Capítulo 14

CORRELACIÓN Y REGRESIÓN

Sumario

Objetivos específicos

Al finalizar el estudio del capítulo, el estudiante será capaz de:

1. Explicar los conceptos de correlación y regresión.
2. Calcular e interpretar el coeficiente de correlación lineal simple.
3. Establecer la distinción entre correlación y causalidad.
4. Calcular e interpretar el coeficiente de correlación parcial.
5. Calcular e interpretar la ecuación de regresión lineal.
6. Determinar la confiabilidad de una línea de regresión ajustada.
7. Utilizar modelos de regresión ya ajustados para hacer estimaciones.
8. Realizar inferencias acerca del coeficiente de correlación lineal poblacional (ϱ).

Resumen

En este capítulo se explican los conceptos de correlación y regresión, se discuten e ilustran las técnicas empleadas para medir la correlación y el ajuste de la línea de regresión simple; se establece la diferencia entre correlación y causalidad, se presentan ideas acerca del uso y limitaciones de esas técnicas.

14.1. NIVELES DE MEDICIÓN

Antes de entrar al tema de la asociación entre variables, es importante hacer una referencia a los "niveles de medición". Tradicionalmente, se ha pensado que solo se puede medir cuando se tiene un sistema métrico, una unidad de medida: centímetros, kilogramos, colones, minutos, grados, etc. Sin embargo, modernamente, y en especial en las ciencias sociales, se utiliza un enfoque mucho más amplio que permite incluir, dentro del concepto de medición, situaciones en las que solo es posible ordenar los elementos o personas, o simplemente determinar si difieren respecto a la característica de interés. Desde este punto de vista, el término medición no solo incluye la operación de apreciar o evaluar el monto o intensidad de una característica, usando un sistema métrico y un instrumento apropiado, sino también una valoración comparativa basada en un ordenamiento o *ranking* (sistema ordinal). La perspectiva también permite incluir el proceso mediante el cual se determina si un evento o situación tiene o no una cierta característica, este tipo de medición descansa en un sistema nominal.

A la luz de lo anterior, una definición amplia de medición es la siguiente:

Medición es cualquier procedimiento mediante el cual se asigna un valor al nivel o estado de una característica de estudio.

Para medir se requiere un instrumento, procedimiento o criterio que deje asignar los valores. Ejemplos típicos de instrumentos de medición son las balanzas, las cintas métricas y los cronómetros; pero hay otros tipos, en las ciencias sociales con frecuencia se denomina instrumentos de medición a los cuestionarios, ya que las preguntas incluidas permiten al investigador o analista medir –en forma individual o combinada– las características de interés en individuos concretos. También existen criterios y procedimientos

como los que usan los jueces para calificar sementales en una feria y la FIFA para preparar su *ranking* de las selecciones de los países.[1]

Se distinguen varios niveles de medición:

Nivel nominal. Cuando se tiene una característica para la cual se pueden definir categorías diferentes, pero no es posible ordenarlas, ni decir en cuánto difiere una de otra, se tiene una escala nominal y se dice que la característica está medida a nivel nominal. En esta escala, por lo tanto, solo es posible decir si dos elementos son iguales o diferentes, pero no es factible asegurar cuál es mayor o menor, ni en qué volumen o cantidad difieren. Las categorías permiten juntar las personas o elementos que son iguales –para la característica en consideración– pero no se puede ordenar las categorías de la clasificación de menor a mayor o a la inversa, ni decir en cuánto se diferencia una de otra. En otras palabras, aparte de mencionar que los elementos dentro de cada categoría difieren de los que están en las otras, no se garantiza nada más, constituye el nivel de medición más simple.

El estado conyugal es un ejemplo de una característica medida en escala nominal. Se puede clasificar a las personas en seis categorías: solteras, casadas, unidas, viudas, divorciadas y separadas. Todas, dentro de una categoría, tienen el mismo estado conyugal, son iguales en ello, y se diferencian de quienes están en cualquiera de los otros estados. Sin embargo, las categorías utilizadas: solteras, casadas, etc., no pueden ordenarse, tampoco es posible decir en qué volumen de estado conyugal se diferencian un viudo de un soltero.

Otro caso de una variable, medida a nivel nominal, es el país de nacimiento. Suponga que se tiene una muestra de adultos residentes en Centroamérica y se han fijado los siguientes códigos para clasificarlos:

1. Guatemala	2. Honduras	3. El Salvador	4. Nicaragua
5. Costa Rica	6. Panamá	7. Belice	8. Otro país

El sistema de clasificación permite agrupar las personas nacidas en un mismo país, así, se tienen 8 grupos internamente homogéneos, pero diferentes entre sí, respecto a la característica de interés. No permite, sin embargo, ordenar las categorías o decir en cuánto difieren dos de estas respecto a la característica "país de nacimiento", pretensión que, por otra parte, carece de sentido.

1. La clasificación mundial que hace la FIFA regularmente, cuyo nombre oficial es Clasificación Mundial FIFA/Coca-Cola, también conocido como ranking FIFA, es un sistema de ordenación de las 208 selecciones de fútbol pertenecientes a la FIFA (en este momento). Se basa en los puntos obtenidos en partidos oficiales y en ciertos criterios sobre la potencialidad de las selecciones que se enfrentan.

Los números que aparecen frente a las categorías son un recurso para facilitar el almacenamiento digital de la información y su posterior procesamiento en la computadora, pero no se interpretan como números que puedan ser ordenados o sometidos a operaciones aritméticas como suma, resta, división, etcétera.

Nivel ordinal. En un sistema de medición nominal solo es posible decir si dos objetos son iguales o diferentes. Hay situaciones, sin embargo, en las que no se tiene una unidad de medida, pero se está en capacidad de determinar si los objetos tienen más o menos de la característica de interés y, por lo tanto, es posible ordenarlos. Cuando esta situación se da, se tiene un sistema de medición ordinal y se dice que la característica está medida a nivel ordinal.

Un ejemplo de sistema ordinal lo es cuando se clasifican prendas de vestir como las camisas, en inglés se habla de *small*, *medium*, *large* y *extralarge*.

Otro ejemplo, es el caso de la jerarquía militar, en la cual se encuentran soldados rasos, cabos, sargentos, tenientes, capitanes, mayores, coroneles y generales, y es sabido que quienes están en las diferentes categorías son diferentes en cuanto poder y mando,, además, son ordenables de acuerdo con esa categoría:

raso - cabo - sargento - teniente - capitán - mayor - coronel - general

Las preguntas de opinión, en las que se pide a las personas evaluar el trabajo realizado por un presidente o un funcionario público electo, también originan a variables de nivel ordinal en una escala del tipo siguiente:

1. Muy bueno 2. Bueno 3. Regular 4. Malo 5. Muy malo

Es claro que todas las personas que dan una cierta respuesta tienen igual opinión y muestran mejor o peor que a las situadas en las otras categorías, y por ello se tiene una típica escala ordinal.

Nivel métrico. Cuando se dispone de una unidad de medida es posible determinar si dos elementos son iguales o diferentes, cuál es mayor o menor y en cuánto difieren, entonces se tiene la denomina escala de medición métrica. Como ejemplos pueden citarse la estatura en centímetros y la temperatura en grados centígrados.

Se presentan dos tipos de escalas métricas: las de intervalo y las de razón. La diferencia entre ellas reside, fundamentalmente, en la forma en que se establece el cero de la escala.

La escala de intervalo. Es aquella en la cual el 0 de la escala es fijado arbitrariamente. Entre los ejemplos más conocidos están las de temperatura (centígrados y fahrenheit) y el coeficiente de inteligencia (CI). El 0 en las escalas de temperatura no indica ausencia de temperatura y que, además, el punto de congelación no es el mismo en ambas escalas. En estas, como el 0 es arbitrario y no indica ausencia de temperatura, no es posible

afirmar, por ejemplo, que una ciudad donde la temperatura promedio en Navidad es 30 °C, es el doble de caliente que otra donde ese promedio es 15 °C. Tampoco se puede decir que una persona con un CI de 120 es el doble de inteligente que una con CI de 60.

La escala de razón. Tiene propiedades muy similares a la de intervalo, pero el 0 es verdadero o absoluto y representa la ausencia de la característica en cuestión. Así, por ejemplo, si se considera peso, el valor 0 kilos indica ausencia de peso y es legítimo decir que una persona de 120 kilos, pesa el doble que una con un peso de 60 kilos.

La escala más fuerte y con mejores propiedades, estadísticamente hablando, es la de razón y la más débil la nominal. Ciertas técnicas estadísticas requieren datos a nivel métrico para ser utilizadas, otras admiten nivel ordinal o nominal. Un paso muy importante, al analizar datos estadísticos, es determinar cuál es el nivel de medición de las variables, de manera que sea posible seleccionar la técnica más adecuada para el nivel de medición de la información que se analiza.

En el capítulo 1, NATURALEZA DE LA ESTADÍSTICA, se hizo referencia a características cuantitativas (o variables) y a cualitativas (o atributos). Desde la perspectiva de esta sección, las primeras corresponden a aquellas medidas a nivel métrico y las segundas a las de nivel ordinal o nominal. Para los efectos de lo que sigue, se usa el término variable para referirse tanto a los atributos como a las variables cuantitativas, también se emplea, en ciertas oportunidades, el término variable cualitativa.

14.2. ASOCIACIÓN ENTRE VARIABLES: LOS CONCEPTOS DE CORRELACIÓN Y REGRESIÓN

En capítulos previos, se consideró una serie de técnicas descriptivas que incluyen la construcción y representación de distribuciones de frecuencias, el cálculo de medidas de posición y de variabilidad y la elaboración de cuadros y gráficos; también se presentaron porcentajes y números relativos. Todas estas técnicas estadísticas tienen un rasgo común: el ser univariables, es decir, el considerar solo una variable o característica de la unidad de estudio. Si se analizan estudiantes, por ejemplo, se toma su peso, se hace la distribución de frecuencias, se construye el histograma, se calculan las medidas de posición y variabilidad, etc. Luego, se toman los datos de estatura y se hace lo mismo, etc. En la práctica, sin embargo, un gran número de problemas estadísticos involucran la consideración simultánea de dos o más variables en la misma unidad de estudio, y es muy importante examinar la forma como se relacionan esas variables. Esto lleva a identificar poblaciones bivariantes o multivariantes, según se consideren dos o más variables en una misma unidad de estudio, en el caso antes mencionado, en lugar de analizar el peso y la estatura de los estudiantes en forma independiente, se analiza la relación que existe

entre esas dos variables. Igualmente, se puede estudiar la relación entre el número de horas dedicadas a prepararse para un examen, el coeficiente de inteligencia y la nota obtenida en el examen.

El enfoque multivariable –dos o más variables– obedece al hecho de que en la mayoría de los problemas estudiados intervienen muchos factores, por ello el análisis simultáneo de varias variables produce una mejor descripción y explicación del fenómeno. Tal es el caso, por ejemplo, del economista interesado en conocer el ingreso de los jefes de familia de una comunidad. En su estudio, él recoge información, no solo del ingreso mensual de los jefes, sino de otras características como sexo, edad, ocupación, estudios, experiencia, etc. Después de describir la distribución del ingreso y sus características, el economista estará en capacidad de indicar, por ejemplo, en qué extensión un cierto nivel de ingreso está asociado con el hecho de ser hombre o mujer, un obrero o un profesional, un nuevo empleado o uno con experiencia, de tener 20 o 60 años. Definitivamente, el investigador podrá contestar a más preguntas y estará mejor capacitado para explicar las características y peculiaridades de la distribución del ingreso y para estimaciones de su nivel, si conoce las variables asociadas al ingreso y el tipo e intensidad de la relación existente entre ellas.

Igualmente, puede postularse que el rendimiento de los estudiantes en una cierta prueba escolar depende de diferentes elementos, entre ellos la inteligencia, el tiempo que dedicaron a estudiar la materia, el interés por esta y la confianza en sus capacidades para resolver el examen.

Como se indicó, las técnicas tratadas hasta el momento no permiten enfrentar, adecuadamente, estos problemas; por ello, en el presente capítulo se describen e ilustran dos de las técnicas multivariables más usadas: la correlación y la regresión. Aunque no es fácil diferenciarlas de forma tajante, en la correlación puede decirse que las variables se estudian para descubrir si existe asociación entre ellas y, en caso de existir, medir su grado o intensidad. En la regresión, por otra parte, se estudia la naturaleza de la relación entre las variables y se trata de establecer una relación funcional que permite predecir una de ellas (variable dependiente), conociendo las otras (variables independientes).[2]

Aquí se hace referencia, básicamente, a la correlación y a la regresión lineal simple, es decir, a métodos y situaciones que involucran únicamente dos variables y en las cuales la relación que se postula entre las dos es lineal, es decir, puede representarse por una línea recta. Sin embargo, los procedimientos presentados pueden extenderse fácilmente, con las modificaciones pertinentes, a situaciones más complejas, como aquellas en las

2. Las técnicas de correlación y regresión suponen nivel métrico de medición de las variables analizadas. Existen otras técnicas específicas para el análisis multivariable de características medidas a nivel ordinal o nominal.

cuales se estudian más de dos variables (regresión y correlación múltiples), y a aquellas en las que la relación funcional supuesta entre las variables es no-lineal.

Para entrar más concretamente en el tema, considere un ejemplo simple:

A un grupo de 20 alumnos universitarios, escogidos al azar de un curso de Estadística, se les entregó una hoja de papel y se les pidió anotar la información indicada seguidamente:

1. Sexo (H: Hombre y M: Mujer).
2. Peso (kilos).
3. Estatura (cm).
4. Opinión a favor o en contra de que una mujer sea electa presidenta del país (F: Favorable y C: Contra).
5. Cantidad gastada en transporte a la semana (colones).
6. Tiempo en minutos que le toma al estudiante –en promedio– llegar a la universidad desde su casa (lugar donde reside).

En el caso de las tres últimas preguntas, se señaló que si no estaban seguros hicieran una estimación y, específicamente, para la suma gastada en transporte, se pidió que tomaran todos los gastos, excepto los extraordinarios; esto hizo, es posible, que quienes estudian y trabajan incluyeran también el traslado de la vivienda al lugar de trabajo. Se pidió, además, que los valores fueran redondeados al entero más próximo.

El sexo fue incluido, dentro de la información solicitada, porque se consideró muy probable que la opinión sobre una mujer presidenta se asocia con el sexo del estudiante.

Los datos recogidos en la forma antes descrita se incluyen a continuación. Note que se presenta la información (dato) de cada característica (variable o atributo) para cada estudiante (unidad de estudio).

Cuadro 14.1

CARACTERÍSTICAS PERSONALES Y OTRA INFORMACIÓN DE UNA MUESTRA DE 20 ALUMNOS DE UN CURSO UNIVERSITARIO DE ESTADÍSTICA

Número de estudiante	Sexo	Peso en kilos	Estatura en centímetros	Opinión mujer presidenta	Distancia a la universidad en kilómetros	Gasto semanal en colones	Tiempo de traslado en minutos
1	H	45	163	C	6	28	35
2	M	59	168	F	5	30	20
3	M	56	175	F	18	42	80
4	M	65	178	C	1	50	5
5	H	65	178	F	1	50	5
6	H	60	180	C	5	8	30
7	M	59	160	F	15	32	60
8	M	48	155	F	2	10	7
9	M	41	162	F	8	25	35
10	M	51	158	F	3	40	15
11	M	53	160	C	2	0	30
12	M	57	160	F	12	20	60
13	M	52	166	F	12	30	25
14	H	58	172	C	14	50	60
15	M	63	167	F	3	25	30
16	H	65	177	F	5	5	30
17	M	50	158	F	10	45	20
18	H	60	177	C	5	80	35
19	H	68	180	C	5	50	30
20	M	65	170	F	14	86	60

Para lograr una primera idea de los datos, pueden construirse las distribuciones de frecuencias y calcular algunas medidas básicas de posición y variabilidad. Esto aparece seguidamente:

1. Sexo		
	Frecuencia	%
Hombres	7	35
Mujeres	13	65
Total	20	100

4. Opinión mujer presidenta		
	Frecuencia	%
A favor	13	65
En contra	7	35
Total	20	100

3. Estatura en centímetros		
	Frecuencia	
155-159	3	$\bar{x} = 7,3$
160-164	5	Me = 168,2
165-169	4	$s_x = 8,41$
170-174	2	$cv = 5,0\%$
175-179	4	
180-184	2	
Total	20	

2. Peso en kilos		
	Frecuencia	
40-44	1	$\bar{x} = 57$
45-49	2	Me = 58,5
50-54	4	$s_x = 7,43$
55-59	5	$cv = 13,0\%$
60-64	4	
65-69	4	
Total	20	

5. Distancia en kilómetros		
	Frecuencia	
0-3	5	$\bar{x} = 7,3$
4-7	6	Me = 5
8-11	3	$s_x = 5,21$
12-15	5	$cv = 71,4\%$
16-19	1	
Total	20	

6. Gasto semanal en colones		
	Frecuencia	
0-19	4	$\bar{x} = 35,3$
20-29	4	Me = 31
30-39	3	$s_x = 22,61$
50-59	3	$cv = 64,0\%$
50-59	3	
80-89	3	
Total	20	

7. Tiempo en minutos		
	Frecuencia	
5, 7, 15	3	$\bar{x} = 33,6$
20	2	Me = 30
25	1	$s_x = 20,70$
30	5	$cv = 61,6\%$
35	3	
45	1	
60	4	
80	1	

Nota: las medidas de posición y variabilidad incluidas fueron calculadas con los datos sin agrupar. La moda fue excluida porque no estaba bien definida en la mayoría de las distribuciones.

14.3. ASOCIACIÓN ENTRE VARIABLES CUALITATIVAS

Como fue indicado, la información sobre el sexo se anotó porque la repuesta a la pregunta 4 –opinión sobre una mujer presidenta– podría estar asociada al sexo del entrevistado. Para investigar este punto, se clasifican las respuestas considerando, simultáneamente, la opinión y el sexo; el resultado de esta clasificación cruzada aparece en el cuadro 2. Note que los porcentajes se han calculado hacia abajo, tomando como base los totales para cada sexo; se procede así porque interesa determinar si la opinión (variable dependiente) está relacionada con el sexo (variable independiente).

Cuadro 14.2
ENTREVISTADOS SEGÚN SI ESTÁN A FAVOR O EN CONTRA DE QUE UNA MUJER SEA PRESIDENTE DEL PAÍS, POR SEXO

	Sexo del entrevistado (absolutos)			Sexo del entrevistado (porcentual)	
	Hombre	Mujer	total	Hombre	Mujer
A favor	2	11	13	29%	85%
En contra	5	2	7	71%	15%
Total	7	13	20	100%	100%

Dos características están **asociadas** si la distribución de los valores de una no es la misma para los diferentes valores de la otra. Por el contrario, son **independientes** –no asociada– cuando las distribuciones no difieren. En el presente caso, es evidente la asociación entre las variables, ya que la distribución de las opiniones cambia según el sexo que se considere; así, los hombres tienden a declararse en contra de que una mujer sea presidenta (71%), mientras que las mujeres apoyan la idea (85%).

Una forma muy simple de medir la asociación, en el caso de tablas de 2 X 2 (cuatro casillas) como la considerada en este ejemplo, es usando la **diferencia porcentual**, o sea, la diferencia entre los porcentajes de las dos columnas en un mismo renglón. Cuando las variables no están asociadas, las distribuciones son iguales en cada una de las columnas y la diferencia porcentual para cada fila es cero (caso 1, próxima página). Por el contrario, la diferencia porcentual es de 100 cuando están perfectamente asociadas (caso 2). La diferencia asumirá valores intermedios según sea el grado de asociación entre las variables: bajo, moderado o alto.

En el presente caso (cuadro 14.2), el cálculo de la diferencia porcentual indica que hay una diferencia de 85% – 29% = 56% entre las proporciones de mujeres y de hombres que apoyan la idea de que una mujer sea presidenta, la cual señala una relación moderadamente alta entre la opinión y el sexo.

Caso 1

		Variable x	
		A	B
Variable y	C	50%	50%
	D	50%	50%

Caso 2

		Variable x	
		A	B
Variable y	C	100%	0%
	D	0%	100%

DIF PORCENTUAL 50 –50 = 0 DIF PORCENTUAL: 1 – 100 = – 99

La diferencia porcentual puede definirse únicamente cuando se tienen tablas de 2 X 2. Existen otras medidas para calcular el grado de asociación entre variables cualitativas, en tablas de 2 X 2 y de más dimensiones; estas medidas aparecen en los textos de métodos estadísticos y son suministradas rutinariamente por los programas computacionales especializados en el manejo de datos de encuestas y de estudios estadísticos. Esas medidas no se discutirán en este apartado, pues se concentra totalmente en la asociación entre variables métricas.

14.4. CORRELACIÓN LINEAL SIMPLE

Los problemas de correlación simple son aquellos en los cuales interesa medir la intensidad de la asociación entre dos variables métricas, por ejemplo, entre peso y estatura.

Se sabe, por experiencia, que conforme crece un joven va aumentando su peso y estatura y, en general, cuanto mayor es el peso, mayor es la estatura y viceversa (aunque también hay personas que son altas y flacas y bajas y gordas). Resulta razonable postular, entonces, que entre el peso y la estatura existe una relación directa. Si se toma una muestra de personas y para cada una se anota su peso y estatura, como en el ejemplo, se tiene la oportunidad de examinar, de manera empírica, si en el conjunto de datos se da esa relación esperada. Se asigna arbitrariamente la letra x al peso y la y a la estatura.

El problema concreto es muy claro: se han hecho observaciones sobre peso (x) y estatura (y) para cada uno de los integrantes de la muestra de 20 alumnos y se quiere determinar, en primer término, si existe asociación entre esas variables y luego proceder a la medición de la intensidad de esa asociación. Para ello podría construirse una tabla como la del caso de la opinión sobre una mujer presidenta y el sexo. Sin embargo, como se trata de variables cuantitativas, una forma más apropiada es presentar primero las observaciones en un **diagrama de dispersión**, o sea, representando en un sistema de coordenadas rectangulares los pares de valores x e y correspondientes a los elementos de la muestra.

Para construir el diagrama de dispersión, se define una escala para x (abscisa) y otra para y (ordenada), luego cada estudiante es representado en él por un punto, cuyas

coordenadas corresponden a su peso y estatura. Esto es lo que se ha hecho en el gráfico de la figura 14.1. Note que las escalas no se inician en (0,0), sino en los valores –más cómodos– 38 y 150; en modo alguno, afecta la presentación y análisis de los datos, pues interesa examinar la asociación entre los puntos y no su distancia a los ejes.

El diagrama de dispersión, correspondiente al peso y la estatura de los 20 estudiantes del curso de Estadística, es el siguiente:

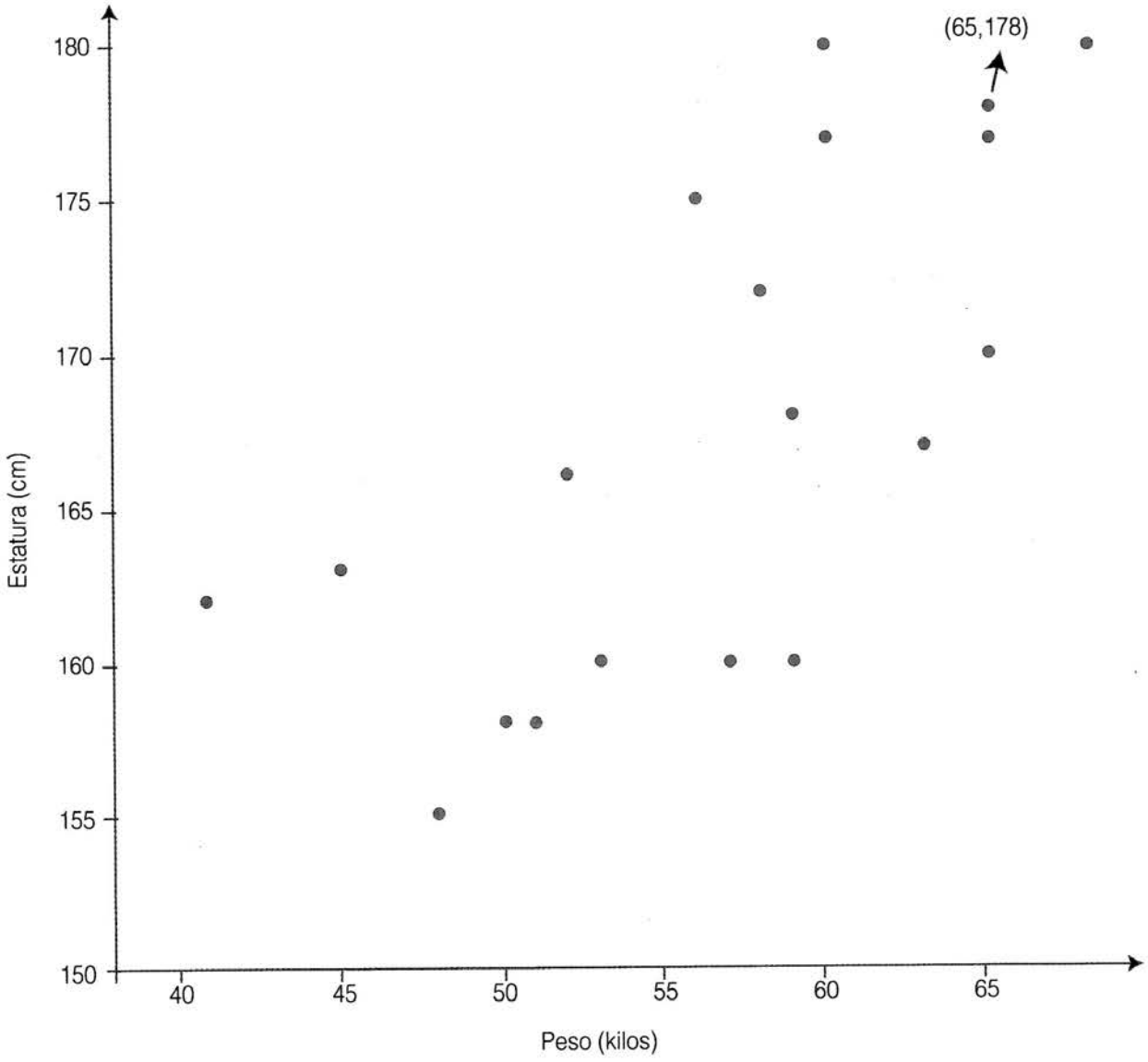


Figura 14.1. Gráfico del diagrama de dispersión para el peso y la estatura de 20 estudiantes
Nota: el punto (65,178) es doble, ya que representa los casos 4 y 5.

Puede observarse, en primer término, que los datos siguen la relación esperada: los estudiantes que pesan más tienden a ser los más altos y, por el contrario, los de menor peso son los de menor estatura. Por otra parte, hay bastante variabilidad; así, por ejemplo, quienes pesan alrededor de 65 kilos varían en su estatura aproximadamente entre 170 y 180 cm. Debe concluirse, por lo tanto, que entre peso y estatura existe una relación lineal directa pero no es perfecta, ya que si lo fuera, los puntos configurarían una línea recta en el gráfico.

¿Cómo puede medirse esa relación lineal? Con el **coeficiente de correlación lineal** r, también conocido como coeficiente de correlación de Pearson, el cual se define de la siguiente forma:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

donde:

s_{xy} : representa la covariancia de xy;

s_x : la desviación estándar de x;

s_y : la desviación estándar de y.

La covariancia de xy mide la asociación lineal entre las variables x e y, puede ser negativa, positiva o 0, dependiendo del grado y tipo de relación que exista entre x e y. Su definición formal es la siguiente:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}\right).$$

Cuando se tienen los datos originales, el coeficiente de correlación lineal puede calcularse fácilmente utilizando alguna de las siguientes expresiones:

$$r_{xy} = \frac{\frac{1}{n-1}\left(\sum x_i y_i - \frac{\sum x_i y_i}{n}\right)}{\sqrt{\left(\frac{1}{n-1}\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]\right)\left(\frac{1}{n-1}\left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]\right)}}$$

$$= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}. \qquad (2)$$

Seguidamente se ilustra, con un pequeño conjunto de datos ($n = 5$), el uso de la fórmula (2), que es la más cómoda, para calcular el valor de r. Esta fórmula depende de los siguientes valores:

n = número de pares de valores x e y con que se cuenta, o sea, tamaño de la muestra;

$\sum x_i$ = suma de todos los valores de x;

$\sum y_i$ = suma de todos los valores de y;

$\sum x_i y_i$ = suma de los productos cruzados de x e y;

$\sum x_i^2$ = suma de los cuadrados de los valores de x;

$\sum y_i^2$ = suma de los cuadrados de los valores de y.

Note también que la información se ha organizado en forma tabular, con el fin de facilitar el cálculo de las sumatorias de productos cruzados y cuadrados que requiere el cómputo del coeficiente de correlación.

Ejemplo 1

i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	2	4	8	4	16
2	3	2	6	9	4
3	3	4	12	9	16
4	4	3	12	16	9
5	5	1	5	25	1
sumas	17	14	43	63	46

$$r = \frac{5 \cdot 43 - 17 \cdot 14}{\sqrt{(5 \cdot 63 - (17)^2)(5 \cdot 46 - (14)^2)}} = \frac{-23}{\sqrt{26 \cdot 34}} = \frac{-23}{29,7} = -0,77.$$

En el ejemplo del peso y la estatura de los estudiantes que se ha estado comentando, se definió x = peso y y = estatura; a partir de los datos de la sección 14.2, cuadro 14.1, se obtiene:

$\sum x_i = 1135$ $\quad \sum x_i^2 = 65\,403$ $\quad \sum x_i y_i = 191\,121$

$\sum y_i = 3354$ $\quad \sum y_i^2 = 563\,710$ $\quad n = 20$

Aplicando la expresión anterior, el valor de r_{xy} es el siguiente:

$$r_{xy} = \frac{20 \cdot 191121 - 1135 \cdot 3354}{\sqrt{(20 \cdot 65\,403 - (1135)^2)(20 \cdot 563\,710 - (3354)^2)}}$$

$$= \frac{15630}{\sqrt{19\,835 \cdot 24\,884}} = \frac{15630}{221 - 22\,216,5} = 0,7035.$$

El valor obtenido $r_{xy} = 0,70$, como se verá en el próximo apartado, revela una correlación lineal directa moderadamente alta entre peso y estatura y es coherente con lo que muestra el diagrama de dispersión. Se usa el término "directa" cuando, al aumentar una variable, la otra también se incrementa; el término "inversa" cuando, al aumentar una, la otra disminuye.[3]

14.5 INTERPRETACIÓN DEL COEFICIENTE DE CORRELACIÓN LINEAL SIMPLE *(r)*

El coeficiente de correlación lineal r solo puede asumir valores entre –1 y +1, ambos inclusive. En símbolos: $-1 \leq r \leq 1$.

La interpretación de r, por lo tanto, descansa en dos elementos: su valor, que indica la intensidad o grado de asociación, y su signo, que señala el tipo de asociación lineal: positiva o directa (+) y negativa o inversa (–). Un coeficiente de correlación igual a + 1, indica una relación lineal perfecta positiva (o directa) entre las variables analizadas: todos los puntos se ubican sobre una línea recta y al aumentar x aumenta y, a cada valor x corresponde un solo valor de y (diagrama A, gráfico de la figura 14.2). Una r igual a + 1 señala también una relación perfecta, pero en este caso negativa (o inversa) (diagrama B, gráfico de la figura 14.2).

La **ausencia de relación lineal** es indicada por un $r = 0$. Debe señalarse, sin embargo, que pueden darse dos situaciones:

a) En la primera no existe relación lineal ni de ningún otro tipo entre las variables.

b) En la segunda, las dos variables no muestran asociación lineal, pero sí tienen otro tipo de relación (cuadrática o cúbica, por ejemplo).

La primera situación se ilustra en el diagrama C; como puede notarse, los puntos se dispersan sin guardar ningún tipo de asociación definida, no puede afirmarse que, al aumentar x, y tienda a aumentar o a disminuir. Un ejemplo de la otra situación lo constituye el diagrama D, aunque r es 0, y el gráfico muestra que no existe asociación lineal entre las variables, eso no implica que no están asociadas, al contrario, los puntos describen una curva de segundo grado y es evidente una asociación perfecta de tipo cuadrática. Debe quedar claro, por lo tanto, que un $r = 0$ indica ausencia de correlación lineal entre las dos variables analizadas, pero no excluye la posibilidad de que haya otro tipo

3. El cálculo del coeficiente de correlación, de las desviaciones estándar y de la covariancia, al igual que de otras medidas comprendidas en las secciones siguientes, puede realizarse muy cómodamente digitando los datos en la hoja de cálculo Excel, incluida en el Microsoft Office, y recurriendo a las funciones estadísticas que aparecen en esa hoja de cálculo.

de asociación. Por ello, un investigador que encuentra un $r = 0$ muy pequeño, no debe concluir, precipitada y confiadamente, que las variables en estudio no están asociadas, más bien debe explorar la posibilidad de que tengan otro tipo de asociación. El diagrama de dispersión cumple, en este aspecto, una función muy útil porque permite decidir rápidamente si debe o no investigarse la existencia de una relación no-lineal.

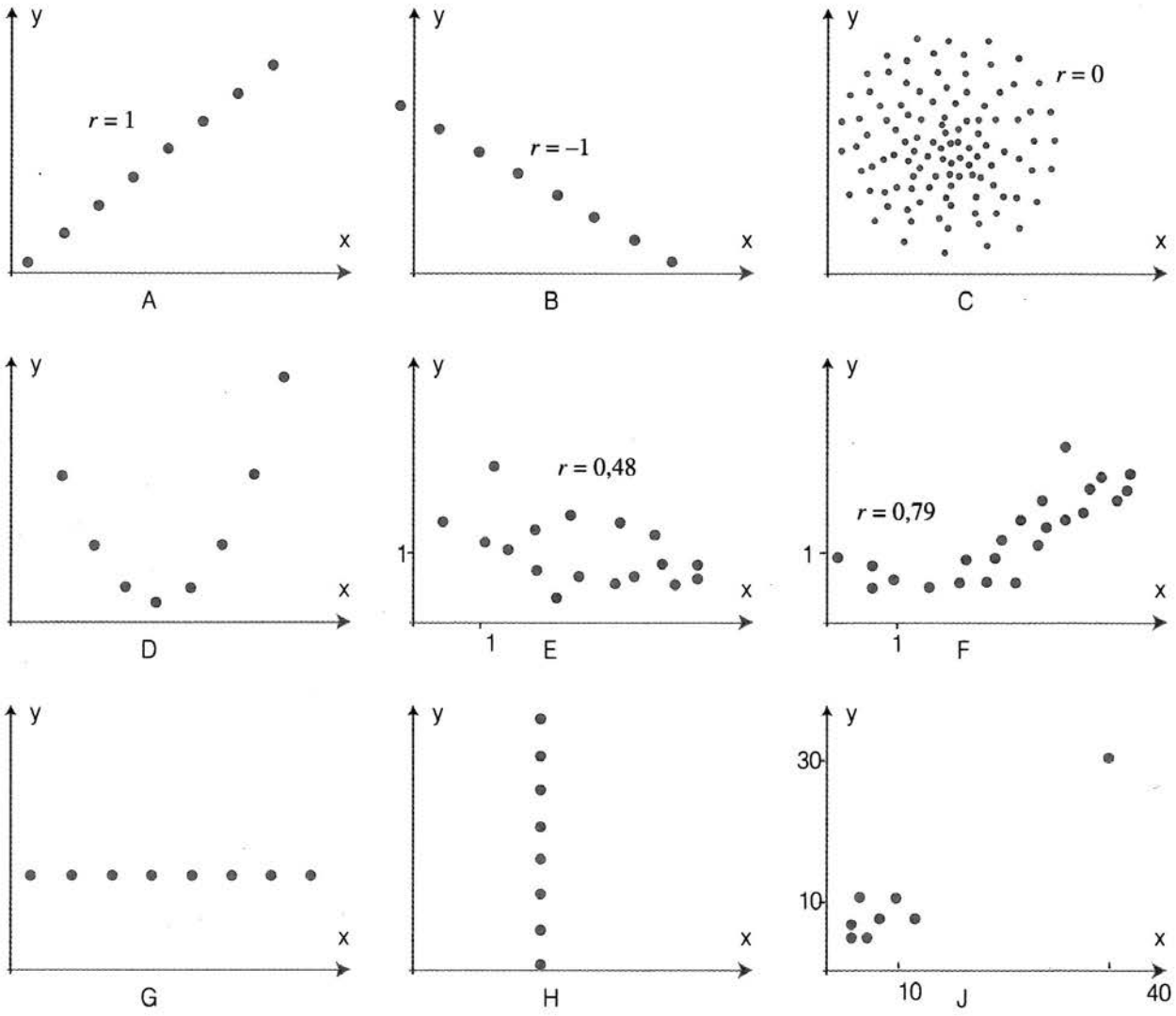


Figura 14.2. Gráfico de los diagramas de dispersión para diferentes valores de r

De las situaciones extremas discutidas anteriormente ($r = 0$ y $r = \pm 1$), se evidencia que, cuanto más cercano a la unidad (en valor absoluto) sea el valor de r, mayor será el grado de relación lineal entre las variables analizadas, y entre más cercano sea a 0, menor será esa asociación.

En los diagramas E y F se muestran dispersiones correspondientes a valores de $r = -0,48$ y $r = 0,79$. El segundo refleja un grado de correlación alto, y el primero uno que puede considerarse moderado; cabe aclarar, sin embargo, que en la realidad no hay reglas fijas acerca de cómo decidir cuándo un coeficiente de correlación puede considerarse alto o bajo. Esto depende de varios factores, entre ellos el grado de correlación encontrado usualmente en un cierto campo del conocimiento y el uso práctico que se pretende dar a la asociación encontrada.[4]

En el ejemplo discutido anteriormente (sección 14.4), al relacionar la estatura con el peso en un grupo de 20 estudiantes, se obtuvo $r = 0,73$. Este valor indica que existe un grado de correlación lineal moderadamente alto entre las variables peso y estatura.

Dos situaciones en las cuales no existe asociación entre las variables analizadas, y r no está definida, son las ilustradas en los diagramas G y H; como se observa, no hay asociación entre x e y. En la situación G, para cualquier valor de x se da un mismo valor de y; mientras que en el caso H, para cualquier valor de y existe un único valor de x.[5]

Conviene señalar, también, que la impresión visual obtenida del diagrama de dispersión puede resultar engañosa, como guía, para estimar el valor de r; por ello, es recomendable calcular siempre el coeficiente de correlación, aun cuando a simple vista parezca que la asociación es muy alta o muy baja. Para ilustrar este punto, considere el diagrama J del gráfico de la figura 14.2, trate de estimar el valor de r con base en su impresión visual y anótelo; luego, proceda a calcular el valor de r según los datos incluidos a continuación, estos son los mismos que se usaron para construir el diagrama de dispersión citado.

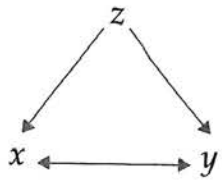
x	y	x	y
5	10	8	7
4	6	10	10
4	4	13	7
6	4	40	30

¿Cómo se compara el valor estimado visualmente, con el calculado? Ahora repita el cálculo eliminando el valor extremo (40,30) y analice de nuevo la situación, ¿a qué conclusiones llega?

4. En los estudios de rendimiento académico, por ejemplo, no son frecuentes correlaciones de más de 0,50; en cambio, en los experimentos físicos son normales valores mucho mayores y aún cercanos a 1.
5. El cálculo aritmético de r por la fórmula correspondiente llevaría a $\frac{0}{0}$, o sea, a una indeterminación, ya que ambos casos, tanto la covariancia como una de las desviaciones estándar, son iguales a cero.

14.6. CORRELACIÓN Y CAUSALIDAD

En ciertos casos, la correlación entre dos variables se debe a que una de ellas es causa de la otra; sin embargo, no es así en otras oportunidades. Por eso, cuando se analizan coeficientes de correlación, siempre debe tenerse en mente que el hecho de que dos variables muestren cierta asociación lineal **no indica, necesariamente**, que una tenga un efecto causal directo o indirecto sobre la otra. En realidad, la interpretación de r, como medida del grado de relación lineal entre dos variables, es una interpretación puramente matemática, desprovista completamente de relación de causa y efecto. La existencia de una correlación elevada entre dos variables x e y no permite concluir que una sea la causa de la otra; la relación lineal (o de otro tipo), puede deberse a una tercera variable z, es causa común de ambas, esto hace, por lo tanto, que x e y muestren correlación. Simbólicamente, puede representarse en la siguiente forma:



Debe notarse que la correlación es real y la relación estadística entre x e y existe, no se infiere que x es la causa de y o a la inversa. El error no está en afirmar que existe correlación, porque eso es cierto, sino en derivar relaciones de causa y efecto con base en una correlación observada. La mayoría de los errores cometidos al aplicar los métodos estadísticos, de correlación y al interpretar sus resultados provienen de la inclinación natural de las personas a inferir relaciones causales a partir de asociaciones estadísticas.

En los textos de estadística se ofrecen ejemplos de asociaciones entre variables causadas por una variable común –tercera variable– sin que exista relación de causa y efecto entre las analizadas. Dos ejemplos muy antiguos, pero ilustrativos, son los siguientes:

El profesor Yule, un distinguido estadístico de la primera parte del siglo XX, al analizar una serie de años, observó una correlación positiva entre la proporción de matrimonios por la iglesia y la tasa de mortalidad en Inglaterra. Una baja proporción de matrimonios por la iglesia coincidía con una tasa de mortalidad baja y una alta proporción de matrimonios por la iglesia coincidía con una tasa de mortalidad elevada. Una inferencia causal, basada en estos datos, llevaría a la conclusión de que la supresión de los matrimonios religiosos reduciría la tasa de mortalidad. Esta conclusión, por supuesto, sería absurda. En realidad, la relación estadística positiva entre ambas variables se debía a que, en Inglaterra, por muchos años, tanto el número de matrimonios como la

mortalidad estuvieron disminuyendo; esa evolución en el tiempo fue la responsable de la relación positiva observada por el profesor Yule.[6]

Una relación positiva fue observada entre el número de cigüeñas y el número de nacimientos en Europa Noroccidental. Solo una persona muy romántica e ingenua podría afirmar, con base en esta evidencia estadística, que la leyenda acerca de que los niños son traídos por las cigüeñas es cierta. Una explicación más prosaica, pero con mayor verosimilitud, es que con el aumento de la población no solo hay más nacimientos, sino mayor construcción de edificios, esto aumenta el número de lugares donde las cigüeñas pueden construir sus nidos.[7]

Una tercera variable también puede actuar para producir una relación que resulte contraria a las expectativas de un investigador o de un lector, basadas en un modelo de relación causal; como ilustración, considere un ejemplo hipotético. Suponga que un profesor universitario de estadística ha recogido datos de diez estudiantes sobre la nota en Filosofía y el número de horas semanales dedicadas a estudiar esta materia. El profesor espera que los datos muestren una correlación positiva (directa), es decir, que quienes dedican más tiempo al curso, obtengan una mejor nota. Sin embargo, al construir el diagrama de dispersión, encuentra lo siguiente:

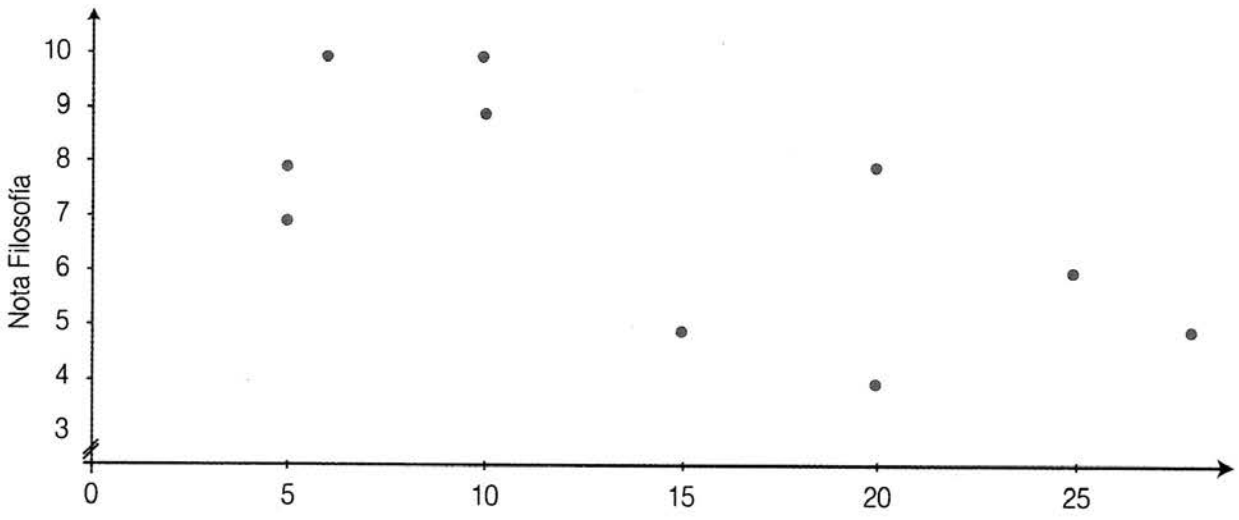


Figura 14.3. Gráfico de la nota en Filosofía y número de horas semanales dedicadas a estudiar esta materia en una muestra de 10 estudiantes

6. G. U. Yule (1921), Why do we sometimes get nonsense correlations? Journal of The Royal Statistical Society, Vol. 85, p. 95.
7. Wallis y Robert. (1956). Statistics: A New Approach. Illinois: Glencoe.

Los resultados son evidentemente sorpresivos: la relación es inversa y fuerte ($r = 0,65$). Las cifras señalan que quienes estudian más horas obtienen notas más bajas. Algún incauto podría sugerir que la mejor manera de salir bien en Filosofía es no estudiando la materia; otro más reflexivo tiene en cuenta que estudiar perjudica porque cansa, el cansancio enferma y la enfermedad reduce la capacidad de estudio y la habilidad para responder correctamente el examen. La anterior paradoja se resuelve si se considera información adicional como el Coeficiente de Inteligencia (CI) (cuadro 14.3) y los diagramas de dispersión y coeficientes de correlación de las variables (gráfico de la figura 14.4).

Cuadro 14.3

NOTA EN FILOSOFÍA, HORAS DE ESTUDIO SEMANALES Y COEFICIENTES DE INTELIGENCIA PARA UNA MUESTRA DE 10 ESTUDIANTES

Número de estudiante	Horas de estudio (x)	Nota de filosofía (y)	Coeficiente de inteligencia (z)
1	10	10	110
2	25	6	85
3	6	10	120
4	15	5	90
5	28	5	80
6	20	4	85
7	20	8	90
8	5	7	100
9	10	9	100
10	5	8	110

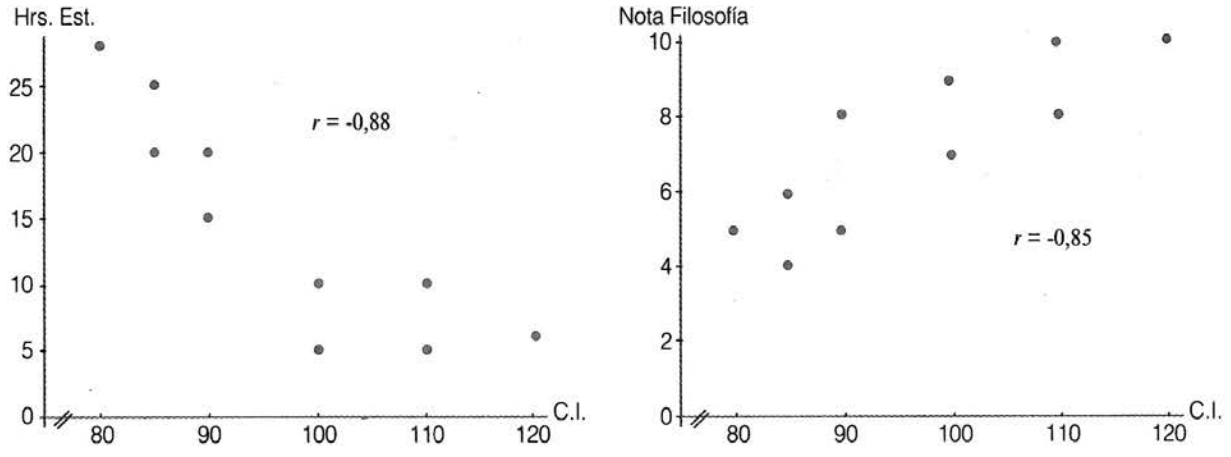


Figura 14.4. Gráfico que muestra el diagrama de dispersión para los datos del cuadro 14.3

Se tiene que la correlación entre el CI y las horas de estudio es fuerte pero inversa $(r = -0,88)$, lo cual parece natural: los estudiantes más inteligentes requieren una menor cantidad de horas de estudio para obtener una buena nota. A su vez, la relación entre la nota en Filosofía y el CI es positiva $(r = 0,85)$, lo que también parece razonable.

Resulta, entonces, que la relación inversa observada entre horas de estudio y nota en Filosofía, se debe, básicamente, a que los estudiantes más inteligentes requieren menos horas de estudio para obtener una buena nota. La variable CI está actuando simultáneamente sobre las horas de estudio (en forma inversa) y sobre la nota en Filosofía (en forma directa), y provocando una asociación contraria a la esperada entre estas variables.

Para terminar estas observaciones sobre correlación y causalidad, conviene reiterar que la existencia de una asociación estadística entre dos variables puede ser una evidencia válida de que las variables están conectadas causalmente, pero la existencia de correlación en modo alguno, implica causalidad y no garantiza, necesariamente, que haya una relación causal entre las variables. En realidad, la correlación observada puede tener diferentes orígenes, tal como se indica:

a) **Relación causal**: las variables (o factores) pueden estar relacionadas causalmente.

   Así, x puede causar y, y puede causar x o ambas pueden causarse en forma recíproca.

   Simbólicamente:

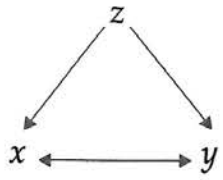
   $x \rightarrow y$ La flecha se utiliza aquí para indicar casualidad.

   $y \leftarrow y$ La variable situada en el origen de la flecha es la causa.

b) **Tercera variable como causa común:** la correlación entre dos variables puede deberse a la presencia de una tercera variable o factor que, al causar x y y, produce la correlación observada. En este caso, la correlación existe, es real, pero no por una relación causal entre x e y sino por la presencia del tercer factor. Las correlaciones de este tipo se han denominado espurias, o sea, no verdaderas; sin embargo, dado que existe, más bien se considera espuria la interpretación causal de la correlación estadística y no la correlación en sí.

   Simbólicamente:

   El símbolo $\leftrightarrow$ se utiliza para indicar la existencia de correlación entre las variables x, y.

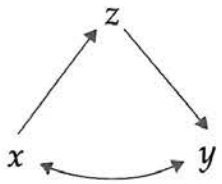


En este caso en el cual la correlación es producida por la tercera variable z, es evidente que si se mantiene constante, es decir, no varía, la correlación entre x e y desaparecerá.

c) **Tercera variable como interviniente**: la relación causal entre x e y se produce, en ciertas oportunidades, a través de una tercera variable z (interviniente). Así, x causa z y z, a su vez, causa y. En este caso, la correlación entre x e y no es ficticia, sino real o verdadera, pero se produce a través de una variable interviniente.

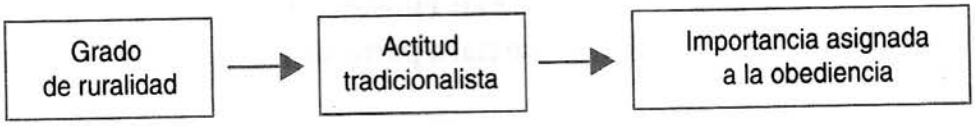
Simbólicamente:

$x \rightarrow z \rightarrow y$ o



En este caso, al igual que en el anterior, si el factor z permanece constante o no se lo deja variar, la correlación entre x e y desaparece. Si el investigador no conoce el orden causal de las variables, le será imposible determinar si z es una variable interviniente o si se trata de una causa común, por lo tanto, tampoco podrá decidir si es una correlación espuria o de una correlación verdadera.

Para ilustrar este punto de la variable interviniente suponga que, en un estudio, se encuentra una relación positiva entre grado de ruralidad del lugar de residencia de la madre e importancia asignada a la obediencia como una cualidad que deben tener los niños. En este caso, puede plantearse la siguiente relación:



Se postula que el grado de ruralidad produce actitudes tradicionalistas y estas tienden, a su vez, a que se asigne una mayor importancia a la obediencia. La relación entre ruralidad e importancia asignada a la obediencia se da a través de la variable interviniente: actitud tradicionalista.

14.7. CORRELACIÓN PARCIAL

En ciertas situaciones, como las mencionadas en la sección anterior, se desea determinar en qué medida una correlación observada entre dos variables se deba a una tercera variable que actúa como causa común. En otras palabras, si la correlación entre x e y es real o si es causada, total o en parte, por una variable común z. Para saberlo, es necesario examinar la correlación entre x e y, manteniendo constante z o, como se dice usualmente, eliminando el efecto de z. Esto se logra al utilizar el **coeficiente de correlación parcial**, el cual se representa con el símbolo r_{xyz} y cuya fórmula es la siguiente:

$$r_{xyz} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

En este caso, r_{xyz} representa la correlación entre x e y después de haber controlado o eliminado el efecto de z. Las dos variables antes del punto indican la correlación que interesa; la situada después del punto, la variable que se controla. Si es más importante la correlación entre x y z, controlada, el coeficiente se escribe r_{xzy}, por lo tanto, la fórmula indicada arriba sufriría las modificaciones pertinentes. El cálculo de la correlación parcial depende únicamente de los coeficientes de correlación simples entre las variables bajo análisis : r_{xy}, r_{xz} y r_{yz}.

Para ilustrar el uso del coeficiente de correlación parcial, considere el ejemplo de la nota de Filosofía, las horas de estudio y el coeficiente de inteligencia (CI) discutido en la sección anterior, en el cual interesaba determinar por qué la correlación entre horas de estudio y nota de Filosofía, contrariando el sentido común, resultaba negativa y elevada $(r = 0,65)$, pues lo esperado es que fuera elevada pero positiva. Un análisis de los diagramas de dispersión y de los coeficientes de correlación simples llevó a la conclusión de que la razón era la existencia de una relación inversa entre el CI y las horas de estudio, y positiva entre el CI y la nota en Filosofía. El problema puede resolverse calculando el coeficiente de correlación parcial a partir de los coeficientes de correlación simples de las variables:

x_1: Nota de filosofía — $r_{13} = -0,65$ (Filosofía y horas estudio)

x_2: Horas de estudio — $r_{13} = 0,85$ (Filosofía y CI)

x_3: Coeficiente inteligencia (CI) — $r_{23} = -0,88$ (Horas de estudio y CI)

$$r_{12,3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} = \frac{-0,65 - (0,85)(-0,88)}{\sqrt{(1 - (0,85)^2)(1 - (-0,88)^2)}}$$

$$= \frac{-0,65 + 0,75}{\sqrt{(0,2775)(0,00626)}} = \frac{0,10}{0,25} = 0,40.$$

El valor obtenido de $r_{12,3} = 0,40$ indica una correlación positiva y moderadamente alta entre la nota de Filosofía y horas de estudio, una vez que se controla o elimina el efecto de la inteligencia (CI). El resultado, como puede notarse, es totalmente coherente con la expectativa de que, al dedicar más tiempo al estudio, se obtienen mejores notas.

La técnica de correlación parcial puede generalizarse a situaciones donde se desea controlar más de una variable. Este tema, sin embargo, queda fuera del alcance del presente texto.[8]

14.8. LA REGRESIÓN LINEAL SIMPLE

Considere ahora la información presentada en el cuadro 14.1 sobre la distancia que hay entre el lugar donde se vive y la universidad y el tiempo que se toma para llegar a ella. Entre estas dos variables cabe esperar una correlación positiva significativa. Como la distancia determina el tiempo, se ha identificado este con y y la distancia con x.

Seguidamente, se presentan los valores pertinentes y el cálculo del coeficiente r.

$$\sum x = 155 \quad \sum y = 712 \quad \sum x^2 = 1681 \quad \sum y^2 = 32\,724 \quad \sum xy = 7119.$$

$$r = \frac{20 \cdot 7119 - 155 \cdot 712}{\sqrt{[20 \cdot 1681 - (155)^2][20 \cdot 32\,724 - (712)^2]}} = \frac{32\,020}{\sqrt{9595 \cdot 147\,536}}$$

$$= \frac{32\,020}{37\,624,6} = 0,851.$$

El valor obtenido indica que entre la distancia y el tiempo existe una relación lineal directa bastante alta, muy cercana a +1. Esto lo confirma el diagrama de dispersión (gráfico de la figura 14.5), en el cual se observa cómo los puntos que representan a los estudiantes se sitúan muy cerca de una línea recta, y que las distancias cortas corresponden a tiempos reducidos y las distancias largas a tiempos mayores.

Resulta de interés desarrollar algún procedimiento o fórmula que permita calcular, aproximadamente, el tiempo que le toma a un estudiante llegar a la universidad, conociendo la distancia a la que vive de ella. Este es un problema típico en el análisis de datos y se denomina –en estadística– con el nombre de **regresión**. En la regresión, como se indicó antes, interesa una variable dependiente (y) y la otra (u otras), llamada variable independiente se considera por la posibilidad que brindan de estimar o predecir la variable dependiente usando un cierto modelo o relación funcional.

8. Los procedimientos correspondientes aparecen en cualquier texto de métodos estadísticos o en los programas computacionales de análisis estadístico.

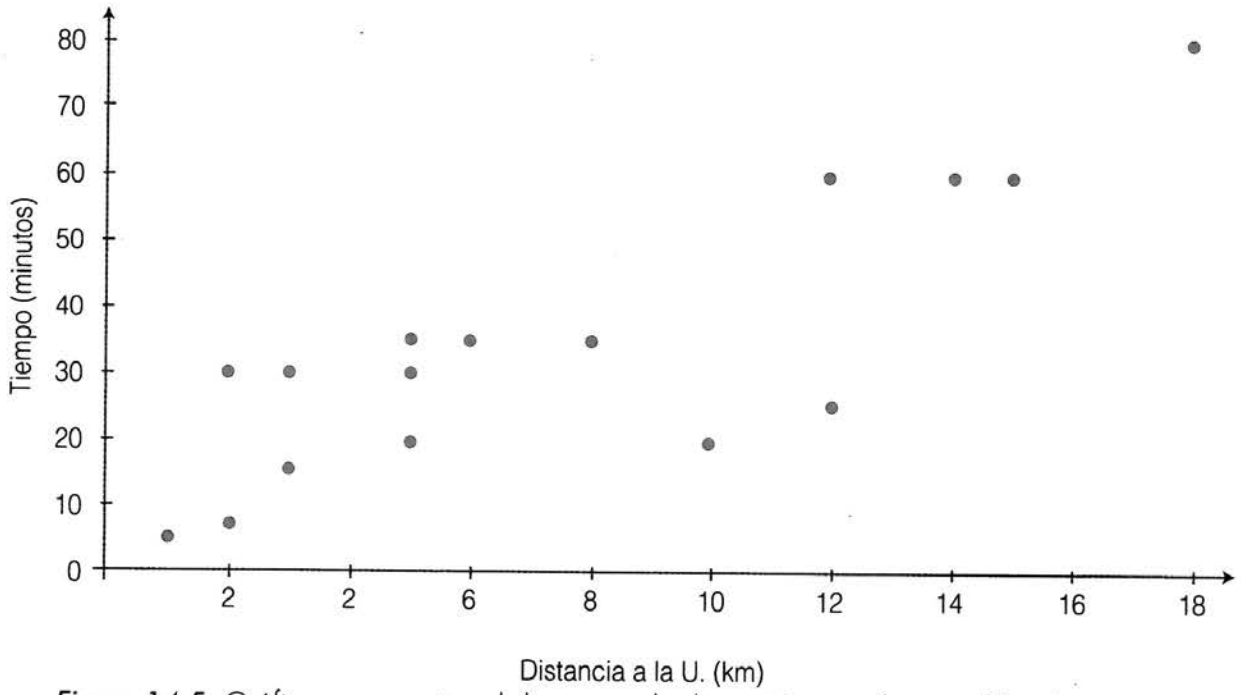


Figura 14.5. Gráfico que muestra el diagrama de dispersión para las variables distancia a la universidad (km) y tiempo que toma llegar a la universidad (min)
Nota: en la posición (5,30) coinciden tres puntos y en la (14,60) dos.

Básicamente, el problema consiste en ajustar a los datos, con un cierto criterio, un modelo que permita describirlos y pronosticar o inferir la variable dependiente y, conociendo la independiente x.

En este caso, dada la alta correlación lineal que hay entre las variables, es totalmente apropiado plantear un modelo de regresión lineal, que exprese el tiempo como función de la distancia a la que se reside de la universidad. La expresión funcional es la siguiente:

$$y = a + bx.$$

Donde:

y : tiempo en minutos que toma llegar a la universidad.

x : distancia, en kilómetros, a la que se vive de la universidad.

a y b: constantes que definen la ecuación lineal de regresión, deben estimarse utilizando los datos de la muestra de 20 alumnos.

Geométricamente, la a es la ordenada del punto $(0,a)$ donde la recta corta al eje de ordenadas y, se denomina intersección; la b es la pendiente de la recta, la cual puede calcularse cuando se tienen dos puntos (x_1, y_1) y (x_2, y_2) con la expresión $b = \dfrac{y_1 - y_2}{x_1 - x_2} = \dfrac{\Delta y}{\Delta x}$.

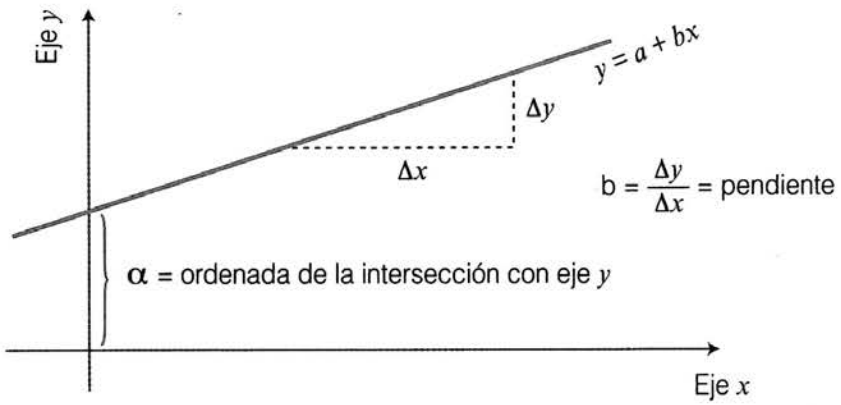


Figura 14.6. Modelo de regresión lineal que expresa el tiempo como función de la distancia

Como en este caso se tiene una muestra de 20 puntos, se debe buscar una forma para obtener *a* y *b* usando esa información.

Una posibilidad para estimar *a* y *b* es trazando una recta que pase entre los puntos del diagrama de dispersión, se procede luego a obtener del gráfico los valores de *a* y *b*. Este método se denomina de *mano alzada* o gráfico, los resultados dependen del juicio y destreza de quien realiza el ajuste, variará según la persona que lo haga. Por esta razón, es preferible hacer el ajuste siguiendo un procedimiento objetivo, matemático, que garantice el mismo resultado sin importar quien haga el ajuste.

El procedimiento más usado se denomina *criterio de cuadrados mínimos*, recibe este nombre porque su condición es que la línea de regresión ajustada haga mínima la suma de los cuadrados de las diferencias entre los valores reales observados (Y_{ob}) y los valores estimados mediante la ecuación de regresión ajustada a los valores muestrales (*Yest*). Simbólicamente, la condición es la siguiente:

$$y_{est} = a + bx \quad \text{tal que} \quad \sum (y_{ob} - y_{est})^2 \quad \text{sea mínima.}^{9}$$

En el caso concreto de la regresión lineal, la aplicación del criterio de cuadrados mínimos produce el siguiente sistema de ecuaciones:

$$\sum y = na + b\sum x.$$

$$\sum xy = a\sum x + b\sum x^2.$$

9. Un criterio, igualmente válido, sería pedir que la suma de los valores absolutos de las diferencias entre los valores observados y los estimados con la ecuación de regresión sea un mínimo. Sin embargo, este criterio no se utiliza porque obligaría a manejar valores absolutos, lo que es incómodo; por otra parte, los cuadrados de las diferencias cumplen la misma función son más cómodos de manejar y tienen otras ventajas matemáticas y estadísticas.

La solución de este sistema produce dos fórmulas muy conocidas para el cálculo de a y b:

$$b = \frac{\sum xy - n\overline{xy}}{\sum x^2 - n\overline{x^2}} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}, \quad a = \overline{y} - b\overline{x}.$$

Si se aplican las fórmulas anteriores a los datos de distancia a que se vive de la universidad (x) y tiempo que tarda en llegar a ella (y), del ejemplo comentado *(cuadro 14.1)* se obtienen los valores de a y b, los cuales definen la línea de regresión.[10]

$$b = \frac{20 \cdot 7119 - 155 \cdot 71,2}{20 \cdot 1681 - (155)^2} = \frac{32\,020}{9595} = 3,337.$$

$$a = 35,6 - 3,337 \cdot 7,75 = 9,738.$$

De acuerdo con estos resultados, la ecuación de la recta de regresión es:

$$y = 9,738 + 3,337x.$$

En el gráfico de la figura 14.7, esta línea de regresión ajustada se representa en el diagrama de dispersión:

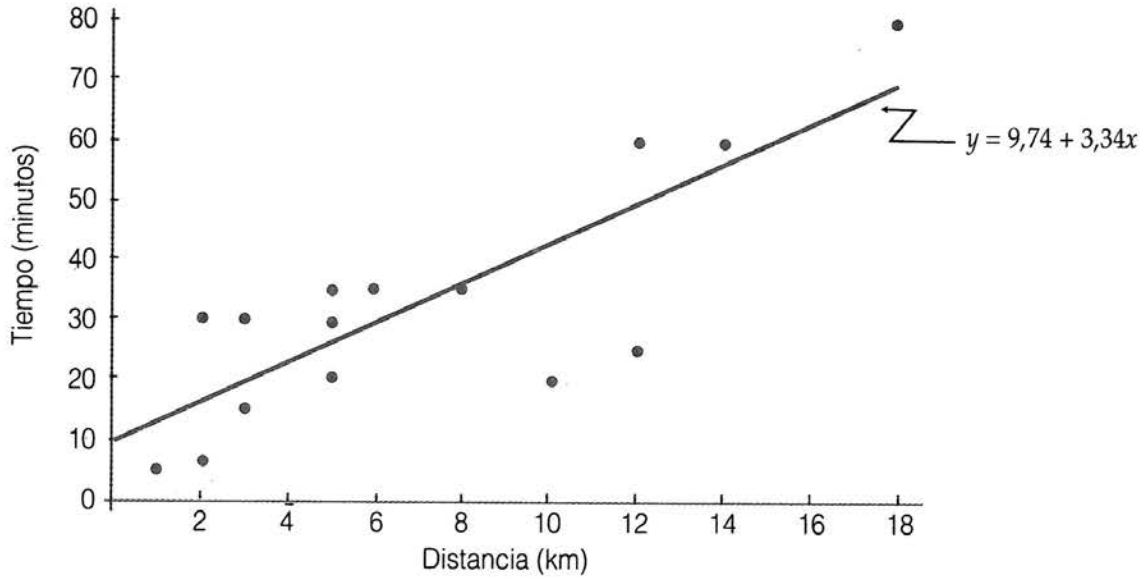


Figura 14.7. Gráfico que muestra la representación de la línea de regresión calculada para el tiempo que se tarda y distancia a la que se vive de la universidad

La ecuación ajustada es un modelo lineal que expresa la relación entre las variables dependiente y la independiente. Por medio de esta es posible estimar el tiempo promedio que le toma al estudiante llegar a la universidad, conociendo la distancia a la que se vive de ella. Por ejemplo, un estudiante que vive a 6 km de la universidad, se espera que

10. Como se advirtió oportunamente, este cálculo puede hacerse con mayor comodidad en forma directa, utilizando las funciones estadísticas de la hoja de cálculo Excel.

tarde, en promedio, alrededor de 30 minutos en llegar. El valor se obtiene sustituyendo x por 6 en la ecuación:

$$y_{est} = 9,74 + 3,34 \cdot 6 = 29,78.$$

Note que 29,78 min es un valor promedio estimado de y, para $x = 6$ km; en la práctica, el tiempo puede ser menor o mayor, pues, como se aprecia en el gráfico, los puntos presentan cierto grado de dispersión alrededor de la tendencia lineal. Una estimación exacta del valor de y requiere que la relación sea perfecta.

La recta de regresión puede utilizarse para estimar valores esperados de y correspondientes a valores de x dentro del intervalo estudiado (distancias entre 1 y 18 km en este caso). Este tipo de estimación se denomina **interpolación**. La recta también puede usarse para estimar valores de y correspondientes a valores de x fuera del intervalo estudiado, este tipo de cálculo se denomina **extrapolación**.

Las extrapolaciones deben realizarse con cautela, pues implican el supuesto de que el comportamiento de y, dentro del intervalo estudiado, se mantiene para valores de x fuera de él. Esta suposición puede no ser correcta y conducir a errores graves. En el ejemplo que se está comentando, si un estudiante vive a 200 kilómetros de la universidad, requeriría aproximadamente 11 horas y 18 minutos en llegar, es decir, pasaría casi todo el día en el viaje de ida y en el regreso. Esto, por supuesto carece de sentido real, porque un estudiante que viva a esa distancia, y aún a una mucho menor, seguramente trasladará su domicilio a un punto cercano a la universidad si desea estudiar en ella.

El valor b representa la pendiente de la recta, se denomina **coeficiente de regresión.** Es de signo positivo cuando, al aumentar x, la y se incrementa, es decir, cuando la relación es directa, y de signo negativo cuando es inversa. En este ejemplo, $b = 3,34$ indica que, en promedio, por cada kilómetro que aumenta la distancia entre la universidad y la casa del estudiante, el tiempo para llegar crece, en promedio, en 3,34 minutos.

En cuanto al valor a, como ya se mencionó, representa la ordenada del punto donde la recta corta el eje de y, y corresponde al valor de y cuando $x = 0$. Se trata de un término constante que produce un mejor ajuste de la recta a los datos. En ciertas oportunidades tiene sentido concreto e importancia, pero usualmente se trata de un valor extrapolado, el cual debe interpretarse con sumo cuidado. En el presente caso, en sentido estricto, si una persona vive a distancia 0 de la universidad, es decir, vive ahí, $(x = 0)$, debe durar un tiempo 0 en llegar $(y = 0)$. Claro que esto no sucede; en cambio, se da el caso de estudiantes que viven frente a la universidad –cruzando la calle– para quienes x es prácticamente 0, también el tiempo para llegar es muy cercano a 0. Por lo anterior, sería incorrecto interpretar que $a = 9,74$ significa que quienes viven a una distancia 0 de la universidad tardan 9,7 min en llegar.

14.9. CONFIABILIDAD DEL MODELO DE REGRESIÓN: EL COEFICIENTE DE DETERMINACIÓN

Para diferentes propósitos, resulta importante conocer el grado de confiabilidad que tiene una línea de regresión ajustada para representar la relación entre x e y. Intuitivamente, se sabe que la confiabilidad será mayor cuanto más marcada sea la relación lineal y más apiñados alrededor de la línea de regresión estén los valores observados, y menor cuando suceda lo contrario. Una forma simple de medir la confiabilidad es utilizando el "coeficiente de determinación", el cual se define como $R^2 = r^2$, por consiguiente, se basa en el grado de asociación lineal entre las variables. Este coeficiente (R^2) indica la proporción de la variancia de la variable y, explicada por su relación lineal con la variable x, es 1 cuando la relación entre las dos variables es perfecta $(r = \pm 1)$, 0 cuando no exista ninguna asociación lineal entre las variables $(r = 0)$.

Este coeficiente equivale a la **razón de determinación** (R^2) presentada en el capítulo 9, "Medidas de variabilidad", aprovecha que la variancia de y se puede pensar como formada de dos partes:

a) La variancia explicada por x o debida a la relación lineal con x.

b) La variancia no explicada o debida a otros factores.

El coeficiente de determinación R^2, por lo tanto, puede definirse como:

$$R^2 = \text{coeficiente de determinanación} = \frac{\text{Variancia explicada de } y}{\text{Variancia total de } y}.$$

En el presente caso, como $r = 0,85$, $R^2 = (0,85)^2 = 0,72$, indica que un 72% de la variabilidad observada en el tiempo para llegar a la universidad, desde la vivienda donde se reside, se explica por las diferentes distancias a las que viven los alumnos. El restante 28% se debe a otros factores.

Es importante señalar que la equivalencia $R^2 = r^2$ solo se da en la regresión lineal simple; cuando es múltiple o no-lineal debe usarse otra expresión para el cálculo de R^2, aunque siempre es válido interpretarlo como la proporción de la variancia de la variable dependiente (y), el cual es explicado por las variables dependientes $(x_1, x_2, x_3, etc.)$ usando el modelo de regresión aplicado.

14.10. UN EJEMPLO ILUSTRATIVO

Para revisar y fijar mejor los conceptos y procedimientos del análisis de regresión, considere los siguientes datos que se refieren a y: número de errores cometidos al digitar un documento, y a t: tiempo, en minutos, consumido al digitarlo. Los datos corresponden

a una muestra al azar de 10 documentos, tomados de un conjunto numeroso de naturaleza similar e igual extensión, pero en los cuales han variado las personas que los digitaron y la exigencia respecto al tiempo en el que debían estar listos. En este caso, la variable dependiente es y, la independiente t.

Tiempo	Errores	Tiempo	Errores			
t	y	t	y	$\sum y = 35$		$\sum t = 80$
4	6	8	4		$\sum ty = 250$	
5	5	9	3			
7	5	10	2	$\sum y^2 = 145$		$\sum t^2 = 692$
7	3	10	1			
8	4	12	2			

a) Cálculo de la recta de regresión $y = a + bt$ que permite estimar el número de errores, conociendo el tiempo consumido en digitar el documento.

$$b = \frac{n\sum ty - \sum t \sum y}{n\sum t^2 - \left(\sum t\right)^2} = \frac{10 \cdot 250 - 80 \cdot 35}{10 \cdot 692 - (80)^2} = \frac{-300}{520} = -0,577 \approx -0,58.$$

$$a = \bar{y} - b\bar{t} = \frac{35}{10} - (-0,577)\frac{80}{10} = 3,5 - (-4,616) = 8,116 \approx 8,12.$$

$$y_{est} = 8,12 - 0,58t \quad (\text{N}^\text{o} \text{ de errores} = 8,12 - 0,58 \cdot \text{min}).$$

b) Existe una relación inversa entre el tiempo de digitación del documento y el número de errores cometido. El valor $b = -0,58$ señala que, en promedio, por cada minuto que aumenta el tiempo consumido en transcribirlo, el número de errores disminuye en 0,58.

Respecto a la constante $a = 8,12$, geométricamente indica que la recta ajustada corta el eje de y un poco arriba de 8, o sea, cuando $t = 0$, el valor de y es 8 (8 errores).

¿Podría interpretarse el valor $a = 8,12$ en el sentido de que, cuando se tarda 0 minutos en digitar un documento, se cometen 8 errores? Por supuesto que no. Como ya se dijo, se trata de un valor extrapolado carente de sentido real. Igualmente incorrecto sería afirmar que como $y_{est} = 8,12 - 0,58 \cdot 14 = 0$, si se dura 14 minutos (o más) en digitarlo no se cometería ningún error.

c) Los puntos comentados antes son reafirmados por el diagrama de dispersión incluido seguidamente, en el cual aparece representada también la recta de regresión.

Para hacer esto último, se le dieron a t los valores 0 y 10, pues bastan dos puntos para representarla:

$$y = 8,12 - 0,58 \cdot 0 = 8,12.$$
$$y = 8,12 - 0,58 \cdot 10 = 2,32.$$

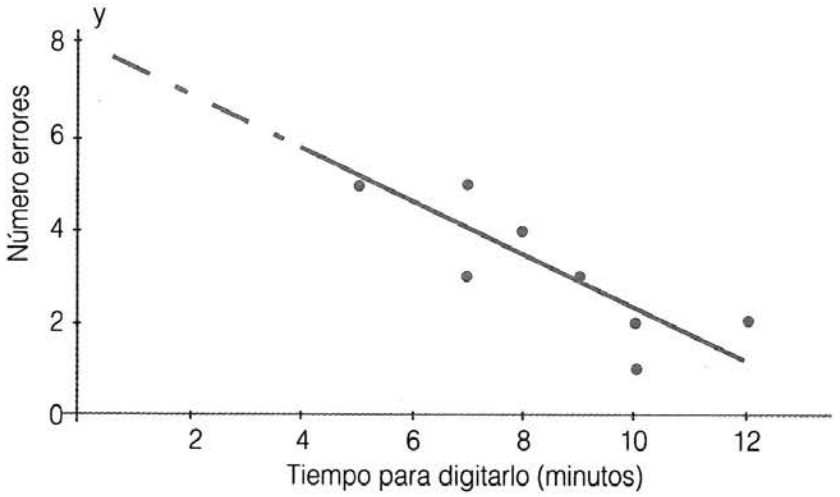


Figura 14.8. Gráfico que muestra la recta de regresión

La cercanía de los puntos de la recta indica que el ajuste es bueno y hay una marcada relación lineal inversa entre las variables. Esto lo confirman el valor del coeficiente de correlación $r = -0,88$ y el de determinación $r^2 = 0,77$. Un 77% de la variancia en el número de errores se explica por la relación lineal con el tiempo que se dura digitando el documento.

$$r = \frac{n\sum ty - \sum t \cdot \sum y}{\sqrt{[n\sum t^2 - (\sum t)^2][n\sum y^2 - (\sum y)^2]}} = \frac{10 \cdot 250 - 80 \cdot 35}{\sqrt{[10 \cdot 692 - (80)^2][10 \cdot 145 - (35)^2]}}$$

$$= \frac{-300}{\sqrt{520 \cdot 225}} = \frac{-300}{342,05} = -0,08$$

$$r^2 = (-0,88)^2 = 0,77.$$

14.11. REGRESIÓN MÚLTIPLE

En las secciones anteriores de este capítulo, se han considerado problemas de regresión en los cuales están involucradas solo dos variables. En la práctica, sin embargo, en muchas situaciones es sabido que la variable que interesa y depende no de una, sino de diversas variables independientes. Esto lleva a la denominada regresión múltiple. Considere un ejemplo sencillo, pero real, en el cual aparecen tres variables: una dependiente y dos independientes.

En 1970, tres investigadoras del Instituto de Investigaciones Psicológicas de la Universidad de Costa Rica (IIP-UCR) se interesaron en examinar la medida cómo el Examen de Admisión de esa universidad permitía predecir o pronosticar la nota del Examen Comprensivo realizado al final del primer año común a los alumnos aceptados (Estudios Generales).[11]

El estudio se basó en un total de 1816 alumnos que tomaron el examen de admisión en 1969 y entraron a la universidad en 1970. Las variables principales examinadas fueron:

y : examen comprensivo

x_1 : puntaje de la Prueba de Aptitud

x_2 : nota de bachillerato

(Variables medidas en escala 0-100).

Para hacer su análisis, plantearon el modelo de regresión lineal múltiple siguiente:

$$y = b_0 + b_1x_1 + b_2x_2.$$

Este modelo postula que la nota en el examen comprensivo (y) depende del puntaje en la prueba de aptitud (x_1) y de la nota de bachillerato (x_2). El coeficiente b_0 es una constante, similar a la a en el modelo de regresión lineal simple, b_1 y b_2 representan coeficientes de regresión de naturaleza e interpretación similar a la b de la regresión lineal simple.

El modelo fue ajustado por las investigadoras a los datos, usando el criterio de cuadrados mínimos, y obtuvieron la siguiente ecuación:

$$y_{est} = 30,589 + 0,316x_1 + 0,335x_2.$$

El coeficiente de determinación del ajuste fue $R^2 = 0,22$ y el de correlación múltiple $R = \sqrt{0,22} = 0,47$.[12]

¿Qué indica este análisis de regresión? En primer término, que un 22% de la variabilidad mostrada en la nota del examen comprensivo, realizado al final del año en Estudios Generales, puede explicarse –estadísticamente– por la prueba de aptitud y la nota de bachillerato. El 78% restante de la variabilidad se debe a otras variables no incluidas en el modelo de regresión, como podría ser el tiempo dedicado al estudio, nivel socioeconómico de la familia, grado de asimilación del estudiante al ambiente universitario, etcétera.

11. Wong, M., León-Páez, O. y León, E. (1971). Validez Predictiva del Examen de Admisión, Instituto de Investigaciones Psicológicas. San Jose: EUCR.

12. El coeficiente de correlación múltiple R es el coeficiente de correlación lineal entre los valores de y observados y los estimados con la ecuación de regresión. Expresa el grado de relación lineal de y con todas las variables independientes, en este caso x_1 y x_2.

En principio, este $R^2 = 0;22$ podría parecer muy alejado de 1 y, por lo tanto, bastante bajo; sin embargo, debe considerarse que el rendimiento académico depende de numerosos factores, por ello es difícil obtener porcentajes elevados de variancia explicada. Es más, los valores $R^2 = 0,22$ y el valor $R = \sqrt{0,22} = 0,47$ indican que las investigadoras alcanzaron, con esa ecuación –relativamente sencilla–, un nivel de explicación bastante elevado, si se la compara con las correlaciones obtenidas usualmente en los estudios de rendimiento académico.

Es importante señalar, también, que el modelo ajustado permite estimar la nota esperada de un estudiante en el examen comprensivo si se conoce su puntaje de aptitud y su nota de bachillerato. Así, por ejemplo, si un estudiante tiene un puntaje de aptitud $x_1 = 60$ y una nota de bachillerato $x_2 = 70$, sustituyendo en la ecuación se logra la nota esperada en el examen comprensivo, que resulta ser 73, como puede notarse seguidamente:

$$y = 30,589 + 0,316 \cdot 60 + 0,335 \cdot 70 = 30,589 + 18,96 + 23,45 = 73,0.$$

En cuanto al valor $b_0 = 30,589$, al igual que el a de la línea de regresión lineal simple, es una constante importante para el ajuste del modelo, pero que no es factible interpretar sustantivamente en esta situación concreta.

El valor $b_2 = 0,335$ indica que, por cada punto porcentual o centésima de aumento en la nota de bachillerato, se espera un incremento de 0,335 centésimas en la nota del examen comprensivo, suponiendo que el puntaje de la prueba de aptitud (x_2) no varía. Así, si se tienen dos estudiantes con un mismo puntaje de aptitud, como 70, y uno de ellos saca un 80 en bachillerato y el otro un 90, los valores esperados en el examen comprensivo serían:

$$y = 30,589 + 0,316 \cdot 70 + 0,335 \cdot 80 = 79,51.$$

$$y = 30,589 + 0,316 \cdot 70 + 0,335 \cdot 90 = 82,86.$$

La diferencia entre las notas es $82,86 - 79,51 = 3,35$, o sea, $10 \cdot b_2 = 10 \cdot 0,335 = 3,35$, debe atribuirse totalmente a los 10 puntos más que obtuvo el segundo estudiante en bachillerato. En forma similar, se interpreta el valor $b_1 = 0,316$.

Es importante examinar qué habría pasado si, en lugar de una regresión múltiple, se hubiera usado únicamente una regresión simple con el puntaje de aptitud (x_1) o con la nota de bachillerato (x_2). Para ello, observe los ajustes de ese modelo simple con cada una de las variables:

$$\text{Aptitud } (x_1)\, y = 55,02 + 0,296 \cdot x_1 \cdot r^2 = 0,13.$$

$$\text{Bachillerato } (x_2)\, y = 41,91 + 0,393 \cdot x_2 \cdot r^2 = 0,10.$$

Es evidente lo provechoso que resultó usar simultáneamente las dos variables: el modelo múltiple permite explicar un 22% de la variabilidad en la nota del examen comprensivo, mientras que, si hubiera usado solo una de las variables, la variancia explicada habría sido solo de un 10 o un 13%.

Se tiene, por otra parte, que los coeficientes de regresión de las ecuaciones simples no coinciden con los de la ecuación de regresión múltiple. Esto se debe a que, en esta última, el ajuste considera simultáneamente la información de x_1 y x_2.

14.12. INFERENCIA ACERCA DEL COEFICIENTE DE CORRELACIÓN POBLACIONAL ρ

Si el cálculo del coeficiente de correlación se realiza incluyendo a todos los N elementos que forman la población, se obtiene el valor poblacional ρ_{xy}, el cual, usualmente, no se conoce y se estima con el valor muestral r_{xy}. Ahora bien, el coeficiente de correlación r, al igual que otras medidas estadísticas calculadas a partir de muestras como: $\bar{x}$, s^2, p, etc., está sujeto a fluctuaciones de muestreo y su confiabilidad como estimador de ρ está ligada al tamaño de la muestra en la cual se basa. El valor dado por una muestra determinada es uno de todos los posibles valores de r que se obtendrían al tomar muestras aleatorias del mismo tamaño de la población de interés; r varía en cada una, por ello debe examinarse la medida en que el valor dado por una cierta muestra es un estimador confiable del valor poblacional ρ. Para este propósito, semejante al caso de otros estimadores, se hace necesario recurrir a una prueba de hipótesis o a un intervalo de confianza para evaluar la confiabilidad que puede asignársele.

Las inferencias acerca de un valor poblacional requieren conocer la distribución de muestreo del estimador y la forma como se comporta al aumentar el tamaño de la muestra. Para el valor muestral r, la teoría estadística permite derivar su distribución bajo ciertos supuestos. Así, cuando la muestra haya sido extraída al azar de una población normal bivariante, es decir, si x e y tienen cada una distribución normal, se tiene que la distribución de r depende de n –tamaño de la muestra– y del valor poblacional ρ. Ahora bien, cuando $\rho = 0$, la distribución de r es simétrica alrededor de 0 (ver gráfico de la figura 14.8), su error estándar lo da la expresión $\sqrt{\frac{1-r^2}{n-2}}$. En esta situación especial, la significancia de r frente a la hipótesis $\rho = 0$ puede probarse utilizando el estadístico:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

el cual tiene la distribución t de *Student* con $n - 2$ grados de libertad.

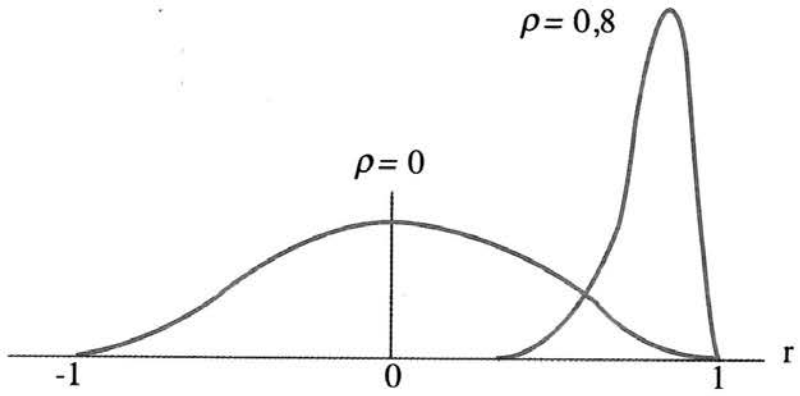


Figura 14.9. Gráfico que muestra la distribución muestral de r cuando el coeficiente de correlación lineal poblacional (ϱ) es 0 y 0,8

Ejemplo 2

En una muestra al azar de 45 jefes de familia, se encontró un valor $r = -0,32$ entre las variables años de estudio y edad. Someta a prueba la hipótesis $H_0: \rho = 0$. Use $\alpha = 0,05$.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,32 \cdot \sqrt{43}}{\sqrt{1-(-0,32)^2}} = \frac{-0,32 \cdot 6,56}{\sqrt{0,8976}} = \frac{-2,0992}{0,9474} = -2,215.$$

La t tabular para 40 grados de libertad, $\alpha = 0,05$, dos colas, es: 2,021. Como $|-2,215| > 2,021$, se rechaza la H_0 y se concluye que existe una correlación inversa entre educación y edad.[13]

En el gráfico 14.9 puede observarse que, cuando ρ se aleja de 0, la distribución del estimador muestral r se vuelve marcadamente asimétrica y el procedimiento antes citado no es apropiado para las pruebas de significancia. Una solución propuesta por Fisher consiste en emplear la transformación:

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

Esta variable z sigue una distribución aproximadamente normal, su media y desviación estándar vienen dadas por las siguientes expresiones:

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \text{ y } \sigma_z = \frac{1}{\sqrt{n-3}}.$$

Usando esta transformación, la hipótesis $H_0: \rho = 0$ se somete a prueba, partiendo del valor r muestral, calculando $z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ y obteniendo seguidamente:

$$z_c = \frac{z - \mu_z}{\sigma_z}.$$

13. Si se suministra a la función DIST T de Excel, el valor $t = 2,215$, el número de grados de libertad, (43), y se le dice que la prueba es de dos colas, esta indicará una probabilidad de 0,032, la cual al ser menor que el valor crítico $\alpha = 0,05$, indica que la hipótesis H_0 debe rechazarse.

Este z_c se compara con el valor de la normal estándar correspondiente al nivel de significancia que se haya escogido, y la hipótesis se rechaza si el valor calculado supera el valor tabular respectivo.

El cálculo de los valores de $\ln\left(\frac{1+r}{1-r}\right)$ y de los de $\ln\left(\frac{1+\rho}{1-\rho}\right)$ puede hacerse fácilmente con la función ln de una calculadora o con una hoja de cálculo. También es posible usar las funciones estadísticas de Excel, las cuales permiten obtener las expresiones anteriores, mediante la función Fisher, y la probabilidad correspondiente a la z_c.

Ejemplo 3

Considere el valor $r = 0,70$ que se obtuvo en la sección 2 de este capítulo y que corresponde a las variables peso y estatura de una muestra de 20 estudiantes. Pruébese la hipótesis $H_0: \rho = 0$ contra la alternativa $H_1: \rho \neq 0$ (use $\alpha = 0,05$).

$$t = \frac{0,70 \cdot \sqrt{18}}{\sqrt{1-(0,69)^2}} = \frac{0,70 \cdot 4,24}{0,724} = 4,10.$$

Como $t_{18(0,025)} = 2,101$ y $|4,16|2,10$, se rechaza H_0.

Se concluye que ρ es mayor que 0, por lo tanto, existe correlación lineal positiva entre peso y estatura.

Ejemplo 4

Estudios de años anteriores señalan que la correlación entre peso y estatura de alumnos universitarios ha sido cercana al 0,60. Someta a prueba la hipótesis $H_0: \rho = 0,60$ contra la alternativa $H_1: \rho \neq 0,60$, usando la $r = 0,70$ entre peso y estatura.

En este caso, como ρ está bastante alejado de cero, es mejor utilizar la transformación z.

$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,70}{1-0,70}\right) = \frac{1}{2}\ln(5,667) = 0,8673.$$

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2}\ln\left(\frac{1+0,60}{1-0,60}\right) = \frac{1}{2}\ln(4) = 0,6931.$$

$$\sigma_z = \frac{1}{\sqrt{n-3}} = \frac{1}{\sqrt{17}} = 0,2425.$$

$$z_c = \frac{z-\mu_z}{\sigma_z} = \frac{0,8673-0,6931}{0,2425} = \frac{0,1742}{0,2425} = 0,7183.$$

Como $0,7183 < 1,96$, se mantiene la $H_0: \rho = 0,60$. Se concluye que el valor muestral observado no permite afirmar que la correlación entre sexo y estatura ha aumentado.

EJERCICIOS DE AUTOEVALUACIÓN

1. En una encuesta telefónica a nivel nacional, realizada en los primeros días de diciembre del 2009, se preguntó a una muestra de 290 trabajadores si consideraban que el próximo presidente de Costa Rica debería gobernar siguiendo las políticas y la forma como lo estaba haciendo Óscar Arias o debería hacerlo diferente. Las respuestas se presentan seguidamente.

Próximo presidente debería gobernar

	Sector público	Sector privado	Total
Siguiendo políticas y forma de Óscar Arias	24	81	105
En forma diferente a Óscar Arias	74	111	185
Total	98	192	290

En su opinión, ¿existe asociación o no entre la opinión y el sector de actividad al que pertenece el entrevistado?

2. Al estudiar una muestra al azar de 124 hombres adultos, un economista encuentra un $r = 0,24$ entre las variables x: años de educación aprobados y y: ingreso mensual en colones. Al ver el resultado, el economista dice: "No hay duda que existe correlación entre educación e ingreso mensual; evidentemente, al tener más educación, los hombres ganan más, la educación es la causa del mayor ingreso".

   Lea cuidadosamente las afirmaciones del economista y luego responda a las preguntas siguientes:

   a) ¿Puede afirmar el economista, con bastante seguridad, que hay correlación entre las variables? Justifique su respuesta. Realice los cálculos e incluya los comentarios que crea necesarios.

   b) ¿Será correcto afirmar, a partir del valor de r y del análisis estadístico únicamente, que la educación es la causa del mayor ingreso?

3. Un investigador participante en un programa de educación y pobreza estudió una muestra aleatoria de familias de una región y al calcular el coeficiente de correlación entre nivel de ingreso familiar y fecundidad (número de hijos tenidos) encontró una $r = -0,60$.

Si se utiliza únicamente la correlación anterior, ¿cuál de las afirmaciones siguientes le parece a usted correcta y por qué?

a) La pobreza es la causa de la alta fecundidad (mayor número de hijos).

b) La alta fecundidad (tener muchos hijos) es la causa de la pobreza.

c) Ninguna de las anteriores. ¿Por qué? (Explique).

4. En un estudio que abarcó cerca de 15 años se correlacionó el Producto Interno Bruto (PIB) anual en dólares de un país con el monto de basura producida en ese mismo periodo por habitante (en kilos). El valor obtenido para el coeficiente fue $r = 0,81$. Escriba una interpretación del valor obtenido.

5. Usando los datos que aparecen en el cuadro 14.1 de este capítulo, correspondientes a la distancia a la que se vive de la universidad (km) y monto en colones que se gasta semanalmente en transporte, realice los ejercicios indicados a continuación:

a) Construya el diagrama de dispersión y calcule el coeficiente de correlación. (Sugerencia: coloque la distancia en el eje x).

b) Ajuste una línea de regresión que permita estimar el gasto de transporte conociendo la distancia entre la residencia y la universidad. Represente la recta en el diagrama.

c) Interprete los valores a y b obtenidos en el ajuste.

d) Estime el gasto para un estudiante que vive a 10 km de la universidad.

e) Calcule el coeficiente de determinación. Comente la confiabilidad que tiene el modelo para estimar el gasto.

6. En un examen final de Estadística, se anotó el tiempo transcurrido entre el momento en que se inició la prueba y cuando el alumno entregó el examen. Esta información, junto con la nota obtenida, aparece a continuación para los 29 alumnos que tomaron el examen.

a) Calcule el coeficiente de correlación lineal a partir de los valores siguientes:

$$s_{xy} = -88,80;\ s_x = 24,43 \text{ y } s_y = 20,73.$$

b) Escriba una interpretación del valor obtenido para r y una conclusión para las variables consideradas.

c) Construya el diagrama de dispersión. ¿Mantendría su interpretación en b, después de exáminar el diagrama de dispersión?, ¿por qué?

Orden de entrega	Tiempo en minutos (x)	Nota (y)	Orden de entrega	Tiempo en minutos (x)	Nota (y)
1	104	33	16	144	84
2	105	13	17	148	72
3	111	42	18	150	46
4	112	69	19	157	84
5	116	48	20	158	53
6	119	70	21	164	56
7	120	48	22	166	59
8	121	54	23	167	28
9	122	49	24	169	39
10	123	70	25	170	58
11	125	76	26	175	32
12	127	87	27	175	20
13	130	61	28	176	46
14	131	54	29	177	29
15	136	97			

Nota: si trata de comprobar los valores de la covariancia, usando las funciones estadísticas del Excel, deberá tomar en cuenta que la definición de la covariancia en este programa utiliza como divisor n y no $n-1$. Por ello, el valor dado en Excel deberá multiplicarse por $\frac{1}{n-1}$ para que coincida con el que aparece al comienzo de este ejercicio.

7. En un estudio se pidió a una muestra al azar de 150 adultos calificar la labor del presidente de un país con una nota, usando un termómetro que varía de 0 a 100, además, se solicitó evaluar la situación económica del país en una escala de 7 puntos, que incluía las siguientes categorías:

   1. Muy mala 2. Mala 3. Algo mala 4. Ni mala ni buena
   5. Algo buena 6. Buena 7. Muy buena

   La correlación entre esas dos variables, "Percepción de la situación económica" y "Evaluación del trabajo del presidente", arrojó un valor de $r = 0,43$.

   a) ¿Puede afirmarse que existe una correlación lineal positiva entre las variables "Percepción de la situación económica" y "Evaluación del trabajo del presidente"? Use $\alpha = 0,05$.

   b) ¿Qué proporción de la variabilidad, en las evaluaciones de la labor del presidente, puede ser explicada por las diferencias en la percepción de la situación económica?

c) ¿Qué otros factores, además de la percepción de la situación económica, pueden estar influyendo en la calificación de la labor del presidente?

d) ¿Cuál es el nivel de medición de las variables?

e) En su opinión ¿es correcto calcular la correlación lineal r entre esas dos variables o no?

RESPUESTA A LOS EJERCICIOS DE AUTOEVALUACIÓN

1. Se procede, en primer término, calcular la distribución porcentual de las columnas (% verticales) para hacer las comparaciones pertinentes y determinar si hay relación entre el sector de actividad del entrevistado y su opinión acerca de cómo debe gobernar el próximo presidente.

	Próximo presidente debería gobernar	
	Sector público	**Sector privado**
Siguiendo políticas y forma de Óscar Arias	24,5%	42,2%
En forma diferente a Óscar Arias	75,5%	57,8%
Total	100	100

Puede apreciarse que las distribuciones porcentuales difieren, pero no marcadamente. Las personas que laboran en el sector público tienden a considerar, mayoritariamente, que el próximo presidente debe gobernar en forma diferente a como lo ha hecho Óscar Arias, mientras que quienes se desenvuelven en el sector privado piensan que igual, pero en forma menos marcada. En correspondencia con lo anterior, la diferencia porcentual es moderada: $75,5 - 57,8 = 17,7$.

2. a) Como el coeficiente de correlación $r = 0,24$ se basa en una muestra $(n = 124)$, existe la posibilidad de que el valor obtenido sea fruto del azar y no refleje la situación real de la población. Por ello, para responder a la pregunta, debe realizarse la prueba de hipótesis $H_0: \rho = 0$. Esto se hace inmediatamente a un nivel de significancia $\alpha = 0,05$.

La teoría estadística indica que la expresión $t = r\sqrt{\frac{N-2}{1-r^2}}$ tiene distribución t de *Student* con $n - 2$ grados de libertad. Por eso, se calcula el valor t y se compara con la t tabular, para 122 grados de libertad, la cual es 1,98.

$$t = r\sqrt{\frac{N-2}{1-r^2}} = 0,24 \cdot \sqrt{\frac{122}{1-(0,24)^2}} = 0,24 \cdot \sqrt{\frac{122}{0,9424}} = 2,73.$$

Como $2,73 > 1,98$, se rechaza la hipótesis nula y se acepta que existe una correlación positiva entre los años de educación aprobados y el ingreso mensual.

b) La correlación estadística indica que existe una asociación positiva, matemática, entre los valores de educación e ingreso, pero esto no permite concluir, necesariamente, que haya una relación de causalidad entre las variables. La correlación bien puede deberse a una tercera variable, la cual influye tanto en

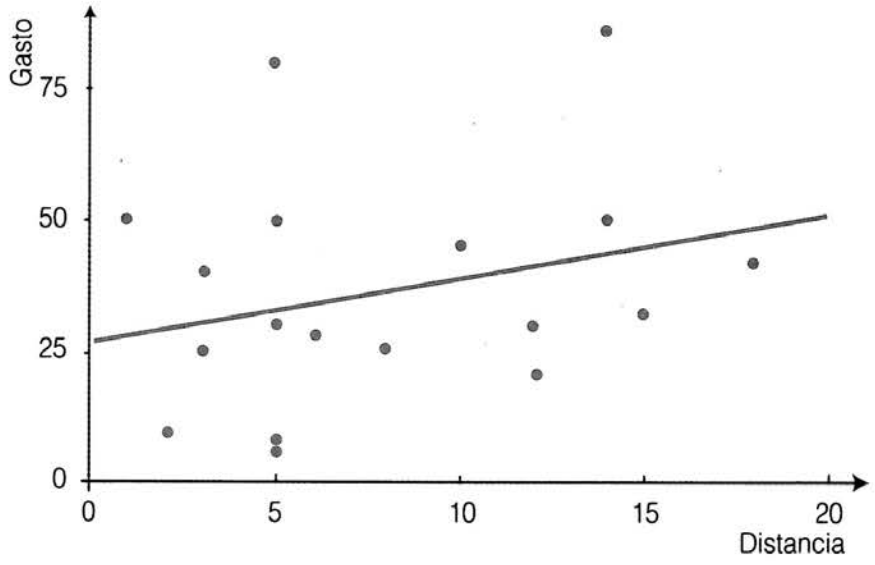
la educación como en el ingreso. La relación de causalidad debe deducirse de otro tipo de elementos de juicio.

3. c) Ninguna de las anteriores, porque la existencia de correlación no implica, necesariamente, relación de causalidad entre las dos variables.

   Solo puede afirmarse que existe una correlación negativa marcada entre número de hijos y nivel de ingreso, las familias con menos hijos tienden a poseer un mayor ingreso.

4. Existe una marcada correlación histórica positiva entre el aumento del PIB y la cantidad de basura producida por habitante. Conforme se incrementa el PIB, también lo hace el volumen per cápita de basura. Como $r^2 = 0,6561$, un 66% de la variabilidad observada en la cantidad anual de basura producida por habitante, estadísticamente, se explica por la relación lineal con el PIB.

5. a) Diagrama de dispersión



   b) $b = 1,13, a = 27,04, y = 27,04 + 1,13x, r = 0,26$.

   c) El valor $b = 1,13$ es el coeficiente de regresión e indica que por cada aumento en un kilómetro en la distancia a la que se vive de la universidad, el gasto aumenta en 1,13 colones.

   $a = 27,04$ es la ordenada del punto donde la recta de regresión corta al eje y, o sea es el valor de y para $x = 0$. Es un valor que ayuda a ajustar mejor la recta a los datos, pero en el presente caso carece de un interpretación concreta razonable. Decir que indica el gasto en transporte en el cual incurriría una persona que vive a cero kilómetros de la universidad (o "enfrente") sería una interpretación errónea.

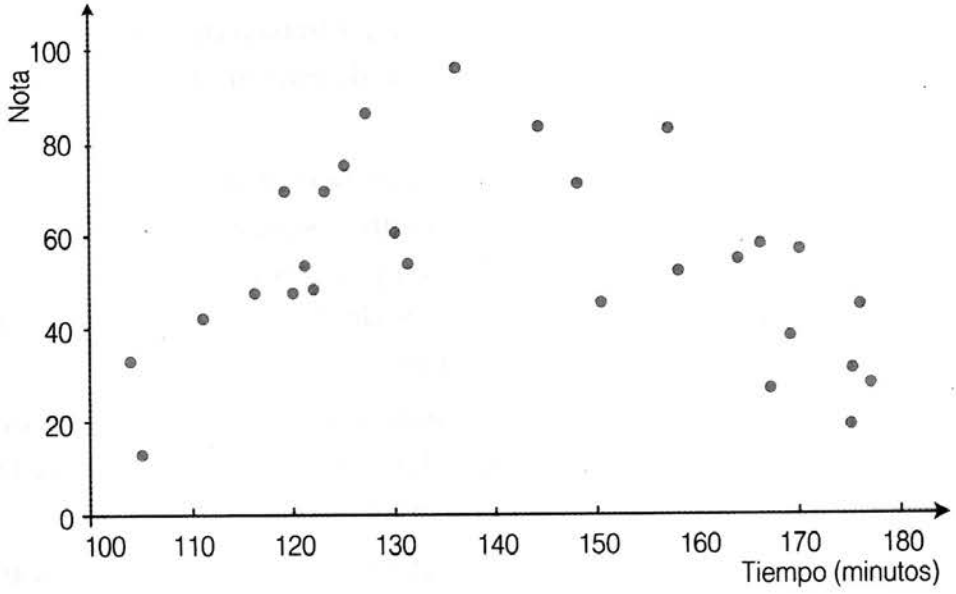
d) Si un estudiante vive a 10 km de la universidad, su gasto se estima en $y = 27,04 + 1,13 \cdot 10 = 27,0 + 11,30 = 38,34$ colones.

e) $r^2 = (0,26)^2 = 0,0676$. Esto implica que un 6,8% de la variabilidad de y (gasto semanal) es explicado, estadísticamente, por la relación lineal con la distancia a la que se vive de la universidad. La correlación entre distancia y gasto es muy baja, el modelo tiene poca confiabilidad para estimar el gasto semanal.

6. a) $S_{xy} = -88,80; S_x = 24,43; S_y = 20,73; r = \frac{S_{xy}}{S_x S_y} = -0,175$.

b) El valor $r = -0,175$ indica que hay una correlación lineal negativa baja entre las variables. Prácticamente, no hay relación lineal entre tiempo dedicado a resolver el examen y la nota obtenida, el r^2 es solo 3,06%.

c) Diagrama de dispersión



El diagrama de dispersión muestra que la asociación entre la nota y el tiempo que se dura haciendo el examen no es lineal sino curvilínea: los estudiantes que lo entregan pronto (duran poco) y los que lo entregan de últimos (duran más) tienden a obtener notas bajas; entretanto, quienes tienen una duración media logran las notas más altas. La conclusión debe ser que sí existe una relación importante entre la nota y el tiempo, pero que esta es curvilínea. El modelo lineal no es apropiado en este caso.

7. $n = 150$ y $r = 0,43$.

a) Para responder a esta pregunta, debe realizarse la prueba de la hipótesis $H_0: \rho = 0$ contra la alternativa $H_1: \rho > 0$. Se usa $\alpha = 0,05$. En primer lugar, se calcula el valor de t correspondiente:

$$t = r\sqrt{\frac{N-2}{1-r^2}} = 0,43 \cdot \sqrt{\frac{148}{1-(0,43)^2}} = 0,43 \cdot \sqrt{\frac{148}{0,8151}} = 5,79.$$

Dado que la muestra es grande y el número de grados de libertad elevado (148), el valor de t tabular es prácticamente igual al valor de z normal para el nivel de significancia escogido: 1,645. Como 5,79 es muy superior a 1,645, se rechaza la hipótesis nula y se concluye que sí existe correlación entre la evaluación del presidente y la percepción de la situación económica.

b) $r^2 = (0,43)^2 = 0,1849 \approx 18\,\%$.

18% de la variabilidad, en la evaluación del trabajo del presidente, es explicada estadísticamente por la relación lineal de esa variable con la percepción de la situación económica del país.

c) Los otros factores que se sabe influyen en la evaluación del trabajo del presidente son: la simpatía política de los entrevistados, quienes pertenecen a su mismo partido tienden a calificarlo bien y los contrarios mal; su labor en otras áreas de acción del gobierno y aspectos de su personalidad como humildad, simpatía, forma de tratar a las personas.

d) La calificación de la labor del presidente se realiza en una escala de 0 a 100 y puede considerarse de nivel métrico. En cuanto a la evaluación de la situación económica del país, la variable es de nivel ordinal.

e) En rigor, el cálculo de la correlación r debe hacerse solo si ambas variables son métricas. En la práctica, sin embargo, especialmente en las ciencias sociales, es corriente que se calculen correlaciones entre variables ordinales y métricas o entre variables ordinales, cuando estas corresponden a escalas como la de evaluación de la situación económica usada en este ejemplo.

REFERENCIAS BIBLIOGRÁFICAS

Croxton, F. y Cowden, D. (1959). *Estadística General Aplicada* (trad. por Teodoro Ortiz y Manuel Bravo). México: Fondo de Cultura Económica.

Dirección General de Estadística y Censos. Anuarios Estadísticos. Censos de Población, Vivienda y Agricultura. San José, Costa Rica.

Haber, A. y Runyon, R. (1973). *Estadística General* (trad. por Ricardo Lassala Mozo). México: Fondo Educativo Interamericano.

Hernández, O. (2009). *Estadística Elemental para Ciencias Sociales*. San José: EUCR.

Kennedy, J. y Neville, A. (1982). *Problemas de Estadística General*. San José: EUCR.

Lininger, C. y Warwick, D. (1978). *La Encuesta por Muestreo: Teoría y Práctica* México: CEQSA.

Lipschutz, S. (1982). *Probabilidad*: Serie Schaum. México: McGraw-Hill.

Ministerio de Planificación Nacional y Política Económica (1982). *Manual para la Presentación de Cuadros y Gráficos Estadísticos*. San José: Fotolitografía. G. N.

Noelle, E. (1970). *Encuestas en la Sociedad de Masas* (trad. por Bloy Fuente Herrero). Madrid: Alianza.

Padua, J. (1978). *Técnicas de Investigación en Ciencias Sociales*. México: Trillas.

Quintana, C. y Segnini, C. (1982). *Problemas de Estadística General*. San José: EUCR.

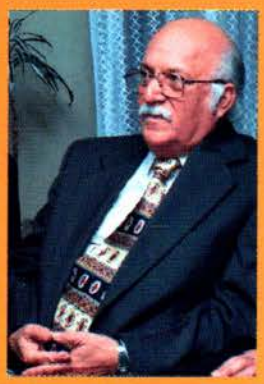
Shao, S. (1980). *Estadística para economistas y Administradores de Empresas* (trad. por Romeo Madrigal). México: Herrero Hnos.

Yamane, T. (1974). *Estadística* (trad. por Nuria Cortado de Kohan). México: Harla.

Zuwaylif, F. (1977). *Estadística General Aplicada* (trad. por Alberto Saenger). México: Fondo Educativo Interamericano.

ELEMENTOS DE **ESTADÍSTICA DESCRIPTIVA**

Miguel Gómez Barrantes



Realizó sus estudios en la Universidad de Costa Rica y en la Universidad de Michigan, Ann Arbor, en las que obtuvo los títulos de licenciado y máster en sociología.

Asistió a cursos de posgrado y especialización sobre métodos estadísticos, demografía, sociología, muestreo y diseño de encuestas en diversos centros universitarios de América.

Catedrático en la Universidad de Costa Rica desde 1973. En la facultad de Ciencias Económicas, se desempeñó como profesor de Estadística General y de Diseño de Encuestas.

Director del Departamento de Investigaciones del Centro de Estudios Sociales y de Población de la Universidad de Costa Rica (1971-1973) y jefe de la Unidad de Investigaciones Socioeconómicas de la Oficina de Información de la Casa Presidencial (1974-1981).

Publicó numerosos ensayos, estudios, diseños de encuestas públicas, monografías y varios libros de texto en el campo de la estadística y la investigación social.

EUNED
EDITORIAL UNIVERSIDAD ESTATAL A DISTANCIA

ISBN: 978-9968-31-31-934-8
9 789968 319348