2. ANÁLISIS DE RELACIONES

JOHN W. TUKEY

John W. Tukey (1915-) empezó como químico, siguió como matemático y finalmente se convirtió en estadístico debido a lo que él mismo denominó "la experiencia de los problemas reales y de los datos reales" que adquirió durante la II Guerra Mundial. En 1937, John W. Tukey fue a la Princeton University a estudiar química pero se doctoró en matemáticas en 1939. Durante la guerra trabajó en temas de precisión de tiro. Después de la guerra, simultaneó su labor en la Princeton University con su trabajo en los Laboratorios Bell, quizá el grupo de investigación industrial más importante del mundo.

Tukey dedicó la mayor parte de su atención al estudio estadístico de problemas especialmente difíciles de resolver, como son la seguridad de las anestésicos, el comportamiento sexual de los seres humanos, la comprobación del cumplimiento de la prohibición de las pruebas nucleares, y la determinación de la calidad del aire y la contaminación ambiental.

Basándose en "la experiencia de los problemas reales y de los datos reales", John Tukey desarrolló el análisis exploratorio de datos. Inventó algunas de las herramientas estadísticas que hemos visto en el capítulo 1 como, por ejemplo, los diagramas de tallos y los diagramas de caja. Tukey cambió el enfoque del análisis de datos, defendiendo un análisis de datos mucho más flexible, más exploratorio, cuyo objetivo no consiste simplemente en dar respuesta a preguntas concretas. El primer propósito es contestar a la pregunta: "¿Qué dicen los datos?". Este capítulo, igual que el capítulo 1, sigue el camino que marcó Tukey, y para ello presentamos más ideas y herramientas para examinar datos.

2.1 Introducción

Un estudio médico halló que las mujeres bajas son más propensas a sufrir ataques al corazón que las mujeres de altura media. Además, las mujeres altas sufren menos ataques al corazón que las de altura media. Por otro lado, un grupo asegurador informa que con coches grandes se producen menos muertes por cada 10.000 vehículos que con coches pequeños. Estos y muchos otros estudios estadísticos buscan relaciones entre dos variables. Para entender este tipo de relaciones, a menudo también tenemos que examinar otras variables. Para llegar a la conclusión de que las mujeres bajas sufren más ataques al corazón, los investigadores tuvieron que eliminar el efecto de otras variables como el peso y los hábitos deportivos. En este capítulo examinaremos la relación entre variables. También veremos que la relación entre dos variables puede verse afectada de forma importante por variables latentes de entorno.

Como la variación está siempre presente, las relaciones estadísticas son tendencias generales, no reglas blindadas. Nuestras relaciones admiten excepciones individuales. A pesar de que como media los fumadores mueren antes que los no fumadores, algunos fumadores que consumen más de tres paquetes diarios llegan a los noventa años. Para estudiar la relación entre dos variables, las medimos en los mismos individuos. A menudo, creemos que una de las variables puede explicar o influye sobre la otra.

VARIABLE RESPUESTA Y VARIABLE EXPLICATIVA

Una **variable respuesta** mide el resultado de un estudio. Una **variable explicativa** influye o explica cambios en la variable respuesta.

Variable independiente

Variable dependiente

A menudo, encontrarás que a las variables explicativas se les llama **variables independientes** y a las variables respuesta, **variables dependientes**. La idea es que el valor de la variable respuesta depende del de la variable explicativa. Como en estadística las palabras "independiente" y "dependiente" tienen otros significados que no están relacionados con lo que acabamos de ver, no utilizaremos esta terminología.

La manera más fácil de distinguir entre variables explicativas y variables respuesta es dar valores a una de ellas y ver lo que ocurre en la otra.

EJEMPLO 2.1. Los efectos del alcohol

El alcohol produce muchos efectos sobre el cuerpo humano. Uno de ellos es la bajada de la temperatura corporal. Para estudiar este efecto, unos investigadores suministraron distintas dosis de alcohol a unos ratones y al cabo de 15 minutos midieron la variación de temperatura de su cuerpo. La cantidad de alcohol es la variable explicativa y el cambio de temperatura corporal es la variable respuesta. ■

Cuando no asignamos valores a ninguna variable, sino que simplemente observamos los valores que adquieren, éstas pueden ser o no variables explicativas y variables respuesta. El que lo sean depende de cómo pensemos utilizar los datos.

EJEMPLO 2.2. Calificaciones en la prueba SAT

Alberto quiere saber qué relación existe entre la media de las calificaciones de Matemáticas y la media de las calificaciones de Lengua obtenidas por estudiantes de los 51 Estados de EE UU (incluyendo el Distrito de Columbia) en la prueba SAT. Inicialmente, Alberto no cree que una variable dependa de los valores que tome la otra. Tiene dos variables relacionadas y ninguna de ellas es una variable explicativa.

Julia, con los mismos datos, se plantea la siguiente pregunta: ¿puedo predecir la calificación de Matemáticas de un Estado si conozco su calificación de Lengua? En este caso, Julia trata la calificación de Lengua como variable explicativa y la de Matemáticas como variable respuesta. ■

En el ejemplo 2.1, el alcohol realmente causa un cambio en la temperatura corporal. No existe ninguna relación causa-efecto entre las calificaciones de Matemáticas y las de Lengua del ejemplo 2.2. De todas formas, como existe una estrecha relación entre las calificaciones de Matemáticas y de Lengua, podemos utilizar la de Lengua para predecir la de Matemáticas. En la sección 2.4 aprenderemos a hacer dicha predicción. Ésta requiere que identifiquemos una variable explicativa y una variable respuesta. Otras técnicas estadísticas ignoran esta distinción. Recuerda que llamar a una variable explicativa y a otra variable respuesta no significa necesariamente que los cambios en una de ellas causen cambios en la otra.

Muchos estudios estadísticos examinan datos de más de una variable. Afortunadamente, los estudios estadísticos de datos de varias variables se basan en las herramientas que hemos utilizado para examinar una sola variable. Los principios en los que se basa nuestro trabajo también son los mismos:

- Empieza con un gráfico; luego, añade resúmenes numéricos.
- Identifica el aspecto general y las desviaciones.
- Cuando el aspecto general sea bastante regular, utiliza un modelo matemático para describirlo.

APLICA TUS CONOCIMIENTOS

2.1. En cada una de las situaciones siguientes, ¿qué es más razonable, simplemente explorar la relación entre dos variables o contemplar una de las variables como variable explicativa y la otra como variable respuesta?

(a) La cantidad de tiempo que un alumno pasa estudiando para un examen de Estadística y la calificación obtenida en el examen.

(b) El peso y la altura de una persona.

(c) La lluvia caída durante un año y el rendimiento de un cultivo.

(d) Las calificaciones de Estadística y de Francés de los estudiantes.

(e) El tipo de trabajo de un padre y el de su hijo.

2.2. ¿Es posible predecir la altura que tiene un niño de 16 años a partir de la altura que tenía a los 6? Una manera de descubrirlo consistiría en medir la altura de un grupo suficientemente numeroso de niños de 6 años, esperar hasta que cumplieran los 16 años y entonces volver a medirlos. En este caso, ¿cuál es la variable explicativa y cuál es la variable respuesta? ¿Estas variables son categóricas o cuantitativas?

2.3. Tratamiento del cáncer de mama. El tratamiento más común para combatir el cáncer de mama consistía en la extirpación completa del pecho. Hoy en día, se suele extirpar únicamente el tumor y los ganglios linfáticos circundantes, aplicando después radioterapia en la zona afectada. El cambio de tratamiento se produjo tras un amplio experimento médico que comparó ambas técnicas: se seleccionó al azar dos grupos de enfermas; cada uno siguió un tratamiento distinto y se realizó un minucioso seguimiento de las pacientes para comprobar el periodo de supervivencia. ¿Cuál es la variable explicativa y cuál es la variable respuesta? ¿Son variables categóricas o cuantitativas?

2.2 Diagramas de dispersión

La manera más común de mostrar gráficamente la relación entre dos variables cuantitativas es un *diagrama de dispersión*. He aquí un ejemplo de diagrama de dispersión.

EJEMPLO 2.3. Notas en la prueba SAT

La tabla 2.1 proporciona datos sobre la educación en los diversos Estados de EE UU. La primera columna identifica los Estados, la segunda indica a qué región censal pertenece cada uno: *East North Central* (ENC), *East South Central* (ESC), *Middle Atlantic* (MA), *Mountain* (MTN), *New England* (NE), *Pacific* (PAC), *South Atlantic* (SA), *West North Central* (WNC) y *West South Central* (WSC). La tercera columna contiene la población de cada Estado en miles de habitantes. Las cinco variables restantes son la media de Lengua y de Matemáticas en la prueba SAT, el porcentaje de alumnos que se presentan a la prueba, el porcentaje de residentes que no se graduaron en secundaria y la media de los salarios de los profesores expresado en miles de dólares.

En EE UU se usan las medias de las calificaciones obtenidas en las pruebas SAT para evaluar los sistemas educativos, tanto estatales como locales. Este sistema de evaluación no es un buen procedimiento, ya que el porcentaje de alumnos de enseñanza media que se presenta a estas pruebas varía mucho según el Estado. Vamos a examinar la relación entre el porcentaje de alumnos que se presentan a estas pruebas en cada Estado y la media de las calificaciones de matemáticas.

Creemos que "el porcentaje de alumnos que se presentan" nos ayudará a entender "la media de los resultados". Por tanto, la variable "el porcentaje de alumnos que se presentan" es la variable explicativa y la variable "la media de las calificaciones de Matemáticas" es la variable respuesta. Queremos ver cómo varía la media de las calificaciones cuando cambia el porcentaje de alumnos que se presentan al examen. Es por este motivo que situaremos el porcentaje de alumnos que se presentan (la variable explicativa) en el eje de las abscisas. La figura 2.1 es el diagrama de dispersión en el que cada punto representa a un Estado. Por ejemplo, en Alabama el 8% de los alumnos se presentó al examen y la media de las calificaciones de Matemáticas fue de 558. Busca el 8 en el eje de las x (eje de las abscisas) y el 558 en el eje de las y (eje de las ordenadas). El Estado de Alabama aparece como el punto (8, 558), encima del 8 y a la derecha de 558. La figura 2.1 muestra cómo localizar el punto correspondiente a Alabama en el diagrama. ■

Tabla 2.1. Datos sobre la educación en EE UU.

Estado*	Región**	Población (1.000)	SAT Lengua	SAT Matemáticas	Porcentaje de alumnos presentados	Porcentaje sin estudios de secundaria	Salario de profesores ($1.000)
AL	ESC	4.273	565	558	8	33,1	31,3
AK	PAC	607	521	513	47	13,4	49,6
AZ	MTN	4.428	525	521	28	21,3	32,5
AR	WSC	2.510	566	550	6	33,7	29,3
CA	PAC	31.878	495	511	45	23,8	43,1
CO	MTN	3.823	536	538	30	15,6	35,4
CT	NE	3.274	507	504	79	20,8	50,3
DE	SA	725	508	495	66	22,5	40,5
DC	SA	543	489	473	50	26,9	43,7
FL	SA	14.400	498	496	48	25,6	33,3
GA	SA	7.353	484	477	63	29,1	34,1
HI	PAC	1.184	485	510	54	19,9	35,8
ID	MTN	1.189	543	536	15	20,3	30,9
IL	ENC	11.847	564	575	14	23,8	40,9
IN	ENC	5.841	494	494	57	24,4	37,7
IA	WNC	2.852	590	600	5	19,9	32,4
KS	WNC	2.572	579	571	9	18,7	35,1
KY	ESC	3.884	549	544	12	35,4	33,1
LA	WSC	4.351	559	550	9	31,7	26,8
ME	NE	1.243	504	498	68	21,2	32,9
MD	SA	5.072	507	504	64	21,6	41,2
MA	NE	6.092	507	504	80	20,0	42,9
MI	ENC	9.594	557	565	11	23,2	44,8
MN	WNC	4.658	582	593	9	17,6	36,9
MS	ESC	2.716	569	557	4	35,7	27,7
MO	WNC	5.359	570	569	9	26,1	33,3
MT	MTN	879	546	547	21	19,0	29,4
NE	WNC	1.652	567	568	9	18,2	31,5
NV	MTN	1.603	508	507	31	21,2	36,2
NH	NE	1.162	520	514	70	17,8	35,8
NJ	MA	7.988	498	505	69	23,3	47,9
NM	MTN	1.713	554	548	12	24,9	29,6
NY	MA	18.185	497	499	73	25,2	48,1
NC	SA	7.323	490	486	59	30,0	30,4
ND	WNC	644	596	599	5	23,3	27,0
OH	ENC	11.173	536	535	24	24,3	37,8
OK	WSC	3.301	566	557	8	25,4	28,4
OR	PAC	3.204	523	521	50	18,5	39,6
PA	MA	12.056	498	492	71	25,3	46,1
RI	NE	990	501	491	69	28,0	42,2
SC	SA	3.699	480	474	57	31,7	31,6
SD	WNC	732	574	566	5	22,9	26,3
TN	ESC	5.320	563	552	14	32,9	33,1
TX	WSC	19.128	495	500	48	27,9	32,0
UT	MTN	2.000	583	575	4	14,9	30,6

Tabla 2.1 *(continuación).*

Estado*	Región**	Población (1.000)	SAT Lengua	SAT Matemáticas	Porcentaje de alumnos presentados	Porcentaje sin estudios de secundaria	Salario de profesores ($1.000)
VT	NE	589	506	500	70	19,2	36,3
VA	SA	6.675	507	496	68	24,8	35,0
WA	PAC	5.533	519	519	47	16,2	38,0
WV	SA	1.826	526	506	17	34,0	32,2
WI	ENC	5.160	577	586	8	21,4	38,2
WY	MTN	481	544	544	11	17,0	31,6

* Para identificar los Estados véase la tabla 1.1.
** Las regiones censadas son East North Central, East South Central, Middle Atlantic, Mountain, New England Pacific, South Atlantic, West North Central y West South Central.
Fuente: *Statistical Abstract of the United States*, 1992.

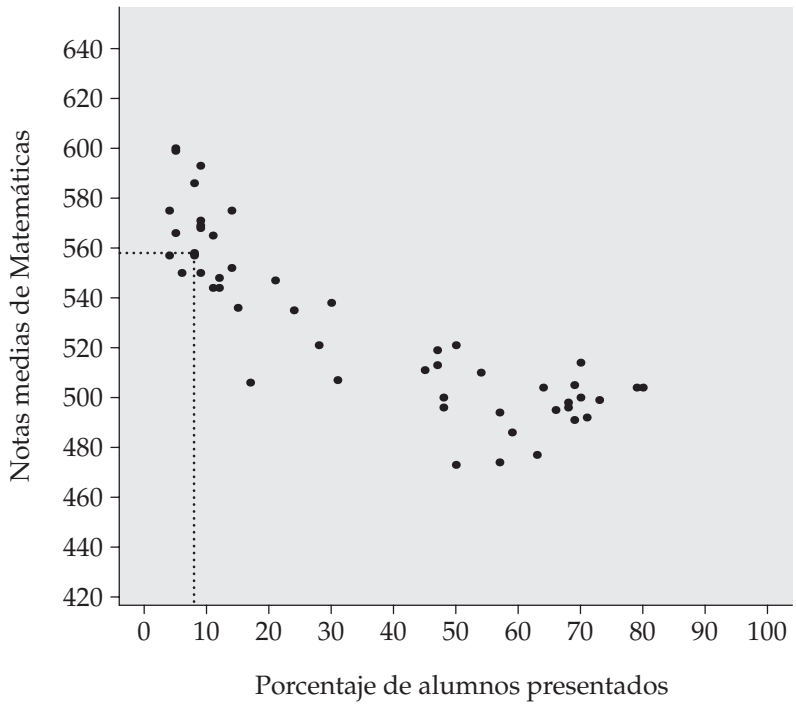


Figura 2.1. Diagrama de dispersión correspondiente a las notas medias de Matemáticas en la prueba SAT en relación con el porcentaje de alumnos que se presentan a dicho examen. La intersección de las líneas discontinuas corresponde al punto (8, 558), el dato del Estado de Alabama.

DIAGRAMA DE DISPERSIÓN

Un **diagrama de dispersión** muestra la relación entre dos variables cuantitativas medidas en los mismos individuos. Los valores de una variable aparecen en el eje de las abscisas y los de la otra en el eje de las ordenadas. Cada individuo aparece como un punto del diagrama. Su posición depende de los valores que toman las dos variables en cada individuo.

Sitúa siempre a la variable explicativa, si una de ellas lo es, en el eje de las abscisas del diagrama de dispersión. En general, llamamos a la variable explicativa x y a la variable respuesta y. Si no distinguimos entre variable explicativa y variable respuesta, cualquiera de las dos se puede situar en el eje de las abscisas.

APLICA TUS CONOCIMIENTOS

2.4. Manatís en peligro. Los manatís son unos animales grandes y dóciles que viven a lo largo de la costa de Florida. Cada año, lanchas motoras hieren o matan muchos manatís. A continuación, se presenta una tabla que contiene el número de licencias para lanchas motoras (expresado en miles de licencias por año) expedidas en Florida y el número de manatís muertos entre los años 1977 y 1990.

Año	Licencias expedidas (1.000)	Manatís muertos	Año	Licencias expedidas (1.000)	Manatís muertos
1977	447	13	1984	559	34
1978	460	21	1985	585	33
1979	481	24	1986	614	33
1980	498	16	1987	645	39
1981	513	24	1988	675	43
1982	512	20	1989	711	50
1983	526	15	1990	719	47

(a) Queremos analizar la relación entre el número de licencias anualmente expedidas en Florida y el número de manatís muertos cada año. ¿Cuál es la variable explicativa?

(b) Dibuja un diagrama de dispersión con estos datos. (Indica en los ejes los nombres de las variables, no te limites a indicar x e y.) ¿Qué nos dice el diagrama de dispersión sobre la relación entre estas dos variables?

2.2.1 Interpretación de los diagramas de dispersión

Para interpretar un diagrama de dispersión, aplica las estrategias de análisis de datos aprendidas en el capítulo 1.

EXAMEN DE UN DIAGRAMA DE DISPERSIÓN

En cualquier **gráfico de datos**, identifica el aspecto general y las **desviaciones** sorprendentes del mismo.

Puedes describir el aspecto general de un diagrama de dispersión mediante la **forma,** la **dirección** y la **fuerza** de la relación.

Un tipo importante de desviación son las **observaciones atípicas,** valores individuales que quedan fuera del aspecto general de la relación.

Grupos

La figura 2.1 muestra una *forma* clara: hay dos **grupos** distintos de Estados. En el grupo situado más a la derecha, el 45% o más de los alumnos se presentó a la prueba y las medias de los resultados estatales son bajas. Los Estados situados en el grupo de la izquierda tienen calificaciones más altas y porcentajes menores de alumnos presentados. No hay observaciones atípicas claras, es decir, no hay puntos situados de forma clara fuera de los grupos.

¿Qué puede explicar la existencia de dos grupos? En EE UU existen dos pruebas principales de acceso a la universidad: la prueba SAT (*Scholastic Assessment Test*) y la prueba ACT (*American College Testing*). En cada Estado predomina una de las dos pruebas. El grupo que aparece a la izquierda en el diagrama de dispersión de la figura 2.1 está constituido por Estados donde predomina la prueba ACT. El grupo de la derecha está formado por Estados en los que predomina la prueba SAT. En los Estados ACT, los alumnos que se presentan a la prueba SAT lo hacen porque quieren acceder a universidades más selectivas, que exigen una nota elevada en la prueba SAT. Este grupo selecto de estudiantes suele obtener unas notas en la prueba SAT superiores a las que obtienen los estudiantes de los Estados donde predomina dicha prueba.

La relación de la figura 2.1 tiene una *dirección* clara: los Estados donde el porcentaje de alumnos que se presentan a la prueba SAT es elevado tienden a tener notas medias más bajas. Tenemos una *asociación negativa* entre dos variables.

ASOCIACIÓN POSITIVA Y ASOCIACIÓN NEGATIVA

Dos variables están **asociadas positivamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores también situados por encima de la media de la otra variable, y cuando valores inferiores a la media también tienden a ocurrir conjuntamente.

Dos variables están **asociadas negativamente** cuando valores superiores a la media de una de ellas tienden a ir acompañados de valores inferiores a la media de la otra variable, y viceversa.

La *fuerza* de la relación del diagrama de dispersión está determinada por lo cerca que quedan los puntos de una determinada curva imaginaria. En general, la relación de la figura 2.1 no es fuerte —Estados con porcentajes similares de alumnos que se presentan a la prueba SAT muestran bastante variación en sus notas medias—. He aquí un ejemplo de una relación fuerte con una forma clara.

Tabla 2.2. Medias de grados-día y consumo de gas de la familia Sánchez.

Mes	Grados-día	Gas (m^3)
Noviembre	13,3	17,6
Diciembre	28,3	30,5
Enero	23,9	24,9
Febrero	18,3	21,0
Marzo	14,4	14,8
Abril	7,2	11,2
Mayo	2,2	4,8
Junio	0	3,4
Julio	0	3,4
Agosto	0,5	3,4
Septiembre	3,3	5,9
Octubre	6,7	8,7
Noviembre	16,7	17,9
Diciembre	17,8	20,2
Enero	28,9	30,8
Febrero	16,7	19,3

EJEMPLO 2.4. Calefacción del hogar

La familia Sánchez está a punto de instalar paneles solares en su casa para reducir el gasto en calefacción. Para conocer mejor el ahorro que puede significar, antes de instalar los paneles los Sánchez han ido registrando su consumo de gas en los últimos meses. El consumo de gas es más elevado cuando hace frío, por lo que debe existir una relación clara entre el consumo de gas y la temperatura exterior.

La tabla 2.2 muestra los datos de 16 meses.[1] La variable respuesta y es la media de los consumos de gas diarios durante el mes, en metros cúbicos (m^3). La variable explicativa x es la media de los grados-día de calefacción diarios durante el mes. (Los grados-día de calefacción son la medida habitual de la demanda de calefacción. Se acumula un grado-día por cada grado que la temperatura media diaria está por debajo de 18,5 ºC. Una temperatura media de 1 ºC, por ejemplo, corresponde a 17,5 grados-día de calefacción.

El diagrama de dispersión de la figura 2.2 muestra una asociación positiva fuerte. Más grados-día indican más frío y, por tanto, más gas consumido. La forma de la relación es **lineal**. Es decir, los puntos se sitúan a lo largo de una recta imaginaria. Es una relación fuerte porque los puntos se apartan poco de dicha recta. Si conocemos las temperaturas de un mes podemos predecir con bastante exactitud el consumo de gas. ■

Relación lineal

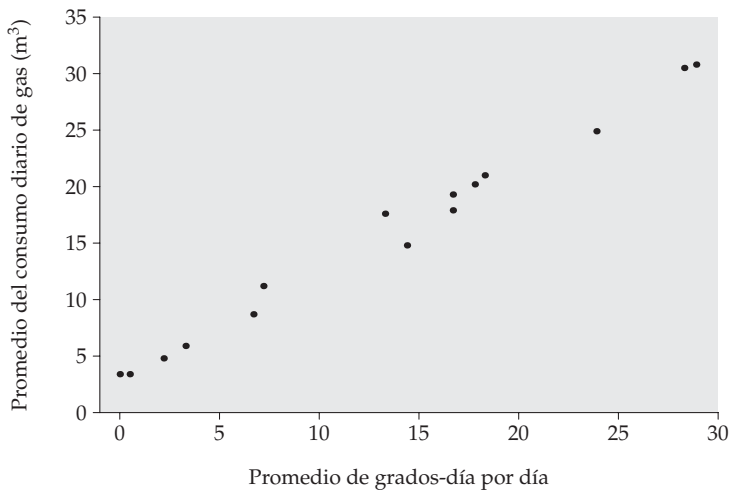


Figura 2.2. Diagrama de dispersión del consumo diario medio de gas de la familia Sánchez durante 16 meses en relación con la media diaria de grados-día en esos meses. Datos de la tabla 2.2.

[1]Datos de Robert Dale, Purdue University.

Por supuesto, no todas las relaciones son de tipo lineal. Es más, no todas las relaciones tienen una dirección clara que podamos describir como una asociación positiva o negativa. El ejercicio 2.6 da un ejemplo de una relación que no es lineal y que no tiene una dirección clara.

APLICA TUS CONOCIMIENTOS

2.5. Más sobre manatís en peligro. En el ejercicio 2.4 dibujaste un diagrama de dispersión del número de licencias para lanchas motoras registradas anualmente en Florida y del número de manatís que matan las lanchas cada año.

(a) Describe la dirección de la relación. Las variables, ¿están asociadas positiva o negativamente?

(b) Describe la forma de la relación. ¿Es lineal?

(c) Describe la fuerza de la relación. ¿Se puede predecir con precisión el número de manatís muertos cada año conociendo el número de licencias expedidas en ese año? Si Florida decidiera congelar el número de licencias en 716.000, ¿cuántos manatís matarían, aproximadamente, las lanchas motoras cada año?

2.6. El consumo, ¿aumenta con la velocidad? ¿Cómo varía el consumo de gasolina de un coche a medida que aumenta su velocidad? Aquí se presentan los datos correspondientes al modelo británico del Ford Escort. La velocidad se ha medido en kilómetros por hora y el consumo de carburante en litros de gasolina por 100 kilómetros.[2]

Velocidad (km/h)	Consumo (litros/100 km)	Velocidad (km/h)	Consumo (litros/100 km)
10	21,00	90	7,57
20	13,00	100	8,27
30	10,00	110	9,03
40	8,00	120	9,87
50	7,00	130	10,79
60	5,90	140	11,77
70	6,30	150	12,83
80	6,95		

[2]T. N. Lam, "Estimating fuel consumption from engine size", *Journal of Transportation Engineering*, 111, 1985, págs. 339-357.

(a) Dibuja un diagrama de dispersión. ¿Cuál es la variable explicativa?

(b) Describe la forma de la relación. ¿Por qué no es lineal? Explica lo que indica la forma de la relación.

(c) ¿Por qué no tiene sentido decir que las variables están asociadas positiva o negativamente?

(d) La relación, ¿es razonablemente fuerte o, por el contrario, es más bien débil? Justifica tu respuesta.

2.2.2 Inclusión de variables categóricas en los diagramas de dispersión

Desde hace tiempo, los resultados de los alumnos de las escuelas del sur de EE UU están por debajo del resto de escuelas del país. De todas formas, los esfuerzos para mejorar la educación han reducido la diferencia. En nuestro estudio sobre las pruebas de acceso a la universidad, los Estados del sur, ¿están por debajo de la media?

EJEMPLO 2.5. ¿El Sur es diferente?

Se han señalado en la figura 2.3 los Estados del Sur del diagrama de dispersión de la figura 2.1 con un símbolo diferente al resto de los Estados. (Consideramos como Estados del Sur los Estados de las regiones East South Central y South Atlantic.) En el diagrama, la mayoría de los Estados del Sur aparecen mezclados con los demás. De todas formas, algunos Estados del Sur se hallan en los bordes inferiores de sus grupos, junto con el Distrito de Columbia, que es más una ciudad que un Estado. Georgia, Carolina del Sur y Virginia Occidental tienen notas SAT inferiores a las que cabría esperar de acuerdo con el porcentaje de alumnos de secundaria que se presentan al examen. ■

Al clasificar los Estados en "Estados del Sur" y "resto de los Estados", hemos introducido una tercera variable en el diagrama de dispersión, una variable categórica que sólo tiene dos valores. Los dos valores se muestran con dos símbolos distintos. **Cuando quieras añadir una variable categórica a un diagrama de dispersión, utiliza colores o símbolos distintos para representar los puntos.**[3]

[3]W. S. Cleveland y R. McGill. "The many faces of a scatterplot", *Journal of the American Statistical Association*, 79, 1984, págs. 807-822.

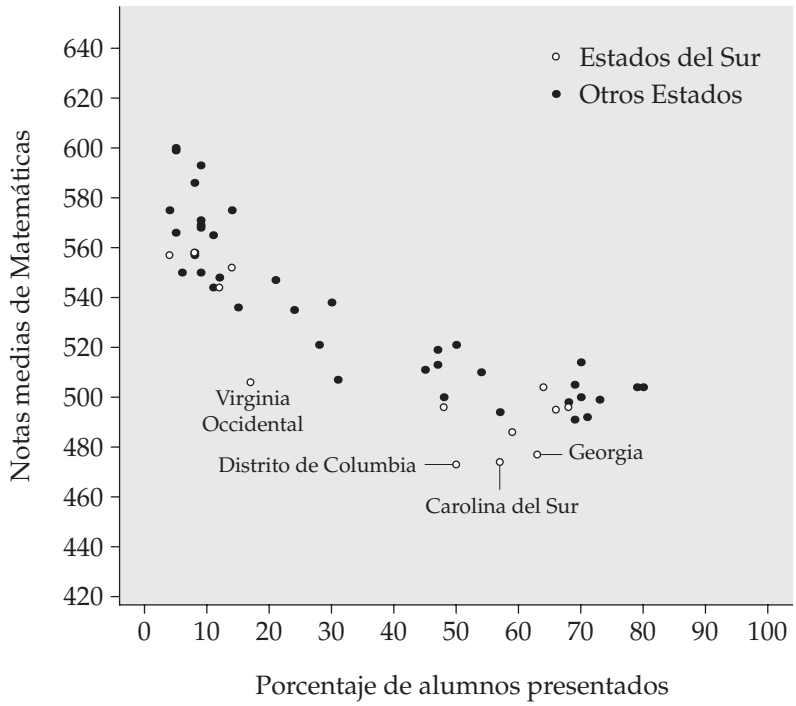


Figura 2.3. Nota media de Matemáticas en la prueba SAT y porcentaje de alumnos que se presenta a la prueba en cada Estado.

EJEMPLO 2.6. Los paneles solares, ¿reducen el consumo de gas?

Al poco tiempo de recopilar los datos que aparecen en la tabla 2.2 y en la figura 2.2, la familia Sánchez decidió instalar paneles solares en su casa. Para determinar el ahorro de gas que podía representar la instalación de estos paneles, los Sánchez registraron su consumo de gas durante 23 meses más. Para ver este efecto, añadimos los nuevos grados-día y el consumo de gas de estos meses en el diagrama de dispersión. La figura 2.4 es el resultado. Utilizamos símbolos distintos para distinguir los datos de "antes" de los de "después". En los meses poco fríos no hay mucha diferencia entre los dos grupos de datos. En cambio, en los meses más fríos el consumo de gas es claramente menor después de instalar los paneles solares. El diagrama de dispersión muestra que se ahorra energía después de instalar los paneles. ■

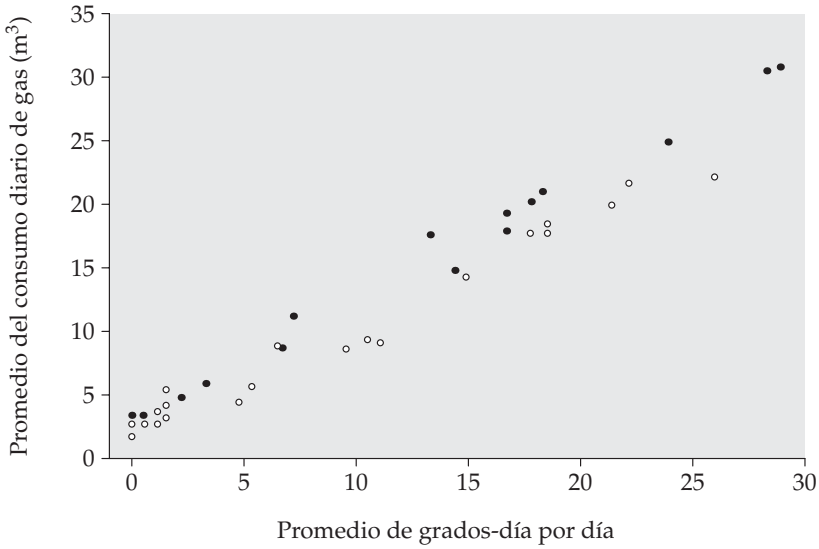


Figura 2.4. Consumo de gas con relación a los grados-día. Los puntos negros corresponden a los 16 meses sin paneles solares, mientras que los blancos corresponden a los meses con paneles solares.

Nuestro ejemplo sobre el consumo de gas tiene un problema que se presenta frecuentemente al dibujar diagramas de dispersión. Puede ser que no te des cuenta del problema cuando hagas el dibujo con un ordenador. Cuando algunos individuos tienen exactamente los mismos valores, ocupan el mismo punto del diagrama de dispersión. Fíjate en los valores de junio y julio de la tabla 2.2. La tabla 2.2 contiene datos de 16 meses, en cambio en la figura 2.2 sólo aparecen 15 puntos: junio y julio ocupan el mismo punto. Puedes utilizar símbolos distintos para señalar los puntos que representan más de una observación. Algunos programas estadísticos lo hacen automáticamente, pero otros no. Te recomendamos que utilices símbolos distintos para las observaciones repetidas cuando representes a mano un número pequeño de observaciones.

APLICA TUS CONOCIMIENTOS

2.7. La gente obesa, ¿consume más energía? El nivel metabólico de una persona, es decir, el ritmo al que su cuerpo consume energía, es un factor importante a

tener en cuenta en estudios de dietética. La tabla 2.3 proporciona datos sobre el sexo, el peso magro (peso total descontando su contenido en grasa) y el nivel metabólico en reposo de 12 mujeres y 7 hombres que eran los sujetos de un estudio de dietética. El nivel metabólico se expresa en calorías consumidas en 24 horas, la misma unidad utilizada para expresar el valor energético de los alimentos. Los investigadores creen que el peso magro corporal tiene una importante influencia en el nivel metabólico.

Tabla 2.3. Peso magro corporal y nivel metabólico.

Sujeto	Sexo	Peso (kg)	Nivel metabólico	Sujeto	Sexo	Peso (kg)	Nivel metabólico
1	H	62,0	1.792	11	M	40,3	1.189
2	H	62,9	1.666	12	M	33,1	913
3	M	36,1	995	13	H	51,9	1.460
4	M	54,6	1.425	14	M	42,4	1.124
5	M	48,5	1.396	15	M	34,5	1.052
6	M	42,0	1.418	16	M	51,1	1.347
7	H	47,4	1.362	17	M	41,2	1.204
8	M	50,6	1.502	18	H	51,9	1.867
9	M	42,0	1.256	19	H	46,9	1.439
10	H	48,7	1.614				

(a) Dibuja un diagrama de dispersión sólo con los datos de las mujeres. ¿Cuál sería la variable explicativa?

(b) La asociación entre estas dos variables, ¿es positiva o negativa? ¿Cuál es la forma de la relación? ¿Cuál es la fuerza de la relación?

(c) Ahora, añade en el diagrama de dispersión los datos de los hombres utilizando un color o un símbolo distinto al utilizado para las mujeres. La relación entre el nivel metabólico y el peso magro de los hombres, ¿es igual al de las mujeres? ¿En qué se distinguen el grupo de hombres y el grupo de mujeres?

RESUMEN DE LA SECCIÓN 2.2

Para estudiar la relación entre variables, tenemos que medir las variables sobre el mismo grupos de individuos.

Si creemos que los cambios de una variable x explican o que incluso son la causa de los cambios de una segunda variable y, a la variable x la **llamaremos variable explicativa** y a la variable y **variable respuesta**.

Un **diagrama de dispersión** muestra la relación entre dos variables cuantitativas, referidas a un mismo grupo de individuos. Los valores de una variable se

sitúan en el eje de las abscisas y los valores de la otra en el de las ordenadas. Cada observación viene representada en el diagrama por un punto.

Si una de las dos variables se puede considerar una variable explicativa, sus valores se sitúan siempre en el eje de las abscisas del diagrama de dispersión. Sitúa la variable respuesta en el eje de las ordenadas.

Para mostrar el efecto de las variables categóricas, representa los puntos de un diagrama de dispersión con colores o símbolos distintos.

Cuando analices un diagrama de dispersión, identifica su aspecto general describiendo la **dirección**, la **forma** y la **fuerza** de la relación, y luego identifica las **observaciones atípicas** y otras desviaciones.

Forma: relaciones lineales cuando los puntos del diagrama de dispersión se sitúan aproximadamente a lo largo de una recta, son una forma importante de relación entre dos variables. Las relaciones curvilíneas y las **agrupaciones** son otras formas en las que también tienes que fijarte.

Dirección: si la relación entre las dos variables tiene una dirección clara, decimos que existe una **asociación positiva** (si valores altos de las dos variables tienden a ocurrir simultáneamente) o una **asociación negativa** (si valores altos de una variable tienden a coincidir con valores bajos de la otra).

Fuerza: la **fuerza** de la relación entre variables viene determinada por la proximidad de los puntos del diagrama a alguna forma simple como, por ejemplo, una recta.

EJERCICIOS DE LA SECCIÓN 2.2

2.8. Inteligencia y calificaciones escolares. Los estudiantes que tienen coeficientes de inteligencia mayores, ¿tienden a ser mejores en la escuela? La figura 2.5 es un diagrama de dispersión correspondiente a las calificaciones medias escolares y a los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato en una escuela rural.[4]

(a) Explica en palabras qué significaría una asociación positiva entre el coeficiente de inteligencia y la nota media escolar. El diagrama, ¿muestra una asociación positiva?

(b) ¿Cuál es la forma de la relación? ¿Es aproximadamente lineal? ¿Es una relación muy fuerte? Justifica tus respuestas.

[4]Datos de Darlene Gordon, Purdue University.

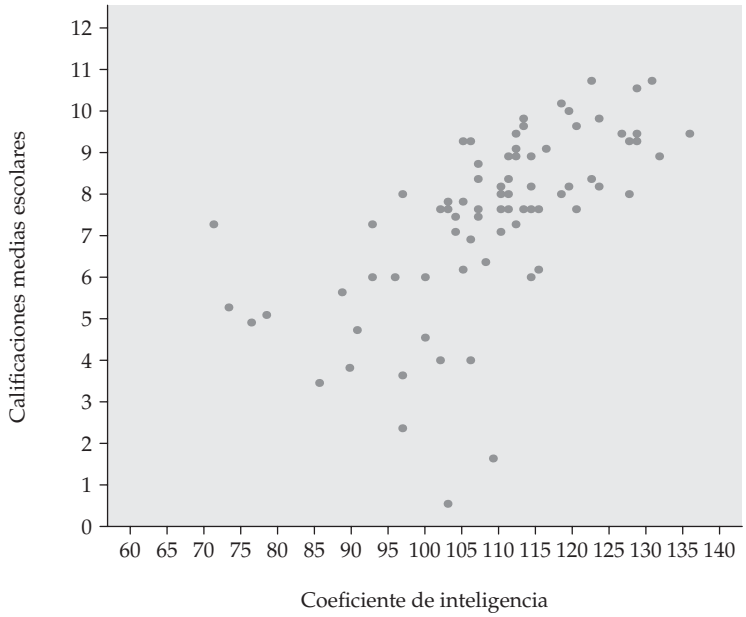


Figura 2.5. Diagrama de dispersión de las calificaciones medias escolares *versus* el coeficiente de inteligencia. Para el ejercicio 2.8.

(c) En la parte baja del diagrama aparecen algunos puntos que podríamos llamar observaciones atípicas. En concreto, un estudiante tiene una nota escolar muy baja, a pesar de tener un coeficiente de inteligencia medio. ¿Cuáles son, de forma aproximada, el coeficiente de inteligencia y la nota media escolar de este estudiante?

2.9. Calorías y sal en salchichas. Las salchichas con un contenido alto en calorías, ¿tienen también un contenido alto en sal? La figura 2.6 es un diagrama de dispersión que relaciona las calorías con el contenido en sal (expresado en miligramos de sodio) de 17 marcas distintas de salchichas elaboradas con carne de ternera.[5]

(a) Di de manera aproximada cuáles son los valores máximo y mínimo del contenido en calorías de las distintas marcas. De forma aproximada, ¿cuáles son los contenidos de sal de las marcas con más y con menos calorías?

[5] *Consumer Reports*, junio 1986, págs. 366-367.

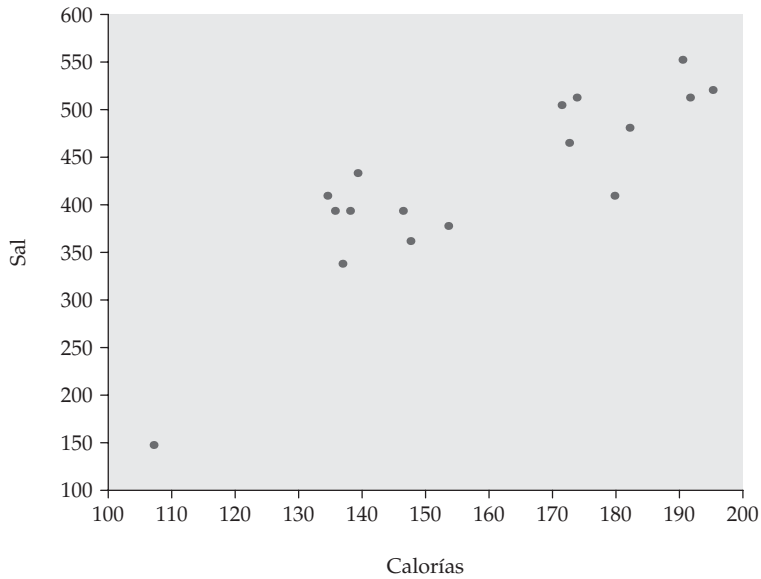


Figura 2.6. Diagrama de dispersión que relaciona las calorías y el contenido de sal de 17 marcas de salchichas. Para el ejercicio 2.9.

(b) El diagrama de dispersión, ¿muestra alguna asociación positiva o negativa clara? Explica con palabras el significado de esta asociación.

(c) ¿Has identificado alguna observación atípica? Prescindiendo de las posibles observaciones atípicas, ¿existe una relación lineal entre estas variables? Si ignoras las observaciones atípicas, ¿crees que existe una asociación fuerte entre ambas variables?

2.10. Estados ricos y Estados pobres. Una medida de la riqueza de un Estado es la mediana de ingresos por hogar. Otra medida de riqueza es la media de ingresos por persona. La figura 2.7 es un diagrama de dispersión que relaciona estas dos variables en EE UU. Ambas variables se expresan en miles de dólares. Debido a que las dos variables se expresan en las mismas unidades, la separación entre unidades es la misma en ambos ejes.[6]

[6]1997 *Statistical Abstract of the United States.*

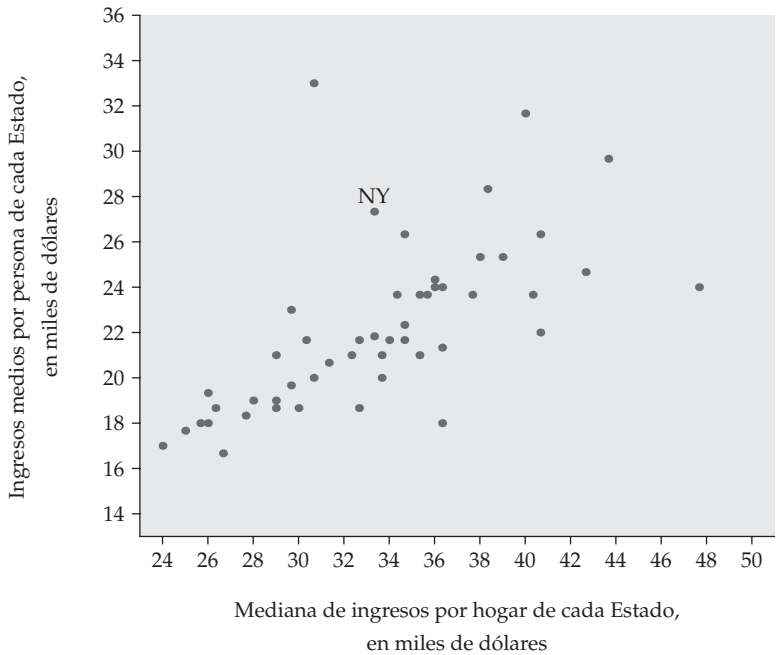


Figura 2.7. Diagrama de dispersión que relaciona los ingresos medios por persona con la mediana de ingresos por hogar. Para el ejercicio 2.10.

(a) En el diagrama de dispersión, hemos señalado el punto correspondiente a Nueva York. ¿Cuáles son, aproximadamente, los valores de la mediana de ingresos por hogar y la media de ingresos por persona?

(b) Explica por qué esperamos que haya una asociación positiva entre estas variables. Explica también, por qué esperamos que los ingresos por hogar sean mayores que los ingresos por persona.

(c) Sin embargo, en un determinado Estado, la media de los ingresos por persona puede ser mayor que la mediana de ingresos por hogar. De hecho, el Distrito de Columbia tiene una mediana de ingresos por hogar de 30.748 $ y una media de ingresos por persona de 33.435 $. Explica por qué esto puede ocurrir.

(d) Alaska es el Estado con la mediana de ingresos por hogar mayor. ¿Cuál es aproximadamente su mediana de ingresos por hogar? Podemos considerar Alaska y el Distrito de Columbia observaciones atípicas.

(e) Obviando las observaciones atípicas, describe la forma, la dirección y la fuerza de la relación.

2.11. El vino, ¿es bueno para tu corazón? Existe alguna evidencia de que tomar vino con moderación ayuda a prevenir los ataques al corazón. La tabla 2.4 proporciona datos sobre el consumo de vino (en litros de alcohol, procedente del vino, por cada 100.000 personas) y sobre las muertes anuales por ataques al corazón (muertos por cada 100.000 personas) en 19 países desarrollados.[7]

Tabla 2.4. Consumo de vino y enfermedades del corazón.

País	Consumo de alcohol*	Tasa de muertes por ataques al corazón**	País	Consumo de alcohol	Tasa de muertes por ataques al corazón
Alemania	2,7	172	Holanda	1,8	167
Australia	2,5	211	Irlanda	0,7	300
Austria	3,9	167	Islandia	0,8	211
Bélgica/Lux.	2,9	131	Italia	7,9	107
Canadá	2,4	191	Noruega	0,8	227
Dinamarca	2,9	220	N. Zelanda	1,9	266
España	6,5	86	Reino Unido	1,3	285
EE UU	1,2	199	Suecia	1,6	207
Finlandia	0,8	297	Suiza	5,8	115
Francia	9,1	71			

* Procedente del vino. ** Por 100.000 personas

(a) Dibuja un diagrama de dispersión que muestre cómo el consumo nacional de vino ayuda a explicar las muertes por ataques al corazón.

(b) Describe la forma de la relación. ¿Existe una relación lineal? ¿Es una relación fuerte?

(c) La dirección de la asociación, ¿es positiva o negativa? Explica de forma llana qué dice la relación sobre el consumo de vino y los ataques al corazón. Estos datos, ¿proporcionan una clara evidencia de que tomar vino causa una reducción de las muertes por ataques al corazón? ¿Por qué?

2.12. El profesor Moore y la natación. El profesor Moore nada 1.800 metros de forma regular. Un intento inútil de contrarrestar el paso de los años. He aquí los tiempos (en minutos) y su ritmo cardíaco después de nadar (en pulsaciones por minuto) en 23 sesiones de natación.

[7] M. H. Criqui, University of California, San Diego. Apareció en el *New York Times*, el 28 de diciembre de 1994.

Minutos	34,12	35,72	34,72	34,05	34,13	35,72	36,17	35,57
Pulsaciones	152	124	140	152	146	128	136	144
Minutos	35,37	35,57	35,43	36,05	34,85	34,70	34,75	33,93
Pulsaciones	148	144	136	124	148	144	140	156
Minutos	34,60	34,00	34,35	35,62	35,68	35,28	35,97	
Pulsaciones	136	148	148	132	124	132	139	

(a) Dibuja un diagrama de dispersión. (¿Cuál es la variable explicativa?)

(b) La asociación entre estas variables, ¿es positiva o negativa? Explica por qué crees que la relación va en este sentido.

(c) Describe la forma y la fuerza de la relación.

2.13. ¿Qué densidad de siembra es excesiva? ¿Cuál debe ser la densidad de siembra del maíz para que un agricultor obtenga el máximo rendimiento? Si se siembran pocas plantas, el suelo estará poco aprovechado y el rendimiento será bajo. Si se siembra muy denso, las plantas competirán por el agua y los nutrientes del suelo, por lo que el rendimiento tampoco será el deseado. Para determinar la densidad de siembra óptima, se hace un experimento que consiste en sembrar plantas de maíz a distintas densidades de siembra en parcelas de fertilidad similar. Los rendimientos obtenidos son los siguientes:[8]

Densidad de siembra (plantas por hectárea)	**Rendimiento (toneladas por hectárea)**			
30.000	10,1	7,6	7,9	9,6
40.000	11,2	8,1	9,1	10,1
50.000	11,1	8,7	9,4	10,1
60.000	9,1	9,3	10,5	
70.000	8,0	10,1		

(a) ¿Cuál es la variable explicativa: el rendimiento o la densidad de siembra?

(b) Dibuja un diagrama de dispersión con los datos del rendimiento y de la densidad de siembra.

(c) Describe el aspecto general de la relación. ¿Es una relación lineal? ¿Existe una asociación positiva, negativa o ninguna de las dos?

(d) Calcula los rendimientos medios de cada una de las densidades de siembra. Dibuja un diagrama de dispersión que relacione estas medias con la densidad de siembra. Une las medias con segmentos para facilitar la interpretación del

[8]W. L. Colville y D. P. McGill, "Effect of rate and method of planting on several plant characters and yield of irrigated corn", *Agronomy Journal*, 54, 1962, págs. 235-238.

diagrama. ¿Qué densidad de siembra recomendarías a un agricultor que quisiera sembrar maíz en un campo de fertilidad similar a la del experimento?

2.14. Salario de profesores. La tabla 2.1 muestra datos sobre la educación en EE UU. Es posible que los Estados con un nivel educativo menor paguen menos a sus profesores. Esto se podría explicar por el hecho de que son más pobres.

(a) Dibuja un diagrama de dispersión que relacione la media de los salarios de los profesores y el porcentaje de residentes que no tienen una carrera universitaria. Considera esta última variable como explicativa.

(b) El diagrama muestra una asociación negativa débil entre las dos variables. ¿Por qué decimos que la relación es negativa? ¿Por qué decimos que es débil?

(c) En la parte superior izquierda de tu diagrama hay una observación atípica. ¿A qué Estado corresponde?

(d) Existe un grupo bastante claro formado por nueve Estados en la parte inferior derecha del diagrama. Estos Estados tienen muchos residentes que no se graduaron en una escuela secundaria y además los salarios de los profesores son bajos. ¿Qué Estados son? ¿Se sitúan en alguna parte concreta del país?

2.15. Transformación de datos. Al analizar datos, a veces conviene hacer una **transformación de datos** que simplifique el aspecto general de la relación. A continuación se presenta un ejemplo de cómo transformando la variable respuesta se puede simplificar el aspecto del diagrama de dispersión. La población europea entre los años 1750 y 1950 creció de la siguiente manera:

Transformación de datos

Año	1750	1800	1850	1900	1950
Población (millones)	125	187	274	423	594

(a) Dibuja el diagrama de dispersión correspondiente a estos datos. Describe brevemente el tipo de crecimiento en el periodo señalado.

(b) Calcula los logaritmos de la población de cada uno de los años (puedes utilizar tu calculadora). Dibuja un nuevo diagrama de dispersión con la variable población transformada. ¿Qué tipo de crecimiento observas ahora?

2.16. Variable categórica explicativa. Un diagrama de dispersión muestra la relación entre dos variables cuantitativas. Vamos a ver un gráfico similar en el que la variable explicativa será una variable categórica en vez de una cuantitativa.

La presencia de plagas (insectos nocivos) en los cultivos se puede determinar con la ayuda de trampas. Una de ellas consiste en una lámina de plástico de distintos colores que contiene en su superficie un material pegajoso. ¿Qué colores

atraen más a los insectos? Para responder a esta pregunta un grupo de investigadores llevó a cabo un experimento que consistió en situar en un campo de avena 24 trampas de las cuales había 6 de color amarillo, 6 blancas, 6 verdes y 6 azules.[9]

Color de la trampa	Insectos capturados					
Amarillo	45	59	48	46	38	47
Blanco	21	12	14	17	13	17
Verde	37	32	15	25	39	41
Azul	16	11	20	21	14	7

(a) Dibuja un gráfico que relacione los recuentos de insectos capturados con el color de la trampa (sitúa el color de las trampas a distancias iguales en el eje de las abscisas). Calcula las medias de insectos atrapados en cada tipo de trampa, añádelas al gráfico y únelas con segmentos.

(b) ¿Qué conclusión puedes obtener de este gráfico sobre la atracción de estos colores sobre los insectos?

(c) ¿Tiene sentido hablar de una asociación positiva o negativa entre el color de la trampa y el número de insectos capturados?

2.3 Correlación

Un diagrama de dispersión muestra la forma, la dirección y la fuerza de la relación entre dos variables cuantitativas. Las relaciones lineales son especialmente importantes, ya que una recta es una figura sencilla bastante común. Decimos que una relación lineal es fuerte si los puntos del diagrama de dispersión se sitúan cerca de la recta, y débil si los puntos se hallan muy esparcidos respecto de la recta. De todas maneras, a simple vista, es difícil determinar la fuerza de una relación lineal. Los dos diagramas de dispersión de la figura 2.8 representan exactamente los mismos datos, con la única diferencia de la escala de los ejes. El diagrama de dispersión inferior da la impresión de que la asociación entre las dos variables es más fuerte. Es fácil engañar a la vista cambiando la escala.[10] Por ello, necesitamos seguir nuestra estrategia para el análisis de datos y utilizar una medida numérica que complemente el gráfico. La *correlación* es la medida que necesitamos.

[9] Adaptado de M. C. Wilson y R. E. Shade, "Relative attractiveness of various luminescent colors to the cereal leaf beetle and the meadow spittlebug", *Journal of Economic Entomology*, 60, 1967, págs. 578-580.

[10] W. S. Cleveland, P. Diaconis y R. McGill, "Variables on scatterplots look more highly correlated when the scales are increased", *Science*, 216, 1982, págs. 1138-1141.

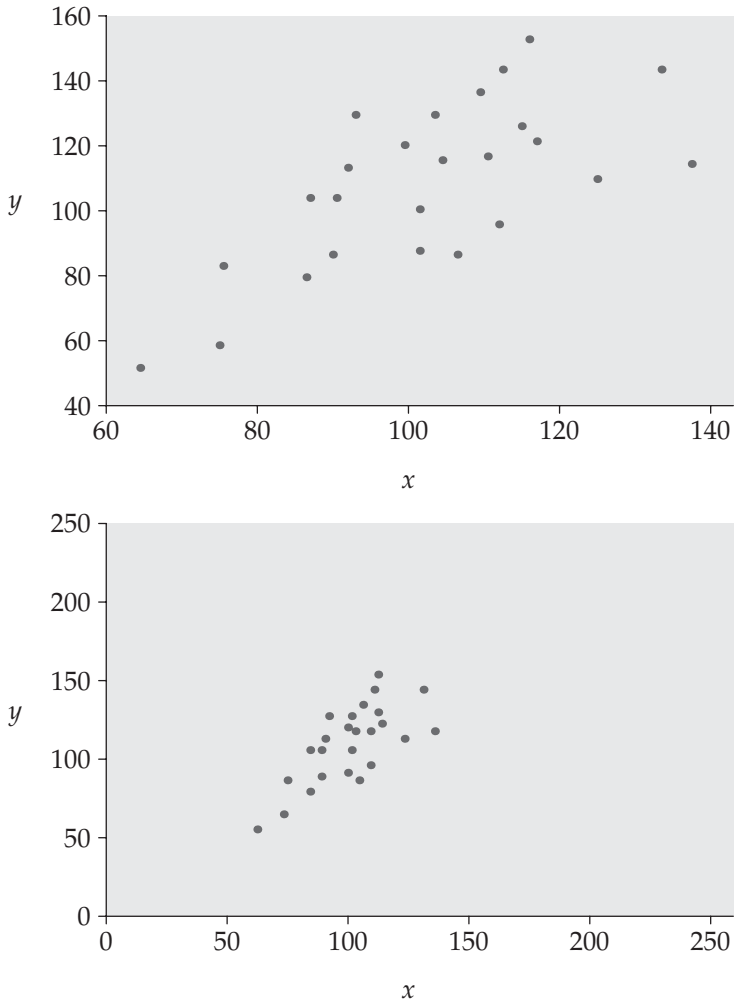


Figura 2.8. Dos diagramas de dispersión con los mismos datos. Debido a las diferentes escalas utilizadas, la fuerza de la relación lineal parece mayor en el gráfico inferior.

2.3.1 Correlación *r*

CORRELACIÓN

La **correlación** mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. La correlación se simboliza con la letra r.

Supón que tenemos datos de dos variables x e y para n individuos. Los valores para el primer individuo son x_1 e y_1, para el segundo son x_2 e y_2, etc. Las medias y las desviaciones típicas de las dos variables son $\bar{x}$ y s_x para los valores de x, e $\bar{y}$ y s_y para los valores de y. La correlación r entre x e y es

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Como siempre, $\sum$ (la letra sigma mayúscula del alfabeto griego) indica "suma estos términos para todos los individuos". La fórmula de la correlación r es algo complicada. Nos ayuda a entender qué es la correlación pero, en la práctica, conviene utilizar un programa estadístico o una calculadora para hallar r a partir de los valores de las dos variables x e y. Con el objetivo de consolidar tu comprensión del significado de la correlación, en el ejercicio 2.17 tienes que calcular la correlación paso a paso a partir de la definición.

La fórmula de r empieza estandarizando las observaciones. Supón, por ejemplo, que x es la altura en centímetros e y el peso en kilogramos y que tenemos las alturas y los pesos de n personas. Por tanto, $\bar{x}$ y s_x son la media y la desviación típica de las n alturas, ambas expresadas en centímetros. El valor

$$\frac{x_i - \bar{x}}{s_x}$$

es la altura estandarizada de la i-ésima persona, tal como vimos en el capítulo 1. La altura estandarizada nos indica a cuántas desviaciones típicas se halla la altura de un individuo con respecto a la media. Los valores estandarizados no tienen unidades de medida —en este ejemplo, las alturas estandarizadas ya no se expresan en centímetros—. Estandariza también los pesos. La correlación r es como una media de los productos de las alturas estandarizadas y de los pesos estandarizados para las n personas.

APLICA TUS CONOCIMIENTOS

2.17. Clasificación de fósiles. El *Archaeopteryx* es una especie extinguida que tenía plumas como un pájaro, pero que también tenía dientes y cola como un reptil. Sólo se conocen seis fósiles de estas características. Como estos especímenes difieren mucho en su tamaño, algunos científicos creen que pertenecen a especies distintas. Vamos a examinar algunos datos. Si los fósiles pertenecen a la misma especie y son de tamaños distintos porque unos son más jóvenes que otros, tiene que haber una relación lineal positiva entre las longitudes de algunos de los huesos en todos los individuos. Una observación atípica en esta relación sugeriría una especie distinta. He aquí los datos de las longitudes en centímetros del fémur y del húmero de cinco fósiles que conservan ambos huesos.[11]

Fémur	38	56	59	64	74
Húmero	41	63	70	72	84

(a) Dibuja un diagrama de dispersión. ¿Crees que los 5 fósiles pertenecen a la misma especie?

(b) Halla la correlación r, paso a paso. Es decir, halla la media y la desviación típica de las longitudes de los fémures y de los húmeros. (Utiliza tu calculadora para calcular las medias y las desviaciones típicas.) Halla los valores estandarizados de cada valor. Calcula r a partir de su fórmula.

(c) Ahora entra los datos en tu calculadora y utiliza la función que permite calcular directamente r. Comprueba que obtienes el mismo valor que en (b).

2.3.2 Características de la correlación

La fórmula de la correlación ayuda a ver que r es positivo cuando existe una asociación positiva entre las variables. Por ejemplo, el peso y la altura están asociados positivamente. La gente que tiene una altura superior a la media tiende también a tener un peso superior a la media. Para esta gente los valores estandarizados de altura y peso son positivos. La gente que tiene una altura inferior a la media

[11] M. A. Houck *et al*. "Allometric scaling in the earliest fossil bird, *Archaeopteryx lithographica*", *Science*, 247, 1900, págs. 195-198. Los autores llegan a la conclusión de que todos los especímenes corresponden a la misma especie.

también tiende a tener un peso inferior a la media. Los dos valores estandarizados son negativos. En ambos casos los productos de la fórmula de r son en su mayor parte positivos y, por tanto, también lo es r. De la misma manera, podemos ver que r es negativa cuando la asociación entre x e y es negativa. Un estudio más detallado de la fórmula proporciona más propiedades de r. A continuación tienes las siete ideas que necesitas conocer para poder interpretar correctamente la correlación.

1. La correlación no hace ninguna distinción entre variables explicativas y variables respuesta. Da lo mismo llamar x o y a una variable o a otra.
2. La correlación exige que las dos variables sean cuantitativas para que tenga sentido hacer los cálculos de la fórmula de r. No podemos calcular la correlación entre los ingresos de un grupo de personas y la ciudad en la que viven, ya que la ciudad es una variable categórica.
3. Como r utiliza los valores estandarizados de las observaciones, no varía cuando cambiamos las unidades de medida de x, de y o de ambas. Si en vez de medir la altura en centímetros lo hubiéramos hecho en pulgadas, o si en lugar de medir el peso en kilogramos lo hubiéramos hecho en libras, el valor de r sería el mismo. La correlación no tiene unidad de medida. Es sólo un número.
4. Una r positiva indica una asociación positiva entre las variables. Una r negativa indica una asociación negativa.
5. La correlación r siempre toma valores entre -1 y 1. Valores de r cercanos a 0 indican una relación lineal muy débil. La fuerza de la relación lineal aumenta a medida que r se aleja de 0 y se acerca a 1 o a -1. Los valores de r cercanos a -1 o a 1 indican que los puntos se hallan cercanos a una recta. Los valores extremos $r = -1$ o $r = 1$ sólo se dan cuando existe una relación lineal perfecta y los puntos del diagrama de dispersión están exactamente sobre una recta.
6. La correlación sólo mide la fuerza de una relación lineal entre dos variables. La correlación no describe las relaciones curvilíneas entre variables aunque sean muy fuertes.
7. Al igual que ocurre con la media y la desviación típica, la correlación se ve fuertemente afectada por unas pocas observaciones atípicas. La correlación de la figura 2.7 es $r = 0{,}634$ cuando se incluyen todas las observaciones, de todas formas aumenta hasta $r = 0{,}783$ cuando obviamos Alaska y el Distrito de Columbia. Cuando detectes la presencia de observaciones atípicas en el diagrama de dispersión, utiliza r con precaución.

Los diagramas de dispersión de la figura 2.9 ilustran cómo los valores de r cercanos a 1 o a -1 corresponden a relaciones lineales fuertes. Para dejar más claro el significado de r, las desviaciones típicas de ambas variables en estos diagramas son iguales, y también son iguales las escalas en los ejes de las abscisas y de las ordenadas. No es fácil, en general, estimar el valor de r a partir de la observación del diagrama de dispersión. Recuerda que un cambio de escala puede engañar tu vista, pero no modifica la correlación.

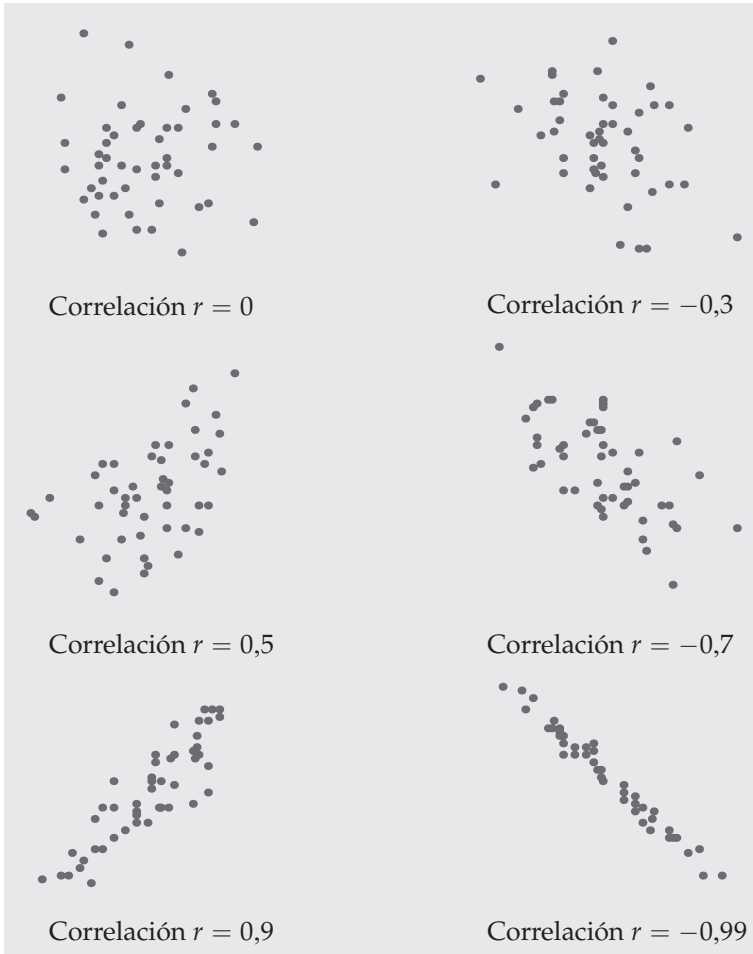


Figura 2.9. El coeficiente de correlación mide la fuerza de la asociación lineal. Cuando los puntos están muy cerca de la recta, los valores de r están más próximos a 1 o a -1.

Los datos reales que hemos examinado también ilustran cómo la correlación mide la fuerza y la dirección de relaciones lineales. La figura 2.2 muestra una relación lineal positiva muy fuerte entre los grados-día y el consumo de gas. La correlación es $r = 0{,}9953$. Compruébalo con tu calculadora utilizando los datos de la tabla 2.2. La figura 2.1 muestra una asociación negativa clara, aunque más débil, entre el porcentaje de estudiantes que se presentan a la prueba SAT y la nota media de Matemáticas en la prueba SAT de cada Estado de EE UU. En este caso, la correlación es $r = -0{,}8581$.

Recuerda que **la correlación no es una descripción completa de los datos de dos variables**, incluso cuando la relación entre las variables es lineal. Junto con la correlación tienes que dar las medias y las desviaciones típicas de x e y. (Debido a que la fórmula de la correlación utiliza las medias y las desviaciones típicas, estas medidas son las adecuadas para acompañar la correlación.) Conclusiones basadas sólo en las correlaciones puede que tengan que ser revisadas a la luz de una descripción más completa de los datos.

EJEMPLO 2.7. Puntuaciones para submarinistas

La condición física de los submarinistas profesionales se determina mediante las puntuaciones dadas por un grupo de jueces que utilizan una escala que va de 0 a 10. Existe alguna controversia sobre la objetividad de este método.

Tenemos las puntuaciones dadas por dos jueces, los señores Hernández y Fernández, a un grupo numeroso de submarinistas. ¿Concuerdan las puntuaciones de los dos jueces? Calculamos r y vemos que su valor es 0,9. Pero la media de las puntuaciones del Sr. Hernández está 3 puntos por debajo de la media del Sr. Fernández.

Estos hechos no se contradicen. Simplemente, son dos tipos diferentes de información. Las puntuaciones medias muestran que el Sr. Hernández puntúa más bajo que el Sr. Fernández. De todas formas, como el Sr. Hernández puntúa a todos los submarinistas con 3 puntos menos que el Sr. Fernández, la correlación permanece alta. Sumar o restar un mismo valor a todos los valores de x o de y no modifica la correlación. Aunque las puntuaciones de los jueces Hernández y Fernández son distintas, los submarinistas mejor puntuados por el juez Hernández son también los mejor puntuados por el juez Fernández. La r alta muestra la concordancia. Pero si el Sr. Hernández puntúa a un submarinista y el Sr. Fernández a otro, tenemos que añadir tres puntos a las puntuaciones del Sr. Hernández para que la comparación sea justa. ■

APLICA TUS CONOCIMIENTOS

2.18. Reflexiones sobre la correlación. La figura 2.5 es un diagrama de dispersión que relaciona las notas medias escolares y los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato.

(a) La correlación r de estos datos, ¿es próxima a -1, claramente negativa aunque no próxima a -1, próxima a 0, próxima a 1, claramente positiva pero no próxima a 1? Justifica tu respuesta.

(b) La figura 2.6, muestra las calorías y los contenidos de sodio de 17 marcas de salchichas. En esta ocasión, la correlación ¿es más próxima a 1 que la correlación de la figura 2.5? ¿es más próxima a 0? Justifica tu respuesta.

(c) Tanto la figura 2.5 como la figura 2.6 contienen observaciones atípicas. La eliminación de estas observaciones, ¿*aumentará* el coeficiente de correlación de una figura y lo *disminuirá* en la otra? ¿Qué ocurre en cada figura? ¿Por qué?

2.19. Si las mujeres siempre se casaran con hombres que fueran 2 años mayores que ellas, ¿cuál sería la correlación entre las edades de las esposas y las edades de sus maridos? (Sugerencia: dibuja un diagrama de dispersión con varias edades.)

2.20. Falta de correlación, pero asociación fuerte. A medida que aumenta la velocidad, el consumo de un automóvil disminuye al principio y luego aumenta. Supón que esta relación es muy regular, tal como muestran los siguientes datos de la velocidad (kilómetros por hora) y el consumo (litros por 100 km).

Velocidad (km/h)	30	45	55	70	85
Consumo (litros/100 km)	9,8	8,4	7,8	8,4	9,8

Dibuja un diagrama de dispersión del consumo con relación a la velocidad. Muestra que la correlación es $r = 0$. Explica por qué r es 0, a pesar de que existe una fuerte relación entre la velocidad y el consumo.

RESUMEN DE LA SECCIÓN 2.3

La **correlación** r mide la fuerza y la dirección de la asociación lineal entre dos variables cuantitativas x e y. Aunque puedes calcular r para cualquier diagrama de dispersión, r sólo mide la relación lineal.

La correlación indica la dirección de una relación lineal con su signo: $r > 0$ para asociaciones positivas y $r < 0$ para asociaciones negativas.

La correlación siempre cumple que $-1 \leq r \leq 1$. Valores de r cercanos a -1 o a 1 indican una fuerte asociación. Cuando los puntos de un diagrama de dispersión se sitúan exactamente a lo largo de una recta $r = \pm 1$.

La correlación ignora la distinción entre variables explicativas y variables respuesta. El valor de r no se ve afectado por cambios en las unidades de medida de cada una de las variables. De todas formas, r se puede ver muy afectada por las observaciones atípicas.

EJERCICIOS DE LA SECCIÓN 2.3

2.21. El profesor Moore y la natación. El ejercicio 2.12 proporciona datos sobre el tiempo que el profesor Moore, un hombre de mediana edad, tarda en nadar 1.800 metros y su ritmo cardíaco posterior.

(a) Si no lo hiciste en el ejercicio 2.12, calcula el coeficiente de correlación r. Explica, después de analizar el diagrama de dispersión, por qué el valor de r es razonable.

(b) Supón que los tiempos se hubieran medido en segundos. Por ejemplo, 34,12 minutos serían 2.047 segundos. ¿Cambiaría el valor de r?

2.22. Peso corporal y nivel metabólico. La tabla 2.3 proporciona datos sobre el nivel metabólico y el peso magro de 12 mujeres y 7 hombres.

(a) Dibuja un diagrama de dispersión si no lo hiciste en el ejercicio 2.7. Utiliza colores o símbolos distintos para las mujeres y para los hombres. ¿Crees que la correlación será aproximadamente igual para los hombres y las mujeres, o bastante distinta para los dos grupos? ¿Por qué?

(b) Calcula r para el grupo de las mujeres y también para el grupo de los hombres. (Utiliza la calculadora.)

(c) Calcula el peso magro medio de las mujeres y de los hombres. El hecho de que, como media, los hombres sean más pesados que las mujeres, ¿influye en las correlaciones? Si es así, ¿por qué?

(d) El peso magro se midió en kilogramos. ¿Cuál sería la correlación si lo hubiéramos medido en libras? (2,2 libras equivalen a 1 kilogramo.)

2.23. ¿Cuantas calorías? Una industria agroalimentaria solicita a un grupo de 3.368 personas que estimen el contenido en calorías de algunos alimentos. La tabla 2.5 muestra las medias de sus estimaciones y el contenido real en calorías.[12]

[12] De una encuesta de la Wheat Industry Council aparecida el 20 de octubre de 1983 en *USA Today*.

Tabla 2.5. Calorías estimadas y reales de 10 alimentos.

Alimento	Calorías estimadas	Calorías reales
225 g de leche entera	196	159
142 g de espaguetis con salsa de tomate	394	163
142 g de de macarrones con queso	350	269
Una rebanada de pan de trigo	117	61
Una rebanada de pan blanco	136	76
57 g de caramelos	364	260
Una galleta salada	74	12
Una manzana de tamaño medio	107	80
Una patata de tamaño medio	160	88
Una porción de pastel de crema	419	160

(a) Creemos que el contenido real en calorías de los alimentos, puede ayudar a explicar las estimaciones de la gente. Teniendo esto presente, dibuja un diagrama de dispersión con estos datos.

(b) Calcula la correlación r (utiliza tu calculadora). Explica, basándote en el diagrama de dispersión, por qué r es razonable.

(c) Las estimaciones son todas mayores que los valores reales. Este hecho, ¿influye de alguna manera en la correlación? ¿Cómo cambiaría r si todos los valores estimados fuesen 100 calorías más altos?

(d) Las estimaciones son demasiado altas para los espaguetis y los pasteles. Señala estos puntos en el diagrama de dispersión. Calcula r para los ocho alimentos restantes. Explica por qué r cambia en el sentido en que lo hace.

2.24. Peso del cerebro y coeficiente de inteligencia. La gente que tiene un cerebro mayor, ¿tiene también un coeficiente de inteligencia mayor? Un estudio realizado con 40 sujetos voluntarios, 20 hombres y 20 mujeres, proporciona una explicación. El peso del cerebro se determinó mediante una imagen obtenida por resonancia magnética (IRM). (En la tabla 2.6 aparecen estos datos. IRM es el recuento de "pixels" que el cerebro genera en la imagen. El coeficiente de inteligencia (CI) se midió mediante la prueba Wechsler.[13])

(a) Haz un diagrama de dispersión para mostrar la relación entre el coeficiente de inteligencia y el recuento de IRM. Utiliza símbolos distintos para hombres y mujeres. Además, halla la correlación entre ambas variables para los 40 sujetos, para los hombres y para las mujeres.

[13]L. Willerman, R. Schultz, J. N. Rutledge y E. Bigler, "In vivo brain size and intelligence", *Intelligence*, 15, 1991, págs. 223-228.

Tabla 2.6. Tamaño del cerebro y coeficiente de inteligencia.

Hombres				Mujeres			
IRM	**CI**	**IRM**	**CI**	**IRM**	**CI**	**IRM**	**CI**
1.001.121	140	1.038.437	139	816.932	133	951.545	137
965.353	133	904.858	89	928.799	99	991.305	138
955.466	133	1.079.549	141	854.258	92	833.868	132
924.059	135	945.088	100	856.472	140	878.897	96
889.083	80	892.420	83	865.363	83	852.244	132
905.940	97	955.003	139	808.020	101	790.619	135
935.494	141	1.062.462	103	831.772	91	798.612	85
949.589	144	997.925	103	793.549	77	866.662	130
879.987	90	949.395	140	857.782	133	834.344	83
930.016	81	935.863	89	948.066	133	893.983	88

(b) En general, los hombres son más corpulentos que los mujeres, por tanto sus cerebros suelen ser más grandes. ¿Cómo se muestra este efecto en tu diagrama? Halla la media del recuento de IRM para hombres y para mujeres para comprobar si existe diferencia.

(c) Tus resultados en (b) sugieren que para analizar la relación entre el coeficiente de inteligencia y el peso del cerebro, es mejor separar a hombres y mujeres. Utiliza tus resultados en (a) para comentar la naturaleza y la fuerza de esta relación para hombres y mujeres de forma separada.

2.25. Un cambio en las unidades de medida puede alterar drásticamente el aspecto de un diagrama de dispersión. Considera los siguientes datos:

x	-4	-4	-3	3	4	4
y	0,5	$-0,6$	$-0,5$	0,5	0,5	$-0,6$

(a) Dibuja un diagrama de dispersión con los datos anteriores en el que la escala de las ordenadas y la de las abscisas vayan de -6 a 6.

(b) Calcula, a partir de x e y, los valores de las nuevas variables: $x^* = \frac{x}{10}$ e $y^* = 10y$. Dibuja y^* en relación con x^* en el mismo diagrama de dispersión utilizando otros símbolos. El aspecto de los dos diagramas es muy diferente.

(c) Utiliza una calculadora para hallar la correlación entre x e y. Luego, halla la correlación entre x^* e y^*. ¿Cuál es la relación entre las dos correlaciones? Explica por qué este resultado no es sorprendente.

2.26. Docencia e investigación. Un periódico universitario entrevista a un psicólogo a propósito de las evaluaciones que hacen los estudiantes de sus profesores. El psicólogo afirma: "La evidencia demuestra que la correlación entre la

capacidad investigadora de los profesores y la evaluación docente que hacen los estudiantes es próxima a cero". El titular del periódico dice: "El profesor Cruz dice que los buenos investigadores tienden a ser malos profesores y viceversa". Explica por qué el titular del periódico no refleja el sentido de las palabras del profesor Cruz. Escribe en un lenguaje sencillo (no utilices la palabra "correlación") lo que quería decir el profesor Cruz.

2.27. Diversificación de inversiones. Un artículo en una revista de una asociación dice: "Una cartera bien diversificada incluye asientos con correlaciones bajas". El artículo incluye una tabla de correlaciones entre los rendimientos de varios tipos de inversiones. Por ejemplo, la correlación entre unos bonos municipales y acciones de grandes empresas es 0,50 y la correlación entre los bonos municipales y acciones de pequeñas empresas es 0,21.[14]

(a) María invierte mucho en bonos municipales y quiere diversificar sus inversiones añadiendo unas acciones que tengan unos rendimientos que no sigan la misma tendencia que los rendimientos de sus bonos. Para conseguir su propósito, ¿qué tipo de acciones debe escoger María, las acciones de grandes empresas o acciones de pequeñas empresas? Justifica tu respuesta.

(b) Si María quiere una inversión que tienda a aumentar cuando los rendimientos de sus bonos tiendan a disminuir, ¿qué tipo de correlación debe buscar?

2.28. Velocidad y consumo de gasolina. Los datos del ejercicio 2.20 se presentaron para mostrar un ejemplo de una relación curvilínea fuerte para la cual, sin embargo, $r = 0$. El ejercicio 2.6 proporciona datos sobre el consumo del Ford Escort con relación a la velocidad. Dibuja un diagrama de dispersión si no lo hiciste en el ejercicio 2.6. Calcula la correlación y explica por qué r está cerca de 0 a pesar de la fuerte relación entre la velocidad y el consumo.

2.29. ¿Dónde está el error? Cada una de las siguiente afirmaciones contiene un error. Explica en cada caso dónde está la incorrección.

(a) "Hay una correlación alta entre el sexo de los trabajadores y sus ingresos."

(b) "Hallamos una correlación alta ($r = 1{,}09$) entre las evaluaciones de los profesores hechas por los estudiantes y las hechas por otros profesores."

(c) "La correlación hallada entre la densidad de siembra y el rendimiento del maíz fue de $r = 0{,}23$ hectolitros."

[14] *T. Rowe Price Report*, invierno 1997, pág. 4.

2.4 Regresión mínimo-cuadrática

La correlación mide la fuerza y la dirección de la relación lineal entre dos variables cuantitativas. Si un diagrama de dispersión muestra una relación lineal, nos gustaría resumirla dibujando una recta a través de la nube de puntos. La regresión mínimo-cuadrática es un método para hallar una recta que resuma la relación entre dos variables, aunque sólo en una situación muy concreta: una de las variables ayuda a explicar o a predecir la otra. Es decir, la regresión describe una relación entre una variable explicativa y una variable respuesta.

RECTA DE REGRESIÓN

La **recta de regresión** es una recta que describe cómo cambia una variable respuesta y a medida que cambia una variable explicativa x.

A menudo, utilizamos una recta de regresión para predecir el valor de y a partir de un valor dado de x.

EJEMPLO 2.8. Predicción del consumo de gas

El diagrama de dispersión de la figura 2.10 muestra que existe una fuerte relación lineal entre la temperatura exterior media de un mes (medida en grados-día de calefacción diarios) y el consumo medio diario de gas de ese mes en casa de los Sánchez. La correlación es $r = 0{,}9953$, cerca de $r = 1$ que corresponde a los puntos situados sobre la recta. La recta de regresión trazada a través de los puntos de la figura 2.10 describe los datos muy bien.

La familia Sánchez quiere utilizar dicha relación para predecir su consumo de gas. "Si un mes tiene una media diaria de 10 grados-día, ¿cuánto gas utilizaremos?"

Predicción

Para **predecir** el consumo de gas de los días de un mes con una media de 10 grados-día, en primer lugar localiza el valor 10 en el eje de las abscisas. Luego, ves "hacia arriba y hacia la izquierda", como en la figura, para hallar el consumo de gas y que corresponde a $x = 10$. Predecimos que los Sánchez consumirán aproximadamente 12,5 metros cúbicos de gas cada día de ese mes. ■

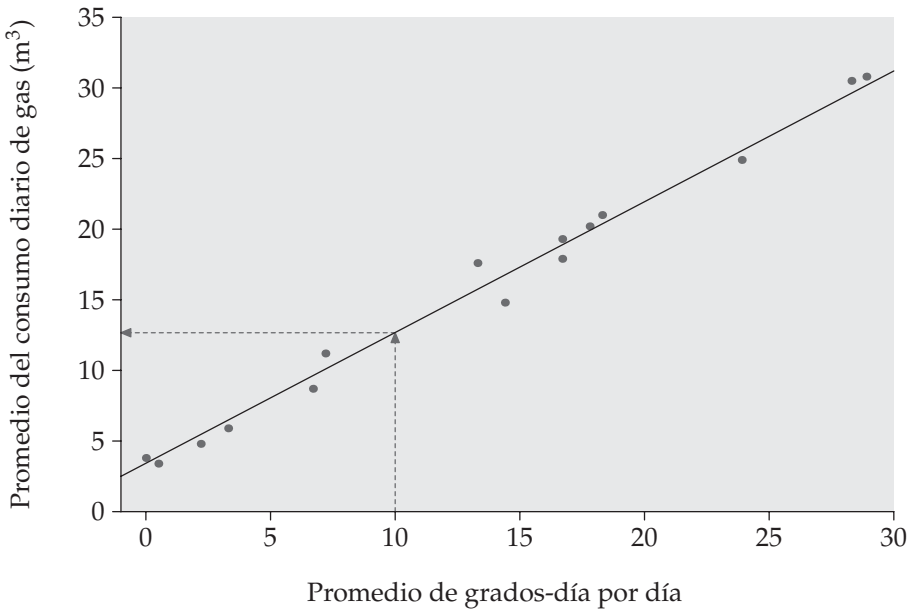


Figura 2.10. Datos del consumo de gas de la familia Sánchez, con una recta de regresión para predecir el consumo de gas a partir de los grados-día. Las líneas discontinuas muestran cómo predecir el consumo de gas en un mes con una media diaria de 10 grados-día.

2.4.1 Recta de regresión mínimo-cuadrática

Diferentes personas dibujarían, a simple vista, diferentes rectas en un diagrama de dispersión. Esto es especialmente cierto cuando los puntos están más dispersos que los de la figura 2.10. Necesitamos una manera de dibujar la recta de regresión que no dependa de nuestra intuición de por dónde tendría que pasar dicha recta. Utilizaremos la recta para predecir y a partir de x; en consecuencia, los errores de predicción estarán en y, el eje de las ordenadas del diagrama de dispersión. Si predecimos un consumo de 12,5 m^3 para un mes con 10 grados-día y el consumo real resulta ser de 13,45 m^3, nuestro error es

$$\begin{aligned} \text{error} &= \text{valor observado} - \text{valor predicho} \\ &= 13{,}45 - 12{,}5 = 0{,}95 \end{aligned}$$

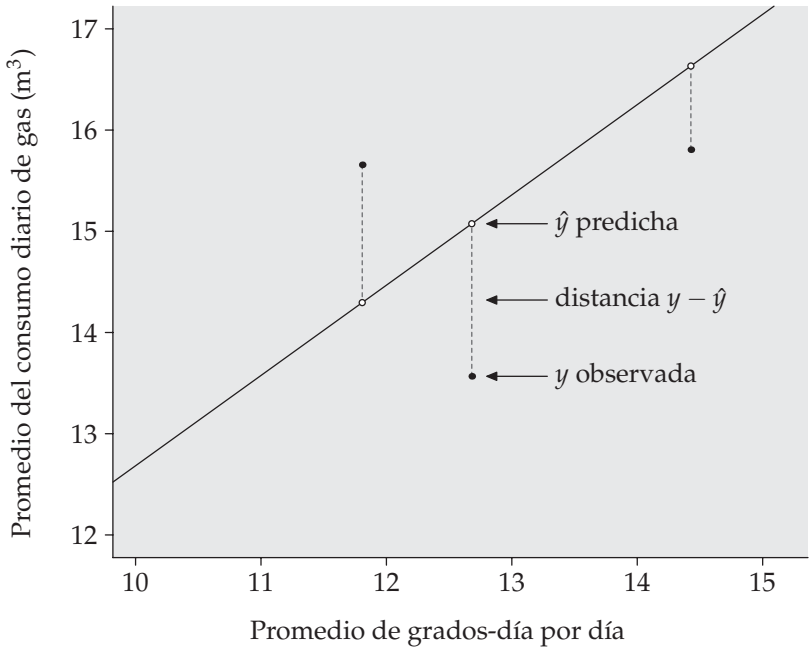


Figura 2.11. La idea de los mínimos cuadrados. Para cada observación, halla la distancia vertical de cada punto del diagrama de dispersión a la recta. La regresión mínimo-cuadrática hace que la suma de los cuadrados de estas distancias sea lo más pequeña posible.

Ninguna recta podrá pasar exactamente por todos los puntos del diagrama de dispersión. Queremos que las distancias *verticales* de los puntos a la recta sean lo más pequeñas posible. La figura 2.11 ilustra esta idea. El diagrama muestra sólo 3 puntos de la figura 2.10 conjuntamente con la recta y en una escala ampliada. La recta pasa por encima de dos de los puntos y por debajo de uno de ellos. Las distancias verticales de los puntos a la recta de regresión se han señalado con líneas discontinuas. Existen muchos procedimientos para conseguir que las distancias verticales "sean lo más pequeñas posible". El más común es el método de *mínimos cuadrados*.

RECTA DE REGRESIÓN MÍNIMO-CUADRÁTICA

La **recta de regresión mínimo-cuadrática** de y con relación a x es la recta que hace que la suma de los cuadrados de las distancias verticales de los puntos observados a la recta sea lo más pequeña posible.

Una de las razones de la popularidad de la recta de regresión mínimo-cuadrática es que el procedimiento para encontrar dicha recta es sencillo: se calcula a partir de las medias, las desviaciones típicas de las dos variables y su correlación.

ECUACIÓN DE LA RECTA DE REGRESIÓN MÍNIMO-CUADRÁTICA

Tenemos datos de la variable explicativa x y de la variable respuesta y para n individuos. A partir de los datos, calcula $\bar{x}$ e $\bar{y}$, las desviaciones típicas s_x y s_y de las dos variables y su correlación. La recta de regresión mínimo-cuadrática es

$$\hat{y} = a + bx$$

con **pendiente**

$$b = r\frac{s_y}{s_x}$$

y **ordenada en el origen**

$$a = \bar{y} - b\bar{x}$$

Escribimos $\hat{y}$ en la ecuación de la recta de regresión para subrayar que la recta *predice* una respuesta $\hat{y}$ para cada x. Debido a la dispersión de los puntos a lo largo de la recta, la respuesta predicha no coincidirá, por regla general, con la respuesta realmente *observada* y. En la práctica, no necesitas calcular primero las medias, las desviaciones típicas y la correlación. Cualquier programa estadístico, o tu calculadora, te dará la pendiente b y la ordenada en el origen a de la recta de regresión mínimo-cuadrática a partir de los valores de las variables x e y. Por tanto, puedes concentrarte en comprender y utilizar la recta de regresión.

EJEMPLO 2.9. Utilización de la recta de regresión

La recta de la figura 2.10 es de hecho la recta de regresión mínimo-cuadrática del consumo de gas con relación a los grados-día. Introduce los datos de la tabla 2.2 en tu calculadora y comprueba que la recta de regresión es

$$\hat{y} = 3{,}0949 + 0{,}94996x$$

Pendiente

La **pendiente** de una recta de regresión es importante para interpretar los datos. Esta pendiente es la tasa de cambio, la cantidad en que varía $\hat{y}$ cuando x aumenta en una unidad. La pendiente $b = 0{,}94996$ de este ejemplo dice que, como media, cada grado-día adicional predice un aumento diario del consumo de 0,94996 m^3 de gas.

Ordenada en el origen

La **ordenada en el origen** de la recta de regresión es el valor de $\hat{y}$ cuando $x = 0$. Aunque necesitamos el valor de la ordenada en el origen para dibujar la recta de regresión, sólo tiene significado estadístico cuando x toma valores cercanos a 0. En nuestro ejemplo, $x = 0$ ocurre cuando la temperatura exterior media es de al menos 18,5 °C. Predecimos que los Sánchez utilizarán una media de $a = 3{,}0949$ m^3 de gas diarios con 0 grados-día. Utilizan este gas para cocinar y para calentar el agua, y este consumo se mantiene incluso cuando no hace frío.

Predicción

La ecuación de la recta de regresión facilita la **predicción**. Tan sólo sustituye x por un valor concreto en la ecuación. Para predecir el consumo de gas a 10 grados-día, sustituye x por 10.

$$\begin{aligned} \hat{y} &= 3{,}0949 + (0{,}94996)(10) \\ &= 3{,}0949 + 9{,}4996 = 12{,}5945 \end{aligned}$$

Trazado de la recta

Para **trazar la recta** en el diagrama de dispersión, utiliza la ecuación para hallar $\hat{y}$ de dos valores de x que se encuentren en los extremos del intervalo determinado por los valores de x de los datos. Sitúa cada $\hat{y}$ sobre su respectiva x y traza la recta que pase por los dos puntos. ■

La figura 2.12 muestra los resultados de la regresión de los datos de consumo de gas obtenidos con una calculadora con funciones estadísticas y con dos programas estadísticos. Cada resultado da la pendiente y la ordenada en el origen de la recta mínimo-cuadrática, calculadas con más decimales de los que necesitamos. Los programas también proporcionan información que no necesitamos —la gracia de utilizar programas es saber prescindir de la información extra que siempre se proporciona—. En el capítulo 10, utilizaremos la información adicional de estos resultados.

```
LinReg
y = ax + b
a = .94996152
b = 3.09485166
r² = .99041538
r = .99519615
```

(a)

The regression equation is
Consumo-Gas = 3.09 + 0.95 G-dia

Predictor	Coef	Stdev	t-ratio	p
Constant	3.0949	0.3906	7.92	0.000
G-dia	0.94996	0.0250	38.04	0.000

s = 0.9539 R-sq = 99.0% R-sq(adj) = 99.0%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	1316.26	1316.26	1446.67	0.000
Error	14	12.74	0.91		
Total	15	1329.00			

(b)

Dependent variable is: Consumo-Gas
No Selector
R squared = 99.0% R squared (adjusted) = 99.0%
s = 0.9539 with 16-2 = 14 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	1316.260	1	1316.260	1447
Residual	12.738	14	0.910	

Variable	Coefficient	s.e of Coeff	t-ratio	prob
Constant	3.094852	0.3906	7.92	$\leq$0.0001
G-dia	0.949962	0.0250	38.04	$\leq$0.0001

(c)

Figura 2.12. Resultados de la regresión mínimo-cuadrática del consumo de gas obtenidos con una calculadora y con dos programas estadísticos. (a) Calculadora TI-83. (b) Minitab. (c) Data Desk.

APLICA TUS CONOCIMIENTOS

2.30. El ejemplo 2.9 da la ecuación de la recta de regresión del consumo de gas y con relación a los grados-día x de los datos de la tabla 2.2 como

$$\hat{y} = 3{,}0949 + 0{,}94966x$$

Entra los datos de la tabla 2.2 en tu calculadora.

(a) Utiliza la función de regresión de la calculadora para hallar la ecuación de la recta de regresión mínimo-cuadrática.

(b) Utiliza tu calculadora para hallar la media y la desviación típica de x e y, y su correlación r. Halla la pendiente b y la ordenada en el origen a de la recta de regresión a partir de esos valores, utilizando las ecuaciones del recuadro *Ecuación de la recta de regresión mínimo-cuadrática*. Comprueba que en (a) y en (b) obtienes la ecuación del ejemplo 2.9. (Los resultados pueden ser algo distintos debido a los errores de redondeo.)

2.31. Lluvia ácida. Unos investigadores determinaron, durante 150 semanas consecutivas, la acidez de la lluvia en una zona rural de Colorado, EE UU. La acidez se determina mediante el pH. Valores de pH bajos indican una acidez alta. Los investigadores observaron una relación lineal entre el pH y el paso del tiempo e indicaron que la recta de regresión mínimo-cuadrática

$$\text{pH} = 5{,}43 - (0{,}0053 \times \text{semanas})$$

se ajustaba bien a los datos.[15]

(a) Dibuja esta recta. ¿La asociación es positiva o negativa? Explica de una manera sencilla el significado de esta asociación.

(b) De acuerdo con la recta de regresión, ¿cuál era el pH al comienzo del estudio (semana $= 1$)? ¿Y al final (semana $= 150$)?

(c) ¿Cuál es la pendiente de la recta de regresión? Explica claramente qué indica la pendiente respecto del cambio del pH del agua de lluvia en esta zona rural.

2.32. Manatís en peligro. El ejercicio 2.4 proporciona datos sobre el número de lanchas registradas en Florida y el número de manatís muertos por las lanchas motoras entre 1977 y 1990. La recta de regresión para predecir los manatís muertos a partir del número de lanchas motoras registradas es

$$\text{muertos} = -41{,}4 + (0{,}125 \times \text{lanchas})$$

[15] W. M. Lewis y M. C. Grant, "Acid precipitation in the western United States", *Science*, 207, 1980, págs. 176-177.

(a) Dibuja un diagrama de dispersión y añádele la recta de regresión. Predice el número de manatís que matarán las lanchas en un año en que se registraron 716.000 lanchas.

(b) He aquí nuevos datos sobre los manatís muertos durante cuatro años más.

Año	Licencias expedidas (1.000)	Manatís muertos
1991	716	53
1992	716	38
1993	716	35
1994	735	49

Añade estos puntos al diagrama de dispersión. Durante estos cuatro años, Florida tomó fuertes medidas para proteger a los manatís. ¿Observas alguna evidencia de que estas medidas tuvieron éxito?

(c) En el apartado (a) predijiste el número de manatís muertos en un año con 716.000 lanchas registradas. En realidad, el número de lanchas registradas se mantuvo en 716.000 durante los siguientes tres años. Compara las medias de manatís muertos en estos años con tu predicción en (a). ¿Qué nivel de exactitud has alcanzado?

2.4.2 Características de la regresión mínimo-cuadrática

La regresión mínimo-cuadrática tiene en cuenta las distancias de los puntos a la recta sólo en la dirección de y. Por tanto, en una regresión las variables x e y juegan papeles distintos.

Característica 1. La distinción entre variable explicativa y variable respuesta es básica en regresión. La regresión mínimo-cuadrática considera sólo las distancias verticales de los puntos a la recta. Si cambiamos los papeles de las dos variables, obtenemos una recta de regresión-mínimo cuadrática distinta.

EJEMPLO 2.10. El universo se expande

La figura 2.13 es un diagrama de dispersión dibujado con los datos que sirvieron de base para descubrir que el Universo se está expandiendo. Son las distancias a la Tierra de 24 galaxias y las velocidades con que éstas se alejan de nosotros, proporcionadas por el astrónomo Edwin Hubble en 1929.[16] Existe una relación

[16]E. P. Hubble, "A relation between distance and radial velocity among extra-galactic nebulae", *Proceedings of the National Academy of Sciences*, 15, 1929, págs. 168-173.

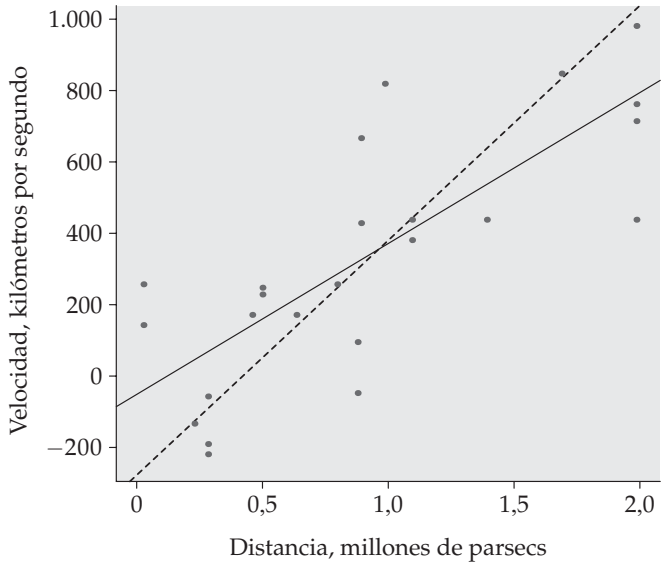


Figura 2.13. Diagrama de dispersión de los datos de Hubble sobre la distancia a la Tierra de 24 galaxias y la velocidad con la que éstas se alejan de nosotros. Las dos rectas son de regresión mínimo-cuadrática: la de la velocidad con relación a la distancia (línea continua) y la de la distancia con relación a la velocidad (línea discontinua).

lineal positiva, $r = 0{,}7842$, de manera que las galaxias que se hallan más lejos se alejan más rápidamente. De hecho, los astrónomos creen que la relación es perfectamente lineal y que la dispersión se debe a errores de medición.

Las dos rectas del dibujo son rectas de regresión mínimo-cuadrática. La recta de trazado continuo es la regresión de la velocidad con relación a la distancia, mientras que la de trazado discontinuo es la regresión de la distancia con relación a la velocidad. *La regresión de la velocidad con relación a la distancia y la regresión de la distancia con relación a la velocidad dan rectas distintas.* Al determinar la recta de regresión, debes saber cuál es la variable explicativa. ■

Característica 2. Existe una estrecha conexión entre la correlación y la regresión. La pendiente de la recta de regresión mínimo-cuadrática es

$$b = r\frac{s_y}{s_x}$$

Esta ecuación indica que, a lo largo de la recta de regresión, a **un cambio de una desviación típica de x le corresponde un cambio de r desviaciones típicas de y.** Cuando las variables están perfectamente correlacionadas ($r = 1$ o $r = -1$), el cambio en la respuesta predicha $\hat{y}$ es igual al cambio de x (expresado en desviaciones típicas). En los restantes casos, como $-1 \leq r \leq 1$, el cambio de $\hat{y}$ es menor que el cambio de x. A medida que la correlación es menos fuerte, la predicción $\hat{y}$ se mueve menos en respuesta a los cambios de x.

Carecterística 3. **La recta de regresión mínimo-cuadrática siempre pasa por el punto $(\bar{x}, \bar{y})$** del diagrama de dispersión de y con relación a x. Por tanto, la recta de regresión mínimo-cuadrática de y con relación a x es la recta de pendiente $r\frac{s_y}{s_x}$ que pasa a través del punto $(\bar{x}, \bar{y})$. Podemos describir completamente la regresión con $\bar{x}$, s_x, $\bar{y}$, s_y y r.

Característica 4. La correlación r describe la fuerza de la relación lineal. En este contexto se expresa de la siguiente manera: **el cuadrado de la correlación, r^2, es la fracción de la variación de las y que explica la recta de regresión mínimo-cuadrática de y con relación a x.**

La idea de la regresión es la siguiente: cuando existe una relación lineal, parte de la variación de y se explica por el hecho de que cuando x cambia, arrastra consigo a y. Mira otra vez la figura 2.10. Hay mucha variación en los valores observados de y, los datos de consumo de gas. Los valores de y toman valores que van de 3 a 31. El diagrama de dispersión muestra que la mayor parte de la variación de y se explica por la variación de la temperatura exterior (medida en grados-día x) que arrastra consigo el consumo de gas. Sólo existe una pequeña variación residual de y que aparece en la dispersión de los puntos a lo largo de la recta. Por otro lado, los puntos de la figura 2.13 están mucho más dispersos. La dependencia lineal de la velocidad con relación a la distancia explica sólo una parte de la variación observada en la velocidad. Podrías adivinar, por ejemplo, que cuando $x = 2$ el valor de y será mayor que cuando $x = 0$. De todas formas, existe todavía una variación considerable de y cuando x se mantiene fija —mira los cuatro puntos de la figura 2.13 cuando $x = 2$—. Esta idea se puede expresar algebraicamente, aunque no lo haremos. Es posible dividir la variación total de los valores observados de y en dos partes. Una de ellas es la variación que esperamos obtener de $\hat{y}$ a medida que x se mueve a lo largo de la recta de regresión. La otra mide la variación de los datos con relación a la recta. El cuadrado de la correlación r^2 es el primero de estos dos componentes expresado como fracción de la variación total.

$$r^2 = \frac{\text{variación de } \hat{y} \text{ junto con } x}{\text{variación total de las } y \text{ observadas}}$$

EJEMPLO 2.11. Utilización de r^2

En la figura 2.10, $r = 0{,}9953$ y $r^2 = 0{,}9906$. Más del 99% de la variación del consumo de gas se explica por la relación lineal con los grados-día. En la figura 2.13, $r = 0{,}7842$ y $r^2 = 0{,}6150$. La relación lineal entre la distancia y la velocidad explica el 61,5% de la variación de las dos variables. Hay dos rectas de regresión, pero existe sólo una correlación y r^2 ayuda a interpretar ambas regresiones. ■

Cuando presentes los resultados de una regresión, da el valor de r^2 como una medida de lo buena que es la respuesta que proporciona la regresión. Todos los resultados de programas estadísticos de la figura 2.12 incluyen r^2, en tanto por uno o en tanto por ciento. Cuando tengas una correlación, elévala al cuadrado para tener una idea más precisa de la fuerza de la asociación. Una correlación perfecta ($r = -1$ o $r = 1$) significa que los puntos se hallan perfectamente alineados a lo largo de una recta. En este caso $r^2 = 1$, es decir, toda la variación de una variable se explica por la relación lineal con la otra variable. Si $r = -0{,}7$ o $r = 0{,}7$, entonces $r^2 = 0{,}49$. Es decir, aproximadamente la mitad de la variación se explica con la relación lineal. En la escala de la r^2, una correlación de $r = \pm 0{,}7$ se halla a medio camino entre 0 y ± 1.

Las características anteriores son propiedades especiales de la regresión mínimo-cuadrática. No son ciertas para otros métodos de ajuste de una recta a unos datos. Otra razón por la cual el método de los mínimos cuadrados es el más común para ajustar una recta de regresión a unos datos es que tiene muchas propiedades interesantes.

APLICA TUS CONOCIMIENTOS

2.33. El profesor Moore y la natación. He aquí los tiempos (en minutos) que tarda el profesor Moore en nadar 1.800 metros y su ritmo cardíaco después de bracear (en pulsaciones por minuto) en 23 sesiones de natación.

Minutos	34,12	35,72	34,72	34,05	34,13	35,72	36,17	35,57
Pulsaciones	152	124	140	152	146	128	136	144
Minutos	35,37	35,57	35,43	36,05	34,85	34,70	34,75	33,93
Pulsaciones	148	144	136	124	148	144	140	156
Minutos	34,60	34,00	34,35	35,62	35,68	35,28	35,97	
Pulsaciones	136	148	148	132	124	132	139	

(a) Un diagrama de dispersión muestra una relación lineal negativa relativamente fuerte. Utiliza tu calculadora o un programa informático para comprobar que la recta de regresión mínimo-cuadrática es

$$\text{pulsaciones} = 479{,}9 - (9{,}695 \times \text{minutos})$$

(b) Al siguiente día el profesor tardó 34,30 minutos. Predice su ritmo cardíaco. En realidad su pulso fue 152. ¿Cómo de exacta es tu predicción?

(c) Supón que sólo conociéramos que las pulsaciones fueron 152. Ahora quieres predecir el tiempo que el profesor estuvo nadando. Halla la recta de regresión mínimo-cuadrática apropiada para la ocasión. ¿Cuál es tu predicción? ¿Es muy exacta?

(d) Explica de forma clara, a alguien que no sepa estadística, por qué las dos rectas de regresión son distintas.

2.34. Predicción del comportamiento de mercados de valores. Algunas personas creen que el comportamiento de un mercado de valores en enero permite predecir el comportamiento del mercado durante el resto del año. Toma como variable explicativa x el porcentaje de cambio en el índice del mercado de valores en enero y como variable respuesta y la variación del índice a lo largo de todo el año. Creemos que existe una correlación positiva entre x e y, ya que el cambio de enero contribuye al cambio anual. Cálculos a partir de datos del periodo 1960-1997 dan

$$\bar{x} = 1{,}75\% \qquad s_x = 5{,}36\% \qquad r = 0{,}596$$
$$\bar{y} = 9{,}07\% \qquad s_y = 15{,}35\%$$

(a) ¿Qué porcentaje de la variación observada en los cambios anuales del índice se explica a partir de la relación lineal con el cambio del índice en enero?

(b) ¿Cuál es la ecuación de la recta mínimo-cuadrática para la predicción del cambio en todo el año a partir del cambio en enero?

(c) En enero el cambio medio es $\bar{x} = 1{,}75\%$. Utiliza tu recta de regresión para predecir el cambio del índice en un año para el cual en enero sube un 1,75%. ¿Por qué podías haber conocido este resultado (hasta donde te permite el error de redondeo) sin necesidad de hacer ningún cálculo?

2.35. Castores y larvas de coleóptero. A menudo los ecólogos hallan relaciones sorprendentes en nuestro entorno. Un estudio parece mostrar que los castores pueden ser beneficiosos para una determinada especie de coleóptero. Los investigadores establecieron 23 parcelas circulares, cada una de ellas de 4 metros de diámetro, en una zona en la que los castores provocaban la caída de álamos al

alimentarse de su corteza. En cada parcela, los investigadores determinaron el número de tocones resultantes de los árboles derribados por los castores y el número de larvas del coleóptero. He aquí los datos:[17]

Tocones	2	2	1	3	3	4	3	1	2	5	1	3
Larvas	10	30	12	24	36	40	43	11	27	56	18	40
Tocones	2	1	2	2	1	1	4	1	2	1	4	
Larvas	25	8	21	14	16	6	54	9	13	14	50	

(a) Haz un diagrama de dispersión que muestre cómo el número de tocones debidos a los castores influye sobre el de larvas. ¿Qué muestra tu diagrama? (Los ecólogos creen que los brotes que surgen de los tocones resultan más apetecibles para las larvas ya que son más tiernos que los de los árboles mayores.)

(b) Halla la recta de regresión mínimo-cuadrática y dibújala en tu diagrama.

(c) ¿Qué porcentaje de la variación observada en el número de larvas se puede explicar por la dependencia lineal con el número de tocones?

2.4.3 Residuos

Una recta de regresión es un modelo matemático que describe una relación lineal entre una variable explicativa y una variable respuesta. Las desviaciones de la relación lineal también son importantes. Cuando se dibuja una recta de regresión, se ven las desviaciones observando la dispersión de los puntos respecto a dicha recta. Las distancias verticales de los puntos a la recta de regresión mínimo-cuadrática son lo más pequeñas posible, en el sentido de que tienen la menor suma de cuadrados posible. A estas distancias les damos un nombre: *residuos*.

RESIDUOS

Un **residuo** es la diferencia entre el valor observado de la variable respuesta y el valor predicho por la recta de regresión. Es decir,

$$\begin{aligned} \text{residuo} &= y \text{ observada} - y \text{ predicha} \\ &= y - \hat{y} \end{aligned}$$

[17] G. D. Martinsen, E. M. Driebe y T. G. Whitham, "Indirect interactions mediated by changing plant chemistry: beaver browsing benefits beetles", *Ecology*, 79, 1998, págs. 192-200.

Tabla 2.7. Edad de la primera palabra y puntuación en la prueba Gesell.

Niño	Edad	Puntuación	Niño	Edad	Puntuación
1	15	95	12	9	96
2	26	71	13	10	83
3	10	83	14	11	84
4	9	91	15	11	102
5	15	102	16	10	100
6	20	87	17	12	105
7	18	93	18	42	57
8	11	100	19	17	121
9	8	104	20	11	86
10	20	94	21	10	100
11	7	113			

EJEMPLO 2.12. Predicción de la inteligencia

¿Predice su inteligencia posterior la edad a la que un niño empieza a hablar? Un estudio del desarrollo de 21 niños, registró la edad, en meses, a la que cada niño pronunciaba la primera palabra y su puntuación en la prueba Gesell (*Gesell Adaptative Score*), una prueba de aptitud llevada a cabo mucho más tarde. Los datos aparecen en la tabla 2.7.[18]

La figura 2.14 es un diagrama de dispersión en el que se toma la edad en que se pronunció la primera palabra como variable explicativa x y la puntuación en la prueba Gesell como variable respuesta y. Los niños 3 y 13, y los niños 16 y 21 tienen valores idénticos para ambas variables, por lo que se utiliza un símbolo distinto para mostrar que estos puntos representan a dos individuos diferentes. El diagrama muestra una asociación negativa, es decir, los niños que empiezan a hablar más tarde tienden a tener puntuaciones más bajas en la prueba que los niños que hablan antes. El aspecto general de la relación es moderadamente lineal. La correlación describe la dirección y la fuerza de la relación lineal, $r = -0{,}640$.

La recta que se ha trazado en el diagrama es la recta de regresión mínimo-cuadrática de la puntuación Gesell con relación a la edad de la primera palabra. Su ecuación es

$$\hat{y} = 109{,}8738 - 1{,}1270x$$

[18]M. R. Mickey, O. J. Dunn y V. Clark, "Note on the use of stepwise regression in detecting outliers", *Computers and Biomedical Research*, 1, 1967, págs. 105-111. Estos datos han sido utilizados por varios autores; yo los he hallado en N. R. Draper y J. A. John, "Influential observations and outliers in regression", *Technometrics*, 23, 1981, págs. 21-26.

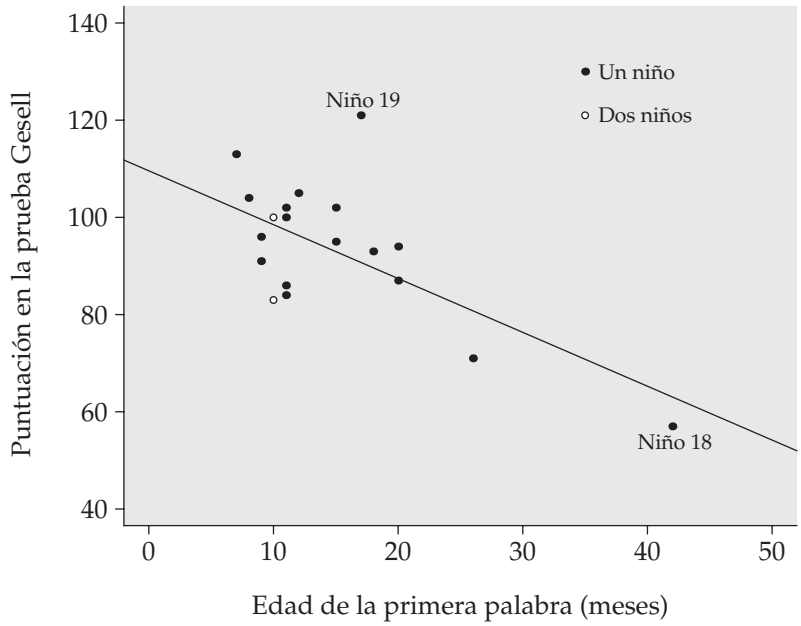


Figura 2.14. Diagrama de dispersión de las puntuaciones de la prueba Gesell con relación a la edad de la primera palabra de 21 niños. La recta es la recta de regresión mínimo-cuadrática para predecir la puntuación Gesell a partir de la primera palabra.

Para el primer niño, que empezó a hablar a los 15 meses, predecimos la puntuación

$$\hat{y} = 109{,}8738 - (1{,}1270)(15) = 92{,}97$$

La puntuación real de este niño fue de 95. El residuo es

$$\begin{aligned} \text{residuo} &= y \text{ observada} - y \text{ predicha} \\ &= 95 - 92{,}97 = 2{,}03 \end{aligned}$$

El residuo es positivo porque el punto se halla por encima de la recta. ■

Existe un valor residual para cada punto. Hallar los valores residuales con una calculadora es bastante laborioso, ya que primero tienes que hallar la respuesta predicha para cada x. Los programas estadísticos te dan todos los residuos

a la vez. He aquí los 21 residuos de los datos de la prueba Gesell obtenidos con un programa estadístico:

```
2.0310   -9.5721   -15.6040   -8.7309    9.0310    -0.3341     3.4120
2.5230    3.1421     6.6659   11.0151   -3.7309   -15.6040   -13.4770
4.5230    1.3960     8.6500   -5.5403   30.2850   -11.4770     1.3960
```

Debido a que los residuos muestran a qué distancia se hallan los datos de nuestra recta de regresión, el examen de los residuos nos ayuda a valorar en que medida la recta describe la distribución de los datos. A pesar de que los residuos se pueden calcular a partir de cualquier modelo que se haya ajustado a los datos, los de la recta de regresión mínimo-cuadrática tienen una propiedad especial: **la media de los residuos es siempre cero**.

Compara el diagrama de dispersión de la figura 2.14 con el *diagrama de residuos* correspondiente a los mismos datos de la figura 2.15. En dicha figura, la recta horizontal que pasa por cero nos ayuda a orientarnos. Esta recta corresponde a la recta de regresión de la figura 2.14.

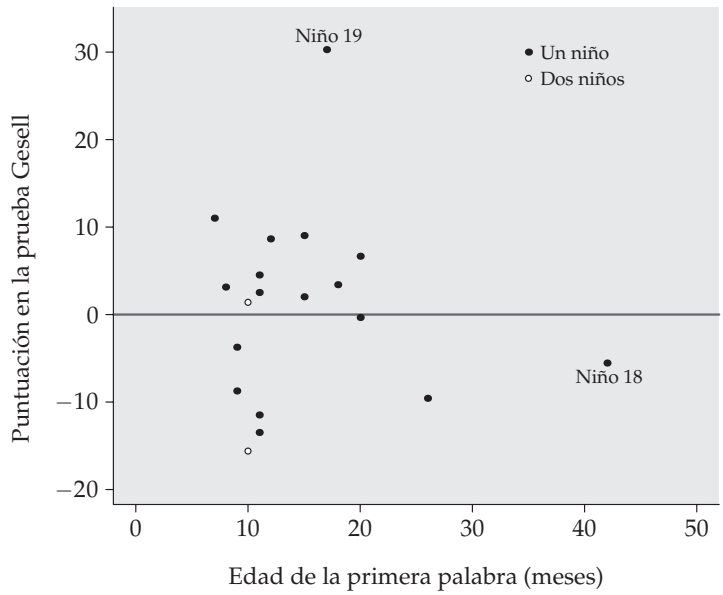


Figura 2.15. Diagrama de residuos para la regresión de las puntuaciones en la prueba Gesell en relación con la edad de la primera palabra. El niño 19 es una observación atípica. El niño 18 es una observación influyente que no tiene un residuo grande.

DIAGRAMA DE RESIDUOS

Un **diagrama de residuos** es un diagrama de dispersión de los residuos de la regresión con relación a la variable explicativa. Los diagramas de residuos nos ayudan a valorar el ajuste de la recta de regresión.

Si la recta de regresión se ajusta bien a la relación entre x e y, los residuos no tienen que tener ninguna distribución especial. En dicho caso, la distribución de residuos se parecerá a la distribución que de forma simplificada se muestra en la figura 2.16(a). Este diagrama muestra que la distribución de los residuos es uniforme a lo largo de la recta, no se detectan observaciones atípicas. Cuando examines los residuos en el diagrama de dispersión o en el diagrama de residuos, has de fijarte en algunos detalles:

- **Una forma curva** de la distribución de los residuos indica que la relación no es lineal. La figura 2.16(b) es un ejemplo ilustrativo. La recta no es una buena descripción para estos datos.
- **Un crecimiento o decrecimiento de la dispersión de los residuos** a medida que aumentan las x. La figura 2.16(c) es un ejemplo. En él, la predicción de y será menos precisa para valores de x mayores.
- **Los puntos individuales con residuos grandes**, como el del niño 19 de las figuras 2.14 y 2.15. Estos puntos son observaciones atípicas, ya que no encajan en el aspecto lineal de la nube de puntos.
- **Los puntos individuales que son extremos en el eje de las abscisas**, como el niño 18 de las figuras 2.14 y 2.15. Estos puntos pueden no tener grandes residuos, pero pueden ser muy importantes. Más adelante estudiaremos este tipo de puntos.

2.4.4 Observaciones influyentes

Los niños 18 y 19 del ejemplo Gesell son poco frecuentes, pero por motivos distintos. El niño 19 está lejos de la recta de regresión. Este niño empezó a hablar mucho más tarde que los demás. Su valor Gesell es tan alto que tendríamos que comprobar que no se trata de un error de transcripción de los datos. De todas

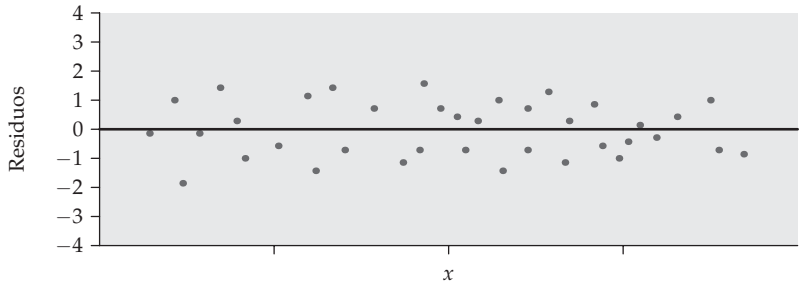


Figura 2.16(a). Distribuciones idealizadas de diagramas de residuos de la recta de regresión mínimo-cuadrática. El gráfico (a) indica un buen ajuste de la recta de regresión.

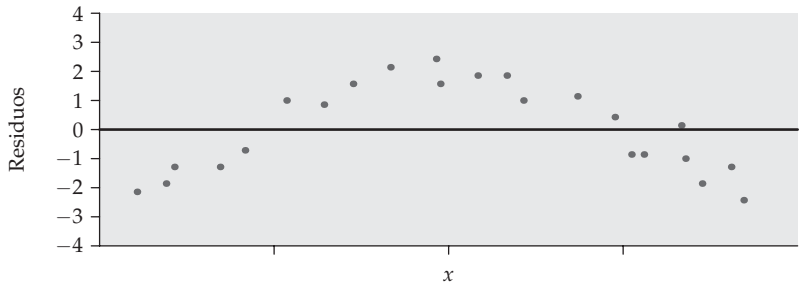


Figura 2.16(b). El gráfico (b) muestra una forma curva, por tanto, la recta se ajusta mal.

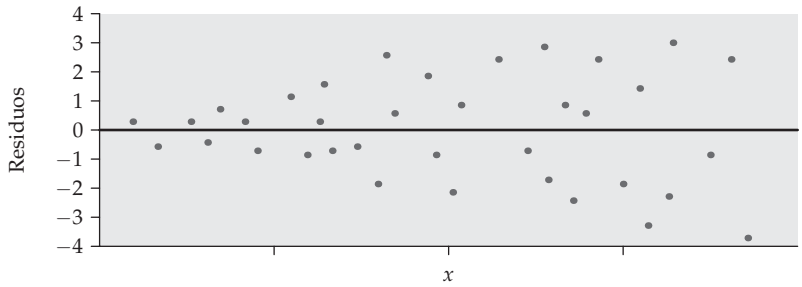


Figura 2.16(c). La variable respuesta y del gráfico (c) presenta más dispersión para los valores mayores de la variable explicativa x. Por tanto, la predicción será menos precisa cuanto mayor sea x.

formas, el valor Gesell es correcto. El punto correspondiente al niño 18 se halla cerca de la recta, sin embargo se encuentra alejado en la dirección de las abscisas. El niño 18 fue el que empezó a hablar más tarde. *Debido a su posición extrema en el eje de las abscisas tiene una gran influencia sobre la posición de la recta de regresión*. La figura 2.17 añade una segunda recta de regresión, calculada tras excluir al niño 18. Puedes ver que sin esta observación la posición de la recta se ha modificado. A estos puntos los llamamos *influyentes*.

OBSERVACIONES ATÍPICAS Y OBSERVACIONES INFLUYENTES EN REGRESIÓN

Una **observación atípica** es aquélla que queda separada de las restantes observaciones.

Una observación es **influyente** con relación a un cálculo estadístico si al eliminarla cambia el resultado del cálculo. En regresión mínimo-cuadrática, las observaciones atípicas en la dirección del eje de las abscisas son, en general, observaciones influyentes.

Los niños 18 y 19 son ambos observaciones atípicas de la figura 2.17. El niño 18 es una observación atípica en la dirección del eje de las abscisas y es también una observación influyente para la recta de regresión mínimo-cuadrática. El niño 19, en cambio, es una observación atípica en la dirección del eje de las ordenadas. Tiene menos influencia en la posición de la recta de regresión porque hay muchos puntos con valores de x similares que retienen la recta por debajo de la observación atípica. Los puntos influyentes suelen tener residuos pequeños, ya que tiran de la recta hacia su posición. Si sólo te fijas en los residuos, pasarás por alto los puntos influyentes. Las observaciones influyentes pueden modificar en gran manera la interpretación de unos datos.

EJEMPLO 2.13. Una observación influyente

La fuerte influencia del niño 18 hace que la recta de regresión de la puntuación Gesell con relación a la primera palabra sea engañosa. Los datos originales tienen $r^2 = 0{,}41$. Es decir, un 41% de la variación total observada en la prueba Gesell

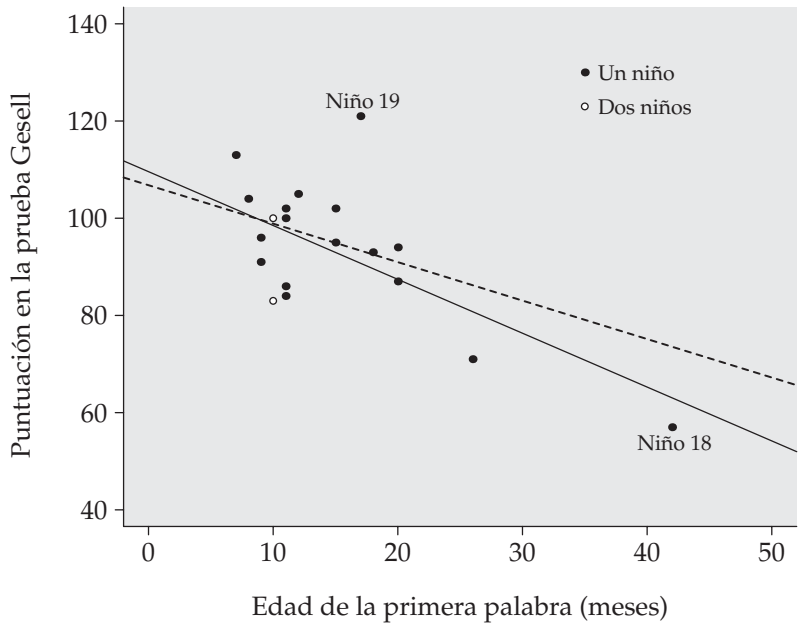


Figura 2.17. Dos rectas de regresión mínimo-cuadráticas de las puntuaciones Gesell en relación con la edad de la primera palabra. La recta de trazo continuo se ha calculado a partir de todos los datos. La de trazo discontinuo se ha calculado excluyendo al niño 18. El niño 18 es una observación influyente, ya que cuando se elimina este punto, la posición de la recta cambia.

se puede explicar a partir de la edad a la que los niños empiezan a hablar. Esta relación es suficientemente fuerte para que sea de interés para los padres. Pero si dejamos fuera al niño 18, r^2 cae al 11%. La fuerza aparente de la asociación se debía en gran medida a una sola observación influyente.

¿Qué debe hacer un investigador? Debe decidir si el desarrollo del niño 18 fue tan lento que no se debería permitir que influyera en el resultado del análisis. Si excluye al niño 18, desaparece en gran parte la evidencia de la relación entre la edad a la que un niño empieza a hablar y su posterior puntuación en la prueba. Si mantiene esta observación, necesitará datos adicionales de niños que hayan empezado a hablar tardíamente, de manera que el análisis no dependa tanto de una sola observación. ■

APLICA TUS CONOCIMIENTOS

2.36. Consumo de gasolina y velocidad. El ejercicio 2.6 proporciona datos sobre el consumo de gasolina y de un automóvil a distintas velocidades x. El consumo de carburante se ha medido en litros de gasolina por 100 kilómetros y la velocidad en kilómetros por hora. Con la ayuda de un programa estadístico hemos obtenido la recta de regresión mínimo-cuadrática y también los residuos. La recta de regresión es

$$\hat{y} = 11{,}058 - 0{,}01466x$$

Los residuos, en el mismo orden que las observaciones, son

10,09	2,24	−0,62	−2,47	−3,33	−4,28	−3,73	−2,94
−2,17	−1,32	−0,42	0,57	1,64	2,76	3,97	

(a) Dibuja un diagrama de dispersión con las observaciones y traza la recta de regresión en tu diagrama.

(b) ¿Utilizarías la recta de regresión para predecir y a partir de x? Justifica tu respuesta.

(c) Comprueba que la suma de los residuos es 0 (o muy cercana a 0, teniendo en cuenta los errores de redondeo).

(d) Dibuja un diagrama de residuos con relación a los valores de x. Traza una recta horizontal a la altura del valor 0 del eje de las ordenadas. Comprueba que la distribución de los residuos a lo largo de esta recta es similar a la distribución de los puntos a lo largo de la recta de regresión del diagrama de dispersión en (a).

2.37. ¿Cuántas calorías? La tabla 2.5 proporciona datos sobre el contenido real en calorías de diez alimentos y la media de los contenidos estimados por un numeroso grupo de personas. El ejercicio 2.23 explora la influencia de dos observaciones atípicas sobre la correlación.

(a) Dibuja un diagrama de dispersión adecuado para predecir la estimación de las calorías a partir de los valores reales. Señala los puntos correspondientes a los espaguetis y a los pasteles en tu diagrama. Estos dos puntos quedan fuera de la relación lineal de los ocho puntos restantes.

(b) Utiliza tu calculadora para hallar la recta de regresión de las calorías estimadas con relación a las calorías reales. Hazlo dos veces, primero, con todos los puntos y luego, dejando fuera los espaguetis y los pasteles.

(c) Dibuja las dos rectas de regresión en tu diagrama (una de trazo continuo y la otra con trazo discontinuo). Los espaguetis y los pasteles, tomados conjuntamente, ¿son observaciones influyentes? Justifica tu respuesta.

2.38. ¿Influyentes o no? Hemos visto que el niño 18 de los datos Gesell de la tabla 2.7 es una observación influyente. Ahora vamos a examinar el efecto del niño 19, que también es una observación atípica en la figura 2.14.

(a) Halla la recta de regresión mínimo-cuadrática de la puntuación en la prueba Gesell respecto a la edad a la cual un niño empieza a hablar, dejando fuera al niño 19. El ejemplo 2.12 da la recta de regresión con todos los niños. Dibuja ambas rectas en el mismo gráfico (no es necesario que lo hagas sobre un diagrama de dispersión; tan sólo dibuja las rectas). ¿Calificarías al niño 19 como muy influyente? ¿Por qué?

(b) La exclusión del niño 19, ¿qué efecto tiene sobre el valor r^2 de esta regresión? Explica por qué cambia r^2 al excluir al niño 19.

RESUMEN DE LA SECCIÓN 2.4

Una **recta de regresión** es una recta que describe cómo cambia una variable respuesta y al cambiar una variable explicativa x.

El método más común para ajustar una recta en un diagrama de dispersión es el método de mínimos cuadrados. La **recta de regresión mínimo-cuadrática** es la recta de la ecuación $\hat{y} = a + bx$ que minimiza la suma de cuadrados de las distancias verticales de los valores observados de y a la recta de regresión.

Puedes utilizar una recta de regresión para **predecir** el valor de y a partir de cualquier valor de x, sustituyendo esta x en la ecuación de la recta.

La **pendiente** b de una recta de regresión $\hat{y} = a + bx$ indica el cambio de la variable respuesta predicha $\hat{y}$ a lo largo de la recta de regresión, al cambiar la variable explicativa x. En concreto, b es el cambio de $\hat{y}$ al aumentar x en una unidad.

La **ordenada en el origen** a de una recta de regresión $\hat{y} = a + bx$ es la respuesta predicha $\hat{y}$ cuando la variable explicativa es $x = 0$. Esta predicción no tiene significado estadístico a no ser que x pueda tomar valores cercanos a 0.

La recta de regresión mínimo-cuadrática de y con relación a x es la recta de pendiente $r\frac{s_y}{s_x}$ y ordenada en el origen $a = \bar{y} - b\bar{x}$. Esta recta siempre pasa por el punto $(\bar{x}, \bar{y})$.

La **correlación y la regresión** están íntimamente relacionadas. Cuando las variables x e y se miden en unidades estandarizadas, la correlación r es la pendiente de la recta de regresión mínimo-cuadrática. El cuadrado de la correlación r^2 es la proporción de la variación de la variable respuesta explicada por la regresión mínimo-cuadrática.

Puedes examinar el ajuste de una recta de regresión estudiando los **residuos**, que son las diferencias entre los valores observados y los valores predichos de y.

Vigila los puntos atípicos con residuos anormalmente grandes, y también las distribuciones no lineales y desiguales de los residuos.

Fíjate también en las **observaciones influyentes**, que son los puntos aislados que cambian de forma sustancial la posición de la recta de regresión. Las observaciones influyentes suelen ser observaciones atípicas en la dirección de las abscisas y no tienen por qué tener residuos grandes.

EJERCICIOS DE LA SECCIÓN 2.4

2.39. Repaso sobre relación lineal. Antonio guarda sus ahorros en un colchón. Empezó con 500 € que le dio su madre y cada año fue añadiendo 100 €. Sus ahorros totales y después de x años vienen dados por la ecuación

$$y = 500 + 100x$$

(a) Representa gráficamente esta ecuación. (Escoge dos valores de x, tales como 0 y 10. Calcula los valores correspondientes de y a partir de la ecuación. Dibuja estos dos puntos en el gráfico y dibuja la recta uniéndolos.)

(b) Después de 20 años, ¿cuánto tendrá Antonio en su colchón?

(c) Si Antonio hubiera añadido cada año 200 € a sus 500 € iniciales, en vez de 100, ¿cuál sería la ecuación que describiría sus ahorros después de x años?

2.40. Repaso sobre relación lineal. En el periodo posterior a su nacimiento, una rata blanca macho gana exactamente 40 gramos (g) por semana. (Esta rata es extrañamente regular en su crecimiento, pero un crecimiento de 40 g por semana es un valor razonable.)

(a) Si la rata pesaba 100 gramos al nacer, da una ecuación para predecir su peso después de x semanas. ¿Cuál es la pendiente de esta recta?

(b) Dibuja un gráfico de esta recta para valores de x entre el nacimiento y las 10 semanas de edad.

(c) ¿Utilizarías esta recta para predecir el peso de la rata a los 2 años? Haz la predicción y medita sobre si el resultado es razonable.

2.41. Coeficiente de inteligencia y notas escolares. La figura 2.5 muestra las notas escolares medias y los coeficientes de inteligencia de 78 estudiantes de primero de bachillerato. La media y la desviación típica de los coeficientes de inteligencia son

$$\bar{x} = 108{,}9 \qquad\qquad s_x = 13{,}17$$

Para las notas medias escolares

$$\bar{y} = 7{,}447 \qquad\qquad s_y = 2{,}10$$

La correlación entre los coeficientes de inteligencia y las notas escolares medias es $r = 0{,}6337$.

(a) Halla la ecuación de la recta de regresión mínimo-cuadrática que permita predecir las notas escolares a partir de los coeficientes de inteligencia.

(b) ¿Qué porcentaje de la variación observada en las notas escolares se puede explicar por la relación lineal entre las notas escolares y los coeficientes de inteligencia?

(c) Un estudiante tiene un coeficiente de inteligencia de 103 y una nota media escolar de sólo 0,53. ¿Cuál es la predicción de la nota media escolar de un estudiante con un coeficiente de inteligencia de 103? ¿Cuál es el valor residual de este estudiante?

2.42. Llévame a ver un partido de baloncesto. ¿Qué relación existe entre el precio de los bocadillos de salchicha y el de los refrescos de cola en los estadios de baloncesto de EE UU? He aquí algunos datos:[19]

Estadio	Bocadillo	Refrescos de cola	Estadio	Bocadillo	Refrescos de cola	Estadio	Bocadillo	Refrescos de cola
Angels	2,50	1,75	Giants	2,75	2,17	Rangers	2,00	2,00
Astros	2,00	2,00	Indians	2,00	2,00	Red Sox	2,25	2,29
Braves	2,50	1,79	Marlins	2,25	1,80	Rockies	2,25	2,25
Brewers	2,00	2,00	Mets	2,50	2,50	Royals	1,75	1,99
Cardinals	3,50	2,00	Padres	1,75	2,25	Tigers	2,00	2,00
Dodgers	2,75	2,00	Phillies	2,75	2,20	Twins	2,50	2,22
Expos	1,75	2,00	Pirates	1,75	1,75	White Sox	2,00	2,00

(a) Dibuja un diagrama de dispersión que sea adecuado para predecir el precio del refresco de cola a partir del precio del bocadillo. Describe la relación que observas. ¿Hay observaciones atípicas?

(b) Halla la correlación entre el precio de los bocadillos y el precio de los refrescos de cola. ¿Qué porcentaje de la variación del precio del refresco se explica a partir de la relación lineal?

[19] Apareció en *Philadelphia City Paper*, 23-29 de mayo de 1997.

(c) Halla la ecuación de la recta de regresión mínimo-cuadrática para predecir el precio del refresco a partir del precio del bocadillo. Dibuja la recta en tu diagrama de dispersión. A partir de tus resultados en (b), explica por qué no es sorprendente que la recta sea casi horizontal (pendiente próxima a cero).

(d) Señala la observación que potencialmente es más influyente. ¿A qué estadio corresponde? Halla la recta de regresión mínimo-cuadrática sin esta observación y dibújala en tu diagrama de dispersión. Esta observación, ¿es realmente una observación influyente?

2.43. Análisis de agua. Las empresas suministradoras de agua la analizan regularmente para detectar la posible presencia de contaminantes. La determinación de éstos se hace de forma indirecta, por ejemplo, por colorimetría, que consiste en añadir un reactivo que da color al reaccionar con el contaminante a determinar. Posteriormente se hace pasar un haz de luz por la solución coloreada y se determina su "absorbancia". Para calibrar este método de análisis, los laboratorios disponen de patrones con concentraciones conocidas del producto a determinar. Suele existir una relación lineal entre la concentración del producto a determinar y su absorbancia una vez ha tenido lugar la reacción anteriormente comentada. He aquí una serie de datos sobre la absorbancia y la concentración de nitratos. Los nitratos se expresan en miligramos por litro de agua.[20]

Nitratos	50	50	100	200	400	800	1.200	1.600	2.000	2.000
Absorbancia	7,0	7,5	12,8	24,0	47,0	93,0	138,0	183,0	230,0	226,0

(a) Teóricamente estos datos deben mantener una relación lineal. Si el coeficiente de correlación no es al menos 0,997, hay que suponer que algo fue mal y hay que repetir el proceso de calibración. Representa gráficamente los datos y halla su correlación. ¿Se debe repetir la calibración?

(b) Determina la ecuación de la recta de regresión mínimo-cuadrática que nos permita predecir la absorbancia a partir de la concentración de nitratos. Si el laboratorio analiza un patrón con 500 miligramos de nitratos por litro, ¿qué valor de absorbancia obtendrás? Basándote en tu dibujo y en la correlación, ¿crees que la estimación de la absorbancia será muy exacta?

[20]De una conferencia de Charles Knauf del Environmental Health Laboratory, Monroe County (New York).

2.44. Crecimiento de una niña. Los padres de Sara están preocupados porque creen que es baja para su edad. Su médico ha ido registrando las siguientes alturas de Sara:

Edad (meses)	36	48	51	54	57	60
Altura (cm)	86	90	91	93	94	95

(a) Dibuja un diagrama de dispersión con estos datos. Fíjate en la fuerte relación lineal.

(b) Usando la calculadora, halla la ecuación de la recta de regresión mínimo-cuadrática de la altura en relación con la edad.

(c) Predice la altura de Sara a los 40 y a los 60 meses. Utiliza tus resultados para dibujar la recta de regresión en tu diagrama de dispersión.

(d) ¿Cuál es el ritmo de crecimiento de Sara en centímetros por mes? Las niñas con crecimiento normal ganan unos 6 cm de altura entre los 4 (48 meses) y los 5 años (60 meses). En este último caso, ¿qué valor toma el ritmo de crecimiento expresado en centímetros por mes? ¿Crece Sara más despacio de lo normal?

2.45. Invertir en y fuera de EE UU. Unos inversores quieren saber qué relación existe entre los rendimientos de las inversiones en EE UU y las inversiones fuera de EE UU. La tabla 2.8 proporciona datos sobre los rendimientos totales de los valores bursátiles en EE UU y fuera de EE UU, durante un periodo de 26 años. (Los rendimientos totales se calculan a partir de los cambios de cotización más los dividendos pagados, expresados en dólares. Ambos rendimientos son medias de muchos valores individuales.[21])

(a) Haz un diagrama de dispersión adecuado para predecir los rendimientos de los valores bursátiles fuera de EE UU a partir de los rendimientos en EE UU.

(b) Halla la correlación y r^2. Describe con palabras la relación entre los rendimientos en y fuera de EE UU. Utiliza r y r^2 para hacer más precisa tu descripción.

(c) Halla la recta de regresión mínimo-cuadrática de los rendimientos fuera de EE UU en función de los rendimientos en EE UU. Traza la recta en el diagrama de dispersión.

(d) En 1997, el rendimiento de las acciones en EE UU fue del 33,4%. Utiliza la recta de regresión para predecir el rendimiento de las acciones fuera de EE UU. El

[21] Los rendimientos en EE UU corresponden al índice de 500 valores Standard & Poor's. Los rendimientos fuera de EE UU corresponden a Morgan Stanley Europe, Australasia, índice Far East (EAFE).

Tabla 2.8. Rendimientos anuales en y fuera de EE UU.

Año	Rendimiento fuera de EE UU	Rendimiento en EE UU	Año	Rendimiento fuera de EE UU	Rendimiento en EE UU
1971	29,6	14,6	1985	56,2	31,6
1972	36,3	18,9	1986	69,4	18,6
1973	−14,9	−14,8	1987	24,6	5,1
1974	−23,2	−26,4	1988	28,5	16,8
1975	35,4	37,2	1989	10,6	31,5
1976	2,5	23,6	1990	−23,0	−3,1
1977	18,1	−7,4	1991	12,8	30,4
1978	32,6	6,4	1992	−12,1	7,6
1979	4,8	18,2	1993	32,9	10,1
1980	22,6	32,3	1994	6,2	1,3
1981	−2,3	−5,0	1995	11,2	37,6
1982	−1,9	21,5	1996	6,4	23,0
1983	23,7	22,4	1997	2,1	33,4
1984	7,4	6,1			

rendimiento fuera de EE UU fue del 2,1%. ¿Estás seguro de que las predicciones basadas en la recta de regresión serán suficientemente precisas? ¿Por qué?

(e) Señala el punto que tenga el mayor residuo (positivo o negativo). ¿Qué año es? ¿Parece probable que existan puntos que sean observaciones muy influyentes?

2.46. Representa gráficamente tus datos, ¡siempre! La tabla 2.9 presenta cuatro conjuntos de datos preparados por el estadístico Frank Anscombe para ilustrar los peligros de hacer cálculos sin antes representar los datos.[22]

(a) Sin dibujar un diagrama de dispersión, halla la correlación y la recta de regresión mínimo-cuadrática para los cuatro grupos de datos. ¿Qué observas? Utiliza la recta de regresión para predecir y cuando $x = 10$.

(b) Dibuja un diagrama de dispersión para cada uno de los conjuntos de datos con las rectas de regresión correspondientes.

(c) ¿En cuál o cuáles de los cuatro casos utilizarías la recta de regresión para describir la dependencia de y en relación a x? Justifica tu respuesta en cada caso.

2.47. ¿Cuál es mi nota? En el curso de economía del profesor Marcet, la correlación entre la calificación acumulada por los estudiantes antes de examinarse y la

[22]Frank J. Anscombe, "Graphs in statistical analysis", *The American Statistician*, 27, 1973, págs. 17-21.

Tabla 2.9. Correlaciones y regresiones con cuatro conjuntos de datos.

Conjunto de datos A											
x	10	8	13	9	11	14	6	4	12	7	5
y	8,04	6,95	7,58	8,81	8,33	9,96	7,24	4,26	10,84	4,82	5,68
Conjunto de datos B											
x	10	8	13	9	11	14	6	4	12	7	5
y	9,14	8,14	8,74	8,77	9,26	8,10	6,13	3,10	9,13	7,26	4,74
Conjunto de datos C											
x	10	8	13	9	11	14	6	4	12	7	5
y	7,46	6,77	12,74	7,11	7,81	8,84	6,08	5,39	8,15	6,42	5,73
Conjunto de datos D											
x	8	8	8	8	8	8	8	8	8	8	19
y	6,58	5,76	7,71	8,84	8,47	7,04	5,25	5,56	7,91	6,89	12,50

calificación del examen final es $r = 0{,}6$. La media de las calificaciones acumuladas antes del examen final es 280 y la desviación típica, 30, mientras que la media de las notas del examen final es 75 y la desviación típica, 8. Al profesor Marcet se le extravió el examen final de Julia, pero sabe que su calificación acumulada es 300; por ello, decide predecir la calificación del examen final de Julia a partir de las calificaciones acumuladas por ésta.

(a) ¿Cuál es la pendiente de la recta de regresión mínimo-cuadrática de la calificación del examen final con relación a la calificación acumulada antes del examen final de ese curso? ¿Cuál es la ordenada en el origen?

(b) Utiliza la recta de regresión para predecir la calificación del examen final de Julia.

(c) Julia no cree que el método del profesor Marcet para predecir la calificación de su examen sea muy bueno. Calcula r^2 para argumentar que la calificación real del examen final de Julia podía haber sido mucho más alta (o mucho más baja) que el valor predicho.

2.48. Predicción sin sentido. Utiliza la regresión mínimo-cuadrática con los datos del ejercicio 2.44 para predecir la altura de Sara a los 40 años (480 meses).

La predicción es absurdamente grande. No es razonable utilizar datos con valores entre 36 y 60 meses para predecir la altura a los 480 meses.

2.49. Invertir en y fuera de EE UU. El ejercicio 2.45 examinó la relación entre los rendimientos de los valores bursátiles en EE UU y fuera de EE UU. Los inversores también quieren saber cuáles son los rendimientos típicos y cuál es la variabilidad entre años (en términos financieros llamada *volatilidad*). La regresión y la correlación no dan información sobre el centro y la dispersión.

(a) Halla los cinco números resumen de los rendimientos tanto en EE UU como fuera de ellos, y dibuja los correspondientes diagramas de caja en un mismo gráfico para comparar las dos distribuciones.

(b) Durante este periodo, ¿los rendimientos fueron mayores en EE UU o fuera? Justifica tu respuesta.

(c) En este periodo, ¿los rendimientos fueron más volátiles (más variables) en o fuera de EE UU? Razona tu respuesta.

2.50. Un estudio sobre la relación entre la asistencia a clase y las calificaciones de los estudiantes de primer curso de la Universidad Pompeu Fabra mostró que, en general, los alumnos que asisten a un mayor porcentaje de clases obtienen mejores calificaciones. Concretamente, la asistencia a clase explicaba el 16% de la variación en la media de las calificaciones obtenidas. ¿Cuál es el valor de la correlación entre el porcentaje de asistencia a clase y la media de las calificaciones obtenidas?

2.51. ¿Suspenderé el examen final? Creemos que los estudiantes que obtienen buenas calificaciones en el examen parcial de un determinado curso de estadística, también obtendrán buenas calificaciones en el examen final. El profesor Smith analizó las calificaciones de 346 estudiantes que se matricularon en ese curso de estadística durante un periodo de 10 años.[23] La recta de regresión mínimo-cuadrática para la predicción de la calificación del examen final a partir de la calificación del examen parcial era $\hat{y} = 46{,}6 + 0{,}41x$.

La calificación del examen parcial de María está 10 puntos por encima de la media de los estudiantes analizados. ¿Cuál habría sido tu predicción sobre el número de puntos por encima de la media del examen final de María? (Sugerencia: utiliza el hecho de que la recta de regresión mínimo-cuadrática pasa por el punto $(\bar{x}, \bar{y})$ y el hecho de que la calificación del examen parcial de María es $\bar{x} + 10$. Estamos ante un ejemplo de un tipo de fenómeno que dio a la "regresión"

[23]Gary Smith, "Do statistics test scores regress toward the mean?", *Chance*, 10, nº 4, 1997, págs. 42-45.

su nombre: los estudiantes que obtienen buenas calificaciones en el examen parcial, obtienen como media calificaciones no tan buenas en el examen final, pero todavía por encima de la media.)

2.52. Predicción del número de estudiantes matriculados. A la Facultad de Matemáticas de una gran universidad le gustaría utilizar el número de estudiantes x recién llegados a la universidad, para predecir el número de estudiantes y que se matriculará en el curso de Introducción al Análisis Matemático del semestre de otoño. He aquí los datos de los últimos años.[24]

Año	1991	1992	1993	1994	1995	1996	1997	1998
x	4.595	4.827	4.427	4.258	3.995	4.330	4.265	4.351
y	7.364	7.547	7.099	6.894	6.572	7.156	7.232	7.450

Un programa estadístico halla la correlación $r = 0{,}8333$ y la recta de regresión mínimo-cuadrática

$$\hat{y} = 2.492{,}69 + 1{,}0663x$$

El programa estadístico también da la tabla de residuos:

Año	1991	1992	1993	1994	1995	1996	1997	1998
Residuo	−28,44	−92,83	−114,30	−139,09	−180,65	46,13	191,44	317,74

(a) Dibuja un diagrama de dispersión con la recta de regresión. Ésta no da una buena predicción. ¿Qué porcentaje de la variación en las matrículas para el curso de matemáticas se explica a partir de la relación entre éstas y el recuento de recién ingresados en la universidad?

(b) Comprueba que los residuos suman cero (o aproximadamente cero si se tiene en cuenta el error de redondeo).

(c) Los diagramas de residuos son a menudo reveladores. Dibuja los residuos con relación al año. Una de las facultades de la universidad ha cambiado recientemente su programa docente. Ahora exige a sus estudiantes que tomen un curso de matemáticas. ¿Cómo muestra el diagrama de residuos este cambio? ¿En qué años tuvo lugar dicho cambio?

[24] Datos de Peter Cook, Purdue University.

2.5 Precauciones con la correlación y la regresión

La correlación y la regresión son dos potentes instrumentos para describir la relación entre dos variables. Cuando los utilices tienes que recordar sus limitaciones, empezando por el hecho de que **la correlación y la regresión sólo describen relaciones lineales**. Recuerda también que tanto **la correlación *r* como la recta de regresión mínimo-cuadrática pueden estar muy influenciadas por unas pocas observaciones extremas**. Una observación influyente o un error en la transcripción de un dato puede cambiar mucho sus valores. Por consiguiente, representa siempre tus datos antes de interpretar una correlación o una regresión. A continuación, vamos a ver otras precauciones que conviene tomar cuando se aplica la correlación y la regresión, o se leen trabajos en los que se hayan utilizado.

2.5.1 Extrapolación

Supón que tienes datos sobre el crecimiento de los niños entre los 3 y los 8 años de edad, y hallas una fuerte relación lineal entre la edad x y la altura y. Si ajustas una recta de regresión a estos datos y la utilizas para predecir la altura a la edad de 25 años, acabarás pronosticando que el niño tendrá una altura de 2,43 metros, cuando en realidad el crecimiento disminuye a partir de cierta edad y se detiene al llegar a la madurez. Por tanto, extrapolar la relación lineal más allá de la madurez no tiene ningún sentido. Pocas relaciones son lineales para todos los valores de x. Por consiguiente, no extiendas la predicción más allá del intervalo de valores de x para los que tienes datos.

EXTRAPOLACIÓN

La **extrapolación** es la utilización de una recta de regresión para la predicción fuera del intervalo de valores de la variable explicativa x que utilizaste para obtener la recta. Este tipo de predicciones no son fiables.

APLICA TUS CONOCIMIENTOS

2.53. Disminución de la población rural. En Estados Unidos la población rural ha ido disminuyendo de forma constante a lo largo de este siglo. He aquí datos sobre esta población (expresado en millones de personas) desde 1935 hasta 1980.

Año	1935	1940	1945	1950	1955	1960	1965	1970	1975	1980
Población rural	32,1	30,5	24,4	23,0	19,1	15,6	12,4	9,7	8,9	7,2

(a) Dibuja un diagrama de dispersión con estos datos. Halla la recta de regresión que exprese la relación entre la población rural en EE UU y el año.

(b) De acuerdo con la recta de regresión, ¿cuánto disminuye, como media, la población rural cada año durante este periodo? ¿Qué porcentaje de la variación observada se explica con la recta de regresión?

(c) Utiliza la recta de regresión para predecir la población rural en el año 1990. ¿Te parece un valor razonable? ¿Por qué?

2.5.2 Utilización de medias

En muchos estudios de correlación y de regresión se trabaja con medias o con diversas medidas que combinan la información de muchos individuos. Tienes que ser muy cuidadoso y resistir la tentación de aplicar los resultados de este tipo de estudios a individuos. En la figura 2.2, hemos visto una asociación muy fuerte entre la temperatura exterior y el consumo de gas de los Sánchez. Cada punto del diagrama de dispersión representa un mes. Los grados-día y el consumo de gas son medias de todos los días de un mes. Los datos de cada uno de los días hubieran mostrado una mayor dispersión con relación a la recta de regresión y una menor correlación. Calcular medias de todos los días de un mes permite reducir las variaciones diarias debidas, por ejemplo, a que un día se deja una puerta abierta, o al mayor consumo de agua caliente un día que hay invitados, etc. **Las correlaciones basadas en medias habitualmente son demasiado altas cuando se aplican a observaciones individuales**. Por ello, en todo estudio estadístico es importante fijarse exactamente en cómo se han medido las variables.

APLICA TUS CONOCIMIENTOS

2.54. Índices del mercado de valores. El índice bursátil Standard & Poor's es la media de los valores de 500 acciones. Existe una correlación moderadamente fuerte (aproximadamente $r = 0{,}6$) entre la variación de este índice en enero y su variación en todo el año. De todas formas, si nos fijáramos en las variaciones individuales de 500 acciones encontraríamos una correlación bastante distinta. ¿Esta correlación sería mayor o menor que la obtenida con las medias? ¿Por qué?

2.5.3 Variables latentes

En nuestro estudio sobre la correlación y la regresión lineal sólo nos fijamos en dos variables a la vez. A menudo, la relación entre dos variables está muy influida por otras. Métodos estadísticos más avanzados permiten el estudio de muchas variables simultáneamente, por lo que podemos tenerlas en cuenta. A veces, la relación entre dos variables se encuentra muy influida por otras variables que no medimos o de las que ni siquiera sospechábamos su existencia. A estas últimas variables las llamamos *variables latentes*.

VARIABLE LATENTE

Una **variable latente** es una variable que no se incluye entre las variables estudiadas y que, sin embargo, tiene un importante efecto sobre la relación que existe entre ellas.

Una variable latente puede enmascarar una relación entre x e y, o puede sugerir una falsa relación entre dos variables. Veamos algunos ejemplos de cada uno de estos efectos.

EJEMPLO 2.14. ¿Discriminación de género en los tratamientos médicos?

Unos estudios muestran que es más probable que se les haga pruebas específicas y tratamientos contundentes a los hombres que sufren problemas cardíacos, como por ejemplo un *bypass*, que a las mujeres con dolencias similares. Esta relación entre tratamiento recibido y sexo, ¿se debe a una discriminación de género?

Puede ser que no. Los hombres y las mujeres sufren problemas en el corazón a edades diferentes —en general las mujeres son de 10 a 15 años mayores que los hombres—. Los tratamientos contundentes son más peligrosos en el caso de pacientes de más edad; por tanto, los médicos puede ser que duden al recomendarlos a este tipo de pacientes. La relación entre el sexo y las decisiones de los médicos se podría explicar a partir de variables latentes —la edad y la condición general del paciente—. Tal como comentaba el autor de un estudio sobre este tema: "Cuando hombres y mujeres están en condiciones similares, siendo la única diferencia el sexo, los tratamientos también son similares".[25] ■

[25]Daniel Mark, "Age, not bias, may explain differences in treatment", *New York Times*, 26 de abril de 1994.

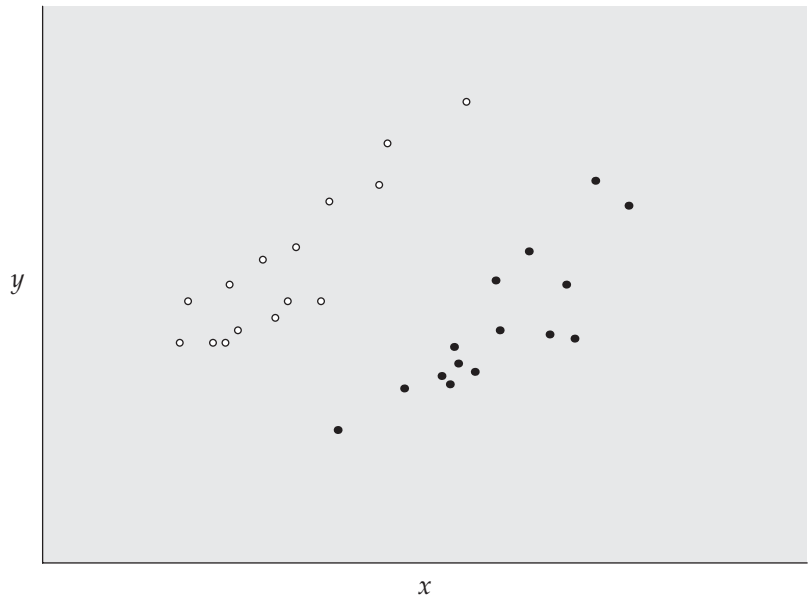


Figura 2.18. Las variables de este diagrama de dispersión tienen una correlación muy pequeña. Sin embargo, dentro de cada grupo la correlación entre las dos variables es fuerte.

EJEMPLO 2.15. Condiciones de vivienda inadecuadas

Un estudio sobre la vivienda en la ciudad de Hull, Inglaterra, midió algunas variables en cada uno de los barrios de la ciudad. La figura 2.18 es una simplificación de los hallazgos del estudio. La variable x es una medida de la densidad excesiva de población. La variable y es la proporción de viviendas que no tienen sanitarios. Debido a que x e y miden condiciones inadecuadas de las viviendas, esperamos encontrar una correlación alta entre ambas variables. En realidad, la correlación fue sólo $r = 0{,}08$. ¿Cómo se puede explicar esto?

La figura 2.18 muestra que hay dos tipos de barrios. Los del grupo inferior de la figura tienen una elevada proporción de viviendas de iniciativa pública. Estos barrios tienen valores elevados de x pero valores pequeños de y, ya que las viviendas públicas siempre tienen sanitarios. En cambio, los del grupo superior de la figura carecen de viviendas públicas y tienen valores elevados tanto de x como de y. Dentro de cada tipo de barrio, existe una fuerte asociación positiva entre x e y. En la figura 2.18, $r = 0{,}85$ y $r = 0{,}91$ dentro de cada grupo. Sin embargo,

debido a que a valores similares de x les corresponden valores bastante distintos de y en los dos grupos, es difícil predecir y sólo a partir de x. Cuando analizamos todos los barrios conjuntamente, ignorando por tanto la variable latente —la proporción de vivienda de promoción pública— se enmascara la verdadera naturaleza de la relación entre x e y.[26] ■

APLICA TUS CONOCIMIENTOS

2.55. Televisión y notas escolares. Los niños que pasan muchas horas delante del televisor obtienen, como media, peores notas en la escuela que los niños que pasan menos horas. Sugiere variables latentes que puedan afectar a la relación entre estas variables debido a que influyen tanto sobre el hecho de pasar muchas horas delante de la televisión como sobre las notas escolares.

2.56. Educación e ingresos. Existe una fuerte correlación positiva entre los años de formación y los ingresos de los economistas empleados en empresas. En especial, los economistas doctorados ganan más que los que sólo son licenciados. Hay también una fuerte correlación positiva entre los años de formación y los ingresos de los economistas empleados en las universidades. Sin embargo, cuando se considera conjuntamente a todos los economistas, existe una correlación *negativa* entre la educación y los ingresos. La explicación es que las empresas pagan salarios altos y emplean principalmente a economistas que son sólo licenciados, mientras que las universidades pagan salarios bajos y emplean principalmente a economistas con doctorado. Haz un diagrama de dispersión con dos tipos de observaciones (de la empresa y de la universidad), para ilustrar cómo se puede tener al mismo tiempo una correlación positiva fuerte dentro de cada grupo y una correlación conjunta negativa. (Consejo: empieza estudiando la figura 2.18.)

2.5.4 Asociación no implica causalidad

Cuando estudiamos la relación entre dos variables, a menudo queremos mostrar qué cambios en la variable explicativa *causan* cambios en la variable respuesta. Que exista una fuerte asociación entre dos variables no es suficiente para sacar conclusiones sobre las relaciones causa-efecto. A veces, una asociación observada refleja una relación causa-efecto. La familia Sánchez consume más gas en los

[26]M. Goldstein, "Preliminary inspection of multivariate data", *The American Statistician*, 36, 1982, págs. 358-362.

meses más fríos, ya que cuando hace más frío se necesita más gas para mantener la casa caliente. En otros casos, una asociación se explica por variables latentes y la conclusión de que x causa y es errónea o no está demostrada.

EJEMPLO 2.16. Los televisores, ¿alargan la vida?

Considera el número de televisores por persona x y la esperanza media de vida y de los países del mundo. Existe una correlación positiva fuerte: los países con muchos televisores tienen esperanzas de vida mayores.

El hecho básico que explica una relación de causalidad es que cuando cambiamos el valor de x, cambia el valor de y. ¿Podemos aumentar la esperanza de vida de la gente de Ruanda enviándoles televisores? Evidentemente no. Aunque los países ricos tienen más televisores que los países pobres, de hecho tienen una esperanza de vida mayor debido a una alimentación mejor, a una mejor calidad del agua y una mejor asistencia médica. No existe ninguna relación de causa-efecto entre el número de televisores y la esperanza de vida. ■

Correlaciones como la del ejemplo 2.16 se conocen a menudo como "correlaciones sin sentido". La correlación es real. Lo que no tiene sentido es la conclusión de que un cambio en el valor de una variable es la causa de un cambio en la otra. Una variable latente —como por ejemplo la riqueza nacional del ejemplo 2.16— que influye tanto en x como en y puede provocar una fuerte correlación aunque no exista ninguna conexión directa entre x e y.

EJEMPLO 2.17. Peligrosidad de las anestesias

El *National Halothane Study* fue un importante estudio sobre la peligrosidad de las anestesias utilizadas en cirugía. Datos de más de 850.000 operaciones realizadas en los 34 hospitales más importantes de EE UU mostraron las siguientes tasas de mortalidad para las cuatro anestesias más utilizadas:[27]

Anestesia	A	B	C	D
Tasa de mortalidad	1,7%	1,7%	3,4%	1,9%

[27] L. E. Moses y F. Mosteller, "Safety of anesthetics", en J. Tanur *et al*. (eds.), *Statistics: A Guide to the Unknown*, 3ª ed. Wadsworth, Belmont, Calif., 1989, págs. 15-24.

Existe una asociación clara entre la anestesia utilizada y la tasa de mortalidad de los pacientes. Parece ser que la anestesia C es peligrosa. De todas formas es claro que existen variables latentes como la edad, el estado de salud general del paciente o la importancia de la operación, que hay que tener en cuenta. De hecho, la anestesia C se utilizó con más frecuencia en operaciones importantes con pacientes de más edad y en un estado general de salud delicado. La tasa de mortalidad sería mayor para este tipo de pacientes independientemente de la anestesia utilizada. Después de considerar estas variables latentes y tener en cuenta su efecto, la relación aparente entre la anestesia y la tasa de mortalidad es mucho más débil. ■

Los ejemplos que acabamos de ver, y otros, sugieren que tenemos que tener precaución sobre la correlación, la regresión y en general sobre la asociación entre dos variables.

ASOCIACIÓN NO IMPLICA CAUSALIDAD

Una asociación entre una variable explicativa x y una variable respuesta y, incluso si es muy fuerte, no es por sí misma una evidencia suficiente de que cambios de x realmente causen cambios de y.

Experimentos

La mejor manera de obtener una buena evidencia de que x causa y es realizar un **experimento** en el que x tome distintos valores y las variables latentes se mantengan bajo control. En el capítulo 3 trataremos sobre los experimentos. Cuando no se pueden realizar experimentos, es difícil y controvertido hallar la explicación de una asociación observada. Muchas disputas en las que intervienen estadísticos hacen referencia a relaciones de causalidad que no se han podido demostrar mediante un experimento. ¿Fumar causa cáncer de pulmón? ¿Qué ocurre con los fumadores pasivos? ¿Vivir cerca de una línea de alta tensión causa leucemia? ¿Ha disminuido la diferencia de salarios entre los trabajadores con más formación y los que tienen menos? Todos estos temas forman parte de debates en medios de comunicación. Todos ellos hacen referencia a la relación entre variables. Y todos ellos tienen en común que tratan de esclarecer relaciones de causalidad en situaciones en las que interactúan muchas variables.

EJEMPLO 2.18. Fumar, ¿provoca cáncer de pulmón?

A pesar de las dificultades, a veces es posible establecer relaciones fuertes de causalidad sin necesidad de hacer experimentos. La evidencia de que fumar provoca cáncer de pulmón es lo más fuerte que puede ser una evidencia no experimental.

Los médicos han observado, durante mucho tiempo, que los enfermos de cáncer de pulmón eran fumadores. La comparación de fumadores con sujetos "similares" no-fumadores muestra una asociación muy fuerte entre el fumar y la muerte por cáncer de pulmón. Esta asociación, ¿se podría explicar mediante variables latentes? Podría existir, por ejemplo, un factor genético que predispusiera a la gente tanto a la adicción a la nicotina como al cáncer de pulmón? El fumar y el cáncer de pulmón podrían estar positivamente asociados incluso si el fumar no tuviera un efecto directo sobre los pulmones. ¿Cómo responder a estas preguntas? ■

Vamos a responder de forma general: ¿cómo podemos establecer una relación de causalidad sin hacer un experimento?

- *La asociación es fuerte*. La asociación entre fumar y el cáncer de pulmón es muy fuerte.
- *La asociación es consistente*. Muchos estudios en diferentes lugares y con diferente tipo de gente, relacionan el fumar con el cáncer de pulmón. Este hecho reduce las posibilidades de que una variable latente que actúa en unas condiciones o para un grupo de gente determinado explique la asociación.
- *Dosis mayores están asociadas a respuestas mayores*. La gente que fuma más cigarrillos por día o que fuma durante más tiempo padecen más a menudo cáncer de pulmón. La gente que deja de fumar reduce este riesgo.
- *La supuesta causa precede al efecto en el tiempo*. El cáncer se desarrolla después de años de fumar. El número de hombres que mueren de cáncer de pulmón crece a medida que este hábito es más común, el desfase es de unos 30 años. El cáncer de pulmón mata más hombres que ninguna otra forma de cáncer. El cáncer de pulmón era raro entre las mujeres hasta que éstas empezaron a fumar y ha ido creciendo con este hábito. Otra vez con un desfase de unos 30 años. Entre las mujeres, el cáncer de pulmón es ahora más importante como causa de muerte que el cáncer de mama.
- *La supuesta causa es plausible*. Experimentos con animales muestran que los alquitranes del humo de los cigarrillos causan cáncer.

Las autoridades sanitarias no dudan en absoluto al decir que fumar causa cáncer de pulmón. De hecho, en los países occidentales, "la causa evitable más importante de muerte y discapacidad es el tabaco".[28] La evidencia sobre esta relación causa-efecto es abrumadora —pero no es tan fuerte como sería la evidencia proporcionada por experimentos bien diseñados—.

APLICA TUS CONOCIMIENTOS

2.57. Los bomberos, ¿causan mayores incendios? Alguien afirma: "Existe una fuerte correlación positiva entre el número de bomberos que actúan en la extinción de un incendio y la importancia del daño que éste ocasiona. Por tanto, el hecho de enviar muchos bomberos sólo ocasiona más daños". Explica por qué este razonamiento es incorrecto.

2.58. ¿Cómo está tu autoestima? Las personas que tienen éxito tienden a estar satisfechas con ellas mismas. Es posible que ayudar a la gente para que se sienta satisfecha les pueda ayudar a tener más éxito en la escuela y en general en la vida. Aumentar la autoestima de los estudiantes fue durante un tiempo uno de los objetivos de muchas escuelas. ¿A qué se debe la asociación entre la autoestima y el éxito escolar? ¿Qué podemos decir aparte de que una autoestima alta es la causa de un mejor éxito escolar?

2.59. Los grandes hospitales, ¿son malos? Un estudio muestra que existe una correlación positiva entre el tamaño de un hospital (medido como número de camas x) y el número medio de días y que los enfermos permanecen en él. ¿Significa esto que se puede reducir la estancia en un hospital si se escogen hospitales pequeños? ¿Por qué?

RESUMEN DE LA SECCIÓN 2.5

La correlación y la regresión tienen que **interpretarse con precaución**. **Representa gráficamente** los datos para estar seguro de que la relación es al menos aproximadamente lineal y para detectar observaciones atípicas y observaciones influyentes.

[28] *The Health Consequences of Smoking: 1983*, U.S. Public Health Service, Washington, D.C., 1983.

Evita la **extrapolación**, que consiste en emplear una recta de regresión para predecir valores de la variable explicativa que quedan fuera del intervalo de valores a partir del cual se calculó la recta.

Recuerda que las **correlaciones basadas en medias** suelen ser demasiado altas cuando se aplican a los datos individuales.

Las **variables latentes** que no mediste pueden explicar la relación entre las variables que mediste. La correlación y la regresión pueden ser engañosas si ignoras variables latentes importantes.

Sobre todo, procura no concluir que existe una relación causa-efecto entre dos variables sólo porque están fuertemente asociadas. **Una correlación alta no implica causalidad**. La mejor evidencia de que una asociación se debe a la causalidad se obtiene mediante un **experimento** en el cual la variable explicativa se va modificando mientras se controlan las demás variables que pueden influir en la variable respuesta.

EJERCICIOS DE LA SECCIÓN 2.5

2.60. Para tener éxito en la universidad, ¿hay que estudiar matemáticas? He aquí un fragmento de un artículo que apareció en un periódico sobre un estudio llevado a cabo con 15.941 estudiantes de secundaria estadounidenses:

En EE UU los estudiantes universitarios pertenecientes a minorías raciales que escogieron en secundaria como asignaturas optativas álgebra y geometría se graduaron en la misma proporción que los hijos de anglosajones.

La relación entre las matemáticas estudiadas en secundaria y la graduación en la universidad es "algo mágico" dice el rector de una universidad, sugiriendo de alguna manera que "las matemáticas son la clave del éxito en la universidad".

Estos hallazgos, dice el rector, "justificarían considerar muy seriamente la posibilidad de llevar a cabo una política que asegure que todos los estudiantes de secundaria pasen por un curso de álgebra y geometría".[29]

¿Qué variables latentes podrían explicar la asociación entre pasar por diversos cursos de matemáticas y el éxito en la universidad? Explica por qué exigir

[29]De un artículo de Gannett News Service que apareció el 23 de abril de 1994 en el *Journal and Courier*, Lafayette, Indiana.

haber estudiado álgebra y geometría seguramente tendría poco efecto sobre los estudiantes universitarios que tienen éxito.

2.61. Comprensión de textos escritos y tamaño del pie. Un estudio con niños de 6 a 11 años que asisten a una escuela de primaria halla una fuerte correlación positiva entre el número de calzado x y la nota obtenida en una prueba de comprensión de textos escritos. ¿Qué explica esta correlación?

2.62. Los edulcorantes artificiales, ¿provocan un aumento de peso? La gente que utiliza edulcorantes artificiales en vez de azúcar tiende a tener más peso que la gente que toma azúcar. ¿Significa esto que los edulcorantes artificiales provocan un aumento de peso? Da una explicación más plausible para esta asociación.

2.63. Calificaciones de Lengua y Matemáticas. La tabla 2.1 proporciona datos sobre la educación en los diversos Estados de EE UU. La correlación en cada Estado entre la media de las calificaciones de Matemáticas y la media de las calificaciones de Lengua en la prueba SAT es $r = 0{,}970$.

(a) Halla r^2 y explica con palabras sencillas qué nos indica este número.

(b) Si calcularas la correlación entre las calificaciones de Matemáticas y las de Lengua en la prueba SAT de un gran número de estudiantes individuales, ¿crees que la correlación sería 0,97 o bastante distinta? Justifica tu respuesta.

2.64. El té, ¿beneficia a los ancianos? Un grupo de estudiantes universitarios cree que el té tiene efectos muy beneficiosos para la salud. Para verificarlo, los estudiantes decidieron hacer una serie de visitas semanales a una residencia de ancianos. En cada una de estas visitas los estudiantes servían té a los residentes. El personal que los atendía se percató de que al cabo de unos meses muchos de los residentes se mostraban más alegres y tenían un aspecto más saludable. Un sociólogo, algo escéptico, felicita a los estudiantes por sus buenas intenciones pero no acaba de creerse que el té ayudara a los ancianos. Identifica las variables explicativa y respuesta de este estudio. Explica qué variables latentes pueden explicar la asociación observada.

2.65. ¿Es interesante estudiar idiomas? Los miembros del seminario de idiomas de una escuela de secundaria creen que el estudio de una lengua extranjera mejora el dominio de la lengua propia de los estudiantes. De los archivos de la escuela, los investigadores obtienen las calificaciones de los exámenes de Lengua de los estudiantes de los últimos cursos. La media de las calificaciones de los

alumnos que estudiaron una lengua extranjera durante al menos dos años es mucho más alta que la de los alumnos que no la estudiaron. El director de la escuela argumenta que estos datos no constituyen una buena evidencia de que el estudio de lenguas extranjeras aumente el dominio de la lengua propia. Identifica las variables explicativa y respuesta de este estudio. Luego, explica qué variable latente anula la conclusión de que el estudio de lenguas mejora el dominio de la lengua propia.

2.66. Formación e ingresos. Existe una fuerte correlación positiva entre los años de escolarización x y los ingresos a lo largo de la vida y de los hombres en Europa. Una posible razón de esta asociación es causal: más educación conduce a empleos mejor pagados. De todas formas, variables latentes podrían explicar una parte de la correlación. Sugiere algunas variables latentes que explicarían por qué los hombres con más formación ganan más.

2.67. Las líneas de alta tensión, ¿provocan cáncer? Se ha sugerido que los campos electromagnéticos como los que se hallan junto a las líneas de alta tensión pueden causar leucemia en los niños. Estudios minuciosos sobre el tema no han hallado ninguna asociación entre la exposición a campos electromagnéticos y la leucemia infantil.[30]

Sugiere algunas variables latentes sobre las que quisieras información con el objetivo de investigar la afirmación de que vivir junto a una línea de alta tensión está asociado con el cáncer.

2.6 Relaciones entre variables categóricas *

Hasta ahora, nos hemos concentrado en relaciones en las que al menos la variable respuesta era cuantitativa. Ahora nos interesaremos en relaciones entre dos o más variables categóricas. Algunas variables —como son el sexo, la raza o la profesión— son intrínsecamente categóricas. Otras variables categóricas se crean agrupando valores de variables cuantitativas en clases. Cuando se publican datos, a menudo se presentan en forma agrupada para ahorrar espacio. Para analizar datos categóricos utilizamos recuentos o porcentajes de los individuos que componen las distintas clases o categorías.

[30]Gary Taubes, "Magnetic field-cancer link: will it rest in peace?", *Science*, 277, 1997, pág. 29.

*El contenido de este apartado es importante en estadística, pero en este libro no se necesita hasta el capítulo 8. Puedes omitirlo si no tienes pensado leer el capítulo 8, o retrasar su lectura hasta que no llegues a dicho capítulo.

EJEMPLO 2.19. Edad y educación

La tabla 2.10 presenta datos sobre el número de años de escolarización de ciudadanos estadounidenses de distintas edades. Muchos menores de 25 años todavía no han completado su educación, por lo que no se encuentran en la tabla. Las variables edad y educación se han agrupado en categorías. La tabla 2.10 es una **tabla de contingencia**, ya que describe dos variables categóricas. La educación es la **variable fila**, puesto que cada fila de la tabla describe a personas con un determinado nivel educativo. La edad es la **variable columna**, porque cada columna de la tabla describe a un grupo de edad distinto. Los valores de la tabla son los recuentos del número de personas que pertenecen a cada una de las categorías combinadas: edad y educación. Aunque las dos variables de la tabla son categóricas, las categorías de cada una de ellas se pueden ordenar de manera natural de menor a mayor. El orden de las columnas y de las filas de la tabla 2.10 refleja una ordenación natural de las distintas categorías. ■

Tablas de contingencia
Variables fila y variables columna

Tabla 2.10. Años de escolarización según edad, datos de 1995 (en miles de personas).

	Grupo de edad			
Educación	**25 a 34**	**35 a 54**	**Mayores de 55**	**Total**
No completaron secundaria	5.325	9.152	16.035	30.512
Completaron secundaria	14.061	24.070	18.320	56.451
De 1 a 3 cursos en la universidad	11.659	19.926	9.662	41.247
4 o más cursos en la universidad	10.342	19.878	8.005	38.225
Total	41.388	73.028	52.022	166.438

2.6.1 Distribuciones marginales

¿Cómo podemos captar mejor la información contenida en la tabla 2.10? En primer lugar, *fíjate en la distribución de cada variable de forma separada*. La distribución de una variable categórica tan sólo dice con qué frecuencia ha ocurrido cada resultado. La columna "Total", situada a la derecha de la tabla, contiene los totales de cada fila. Estos totales de las filas dan la distribución de la educación (la variable fila) entre toda la gente mayor de 25 años: 30.512.000 personas no completaron sus estudios de secundaria, 56.451.000 terminaron secundaria pero no fueron a la universidad, etc. De la misma manera, la fila "Total" en la parte inferior de la tabla da la distribución según la edad. Si la columna y la fila de totales no están,

lo primero que hay que hacer al analizar una tabla de contingencia es calcularlas. Las distribuciones de la variable fila y de la variable columna, de forma separada, se llaman **distribuciones marginales**, ya que aparecen en los márgenes derecho e inferior de la tabla de contingencia.

Distribuciones marginales

Si compruebas el cálculo de la fila y de la columna de totales de la tabla 2.10, encontrarás algunas discrepancias. Por ejemplo, la suma de los valores de la columna "25 a 34" es de 41.387 personas. El valor de la columna de totales para esta columna es de 41.388 personas. La explicación es el **error de redondeo**. Los valores de la tabla se expresan en miles de personas y cada valor se ha redondeado hasta el millar más próximo. Los técnicos que obtuvieron los totales lo hicieron a partir de los números exactos de personas entre 25 y 34 años, y posteriormente los redondearon. El resultado fue de 41.388.000 personas. Si se suman los valores ya redondeados que aparecen en la fila, se obtiene un valor ligeramente distinto.

Error de redondeo

A menudo, los porcentajes se captan más fácilmente que los recuentos. Podemos expresar la distribución marginal de la educación en forma de porcentajes dividiendo los valores de la columna de totales por el total de la tabla y multiplicando por cien.

EJEMPLO 2.20. Cálculo de la distribución marginal

El porcentaje de personas mayores de 25 años que completaron al menos 4 cursos universitarios es

$$\frac{\text{total con 4 años de universidad}}{\text{total}} = \frac{38{,}225}{166{,}438} = 0{,}230 = 23{,}0\%$$

Haremos tres cálculos más para obtener la distribución marginal de la educación en porcentajes. Aquí los tienes.

Educación	Porcentaje
No completaron secundaria	18,3
Completaron secundaria	33,9
De 1 a 3 cursos en la universidad	24,8
4 o más cursos en la universidad	23,0

El total es el 100%, ya que cada individuo pertenece a uno de los cuatro grupos educativos. ■

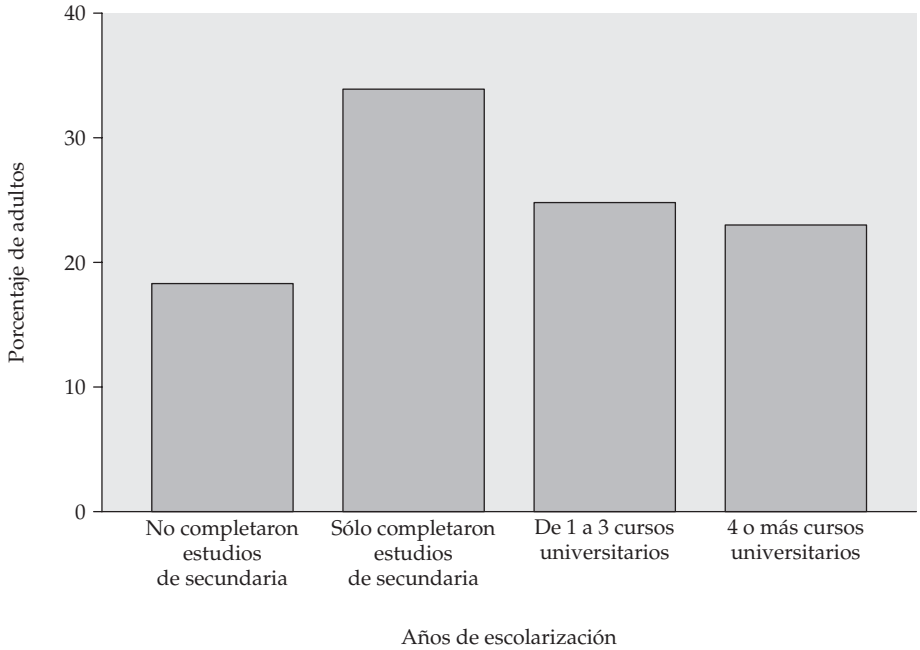


Figura 2.19. Diagrama de barras de la distribución de la educación entre la gente con más de 25 años. Este diagrama corresponde a una de las distribuciones marginales de la tabla 2.10.

Cada distribución marginal de una tabla de contingencia es la distribución de una sola variable categórica. Tal como vimos en el capítulo 1, podemos utilizar un diagrama de barras o un diagrama de sectores para mostrar esta distribución. La figura 2.19 es un diagrama de barras sobre la distribución de la escolarización. Vemos que la gente con al menos algún año en la universidad constituye casi la mitad de la población con más de 25 años.

Cuando trabajes con tablas de contingencia, tendrás que calcular porcentajes, muchos porcentajes. He aquí una orientación para ayudar a decidir qué fracción da el porcentaje que quieres. Pregunta: "¿qué grupo representa el total sobre el que quiero el porcentaje?". El recuento de este grupo es el denominador del cociente que permite obtener el porcentaje. En el ejemplo 2.20, queríamos el porcentaje "de gente mayor de 25 años", por tanto, el recuento de gente mayor de 25 años (el total de la tabla) es el denominador.

APLICA TUS CONOCIMIENTOS

2.68. Los recuentos de la columna "Total" situada a la derecha de la tabla 2.10 son recuentos de gente en cada grupo educativo. Explica por qué la suma de estos recuentos no es igual a 166.438, el total que aparece a la derecha de la última fila de la tabla.

2.69. A partir de los recuentos de la tabla 2.10, halla, en forma de porcentajes, la distribución marginal de la edad para la gente mayor de 25 años.

2.70. Hábitos fumadores de padres e hijos. Tenemos datos de ocho escuelas de secundaria sobre el consumo de tabaco entre los estudiantes y entre sus padres.[31]

Hábitos de los padres	Estudiantes fumadores	Estudiantes no fumadores
Los dos padres fuman	400	1.380
Sólo uno de los padres fuma	416	1.823
Ninguno de los dos padres fuma	188	1.168

(a) ¿A cuántos estudiantes describen estos datos?

(b) ¿Qué porcentaje de estos estudiantes son fumadores?

(c) Halla la distribución marginal del consumo de tabaco de los padres de dos maneras, con recuentos y en porcentajes.

2.6.2 Descripción de relaciones

La tabla 2.10 contiene, aparte de las dos distribuciones marginales de la edad y la educación, mucha más información. La naturaleza de la relación entre la edad y la educación no se puede deducir a partir de las distribuciones marginales; es necesaria toda la tabla. **Para describir las relaciones entre variables categóricas, calcula los porcentajes apropiados a partir de los recuentos**. Utilizamos porcentajes porque los recuentos suelen ser difíciles de comparar directamente. Por ejemplo, 19.878.000 personas entre 35 y 54 años completaron sus estudios universitarios, mientras que sólo 8.005.000 personas con al menos 55 los completaron.

[31] S. V. Zagona (ed.), *Studies and Issues in Smoking Behavior*, University of Arizona Press, Tucson, 1967, págs. 157-180.

El grupo más joven es mucho mayor; por tanto, no podemos comparar directamente estos recuentos.

EJEMPLO 2.21. Educación universitaria

¿Qué porcentaje de personas entre 25 y 34 años completaron 4 cursos de estudios universitarios? Se trata del recuento de personas entre 25 y 34 años que completaron 4 cursos universitarios expresado como porcentaje del total de personas de este grupo de edad:

$$\frac{10{,}342}{41{,}388} = 0{,}250 = 25{,}0\%$$

"Las personas entre 25 y 34" es el grupo del cual queremos un porcentaje; por tanto, el recuento de este grupo es el denominador. De la misma manera, halla el porcentaje de gente de cada grupo de edad con al menos 4 cursos en la universidad. La comparación de los tres grupos es

Grupos de edad	Porcentaje de alumnos con al menos 4 cursos universitarios
25-34	25,0
35-54	27,2
55 o más	15,4

Estos porcentajes nos ayudan a ver cómo la educación universitaria es menos frecuente entre los estadounidenses de 55 o más años que entre adultos más jóvenes. Éste es una aspecto importante de la asociación entre edad y educación. ■

APLICA TUS CONOCIMIENTOS

2.71. Utilizando los recuentos de la tabla 2.10, halla el porcentaje de gente de cada grupo de edad que no terminó la secundaria. Dibuja un diagrama de barras para comparar estos porcentajes. Explica lo que muestran los datos.

2.72. Huevos de serpientes de agua. ¿Cómo influye la temperatura sobre la eclosión de los huevos de serpiente de agua? Unos investigadores distribuyeron huevos recién puestos a tres temperaturas: caliente, templada y fría. La temperatura del agua caliente era el doble de la temperatura de la hembra de serpiente de

agua y la temperatura del agua fría era la mitad de la temperatura corporal de la hembra de serpiente de agua. He aquí los datos sobre el número de huevos y el número de huevos que eclosionaron:[32]

	Fría	**Templada**	**Caliente**
Número de huevos	27	56	104
Huevos eclosionados	16	38	75

(a) Construye una tabla de contingencia con la temperatura y el resultado de la eclosión (sí o no).

(b) Calcula el porcentaje de huevos de cada grupo que eclosionó. Los investigadores opinaban que los huevos no eclosionarían en agua fría. Los datos, ¿apoyan esta opinión?

2.6.3 Distribuciones condicionales

El ejemplo 2.21 no compara las distribuciones de la educación de los tres grupos de edad. Sólo compara los porcentajes de la gente con al menos 4 cursos universitarios. Vamos a ver la situación completa.

EJEMPLO 2.22. Cálculo de distribuciones condicionales

La información sobre el grupo de edad entre 25 y 34 años aparece en la primera columna de la tabla 2.10. Para hallar la distribución completa de la educación en este grupo, mira sólo esta columna. Transforma cada recuento en un porcentaje en relación con el total de la columna, 41.338. He aquí la distribución:

	No terminaron secundaria	Terminaron secundaria	De 1 a 3 cursos universitarios	4 o más cursos universitarios
Porcentaje	12,9	34,0	28,2	25,0

La suma de estos porcentajes tiene que ser 100, ya que todos los individuos de 25 a 34 años pertenecen a alguna de las categorías educativas (de hecho, la suma

[32] R. Shine, T. R. L. Madsen, M. J. Elphick y P. S. Harlow, "The influence of nest temperatures and maternal brooding on hatchling phenotypes in water pythons", *Ecology*, 78, 1997, págs. 1.713-1.721.

Distribución condicional

es 100,1 debido al error de redondeo). Estos cuatro porcentajes son la **distribución condicional** de la educación dado que una persona tiene entre 25 y 34 años. Utilizamos el término "condicional" porque la distribución se refiere sólo a las personas que satisfacen la condición de tener entre 25 y 34 años. ■

Ahora fíjate en la segunda (gente de 35 a 54 años) y en la tercera columna (gente de 55 o más años) de la tabla 2.10 para hallar dos distribuciones condicionales más. Los programas estadísticos pueden expresar rápidamente los valores de cada columna como porcentajes en relación con el total de la columna. La figura 2.20 muestra este resultado. El programa halló los totales de filas y columnas a partir de los valores de la tabla; pueden ser distintos de los de la tabla 2.10.

Cada celda de esta tabla contiene el recuento de la tabla 2.10, así como este recuento expresado como un porcentaje del total de la columna. Los porcentajes de cada columna constituyen la distribución condicional de los años de escolarización de cada grupo de edad. Los porcentajes de cada columna suman 100%, ya que se tiene en cuenta a toda la gente de cada grupo de edad. La comparación de las distribuciones condicionales pone de manifiesto el tipo de asociación existente entre edad y educación. La distribución de la educación en los dos grupos más jóvenes es bastante similar; sin embargo, la educación superior es menos común en el grupo de gente de 55 o más años.

TABLA DE EDUCACION POR EDAD

EDUCACION EDAD

Frequency Col Pct	25-34	35-54	55 over	Total
NoSecund	5325	9152	16035	30512
	12.87	12.53	30.82	
SoloSecund	14061	24070	18320	56451
	33.97	32.96	35.22	
Univ_1_3	11659	19926	9662	41247
	28.17	27.29	18.57	
Univ_sup4	10342	19878	8005	38225
	24.99	27.22	15.39	
Total	41387	73026	52022	166435

Figura 2.20. Resultados del SAS de la tabla de contingencia de edad por educación, con las distribuciones condicionales de la educación en cada grupo de edad. Los porcentajes de cada columna suman 100%.

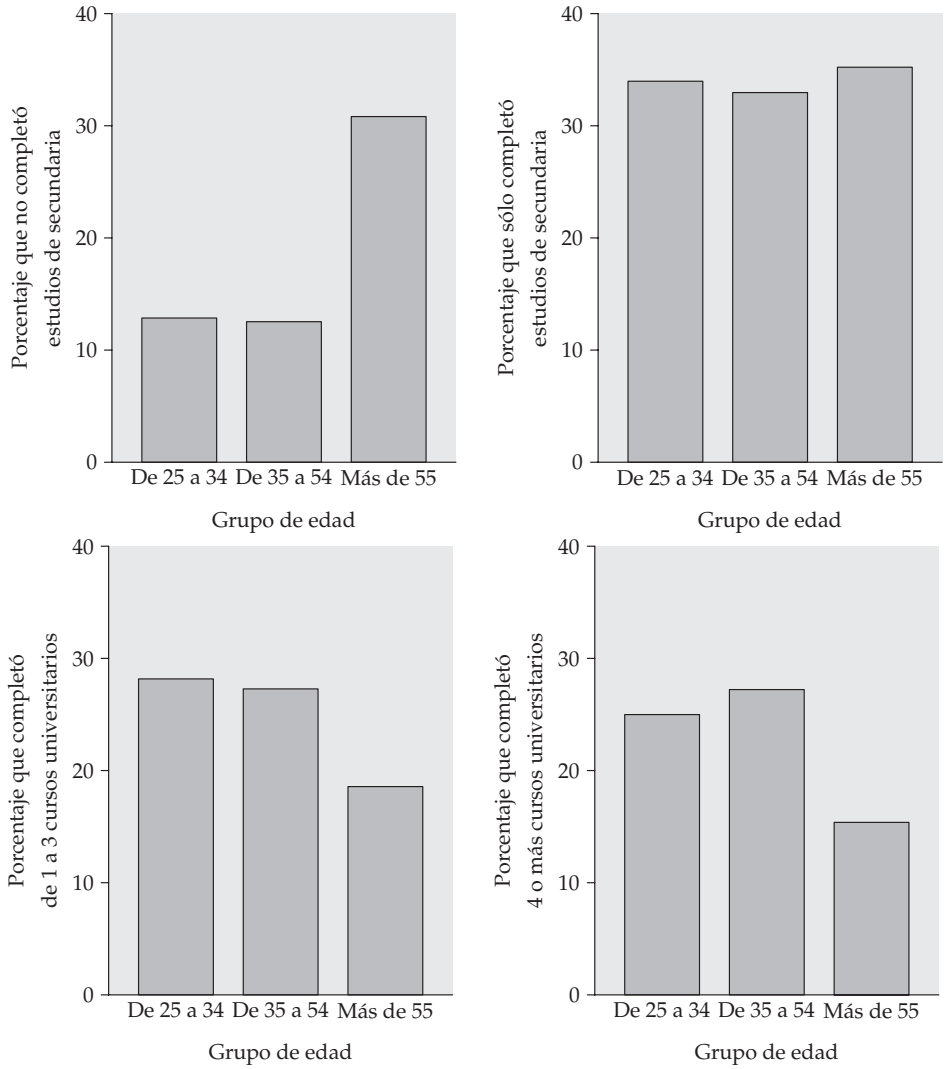


Figura 2.21. Diagrama de barras para comparar los niveles educativos de tres grupos de edad. Dentro de cada nivel educativo, cada barra compara los porcentajes de tres grupos de edad.

Los diagramas de barras pueden ayudar a visualizar una asociación. Dibujaremos tres diagramas de barras en un mismo gráfico, cada uno de ellos similar al de la figura 2.19, para mostrar las tres distribuciones condicionadas. La figura 2.21 muestra una forma alternativa de diagrama de barras. Cada conjunto de tres barras compara los porcentajes en cada uno de los grupos de edad que ha alcanzado un determinado nivel educativo. Es fácil ver que las barras de "25 a 34"

y de "35 a 54" son similares en los cuatro niveles de educación, y que las barras de "55 o más años" muestran que mucha más gente de este grupo no terminó secundaria y que muy pocos llegaron a la universidad.

No existe ningún gráfico (como por ejemplo los diagramas de dispersión) que visualice la forma de la relación entre variables categóricas. Tampoco existe ningún resumen numérico (como por ejemplo la correlación) que valore la fuerza de la asociación. Los diagramas de barras son lo suficientemente flexibles como para visualizar las comparaciones que quieras mostrar. Como resúmenes numéricos utilizaremos los porcentajes adecuados. Debes decidir qué porcentajes necesitas. He aquí una sugerencia: compara las distribuciones condicionales de la variable respuesta (educación) correspondientes a cada uno de los valores de la variable explicativa (edad). Es lo que hicimos en la figura 2.20.

APLICA TUS CONOCIMIENTOS

2.73. Halla la distribución condicional de la edad entre la gente con al menos 4 cursos universitarios. Parte de los recuentos de la tabla 2.10. (Para hacerlo, fíjate sólo en la fila de "4 o más cursos en la universidad" de la tabla.)

2.74. Planes profesionales de hombres y mujeres. Un estudio sobre los planes profesionales de mujeres y hombres jóvenes envió cuestionarios a los 722 alumnos de una clase de último curso de Administración de Empresas de la University of Illinois. Una de las preguntas formuladas era qué especialidad habían escogido. He aquí datos sobre lo que contestaron los alumnos.[33]

	Mujeres	Hombres
Contabilidad	68	56
Administración	91	40
Economía	5	6
Finanzas	61	59

(a) Halla la distribución de la especialidad condicionada al sexo de los estudiantes. A partir de tus resultados describe las diferencias entre hombres y mujer con un gráfico y con palabras.

[33]F. D. Blau y M. A. Ferber, "Career plans and expectations of young women and men", *Journal of Human Resources*, 26, 1991, págs. 581-607.

(b) ¿Qué porcentaje de estudiantes no respondió el cuestionario? La falta de respuesta debilita los resultados obtenidos.

2.75. He aquí los totales de filas y columnas de una tabla de contingencia con dos filas y dos columnas.

a	b	50
c	d	50
60	40	100

Halla *dos diferentes* conjuntos de recuentos a, b, c y d que den los mismos totales. Este ejercicio muestra que la relación entre dos variables no se puede obtener a partir de las distribuciones individuales de las variables.

2.6.4 Paradoja de Simpson

Tal como ocurre con las variables cuantitativas, los efectos de las variables latentes pueden cambiar e incluso invertir las relaciones observadas entre dos variables categóricas. He aquí un ejemplo hipotético que muestra las sorpresas que pueden aguardar a quien utiliza, confiado, datos estadísticos.

EJEMPLO 2.23. ¿Qué hospital es más seguro?

Para ayudar a los ciudadanos a tomar decisiones sobre el cuidado de su salud, en Estados Unidos se publican datos sobre los hospitales del país. Imagínate que quieres comparar el Hospital A con el Hospital B de una misma ciudad. He aquí una tabla de contingencia con los datos sobre la supervivencia de los enfermos después de ser operados en estos dos hospitales. Todos los pacientes que han sido operados últimamente están incluidos. "Sobrevivió" significa que el paciente vivió al menos durante las 6 semanas siguientes a la operación.

	Hospital A	**Hospital B**
No sobrevivieron	63	16
Sobrevivieron	2.037	784
Total	2.100	800

La evidencia parece clara: el Hospital A pierde un 3% ($\frac{63}{2.100}$) de los pacientes que han sido operados en él, mientras que el Hospital B pierde sólo un 2% ($\frac{16}{800}$). Parece que te conviene escoger el Hospital B si necesitas operarte.

No obstante, no todas las operaciones son igual de complejas. Más adelante, en este mismo informe, los pacientes que ingresan en cada hospital aparecen clasificados de acuerdo con el estado general de salud que tenían antes de la operación. Se clasifican en dos grupos, los de buena salud y los de salud delicada. He aquí estos datos más detallados. Comprueba que los valores de la tabla de contingencia original no son más que la suma de los dos tipos de pacientes de las dos tablas siguientes.

Buena salud

	Hospital A	Hospital B
No sobrevivieron	6	8
Sobrevivieron	594	592
Total	600	600

Salud delicada

	Hospital A	Hospital B
No sobrevivieron	57	8
Sobrevivieron	1.443	192
Total	1.500	200

¡Ajá! El Hospital A gana al Hospital B en el caso de los pacientes con buena salud: sólo el 1% ($\frac{6}{600}$) de los pacientes ingresados en el Hospital A fallecieron, mientras que en el Hospital B esta cifra fue del 1,3% ($\frac{8}{600}$). El Hospital A también es mejor para los pacientes delicados de salud; este hospital pierde el 3,8% ($\frac{57}{1.500}$) mientras que el Hospital B pierde el 4% ($\frac{8}{200}$) de los pacientes. Por tanto, el Hospital A es más seguro para los dos tipos de pacientes, los de buena salud y los de salud delicada. Si tienes que operarte, te conviene escoger el Hospital A. ■

El estado general de salud de los pacientes es una variable latente cuando comparamos las tasas de mortalidad en los dos hospitales. Cuando ignoramos esta variable latente, el Hospital B parece más seguro, a pesar de que el Hospital A es mejor para los dos tipos de pacientes. ¿Cómo es posible que A sea mejor en cada grupo y peor en conjunto? Mira los datos. El Hospital A es un centro médico que atrae a pacientes muy enfermos de toda la región. Tenía 1.500 pacientes con salud delicada. El Hospital B tenía sólo 200 de este tipo. Como los pacientes con salud delicada tienen más probabilidades de morir, el Hospital A tiene una tasa de mortalidad mayor a pesar de los buenos resultados que obtiene con los dos tipos de pacientes. La tabla de contingencia original, que no tenía en cuenta el estado general de salud de los pacientes, era engañosa. El ejemplo 2.23 ilustra la *paradoja de Simpson*.

PARADOJA DE SIMPSON

La **paradoja de Simpson** se refiere al cambio de sentido de una comparación o de una asociación cuando datos de distintos grupos se combinan en un solo grupo.

Las variables latentes de la paradoja de Simpson son categóricas. Es decir, clasifican a los individuos en grupos, como cuando los pacientes operados se clasificaron en pacientes con buena salud y pacientes con salud delicada. La paradoja de Simpson no es más que un caso extremo del hecho de que asociaciones observadas pueden ser engañosas cuando hay variables latentes.

APLICA TUS CONOCIMIENTOS

2.76. Retrasos en los aeropuertos. He aquí el número de vuelos que llegaron a la hora prevista y el número de vuelos que llegaron con retraso de dos compañías aéreas en cinco aeropuertos de EE UU en un determinado mes. A menudo, los medios de comunicación dan a conocer los porcentajes de vuelos, de las distintas compañías, que llegan a la hora. El aeropuerto de procedencia es una variable latente que puede hacer que los datos que dan los medios de comunicación sean engañosos.[34]

	Alaska Airlines		America West	
	A la hora	**Con retraso**	**A la hora**	**Con retraso**
Los Angeles	497	62	694	117
Phoenix	221	12	4.840	415
San Diego	212	20	383	65
San Francisco	503	102	320	129
Seattle	1.841	305	201	61

(a) ¿Qué porcentaje de vuelos de Alaska Airlines llegan con retraso? ¿Qué porcentaje de vuelos de America West llegan con retraso? Estos son los datos que, en general, dan a conocer los medios de comunicación.

[34] A. Barnett, "How numbers can trick you", *Technology Review*, octubre 1994, págs. 38-45.

(b) Ahora considera los datos de cada aeropuerto por separado, ¿qué porcentaje de vuelos de Alaska Airlines llegan con retraso? ¿Y de America West?

(c) Considerando los aeropuertos por separado, America West es la peor compañía. Sin embargo, considerando todos los aeropuertos conjuntamente es la mejor. Parece una contradicción. Explica cuidadosamente, basándote en los datos, cómo se puede explicar. (Los climas de Phoenix y Seattle pueden explicar este ejemplo de paradoja de Simpson.)

2.77. Raza y condena a muerte. El hecho de que un acusado de asesinato sea condenado o no a muerte parece estar influenciado por la raza de la víctima. Tenemos datos de 326 casos en los que el acusado fue declarado culpable de asesinato:[35]

	Acusado blanco Pena de muerte		**Acusado negro Pena de muerte**	
	Sí	No	Sí	No
Víctima blanca	19	132	11	52
Víctima negra	0	9	6	97

(a) Utiliza estos datos para construir una tabla de contingencia que relacione la raza del acusado (blanco o negro) con la pena de muerte (sí o no).

(b) Constata que se cumple la paradoja de Simpson: en conjunto, un mayor porcentaje de acusados blancos son condenados a pena de muerte; en cambio, considerando de manera independiente a las víctimas blancas y a las negras, el porcentaje de acusados negros condenados a muerte es mayor que el de blancos.

(c) Utiliza los datos para explicar, en un lenguaje que pueda entender un juez, por qué se da la paradoja.

RESUMEN DE LA SECCIÓN 2.6

Una **tabla de contingencia** de recuentos describe la relación entre dos variables categóricas. Los valores de la **variable fila** identifican las filas de la tabla. Los valores de la **variable columna** identifican las columnas. A menudo, las tablas de contingencia se utilizan para resumir grandes cantidades de datos agrupando los resultados en categorías.

[35]M. Radelet, "Racial characteristics and imposition of the death penalty", *American Sociological Review*, 46, 1981, págs. 918-927.

La **columna de totales** y la **fila de totales** de una tabla de contingencia dan las **distribuciones marginales** de las dos variables de forma separada. Las distribuciones marginales no dan información sobre la relación entre las variables.

Para hallar la **distribución condicional** de la variable fila con relación a un valor determinado de la variable columna, fíjate sólo en esa columna de la tabla. Expresa cada valor de la columna como un porcentaje del total de la columna.

Existe una distribución condicional de la variable fila para cada columna de la tabla. La comparación de estas distribuciones condicionales es una manera de mostrar la asociación entre la variable fila y la variable columna. Es especialmente útil cuando la variable columna es la variable explicativa.

Los **diagramas de barras** son una manera flexible de presentar las variables categóricas. No existe una sola manera de representar la asociación entre dos variables categóricas.

Una comparación entre dos variables que se cumple para cada uno de los valores individuales de una tercera variable, puede cambiar o incluso invertirse cuando se agregan los datos correspondientes a todos los valores de la tercera variable. Esto constituye la **paradoja de Simpson**. Esta paradoja es un ejemplo del efecto de las variables latentes sobre una determinada asociación.

EJERCICIOS DE LA SECCIÓN 2.6

Graduados universitarios. Los ejercicios 2.78 a 2.82 se basan en la tabla 2.11. Esta tabla de contingencia proporciona datos sobre los estudiantes matriculados en el otoño de 1995, tanto en universidades estadounidenses que ofrecen sólo primer ciclo, como en universidades que ofrecen primer y segundo ciclo.[36]

2.78. (a) ¿Cuántos estudiantes están matriculados en primer o en segundo ciclo?

(b) ¿Qué porcentaje de estudiantes de entre 18 y 24 años se matricularon?

(c) Halla los porcentajes de los estudiantes con edades entre 18 y 24 años que están matriculados en las opciones que aparecen en la tabla 2.11. Dibuja un diagrama de barras para comparar estos porcentajes.

(d) El grupo de estudiantes con edades entre 18 y 24 años es el grupo de edad tradicional para estudiantes universitarios. Resume brevemente lo que has aprendido a partir de los datos sobre el predominio de este tipo de estudiantes en los distintos estudios universitarios.

[36] *Digest of Education Statistics 1997*, National Centre for Education Statistics, página web <http://www.ed.gov/NCES>.

Tabla 2.11. Edad de los estudiantes universitarios de EE UU, otoño de 1995 (en miles de estudiantes).

	Primer ciclo		Segundo ciclo	
Edad	**Tiempo completo**	**Tiempo parcial**	**Tiempo completo**	**Tiempo parcial**
Menor de 18	41	125	75	45
Entre 18 a 24	1.378	1.198	4.607	588
Entre 25 a 39	428	1.427	1.212	1.321
Mayor de 40	119	723	225	605
Total	1.966	3.472	6.119	2.559

2.79. (a) Una asociación de alumnos de primer ciclo pregunta: "¿Qué porcentaje de estudiantes de primer ciclo a tiempo parcial, tiene entre 25 y 39 años?"

(b) Un banco que proporciona préstamos a adultos para estudios pregunta: "¿Qué porcentaje de estudiantes que tienen entre 25 y 39 años están matriculados en primer ciclo?"

2.80. (a) Halla la distribución marginal de la edad entre todos los estudiantes; primero, en forma de recuentos y luego en forma de porcentajes. Dibuja un diagrama de barras con estos porcentajes.

(b) Halla la distribución condicional de la edad (en porcentajes) entre los estudiantes matriculados a tiempo parcial en primer ciclo.

(c) Describe brevemente las diferencias más importantes entre las dos distribuciones de edad.

(d) La suma de todos los valores de la columna "Primer ciclo. Tiempo parcial" no es la misma que el total que aparece en la tabla. ¿Por qué?

2.81. Llama a los estudiantes de 40 o más años "estudiantes mayores". Compara la presencia de estos estudiantes en los 4 tipos de matriculación, de forma numérica y con un gráfico. Resume tus hallazgos.

2.82. Pensando un poco puedes obtener más información de la tabla 2.11 que las distribuciones marginales y las distribuciones condicionales. En general, la mayoría de estudiantes universitarios tienen entre 18 y 24 años.

(a) ¿Qué porcentaje de universitarios se encuentran en este grupo de edad?

(b) ¿Qué porcentaje de estudiantes de primer ciclo se hallan en ese grupo?

(c) ¿Y de estudiantes a tiempo parcial?

2.83. Muertes por armas de fuego en EE UU. Después de los accidentes de tráfico, las muertes por armas de fuego constituyen la segunda causa de mortalidad no debida a enfermedades en EE UU. He aquí un estudio sobre las muertes relacionadas con armas de fuego en Milwaukee, Wisconsin, entre 1990 y 1994.[37] Queremos comparar el tipo de arma utilizada en los homicidios y en los suicidios. Sospechamos que a menudo, para los suicidios, se utilizan armas de caza (escopetas y rifles) que se tienen en casa. Compara con un diagrama de barras el tipo de armas utilizadas en suicidios y homicidios. ¿Qué diferencia existe entre las armas utilizadas para cazar (escopetas y rifles) y las pistolas?

	Pistola	**Escopeta**	**Rifle**	**Desconocido**	**Total**
Homicidios	468	28	15	13	524
Suicidios	124	22	24	5	175

2.84. No-respuesta en una encuesta. Una escuela de empresariales realizó una encuesta sobre las empresas de su entorno geográfico. La escuela envió un cuestionario a 200 empresas pequeñas, a 200 empresas medianas y a 200 empresas grandes. La proporción de no-respuesta es importante para decidir la fiabilidad de los resultados. Los datos sobre las respuestas a esta encuesta son

	Pequeñas	**Medianas**	**Grandes**
Respuesta	125	81	40
No-respuesta	75	119	160
Total	200	200	200

(a) ¿Cuál fue el porcentaje global de no-respuesta?

(b) Describe la relación que existe entre las no-respuestas y el tamaño de la empresa. (Utiliza los porcentajes para que tu descripción sea precisa.)

(c) Haz un diagrama de barras para comparar los porcentajes de no-respuesta en los tres tipos de empresas.

[37] S. W. Hargarten *et al*. "Characteristics of firearms involved in fatalities", *Journal of the American Medical Association*, 275, 1996, págs. 42-45.

2.85. Ayuda a adictos a la cocaína. La adicción a la cocaína es difícil de superar. Es posible que la administración de antidepresivos pudiera ayudar a los adictos a abandonar su hábito. Un estudio de tres años de duración con 72 adictos crónicos a la cocaína comparó un antidepresivo llamado desipramina con el litio y un placebo. (El litio es un medicamento habitual para tratar la adicción a la cocaína. Un placebo es un falso medicamento utilizado para ver el efecto de tratamientos sin medicación.) Cada uno de estos tres medicamentos fue administrado al azar a un tercio de los sujetos. He aquí los resultados:[38]

	Reincidencia	
	Sí	**No**
Desipramina	10	14
Litio	18	6
Placebo	20	4

(a) Compara la efectividad de cada uno de los tratamientos para prevenir la reincidencia en el hábito. Utiliza porcentajes y dibuja un diagrama de barras.

(b) ¿Crees que este estudio proporciona una evidencia sólida de que la desipramina causa realmente una reducción de la reincidencia?

2.86. Edad y estado civil de las mujeres. La siguiente tabla de contingencia describe la edad y el estado civil de las mujeres adultas estadounidenses en 1995. Los valores de la tabla se expresan en miles de mujeres.

	Estado civil				
Edad (años)	**Soltera**	**Casada**	**Viuda**	**Divorciada**	**Total**
18 a 24	9.289	3.046	19	260	12.613
25 a 39	6.948	21.437	206	3.408	32.000
40 a 64	2.307	26.679	2.219	5.508	36.713
≥ 65	768	7.767	8.636	1.091	18.264
Total	19.312	58.931	11.080	10.266	99.588

(a) Calcula la suma de los valores de la columna "Casada". ¿Por qué difiere esta suma del valor que aparece en la columna de totales?

[38]D. M. Barnes, "Breaking the cycle of addiction", *Science*, 241, 1988, págs. 1.029-1.030.

(b) Halla la distribución marginal del estado civil de las mujeres adultas (utiliza porcentajes). Dibuja un diagrama de barras para mostrar la distribución.

(c) Compara las distribuciones condicionales del estado civil de las mujeres con edades entre 18 y 24 años, y de las mujeres entre 40 y 64. Describe brevemente las principales diferencias entre estos dos grupos de mujeres apoyándote en los valores porcentuales.

(d) Imagínate que quieres publicar una revista dirigida a mujeres solteras. Halla la distribución condicional de las edades entre las mujeres solteras. Muestra esta distribución mediante un diagrama de barras. ¿A qué grupo o grupos de edad se debería dirigir tu revista?

2.87. ¿Discriminación? Un Instituto Superior de Empresariales imparte dos titulaciones: una de Dirección de Empresas y otra de Derecho. Los aspirantes a cursar estudios en dicho centro deben superar una prueba de admisión. A continuación, se presentan dos tablas de contingencia en las que se clasifica a los aspirantes a cursar estudios en cada una de las titulaciones en función del sexo y del resultado en la prueba de admisión.[39]

	Dirección de Empresas	
	Admitido	**No admitido**
Hombre	480	120
Mujer	180	20

	Derecho	
	Admitido	**No admitido**
Hombre	10	90
Mujer	100	200

(a) Construye una tabla de contingencia con el sexo y el resultado de la prueba de admisión para las dos titulaciones conjuntamente, sumando los recuentos de cada tabla.

(b) A partir de la tabla anterior, calcula el porcentaje de hombres y de mujeres admitidos. El porcentaje de hombres admitidos es superior al de mujeres.

(c) Calcula de forma independiente el porcentaje de mujeres y de hombres admitidos según se trate de aspirantes a Dirección de Empresas o a Derecho. En ambas titulaciones la proporción de mujeres admitidas es superior a la de hombres.

(d) Se cumple la paradoja de Simpson: en cada una de las dos titulaciones el porcentaje de mujeres admitidas es superior al de hombres. Sin embargo, considerando a todos los alumnos conjuntamente, el porcentaje de hombres admitidos

[39] P. J. Bickel y J. W. O'Connell, "Is there a sex bias in graduate admissions?", *Science*, 187, 1975, págs. 398-404.

es superior al de mujeres. Explica esta paradoja en un lenguaje sencillo, para que lo pueda entender una persona que no tenga una especial formación estadística.

2.88. Obesidad y salud. Estudios recientes han puesto de manifiesto que los primeros trabajos sobre obesidad subestimaron los riesgos para la salud asociados con el sobrepeso. El error se debía a no tener en cuenta determinadas variables latentes. En concreto, fumar tiende a reducir el peso, pero conduce a una muerte más temprana. Con esta variable latente, ilustra de forma simplificada la paradoja de Simpson. Es decir, construye dos tablas de contingencia, una para fumadores y otra para no fumadores, con las variables sobrepeso (Sí o No) y muerte temprana (Sí o No). De manera que:

- Tanto los fumadores como los no fumadores con sobrepeso tiendan a morir antes que los que no tienen sobrepeso.
- Pero que cuando se combinen los fumadores y los no fumadores en una sola tabla de contingencia con las variables sobrepeso y muerte temprana, las personas sin sobrepeso tiendan a morir más tempranamente.

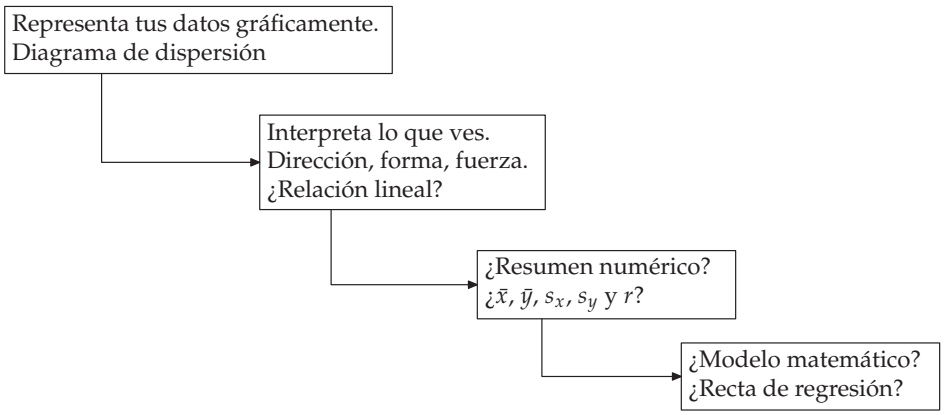
REPASO DEL CAPÍTULO 2

El capítulo 1 trató sobre el análisis de datos de una sola variable. En este capítulo, hemos estudiado el análisis de datos de dos o más variables. El análisis adecuado depende de si las variables son categóricas o cuantitativas, y de si una es una variable explicativa y la otra una variable respuesta.

Cuando tengas una variable categórica explicativa y una variable respuesta cuantitativa, utiliza las herramientas del capítulo 1 para comparar las distribuciones de la variable respuesta según las distintas categorías de la variable explicativa. Dibuja histogramas, diagramas de tallos o diagramas de caja en un mismo gráfico y compara las medianas y las medias. Si las dos variables son categóricas, no existe ningún gráfico satisfactorio (aunque los diagramas de barras pueden ayudar). Describimos su relación numéricamente comparando porcentajes. La sección 2.6, que es optativa, explica cómo hacerlo.

La mayor parte de este capítulo se concentra en la relación entre dos variables cuantitativas. La figura que aparece a continuación organiza las principales ideas de manera que se recalca que nuestras tácticas son las mismas que vimos cuando nos enfrentamos con datos de una sola variable en el capítulo 1. He aquí una lista de lo más importante que tendrías que haber aprendido al estudiar este capítulo.

Análisis de datos de dos variables



A. DATOS

1. Reconocer si una variable es cuantitativa o categórica.
2. Identificar la variable explicativa y la variable respuesta en situaciones donde una variable explica o influye sobre otra.

B. DIAGRAMAS DE DISPERSIÓN

1. Dibujar un diagrama de dispersión de dos variables cuantitativas, situando la variable explicativa (si hay alguna) en el eje de las abscisas.
2. Añadir una variable categórica a un diagrama de dispersión utilizando un símbolo gráfico diferente.
3. Describir la forma, la dirección y la fuerza del aspecto general del diagrama de dispersión. En concreto, reconocer asociaciones positivas o negativas, una relación lineal y las observaciones atípicas en un diagrama de dispersión.

C. CORRELACIÓN

1. Calcular, utilizando una calculadora, el coeficiente de correlación r entre dos variables cuantitativas.

2. Conocer las propiedades básicas de la correlación: r sólo mide la fuerza y la dirección de relaciones lineales; siempre $-1 \leq r \leq 1$; $r = \pm 1$ sólo cuando existe una perfecta relación lineal entre dos variables; a medida que aumenta la fuerza de la relación lineal, r se aleja de 0 y se acerca a ± 1.

D. RECTAS

1. Explicar qué significan la pendiente b y la ordenada en el origen a en la ecuación $y = a + bx$ de una recta.
2. Representar gráficamente una recta a partir de su ecuación.

E. REGRESIÓN

1. Calcular, utilizando una calculadora, la recta de regresión mínimo-cuadrática de una variable respuesta y respecto a una variable explicativa x a partir de los datos.
2. Hallar la pendiente y la ordenada en el origen de la recta de regresión mínimo-cuadrática a partir de las medias, las desviaciones típicas de x y de y, y su correlación.
3. Utilizar la recta de regresión para predecir y con una x dada. Reconocer la extrapolación y ser consciente de sus peligros.
4. Utilizar r^2 para describir qué parte de la variación de una variable se puede explicar a partir de la relación lineal con otra variable.
5. Reconocer las observaciones atípicas y las observaciones influyentes potenciales a partir de un diagrama de dispersión con la recta de regresión dibujada en él.
6. Calcular los residuos y representar su distribución respecto a la variable explicativa x o respecto a otras variables. Reconocer las distribuciones poco comunes.

F. LIMITACIONES DE LA REGRESIÓN Y DE LA CORRELACIÓN

1. Comprender que tanto r como la recta de regresión mínimo-cuadrática pueden estar fuertemente influidas por unas pocas observaciones extremas.

2. Saber que las correlaciones calculadas a partir de las medias de varias observaciones son en general más fuertes que las correlaciones de observaciones individuales.
3. Reconocer posibles variables latentes que puedan explicar la asociación observada entre dos variables x e y.
4. Comprender que incluso una fuerte correlación no significa que exista una relación causa-efecto entre x e y.

G. DATOS CATEGÓRICOS (OPTATIVO)

1. Hallar, a partir de una tabla de contingencia de recuentos, las distribuciones marginales de dos variables obteniendo las sumas de las filas y las sumas de las columnas.
2. Expresar cualquier distribución en porcentajes dividiendo los recuentos de cada categoría por su total.
3. Describir la relación entre dos variables categóricas calculando y comparando los porcentajes. A menudo, esto exige comparar las distribuciones condicionales de una variable para las distintas categorías de la otra variable.
4. Identificar la paradoja de Simpson y ser capaz de explicarla.

EJERCICIOS DE REPASO DEL CAPÍTULO 2

2.89. El vino, ¿es bueno para tu corazón? La tabla 2.4 proporciona datos sobre el consumo de vino y muertes por ataques al corazón en 19 países. Un diagrama de dispersión (ejercicio 2.11) muestra una relación relativamente fuerte.

(a) La correlación para estas variables es $r = -0{,}843$. ¿Por qué la correlación es negativa? ¿Qué porcentaje de la variación de la tasa de mortalidad por ataques al corazón se puede explicar a partir de la relación lineal entre los ataques al corazón y el consumo de vino?

(b) La recta de regresión mínimo-cuadrática para la predicción de la tasa de ataques al corazón a partir del consumo de vino, calculada a partir de los datos de la tabla 2.4, es

$$y = 260{,}56 - 22{,}969x$$

Utiliza esta ecuación para predecir la tasa de mortalidad por ataques al corazón en un país en el que el consumo de alcohol, procedente del vino, de los adultos es de 4 litros anuales.

(c) La correlación en (a) y la pendiente de la recta de regresión mínimo-cuadrática en (b) son ambas negativas. ¿Es posible que estos dos valores tengan signos distintos? Justifica tu respuesta.

2.90. Edad y educación en EE UU. En general, el nivel educativo de la gente mayor es menor que el de la gente más joven; por tanto, podemos sospechar que existe una relación entre el porcentaje de residentes de un Estado de 65 o más años y el porcentaje de población sin estudios universitarios. La figura 2.22 muestra la relación entre estas variables. Los datos son los que aparecen en la tablas 1.1 y 2.1.

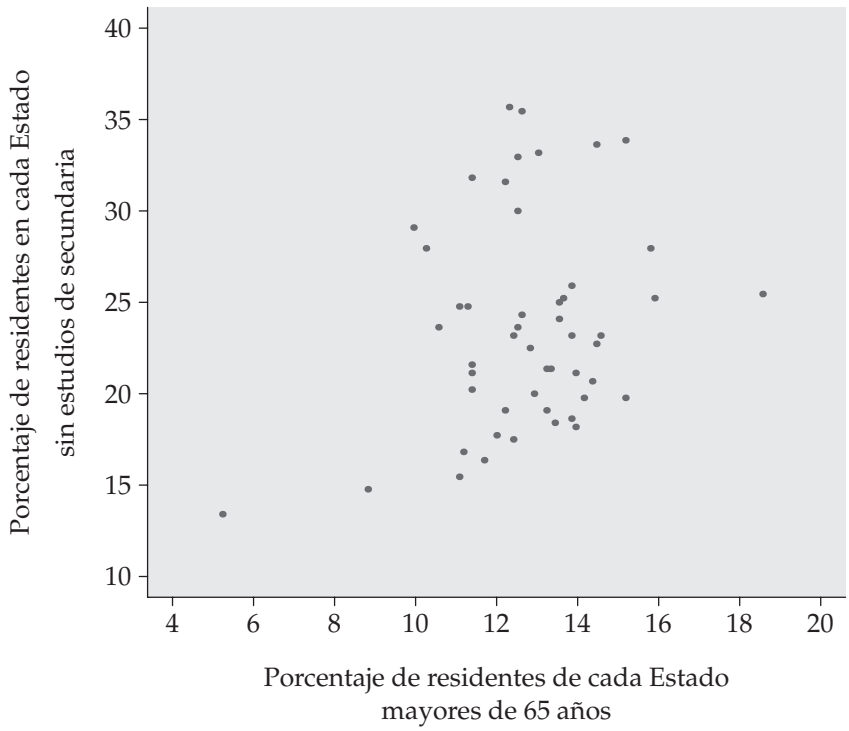


Figura 2.22. Diagrama de dispersión que relaciona el porcentaje de residentes de cada Estado de EE UU sin estudios de secundaria con el porcentaje de residentes de cada Estado mayores o iguales de 65 años.

(a) Explica lo que significa una asociación positiva entre estas variables.

(b) En el diagrama destacan tres observaciones atípicas. Dos de ellas son Alaska y Florida, que ya se identificaron como observaciones atípicas en el histograma de la figura 1.2. La tercera observación atípica, ¿a qué Estado corresponde?

(c) Si ignoramos las observaciones atípicas, ¿la relación entre las dos variables tiene una forma y dirección claras? Justifica tu respuesta.

(d) Si calculamos la correlación, con las tres observaciones atípicas y sin ellas, obtenemos $r = 0{,}054$ y $r = 0{,}259$. ¿Cuál de estos valores corresponde a la correlación sin las observaciones atípicas? Justifica tu respuesta.

2.91. Comida en mal estado. He aquí datos sobre 18 personas que enfermaron después de ingerir comida en mal estado.[40] Los datos dan la edad de cada persona en años, el periodo de incubación (el tiempo en horas desde la ingestión de la comida hasta la aparición de los primeros síntomas) y si la víctima sobrevivió (S) o murió (M).

Persona	Edad	Incubación	Resultado	Persona	Edad	Incubación	Resultado
1	29	13	0	10	30	36	0
2	39	46	1	11	32	48	0
3	44	43	1	12	59	44	1
4	37	34	0	13	33	21	0
5	42	20	0	14	31	32	0
6	17	20	1	15	32	86	1
7	38	18	0	16	32	48	0
8	43	72	1	17	36	28	1
9	51	19	0	18	50	16	0

(a) Dibuja un diagrama de dispersión del tiempo de incubación con relación a la edad. Utiliza símbolos distintos para las personas que murieron y para las que sobrevivieron.

(b) ¿Existe alguna relación entre la edad y el tiempo de incubación? Si existe, descríbela.

(c) Más importante, ¿existe alguna relación entre la edad o el periodo de incubación y si la víctima sobrevivió? Describe cualquier relación que aquí parezca importante.

(d) ¿Existen observaciones atípicas que exijan un investigación aparte?

[40] Datos proporcionados por Dana Quade, University of North Carolina.

2.92. Nematodos y tomateras. Los nematodos son gusanos microscópicos. Tenemos datos de un experimento para estudiar el efecto que producen los nematodos que se encuentran en la tierra en el crecimiento de las plantas. Un investigador preparó 16 contenedores de siembra e introdujo en ellos diferentes cantidades de nematodos. Luego, puso un plantón de tomatera en cada contenedor y a los 16 días midió su crecimiento (en centímetros).[41]

Nematodos	Crecimiento (cm)			
0	10,8	9,1	13,5	9,2
1.000	11,1	11,1	8,2	11,3
5.000	5,4	4,6	7,4	5,0
10.000	5,8	5,3	3,2	7,5

Analiza estos datos y presenta tus conclusiones sobre los efectos de los nematodos en el crecimiento de las plantas.

2.93. ¿Valores calientes? En el mundo de las finanzas es frecuente describir el rendimiento de un determinado valor mediante una recta de regresión que relaciona el rendimiento del valor con el rendimiento general del mercado de valores. Esta representación nos ayuda a visualizar en qué medida el valor sigue la pauta general del mercado. Analizamos el rendimiento mensual y de Philip Morris y los rendimientos mensuales x del índice bursátil Standard & Poor's correspondiente a 500 valores, que representa el mercado en su conjunto, entre julio de 1990 y mayo de 1997. He aquí los resultados:

$$\bar{x} = 1{,}304 \qquad s_x = 3{,}392 \qquad r = 0{,}5251$$

$$\bar{y} = 1{,}878 \qquad s_y = 7{,}554$$

Un diagrama de dispersión muestra que no existen observaciones influyentes destacables.

(a) A partir de esta información, halla la ecuación de la recta mínimo-cuadrática. ¿Qué porcentaje de la variación de Philip Morris se explica por la relación lineal con el mercado en su conjunto?

(b) Explica la información que nos proporciona sobre la pendiente de la recta sobre la respuesta de Philip Morris a las variaciones del mercado. Esta pendiente se llama "beta" en teoría de inversiones.

[41] Datos proporcionados por Matthew Moore.

(c) Los rendimientos de la mayoría de los valores están correlacionados positivamente con el rendimiento general del mercado. Es decir, cuando sube el mercado, las acciones individuales también tienden a subir. Explica por qué un inversor debería preferir valores con beta > 1 cuando el mercado sube y acciones con beta < 1 cuando el mercado baja.

2.94. La epidemia de gripe de 1918. El ejercicio 1.22 proporciona datos sobre la gran epidemia de gripe de 1918. Los diagramas temporales parecen mostrar que las muertes semanales siguen la misma pauta que los nuevos casos semanales con aproximadamente una semana de retraso. Vamos a analizar detalladamente esta relación.[42]

(a) Dibuja tres diagramas de dispersión; en uno de ellos relaciona las muertes semanales con los casos detectados la misma semana, en otro relaciona estas muertes semanales con los casos detectados la semana anterior y en el tercero de ellos relaciona las muertes semanales con los nuevos casos detectados dos semanas antes. Describe y compara las relaciones que observes.

(b) Halla las correlaciones de cada una de las relaciones.

(c) ¿Cuáles son tus conclusiones? ¿Cómo se predice mejor el número de muertes, con los nuevos casos de la misma semana, con los nuevos casos de la semana anterior o los nuevos casos de dos semanas antes?

2.95. Salarios de mujeres. Un estudio de la National Science Foundation[43] de EE UU, halló que la mediana del salario de ingenieras y científicas estadounidenses recién graduadas era sólo un 73% de la mediana de sus homólogos varones. Ahora bien, cuando las recién graduadas se agrupaban por especialidades, la situación era distinta. Las medianas de los salarios de las mujeres, expresadas como un porcentaje de las medianas del salario de los hombres, en 16 campos de estudio eran

94%	96%	98%	95%	85%	85%	84%	100%
103%	100%	107%	93%	104%	93%	106%	100%

¿Cómo es posible que el salario de las mujeres se encuentre muy por debajo del de los hombres cuando se consideran todas las disciplinas conjuntamente y, en cambio, sea prácticamente el mismo cuando se considera por especialidades?

[42] A. W. Crosby, *America's Forgotten Pandemic: The Influenza of 1918*, Cambridge University Press, New York, 1989.

[43] National Science Board, *Science and Engineering Indicators, 1991*, U.S. Government Printing Office, Washington, D.C., 1991. Los datos aparecen en la tabla 3-5 del apéndice.

2.96. Transformación de datos. Los ecólogos recogen datos para estudiar la naturaleza. La tabla 2.12 proporciona datos sobre la media del número de semillas producidas durante un año por algunas especies comunes de árboles y también sobre el peso medio (en miligramos) de éstas. Algunas especies aparecen dos veces, ya que se han considerado en dos localidades. Creemos que los árboles con semillas más pesadas producirán menos semillas, pero, ¿cuál es la forma de la relación?[44]

Tabla 2.12. Peso y recuento del número de semillas producidas por especies arbóreas.

Especies	Número de semillas	Peso de semillas	Especies	Número de semillas	Peso de semillas
Abedul para papel	27.239	0,6	Haya americana	463	247
Abedul amarillo	12.158	1,6	Haya americana	1.892	247
Picea del Canadá	7.202	2,0	Encina	93	1.851
Picea de Engelman	3.671	3,3	Encina escarlata	525	1.930
Picea roja del Canadá	5.051	3,4	Roble rojo americano	411	2.475
Tulipanero	13.509	9,1	Roble rojo americano	253	2.475
Pino ponderosa	2.667	37,7	Avellano de América	40	3.423
Abeto	5.196	40,0	Roble blanco del Canadá	184	3.669
Arce del azúcar	1.751	48,0	Roble blanco americano	107	4.535
Pino	1.159	216,0			

(a) Dibuja un diagrama de dispersión que muestre cómo se puede explicar el número de semillas producidas por un árbol, a partir del peso de éstas. Describe la forma, la dirección y la fuerza de la relación.

(b) Cuando tratamos con tamaños y pesos, los logaritmos de los datos originales son a menudo la forma más adecuada de expresar los datos. Utiliza tu calculadora o un programa informático para calcular los logaritmos de los pesos y recuentos de la tabla 2.12. Dibuja un nuevo diagrama de dispersión utilizando los datos transformados. Ahora, ¿cuál es la forma, la dirección y la fuerza de la relación?

2.97. Hombres y mujeres. La altura media de las mujeres estadounidenses cuando tienen 20 años de edad es de aproximadamente 164 cm, con una desviación típica de unos 6,35 cm. La altura media de los hombres de la misma edad es de aproximadamente 174 cm, con una desviación típica de unos 6,86 cm. Si la

[44]D. F. Greene y E. A. Johnson, "Estimating the mean annual seed production of trees", *Ecology*, 75, 1994, págs. 642-647.

correlación entre las alturas de parejas de hombres y mujeres jóvenes es aproximadamente $r = 0{,}5$, ¿cuál es la pendiente de la recta de regresión de la altura de los hombres con relación a la altura de sus mujeres en las parejas jóvenes? Dibuja un gráfico de esta recta de regresión. Predice la altura de un hombre cuya mujer mide 170 cm de altura.

2.98. Un juego informático. Un sistema multimedia para aprender estadística incluye una prueba para valorar la destreza de los sujetos en la utilización del ratón (*mouse*). El programa informático hace que aparezca, al azar, un círculo en la pantalla. El sujeto tiene que situarse sobre el círculo y *clicar* tan rápido como pueda. Tan pronto como el usuario ha *clicado* sobre el círculo, aparece uno nuevo. La tabla 2.13 proporciona datos sobre los ensayos realizados por un sujeto, 20 con cada mano. "Distancia" es la distancia desde el centro del círculo al punto donde se halla el cursor en el momento del *clicado*, las unidades de medida dependen del tamaño de la pantalla. "Tiempo" es el tiempo transcurrido entre el *clicado* de dos círculos consecutivos, en milisegundos.[45]

Tabla 2.13. Tiempos de respuesta en un juego informático.

Tiempo	Distancia	Mano	Tiempo	Distancia	Mano
115	190,70	derecha	240	190,70	izquierda
96	138,52	derecha	190	138,52	izquierda
110	165,08	derecha	170	165,08	izquierda
100	126,19	derecha	125	126,19	izquierda
111	163,19	derecha	315	163,19	izquierda
101	305,66	derecha	240	305,66	izquierda
111	176,15	derecha	141	176,15	izquierda
106	162,78	derecha	210	162,78	izquierda
96	147,87	derecha	200	147,87	izquierda
96	271,46	derecha	401	271,46	izquierda
95	40,25	derecha	320	40,25	izquierda
96	24,76	derecha	113	24,76	izquierda
96	104,80	derecha	176	104,80	izquierda
106	136,80	derecha	211	136,80	izquierda
100	308,60	derecha	238	308,60	izquierda
113	279,80	derecha	316	279,80	izquierda
123	125,51	derecha	176	125,51	izquierda
111	329,80	derecha	173	329,80	izquierda
95	51,66	derecha	210	51,66	izquierda
108	201,95	derecha	170	201,95	izquierda

[45] P. Velleman, *ActivStats 2.0*, Addison-Wesley Interactive, Reading, Mass., 1997.

(a) Sospechamos que el tiempo depende de la distancia. Dibuja un diagrama de dispersión del tiempo con relación a la distancia. Utiliza símbolos distintos para cada mano.

(b) Describe la relación que observas. ¿Puedes afirmar que el sujeto es diestro?

(c) Halla la recta de regresión del tiempo con relación a la distancia para las dos manos de forma independiente. Dibuja estas rectas en tu diagrama. De las dos regresiones, ¿cuál es mejor para predecir el tiempo a partir de la distancia? Da medidas numéricas que describan la precisión de las dos regresiones.

(d) Debido al aprendizaje, es posible que el sujeto lo haga mejor en los últimos ensayos. También es posible que lo haga peor debido a la fatiga. Dibuja un diagrama de residuos en el que los residuos aparezcan ordenados de acuerdo al orden de realización de los ensayos (de arriba abajo en la tabla 2.12). ¿Existe algún efecto sistemático en el orden de realización de las pruebas?

2.99. La tabla 2.1 proporciona datos sobre la educación en los Estados de EE UU. Utiliza un programa estadístico para examinar la relación entre las calificaciones de Matemáticas y de Lengua en la prueba SAT de la manera siguiente:

(a) Quieres predecir la calificación de Matemáticas en la prueba SAT de un Estado a partir de su calificación de Lengua. Con este fin, halla la recta de regresión mínimo-cuadrática. Sabes que la calificación media de Lengua de un determinado Estado al año siguiente fue 455. Utiliza tu recta de regresión para predecir su calificación media de Matemáticas.

(b) Representa los residuos de tu regresión con relación a la calificación de Lengua en la prueba SAT (un programa estadístico lo puede hacer). Hay un Estado que constituye una observación atípica, ¿cuál es? ¿Tiene dicho Estado una calificación media de Matemáticas más alta o más baja que la que se hubiera predicho a partir de su calificación media de Lengua?

Los siguientes ejercicios hacen referencia a la sección 2.6, que es opcional.

2.100. Aspirina y ataques al corazón. ¿Tomar aspirinas regularmente ayuda a prevenir los ataques al corazón? Un estudio (*Physicians' Health Study*) intentó averiguarlo, tomando como sujetos a 22.071 médicos sanos que tenían al menos 40 años. La mitad de los sujetos, seleccionados al azar, tomó una aspirina un día sí y otro no. La otra mitad tomó un placebo, una píldora falsa que tenía el mismo aspecto y sabor que una aspirina. He aquí los resultados:[46] (La fila "Ninguno de estos" se ha dejado fuera de la tabla.)

[46] Apareció en el 20 de julio de 1989 en el *New York Times*.

	Grupo de la aspirina	Grupo del placebo
Ataques al corazón mortales	10	26
Otro tipo de ataques al corazón	129	213
Embolias	119	98
Total	11.037	11.034

¿Qué indican los datos sobre la relación que existe entre tomar aspirinas, y los ataques al corazón y las embolias? Utiliza porcentajes para hacer más precisos tus razonamientos. ¿Crees que el estudio proporciona suficiente evidencia de que las aspirinas reducen los ataques al corazón (relación causa-efecto)?

2.101. Suicidios. He aquí una tabla de contingencia sobre los suicidios ocurridos en 1993, clasificados según el sexo de la víctima y el método utilizado. Basándote en estos datos, escribe un breve informe sobre las diferencias entre los suicidios de hombres y de mujeres. Asegúrate de que utilizas los recuentos y los porcentajes adecuados para justificar tus afirmaciones.

	Hombres	Mujeres
Arma de fuego	16.381	2.559
Veneno	3.569	2.110
Ahorcamiento	3.824	803
Otros	1.641	623

2.102. Permanecer vivo y fumar. A mediados de los años setenta, un estudio médico contactó al azar con gente de un distrito de Inglaterra. He aquí los datos sobre 1.314 mujeres que eran fumadoras habituales y mujeres que nunca habían fumado. La tabla clasifica a estas mujeres según su edad en el momento inicial de realización del estudio, según su situación con relación al tabaco y según si permanecían vivas al cabo de 20 años.[47]

	De 18 a 44 años		De 45 a 64 años		Mayores de 65 años	
	Fumadora	No fumadora	Fumadora	No fumadora	Fumadora	No fumadora
Fallecidas	19	13	78	52	42	165
Vivas	269	327	167	147	7	28

[47] D. R. Appleton, J. M. French y M. P. J. Vanderpump, "Ignoring a covariate: an example of Simpson's paradox", *The American Statistician*, 50, 1996, págs. 340-341.

(a) A partir de estos datos, construye una sola tabla de contingencia que relacione fumar (sí o no) con fallecer o vivir. ¿Qué porcentaje de fumadoras permaneció con vida durante 20 años? ¿Qué porcentaje de no fumadoras sobrevivió? Parece sorprendente que el porcentaje de mujeres que permaneció con vida fuera mayor entre las fumadoras.

(b) La edad de la mujer en el momento inicial de realización del estudio es una variable latente. Muestra que dentro de cada uno de los tres grupos de edad, el porcentaje de mujeres que permaneció con vida después de 20 años fue mayor entre las no fumadoras. Estamos ante otro ejemplo de la Paradoja de Simpson.

(c) Los autores del estudio dieron la siguiente explicación: "Entre las mujeres mayores (de 65 o más años al inicio del estudio), pocas eran fumadoras; sin embargo, muchas de ellas murieron durante el tiempo de seguimiento del estudio". Compara el porcentaje de fumadoras en cada uno de los tres grupos de edad para verificar esta explicación.