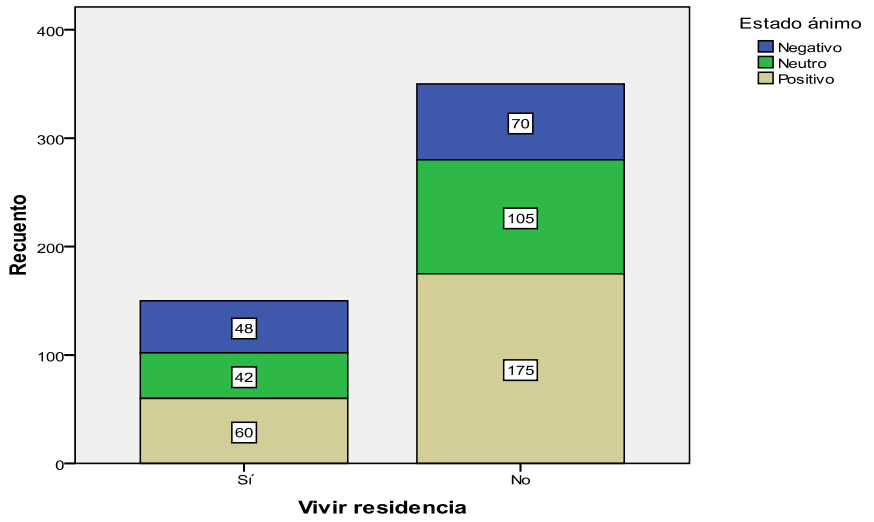


• El diagrama de rectángulos partidos (también denominado de barras apiladas)

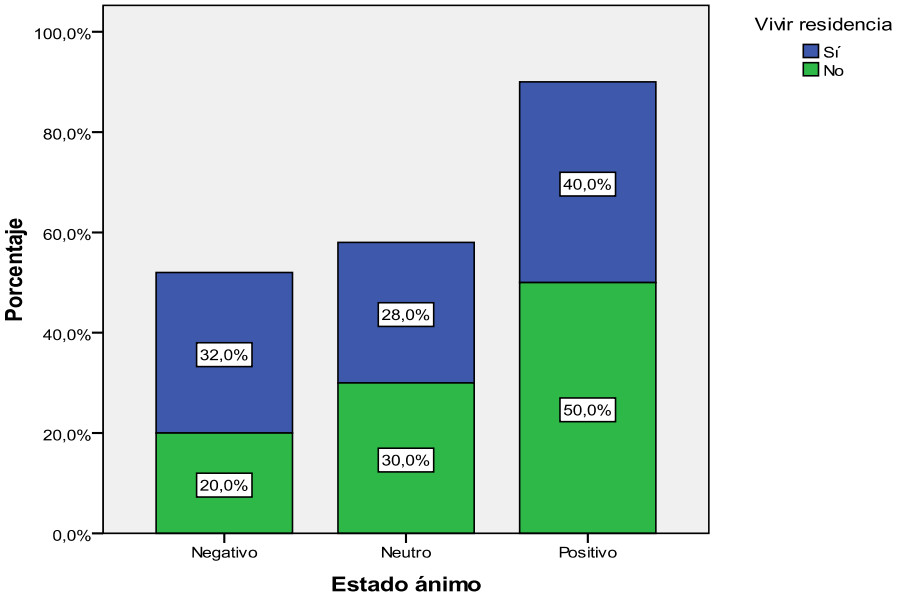
Ejemplos de diagrama de rectángulos partidos de la distribución conjunta de frecuencias absolutas de ambas variables, situando la variable "Estado ánimo" en el eje de abscisas (eje de categorías) en el primero, y la variable "Vivir residencia" en el segundo. Nótese que la altura de cada barra

corresponde a la distribución de frecuencias marginal de la variable situada en el eje de abscisas. En ambos casos, las barras suman el total del tamaño de la muestra.





De nuevo, este diagrama de rectángulos partido concreto sólo puede ser representado en SPSS para frecuencias absolutas, pero no para porcentajes. En el caso de que se solicite la representación con porcentajes lo que se representa, al igual que en el caso del diagrama de barras son un tipo de porcentajes condicionales, tal como se muestra en el ejemplo siguiente obtenido con SPSS.



Ejercicio 2: Sean las variables *X* (Aplicación de un programa de intervención para favorecer la interacción social [Sí (1), No (0)]) e *Y* (Grado de interacción en la hora de recreo [Bajo (1), Medio (2), Alto (3)]), de las que tenemos datos para un grupo de 20 alumnos de una clase en la que se evaluó la eficacia del citado programa de intervención.

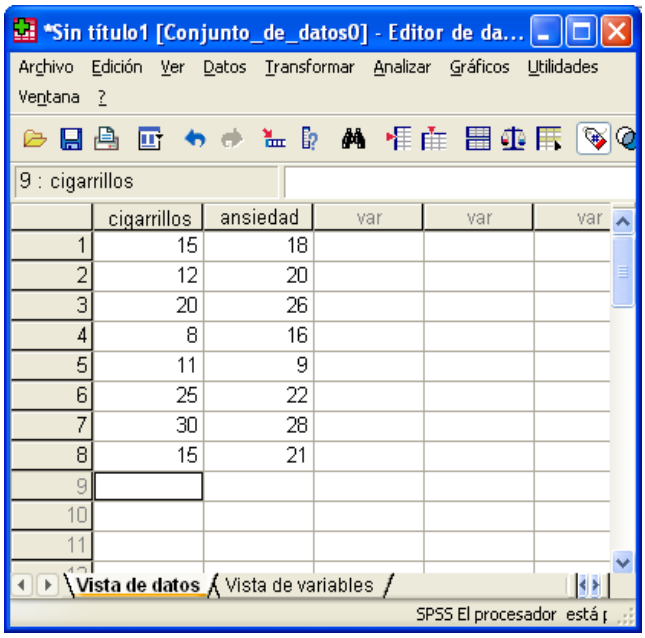
ID	*X*	*Y*
1	1	2
2	1	3
3	0	2
4	1	2
5	1	1
6	0	1
7	0	2
8	0	2
9	1	3
10	0	2
11	1	2
12	1	1
13	1	3
14	0	2
15	0	1
16	1	2
17	0	3
18	0	1
19	0	2
20	1	2

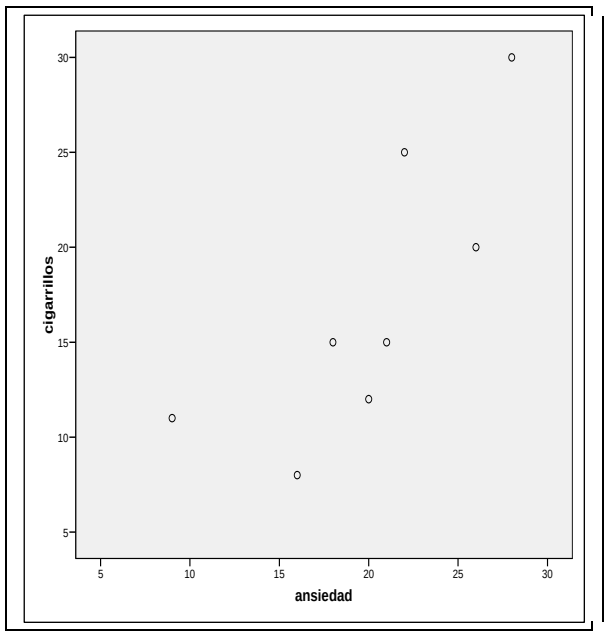
1. Organizar los datos anteriores a través de una distribución conjunta de frecuencias y de una tabla de contingencia (frecuencias absolutas).
2. Rehacer la tabla de contingencia utilizando frecuencias relativas
3. Representar gráficamente la distribución conjunta de ambas variables (frecuencias absolutas) utilizando un gráfico apropiado para tal fin.
4. Obtener los siguientes valores: n_{21}, n_{12} , n_{23} , p_{21} , p_{23} , n_{+2} , n_{1+} , p_{+3} , p_{2+}

2.2. El caso de dos variables cuantitativas

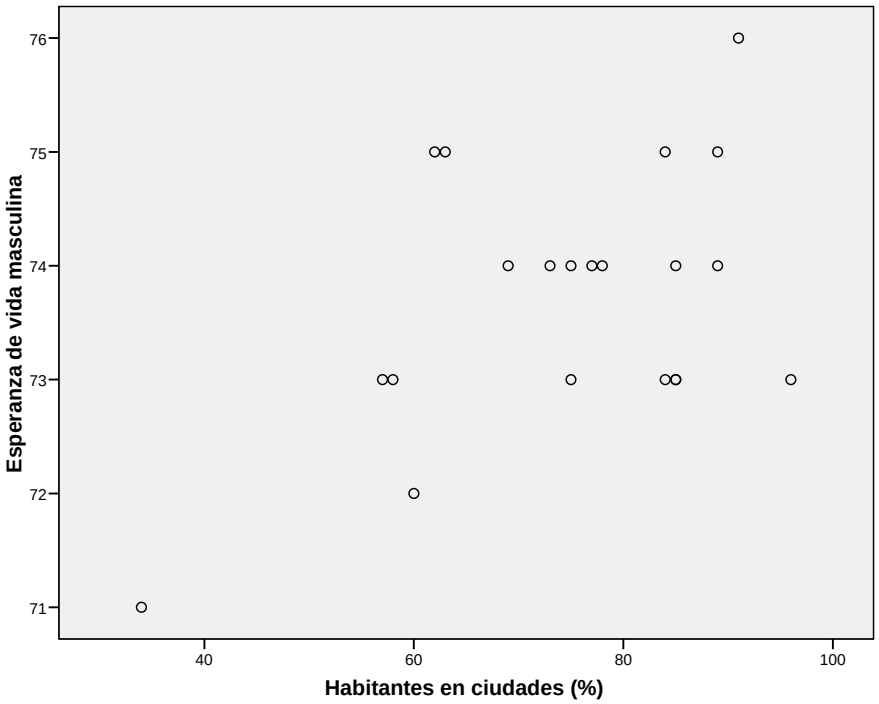
• El diagrama de dispersión (bivariado)

Ejemplo de diagrama de dispersión obtenido a partir de los datos de una muestra de 8 fumadores en las variables "Nº de cigarrillos que, en promedio, se fuma al día" y "Puntuación en un test de ansiedad [0, ..., 30]". Se muestran también los datos a partir de los que ha sido obtenido el mismo con el programa SPSS:

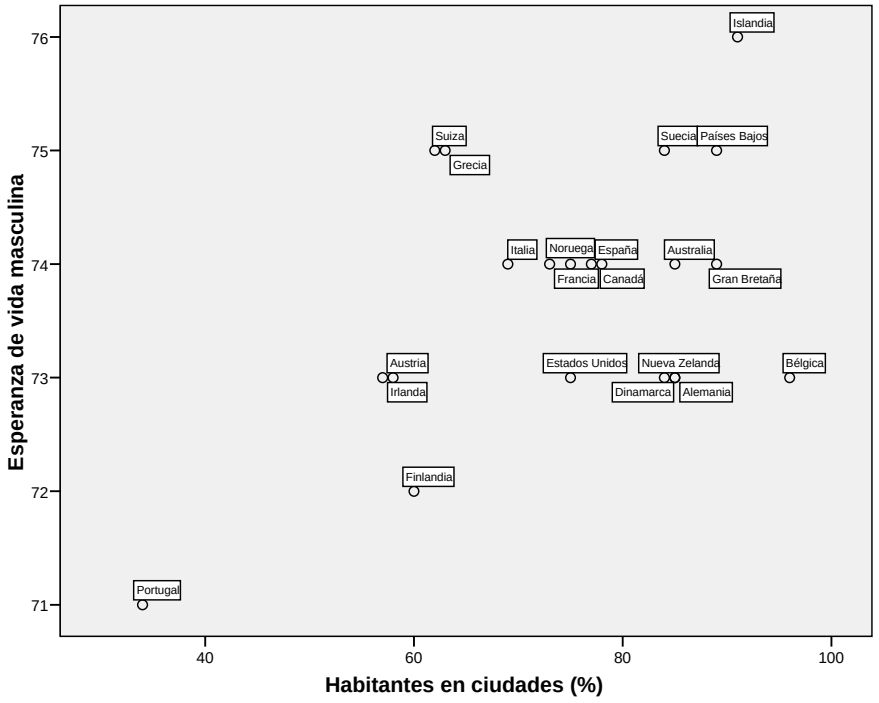




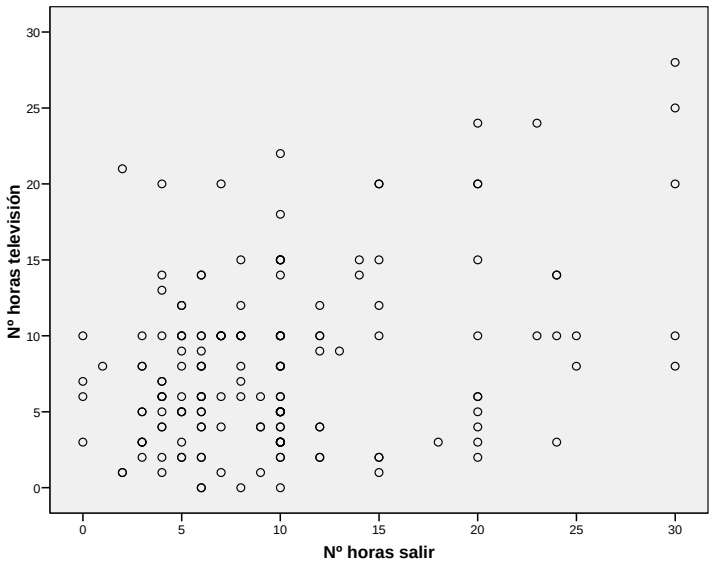
Otro **ejemplo** de diagrama de dispersión obtenido, en este caso, a partir de los datos de esperanza de vida masculina y el porcentaje de población que viven en ciudades en un conjunto de países de la OCDE en el año 1995:



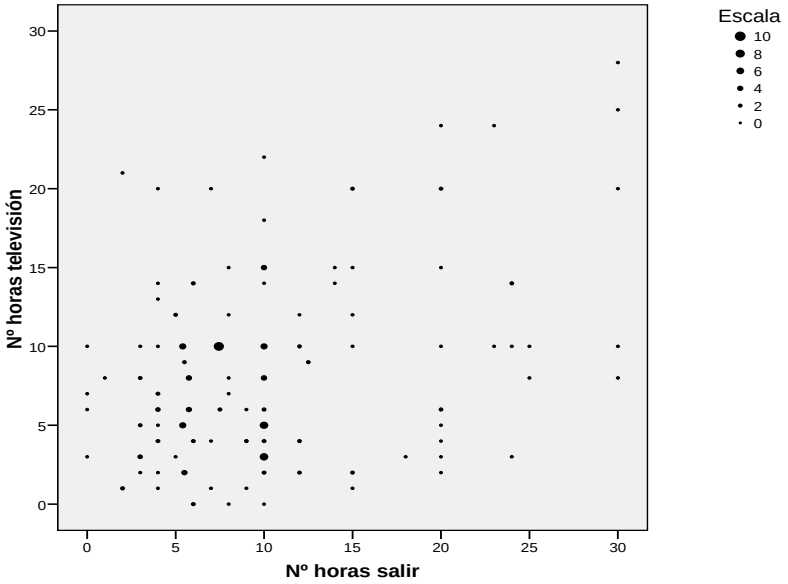
A continuación se muestra el mismo diagrama de dispersión con los puntos etiquetados, a fin de facilitar la posición de los casos (en este ejemplo, países) en la representación gráfica. Esta estrategia sólo resulta conveniente en el caso en que el número de casos en el archivo de datos no sea muy numeroso pues, en caso contario, resulta complicado visualizar la información presentada:



Un problema que se puede presentar en la representación de un diagrama de dispersión es el de la superposición de los puntos, esto es, que haya casos con los mismos valores en ambas variables, algo que no es extraño en archivos de datos con información para muchos casos. Véase por ejemplo el siguiente diagrama de dispersión de las variables "Nº horas salir" y "Nº horas televisión" dedicadas en promedio a la semana, a partir de los datos de una muestra de 174 estudiantes:

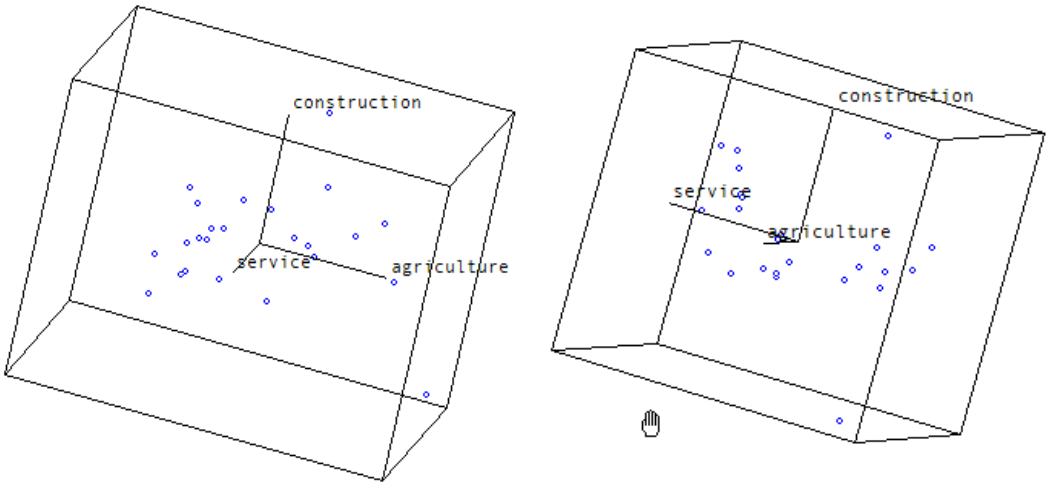


La no presencia de 174 puntos evidencia la superposición de algunos casos en ciertas posiciones. Algunos programas como SPSS permiten dimensionar los puntos en función del número de sujetos que coinciden en la misma posición, lo cual permite obtener una visualización más realista de la distribución conjunta de los datos. Véase como queda el diagrama de dispersión del ejemplo previo con los puntos dimensionados (donde pone 0 en la escala de los puntos, se supone que es 1):



• El diagrama de dispersión con 3 variables:

Ejemplo de diagrama de dispersión con el porcentaje de población activa en tres sectores productivos (agricultura, servicios y construcción) de un conjunto de países europeos (dos instantáneas del mismo obtenidas a partir de la rotación del mismo con el programa ViSta):



La pobre visualización de este tipo de diagrama de dispersión sobre el papel puede verse facilitada si se utiliza un programa que permita una fácil e inmediata rotación del gráfico en cualquier dirección, pues ello permite hacerse una idea más real de cómo es la nube de puntos tridimensional.

Ejercicio 3: Los siguientes datos proceden de un estudio en que se obtuvieron datos de 16 sujetos acerca del nº de horas de deporte que practicaban semanalmente (*X*) y la percepción que tenían sobre su estado de salud general (*Y*) en una escala de 1 a 10, indicando una mayor puntuación una percepción más positiva de la propia salud. Realizar una representación gráfica de la distribución de frecuencias conjunta de ambas variables.

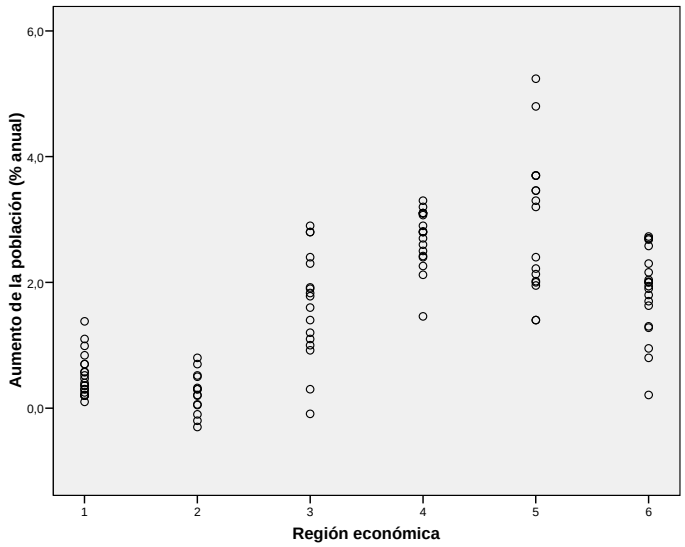
ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
X	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8	8
Y	4	3	3	5	6	4	4	6	5	2	7	9	6	8	9	8

2.3. El caso de una variable categórica y una variable cuantitativa

• El diagrama de dispersión también puede ser aplicado en la representación conjunta de la distribución de frecuencias absolutas de una variable categórica y una variable cuantitativa. A este

tipo de gráfico se le denomina en algunos textos como diagrama de puntos y es habitual que aparezca representada la variable categórica en el eje de abscisas y la variable cuantitativa en el eje de ordenadas.

Ejemplo de diagrama de dispersión de la distribución conjunta de las variables "Región económica" [1:OCDE; 2: Europa oriental; 3: Asia/Pacífico; 4: África; 5: Oriente Medio; 6: America latina]" y "% anual de crecimiento de la población" obtenida a partir de los datos recogidos para un total de 109 países de todo el mundo (N = 109):

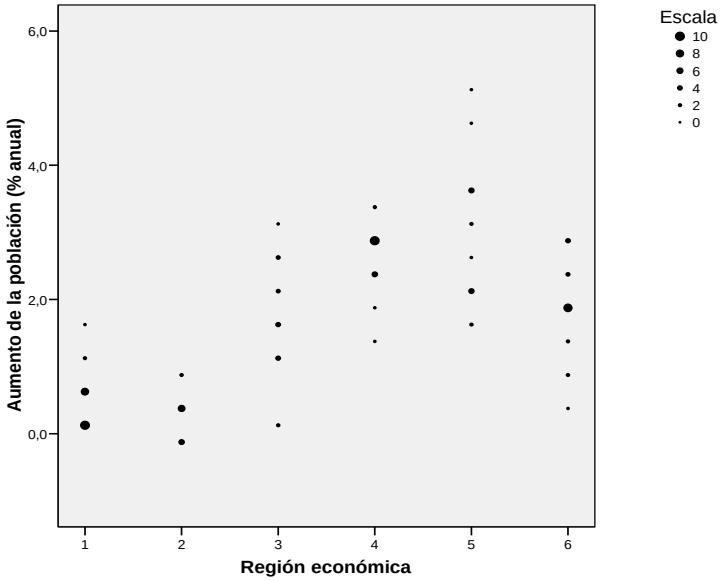


Piénsese acerca de la tabla de datos a partir de la que se ha obtenido este gráfico: ¿cuántas filas?; ¿cuántas columnas?; ¿qué tipo de variables?...

Fragmento de la tabla de datos original:

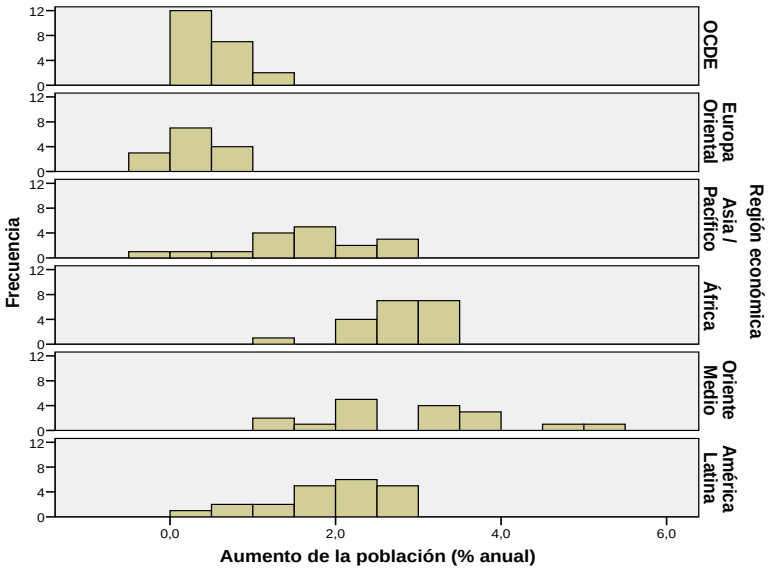
1	Acerbaján	Oriente Medio	1,4
2	Afganistán	Asia / Pacífico	2,8
3	Alemania	OCDE	,4
4	Arabia Saudí	Oriente Medio	3,2
5	Argentina	América Latina	1,3
6	Armenia	Oriente Medio	1,4
7	Australia	OCDE	1,4
8	Austria	OCDE	,2
9	Bahrein	Oriente Medio	2,4
10	Bangladesh	Asia / Pacífico	2,4
11	Barbados	América Latina	,2
12	Bélgica	OCDE	,2
13	Bielorusia	Europa Oriental	,3
14	Bolivia	América Latina	2,7
15	Bosnia	Europa Oriental	,7
16	Botswana	África	2,7
17	Brasil	América Latina	1,3
18	Bulgaria	Europa Oriental	-,2
19	Burkina Faso	África	2,8

Como ya se comentó al hablar del diagrama de dispersión, la posible superposición de puntos puede ocultar cuál es el número real de casos representado por cada punto en el gráfico. Tal inconveniente, cuando se dé el caso, se puede soslayar si se dimensionan los puntos, tal como se ha hecho en el siguiente **ejemplo** (donde pone 0 en la escala de los puntos, se supone que es 1):



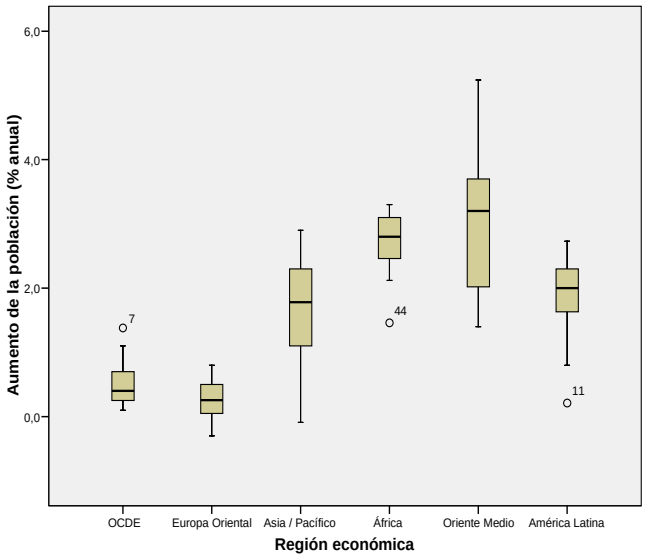
• El panel de histogramas ofrece la visualización en forma de histograma de la variable cuantitativa agrupada en función de los valores de la variable categórica.

Ejemplo para las variables las variables "Región económica" [1:OCDE; 2: Europa oriental; 3: Asia/Pacífico; 4: África; 5: Oriente Medio; 6: America latina]" y "% anual de crecimiento de la población":



• El diagrama de caja y bigotes agrupado ofrece la visualización en forma de gráfico de caja y bigotes de la variable cuantitativa agrupada en función de los valores de la variable categórica.

Ejemplo de diagrama de caja y bigotes de la variable "% anual de crecimiento de la población" agrupada en función de la variable "Región económica":



Ejercicio 4: Sean las variables *X* (Aplicación de un programa de intervención para favorecer la interacción social [Sí (1), No (0)]) e *Y* (Grado de interacción en la hora de recreo, medida por el nº de minutos en que se ha participado en actividades con otros compañeros), de las que tenemos datos para un grupo de 20 alumnos de una clase en la que se evaluó la eficacia del citado programa de intervención. Realiza una representación gráfica de los datos recogidos.

ID	X	Y
1	1	22
2	1	13
3	0	12
4	1	27
5	1	19
6	0	16
7	0	20
8	0	12
9	1	23
10	0	17
11	1	29
12	1	16
13	1	30
14	0	20
15	0	15
16	1	24
17	0	23
18	0	18
19	0	20
20	1	18