T. 8 – Estadísticos de asociación entre variables

1. Concepto de asociación entre variables

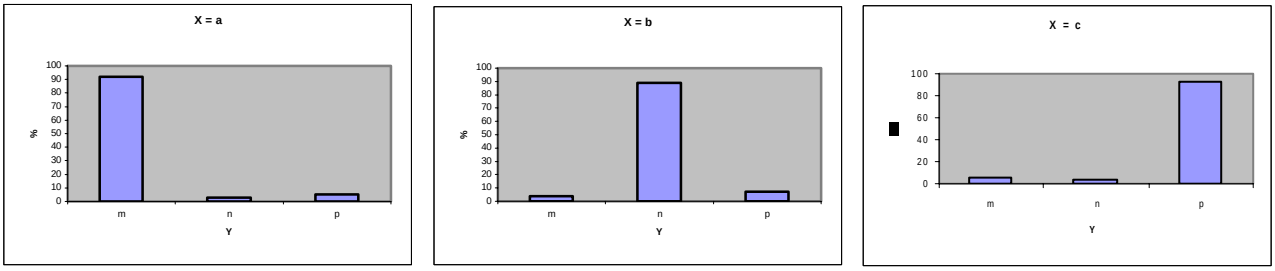
• El análisis estadístico de la asociación (relación, covarianza, correlación) entre variables representa una parte básica del análisis de datos en cuanto que muchas de las preguntas e hipótesis que se plantean en los estudios que se llevan a cabo en la práctica implican analizar la existencia de relación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

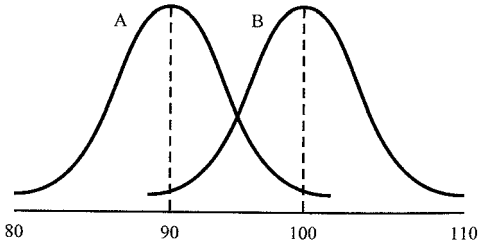
> A modo de **ejemplo** esquemático, si tenemos una variable X [*a, b, c*] y otra variable Y [*m, n, p*], de modo que los datos empíricos evidencian que las entidades que en X son *a* en Y tienden a ser *n* (o viceversa), que las que son *b* tienden a ser *p*, y que las que son *c* tienden a ser *m*, se pone de manifiesto que existe cierta 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, de las distribuciones de frecuencias (expresadas en %) de la variable *Y* para aquellos casos que en la variable *X* son 'a', 'b' y 'c', respectivamente:

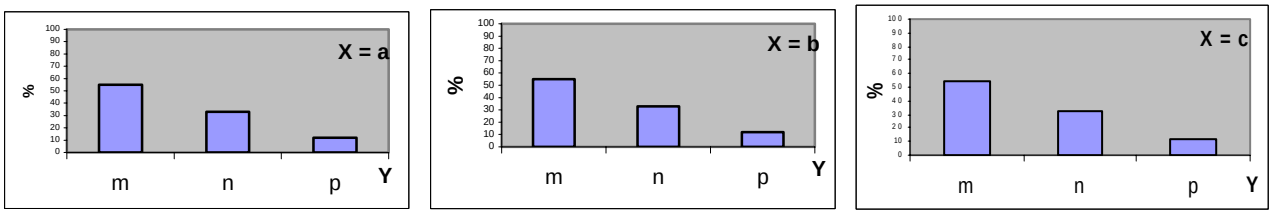


Otro **ejemplo** de la presencia de asociación entre 2 variables, una, la puntuación en un test de aptitud lingüística [0 a 150] y, la otra, la variable sexo [A: hombre; B: mujer]. Véase a continuación, para un conjunto de datos de estas dos variables, la diferencia existente entre las distribuciones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



• Complementariamente, se habla de <u>independencia</u> 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) al nivel máximo de relación entre las variables. Este grado máximo se plasmaría en 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")?

	−	±	+	Total
Sí	48	42	60	150
No	70	105	175	350
Total	118	147	235	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 condicionadas de una de las dos variables agrupada en función de los valores de la otra. Si no hay relación entre las variables estas distribuciones deberían ser iguales. Por ejemplo, podemos comparar las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a vivir en una residencia (48, 42, 60) y a no vivir en una residencia (70, 105, 175).

• Si nos fijamos en las distribuciones de frecuencias absolutas de "Estado de ánimo" condicionadas a los valores de "Vivir en residencia", 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 condicionadas en frecuencias absolutas si el número de casos difiere en las categorías de la variable condicionante o agrupadora.

• La asociación entre dos variables categóricas aparece más explícita en una tabla 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 total de columna. Ambas tablas permitirán llegar al mismo tipo de conclusiones respecto a la asociación entre las 2 variables.

• Si la relación entre las variables es asimétrica, la variable agrupadora o condicionante sería la que sea considerada la variable explicativa (predictora, independiente). Por ejemplo, en un estudio en que se evalúa 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 sería la variable explicativa, deberíamos comparar las distribuciones de la percepción de la influencia de la ciencia condicionadas al nivel de estudios, es decir, en cada categoría de nivel de estudios. En nuestro ejemplo sobre "Estado de ánimo" y "Vivir en residencia", dado que la relación es asimétrica y la variable explicativa es "Vivir en residencia" debemos comparar las distribuciones de "Estado de ánimo" condicionadas a "Vivir en residencia":

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

• En la tabla anterior, la comparación de las distribuciones de frecuencias relativas condicionales con la distribución marginal de la variable de respuesta, nos permitirá comprobar la existencia de asociación entre las dos variables y, en el caso de que exista, la naturaleza de la misma.

A modo de **ejemplo**, si no hubiera relación entre ambas variables, las distribuciones de frecuencias relativas de "Estado de ánimo" condicionadas a "Vivir en residencia" serían iguales a la distribución marginal de la variable "Estado de ánimo", esto es:

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

Como se puede comprobar, las distribuciones de frecuencias relativas de "Estado de ánimo" condicionadas a "Vivir en residencia" difieren bastante de las de la tabla anterior. Así, por ejemplo, se observa 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). Esto parece indicar que sí existe una relación entre ambas variables.

• Si la relación entre las variables es simétrica es indiferente qué variable se elige como agrupadora o 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 variables sea la variable explicativa, podríamos comparar, o bien, las distribuciones de frecuencias relativas de "Lugar de residencia" condicionadas a "Bachiller", o bien, las distribuciones de frecuencias relativas de "Bachiller" condicionadas a "Lugar de residencia".

Ejercicio 1: Analizar la asociación entre las dos variables dicotómicas siguientes: "Participación en un programa de intervención escolar [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.
(Nota: tener en cuenta cuál es la variable explicativa).

C_1	Buenos	Malos
Si	18	12
No	42	28

En un segundo colegio 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.

C_2	Buenos	Malos
Si	24	16
No	31	29

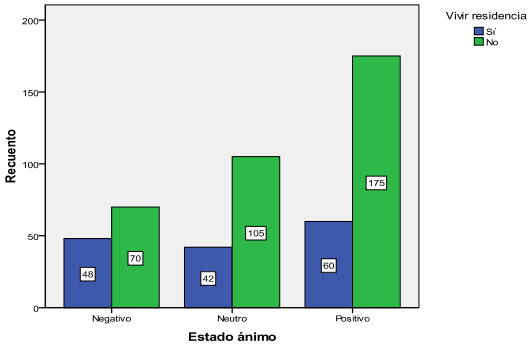
Finalmente, los datos recogidos en un tercer colegio se muestran resumidos en la siguiente tabla de contingencia. Analizar e interpretar la asociación existente entre ambas variables en este caso.

C_3	Buenos	Malos
Si	15	42
No	33	10

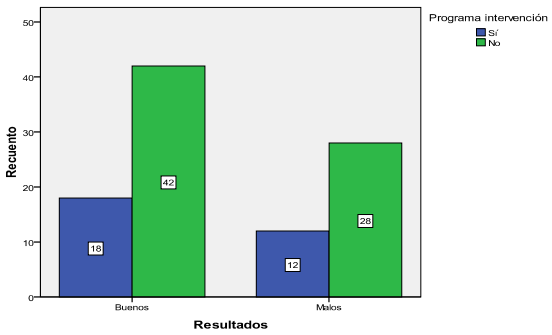
• El análisis gráfico de la asociación entre 2 variables categóricas puede intuirse a partir de un diagrama de barras conjunto de frecuencias absolutas si nos fijamos en la forma de las distribuciones condicionadas, si bien, resultará más fácil visualizar esta información si lo que se representa son frecuencias relativas condicionadas, pues así se elimina el efecto del posible distinto tamaño de los subgrupos. El diagrama de rectángulos partidos obtenido a partir de las frecuencias relativas condicionadas puede resultar también apropiado para evaluar esa asociación.

- **Ejemplos** para los datos de las variables "Estado de ánimo" y "Vivir en residencia", así como para los datos de las variables "Programa de intervención" y "Resultados académicos" en el Colegio 1:

- Diagramas de barras de frecuencias absolutas:

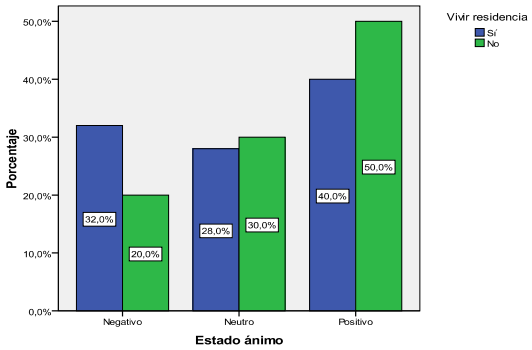


"Estado de ánimo" condicionada a "Vivir en residencia"

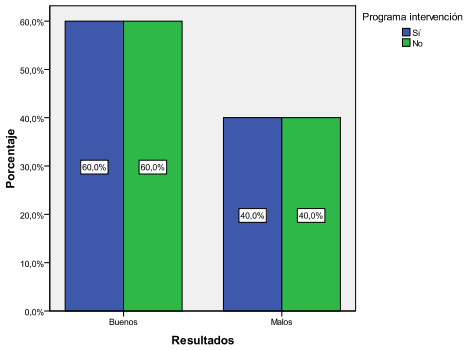


"Resultados académicos" condicionada a "Programa intervención"

- Diagramas de barras con frecuencias relativas (%) condicionadas:

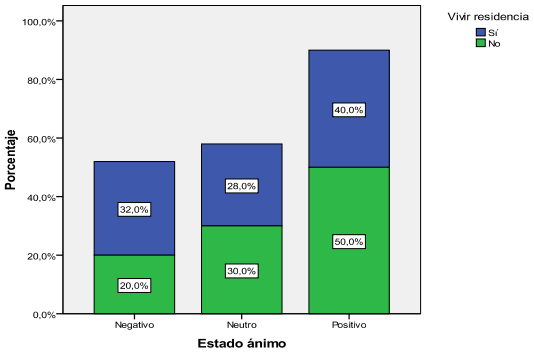


"Estado de ánimo" condicionada a "Vivir en residencia"

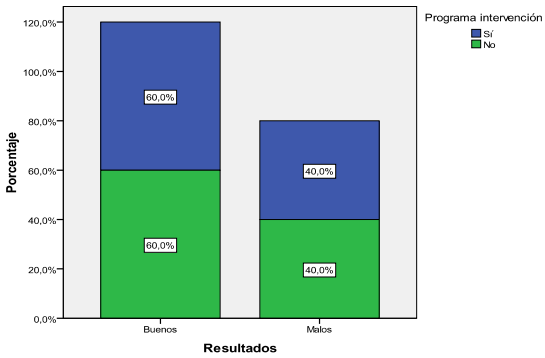


"Resultados académicos" condicionada a "Programa intervención"

- Diagramas de rectángulos partidos con frecuencias relativas (%) condicionadas:



"Estado de ánimo" condicionada a "Vivir en residencia"



"Resultados académicos" condicionada a "Programa intervención"

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

• Existen diferentes índices estadísticos orientados a resumir de forma cuantitativa la asociación entre dos variables categóricas. Aquí nos vamos a centrar en los dos siguientes:

- El índice *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 dos variables son independientes, siendo mayor que 0 cuando exista asociación entre ellas, tanto mayor cuanto más intensa sea esa correlación. Ahora bien, no tiene un límite máximo, lo cual supone una dificultad a nivel interpretativo.

· Sí que puede utilizarse para comparar la asociación entre variables en tablas de contingencia del mismo tamaño (I x J) y con el mismo n.

· 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.

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

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

· 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 esta circunstancia principalmente en la práctica, 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: $\varphi \leq 0{,}3$ => nivel bajo de asociación; $0{,}3 < \varphi \leq 0{,}5$ => nivel medio de asociación; $\varphi > 0{,}5$ => nivel alto de asociación.

- El coeficiente de contingencia de Cramer (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, ello indicará mayor intensidad en la asociación de las variables.

Ejercicio 3: Obtener los índices χ^2, φ y V de Cramer a partir de las tres tablas de contingencia presentadas anteriormente para los tres colegios y, también, para las variables "Estado de ánimo" y "Vivir en residencia".

2.2. El caso de una variable categórica y una cuantitativa

• De nuevo, el análisis de este tipo de asociación supone comparar las distribuciones condicionales de una variable para los distintos valores que toma la otra. Normalmente, se suele tomar como condicionada a la cuantitativa y como condicionante a la categórica, si bien, las conclusiones a las

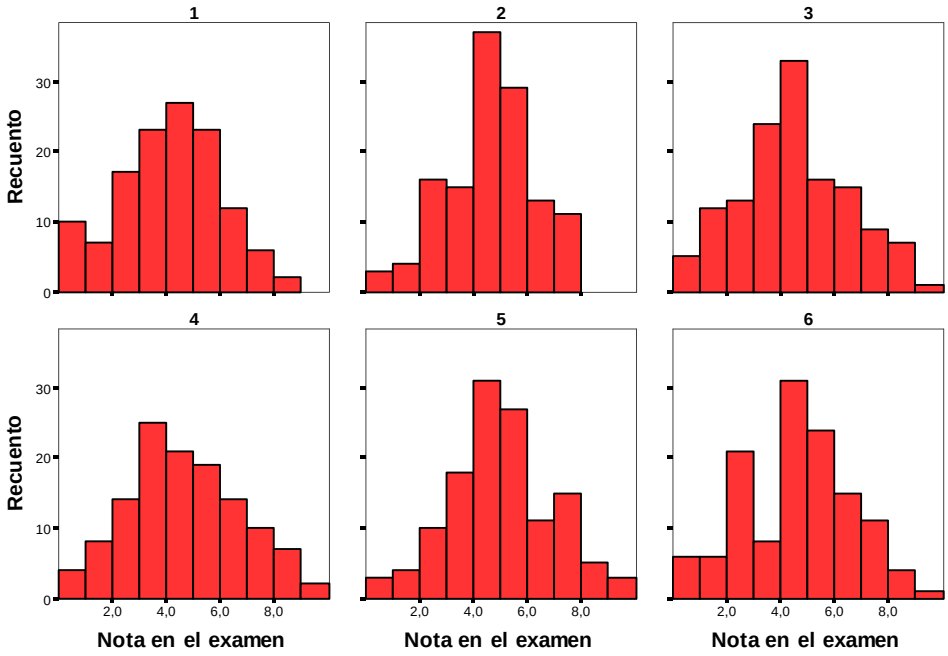
que llegaríamos serían las mismas si se hiciese al revés. Si no hay diferencias entre las distribuciones condicionales, ello indicará que no hay asociación entre ambas variables.

- **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:
-

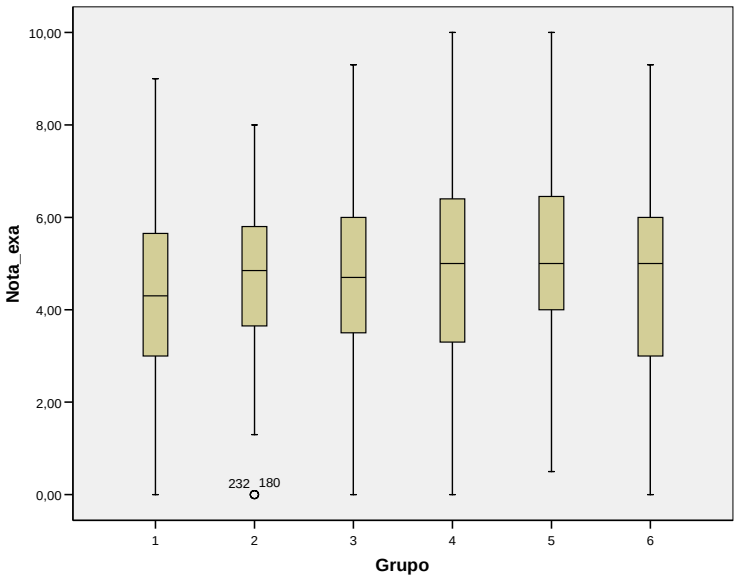
Grupo 1		Grupo 2		Grupo 3		Grupo 4		Grupo 5		Grupo 6	
x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i	x_i	n_i
0	2	0	3	0	2	…	…	…	…	…	…
,3	3	1,3	1	,5	1						
,5	1	1,8	1	,8	2						
,8	3	2,0	2	1,3	2						
1,0	1	2,3	4	1,5	2						
1,3	1	2,5	2	1,8	1						
1,5	2	2,8	5	2,0	7						
1,8	2	2,9	1	2,3	3						
2,0	2	3,0	4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

- Dada la dificultad que puede representar comparar las distribuciones condicionales de una variable cuantitativa, se puede recurrir a representaciones gráficas que faciliten la realización de este tipo de comparación. A modo de ejemplo, las dos siguientes obtenidas para los datos anteriores con el

paquete estadístico SPSS o, también, el diagrama de dispersión presentado en el capítulo anterior para estos mismos datos:



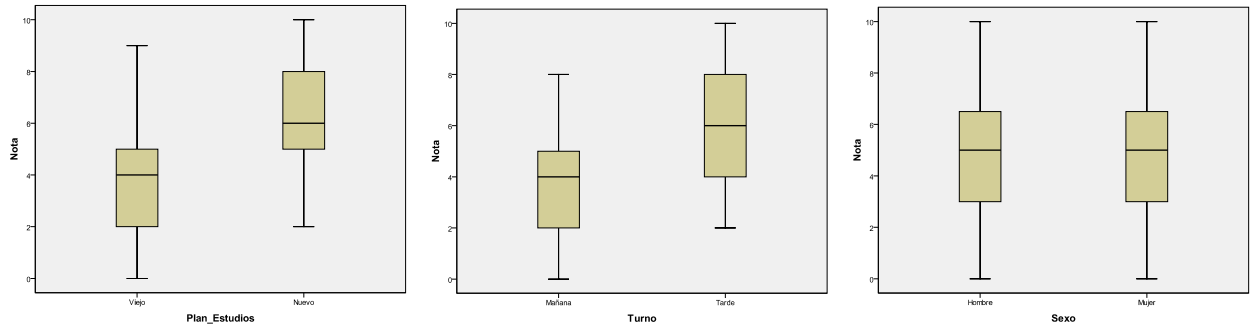
Ejemplo de diagrama de caja y bigotes con la distribución de la variable "Nota en un examen de una asignatura" condicionada a la variable "Grupo en el que se está matriculado":



• Estos gráficos nos permiten comparar el grado de solapamiento (coincidencia) de las distribuciones condicionales. En general, cuanto mayor sea el solapamiento, menor será la intensidad de la

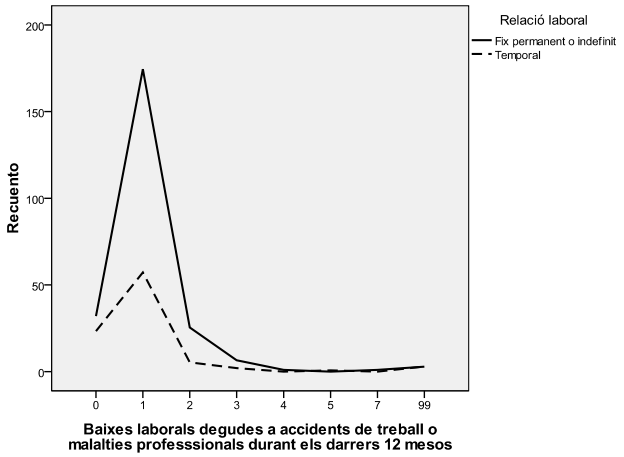
asociación entre las dos variables y, viceversa, cuanto menor sea el solapamiento, mayor será el tamaño del efecto de la relación. En el ejemplo anterior existe bastante solapamiento entre las 6 distribuciones condicionales, poniendo de manifiesto una escasa relación entre ambas variables.

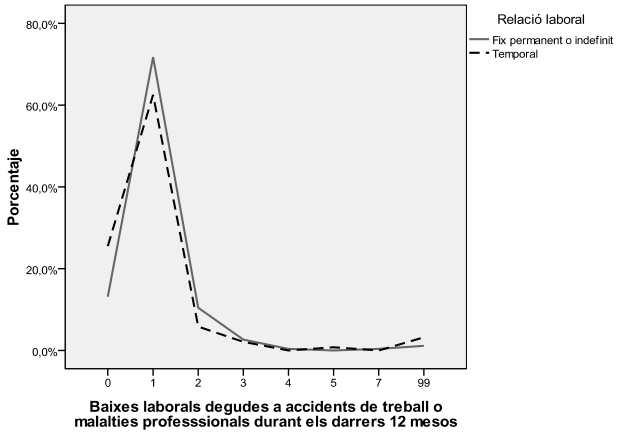
Ejemplo de diferente intensidad de asociación en 3 pares de variables (cada par constituido por una variable categórica dicotómica y una misma variable cuantitativa) basado en la visualización de las distribuciones condicionales mediante diagramas de caja y bigotes. Obsérvese como la relación entre las variables aumenta desde el gráfico de la izquierda (mayor solapamiento) hasta el de la derecha (menor solapamiento).



• Otro gráfico que se suele utilizar para comparar el grado de solapamiento de las distribuciones condicionadas, y que ya se ha visto en temas anteriores, es el polígono de frecuencias superpuesto.

Ejemplo de polígono de frecuencias superpuesto de las distribuciones de la variable nº de bajas laborales en un grupo de trabajadores durante los últimos 12 meses condicionadas al tipo de relación laboral de los trabajadores. Recuérdese que cuando el tamaño de los grupos de la variable condicionante es desigual no se deben representar las frecuencias absolutas sino las 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 puede resultar engañoso al dar la sensación de que ambas distribuciones son bastante diferentes, sin embargo, este efecto es debido a que el nº de trabajadores fijos es muy superior al de trabajadores temporales. En el gráfico de la derecha, donde se representan las distribuciones de frecuencias relativas condicionadas, se puede comprobar que ambas distribuciones son, en realidad, bastante similares.

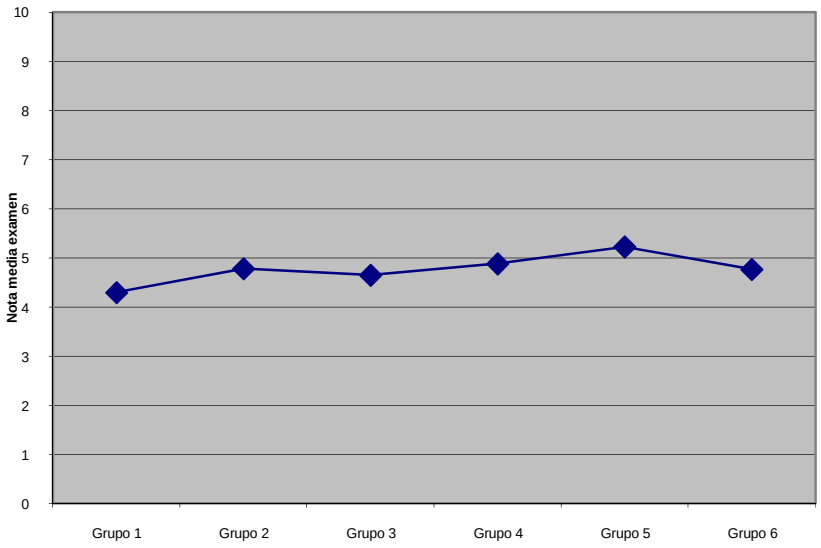




• Mientras que los gráficos anteriores representan la distribución de frecuencias de la variable cuantitativa para cada modalidad de la variable categórica, los gráficos que aparecen a continuación no representan las distribuciones de frecuencias como tal, sino los valores de determinados estadísticos que resumen, o bien la tendencia central, o bien la tendencia central y dispersión de dichas distribuciones condicionadas.

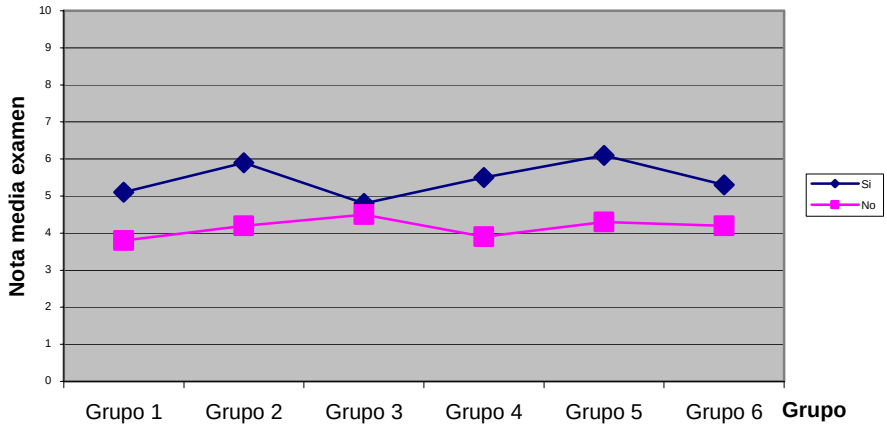
- El diagrama de medias:

Ejemplo de diagrama de medias de la variable "Nota en un examen de una asignatura" condicionada a la variable "Grupo en el que se está matriculado [Grupo 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.

Ejemplo de diagrama de medias de la variable "Nota en un examen de una asignatura" condicionada a las variables "Grupo en el que se está matriculado [1 a 6]" y "Asistencia regular a las clases [Si, No]":



• A la hora de comparar las distribuciones condicionales de la variable cuantitativa para cada uno de los valores de la variable categórica, la atención en la literatura estadística se suele centrar en un aspecto específico de las mismas, la tendencia central. Más concretamente, en el grado de discrepancia de las medias aritméticas de esas distribuciones condicionales (cuanto mayor la discrepancia, mayor la intensidad de la relación), tal como se va a poner de manifiesto en los índices estadísticos que a continuación se presentan, todos basados en las diferencias entre las medias en los subgrupos definidos por la variable categórica.

• Existen diferentes índices estadísticos orientados a cuantificar la intensidad de la asociación entre una variable categórica y una variable cuantitativa. Aquí nos vamos a centrar en los siguientes:

(1) Dada una variable categórica *X* dicotómica [a, b] y una variable cuantitativa *Y*, el índice de asociación *d de Cohen* se obtiene a través de la siguiente expresión:

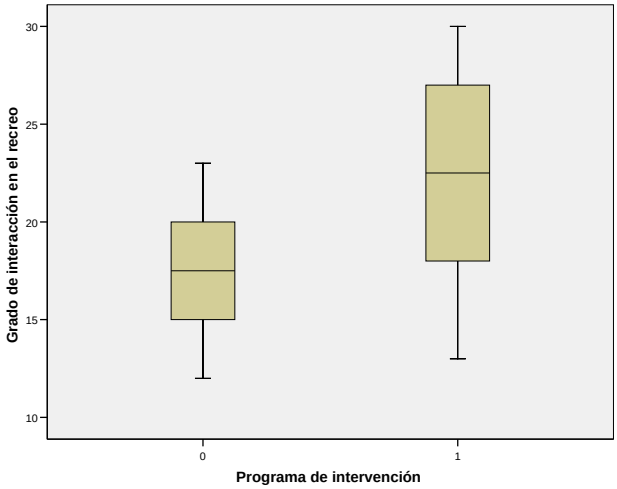
$$d = \frac{\overline{Y}_a - \overline{Y}_b}{s_Y}$$

Los valores que puede tomar *d* no están acotados a un rango, pudiendo ser tanto positivos como negativos. Si las dos variables consideradas son independientes entonces *d* será igual a 0, mientras que cuanto mayor sea la asociación entre ellas, mayor será el valor de *d* en términos absolutos. Cohen sugiere las siguientes normas interpretativas, aunque el propio autor afirma que se deben utilizar sólo en el caso que no se tenga ningún criterio sustantivo que sirva de base

interpretativa: valores absolutos de *d* entre 0,2 y 0,5 indicarían una intensidad de la asociación (tamaño del efecto) baja; entre 0,5 y 0,8 media; mientras que a partir de 0,8, alta.

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. Analizar la asociación entre ambas variables e interpretar los resultados

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) Como caso particular del índice de asociación *d de Cohen* , el índice d_s se obtendría cuando la variable categórica determina dos momentos temporales, esto es, aquellos casos en que una misma variable cuantitativa es medida en un mismo grupo de sujetos antes y después de

determinado evento, situación típica en las recogidas de datos que vienen guiadas por un diseño de medidas repetidas (o intra-sujetos). Fórmula de d_s:

$$d_S = \frac{\overline{DIF}}{s_{DIF}}$$

donde el numerador representa la media de las diferencias para cada sujeto entre la variable 'después' y la variable 'antes', y el denominador la desviación típica de esas diferencias.

Ejercicio 5: Tomando como punto de partida el caso planteado en el ejercicio anterior, supóngase que a los 10 estudiantes a los que se aplicó el programa de intervención para favorecer la interacción social se les midió también, previó al tratamiento, su grado de interacción en la hora de recreo (nº de minutos en que se participa en actividades con otros compañeros). Analizar la relación entre la variable "Grado de interacción" y "Momento temporal en que se realiza la medición de la anterior variable [pre- y post- tratamiento]" e interpretar los resultados obtenidos.

ID	*GI-pre*	*GI-post*
1	21	22
2	14	13
3	25	27
4	21	19
5	21	23
6	27	29
7	20	16
8	30	30
9	21	24
10	20	18

(3) 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 s_Y 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 6: En un ejemplo previo se analizó gráficamente 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. Analizar esos mismos datos haciendo uso del índice *f* de Cohen, teniendo en cuenta los siguientes resultados parciales:

$\overline{Y}_1 = 4,30;\ \overline{Y}_2 = 4,78;\ \overline{Y}_3 = 4,65;\ \overline{Y}_4 = 4,88;\ \overline{Y}_5 = 5,23;\ \overline{Y}_6 = 4,76$

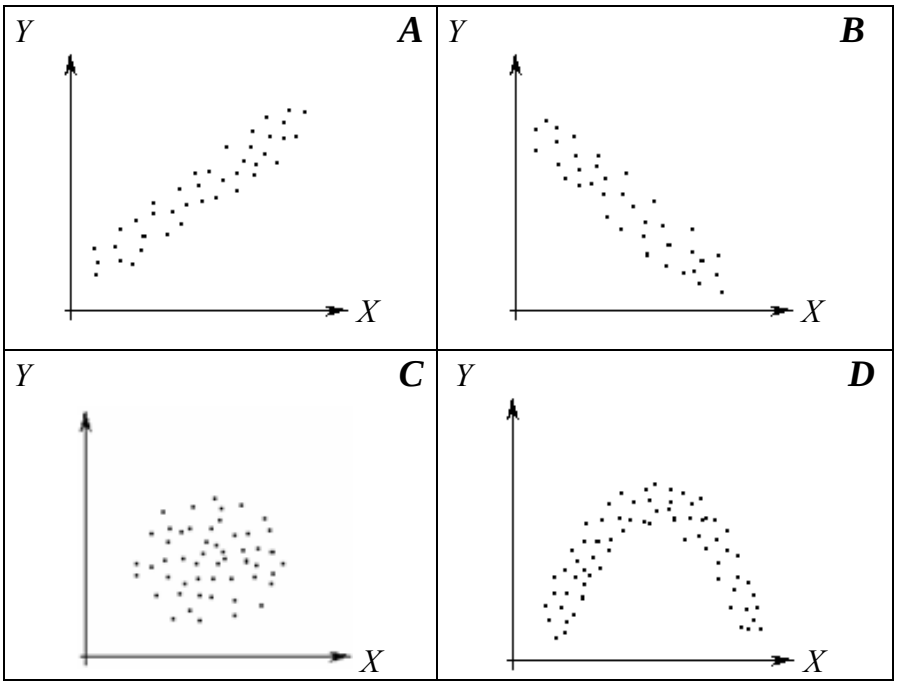
$n_1 = 127;\ n_2 = 128;\ n_3 = 135;\ n_4 = 124;\ n_5 = 127;\ n_6 = 127$

$\overline{Y} = 4,76;\ s_Y = 1,97$

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, es más habitual 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 arriba, 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.

• 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 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: $s_{XY} = SP_{XY} / n$

- Desarrollando algebraicamente la fórmula de la covarianza se puede llegar a una fórmula que se considera más conveniente cuando el cálculo de la misma se ha de realizar de forma manual:

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

Ejemplo para las variables *Calificaciones en música* (*X*) y *Calificaciones en matemáticas* (*Y*) obtenidas por un grupo de 10 niños.

X	*Y*	*X*Y*
5	6	30
7	8	56
8	7	56
5	6	30
9	10	90
4	5	20
5	5	25
5	7	35
7	6	42
8	9	72
$\bar{X} = 6{,}3$	$\bar{Y} = 6{,}9$	Σ (*X*Y*)= 456

$$S_{XY} = \frac{\sum X_i Y_i}{n} - \overline{X}\overline{Y} = \frac{456}{10} - (6{,}3 \cdot 6{,}9) = 2{,}13$$

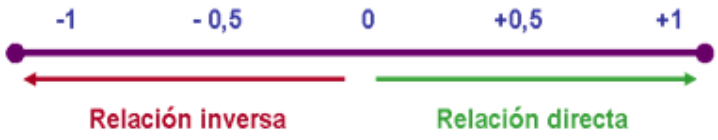
- La covarianza puede tomar valores tanto positivos como negativos. A nivel interpretativo, un mayor valor de la covarianza en valor absoluto indicará una relación lineal más intensa entre las dos variables. Un valor positivo pone de manifiesto una relación lineal directa; uno negativo, una relación 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 de Pearson:

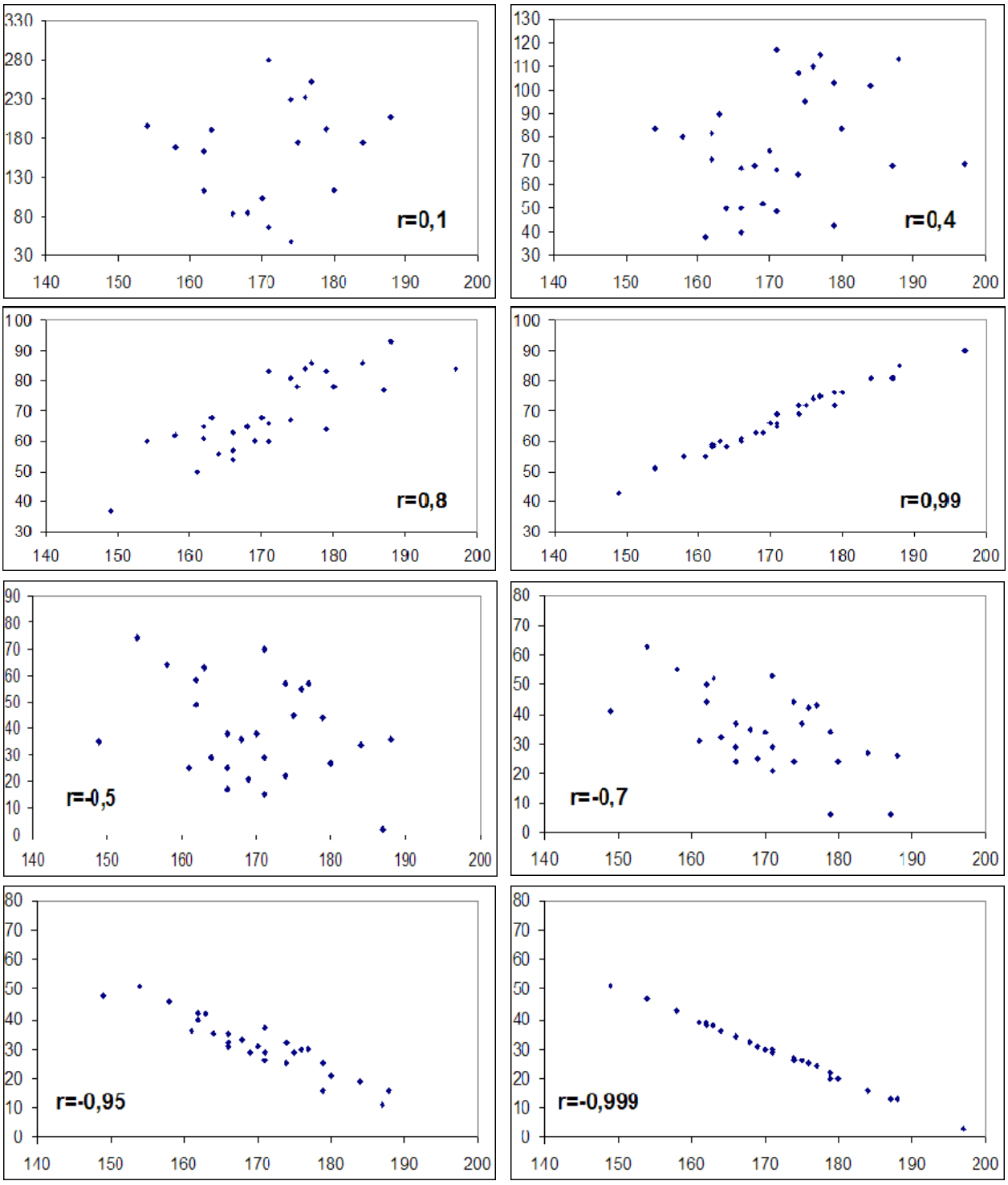
$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$$

– 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 sino de una constante.

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 un conjunto de materias:

Rendimiento en música (*A*)

Rendimiento en matemáticas (*B*)

Rendimiento en lenguaje (*C*)

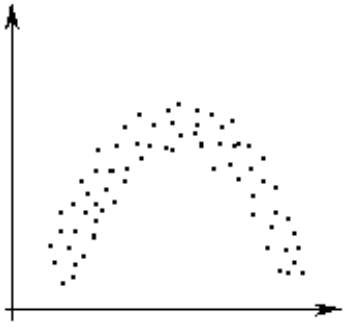
Rendimiento en deporte (*D*)

Rendimiento en 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) En ocasiones la presencia de correlación entre 2 variables puede ser espuria y venir motivada, en realidad, por una tercera variable que está relacionada con ambas. Por ejemplo, si se observa una relación entre el consumo de frutos secos y el infarto de miocardio (a mayor consumo de frutos secos mayor probabilidad de padecer infarto), ésta puede ser espuria debido a que el elevado consumo de frutos secos está asociado a la obesidad, que es la variable que sí está relacionada con el infarto de miocardio.

(d) La presencia de correlación entre 2 variables no debe interpretarse como que existe una relación de causa-efecto entre ambas. Es cierto que el que una variable *X* sea causa de una variable *Y* implica que exista correlación entre ambas, pero lo contrario no es necesariamente cierto; para ello es necesario que se satisfagan otros requisitos asociados a la forma en que se plantea la recogida de datos que se verán en la asignatura de Diseños de Investigación.
(Este comentario se hace extensivo a todos los índices de asociación tratados en este tema).

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

$$R^2_{XY} = r^2_{XY}$$

- 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 $r_{MN} =$

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 7: 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 8: 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 (s_{XY}, r_{XY}, R^2_{XY}).

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

Referencias:

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

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