

Estadística Descriptiva Bivariada (I): La correlación de Pearson- r_{xy}

Eduardo López
Dpto. MIDE. UCM

Introducción

Tal como se dijo en temas anteriores, la Estadística Descriptiva se dividía en univariada y bivariada. De la primera ya se ha hablado. En efecto, se ha estudiado una distribución de datos, que son resultado de medición en una variable. De dicha variable, medida en una muestra, se han estudiado las puntuaciones individuales y las de grupo en sus distintas posibilidades. De las puntuaciones referidas a individuos se ha hablado de la puntuación directa, puntuación diferencial, percentil y puntuación típica. Los datos referidos a grupos se han analizado tanto gráfica como numéricamente. Se han estudiado las medidas de tendencia central (media aritmética, media geométrica, media cuadrática, media armónica), las medidas de posición (mediana, percentiles, deciles, cuartiles) y las medidas de frecuencia (moda). Asimismo se ha prestado especial atención a las medidas de variabilidad (recorrido, desviación media, varianza, desviación típica y coeficiente de variación).

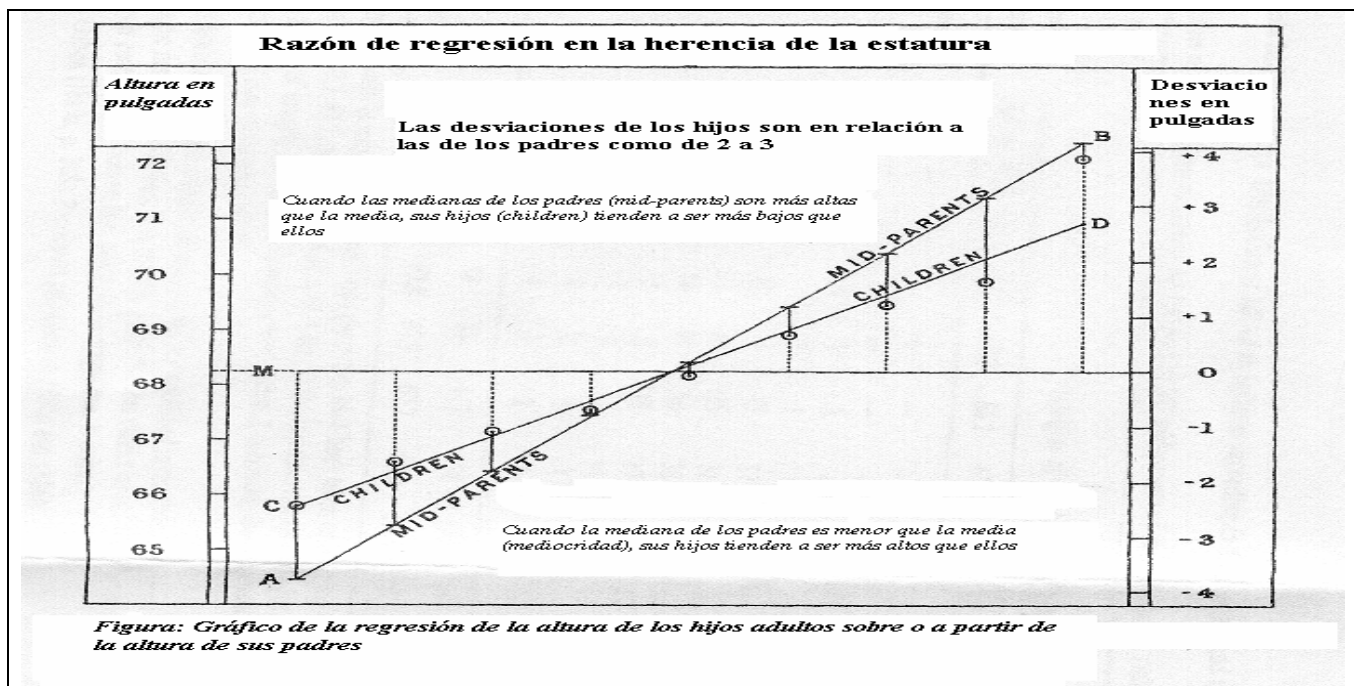
Pero en cualesquiera de los casos dichas puntuaciones se refieren, tal como se ha dicho, a una variable medida una sola vez a una muestra. Ahora se hablará de las distribuciones bivariadas. Se verá de dónde vienen los datos correspondientes a dos distribuciones, que son el objeto de la Estadística descriptiva bivariada. Se tabularán los datos, se obtendrá la covarianza, se hallará la correlación lineal entre dos variables representando gráfica y numéricamente el coeficiente de correlación de Pearson y será interpretado. Se calcularán otros índices de correlación y asociación y se identificará cuándo se utiliza cada uno de ellos.

Francis Galton, como fruto de su experiencia, se dio cuenta de que los hijos de padres altos tendían a ser altos pero como *media* eran más bajos que éstos. Igualmente, los hijos de padres bajos tendían a ser bajos pero como *media* no tan bajos como sus padres.

Es sabido que Francis Galton hizo estudios antropométricos interesándose por la predicción de las características físicas de los descendientes a partir del conocimiento de las de sus progenitores. Dentro de ellas quiso estudiar la relación entre la altura de los padres y de sus hijos. En concreto examinó 928 hijos adultos de 205 padres; es decir, se examinaron las estaturas de 205 padres, cada uno de los cuales tenía de promedio entre 4 y 5 hijos. Se trataba, pues, de comprobar la relación entre la altura de los padres para cada altura *media* de los hijos. Se incluyen los datos que él analizó (Galton, 1889, 208; Anders, 1998, 610/1) y que incluyó en una tabla y gráfico de distribución de 928 hijos adultos según su estatura y la estatura de la *media* (mediana) de sus padres.

Altura-Mdn padres (inchs)	Alturas de los hijos adultos														Número total		Media na
	<	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>	Hijos	Padres	
>72.5											1	3		4	5(4)		
72.5							1	2	1	2	7	2	4	19	6	72.2	
71.5				1	3	4	3	5	10	4	9	2	2	43	11	69.9	
70.5	1		1	1	1	3	12	18	14	7	4	3	3	68	22	69.5	
69.5			1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1		7	11	16	25	31	34	48	21	18	4	3	219	49	68.2	
67.5		3	5	14	15	36	38	28	38	19	11	4		211	33	67.6	
66.5		3	3	5	2	17	17	14	13	4				78	20	67.2	
65.5	1		9	5	7	11	11	7	7	5	2	1		66	12	66.7	
64.5	1	1	4	4	1	5	5		2					23	5	65.8	
<64.5	1		2	4	1	2	2	1	1					14	1		
Tots	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	
Mdn			66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0					

A partir de la contemplación de la tabla se observa que el alejamiento de la *media* de la altura de los hijos es menor que la de los padres en los extremos de la distribución en una proporción de un tercio aproximadamente. Dicho de otro modo, los hijos de padres altos o bajos como *media* tendían a *regresar* a la *media*, es decir, ser más bajos de promedio los hijos de padres altos y más altos los de padres bajos. Si hubiera que predecir la estatura de los hijos a partir de la estatura de los padres se podría afirmar que es menos extrema, esto es, más próxima a la *media*. Es decir, regresan a la *media*.



A partir del conocimiento de este hecho se pueden elevar ciertos principios o leyes con un carácter de mayor o menor generalidad. Y se puede, y esto es lo que interesa aquí, traducir a datos empíricos la supuesta relación entre estatura de los padres y estatura de los hijos, en último término, relaciones entre las características físicas humanas, que es lo que pretendía Galton.

El estudio de las correlaciones es de una importancia extrema en el campo de las ciencias. En concreto, gran parte de la investigación en las ciencias sociales, y el de la educación no es una excepción, se dedica al estudio de la *variabilidad* en alguna variable o realidad entre grupos, individuos o culturas a lo largo del tiempo y a través de diversos entornos y lugares. Sin variabilidad no hay ciencia experimental. Es la variabilidad la que despierta la curiosidad de todos y mueve a algunos a investigar cuál es su origen en orden a *explicarla*. Para intentar explicar la variabilidad de tales variables se recurre al estudio de las covariaciones de éstas con otras variables. Es decir, el punto de arranque de la ciencia es la constatación de la variabilidad de las variables en primer término y su co-variabilidad después.

Afirma Kerlinger (1975, 62) que las relaciones son la esencia de la ciencia. Y las relaciones en la ciencia lo son entre clases o conjuntos de objetos, pues no puede conocerse la relación entre variables midiendo solamente a un sujeto. Es solamente a partir del estudio de la relación entre variables o fenómenos como surgen hipótesis, que luego se confirmarán. Y es precisamente a partir del conocimiento de la relación entre variables como puede llegarse a formular predicciones de una a partir de otra.

Pero tampoco se ha de pensar que la relación se establece solamente entre dos variables. La relación se puede establecer entre muchas variables o fenómenos. Es muy frecuente en la investigación tratar con muchas variables, dado que los fenómenos no son simple relación entre dos. En este momento las relaciones que se examinarán son las existentes entre dos variables.

La correlación

Cuando se dice que las relaciones que se examinarán son las existentes entre *dos variables* no se quiere decir que sean dos variables necesariamente sino que, además de correlacionar dos variables, también puede ser una sola variable *dos veces*. Es decir, el método que pone en relación esas variables es el método de regresión, a través del cual se ponen en relación dos dimensiones de datos, esto es, datos emparejados, y se utiliza fundamentalmente en tres situaciones, a saber, cuando se desea saber si existe relación:

1. Entre dos variables medidas a una misma muestra en orden a comprobar si existe relación entre dichas dos variables;
2. De una variable en dos grupos de sujetos -supuestamente relacionados por algún vínculo ya sea genético o ambiental- en orden a comprobar el grado de relación de dichas muestras en esa variable; y
3. En una variable medida en dos momentos distintos a una misma muestra en orden a comprobar el grado de estabilidad de la variable al cabo del tiempo, ya sea no

mediando algún tipo de intervención o cuando media alguna intervención, en nuestro caso educativa, entre los dos momentos.

Esto mismo se puede decir afirmando que en las correlaciones se estudia:

- 1) Una muestra medida en:
 - dos variables para ver si hay relación entre ellas;
 - una variable en dos momentos distintos, para ver su grado de estabilidad.
- 2) Dos muestras en una variable para ver si entre ellas en esa variable hay algún tipo de comunidad, sea genética y/o ambiental.

Así, pues, en toda correlación existen dos puntuaciones de datos, X_i e Y_i , ya provengan, como se ha dicho, de dos variables medidas a un mismo grupo, de una variable medida a dos grupos de sujetos supuestamente relacionados o de una sola variable medida a un mismo grupo de sujetos en dos momentos distintos. En cualquier caso, en las expresiones que se utilizarán se hablará ordinariamente de correlación *entre dos variables* sin hacer distinción sobre si son dos variables medidas en el mismo grupo o una misma variable medida en dos muestras distintas supuestamente relacionadas o la misma muestra medida en dos momentos distintos.

Representación gráfica

¿Qué es existir correlación entre dos variables? La correlación se puede expresar de dos formas, *gráfica* y *numéricamente*. Si se contemplan dos distribuciones de datos en una variable X_{ij} y en una variable Y_{ij} , respectivamente, se pueden extraer de su contemplación algunas conclusiones. Véanse las puntuaciones de las columnas 1 y 2 de la tabla de datos insertada más adelante, sean cuales fueren las variables medidas. En el campo de la educación y de la enseñanza las variables de las que se habla ordinariamente son del rendimiento en distintas áreas, de la actitud hacia determinados campos, del autoconcepto académico, del tiempo en la tarea, del nivel de aspiraciones, de los hábitos de estudio, de la inteligencia general, de las aptitudes específicas, de los rasgos de personalidad, de los valores, ...

Si se contemplan, se decía, los datos de las dos primeras columnas y se observan con cierto detenimiento, se podrá concluir que entre las parejas de datos de las dos distribuciones parece haber cierta concomitancia: Si estamos hablando de alumnos, parece que los que obtienen puntuación baja en una variable también en su mayoría son alumnos de puntuación baja en la otra, del mismo modo que los que obtienen puntuación alta en una suelen ser los que obtienen resultados altos en la otra, ... Es decir, parece existir una relación directamente proporcional: Si hay bajos resultados, si medios y si altos en una variable también parece haber rendimientos bajos, medios o altos en la otra. Si esa relación se representara gráficamente resultaría aproximadamente la figura **(b)** de las cuatro que se indican, que expresa una relación positiva, dado que una relación directamente proporcional se representa aritméticamente con signo positivo.

Si manipulamos las dos distribuciones emparejando sus datos en orden creciente, tal como se indica en las columnas 3 y 5, se comprueba que el sujeto con puntuación más baja en X_{ij} también

es el sujeto con la puntuación más baja en Y_{ij} , el más mediano en una igualmente es medio en la otra y el más alto en una es el más alto en la otra. En esas condiciones esa relación sería perfecta o casi perfecta.

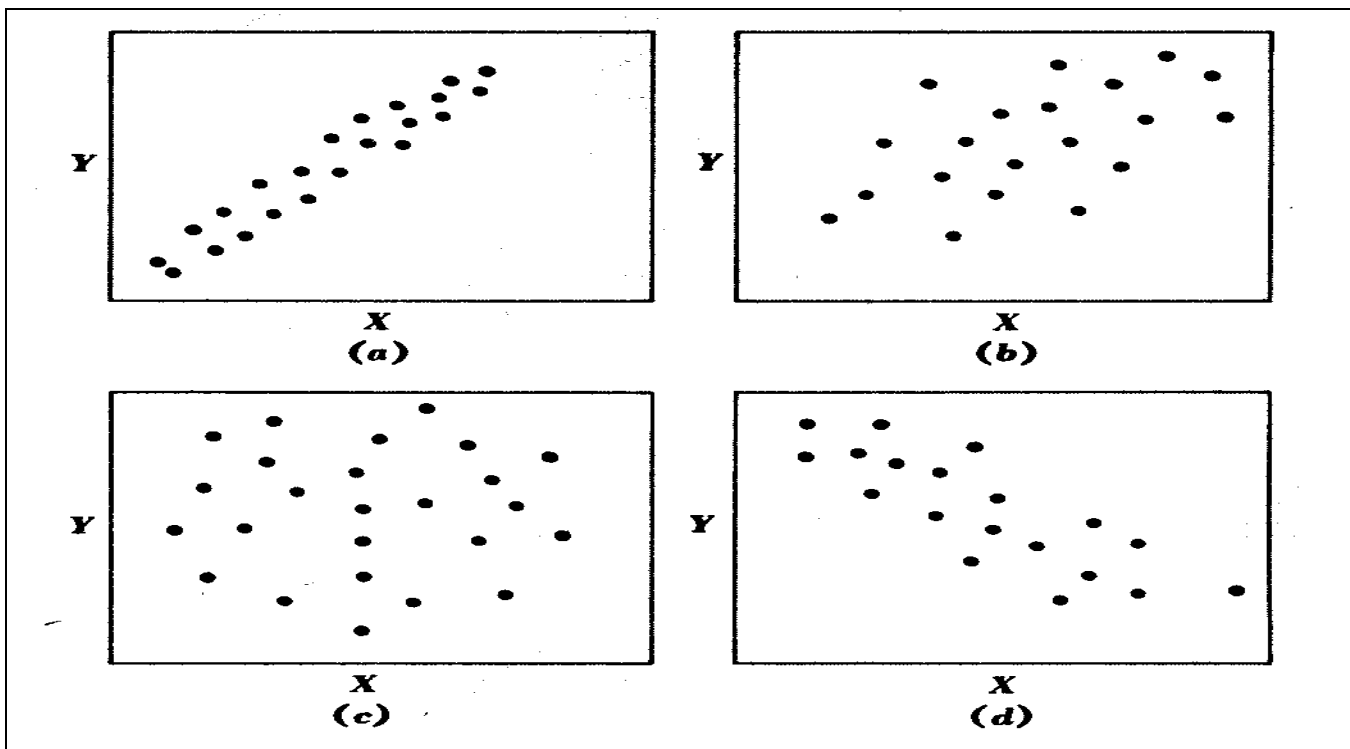


Figura: Representación de una correlación alta y positiva (a), *media* y positiva (b), nula (c) y *media*-alta negativa (d).

Si esa relación se representara gráficamente, resultaría aproximadamente la figura (a). Idéntica puntuación se obtendría con los datos de las columnas 4 y 6 dado que son idénticas a las anteriores, únicamente que en orden decreciente.

Contémpense ahora las columnas 3 y 6, así como las columnas 4 y 5. De la simple contemplación de ambas parejas de columnas se extraen conclusiones claras. Así, en la primera pareja (columnas 3 y 6) se observa que mientras los datos de la primera columna son ascendentes, son descendentes en la segunda. Es decir, existe una relación inversamente proporcional.

Dicho de otro modo, cuanto más baja puntuación tenga un alumno en una variable más alta la tiene en la otra. Otro tanto ocurre en la segunda de las parejas de columnas (4 y 5). Esta relación inversamente proporcional se representa aritméticamente en forma de signo negativo. La distribución gráfica es similar a la figura (d).

Sin embargo, de estas comparaciones es preciso dejar sentado algo bien claro. Sea que contemplemos las columnas 3 y 5, las 4 y 6, las 3 y 6 ó las 4 y 5, incluso las 1 y 2, en todos los casos existe relación; menos perfecta, como se verá en las dos primeras columnas (1 y 2), pero en todos los casos relación. Es decir, el signo de la relación afecta al tipo de relación, pero no a la existencia. No ocurre así, sin embargo, en los datos de la figura (c). En ella la representación de

los puntos, que expresan la confluencia de dos puntuaciones, no sigue una regla determinada sino simplemente un encuentro aleatorio de parejas de puntuaciones.

Y, en efecto, para poder llegar a obtener una representación de distribuciones de ese tipo la forma de hacerlo consiste en la elección o formación de parejas de datos sino ninguna regla, sino simplemente al azar. No existe relación ni directa ni inversamente proporcional sino simplemente no existe relación.

Así, pues, las correlaciones se pueden representar gráfica y numéricamente. De la segunda forma se hablará a continuación.

Cálculo aritmético

Supuestos al uso de r_{xy} . Son varios los supuestos que subyacen al uso de la correlación de Pearson. En concreto el coeficiente de correlación, se dirá en una tabla, se calcula cuando se dan algunas condiciones -no se indican todas: Las dos variables son de naturaleza continua, vienen expresadas en escala de intervalo, la relación que describe es lineal y cumplen los supuestos del contraste paramétrico. Sin embargo, algunos no son igualmente importantes que otros. Aunque los demás se dan por supuestos, el que sirve de elemento discernidor es el que las dos variables son de naturaleza continua y vienen expresadas en escala de intervalo. A pesar de ello, se va a centrar la atención en dos:

- 1) *Distribución normal.* La distribución implicada es normal bivariada, es decir, las puntuaciones de Y_{ij} se distribuyen normalmente para cada valor de X_{ij} . Igualmente, los valores de X_{ij} se distribuyen normalmente para cada valor de Y_{ij} . El incumplimiento de este supuesto tiene poca influencia sobre la validez de la prueba cuando los grados de libertad (tamaño de la muestra menos dos unidades) son más de 25 ó 30, o sea, para muestras por encima de 27 a 32 sujetos.
- 2) *Relación lineal entre X_{ij} e Y_{ij} .* Si dos variables mantienen una relación curvilínea, la utilización del coeficiente de correlación de Pearson será inútil para detectar la relación.

Supónganse dos variables, rendimiento previo y rendimiento posterior. Entre estas dos variables parece no existir inconveniente en admitir que la relación es lineal, según la cual cuanto mayor sea el rendimiento previo mayor será el rendimiento posterior y a la inversa. Ambas dimensiones crecen y decrecen en similar proporción.

Sin embargo, piénsese en estas otras dos variables, típicas del campo de la educación física, velocidad en la carrera y edad de los corredores. Si en el eje X del diagrama se va representando la variable edad y en el eje de ordenadas se representa la velocidad, resulta una representación curvilínea como la del gráfico. En este y en otros casos en los que la relación no sea lineal no se debe utilizar el coeficiente de correlación de Pearson, sino otro alternativo que, como se verá al final del tema, es el coeficiente de llamado *razón de correlación*.

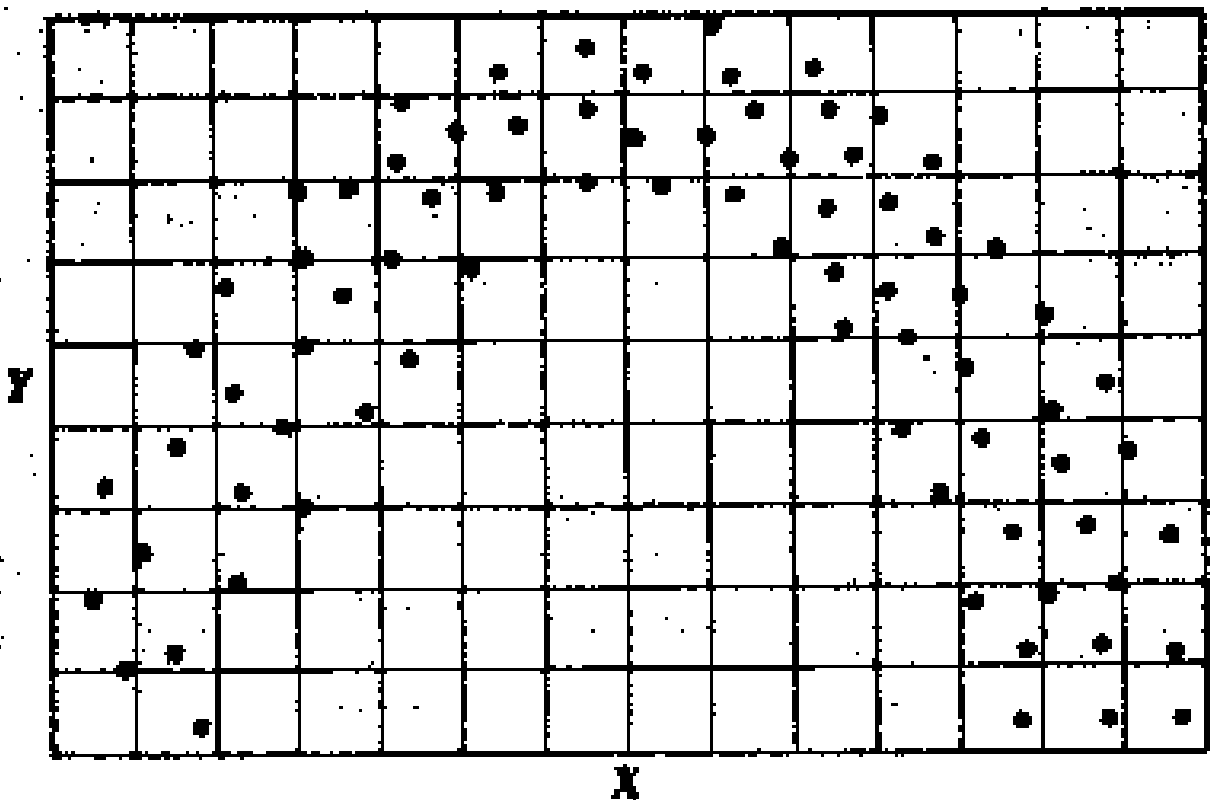


Diagrama de dispersión que muestra una relación curvilínea

La existencia de linealidad se puede detectar fácilmente o cuando se representan gráficamente los datos o cuando se hace una tabla de distribución de frecuencias, siempre que el tamaño de N sea grande; si es pequeño es difícil detectarlo.

Cálculo de r_{xy} . A continuación se indicará cuál es el razonamiento subyacente a la definición del coeficiente de correlación. Se recurrirá a la explicación del coeficiente de *correlación* llamado *producto-momento* de Pearson, dado que es el más usual, el que sirve de base para muchos otros coeficientes de correlación puesto que éstos no son sino una extensión de aquél a variables de otra naturaleza y escala.

Para entender, pues, *numéricamente* lo que es este índice de correlación me permito formular cuatro afirmaciones, que se irán explicando:

Principios y pasos en su cálculo:

1. Las puntuaciones de los sujetos en las dos distribuciones han de variar respecto de sus medias: $(X_{ij} - \text{Media}_x)$ y $(Y_{ij} - \text{Media}_y)$: Puntuaciones diferenciales.
2. Los pares de puntuaciones han de covariar: $(X_{ij} - \text{Media}_x) \cdot (Y_{ij} - \text{Media}_y)$. Multiplicar parejas de puntuaciones respetando el signo.
3. Se han de promediar las covariaciones para obtener un estimador promedio de la covariación: $\Sigma(X_{ij} - \text{Media}_x) \cdot (Y_{ij} - \text{Media}_y) / N_{\text{parejas}}$. Es la Covarianza: CoV_{xy} . También: $\text{CoV}_{xy} = r_{xy} \cdot s_x \cdot s_y$.

4. Es preciso tipificar la covarianza, esto es, reducirla a una misma unidad de interpretación de uno a otro estudio (de 0-nula, a 1-máxima): $\{\Sigma(X_{ij}-Media_x).(Y_{ij}-Media_y) / N_{parejas}\} / s_x.s_y = r_{xy}$.

Vayamos por partes.

1. Para que exista correlación es preciso que las puntuaciones de los sujetos en las dos distribuciones han de variar respecto de sus medias: $(X_{ij}-Media_x)$ y $(Y_{ij}-Media_y)$: Puntuaciones diferenciales.

En la práctica, para que se pueda estudiar la existencia de correlación entre variables, es necesario calcular las puntuaciones diferenciales de todos los sujetos en los dos valores:

Por una parte: $(X_{ij} - \text{media aritmética de } X_{ij})$ (columna 7); y
por otra parte: $(Y_{ij} - \text{media aritmética de } Y_{ij})$ (columna 8).

$X_{ij}(1)$	$(X_{ij} - \text{media de } X_{ij}) (7)$	$(X_{ij} - \text{media } X_{ij}) (Y_{ij} - \text{media } Y_{ij}) (9)$	$(Y_{ij} - \text{media de } Y_{ij}) (8)$	$Y_{ij} (2)$	$X_{ij} (3)$	$X_{ij} (4)$	$Y_{ij} (5)$	$Y_{ij} (6)$
20	+1.2	-5.88	-4.9	15	7	34	5	34
17	-1.8	-3.78	+2.1	22	8	32	9	31
15	-3.8	+11.02	-2.9	17	8	30	10	30
32	+13.2	+133.32	+10.1	30	9	26	12	28
20	+1.2	+2.52	+2.1	22	11	24	14	27
26	+7.2	+51.12	+7.1	27	14	23	15	24
21	+2.2	-4.18	-1.9	18	15	21	16	23
18	-0.8	-0.08	+0.1	20	17	20	17	22
9	-9.8	+57.82	-5.9	14	18	20	18	22
23	+4.2	+17.22	+4.1	24	19	20	20	21
34	+15.2	+214.32	+14.1	34	20	19	21	20
24	+5.2	+42.12	+8.1	28	20	18	22	18
14	-4.8	-5.28	+1.1	21	20	17	22	17
7	-11.8	+175.82	-14.9	5	21	15	23	16
11	-7.8	+77.22	-9.9	10	23	14	24	15
8	-10.8	+117.72	-10.9	9	24	11	27	14
19	+0.2	-1.58	-7.9	12	26	9	28	12
20	+1.2	+3.72	+3.1	23	30	8	30	10
8	-10.8	+42.12	-3.9	16	32	8	31	9
30	+11.2	+124.32	+11.1	31	34	7	34	5

$\Sigma X_{ij} (Y_{ij}) = 8532$ (1)(2)	$\Sigma(X_{ij} - \text{media } X_{ij}) (Y_{ij} - \text{media } Y_{ij}) = +$ 1049.6	$\Sigma X_{ij} \cdot Y_{ij} = 8651$ (3) (5)		
	$\Sigma X_{ij} \cdot Y_{ij} = 8651: (4)(6)$			
	$\Sigma X_{ij} \cdot Y_{ij} = 6312 : (3)(6)$			
ΣX_{ij}^2	(1): 8276	(2)9084	$\Sigma X_{ij} \cdot Y_{ij} = 6312$ (4)(5)	$r_{12} = 0.885$
ΣX_{ij}	376	398		
S_{n-1}	7.97	7.826		
Media	18.8	19.9		

2. Los pares de puntuaciones han de covariar: $(X_{ij} - \text{Media}_x) \cdot (Y_{ij} - \text{Media}_y)$. Multiplicar parejas de puntuaciones respetando el signo.

Esto se consigue multiplicando las puntuaciones diferenciales en todos los pares de datos:

$$(X_{ij} - \text{media de } X_{ij}) (Y_{ij} - \text{media de } Y_{ij}) \text{ (columna 9)}$$

Para *covariar* las puntuaciones es preciso que sean idénticos los signos de las parejas de puntuaciones (columnas 7 y 8) correspondientes a cada sujeto, de tal modo que habrá correlación, más o menos alta, en la medida en que predominen los pares de puntuaciones de idéntico signo, sea éste positivo o negativo.

De la contemplación de las mencionadas columnas se constata que en su mayoría los pares de puntuaciones son de idéntico signo, lo cual se evidencia en la columna 9, que resulta de multiplicar las distintas parejas de puntuaciones diferenciales. Hay en total 6 parejas de puntuaciones que tienen distinto signo, siendo éstas de un valor reducido en comparación con el resto de las desviaciones. La suma de tales productos expresa el valor total de las covariaciones, lo que los sujetos covarían en conjunto.

3. Se han de promediar las covariaciones para obtener un estimador promedio de la covariación: $\Sigma(X_{ij} - \text{Media}_x) \cdot (Y_{ij} - \text{Media}_y) / N_{\text{parejas}}$. Es la Covarianza: CoV_{xy} . También: $\text{CoV}_{xy} = r_{xy} \cdot s_x \cdot s_y$.

Sin embargo, es preciso llegar a un valor que exprese lo que las parejas de puntuaciones covarían por término medio; esto es, este valor no da una apreciación de lo que las parejas de datos covarían en conjunto y promediadamente. Para ello, es preciso que la suma de las covariaciones sea dividida por el número de parejas de datos que haya. Al índice que expresa lo que los sujetos por término medio covarían se denomina **covarianza**, que es una medida de la variación conjunta que tienen dos variables:

$$\Sigma(X_{ij} - \text{media de } X_{ij}) (Y_{ij} - \text{media de } Y_{ij}) / N \text{ parejas} = \text{CoV}_{xy} = S_{xy}$$

En nuestros datos el valor es de 52.48, el cual es un índice de *co/alejamiento* promedio. Se simboliza por: V_{xy} , S_{xy} ó CoV_{xy} . La fórmula anterior se denomina covarianza y es equivalente a otras fórmulas, en las que ahora no nos detenemos, la más conocida de las cuales es ésta:

$$\text{CoV}_{xy} = r_{xy} s_x s_y$$

También viene simbolizada la covarianza por: S_{xy} , la cual es una medida de la relación entre dos variables de vital importancia para el estudio de la dependencia estadística entre variables.

Dice Kerlinger (1975, 89) que este índice de la covarianza, es insatisfactorio

"porque su tamaño fluctúa con las extensiones y escalas de las diferentes X_i e Y_i . Es decir, podría ser 1.00 en este caso y 8.75 en otro caso, haciendo difíciles y engorrosas las comparaciones de un caso con otro".

4. Es preciso tipificar la covarianza, esto es, reducirla a una misma unidad de interpretación de uno a otro estudio (de 0-nula, a 1-máxima): $\{\Sigma(X_{ij}-Media_x).(Y_{ij}-Media_y) / N_{parejas}\} / S_x \cdot S_y = r_{xy}$.

Por ello, es preciso convertir todos los índices de covarianza a una unidad común de interpretación para que sean comparables. Se ha buscado una que tiene un recorrido que va desde +1.00 hasta -1.00. ¿Cómo se consigue? Respondiendo breve y lisamente, *tipificando la covarianza*.

Para entender lo que se quiere decir, es conveniente recurrir a las puntuaciones típicas. Para *tipificar* una puntuación se compara la puntuación diferencial de un sujeto con la desviación típica. Es decir, se compara lo que un sujeto se aleja (puntuación diferencial) con lo que se aleja el grupo (desviación típica). Hecho esto para todos los sujetos, nos permite compararlos situándolos en un punto del continuo de una distribución normal *standard*.

Aquí sucede algo similar: Se puede comparar lo que un grupo *covaría* por término medio (covarianza) con lo que *coaleja* por término medio el grupo en las dos variables. ¿Qué índice es el que sirve para expresar el coalejamiento del grupo? Simplemente: El producto de las desviaciones típicas en las dos distribuciones. Esto es: $S_{xi} (S_{yi})$:

$$\{\Sigma(X_{ij} - media de X_{ij}) (Y_{ij} - media de Y_{ij}) / N \text{ parejas}\} / S_{xi} (S_{yi}) = r_{xy}$$

Al índice o cociente que resulta de dividir la covarianza entre el producto de las desviaciones típicas de X_{ij} e Y_{ij} se lo denomina **correlación** (de Pearson). En nuestro ejemplo es igual a: $52.48 / (7.769)7.628 = 0.885$. Es preciso advertir que, aunque en algún caso se incluyan más de dos dígitos decimales en las correlaciones, lo usual es que se incluyan solamente dos.

Otras formas de cálculo

La fórmula anterior representa un medio adecuado para comprender el sentido de la correlación, si bien no es una fórmula recomendable para el cálculo. Existen otras, aunque no todas igualmente recomendables. Mediante sustitución algebraica de la fórmula anterior se pueden escribir un gran número de fórmulas alternativas.

Si en la fórmula anterior se comprueba que en realidad lo que se tiene, en parte, no es sino la fórmula de las puntuaciones típicas de X_{ij} y de Y_{ij} , entonces la fórmula del coeficiente de correlación no corresponde sino a la división de la suma del producto de las puntuaciones típicas de X_{ij} y de Y_{ij} por el número de parejas. Para corregir posibles sesgos, se puede dividir por $(N \text{ parejas} - 1)$:

$$r_{xy} = \Sigma z_{xi} \cdot z_{yi} / (N \text{ parejas} - 1).$$

Sin duda que no es una fórmula recomendable para el cálculo, puesto que es preciso calcular las puntuaciones típicas de las dos variables, tarea harto laboriosa. Una fórmula alternativa, tampoco recomendable para el cálculo, es la que sirve originariamente a la anterior, es decir, la que expresa las puntuaciones típicas en puntuaciones diferenciales:

$$r_{xy} = \frac{\sum (X_{ij} - \text{media de } X)(Y_{ij} - \text{media de } Y)}{(N \text{ parejas} - 1) s_{xi}s_{yi}}$$

¿Qué fórmula puede ser útil desde el punto de vista de la facilidad de cálculo? La más útil, especialmente cuando se tiene una calculadora, es la que parte de las puntuaciones directas. Teniendo los siguientes totales, que proporciona la función estadística de una calculadora, una vez que se han introducido todos los datos, se puede fácilmente calcular la correlación: $\sum X_{ij}Y_{ij}$, $\sum X_{ij}$, $\sum Y_{ij}$, $\sum X_{ij}^2$, $\sum Y_{ij}^2$:

$$r_{xy} = \frac{(N \sum X_{ij}Y_{ij} - \sum X_{ij} (\sum Y_{ij}))}{\sqrt{\{N\sum X_{ij}^2 - (\sum X_{ij})^2\} \{N\sum Y_{ij}^2 - (\sum Y_{ij})^2\}}}$$

Como se dirá, la correlación obtenida en el problema que se viene comentando es en términos absolutos positiva y alta. Pero ya se dijo, al hablar de la representación gráfica de los valores de las dos distribuciones en X_{ij} e Y_{ij} , que había otra disposición de las puntuaciones, que podían dar lugar a otros valores de correlación y de signo en ocasiones diferente, según fuera la naturaleza de la relación, directa o inversamente proporcional.

Como se varió la disposición de las puntuaciones en ambas variables, las correlaciones que se obtienen son diversas. Veamos. Si nos atenemos a los datos de las variables X_{ij} e Y_{ij} en las columnas 3 y 5, de una simple inspección ocular se observa que la relación es directamente proporcional y casi simétrica. Por ello, el índice de correlación esperable es positivo (directamente proporcional) y alto o altísimo, si no es perfecto (correlación 1). En la tabla de datos se encuentra la información suficiente para calcular el índice de correlación según la fórmula de la calculadora incluída ineditamente arriba:

$$r_{xy} = + 0.9859$$

Es ésta, como era de esperar, una correlación positiva y casi perfecta. Si ahora nos fijamos en los valores de las columnas 4 y 6 se constata que son idénticos a los de las columnas 3 y 5 aunque invertidos, es decir, comenzando por las puntuaciones superiores. En términos estadísticos esto es idéntico porque la correlación expresa ordenamiento de las puntuaciones. La correlación, si se hacen los cálculos es idéntica a la obtenida arriba: $r_{xy} = + 0.9859$.

Interpretación de r_{xy}

Welkowitz *et al.* (1981, 208) indican que el coeficiente de correlación posee algunas características muy apreciables:

1. El valor cero indica ausencia de relación lineal entre las variables;
2. El valor numérico del coeficiente, independientemente del signo de la correlación, indica la fuerza de la relación;
3. El signo indica la dirección de la relación;
4. El valor positivo más elevado es +1, mientras que el negativo es -1.

Es decir, hay algunos aspectos de relieve a la hora de interpretar un índice de correlación. Los mencionados se refieren a la interpretación del número, pero existen otros referidos al método y al concepto.

1. Relación: Asociación y correlación. Una primera idea es que hay dos tipos de índices de relación: Los índices de asociación y los de correlación.

Afirma Fox (1981) que cuando se trabaja con datos nominales y ordinales verbales expresados en un conjunto de categorías limitadas no se permite formular preguntas sobre la correlación, tal como se acaba de ver. Este tipo de datos es el adecuado para las medidas de relación expresadas en forma de asociación, que refleja la tendencia de los datos a aparecer sistemáticamente en ciertas combinaciones de categorías y, por tanto, aparecen en casillas.

Los índices de asociación presentan semejanzas y diferencias respecto de los de correlación: La semejanza reside en que ambos arrojan un valor cero cuando no hay relación entre las variables. La diferencia reside en que no todos los índices de asociación expresan la correlación perfecta mediante el número uno. Es decir, mientras la correlación presenta un recorrido de cero a uno, por el contrario, la relación perfecta en un índice de asociación no siempre alcanza el valor de 1, se queda en valores inferiores a uno.

Para poder interpretar el índice de asociación existe una fórmula que permite calcular cuánta sería la cantidad de relación en el supuesto de que la relación fuera perfecta (índice perfecto máximo), el cual sirve de punto de comparación del índice empírico. Es decir, aun en el caso de una relación perfecta, el índice es inferior a 1 (asociación máxima), que actúa como un valor máximo teórico, el cual sirve para interpretar el valor empírico.

Existen tres medidas de asociación suficientemente generalizadas. Se trata de Chi Cuadrado (χ^2), del Coeficiente de Contingencia (C) y del Coeficiente Phi (ϕ).

2. Relación, no causalidad. Una segunda observación por hacer es que los índices de correlación no expresan relación de causalidad, en una sola dirección, relación unívoca, sino relación biunívoca.

Un coeficiente de correlación no permite determinar la causa de la relación. La correlación es una medida de la estabilidad de la ordenación de dos conjuntos de datos, refleja orden de las puntuaciones de una a otra distribución, tal como se ha podido comprobar en la columna (3), covariación de unas puntuaciones y no dirección de la covariación.

Si entre dos variables existe una elevada correlación, ésta puede darse por alguna de las siguientes razones:

- X_{ij} es la causa de Y_{ij} ;
- Y_{ij} es la causa de X_{ij} ;
- tanto X_{ij} como Y_{ij} son causadas por una tercera, cuarta, ... variables.

Citan Welkowitz *et al.* (1981, 212/3) una historia muy reveladora de lo que se quiere decir y que no resisto la tentación de incluirla:

"se refiere a un estudio que reveló la existencia de una fuerte correlación positiva entre el número de cigüeñas y el de nacimientos en ciudades europeas (es decir, cuantas más cigüeñas en la ciudad, más nacimientos). En lugar de lanzar un dramático pregón, confirmando los míticos poderes de las cigüeñas, se prosiguieron las investigaciones".

Se da una explicación de la inferencia equivocada:

"Se descubrió que las cigüeñas anidaban en las chimeneas, lo cual, a su vez, dio lugar a la conclusión de que la responsable de la relación entre cigüeñas y nacimientos era una tercera variable, el tamaño de la ciudad. En las grandes ciudades hay más habitantes y por ello más nacimientos; y más casas, por tanto más chimeneas, y de ahí, más cigüeñas. Las ciudades pequeñas poseen pocos habitantes, y por tanto se dan pocos nacimientos, casas, chimeneas y cigüeñas. Así, la atribución de causalidad es un problema lógico o científico, no estadístico".

3. Relación tipificada. Una tercera observación se refiere al hecho de que la correlación es una medida tipificada de la relación entre variables.

Pues bien, precisamente por dicha tipificación es posible comparar los índices de correlación e incluso promediar varios en orden a tener una visión de conjunto sobre la relación entre dos variables cuando hay varios estudios sobre el mismo campo. Es decir, cuando hay varios índices de correlación sobre la relación entre dos mismas variables, es útil promediarlas para tener una visión de conjunto sobre dicha relación. Es la *correlación promedio*. Este procedimiento se sigue en el metaanálisis o síntesis cuantitativas de correlaciones.

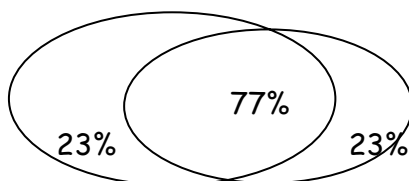
Veamos una síntesis, la de Bloom (1976), que pretendía expresar cuánta era la correlación entre el rendimiento previo y el rendimiento posterior. Este autor consultó 27 estudios o artículos publicados, los cuales contenían en total 82 índices de correlación. No tiene nada de extraño que en un estudio haya varios índices de correlación, porque se puede poner en relación el rendimiento previo de varias asignaturas con sus respectivos rendimientos posteriores y esto se refleja en tantos índices de correlación como materias se hayan examinado. Así, pues, Bloom promedió los 82 índices de correlación obteniendo una correlación en nuestro caso de 0.75, que es la correlación promedio que existe entre el rendimiento previo y el rendimiento posterior. Estos son los datos:

Variables en relación	Estudios	Relaciones (r)	r promedio	Referencia
Rdto. previo con rdto. posterior	27	82	0.75	Bloom (1976)

4. **Fuerza de la relación o varianza compartida.** Un índice de correlación no es fácilmente interpretable. Para ello es preciso que dicho coeficiente sea elevado al cuadrado, el cual expresa la proporción de variabilidad -no se olvide que la correlación es variabilidad compartida por dos variables- que ambas variables comparten.

Indica, pues, cuánta variabilidad o varianza *comparten* ambas distribuciones. El índice que resulta de elevar al cuadrado la correlación se denomina *coeficiente de determinación* e indica la proporción común de varianza de las variables, la que es compartida. Se simboliza por d , que es igual a r_{xy}^2 , es decir: $d = r_{xy}^2$.

Para una mejor interpretación de dicho coeficiente, éste se multiplica por 100, con lo cual se nos dice qué porcentaje de varianza de ambas variables es común o compartido. En nuestro caso el coeficiente de determinación es $0.88^2 = 0.77$. En términos de porcentaje vale: $0.77 (100) = 77\%$. Esto significa que ambas variables comparten aproximadamente un 77 por ciento de su variabilidad.



Expresándolo gráficamente se puede ver con más claridad. Cada variable tiene su propia variabilidad, la de la primera es $S_x^2 = 60.36$ y la de la segunda es $S_y^2 = 58.19$, que son distintas, pero ambas comparten el 77 de sus propias varianzas, es decir, 46.48 la primera y 44.80 la segunda.

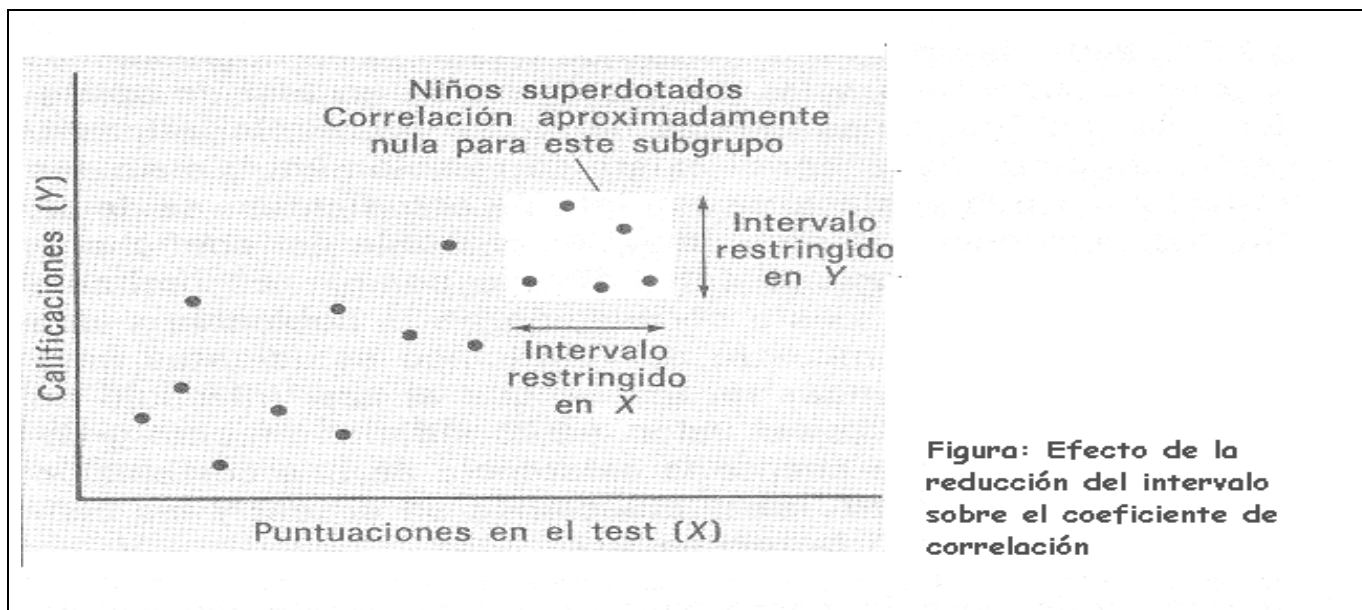
Este valor es enormemente importante, porque una de las funciones que puede cumplir es la de la predicción. Esto es, el cuadrado de la correlación nos permite el conocimiento de una variable (criterio) a partir del conocimiento de otra (predictora) en la misma proporción del coeficiente de determinación.

Del mismo modo que r_{xy}^2 es el coeficiente de determinación, asimismo se puede calcular el *coeficiente de alienación*, que viene definido por la fórmula: $k = \sqrt{1 - r_{xy}^2}$ el cual expresa la proporción de varianza de Y que no se puede predecir a partir de X. Complementariamente, el *coeficiente de valor predictivo* viene definido por la fórmula: $E = 1 - k = 1 - \sqrt{1 - r_{xy}^2}$, que se debe interpretar como el grado de seguridad en la predicción.

5. **Variabilidad y restricción del rango de la distribución.** Se ha mencionado en los cálculos de la correlación que la primera condición para que ésta exista es que haya variabilidad en las puntuaciones.

Esta condición tiene una importancia básica porque, si se da a las puntuaciones en una distribución la posibilidad de que sus puntuaciones varíen, esa variable podrá correlacionar con otra. Es decir, cuando se incrementa la variabilidad de las puntuaciones en ambas variables, es lógico esperar que se incremente el valor absoluto de la correlación. El problema que subyace, como afirma Fox (1981, 263/4) al hablar de la interpretación del coeficiente de correlación, es el del grado en que dos variables correlacionan en una distribución total y completa de la muestra. Esta sería la correlación que mejor reflejara la relación entre variables.

Sin embargo, no son inusuales las situaciones en que la obtención de correlaciones se hace a partir de muestras restringidas, situación que puede dar lugar a correlaciones más altas o más bajas de las que se obtendrían con muestras que cubren toda la gama de la distribución.



Un caso nada inusual es el del investigador o práctico de la educación que tratan con muestras restringidas, seleccionadas de una parte o extremo de la distribución. Las más usuales son las que trabajan con el extremo superior de la distribución, con el inferior o con muestras selectivas, que el mismo sistema educativo proporciona. Muchos programas educativos trabajan con un grupo limitado de la muestra de alumnos, sean deficientes, superdotados, universitarios, ... Es decir, son muestras que tienen como resultado una selección. Al proceso de trabajar con datos de grupos selectivos o de uno de los extremos de la distribución se denomina *restricción*.

Cuando se intenta correlacionar dos variables en sujetos que están en uno de esos subgrupos, se está pidiendo al método de las correlaciones que haga distinciones en ese subgrupo. Y, sin embargo, el método, al reducir considerablemente el intervalo de variación en ambas variables o en una de ellas, se reduce asimismo la correlación entre dichas variables en valor absoluto.

Welkowitz *et al.* (1981, 215) incluyen una figura en la que puede verse que, reduciendo el intervalo en las dos variables, en este caso identificando en el grupo más grande una submuestra de alumnos superdotados, la distribución que las puntuaciones describen es la correspondiente a

una correlación nula, mientras que si se incluyen puntuaciones de un amplio rango en conjunto vienen a indicar que la nube de puntos describe una correlación positiva y no baja.

Afirman estos autores (Welkowitz *et al.*, 1981, 215):

"es mucho más difícil establecer diferencias finas entre casos que están muy próximos en ambas variables que hacerlo entre casos que difieren mucho en dichas variables. Es difícil predecir si un muchacho superdotado obtendrá un sobresaliente alto o bajo, mientras que resulta mucho más sencillo distinguir en un amplio intervalo de estudiantes entre aquellos que van del suspenso al sobresaliente alto".

Alguien podría pensar con buen criterio que entre inteligencia y rendimiento existiera correlación alta y positiva, y no le falta razón. En efecto, cuando se miden ambas variables en una muestra amplia y heterogénea de sujetos, p.e. alumnos de la enseñanza obligatoria, la correlación promedio que se suele obtener es en términos generales alta, tal como han recogido de los estudios de síntesis Fraser *et al.* (1987):

Relación del rdto. con:	n estudios	n relaciones	r promedio	Fuente
Inteligencia	72	503	0.51	Hattie y Hansford (1982)
Capacidad gral.	169	42	0.43	Fleming y Malone (1983)
Capacidad gral.	34	62	0.48	Boulanger (1981)
Capacidad cognitiva	66	58	0.34	Steinkamp y Maehr (1983)

Sin embargo, cuando se correlaciona la inteligencia con el rendimiento en la enseñanza universitaria, dicha correlación en el mejor de los casos (curso primero) es significativa pero baja. La explicación parece ser atribuida a la alta homogeneidad de los sujetos en las puntuaciones en inteligencia:

"Muy probablemente hayamos llegado a un nivel universitario, en el que la homogeneidad intelectual de los sujetos es mayor y en consecuencia sean otros factores no intelectuales quienes más contribuyan a la predicción del rendimiento" (González Galán y López López, 1985, 505).

Así, pues, para que se pueda manifestar la correlación entre dos variables es preciso que tengan o arrojen un amplio rango en sus puntuaciones, es decir, que su intervalo no sea reducido. Cuando esto sucede, cuando hay restricción de la muestra, la correlación obtenida es una *subestimación* de la que se obtendría si se emplearan datos procedentes de todo el grupo.

El segundo caso, por el contrario, conduce a una *supraestimación* de las correlaciones. Además de la selección de uno de los muchos subgrupos restringidos que se dan en educación, sucede que para fines de contraste entre grupos sean seleccionados subgrupos extremos en alguna variable. Pues bien, cuando esto ocurre, es más alta la correlación entre dos variables que si se hubiesen tenido en cuenta los subgrupos intermedios.

En cualquier caso, como afirma Fox (1981, 264/5),

"salvo cuando las correlaciones están basadas en la distribución completa, su interpretación es peligrosa y se debe aplazar hasta que se hayan realizado las correcciones".

6. Dirección de la correlación. Otra de las características de la correlación se refiere a la naturaleza de la relación entre las variables, que se refleja en el signo de la correlación. Dado que el recorrido de una correlación es desde 0 a /1/, el signo es de vital importancia.

Recuérdese que el signo de la correlación deriva del producto de las desviaciones o puntuaciones diferenciales de las dos variables que se quiere poner en relación. En efecto, un signo negativo indica que cuando la desviación de un sujeto en una variable es positiva en la otra es negativa, o viceversa; por el contrario, cuando es positivo indica que la desviación puede ser en ambas columnas positiva o negativa y en consecuencia la correlación es positiva. Esto significa en el primer caso que la relación entre las variables es inversamente proporcional y en el segundo que es directamente proporcional. Esto se evidencia mediante la contemplación de los datos de las columnas incluídas, que con algunas adaptaciones -las adecuadas al objetivo de este apartado- se repiten en este lugar.

Suje- tos	X _{ij} (1)	X _{ij} (2)	Y _{ij} (3)	Y _{ij} (4)	x _{ij} (1)	y _{ij} (3)	x _{ij} (2)	y _{ij} (4)	(1)(3) (x _{ij})(y _{ij})	(1)(4) (x _{ij})(y _{ij})	(2)(3) (x _{ij})(y _{ij})	(2)(4) (x _{ij})(y _{ij})
1	7	34	5	34	-11.8	-14.9	+15.2	+14.1	+	-	-	+
2	8	32	9	31	-10.8	-10.9	+13.2	+11.1	+	-	-	+
3	8	30	10	30	-10.8	-09.9	+11.2	+10.1	+	-	-	+
4	9	26	12	28	-09.8	-07.9	+07.2	+08.1	+	-	-	+
5	11	24	14	27	-07.8	-05.9	+05.2	+07.1	+	-	-	+
6	14	23	15	24	-04.8	-04.9	+04.2	+04.1	+	-	-	+
7	15	21	16	23	-03.8	-03.9	+02.2	+03.1	+	-	-	+
8	17	20	17	22	-01.8	-02.9	+01.2	+02.1	+	-	-	+
9	18	20	18	22	-00.8	-01.9	+01.2	+02.1	+	-	-	+
10	19	20	20	21	+00.2	+00.1	+01.2	+01.1	+	+	+	+
11	20	19	21	20	+01.2	+01.1	+00.2	+00.1	+	+	+	+
12	20	18	22	18	+01.2	+02.1	-00.8	-01.9	+	-	-	+
13	20	17	22	17	+01.2	+02.1	-01.8	-02.9	+	-	-	+
14	21	15	23	16	+02.2	+03.1	-03.8	-03.9	+	-	-	+
15	23	14	24	15	+04.2	+04.1	-04.8	-04.9	+	-	-	+
16	24	11	27	14	+05.2	+07.1	-07.8	-05.9	+	-	-	+
17	26	9	28	12	+07.2	+08.1	-09.8	-07.9	+	-	-	+
18	30	8	30	10	+11.2	+10.1	-10.8	-09.9	+	-	-	+
19	32	8	31	9	+13.2	+11.1	-10.8	-10.9	+	-	-	+
20	34	7	34	5	+15.2	+14.1	-11.8	-14.9	+	-	-	+

Medias	18.8	18.9	19.9	19.9		+	-	-	+
Correlaciones entre las columnas (1) a (4), que se indican:	$r_{x_1y_3}=.98$		$r_{x_1y_4}=-.98$		$r_{x_2y_3}=-.98$		$r_{x_2y_4}=.98$		

De su contemplación se desprenden unas conclusiones suficientemente claras. En la primera columna de signos (1-3) el predominio de los signos positivos es total, así como en la columna de

los signos 2-4, dado que en la primera las puntuaciones eran ascendentes en ambas variables y en la segunda eran descendentes, lo cual da lugar a que las puntuaciones diferenciales sean idénticas, aunque invertidas en el orden de colocación. En cualquier caso, la relación de las variables es directamente proporcional: De baja-baja a alta-alta en la columna 1-3 y de alta-alta a baja-baja en la columna 2-4. En estos dos casos la relación es directamente proporcional, que se expresa en signo positivo. Las correlaciones son ambas positivas: $r_{x_1y_3}=.98$ y $r_{x_2y_4}=.98$.

Por el contrario, en la columna de signos 1-4 y 2-3 el predominio de los signos negativos es casi total, dado que en la columna 1-4 se reflejan puntuaciones de signo ascendente en una columna (columna 1) y de signo descendente en otra (columna 4); igualmente ocurre en la columna 2-3, que es reflejo de las columnas 2, cuyas puntuaciones son descendentes, y 3, cuyas puntuaciones son ascendentes. Es decir, las puntuaciones son ascendentes-descendentes en un caso (columna 1-4) y descendentes-ascendentes en el otro (columnas 2-3).

En las dos situaciones la relación de las variables es inversamente proporcional: De baja-alta a alta-baja en la columna 1-4 y de alta-baja a baja-alta en la columna 2-3. En estos dos casos la relación es inversamente proporcional, que se expresa en signo negativo. Las correlaciones son ambas, en consecuencia, negativas: $r_{x_1y_4} = -.98$ y $r_{x_2y_3} = -.98$.

7. Significación de la correlación. Merece la atención detenerse en la significación del valor de correlación.

Toda correlación igual a 0 es una correlación nula. Si el valor de correlación es distinto de 0, entonces podemos establecer matices. En efecto, si en una primera apreciación visual se ve que una correlación es próxima a 0 -próxima a 0 se entiende que es $< /0.20/-$, entonces se dice que estadísticamente la correlación es inexistente, tanto si es positiva como negativa. Decir que estadísticamente es inexistente significa que dicha correlación puede ser fácilmente (probablemente) explicable por azar; y toda correlación que es explicable por azar se considera una correlación inexistente.

La anterior apreciación es aproximativa e inexacta. Existen otros medios más precisos de saber si la correlación es significativa estadísticamente:

- Pueden consultarse unas tablas de significación del coeficiente de correlación para unos determinados valores de muestras (grados de libertad) y niveles de probabilidad;
- o puede hacerse estadísticamente un contraste de significación de r .

Si la correlación cabe dentro de la catalogación de significativa, entonces ¿cómo valorar un índice significativo de correlación? Es preciso distinguir dos tipos de valoración de la significación, uno absoluto y otro relativo. El *absoluto* se refiere al juicio que merecen distintos valores o cantidades de correlación independientemente de las variables puestas en relación.

Sin embargo, la interpretación *relativa* hace alusión a un juicio comparativo entre la correlación obtenida en un determinado caso y la que ordinariamente se obtiene en los estudios de investigación y que aparecen en la consulta en la bibliografía.

Es decir,

"para evaluar las correlaciones en el contexto de los datos concretos, el investigador tiene que conocer suficientemente bien la bibliografía para saber qué valores de las correlaciones se han obtenido en trabajos anteriores" (Fox, 1981, 265).

En efecto, se puede establecer una tabla de este tipo (puede ser distinta):

Valoración convencional		Valoración más exigente según Fox (1981, 265)	
Intervalos de correlación	Valoración	Valores de correlación	Valoración
0.20- 0.40	Correlación baja	Desde real hasta ± 0.50	Baja (25% varianza común)
0.41- 0.60	Correlación media	Desde ± 0.51 hasta ± 0.70	Moderada (50% varianza)
0.61- 0.80	Correlación alta	Desde ± 0.71 hasta ± 0.86	Alta (50-75 % varianza)
0.81- 0.99	Correlación altísima	Superior a ± 0.86	Muy alta (>75 % varianza)

Con el ejemplo de la correlación entre rendimiento previo y rendimiento posterior, visto anteriormente, se entenderá mejor lo que se quiere decir: Si habitualmente se obtiene un índice de correlación de 0.75 aproximadamente -dato que se sabe por la consulta a la investigación- entre ambas variables, una correlación de 0.59 sería una correlación muy baja, aunque sea una correlación media, incluso rozando con alta. Del mismo, una correlación de 0.83 sería una correlación, además de ser absolutamente hablando altísima, sería inusualmente alta. Es decir, el juicio sobre un índice de correlación es relativo al que se obtiene habitualmente en la investigación, la cual es preciso conocer para establecerlo.

Conclusión

Criterios de selección de los distintos índices de correlación. Se han comentado anteriormente algunas condiciones o supuestos para el cálculo del coeficiente de correlación de Pearson. Como introducción al tema siguiente, aunque algunos índices de correlación no se expliquen por el momento, se va a incluir una tabla, en la que se van a ir repasando los distintos índices de correlación y asociación, desde el punto de vista de los requisitos para su uso.

Ya se habló de algunos supuestos, en especial de la normalidad de las distribuciones de las variables en la población y de la linealidad/curvilinealidad. Sin embargo, en condiciones normales los indicadores de uso de uno u otro índice de correlación/asociación son básicamente dos, la naturaleza de las variables que se ponen en relación y la escala en que dichas variables se presentan. Se indican, no obstante, en ocasiones algunas condiciones o suposiciones adicionales.

Así, pues, se van a incluir en una tabla los requisitos para el uso de distintos coeficientes de relación entre dos o más de dos conjuntos de datos. Se va a indicar el nombre, el símbolo, de qué naturaleza es cada variable y la escala en que se expresa, así como algunos supuestos o condiciones adicionales.

Glosario⁽¹⁾

Asociación	Predicción
Causalidad	Razón de correlación
Coefficiente de alienación	Regresión
Coefficiente de correlación	Regresión de X sobre Y
Coefficiente de correlación en puntuaciones diferenciales	Regresión de Y sobre X
Coefficiente de correlación en puntuaciones típicas	Relación
Coefficiente de correlación producto-momento	Relación curvilínea
Coefficiente de determinación d (r^2_{xy})	Relación directamente proporcional
Correlación	Relación imperfecta
Correlación de Pearson	Relación inversamente proporcional
Correlación significativa	Relación lineal
Covariación	Relación negativa
Covarianza	Relación nula
Diagrama de dispersión	Relación perfecta
Distribución de frecuencias bivariada	Relación positiva
Ecuación de la recta	Restricción de la amplitud
Error típico de estimación	r_{xy}
No linealidad de la regresión	Supuestos
Pares de observaciones	Variabilidad
Pendiente	Varianza de diferencias
Pendiente en puntuaciones típicas	

⁽¹⁾ El glosario de este tema es válido para el de la regresión.