5.2 – Asociación: estadísticos de asociación entre variables

1. Concepto de asociación entre variables

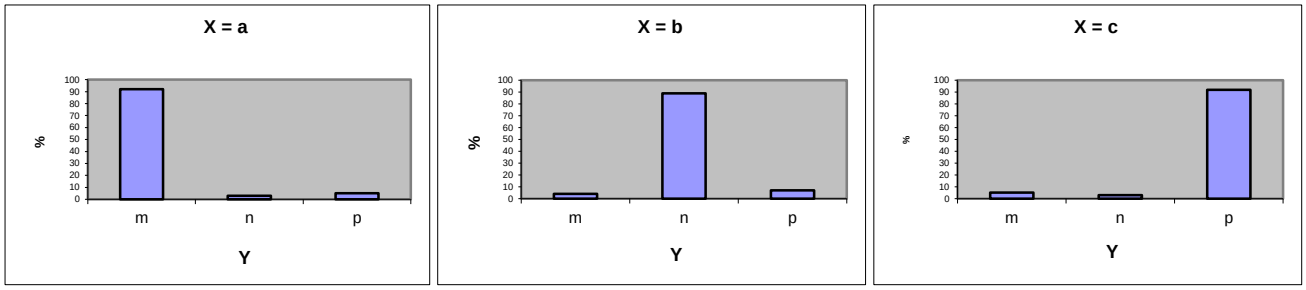
• El análisis estadístico de la asociación (relación, covarianza, correlación) entre variables representa una parte básica del análisis de datos en cuanto que una buena parte de las preguntas e hipótesis que se plantean en los estudios que se llevan a cabo en la práctica, implican analizar la existencia de relación entre variables.

• La existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables.

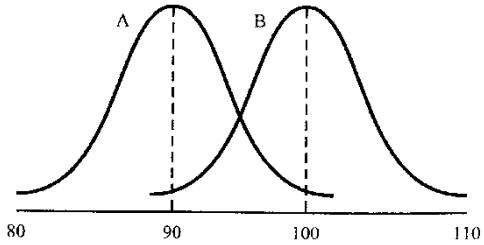
> A modo de **ejemplo** esquemático, si tenemos una variable X [*a*, *b*, *c*] y otra variable Y [*m*, *n*, *p*], de modo que los datos empíricos evidencian que los casos que en X son *a* en Y tienden a ser *m*, que los que son *b* tienden a ser *n*, y que los que son *c* tienden a ser *p*, ello pone de manifiesto la existencia de asociación entre ambas variables.

• Más formal que ésta, Solanas et al. (2005) ofrecen otra propuesta de definición general de lo que significa la asociación entre 2 variables: la existencia de asociación entre dos variables indicaría que la distribución de los valores de una de las dos variables difiere en función de los valores de la otra.

Sean el caso para nuestro **ejemplo** anterior, las distribuciones de frecuencias (expresadas en %) de la variable *Y* para aquellos casos que en la variable *X* son 'a', 'b' y 'c', respectivamente:

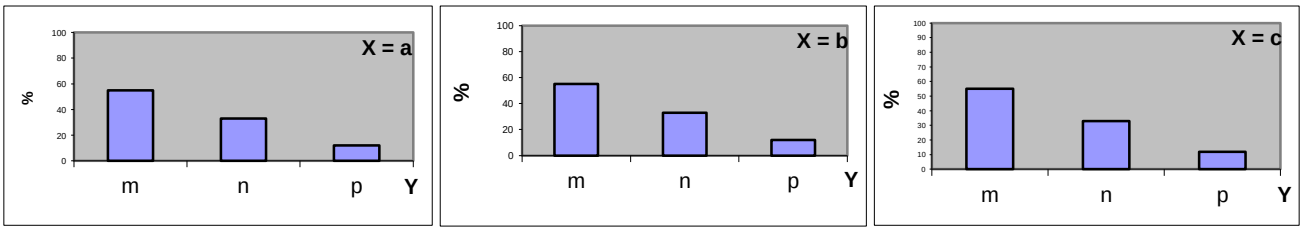


Otro **ejemplo** de la presencia de asociación entre 2 variables, una, la puntuación en un test de aptitud lingüística [0 a 150] (cuantitativa) y, la otra, la variable sexo [A: Varón; B: Mujer] (categórica). Véase a continuación, para un conjunto dado de datos de estas dos variables, la diferencia existente entre las distribuciones de frecuencias de la variable "Aptitud lingüística" condicionada a la variable "Sexo":



• Complementariamente, se habla de independencia entre variables cuando no existe tal patrón de relación entre los valores de las mismas.

Siguiendo el **ejemplo** anterior, sería el caso en que los sujetos que en X son *a*, en Y tienen una distribución que es igual o muy similar a la que tienen los que en X son *b* y a la que tienen los que en X son *c*, tal como ocurre en los siguientes ejemplos gráficos:



Para el **ejemplo** de las puntuaciones en el test de aptitudes numéricas sería el caso en que ambas distribuciones aparecieran superpuestas, poniendo de manifiesto que no hay diferencias en la distribución de las puntuaciones del test en función del sexo.

• La asociación entre variables no debe entenderse como una cuestión de todo o nada, sino como un continuo que iría desde la ausencia de relación (independencia) a la relación total entre las variables. Esto último representaría una relación determinista, esto es, el caso en que a partir del valor de un sujeto cualquiera en una variable, se puede afirmar cual será su valor en la otra variable.

• Señalar que, en ciertos contextos, es común utilizar la expresión tamaño del efecto para hacer referencia a la intensidad de la relación entre 2 variables.

2. Midiendo la asociación entre 2 variables

2.1. El caso de dos variables categóricas

• ¿Qué se puede decir acerca de la asociación entre las dos variables de la tabla de contingencia, "Estado de ánimo" y "Vivir en residencia", de las que se recogió datos en una muestra de 500 personas mayores de 70 años? (La variable "Estado de ánimo" se midió utilizando una escala que reflejaba 3 categorías ordenadas de estado de ánimo: malo, regular y bueno).

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

• Para evaluar si ambas variables están relacionadas hay que observar si la distribución de los valores de una de las variables difiere en función de los valores de la otra, esto es, hay que comparar las distribuciones condicionales de una de las dos variables. Si no hay relación entre las variables estas distribuciones deberían ser iguales. Por ejemplo, podemos comparar las distribuciones condicionales de frecuencias absolutas de "Estado de ánimo", esto es, *Sí* vivir en una residencia (48, 42, 60) frente a *No* vivir en una residencia (70, 105, 175).

• Si nos fijamos en las distribuciones condicionales de frecuencias absolutas de "Estado de ánimo" para cada uno de los dos valores de "Vivir en residencia" [Si; No], se observa que estas

distribuciones no son iguales, sin embargo, esto puede ser debido a que hay más sujetos que no viven en una residencia (350) que sujetos que sí viven en ella (150). En conclusión, no se deben comparar las distribuciones condicionales en frecuencias absolutas si el número de casos difiere en las categorías de la variable agrupadora.

• La asociación entre dos variables categóricas aparece más explícita en una tabla de contingencia de frecuencias relativas condicionadas, pues de ese modo se relativiza el posible diferente tamaño de los subgrupos definidos por cualquiera de las dos variables. Este tipo de tabla se puede obtener de 2 formas alternativas, bien dividiendo las celdas de cada fila entre el respectivo marginal (total) de fila, bien cada columna entre el respectivo marginal (total) de columna. Ambas tablas permitirán llegar al mismo tipo de conclusiones respecto a la asociación entre las 2 variables.

• En la práctica, si la relación entre las variables es asimétrica, es habitual considerar como variable condicionante a la variable explicativa (predictora, independiente). Por ejemplo, en un estudio en que se evaluó la influencia del "Nivel de estudios" [primarios, secundarios, superiores] sobre la "Percepción de la influencia de la ciencia en la sociedad" [negativa, indiferente, positiva], dado que el nivel de estudios era la variable explicativa, para analizar la relación entre ambas variables deberíamos comparar las distribuciones condicionales de frecuencias relativas de la "Percepción de la influencia de la ciencia" condicionada en función del "Nivel de estudios", es decir, para cada categoría de nivel de estudios. En nuestro **ejemplo** sobre "Estado de ánimo" y "Vivir en residencia", asumiendo que la segunda influye sobre la primera (relación asimétrica), deberemos comparar las distribuciones de frecuencias relativas de "Estado de ánimo" condicionada en función de "Vivir en residencia":

	Sí	**No**	**Total**
–	0,32 (48/150)	0,20 (70/350)	0,236 (118/500)
±	0,28 (42/150)	0,30 (105/350)	0,294 (147/500)
+	0,40 (60/150)	0,50 (175/350)	0,470 (235(500)
Total	1	1	1

El siguiente "output" muestra cómo queda la tabla de contingencia anterior cuando es obtenida con SPSS (se debe solicitar que en las casillas de la tabla aparezcan los % de columna):

Tabla de contingencia Estado ánimo * Vivir residencia

			Vivir residencia		
			Sí	No	Total
Estado ánimo	Negativo	Recuento	48	70	118
		% dentro de Vivir residencia	32,0%	20,0%	23,6%
	Neutro	Recuento	42	105	147
		% dentro de Vivir residencia	28,0%	30,0%	29,4%
	Positivo	Recuento	60	175	235
		% dentro de Vivir residencia	40,0%	50,0%	47,0%
Total		Recuento	150	350	500
		% dentro de Vivir residencia	100,0%	100,0%	100,0%

• En la tabla anterior, la comparación de las distribuciones condicionales de frecuencias relativas de "Estado de ánimo" condicionada en función de "Vivir en residencia" nos permitirá comprobar la existencia de asociación entre las dos variables. En caso afirmativo, como lo es para el ejemplo que nos ocupa, la comparación de esas distribuciones condicionales con la distribución marginal de la variable de respuesta nos permitirá ver claramente cuál es la naturaleza de esa relación.

A modo de **ejemplo**, si no hubiera relación entre ambas variables, las distribuciones de frecuencias relativas de "Estado de ánimo" condicionada en función de "Vivir en residencia" serían iguales a la distribución marginal de frecuencias relativas condicionadas (columna 'Total'):

	Sí	**No**	**Total**
−	0,236	0,236	0,236
±	0,294	0,294	0,294
+	0,470	0,470	0,470
Total	1	1	1

• La tabla de contingencia con las verdaderas distribuciones de frecuencias relativas de "Estado de ánimo" condicionada en función de "Vivir en residencia" (ver tabla inferior) difiere bastante de la tabla de independencia presentada arriba, poniendo de manifiesto la existencia de asociación entre ambas variables. Un análisis más exhaustivo de esta relación nos permite observar, por ejemplo, que la proporción de sujetos que tienen un estado de ánimo negativo entre los que viven en una residencia (0,32) es superior a la que cabría esperar si no hubiera relación entre ambas variables (0,236) o que la proporción de sujetos que tienen un estado de ánimo positivo entre los que no viven en una residencia (0,50) es superior a la que cabría esperar si no hubiera relación entre ambas variables (0,47).

	Sí	**No**	**Total**

−	0,32	0,20	0,236
±	0,28	0,30	0,294
+	0,40	0,50	0,470
Total	1	1	1

• Si la relación entre las variables es simétrica es indiferente qué variable se elige como condicionante. Así, por ejemplo, si deseamos valorar si hay relación entre el lugar de residencia (rural o urbano) y la rama de bachiller cursada (ciencias, sociales, salud o humanidades) y no consideramos a priori que una de las dos variables sea la variable explicativa, podríamos comparar, o bien, las distribuciones de frecuencias relativas de "Lugar de residencia" condicionada en función de "Bachiller", o bien, las distribuciones de frecuencias relativas de "Bachiller" condicionada en función de "Lugar de residencia". Las conclusiones a que se llegue serán las mismas.

Ejercicio 1: Analizar la asociación entre las dos variables dicotómicas siguientes: "Participación en un programa de intervención escolar que pretende favorecer el rendimiento académico [Si, No]" y "Resultados académicos a final de curso [Buenos, Malos]" a partir de los datos obtenidos en una muestra de 100 escolares de un colegio. (*Clg_1*).

Clg_1	Sí	No
Buenos	18	42
Malos	12	28

En un segundo colegio (*Clg_2*) se aplica ese mismo programa de intervención a una muestra de también 100 estudiantes, obteniéndose los datos resumidos en la siguiente tabla de contingencia. Analizar e interpretar la asociación existente entre ambas variables en este segundo colegio.

Clg_2	Sí	No
Buenos	24	31
Malos	16	29

Finalmente, los datos recogidos en un tercer colegio (*Clg_3*) se muestran resumidos en la siguiente tabla. Analizar e interpretar la asociación existente entre ambas variables en este caso.

Clg_3	Sí	No
Buenos	15	33
Malos	42	10

• El análisis gráfico de la asociación entre 2 variables categóricas puede intuirse a partir de un gráfico de barras agrupado de frecuencias absolutas si nos fijamos en la forma de cada una de las distribuciones condicionales (ver grupos de barras del mismo color en los ejemplos que se muestran más abajo): cuanto más similar sea la forma relativa de las mismas, menos relación habrá entre las

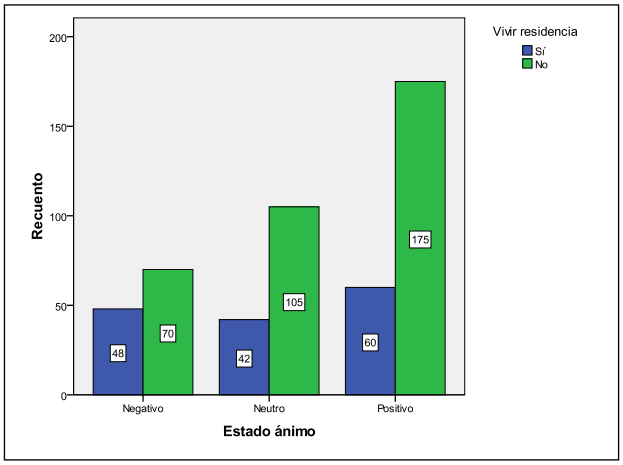
dos variables. Sin duda, nos resultará más fácil visualizar esta información si lo que se representa son frecuencias relativas condicionadas (o porcentajes condicionados), pues así se elimina el efecto del posible distinto tamaño de los subgrupos. Cuanto más similar sea la forma de las distribuciones condicionales (ver grupos de barras del mismo color), menor será la relación existente entre las variables. El gráfico de rectángulos partidos agrupado, cuando es representado con proporciones o porcentajes condicionadas, puede resultar también apropiado para evaluar la existencia de asociación entre dos variables categóricas (ver ejemplos más abajo). SPSS permite obtener tanto el gráfico de barras agrupado como el de rectángulos partidos agrupado, eso sí, con los porcentajes condicionados, no con las frecuencias relativas condicionadas.

• Cuando la relación entre las variables es asimétrica, es práctica habitual situar en el eje horizontal del gráfico de barras agrupado (eje de categorías según el SPSS) a la variable de respuesta, mientras que como variable agrupadora se elige a la variable explicativa.

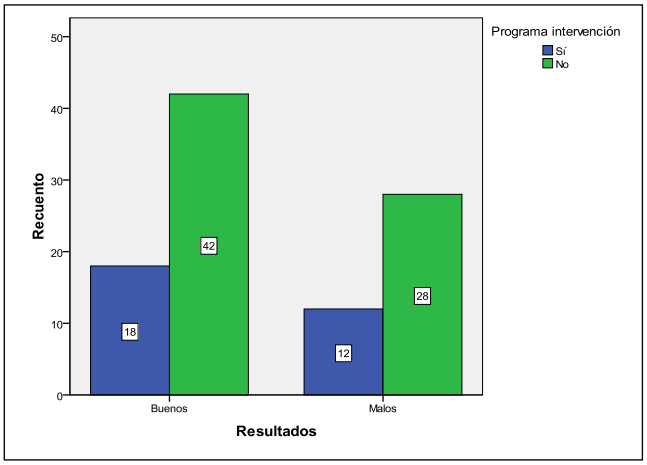
Ejemplos de gráfico de barras agrupado obtenidos con el SPSS para los datos de las variables "Estado de ánimo" y "Vivir en residencia" (izquierda), y para los datos de las variables "Programa de intervención" y "Resultados académicos" en el Colegio 1 (derecha):

(1) Ejemplos de gráfico de barras agrupado con frecuencias absolutas

(adecuado si los subgrupos definidos por la variable agrupadora son de similar tamaño, lo cual no es el caso en estos dos ejemplos):

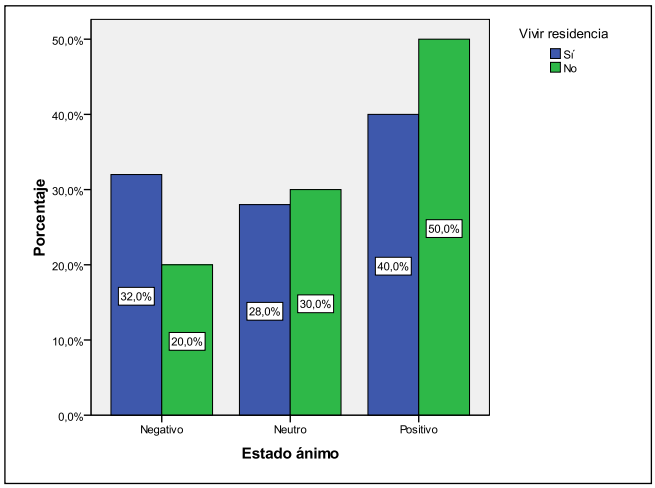


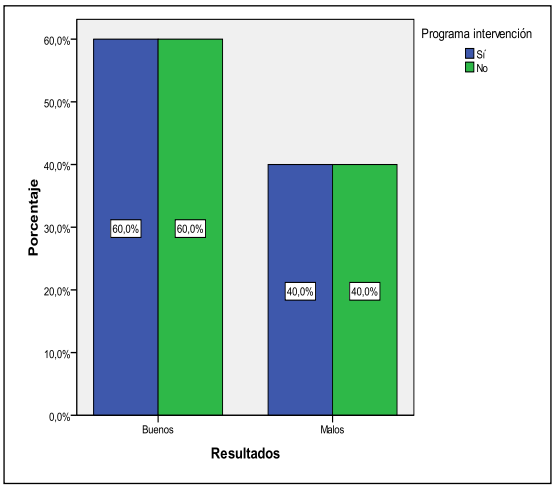
"Estado de ánimo" agrupada por "Vivir en residencia"



"Resultados académicos" agrupada por "Programa de intervención"

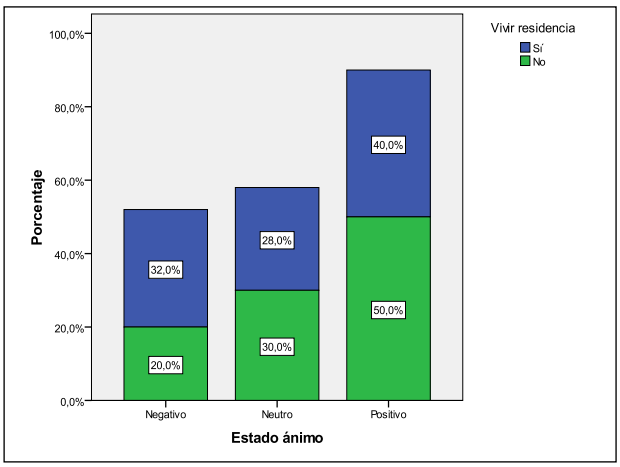
(2) Ejemplos de gráfico de barras agrupado con porcentajes condicionados:

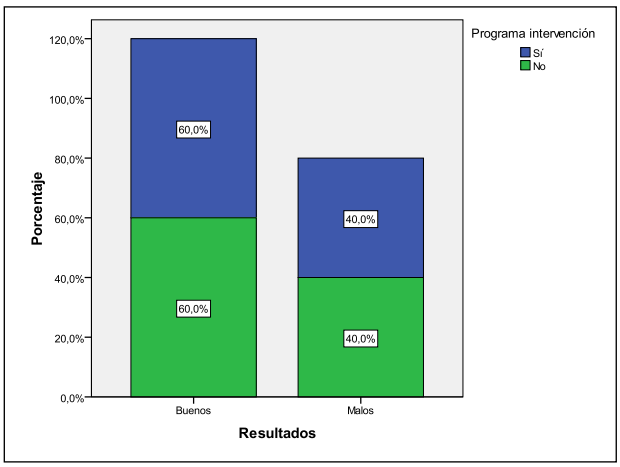




"Estado de ánimo" agrupada por "Vivir en residencia" "Resultados académicos" agrupada por "Programa de intervención"

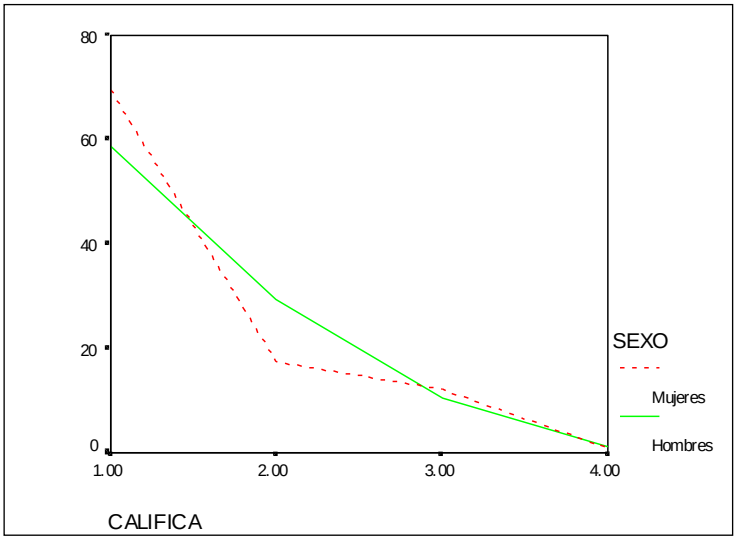
(3) Ejemplos de gráfico de rectángulos partidos agrupado con porcentajes condicionados:





"Estado de ánimo" agrupada por "Vivir en residencia" "Resultados académicos" condicionada a "Programa intervención"

Ejemplo de polígono de frecuencias agrupado con la variable "Calificación en una prueba" [1: Aprobado; 2: Notable; 3: Sobresaliente; 4: Matrícula de Honor] condicionada a la variable "Sexo". Esta representación gráfica ya se presentó con frecuencias absolutas en el capítulo anterior; ahora se muestra con porcentajes, poniéndose de manifiesto una escasa asociación entre ambas variables.



Ejercicio 2: Realizar el gráfico de barras agrupado con frecuencias absolutas, el gráfico de barras agrupado con frecuencias relativas condicionadas y el gráfico de rectángulos partidos agrupado con frecuencias relativas condicionadas para las variables "Programa de intervención" y "Resultados académicos" en el Colegio 2.

Ejercicio 3: Para analizar la asociación entre las variables "Motivación con los estudios de Psicología" y "Disfrutar con las explicaciones" se ha obtenido con SPSS la siguiente tabla de contingencia (datos procedentes de la encuesta sobre la vida académica):

Tabla de contingencia Disfrutar con las explicaciones * Motivación estudios Psicología

			Motivación estudios Psicología			
			alta	media	baja	Total
Disfrutar con las explicaciones	siempre o casi siempre	Recuento	16	8	0	24
		% dentro de Motivación estudios Psicología	???%	10,4%	,0%	???%
	algunas veces	Recuento	74	???	5	144
		% dentro de Motivación estudios Psicología	81,3%	84,4%	83,3%	82,8%
	casi nunca o nunca	Recuento	1	4	1	6
		% dentro de Motivación estudios Psicología	1,1%	???%	16,7%	3,4%
Total		Recuento	91	77	???	174
		% dentro de Motivación estudios Psicología	100,0%	100,0%	100,0%	100,0%

a) Rellena los interrogantes que aparecen en la tabla de contingencia.
b) ¿Cuál es la distribución marginal de frecuencias absolutas de la variable "Motivación…"? ¿y la de "Disfrutar…"?
c) ¿A qué distribución corresponden los valores [16; 74; 1]?
d) Tal como ha sido creada esta tabla, ¿cuál es la variable agrupadora o condicionante?
e) ¿A qué distribución corresponden los valores [10,4; 84,4; 5,2]?

f) ¿Cuáles serían las tres distribuciones condicionales de "Disfrutar…" de porcentajes condicionados si ambas variables fueran independientes?

g) ¿Parece haber relación entre "Motivación…." y "Disfrutar…"?

h) Realiza un gráfico adecuado para evaluar la relación entre ambas variables.

2.1.1 Índices estadísticos orientados a cuantificar la asociación entre dos variables categóricas

• Vamos a presentar tres de los índices estadísticos más utilizados en la práctica para este fin:

(1) El coeficiente ***ji-cuadrado*** **de Pearson (χ^2)**:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{i+} \cdot n_{+j}}{n}\right)^2}{\frac{n_{i+} \cdot n_{+j}}{n}}$$

• El índice χ^2 toma el valor 0 cuando las dos variables son independientes, siendo mayor que 0 cuando exista asociación entre ellas, tanto mayor cuanto más intensa sea esa relación. Ahora bien, no tiene un límite máximo, lo cual supone una dificultad a nivel interpretativo.

• Un problema importante de χ^2 es que su valor no sólo depende de la intensidad de la relación entre las dos variables, sino también del tamaño de la muestra a partir de la que se obtenga, de modo que cuanto mayor sea *n*, mayor será también el valor de χ^2.

• No habrá inconveniente en la interpretación de χ^2 cuando se utilice con fines comparativos, siempre y cuando se use para comparar la asociación entre sendos pares de variables cuyas tablas de contingencia sean del mismo tamaño (*I* x *J*) y el mismo *n*: para aquella tabla que se obtenga un valor de χ^2 más alto, la relación entre ese par de variables será más alta que para el otro par de variables.

• Muchos de los estadísticos que se han propuesto a posteriori a fin de evaluar la asociación entre variables categóricas se basan en el índice χ^2.

(2) El coeficiente ***phi*** **de Pearson (φ)**:

$$\varphi = \sqrt{\frac{\chi^2}{n}}$$

• El coeficiente *phi* puede oscilar entre 0 y $\sqrt{q-1}$, siendo *q* el número de modalidades de la variable que tenga menos de ellas.

• En tablas de contingencia de 2 x 2 oscila entre 0 y 1, por lo que suele utilizarse en la práctica cuando se da esta circunstancia, caso en el que se han extendido las normas interpretativas sugeridas por Cohen a la hora de evaluar la intensidad de la asociación (tamaño del efecto) para este coeficiente: $\varphi < 0{,}3$ => nivel bajo de asociación; $0{,}3 \leq \varphi < 0{,}5$ => nivel medio de asociación; $\varphi \geq 0{,}5$ => nivel alto de asociación.

(3) El coeficiente ***V* de Cramer**:

$$V = \sqrt{\frac{\chi^2}{n(q-1)}} \qquad (q = \min[I, J])$$

• El coeficiente *V* de Cramer oscila entre 0 (independencia) y 1, de modo que cuanto más próximos a 1 sean los valores, mayor intensidad en la asociación de las variables, pudiéndose también extender aquí los criterios interpretativos planteados antes para el coeficiente *phi*.

Ejercicio 4: Obtener los índices φ y *V* de Cramer a partir de las tres tablas de contingencia presentadas anteriormente para los tres colegios y, también, para el ejemplo de las variables "Estado de ánimo" y "Vivir en residencia". Los valores del índice χ^2 son 0, 0,673, 24,97 y 8,78, respectivamente.

2.2. El caso de una variable categórica y una cuantitativa

• De nuevo, el análisis de este tipo de asociación supone comparar la distribuciones de una variable condicionada a los distintos valores que tome la otra. Normalmente, se suele tomar como condicionada a la cuantitativa y como condicionante a la categórica, si bien, las conclusiones a las que llegaríamos serían las mismas si se hiciese al revés. Si no hay diferencias entre las distribuciones condicionales, ello indicará que no hay asociación entre ambas variables.

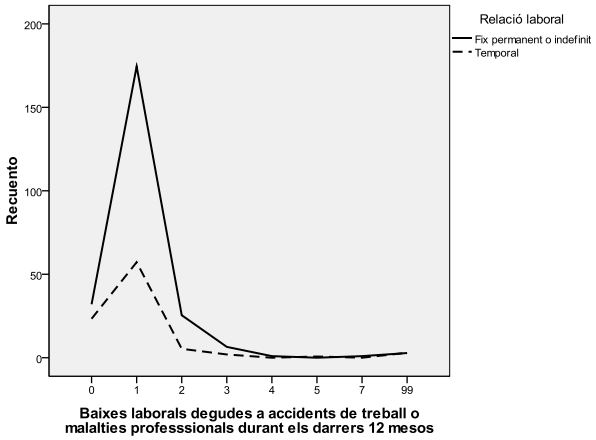
Ejemplo del caso en que se quiera analizar la asociación entre las variables "Nota en un examen de una asignatura [0 a 10]" y "Grupo en el que se está matriculado [1 a 6]", disponiéndose de los datos de un total de 768 estudiantes de 6 grupos:

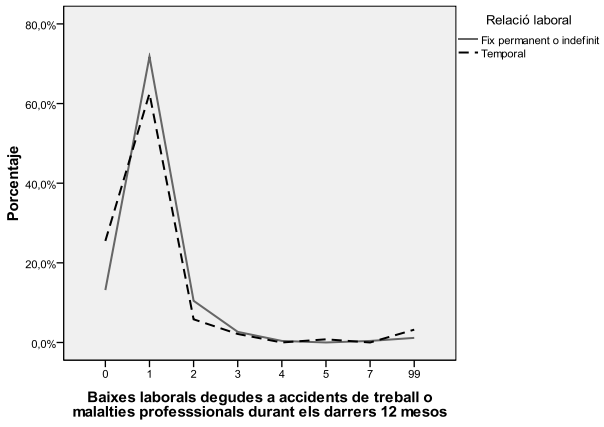
Grupo 1		**Grupo 2**		**Grupo 3**		**Grupo 4**		**Grupo 5**		**Grupo 6**	
x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
0	2	0	3	1,8	1	…	…	…	…	…	…
,3	3	1,3	1	2,0	7						
,5	1	1,8	1	2,3	3						
,8	3	2,0	2	2,5	2						
1,0	1	2,3	4	2,6	3						
1,3	1	2,5	2	2,8	2						
1,5	2	2,8	5	2,9	1						
1,8	2	2,9	1	3,0	2						
2,0	2	3,0	4	3,3	3						
2,3	3	3,3	3	3,5	7						
2,5	2	3,5	6	3,8	9						
2,6	1	3,8	3	3,9	1						
2,8	6	3,9	1	4,0	4						
3,0	5	4,0	2	4,1	1						
3,3	3	4,3	5	4,3	3						
3,5	5	4,5	6	4,5	9						
3,8	7	4,6	1	4,7	6						
4,0	8	4,7	8	4,8	7						
4,3	7	4,8	6	4,9	4						
4,5	5	4,9	6	5,0	3						
4,7	4	5,0	5	5,3	4						
4,8	3	5,1	1	5,5	4						
4,9	5	5,3	5	5,6	1						
5,0	3	5,4	1	5,8	1						
5,1	1	5,5	6	5,9	2						
5,3	6	5,6	1	6,0	4						
5,5	4	5,8	9	6,3	3						
5,8	7	5,9	1	6,5	6						
6,0	5	6,0	5	6,8	2						
6,1	1	6,3	2	7,0	4						
6,3	5	6,5	4	7,1	1						
6,5	3	6,8	4	7,3	2						
6,8	2	6,9	1	7,5	5						
7,0	1	7,0	2	8,0	1						
7,5	3	7,3	4	8,3	2						
7,6	1	7,5	3	8,4	1						
8,0	2	7,8	2	9,0	4						
9,0	2	8,0	2	9,3	1						
		8,1	1	9,5	1						
		8,3	1								

• Tal como ya este ejemplo evidencia, puede resultar bastante dificultoso el comparar las distribuciones condicionales de una variable cuantitativa agrupada en función de los valores de una variable categórica. Para facilitar tal cometido, se puede recurrir a la utilización de algunas representaciones gráficas como las que a continuación se describen.

(1) El polígono de frecuencias agrupado, un tipo de gráfico que ya se presentó en el capítulo anterior, representa una buena opción si tenemos en cuenta un detalle importante: si el tamaño de los subgrupos definidos por la variable condicionante no es el mismo, o bastante similar, es conveniente representar este gráfico con proporciones o porcentajes condicionados, a fin de que los polígonos puedan compararse de un modo equitativo, no distorsionado por el diferente tamaño de los subgrupos.

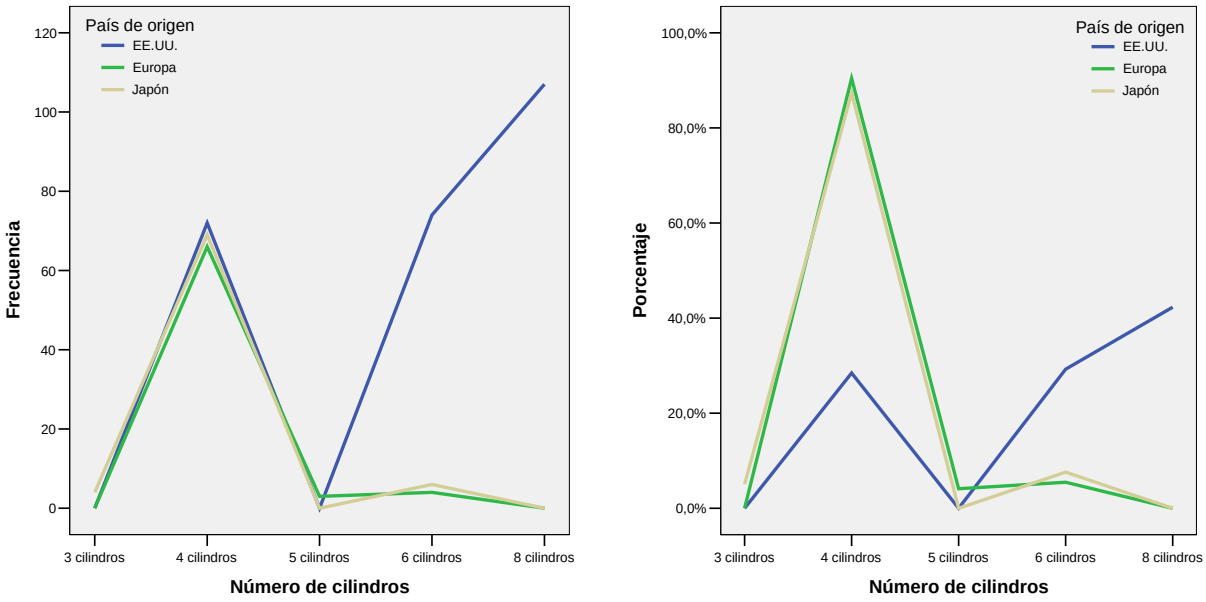
Ejemplo de polígono de frecuencias agrupado para la variable "Número de bajas laborales (durante los últimos 12 meses)" agrupada en función del tipo de "Relación laboral" de los trabajadores [contrato fijo; contrato temporal]. Recuérdese que cuando el tamaño de los grupos definidos por la variable agrupadora es desigual, no se deben representar las frecuencias absolutas sino frecuencias relativas o porcentajes condicionados, es decir, dividiendo la frecuencia absoluta por el tamaño de cada uno de los grupos. Véase en este ejemplo que el gráfico de la izquierda (con frecuencias absolutas) puede resultar engañoso al dar la sensación de que ambas distribuciones son bastante diferentes, sin embargo, este efecto es debido a que el número de trabajadores fijos es bastante superior al de trabajadores temporales. En el gráfico de la derecha, donde se representan las distribuciones de porcentajes condicionados, se puede comprobar que ambas distribuciones son, en realidad, bastante similares, poniendo de manifiesto la casi ausencia de relación entre ambas variables.



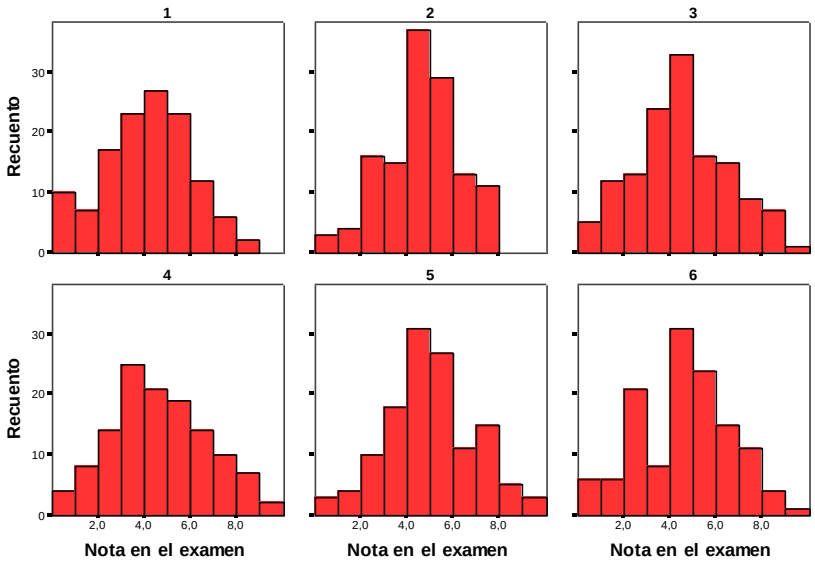


Un **ejemplo** en el que se aprecia más este hecho es el siguiente, con los datos de un estudio que se hizo en los EEUU sobre las características de los diferentes modelos de coches existentes en el mercado. En concreto, a continuación se muestra la información correspondiente a la distribución conjunta de frecuencias de las variables "Nº de cilindros" y "País de origen" para una muestra de 405 vehículos, así como los correspondientes polígonos de frecuencias agrupados obtenidos tanto con frecuencias absolutas como con porcentajes condicionados:

		País de origen			
		EE.UU.	Europa	Japón	Total
Número de cilindros	3 cilindros			4	4
	4 cilindros	72	66	69	207
	5 cilindros		3		3
	6 cilindros	74	4	6	84
	8 cilindros	107			107
Total		253	73	79	405

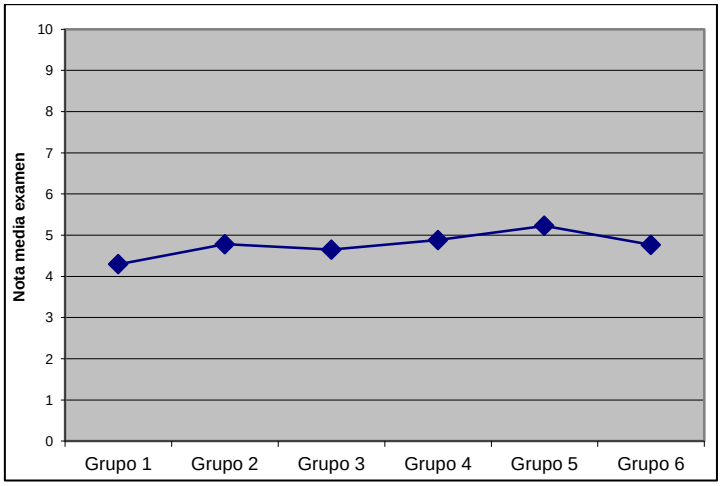


(2) El panel de histogramas muestra un histograma de la variable cuantitativa para cada uno de los subgrupos definidos por la variable categórica. A continuación se muestra un **ejemplo** de este tipo de representación para los datos de las variables "Nota en un examen de una asignatura" [0 a 10] y "Grupo en el que se está matriculado" [1 a 6] presentados ya al principio de esta sección (n = 768). En este ejemplo no se ha optado por representar las frecuencias relativas o los porcentajes condicionados porque los seis grupos eran muy similares en su tamaño.



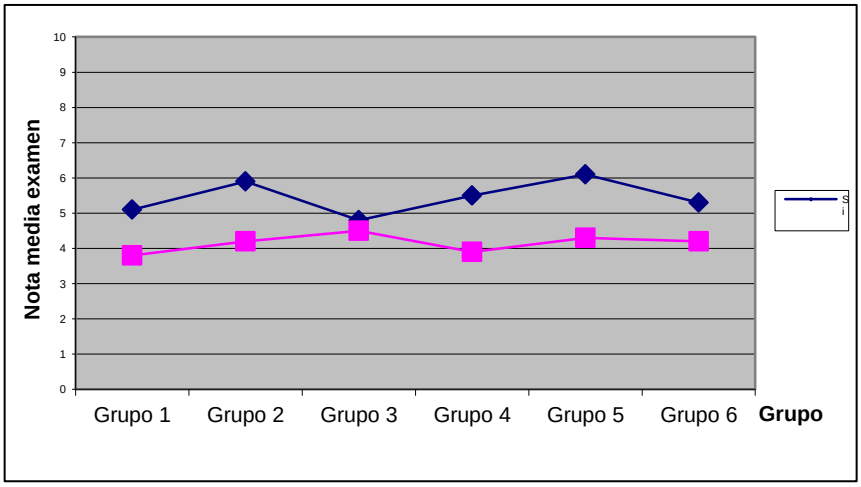
• Mientras que el polígono de frecuencias agrupado y el panel de histogramas representan la distribución de frecuencias de la variable cuantitativa para cada modalidad de la variable categórica, los gráficos que aparecen a continuación lo que representan son determinados estadísticos que caracterizan a esas distribuciones de frecuencias condicionales.

(3) El gráfico de medias: **ejemplo** para la variable "Nota en un examen de una asignatura" agrupada en función de la variable "Grupo en el que se está matriculado [1 a 6]":



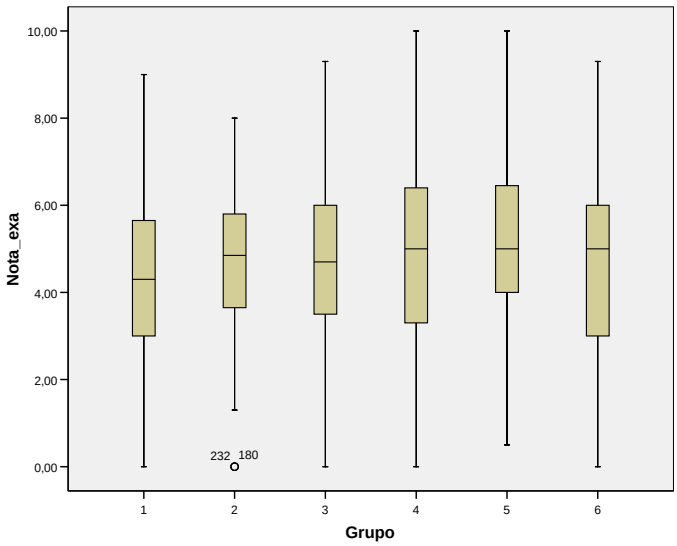
• La agregación de los datos originales en forma de medias hace factible incluir en esta representación gráfica la información de una variable categórica adicional, lo cual nos va a permitir presentar la información correspondiente a 3 variables.

Ejemplo de gráfico de medias de la variable "Nota en un examen de una asignatura" agrupada en función de las variables "Grupo" [1 a 6] y "Asistencia regular a las clases" [Si, No]:



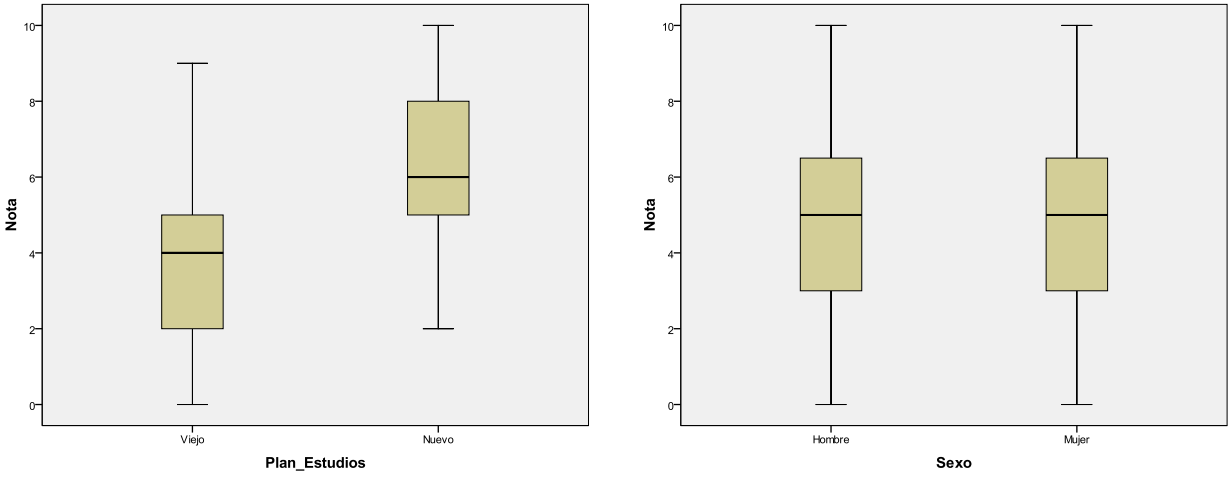
(4) El gráfico de caja y bigotes agrupado, el cual puede ser especialmente apropiado si la variable categórica tiene un número amplio de modalidades, pues resulta más fácil de encajar numerosas

cajas en el mismo espacio gráfico. A continuación se muestra un **ejemplo** para las mismas variables representadas en el panel de histogramas y el gráfico de medias presentados anteriormente.



• Las cuatro representaciones gráficas presentadas nos van a permitir comparar el grado de solapamiento (coincidencia) de las distribuciones condicionales. En general, cuanto mayor sea ese solapamiento, menor será la intensidad de la asociación entre las dos variables y, viceversa, cuanto mayor sea la discrepancia, mayor será el tamaño del efecto de la relación. Así, en los ejemplos anteriores en que aparecían representadas las variables "Grupo" y "Nota" se observa bastante solapamiento entre las 6 distribuciones condicionales, poniendo de manifiesto una escasa relación entre ambas variables.

Ejemplo de diferente intensidad en la asociación en dos pares de variables (cada par constituido por una variable categórica dicotómica y una misma variable cuantitativa, la "Nota"): en este caso, la visualización de las distribuciones condicionales mediante gráficos de caja y bigotes agrupados pone de manifiesto una relación más baja entre las variables "Nota" y "Sexo" (mayor coincidencia), que entre las variables "Nota" y "Plan de estudios" (menor coincidencia).



2.2.1 Índices estadísticos orientados a cuantificar la asociación entre una variable categórica y una variable cuantitativa

• A la hora de captar las diferencias existentes entre las distribuciones condicionales de la variable cuantitativa para cada uno de los valores de la variable categórica, la mayoría de los índices estadísticos que han sido propuestos se han centrado en comparar un aspecto específico de esas distribuciones: su tendencia central y, más comúnmente, la comparación de sus medias.

• Los estadísticos que a continuación se presentan están todos ellos basados en las diferencias entre las medias en los subgrupos definidos por la variable categórica, eso sí, adoptando diferentes estrategias a la hora de estandarizar y hacer más interpretables esas diferencias entre las medias de los subgrupos.

(1) El índice de asociación ***d de Cohen*** es apropiado cuando se tenga una variable cuantitativa *Y* y una variable categórica *X* dicotómica [a, b]. Se calcula a través de la siguiente fórmula:

$$d = \frac{\left|\bar{Y}_a - \bar{Y}_b\right|}{s_Y}$$

El índice *d* de Cohen representa una diferencia de medias tipificada, por lo que su valor puede oscilar entre 0 (variables independientes) y un valor que, aunque no tiene a priori límite máximo, como ocurre para las puntuaciones típicas, es ya infrecuente que adopte valores por encima de 2. Cohen (1992) sugirió las siguientes normas interpretativas, aunque el propio autor remarcó que se deben utilizar sólo en el caso que no se tenga ningún criterio sustantivo que permita realizar la interpretación: valores absolutos de *d* en torno a 0,2 indicarían una intensidad de la asociación (tamaño del efecto) baja; en torno a 0,5, media; mientras que en torno a 0,8 y superiores, alta.

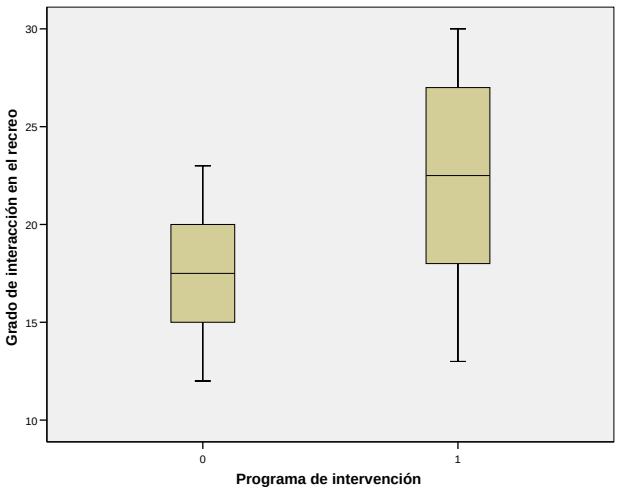
(2) El índice ***f* de Cohen** permite analizar la relación entre una variable cuantitativa (*Y*) y una categórica (*X*) en el caso en que esta última tenga más de dos valores posibles (*k* valores). Se basa para ello en el cálculo de la dispersión de las medias de los diferentes subgrupos definidos por los *k* valores de la variable *X*:

$$f = \frac{s_{\bar{Y}}}{s_Y}, \qquad \text{donde } s_{\bar{Y}} = \sqrt{\frac{\sum_{i=1}^{k} n_i \cdot (\bar{Y}_i - \bar{Y})^2}{n}}$$

• En el caso en que las medias de los subgrupos sean iguales o muy próximas, la desviación típica de las medias será igual o prácticamente igual a 0, denotando la ausencia de asociación entre ambas variables. El valor de la *f* de Cohen será siempre mayor o igual a 0, tanto mayor cuanto más intensa sea la asociación entre las variables.

Ejercicio 5: 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. Analiza la asociación existente entre ambas variables. *Nota*: La desviación estándar de la variable *Y* es igual a 5,09.

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



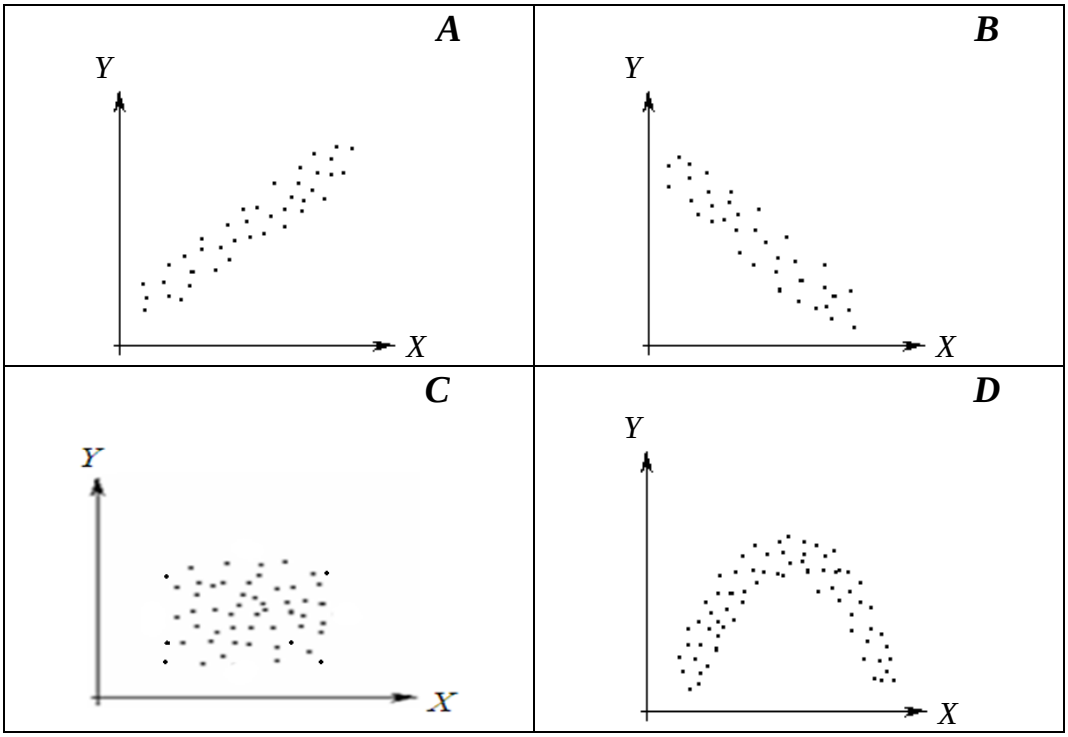
2.3. El caso de dos variables cuantitativas

• Al igual que en los casos anteriores, la existencia de correlación o asociación entre 2 variables cuantitativas viene determinada por la presencia de diferencias en las distribuciones condicionales de una variable para los distintos valores de la otra.

• Sin embargo, dado el número tan amplio de distribuciones condicionales que se pueden llegar a obtener en este caso, lo más habitual es analizar la asociación directamente sobre un diagrama de dispersión, observando la disposición de la nube de puntos que representa la distribución conjunta de ambas variables. Así, ¿qué podríamos decir acerca de la asociación entre los 4 pares de variables cuyos diagramas de dispersión se muestran a continuación?

• Un aspecto relevante del análisis de la correlación entre dos variables cuantitativas es que la presencia de ésta se puede plantear de acuerdo a diferentes modelos o patrones de asociación, por ejemplo, en forma de línea recta, tal como en los ejemplos *A* (relación lineal directa o positiva) y *B*

(relación lineal inversa o negativa) de la figura superior, o en forma curvilínea tal como en *D* (relación parabólica o cuadrática). Así, la forma de evaluar la intensidad de la correlación suele consistir en analizar el ajuste de la nube de puntos al modelo de asociación que se considere que representa más adecuadamente a la distribución conjunta de ambas variables.



2.3.1. Índices estadísticos orientados a cuantificar la asociación entre dos variables cuantitativas

• En la cuantificación de la asociación entre 2 variables cuantitativas nos vamos a ceñir al supuesto de que un modelo de relación lineal subyace a la asociación entre ambas. Subrayar que con frecuencia se obvia en los textos estadísticos que la relación que se analiza es en realidad una relación de tipo lineal. Los índices más utilizados en la práctica estadística a la hora de analizar la intensidad o tamaño del efecto de la relación lineal entre dos variables cuantitativas son los tres siguientes:

(1) La covarianza (S_{XY} *o* σ_{XY}):

$$S_{XY} = \frac{\sum (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{n}$$

• Al numerador de esta expresión se le conoce en la literatura estadística como suma de productos cruzados (SP_{XY}), por lo que la anterior expresión queda como: $s_{XY} = SP_{XY} / n$

• La covarianza puede tomar valores tanto positivos como negativos. A nivel interpretativo, un mayor valor de la covarianza en valor absoluto indicará una relación lineal más intensa entre las dos

variables. Un valor positivo pone de manifiesto una relación lineal directa; uno negativo, una relación lineal inversa; y si igual o muy próximo a 0, la inexistencia de relación lineal entre las dos variables.

(2) El coeficiente de correlación producto-momento de Pearson (R_{XY})

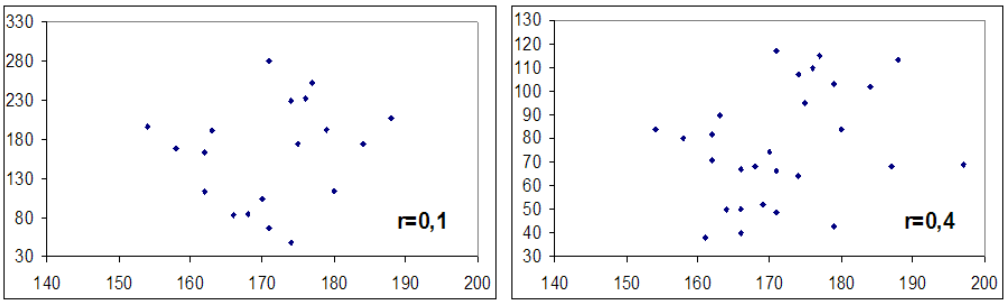
• Los inconvenientes de la covarianza –por una parte, no tiene valores máximo y mínimo y, por otra parte, depende de las unidades de medida de las variables- se resuelven estandarizando este índice al dividirlo por el producto de las desviaciones típicas de ambas variables. Se obtiene así el conocido como coeficiente de correlación producto-momento de Pearson:

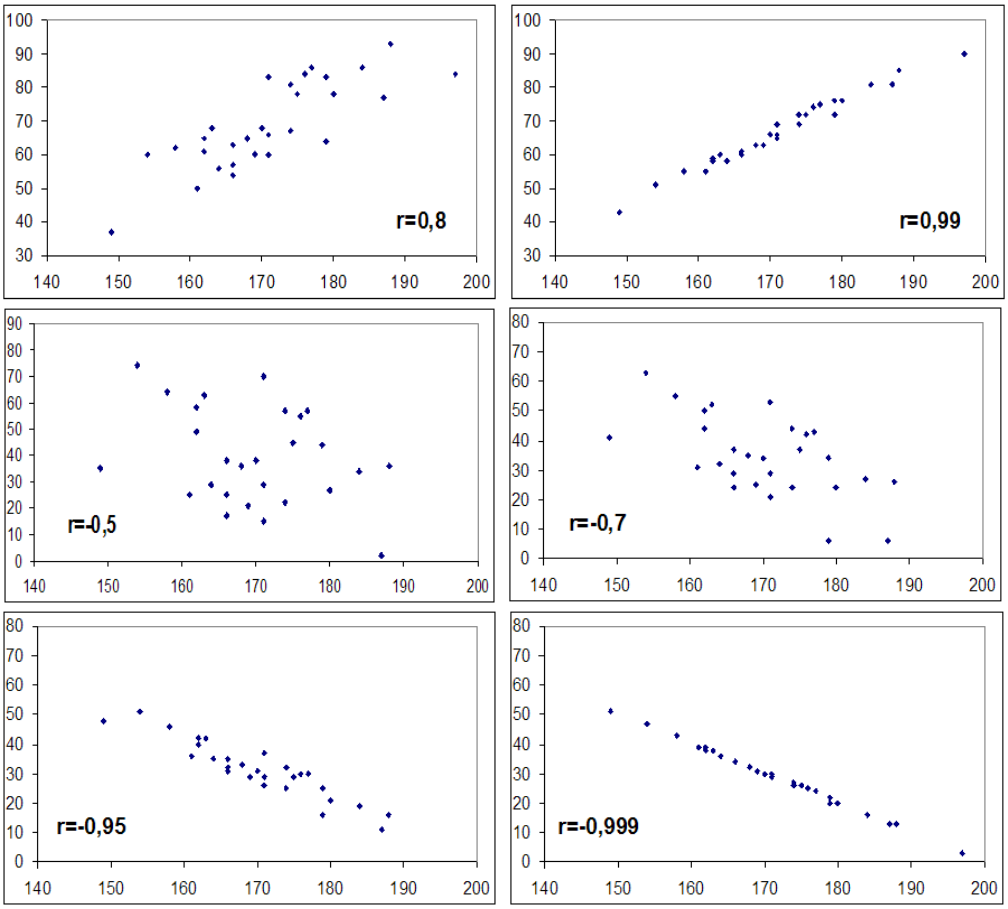
$$R_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

• El coeficiente de correlación de Pearson se interpreta de modo análogo a la covarianza pero, al oscilar entre -1 y 1 como máximo, la interpretación del mismo resulta más intuitiva a la vez que facilita el establecimiento de comparaciones entre los coeficientes obtenidos para conjuntos de datos distintos.

• En el caso en que la desviación típica de una de las dos variables fuera igual a 0, la fórmula de R_{XY} resultaría en una indeterminación, ahora bien, ello ocurrirá en el caso en que todos los valores de esa variable fueran iguales (caso en el que tampoco se puede hablar propiamente de una variable).

Ejemplos del valor de R_{XY} obtenido para diferentes conjuntos de datos (Barón-López, 2005):





• La matriz de correlaciones constituye un tipo de representación en forma de tabla que permite mostrar la asociación existente entre un conjunto de variables por pares. Representadas las mismas variables en filas y columnas, cada casilla de la tabla muestra el valor de la correlación entre la variable fila y columna correspondientes. A tener en cuenta:

(1) Al tratarse de una matriz simétrica, algunos paquetes estadísticos sólo presentan una de las dos mitades de la matriz.

(2) En la diagonal de la matriz se suelen poner unos, dado que esas celdillas representan la correlación de una variable consigo misma.

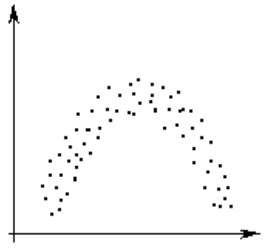
(3) Una matriz de correlaciones podría construirse con variables de cualquier tipo, no obstante, en la literatura suele plantearse sólo para variables cuantitativas.

Ejemplo de matriz de correlaciones a partir del rendimiento de un grupo de niños en 5 materias: música (A), matemáticas (B), lenguaje (C), deporte (D) y ciencias naturales (E).

	A	*B*	*C*	*D*	*E*
A	1				
B	0,23	1			
C	0,36	0,24	1		
D	−0,45	−0,34	−0,29	1	
E	0,07	0,38	0,17	0,13	1

• Algunos comentarios respecto a la interpretación del valor de R_{XY}:

(a) Un valor de R_{XY} (y lo mismo para la covarianza) nulo o próximo a 0 indica que no existe relación lineal entre ambas variables, lo cual no significa que no pueda existir algún otro tipo de patrón de relación entre ellas. (=> importante primero visualizar gráficamente la relación).



=> $R_{XY} \approx 0$

(b) La intensidad de la correlación entre 2 variables puede ser valorada siguiendo diferentes esquemas interpretativos, por ejemplo, algunos autores consideran que un valor absoluto de R_{XY} superior a 0,5 debe ser ya considerado como alto. Sin embargo, otros autores critican este modo de proceder y defienden que, a la hora de valorar un coeficiente de correlación, se debe tener en cuenta el contexto y la información ya existente relativa a la relación entre esas dos variables.

(c) La presencia de correlación entre dos variables no debe interpretarse como que existe una relación de causalidad entre ambas, por muy alto que nos haya dado el índice de asociación. Tal interpretación podría ser acertada en algunos casos pero, en otros muchos, puede representar un error grave. La existencia de correlación es condición necesaria, pero no suficiente, para establecer una relación de causa-efecto entre dos variables. Se deben satisfacer otras condiciones, precisamente, aquéllas cuya aparición se fuerza en la estrategia de recogida de datos asociada a los diseños de investigación experimental. Este tipo de diseño de investigación ya fue introducido en el primer capítulo de este temario, si bien, un tratamiento más en profundidad del mismo tendrá lugar en la asignatura de Diseños de Investigación.
(Este comentario se hace extensivo a todos los índices de asociación tratados en este tema).

> **Ejemplo**: Si se observa una relación positiva entre el nivel de consumo de frutos secos y la prevalencia del infarto de miocardio sería un error afirmar sin más que cuanto mayor sea el consumo de frutos secos, mayor será la ocurrencia del infarto de miocardio. En realidad, es el grado de obesidad -que suele estar relacionado con un alto consumo de frutos secos,

entre otros tipos de alimentos-, la variable que la evidencia científica parece señalar como una verdadera causa de la mayor prevalencia del infarto de miocardio.

(3) El coeficiente de determinación (R^2_{XY}):

• El coeficiente de determinación, al ser el cuadrado del coeficiente de correlación de Pearson, oscila entre 0 (independencia entre las variables) y 1 (relación lineal perfecta).

• Este índice, aparte de útil en otros contextos que se tratarán en temas posteriores, es también el más apropiado a la hora de comparar la relación lineal existente entre 2 pares (o más) de variables (o, también, en un único par de variables medido en 2 momentos temporales o con 2 grupos de sujetos distintos). Por otra parte, resulta inadecuado por razones teóricas inherentes al coeficiente de correlación de Pearson decir, por ejemplo, que la intensidad de la asociación entre *X* e *Y* es el doble que entre *M* y *N* si se han obtenido para ambos pares de variables un $R_{XY} = 0{,}8$ y un $R_{XY} = 0{,}4$, respectivamente. Sin embargo, sí que es posible tal interpretación a partir de los coeficientes de determinación, por ejemplo, si fuese $R^2_{AB} = 0{,}32$ y $R^2_{CD} = 0{,}16$.

Ejercicio 6: Se calculó el coeficiente de correlación entre las puntuaciones en dos tests *X* e *Y* en dos muestras de sujetos pertenecientes a dos países *A* y *B*. Para la muestra *A* se obtuvo un $R_{XY} = 0{,}3$ mientras que para la muestra *B* un $R_{XY} = 0{,}6$. ¿Qué se puede decir en términos comparativos acerca de la asociación entre *X* e *Y* en ambos países?

Ejercicio 7: A partir de los siguientes datos, presentados en el tema anterior, procedentes de un grupo de 16 sujetos sobre el 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, evaluar la asociación entre las dos variables tanto gráficamente (el diagrama de dispersión ya se realizó en el tema anterior) como analíticamente a través de los índices S_{XY}, R_{XY} y R^2_{XY}. (Se recomienda su obtención con un paquete estadístico como el SPSS).

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

Ejercicio 8: Se han obtenido con *SPSS* los siguientes resultados en el análisis de la relación entre las variables "Nº de años de escolarización" y "Puntuación prestigio profesional (escala de 0 a 100)". Calcular a partir de ellos el valor del coeficiente de correlación de Pearson entre ambas variables.

Correlaciones

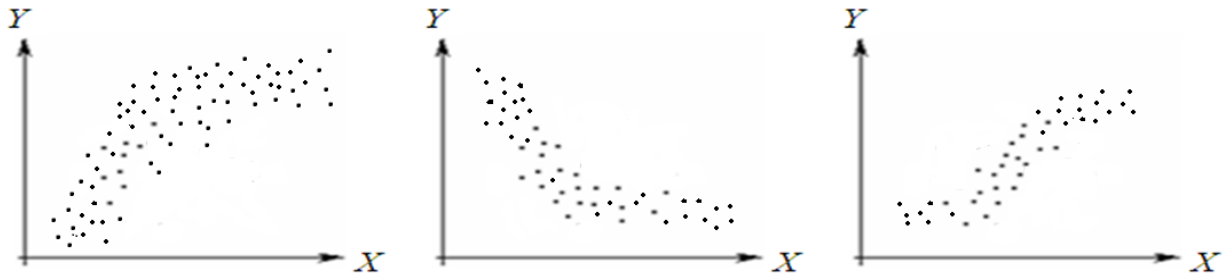
		Número de años de escolarización	Puntuación de prestigio profesional (1980)
Número de años de escolarización	Correlación de Pearson	1	¿?
	Sig. (bilateral)		,000
	Suma de cuadrados y productos cruzados	13436,719	28442,286
	Covarianza	8,904	20,115
	N	1510	1415
Puntuación de prestigio profesional (1980)	Correlación de Pearson	¿?	1
	Sig. (bilateral)	,000	
	Suma de cuadrados y productos cruzados	28442,286	241965,769
	Covarianza	20,115	170,759
	N	1415	1418

Estadísticos descriptivos

	Media	Desviación típica	N
Número de años de escolarización	12,88	2,984	1510
Puntuación de prestigio profesional (1980)	42,93	13,067	1418

(4) El coeficiente de correlación de Spearman (R_S)

• Cuando tengamos dos variables cuantitativas cuya relación diste de ser lineal –aunque sí monótona, ya sea creciente o decreciente (ver ejemplos gráficos a continuación)– es más adecuado aplicar el coeficiente de correlación de Spearman, el cual se basa en reconvertir los valores originales de las variables en valores de orden (al valor más bajo de cada variable se le asigna un 1, al siguiente un 2, y así sucesivamente).



• A continuación se muestra la fórmula para calcular R_S, donde D representa la diferencia, para cada sujeto, entre su valor de orden en una y en otra variable. En cualquier caso, no vamos a incidir aquí en su obtención, pues puede ser calculado fácilmente con un paquete estadístico (e.g., SPSS). La interpretación de R_S es exactamente la misma que la del coeficiente de correlación de Pearson.

$$R_S = 1 - \frac{6 \cdot \sum D^2}{N \cdot \left(N^2 - 1\right)}$$

• La obtención del coeficiente de correlación de Spearman resulta también recomendable en dos situaciones adicionales: (1) cuando tengamos variables cuantitativas en que en alguna de ellas o en ambas haya valores anómalos; (2) cuando para alguna de las variables, o para ambas, no se tenga claro que su escala de medida sea cuantitativa y se prefiera que sean consideradas como variables ordinales.

Ejemplo de obtención del R_S con SPSS para 6 pares de variables resultantes de combinar por pares 3 variables que representan, respectivamente, la valoración de las relaciones con los compañeros, con los profesores y con el personal de administración y servicios (PAS), hecha por una muestra de 174 estudiantes universitarios del Grado de Psicología. La escala de valoración en las 3 variables oscilaba entre 0 (Nada satisfactoria) y 10 (Muy satisfactoria). El analista decidió no considerar la escala de medida de estas variables como cuantitativas y, en consecuencia, decidió aplicar el coeficiente de correlación de Spearman a fin de valorar la asociación entre esas variables.

Correlaciones

			Relación con compañeros	Relación con profesores	Relación con PAS
Rho de Spearman	Relación con compañeros	Coeficiente de correlación	1,000	,192	,070
		Sig. (bilateral)	.	,011	,358
		N	174	174	174
	Relación con profesores	Coeficiente de correlación	,192	1,000	,401
		Sig. (bilateral)	,011	.	,000
		N	174	174	174
	Relación con PAS	Coeficiente de correlación	,070	,401	1,000
		Sig. (bilateral)	,358	,000	.
		N	174	174	174

Referencias

Barón-López, J. (2005). Bioestadística: métodos y aplicaciones. Apuntes y material disponible en http://www.bioestadistica.uma.es/baron/apuntes/

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Solanas, A., Salafranca, L., Fauquet, J. y Núñez, M. I. (2005). *Estadística descriptiva en Ciencias del Comportamiento*. Madrid: Thompson.