XVIII. CORRELACIÓN Y REGRESIÓN [*conclusión*]

EN EL presente capítulo proseguimos el examen de la correlación y la regresión. Se tratarán primero algunas pruebas de significación, a continuación de lo cual pasaremos a las relaciones no lineales, tema que se examinará asimismo brevemente en el capítulo XIX. A continuación estudiaremos los efectos de los errores de medición en las pendientes y las correlaciones. Finalmente, se examinará el tema de la correlación grado-orden.

XVIII.1. *Prueba de significación e intervalos de confianza*

Prueba de significación de r *y* b. Como quiera que r y los coeficientes de mínimos cuadrados a y b sólo describen los datos de las muestras, nuestro interés se centra por lo regular en los parámetros correspondientes de las poblaciones, ρ, α y β. En particular, desearíamos probar la hipótesis nula de que no hay relación (lineal) alguna en la población, o podemos querer obtener intervalos de confianza para ρ o para los coeficientes de regresión. Examinaremos primero la prueba de la hipótesis nula en el sentido de que no se da relación en la población. Según veremos, si podemos suponer una distribución normal de Y acerca de X y homoscedasticidad, podemos también servirnos del análisis de la variancia para verificar la hipótesis de que $\rho = \beta = 0$.

Sirvámonos del hecho de que, toda vez que r y b (y, por consiguiente, también ρ y β) tienen los mismos numeradores, una verificación de la hipótesis de que $\rho = 0$ lo es asimismo de la hipótesis $\beta = 0$ y viceversa. En otros términos: si no se da asociación lineal en la población, la pendiente de la ecuación de regresión será cero y, por tanto, la línea será horizontal. Recordando que la ecuación de regresión representa el camino de las medias de las Y para valores fijos de X, vemos inmediatamente que siempre que $\beta = 0$, las medias de las Y han de ser las mismas para todos los valores de X (véase figura XVIII.1). Esto implica, por supuesto, que la ecuación de regresión sea realmente de forma lineal. En particular, si dividiéramos el eje de las X en cierto número de categorías, encontraríamos que las medias de las categorías de la población son exactamente iguales. Así, pues, podemos traducir la hipótesis de que $\beta = \rho = 0$ en el enunciado de que las medias de Y serán iguales para cada una de las categorías de X. Si nos imaginamos una población infinita, como habrá que hacerlo para satisfacer el supuesto de normalidad, podemos concebir el eje de las X como dividido en un número indefinido de categorías, cada una de las cuales tenga medias idénticas en Y. En esta forma, nuestra hipótesis cero se con-

vierte en $\mu_{y^1} = \mu_{y^2} = \mu_{y^3} = \ldots$, en donde nos servimos del subíndice doble para recalcar que son las medias de las Y las que nos interesan y que tenemos un número indefinidamente grande de categorías X.

El curso del razonamiento anterior sugiere obviamente una extensión de la prueba de análisis de variancia para abarcar un

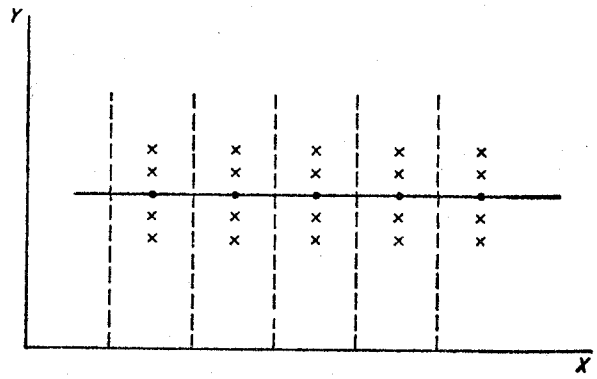

FIG. XVIII.1. *Representación geométrica del hecho de que la hipótesis de* $\beta = 0$ *es equivalente a la hipótesis* $\mu_1 = \mu_2 = \ldots = \mu_k$.

número indefinidamente grande de categorías de la variable de escala nominal (ahora X). Recordemos los supuestos requeridos en el análisis de variancia. Además de la hipótesis nula y del supuesto de que los casos se han muestreado aleatoria e independientemente de cada una de las categorías, hemos de suponer también poblaciones normales y variancias iguales dentro de cada categoría. A condición, pues, de que podamos suponer también muestreo aleatorio, vemos que todos estos supuestos pueden cumplirse si suponemos que la distribución conjunta de X y Y sea normal bivariable. El lector recordará que este último supuesto nos asegura simultáneamente una ecuación de regresión lineal, normalidad de las Y para cada valor fijo de X e iguales variancias para todos los valores de esta variable. De hecho, pues, los supuestos de muestreo al azar y de normalidad bivariable nos capacitan para servirnos del análisis de variancia con objeto de verificar la hipótesis de que $\rho = \beta = 0$, aun cuando no se requiere la normalidad de las X en tanto las ε_i tengan una distribución aproximadamente normal.

Anteriormente encontramos que era necesario obtener las sumas totales de cuadrados y la de entre clases y restarlas, con objeto de obtener la suma de cuadrados dentro. Sin embargo, al verificar la hipótesis de que $\rho = 0$, el proceso se simplifica con-

siderablemente. Ya vimos, en efecto, que la *proporción* de la suma de cuadrados total de la Y explicada por X nos es dada por r^2. Y en forma análoga, la proporción que dejamos inexplicada por X será $1 - r^2$. Como quiera que la suma total de cuadrados puede simbolizarse con Σy^2, las sumas de cuadrados explicada e inexplicada se convierten en $r^2\Sigma y^2$ y $(1 - r^2)\Sigma y^2$ respectivamente.

Los grados de libertad asociados a la suma total de cuadrados son, por supuesto, $N - 1$. Al calcular la suma inexplicada de cuadrados, tomamos la suma de las desviaciones al cuadrado respecto de la línea de mínimos cuadrados, y no respecto de la gran media de las Y. Pero, con objeto de obtener la línea de los mínimos cuadrados, hemos de servirnos de los dos coeficientes a y b. Por consiguiente, hemos perdido 2 grados de libertad, o sea uno más de los que perdimos al tomar las desviaciones respecto del valor particular de $\overline{Y}$. Podemos, pues, asociar $N - 2$ grados con la suma inexplicada de cuadrados y, restando, vemos que hay que asociar un grado de libertad a la suma de cuadrados explicada.

CUADRO XVIII.1. *Prueba de análisis de variancia de la hipótesis* $\rho = 0$

	Suma de cuadrados	*Grados de libertad*	*Apreciaciones de la variancia*	F
Total	Σy^2	$N - 1$		
Explicada	$r^2\Sigma y^2$	1	$\dfrac{r^2\Sigma y^2}{1}$	$\dfrac{r^2(N-2)}{(1-r^2)}$
Inexplicada	$(1 - r^2)\Sigma y^2$	$N - 2$	$\dfrac{(1-r^2)\Sigma y^2}{N-2}$	

Los resultados pueden resumirse ahora como en el cuadro XVIII.1. La ventaja de insertar símbolos en lugar de números en tabla está en que vemos inmediatamente que la cantidad Σy^2 desaparece cuando formamos la razón de las apreciaciones explicadas a las inexplicadas. En otros términos: la suma total de cuadrados se elimina, y podemos escribir una fórmula de F en términos de las *proporciones* de las sumas de cuadrados explicada e inexplicada. De este modo, la fórmula de F sólo comporta las cantidades r^2 y $1 - r^2$, junto con los grados de libertad de $N - 2$ y 1. Podemos, por consiguiente, servirnos de la fórmula:

$$F_{1,N-2} = \frac{r^2}{1 - r^2}(N - 2) \qquad \text{(XVIII.1)}$$

sin tener que ocuparnos en construir una tabla de análisis de

variancia, como fue el caso en el capítulo anterior. Como los cuadros para F sólo admiten pruebas a los niveles de .05, .01 y .001, puede resultar preferible tomar la raíz cuadrada positiva de (XVIII.1) y utilizar la distribución t, con $N - 2$ grados de libertad.

Podemos ilustrar el empleo de esta prueba de análisis de variancia para la significancia de r con los datos del cuadro XVII.1. Obtuvimos allí una correlación de $r = .301$ entre el porcentaje de negros y nuestro índice de discriminación. Al verificar en relación con el significado de r hacemos en realidad la importante pregunta: "¿Con qué probabilidad obtendríamos una r de .301 o mayor (en valor absoluto) si no hubiera efectivamente asociación lineal alguna en la población?" Con objeto de efectuar la prueba F, calculamos simplemente r^2 y $1 - r^2$ y nos servimos de la ecuación XVIII.1. Así, pues, ya que r se basaba en 13 casos, tenemos:

$$F_{1,11} = \frac{(.301)^2}{[1 - (.301)^2]}11 = \frac{.0906}{.9094}11 = 1.10$$

Refiriéndonos a la tabla F, vemos que para 1 y 11 grados de libertad necesitamos una F de 4.84 o mayor para descartar al nivel de .05 suponiendo que la dirección no hubiese sido establecida con anticipación. Decidimos, por consiguiente, no descartar la hipótesis nula de que $\rho = 0$. Aparentemente podríamos haber obtenido una r de .301 o mayor, simplemente por casualidad, aun si no se diera asociación alguna en la población.

Una vez más, es necesario insistir en la diferencia entre una prueba de significación y una medida del grado de relación. Si hubiéramos obtenido una r de .301 con un tamaño de muestra de 50, habríamos tenido:

$$F_{1,48} = \frac{.0906}{.9094}48 = 4.78$$

o sea un valor significativo al nivel de .05. En ambos casos hemos explicado aproximadamente el 9 por ciento de la variación total de la muestra, pero en el último de ellos tenemos más confianza, aunque ligeramente, de que se da una relación en la población.

Intervalos de confianza. Siempre que pueda presuponerse o apreciarse aproximadamente una población normal bivariable, es posible construir intervalos de confianza para ρ y β, así como la línea de regresión. El error estándar de r nos está dado por la fórmula.

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{N - 1}}$$

Por desgracia, la distribución de muestreo de r no será por lo regular simétrica, excepto en el caso especial en que $\rho = 0$. En efecto, la distribución de selección se distorsiona más y más a medida que el valor absoluto de ρ se aproxima a la unidad. Además, observamos que, para poder servirnos de la fórmula anterior del error estándar de r, necesitaríamos conocer o poder apreciar el valor de ρ. Estas dos complicaciones hacen que sea difícil obtener intervalos de confianza para ρ en forma abreviada.

Al calcular un intervalo de confianza respecto de r, convertimos primero r en una nueva estadística z que tiene una distribución de muestreo aproximadamente normal. Ponemos luego un intervalo de confianza alrededor de z en la forma habitual. Finalmente, una vez anotados los límites superior e inferior de confianza de z, reconvertimos estos valores particulares de z en r, con lo que obtenemos los límites de confianza de esta última.

Transformamos r en z por medio de la fórmula:

$$z = 1.151 \log \frac{1+r}{1-r}$$

en donde z puede tomar valores de cero al infinito. Conviene llamar la atención del lector acerca del hecho de que el valor z calculado mediante la fórmula anterior no tiene en absoluto conexión alguna con los valores de Z que utilizamos con la curva normal estándar. Los valores de z pueden obtenerse directamente del cuadro K, Apéndice 2, en lugar de servirse de los logaritmos. Los dos primeros dígitos de r se buscan de arriba abajo en el margen izquierdo, en tanto que el tercero se localiza horizontalmente en la parte superior. Los valores de z correspondientes están dados en el cuerpo del cuadro. Así, por ejemplo, una z de 0.3228 corresponde a una r de .312; una z de 1.3892 corresponde a una r de .883. Al servirnos del cuadro K, prescindimos del signo de r, asignando a z el signo correspondiente una vez hallado su valor numérico. Obsérvese que los valores de z sólo son ligeramente mayores que r cuando $|r| \leq .40$, pero a medida que r crece, z empieza a tomar valores mayores que la unidad.

Podemos servirnos ahora de la transformación de z en un problema de intervalo de confianza. La distribución de selección de z es aproximada a la normal, aun para N pequeñas y desviaciones moderadas de la normalidad bivariada. Su error estándar nos está dado por:

$$\sigma_z = \frac{1}{\sqrt{N-3}} \qquad \text{(XVIII.2)}$$

Y esto no sólo permite servirse de la tabla normal, sino que hemos eliminado además la necesidad de haber estimado ρ, ya que el error estándar de z sólo depende de N. Tomando como ejemplo numérico la correlación de .301 entre el porcentaje de negros y la discriminación, hallamos que el valor correspondiente de z es de 0.3106. Como quiera que no había más que 13 casos, tenemos:

$$\sigma_z = \frac{1}{\sqrt{13-3}} = \frac{1}{\sqrt{10}} = 0.3162$$

Supóngase que deseamos obtener para ρ un intervalo de confianza del 95 por ciento. Primero calculamos dicho intervalo en términos de valores de z. Así, pues, tomaríamos:

$$z \pm 1.96\sigma_z = 0.3106 \pm 1.96(0.3162)$$
$$= 0.3106 \pm 0.6198$$

Por consiguiente, el intervalo de confianza alrededor de z va de $-.3092$ a $+.9304$. Obsérvese que para obtener el límite inferior tuvimos que restar un número mayor, numéricamente, que 0.3106. Esto da un resultado negativo, lo cual significa a su vez que el valor de r correspondiente a dicho límite inferior ha de tomarse también como negativo. Buscando los valores de r correspondientes a los dos límites de confianza de z, obtenemos los valores de $-.300$ y .731 para los límites inferior y superior respectivamente.

Obsérvese que el intervalo no es totalmente simétrico en relación con el valor de .301 obtenido para r. En este caso, el límite superior está algo más cerca de r que el límite inferior. Si hubiéramos hallado una r de .80, el intervalo resultante habría estado todavía más distorsionado en la misma dirección. Puede comprenderse intuitivamente que esto sea así si tenemos presente que, siempre que empezamos a acercarnos al límite superior de la unidad, ponemos también una restricción al límite superior del intervalo de confianza. En esta forma, resultaría imposible, por ejemplo, obtener un intervalo de confianza de .86 ± .16. Si ocurre que r sea negativa, la dirección de la distorsión será opuesta, por supuesto, a la anterior. El intervalo solamente llegará a ser simétrico en relación con r cuando ésta sea igual a cero.

Podemos interpretar este intervalo de confianza en la forma habitual. Nuestro procedimiento es tal que a la larga podemos esperar obtener intervalos que incluyan el valor (fijo) de ρ el 95 por ciento de las veces. Podemos también utilizar tales intervalos de confianza como verificaciones implícitas de hipótesis. En el problema anterior, en efecto, ya hemos observado que el

límite inferior del intervalo es negativo. Y como quiera que cero está incluido en el intervalo, sabemos inmediatamente que no descartaríamos la hipótesis nula de que $\rho = 0$. Y si quisiéramos verificar algún otro valor supuesto de ρ, procederíamos igual. Si por ejemplo hubiéramos anticipado que $\rho = .80$, habríamos descartado al nivel de .05, ya que este valor cae fuera del límite superior de .731.

Sería conveniente también calcular intervalos de confianza a propósito de otras medidas de grados de relación. Por desgracia, se conoce demasiado poco acerca de las distribuciones de muestreo de la mayoría de las medidas de asociación en materia de problemas de contingencia para poder construir intervalos de confianza en relación con ellas. Haggard [11] sugiere un método para computar intervalos de confianza acerca de r_i o correlación interclase, y Goodman y Kruskal [10] discuten la distribución de muestras de varias medidas nominales y ordinales.

Ocasionalmente se quiere poder poner un intervalo de confianza con referencia a b, o se puede tener necesidad de encontrar un cinturón a cuyo interior pueda esperarse que la verdadera ecuación de regresión se encuentre. En ambos casos podemos servirnos de la distribución t en forma relativamente directa. La apreciación del error estándar de b está dada por:

$$\hat{\sigma}_b = \frac{\hat{\sigma}_{y|x}}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2}} \qquad \text{(XVIII.3)}$$

en donde se recordará que:

$$\hat{\sigma}_{y|x} = \sqrt{\sum_{i=1}^{N} \frac{(Y_i - Y_\rho)^2}{N-2}}$$

Con fines de cálculo puede demostrarse algebraicamente que:

$$\hat{\sigma}_{y|x} = \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2 - b \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N-2}} \qquad \text{(XVIII.4)}$$

Podemos servirnos ahora de los cálculos numéricos obtenidos ya para los datos de discriminación del cuadro XVII.1, con lo que obtenemos:

$$\hat{\sigma}_{y|x} = \sqrt{\frac{560\,024 - 19.931(2\,553.77)}{11}} = \sqrt{46\,284} = 215.1$$

y

$$\hat{\sigma}_b = \frac{215.1}{\sqrt{128.131}} = \frac{215.1}{11.32} = 19.00$$

Si deseamos calcular el intervalo de confianza del 99 por ciento, recurrimos directamente a la tabla t y nos servimos de $N-2$ o 11 grados de libertad. Obtenemos en esta forma:

$$b \pm (3.106)(19.00) = 19.931 \pm 59.014$$

* Al apreciar la ecuación de regresión, vemos que nuestra mejor apreciación singular (de "punto") es la línea de los mínimos cuadrados. Como quiera que la cantidad que estamos apreciando ahora ya no es un valor singular, sino una línea entera, nuestra *apreciación del intervalo* ya tampoco será un intervalo, sino una banda a ambos lados de la línea de mínimos cuadrados. De buenas a primeras podría esperarse que dicha banda consistiera en dos líneas paralelas a la de los mínimos cuadrados. Sin embargo, semejante banda implicaría que conocemos la verdadera pendiente y que la única fuente de error está en la apreciación de α. Hemos de recordar que se aprecian ahora dos cantidades (α y β), y, por lo tanto, tenemos dos fuentes de error. El lector ha de percatarse por sí mismo de que toda vez que la pendiente puede habese apreciado asimismo incorrectamente, cuanto más nos vamos alejando del punto $(\bar{X}, \bar{Y})$, tanto mayor resulta la imprecisión. La banda de confianza adopta la forma general de la figura XVIII.2.

* Para trazar esta banda de confianza, será necesario calcular el error estándar de Y_p para varios valores de X. La apreciación del error estándar nos está dada por la fórmula:

$$\hat{\sigma}_{y_p} = \hat{\sigma}_{y|x} \sqrt{\frac{1}{N} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{N} (X_i - \bar{X})^2}} \qquad \text{(XVIII.5)}$$

en donde el valor particular de X a utilizar en $(X - \bar{X})^2$ puede ponerse en cualquier lugar del eje de las X. Obsérvese, de paso, que cuanto más lejos X queda de $\bar{X}$, tanto mayor es el valor numérico del error estándar. Supóngase que deseamos obtener el error estándar estimado cuando $X = 10.0$. Como quiera que $\bar{X}$ 4.837, obtenemos:

$$\hat{\sigma}_{y_p} = 215.1 \sqrt{\frac{1}{13} + \frac{(10.0 - 4.837)^2}{128.131}} = 215.1 \sqrt{.28496} = 114.86$$

* Sirviéndonos nuevamente de la tabla t y de un intervalo del 99 por ciento respecto de Y_p *calculado para este valor fijo de X*, obtendríamos:

$$Y_p \pm (3.106)(114.86) = Y_p \pm 356.8$$

Una vez que hayamos obtenido otros intervalos semejantes de Y_p para otros valores particulares de X, podemos trazar la gráfica de la banda entera. Inútil es decir que el procedimiento en cuestión se haría muy fastidioso si se deseara obtener la banda entera y no se contara con calculadoras.

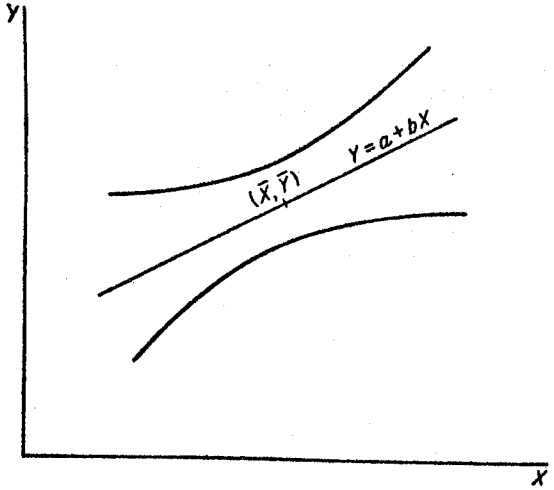

FIG. XVIII.2. *Banda de confianza con respecto de la recta de mínimos cuadrados.*

Probando la diferencia entre dos correlaciones. Como se indicó antes, tiene habitualmente más sentido teórico el comparar dos o más declives que el comparar correlaciones; tal comparación entre declives ocupará nuestra atención en el capítulo XX sobre análisis de covariancia. Sin embargo, ocurre con frecuencia que se han obtenido varias correlaciones y se desea establecer que una de ellas es significativamente más alta que las demás. Mientras nos contentamos en describir relaciones dentro de nuestra muestra particular, podemos comparar simplemente las magnitudes relativas de las dos r y registrar la magnitud de la diferencia. Sin embargo, si deseamos generalizar a una población mayor, plantéase la cuestión de si la diferencia obtenida pueda o no deberse acaso al azar. Supóngase, por ejemplo, que se han obtenido una r de .50 y otra de .30. Puede desearse verificar la hipótesis nula de que las dos correlaciones de las poblaciones son idénticas, esto es, $\rho_1 = \rho_2$.

Cabe imaginar dos situaciones distintas en las que podrían hacerse verificaciones de esta clase. Primero, pueden acaso tenerse *dos muestras independientes* y desearse comparar los grados de relación entre X y Y y dentro de cada una de las muestras. Así, por ejemplo, la relación entre el porcentaje de negros y la discriminación puede acaso no ser la misma en los estados del Sur que en los del Norte. Podría en este caso establecerse la hipótesis de investigación de que ρ_{xy} es más alta en el Sur que en el Norte, verificando la hipótesis nula de que las dos correlaciones son iguales. Un segundo tipo de situación, fácil de confundir con el primero, puede presentarse cuando se dispone de *una sola muestra*. Puede haber en este caso una sola variable dependiente (por ejemplo, la discriminación) y dos variables independientes (por ejemplo, el porcentaje de negros y el porcentaje de mano de obra empleada en la industria). Puede acaso desearse establecer que una de estas variables independientes está más directamente relacionada con la variable dependiente que la otra. Si designamos la segunda variable independiente como Z, podemos tener interés en verificar la hipótesis nula de que $\rho_{xy} = \rho_{zy}$. Veamos primeramente cómo tratamos el primer tipo de situación, para pasar luego a la prueba de una sola muestra.

Si las dos correlaciones se basan en muestras independientes, podemos convertir cada una de las r en z y servirnos de la fórmula del error estándar de la diferencia entre las z, que es análoga a la del error estándar de una diferencia entre medias y se presenta como sigue:

$$\sigma_{z_1 - z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} \qquad \text{(XVIII.6)}$$

Podemos a continuación ya sea establecer un intervalo de confianza relativo a $(z_1 - z_2)$ o buscar el valor de:

$$Z = \frac{(z_1 - z_2) - 0}{\sigma_{z_1 - z_2}}$$

en la tabla normal. El cero figura en la fórmula anterior debido al hecho de que nuestra hipótesis nula adopta la forma $\rho_1 = \rho_2$.

Supóngase que para 17 ciudades del Sur la correlación entre el porcentaje de negros y la discriminación resulta ser de .567, frente a la de .301 de las ciudades del Norte. Así, pues:

$$r_1 = .301 \qquad r_2 = .567$$
$$z_1 = 0.3106 \qquad z_2 = 0.6431$$

y

$$\sigma_{z_1-z_2} = \sqrt{1/10 + 1/14} = \sqrt{.1000 + .0714} = .414$$

Por lo tanto:

$$Z = \frac{.3106 - .6431}{.414} = \frac{-.3325}{.414} = -.803$$

y vemos que esta diferencia de las r no es significativa al nivel de .05. Así, pues, pese a que la correlación sea mayor por lo que se refiere a las ciudades del Sur, esta diferencia puede deberse simplemente al azar.

En el segundo tipo de situación mencionado, no disponemos de dos muestras independientes y no podemos, por consiguiente, servirnos de la misma fórmula del error estándar de $z_1 - z_2$. Se dispone de un método para tratar este tipo de problema, a condición que sólo tengamos interés en generalizar a subpoblaciones de todas las muestras posibles para los que X y Z (las dos variables independientes) tienen las mismas combinaciones de valores que las de la muestra particular que hemos obtenido. En la mayoría de los casos prácticos puede prescindirse impunemente de esta restricción, a menos que exista alguna razón para suponer que el margen de variación es mucho mayor en la población que en la muestra estudiada, en cuyo caso deberemos de todos modos guardarnos de generalizar en un sentido o en otro.

Si verificamos la hipótesis nula de que $\rho_{xy} = \rho_{zy}$, formamos t de la manera siguiente:

$$t = (r_{xy} - r_{zy}) \sqrt{\frac{(N-3)(1+r_{xz})}{2(1 - r_{xy}^2 - r_{xz}^2 - r_{zy}^2 + 2r_{xy}r_{xz}r_{zy})}} \qquad \text{(XVIII.7)}$$

Podemos buscar luego el valor de t en el cuadro, sirviéndonos de $N-3$ grados de libertad. En nuestro ejemplo numérico, supóngase que la correlación entre X y Z para las ciudades del Norte resulta ser de .172 y que la correlación entre Y y Z es de .749. Tendríamos en esta forma:

$$t = (.301 = .749) \sqrt{\frac{10(1 + .172)}{2[1 - .301^2 - .172^2 - .749^2 + 2(.301)(.172)(.749)]}}$$
$$= -1.72.$$

Como tenemos 10 grados de libertad, vemos que no podemos descartar la hipótesis nula de que no hay diferencia entre las

correlaciones de las poblaciones de cada una de las variables independientes con discriminación.

XVIII.2. *Correlación no lineal y regresión*

Hasta aquí hemos venido suponiendo que la ecuación de regresión era de forma lineal. En muchos problemas sociológicos prácticos, el modelo lineal, aunque tal vez no exacto, da con todo una aproximación bastante cercana a la forma verdadera de la ecuación, de modo que no necesitamos ocuparnos de modelos alternativos más complicados. Esto es así, en particular, en relación con los estudios de exploración en los que el grado de adaptación no es excesivamente exacto. Hay casos, sin embargo, en los que la inspección del diagrama de dispersión podrá indicar claramente una relación no lineal, o en los que nuestra teoría ha anticipado una relación de esta clase. Siempre que se dé una relación no lineal semejante, el coeficiente momento-producto dará obviamente una subestimación del grado verdadero de relación, ya que este coeficiente sólo mide el grado de adaptación de la mejor recta singular. Ya vimos que con una curva en forma de U es posible tener una fuerte relación con una r de aproximadamente cero, y se advirtió al lector que era, por lo tanto, incorrecto sacar la conclusión de que dos variables son independientes simplemente porque r sea cero. Si el diagrama de dispersión indica una distribución de puntos más o menos al azar, podemos concluir que no existe relación, pero hemos de estar al acecho al propio tiempo de las relaciones no lineales. Ésta es, por supuesto, una razón más en favor de que el lector debe acostumbrarse a trazar siempre diagramas de dispersión antes de seguir adelante con el análisis.

El tema general de la correlación y la regresión no lineales es demasiado complejo para poder tratarlo adecuadamente en este texto. La razón de la complejidad del análisis no lineal está en que, una vez que progresemos más allá de la ecuación de la recta, hay numerosos tipos de ecuaciones que representan las distintas formas posibles susceptibles de ser adoptadas por las relaciones no lineales. Sólo las más simples de estas ecuaciones pueden tratarse aquí. Afortunadamente, estas ecuaciones relativamente sencillas suelen ser por lo regular adecuadas para la solución de las clases de relaciones que se plantean en la investigación sociológica. Un tipo general de función no lineal puede representarse en términos de polinomios de grado enésimo, que tienen ecuaciones de la forma:

$$Y = a + bX + cX^2 + dX^3 + \ldots + kX^n$$

El examen de las relaciones no lineales de este tipo general lo

dejaremos hasta el próximo capítulo, o sea hasta el momento de emprender el estudio de los problemas de regresión múltiple. En efecto, una vez comprendidos estos problemas de regresión, dispondremos de un método relativamente simple para el tratamiento de aquellos tipos de relaciones no lineales que se dejan describir adecuadamente por medio de polinomios.

Algún otro tipo de relaciones no lineales relativamente sencillo puede tratarse a menudo mediante una transformación de variables que permite el empleo del modelo lineal familiar. Este proceso puede ilustrarse con el caso de las funciones logarítmicas representadas por ecuaciones del tipo:

$$Y = a + b \log X$$

que presentan la forma general de la figura XVIII.3. En una ecuación de este tipo, en efecto, Y es en realidad una función lineal no de la X misma, sino de su logaritmo. Esto sugiere que si podemos transformar cada una de las marcas de X en una nueva variable $Z = \log X$, podemos escribir Y como función lineal de Z. Así, por ejemplo:

$$Y = a + b \log X = a + bZ$$

Podemos calcular ahora la correlación entre Y y Z (o sea de Y y de log X) en la forma habitual. Si damos a conocer la distribución de las marcas a los ejes de las Y y las Z, el resultado habrá de ser aproximadamente de forma lineal. Si queremos, podemos comparar el grado de relación entre Y y Z con el que existe entre Y y X. Si r_{yz} es significativamente mayor que r_{xy}, entonces el modelo logarítmico da una mejor aproximación que el modelo lineal entre X y Y.

Los modelos logarítmicos del tipo anterior se presentan a menudo en casos en que la variable independiente X asume un gran margen de valores, pero en los que, una vez alcanzado cierto valor, los aumentos ulteriores producen cada vez menos efecto sobre la variable dependiente. La magnitud de una ciudad es una variable que presenta con frecuencia esta clase de efecto. Es posible, por tanto, que las ciudades de más de 500 mil habitantes presenten todas ellas marcas de Y muy parecidas. Pero, si se incluye en la muestra a la ciudad de Nueva York, por ejemplo, el valor de X para esta ciudad será tan superior al de las demás ciudades, que el efecto neto consistirá en inclinar la relación en forma muy parecida a la de la figura XVIII.3. En tal caso podrá resultar preferible relacionar Y con log X, ya que el hecho de tomar el logaritmo de la magnitud urbana producirá el efecto de agrupar las marcas extremadamente grandes y de disminuir el "efecto de curvatura" de estas ciudades mayores.

En cierto número de casos el investigador no tendrá tal vez interés en hallar la forma exacta de la ecuación de predicción que mejor se adapte a sus datos. Acaso sólo trate, por ejemplo, de demostrar que la relación es de forma no lineal, o de obtener una medida para el grado de relación, independientemente de su forma. Cuando pueda efectuarse una transformación sencilla

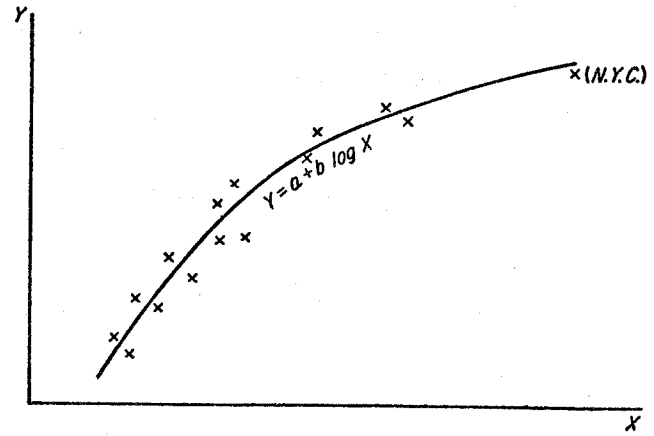

FIG. XVIII.3. *Ecuación logarítmica de mínimos cuadrados de la forma* $Y = a + b \log X$.

como la logarítmica, será indudablemente ventajoso servirse de dicho procedimiento. Pero aun así, el investigador querrá acaso verificar si la medida que ha obtenido constituye o no una buena aproximación del resultado que habría hallado si se hubiera encontrado la mejor adaptación posible. Con objeto de tratar los problemas de esta índole, podemos servirnos de los principios básicos del análisis de variancia y de algunas de las medidas de los grados de asociación desarrolladas en el capítulo sobre análisis de variancia.

El lector recordará que para obtener la suma de cuadrados "dentro" en el análisis de variancia de una forma tomamos la suma de las desviaciones al cuadrado de cada una de las medias de las categorías. Supongamos ahora que las X se han subdividido en cierto número de categorías y que la suma de los cuadrados en Y se analizaban en la forma habitual. Sabemos que para toda categoría dada de X la suma de los cuadrados alrededor de la media de la categoría producirá un resultado numérico inferior al de la suma de los cuadrados alrededor de cualquier otro número. Síguese, en particular, que la suma interior de cuadrados será menor que la suma de las desviaciones cuadradas respecto de aquellos puntos de la línea de mínimos cua-

drados que caen en los puntos medios de los intervalos (véase la figura XVIII.4).

Si ocurre que la ecuación sea de forma lineal, podemos esperar que $\overline{Y}_{.j}$ caerá aproximadamente en la línea de los mínimos cuadrados, de modo que cambiará poco que las desviaciones se tomen respecto de las medias de las categorías o respecto de la lí-

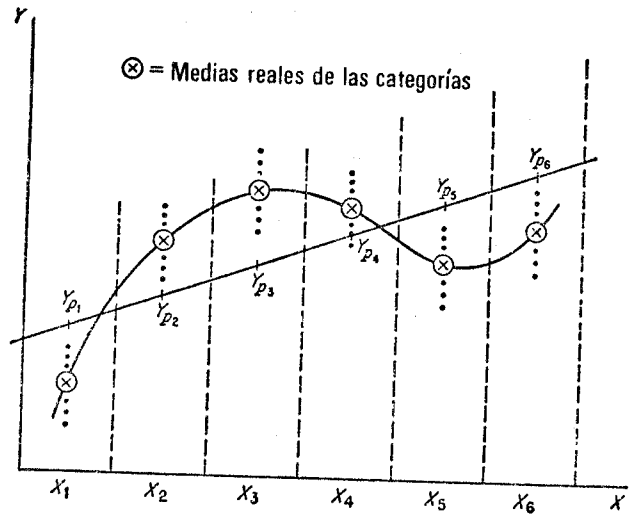

FIG. XVIII.4. *Comparación de las desviaciones respecto de la recta de mínimos cuadrados con las desviaciones respecto de las medias de las categorías.*

nea en cuestión. Por otra parte, si la ecuación es realmente no lineal, entonces, para algunas, al menos, de las categorías, la suma de los cuadrados referidos a la media de la categoría será considerablemente más pequeña que la de los cuadrados referidos a la línea de los mínimos cuadrados. En otros términos: la suma interior o inexplicada de cuadrados será mínima empleando las medias de las categorías y, por consiguiente, la suma de cuadrados entre categorías, o explicada, será máxima. Así, pues, la proporción de variación explicada por las categorías, medida por la razón de correlación E^2, será mayor que la proporción explicada por la línea de mínimos cuadrados, a menos que la verdadera relación sea efectivamente lineal.

Podemos sacar utilidad de este hecho practicando una prueba de no linealidad. Si formamos la cantidad $E^2 - r^2$, obtenemos la proporción de variación explicada en el supuesto de una forma cualquiera de relación no explicada por una relación lineal. Es obvio que para obtener E^2 permitimos que la relación adopte

cualquier forma posible, ya que sólo hemos tomado desviaciones respecto de las medias de las categorías, prescindiendo de dónde estas medias acontezcan encontrarse. Nos estamos preguntando fundamentalmente en cuánto podemos mejorar nuestra posibilidad de predecir valores de Y no restringiéndonos al modelo lineal. Si la mejora es mayor de lo que esperaríamos del azar

CUADRO XVIII.2. *Prueba de análisis de variancia para el caso de no linealidad*

	Sumas de cuadrados	*Grados de libertad*	*Estimaciones de la variancia*	*F*
Total	Σy^2	$N-1$		
Explicada por el modelo lineal	$r^2\Sigma y^2$	1		
Adicional, explicada por el modelo no lineal	$(E^2-r^2)\Sigma y^2$	$k-2$	$\dfrac{(E^2-r^2)\Sigma y^2}{k-2}$	$\dfrac{(E^2-r^2)(N-k)}{(1-E^2)(k-2)}$
Inexplicada	$(1-E^2)\Sigma y^2$	$N-k$	$\dfrac{(1-E^2)\Sigma y^2}{N-k}$	

en el supuesto de que la ecuación de regresión sea efectivamente lineal, entonces podemos concluir que la relación es no lineal.

La prueba de análisis de variancia que emplearemos para verificar la no linealidad asume una forma con la que no tardaremos en familiarizarnos. Hallamos primero la cantidad de variación que puede explicarse sirviéndonos del modelo lineal. Algebraicamente, esta cantidad puede representarse como $r^2\Sigma y^2$. De la variación dejada sin explicar por el modelo lineal, $(1 - r^2)\Sigma y^2$, vemos a continuación qué tanto pueda explicarse por medio del modelo general. Como quiera que $E^2\Sigma y^2$ nos da la suma de cuadrados que puede explicarse por Y cuando no pesa sobre la forma de la relación restricción alguna, la cantidad $(E^2 - r^2)\Sigma y^2$ representa el incremento explicado debido a la no linealidad. Suponiendo que no haya errores de redondeo, esta cantidad habrá de ser siempre positiva.[1] Y como quiera que la cantidad $(1 - E^2)\Sigma y^2$ nos da la suma de cuadrados que no resulta explicada ni siquiera por el modelo mejor adaptado, podemos efectuar una prueba F tal como se indica en el cuadro XVIII.2. Como de costumbre, el denominador de F es el término de error y, como

[1] Siempre que N sea pequeña y sólo pueda, por tanto, usarse un corto número de categorías, resulta poco realista el supuesto de que las puntuaciones de X están agrupadas en los puntos medios de cada intervalo. Esto puede llevar a agrupar los errores, dando un valor a E^2 menor que r^2.

quiera que estamos verificando en relación con desviaciones respecto de la linealidad, tomamos como numerador una apreciación de la variancia basada en $(E^2 - r^2)\Sigma y^2$, o sea la cantidad explicada por el modelo general mejor, que no ha sido explicada todavía por el modelo lineal. Los grados de libertad asociados al numerador pueden obtenerse por sustracción.

Una vez más observamos que la suma total de cuadrados se elimina, dejándonos con la siguiente fórmula para F:

$$F_{k-2,N-k} = \frac{(E^2 - r^2)(N-k)}{(1-E^2)(k-2)} \qquad \text{(XVIII.8)}$$

en donde k representa el número de categorías en las que se ha descompuesto X.

Ilustremos la prueba de no linealidad con los datos que se agruparon en el cuadro XVII.2. Según puede comprobarse fácilmente, las sumas total y entre categorías de cuadrados en Y son como sigue:

$$\text{SC total} = 101\,115.38 - 92\,132.04 = 8\,983.34$$
$$\text{SC entre categorías} = 94\,792.59 - 92\,132.04 = 2\,660.55$$

en donde hemos tratado todas las marcas de Y como si se encontraran en los puntos medios de sus respectivos intervalos y en donde nos hemos servido de los procedimientos para los datos agrupados (véase sec. VI.4). Por lo tanto:

$$E^2 = \frac{\text{SC entre cuadrados}}{\text{SC total}} = \frac{2\,660.55}{8\,983.34} = .2962$$

Toda vez que anteriormente encontramos una r de $-.460$ suponiendo una relación lineal, obtenemos:

$$F_{7,141} = \frac{.2962-(-.460)^2}{1-.2962}\,\frac{150-9}{9-2} = \frac{.0846}{.7038}\,\frac{141}{7} = \frac{11.929}{4.927} = 2.42$$

y vemos que al nivel de .05 podemos descartar la hipótesis nula de una relación lineal entre el porcentaje de personas clasificadas como trabajadoras de granjas rurales y el porcentaje de mujeres que trabajan en la industria.

Si una relación resulta ser no lineal en cuanto a la forma, es muy posible que r no sea significativa estadísticamente, en tanto que E sí lo será. Por supuesto, la significación de E puede comprobarse por medio de un análisis directo de variancia, tomando la razón de las estimaciones explicada e inexplicada de la variancia. Son, pues, así tres las pruebas que pueden efectuarse,

a saber: *1)* la de la significación de r; *2)* la de la significación de las desviaciones respecto de la linealidad $(E^2 - r^2)$, y *3)* la de la significación de E.

Si se encuentra una relación no lineal y se desea una estimación del grado de relación en la población, es preferible servirse de la razón de correlación insesgada ε, examinada en el capítulo XVI y dada por la fórmula:

$$\varepsilon^2 = 1 - \frac{V_w}{V_t}$$

ya que el valor numérico de E es función del número de categorías empleadas y probablemente sobrestimará ligeramente por lo regular la relación en la población. Si ya se ha calculado E, el valor de ε puede también calcularse a partir de la fórmula:

$$\varepsilon^2 = \frac{E^2(N-1) - (k-1)}{N-k} \qquad \text{(XVIII.9)}$$

XVIII.3. *Efectos de los errores de medición*

Si hay mediciones de error en X o Y, bien sean al azar o sistemáticas, puede esperarse una alteración en nuestros resultados. Esto se aplica por supuesto a *todas* las pruebas y mediciones que hemos examinado hasta ahora, incluso los procedimientos no paramétricos. En realidad, uno de los tipos de errores de medición más comunes en sociología, ciencia política y la mayoría de las restantes ciencias sociales, parecería ser consecuencia del uso de dicotomías más bien burdas, tales como *alto* y *bajo* o *presente* y *ausente*. No se comprenden bien las consecuencias que se derivan de los errores de medición, pero la mayor parte del trabajo sistemático sobre el tema se ha llevado a cabo en las escalas de intervalo y en los problemas que implican análisis de correlación y regresión. El tema es por desgracia demasiado técnico para ser tratado en el presente texto, pero resultará conveniente pronunciar por lo menos algunas palabras precautorias.

Si hay una medición de error sistemática, o no aleatoria, cualquier tipo de distorsión resulta posible, siendo así necesario explicar cuáles son las fuentes del error no aleatorio y la forma en que actúan. Si se comparan por ejemplo las medias de tres muestras, y el error de medición es tal que coloque las medias de las muestras segunda y tercera cercanas a la correspondiente a la primera, no se logrará significación estadística cuando, con base en mediciones más exactas, pueda rechazarse fácilmente la hipótesis nula. Pero si los errores de medición son estrictamente al azar, resultará posible tener una mayor claridad acerca de los

efectos de tales errores. En general, las medidas de asociación resultarán atenuadas por los errores aleatorios de medición en cualquier variable. Por ejemplo, en el análisis de las situaciones de variancia, las mediciones aleatorias de error en la escala de intervalos aumentarán las variaciones *dentro* de las categorías, pero no afectarán sistemáticamente las variaciones entre las categorías, lo que hará bajar tanto el valor de F como la correlación interclases.

En el caso de dos escalas de intervalo los errores aleatorios de medición en cualquier variable reducirán la magnitud del coeficiente de correlación. En algunos textos elementales de estadística se examinan los procedimientos correctivos de atenuación, pero se hace basándose en supuestos especiales, inapropiados para uso en la investigación sociológica. (Véase [3].) En general, cuando se cuenta con dos o más medidas de cada variable, resulta posible obtener estimaciones corregidas bajo grupos variables de supuestos. (Véanse [2], [6] y [14].)

Si hay errores aleatorios de medición en Y pero no en X, podemos concebir la situación como una contribución que alcanza sólo al factor de error en la ecuación $Y_i = \alpha + \beta X_i + \varepsilon_i$, pudiendo demostrarse que no habrá efecto sistemático en la estimación b_{yx} del declive, salvo que el error estándar en tal estimación se verá incrementado debido al aumento del error en la variancia. Pero si hay también error aleatorio de medición en X —lo que es muy posible en toda investigación realista—, la estimación b_{yx} del declive se verá asimismo atenuada. En el caso de muestras grandes puede aplicarse una fórmula aproximada para determinar el valor esperado del declive b_{yx}:

$$E(b_{yx'}) = \beta \frac{\sigma_x^2}{\sigma_{x'}^2} = \beta \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}$$

en la que X' representa el valor medido de X, tal como se le representa en la ecuación $X' = X + u$, en donde se supone a u como un componente estrictamente aleatorio, con valor esperado igual a cero, y sin que haya correlación entre u y X. La razón de la atenuación estriba en que la variancia del valor medido X' será mayor que la variancia verdadera de X, según la fórmula:

$$\sigma_{x'}^2 = \sigma_x^2 + \sigma_u^2$$

Vemos así que la atenuación en la estimación de un desnivel es función de la variancia del error de medición, *relativa* a la variancia en X.

Este hecho tiene consecuencias importantes en la práctica. Significa que en cuantos casos haya error aleatorio de medición en una variable independiente, no podemos contar con iguales declives estimados, incluso en el caso de que los declives verdaderos lo sean. Si varias poblaciones (o muestras) difieren con respecto a la cantidad de variación en X, incluso con las mismas variancias de error de medición, las atenuaciones de los declives diferirán. Vale la pena tener esto presente cuando se llevan a cabo comparaciones de los resultados de diferentes estudios. La dificultad señalada se aplica también a todas las medidas de asociación, y no puede ser considerado como un defecto privativo del análisis de regresión.

XVIII.4. *Escalas ordinales: correlación de rangos*

Nos hemos ocupado ahora de medidas de asociación que pueden utilizarse para relacionar dos escalas nominales (ϕ^2, τ_b, etcétera), una escala nominal y una de intervalo (correlación intraclase), y dos escalas de intervalo (r). Las tres medidas que vamos a examinar en esta sección, o sean la r_s de Spearman y la tau y la gamma de Kendall, pueden emplearse para relacionar entre sí dos escalas ordinales. A condición que las dos variables pueden alinearse, cualquiera de estas últimas medidas puede emplearse para dar correlaciones que son algo parecidas a las del momento-producto.

Las medidas ordinales examinadas en esta sección resultan apropiadas cuando la relación entre X y Y es la que se denomina *monotónica en aumento* o bien *monotónica en disminución*. La idea de linealidad es desde luego inapropiada en el caso de las escalas ordinales, como lo es también la idea de una distancia entre valores de X (o de Y). Podemos, sin embargo, hablar de relaciones que se encuentran en aumento (o disminución) constante. Una función de aumento monotónico es aquella que o bien crece siempre o permanece constante, a medida que X aumenta. En otras palabras: cuando X aumenta, Y no disminuye. Una función lineal constituye un caso especial de una función monotónica de aumento (o disminución), pero también lo es una función logarítmica tal como $Y = a + b \log X$. Reconocemos dos clases de relación no lineal, a saber: las que son monotónicas y las que no lo son. El último tipo de relación no lineal tendrá por supuesto una o más curvaturas o inversiones de dirección, como ejemplifica una parábola o ecuación de tercer grado.

Con frecuencia encontramos proposiciones teóricas de la forma "cuanto mayor la X, mayor la Y (o menor la Y)". Estas afirmaciones quieren decir que la relación entre X y Y es monotónica, pero no especifican en qué forma. Las medidas ordinales resultan apropiadas cuando se trata de proposiciones de esta naturaleza. Sería por supuesto preferible refinar nuestras teorías, de modo que se especificase si existe linealidad o alguna clase

particular de no linearidad (por ejemplo, logarítmica), pero si la medición no ha superado el nivel ordinal, resultará imposible distinguir empíricamente entre alternativas lineales o no lineales. (Véase [22].)

La r_s de Spearman. El principio que se halla en la base de la medida de Spearman es muy simple. Comparamos la ordenación de dos grupos de marcas tomando las diferencias de los rangos, cuadrándolas y luego adicionándolas, y tratando finalmente dicha medida de modo que su valor sea +1.0, siempre que los órdenes estén perfectamente de acuerdo, −1.0 si los órdenes discrepan totalmente, y cero si no se da relación alguna. Si simbolizamos la diferencia entre dos lugares cualesquiera como D_i, hallamos el valor de $\sum_{i=1}^{N} D_i^2$ y calculamos r_s por medio de la fórmula:

$$r_s = 1 - \frac{6 \sum_{i=1}^{N} D_i^2}{N(N^2 - 1)} \qquad \text{(XVIII.10)}$$

Esta fórmula para r_s se obtiene tomando la fórmula para una correlación momento-producto y aplicándola a rangos y no a puntuaciones brutas, pudiendo así interpretar la medida de Spearman como la correlación momento-producto entre los rangos de X y los de Y.

Ilustrémosla con algunos datos reunidos por el autor. Los miembros de un campamento de trabajo fueron ordenados de superior a inferior desde los puntos de vista de la popularidad, medida por las amistades y de la participación en las discusiones de grupo. Para ambas variables el orden de clasificación de uno significa una marca elevada. Los órdenes empatados se calculan atribuyendo a cada marca empatada la media aritmética de la puntuación que habría recibido si no hubiera empates. Los valores de D_i se calculan a continuación, tal como se indica en el cuadro XVIII.3. Si el número de empates es pequeño, como en el presente caso, no necesitamos introducir modificación en la fórmula de r_s. Pero si el número de empates es considerable, entonces puede calcularse un factor de corrección (véase [19], pp. 215-220). Obtenemos, pues:

$$r_s = 1 - \frac{6(207.50)}{16(255)} = 1 - .305 = .695$$

Obsérvese que si las clasificaciones concuerdan perfectamente, $\sum_{i=1}^{N} D_i^2$ será cero, y el valor de r_s será la unidad. Si bien la inspección directa de la fórmula no nos da inmediatamente los valores de r_s para la independencia y la asociación perfectamente negativa, resulta que para la asociación negativa perfecta el valor del segundo término será de −2.0 y, por lo tanto, r_s será −1.0.

CUADRO XVIII.3. *Cálculo del coeficiente de Spearman de la correlación de rango*

Personas	*Orden de popularidad*	*Orden de participación*	D_i	D_i^2
Ann	1	5.5	4.5	20.25
Bill	2.5	5.5	3.0	9.00
Jim	2.5	1	—1.5	2.25
Hans	4	2	—2.0	4.00
Marcia	5	3	—2.0	4.00
Joan	6	9.5	3.5	12.25
Ruth	7	5.5	—1.5	2.25
Doris	8	13.5	5.5	30.25
Barbara	9	9.5	0.5	0.25
Cynthia	10	16	6.0	36.00
Hilde	11.5	5.5	—6.0	36.00
Flo	11.5	11.5	0.0	0.00
Nancy	13.5	8	—5.5	30.25
Mart	13.5	15	1.5	2.25
Stan	15	11.5	—3.5	12.25
Sarah	16	13.5	—2.5	6.25
Total			0.0	207.50

Para la no asociación, el segundo factor será exactamente la unidad.

Si N > 10, la distribución de selección de r_s es aproximadamente normal, con una desviación estándar de $1/\sqrt{N-1}$. Por lo tanto, en el ejemplo que estamos examinando, el error estándar será de $1/\sqrt{15}$. Como prueba de la hipótesis nula de que no se da relación en la población, podemos calcular Z como sigue:

$$Z = \frac{r_s - 0}{1/\sqrt{N-1}} = .695\sqrt{15} = 2.69$$

Mirando en la tabla normal vemos que la relación es significativa al nivel de .01.

La tau de Kendall. Al calcular la r_s de Spearman nos servimos de los cuadrados de las diferencias en los rangos. La tau de Kendall, en cambio, que también varía entre −1.0 y 1.0, se basa en una operación algo distinta. En efecto, calculamos primero

una estadística S buscando todos los pares posibles de casos y observando si las puntuaciones están o no en el mismo orden. Así, por ejemplo, supongamos que teníamos las siguientes combinaciones de lugares:

	a	*b*	*c*	*d*
A	1	2	3	4
B	2	3	1	4

Como quiera que las marcas de A se han dado en orden ascendente, podemos calcular S examinando las clasificaciones de B una por una. Fijándonos en el primer valor de la hilera B (individuo a), vemos que la marca de B está en el orden apropiado para los pares (a,b) y (a,d). En otros términos: el individuo a ocupa un lugar inferior a b y d en ambas variables A y B. Por otra parte, la marca de B discrepa (con respecto a la marca de A) para el par (a,c), ya que a ocupa un lugar inferior a c en cuanto a A, pero inversamente en cuanto a B.

Sirvámonos de $+1$ cada vez que un par determinado se halla ordenado igualmente para A y B (lo que se denomina par "concordante") y de -1 cada vez que se halla ordenado al revés (lo que se denomina par "discordante"). El valor de S se obtiene sumando dichos $+1$ y -1 para todos los pares posibles. Por lo tanto, S es igual al número de pares concordantes C, menos el número de pares discordantes D. Por lo tanto, la contribución de los pares (a,b), (a,c) y (a,d) es: $+1-1+1=(2-1)=1$. Con objeto de tener en cuenta los demás pares, recorremos la tabla de izquierda a derecha. Vemos así que la contribución de los pares (b,c) y (b,d) es de $-1+1$, o sea cero. Finalmente, la contribución del par (c,d) es de $+1$. Obsérvese que de hecho podemos obtener el valor total de S disponiendo primero A en el orden apropiado y examinando luego sucesivamente los lugares de la hilera B, contando cada vez el número de lugares de la derecha que están en el orden apropiado y sustrayendo los que están en el orden contrario. De este modo, en este sencillo ejemplo obtenemos:

$$S = C - D = (2-1) + (1-1) + (1-0) = 2$$

Si ahora dividimos S entre el valor máximo posible que podría tener, esto es: $(N-1)+(N-2)+\ldots+2+1=N(N-1)/2$, obtenemos un coeficiente que puede variar de -1 a $+1$. Definimos así el coeficiente tau$_a$ (según Kendall [16]), adecuado cuando no hay empates, como sigue:[2]

[2] Este coeficiente, derivado de los datos de la muestra, se denomina a veces *t*, en tanto que tau se reserva para la contrapartida de la población. Seguiremos, sin embargo, el uso más convencional. La tau de Kendall no debe confundirse con las tau$_a$ y tau$_b$ de Goodman y Kruskal, las que son apropiadas para datos nominales.

$$\tau_a = \frac{S}{\frac{1}{2}N(N-1)} = \frac{C-D}{\frac{1}{2}N(N-1)} \qquad \text{(XVIII.11)}$$

Es obvio que si hay discrepancia perfecta entre los dos sistemas de ordenación (esto es, si B estuviera ordenado como 4, 3, 2, 1), el valor de S será $-\frac{1}{2}N(N-1)$, y τ será -1.0. Y asimismo, si las dos variables no tienen relación alguna entre sí, las contribuciones a S positivas y negativas se invalidarán, y τ será cero.

Con objeto de ilustrar el caso de los órdenes empatados, sirvámonos nuevamente del ejemplo del campamento de trabajo. Dispongamos a los individuos en orden horizontal y reemplacemos los nombres por letras. Nuestra disposición se presenta en esta forma:

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*	*o*	*p*
A	1	2.5	2.5	4	5	6	7	8	9	10	11.5	11.5	13.5	13.5	15	16
B	5.5	5.5	1	2	3	9.5	5.5	13.5	9.5	16	5.5	11.5	8	15	11.5	13.5

Hemos de seguir la regla de que siempre que algún par comporte un empate, ya sea en la marca A o B, su contribución a S será cero. Mirando primero todos los pares que pueden formarse con a, vemos que los pares (a,b), (a,g) y (a,k) no contribuirán con nada a S, ya que las marcas de B para todos dichos individuos están ligadas en 5.5. Por lo tanto, la contribución de todos los demás pares será:

(a,c)	(a,d)	(a,e)	(a,f)	(a,h)	(a,i)	(a,j)	(a,l)	(a,m)	(a,n)	(a,o)	(a,p)	
1	1	−1	+1	+1	+1	+1	+1	+1	+1	+1	+1	$=9-3=6$

A continuación comparamos las marcas de b con cada una de las marcas a su derecha. Obsérvese, sin embargo, que b y c están ligados con respecto a A. Como quiera, por lo tanto, que b y c pudieron haberse dado lo mismo en el orden inverso, hemos de eliminar el par (b,c). Y en forma análoga, los pares (b,g) y (b,k) están ligados en B y, por consiguiente, no harán contribución alguna a S. En esta forma, para los pares de b, obtenemos una suma de $9-2$, o sea 7. Recorriendo la tabla de izquierda a derecha obtenemos finalmente:

$$\begin{aligned} S = C - D = {} & (9-3)+(9-2)+(13-0)+(12-0)+(11-0) \\ & +(6-3)+(8-0)+(2-5)+(5-2)+(0-6) \\ & +(4-0)+(2-1)+(2-0)+(0-2)+(1-0) \\ & =60 \end{aligned}$$

Con objeto de corregir en relación con los empates, hemos de practicar ahora un ajuste en el denominador de tau. Semejante ajuste tiene el efecto de producir un aumento del valor numérico de tau, si bien dicho aumento será ligero, a menos que el número de empates sea muy grande. La fórmula de tau (la que Kendall designó como τ_b) puede generalizarse como sigue:

$$\tau_b = \frac{S}{\sqrt{\tfrac{1}{2}N(N-1)-T}\sqrt{\tfrac{1}{2}N(N-1)-U}} \qquad \text{(XVIII.12)}$$

en donde $T = \frac{1}{2}\Sigma t_i(t_i - 1)$, siendo t_i el número de empates en cada grupo de empates en A, y $U = \frac{1}{2}\Sigma u_i(u_i - 1)$, siendo u_i el número de empates en cada grupo de empates en B. En el ejemplo anterior tenemos tres empates, de dos cada uno, en la variable A (popularidad). Por lo tanto:

$$T = \tfrac{1}{2}[2(1) + 2(1) + 2(1)] = 3$$

Y en forma análoga, hay tres empates, de dos cada uno, y una marca con cuatro empates en la variable B (participación). Por consiguiente:

$$U = \tfrac{1}{2}[2(1) + 2(1) + 2(1) + 4(3)] = 9$$

De donde:

$$r_b = \frac{60}{\sqrt{[8(15)-3][8(15)-9]}} = \frac{60}{\sqrt{(117)(111)}} = \frac{60}{114.0} = .526$$

Prueba de significación para tau. Kendall [16] ha demostrado que para tamaños de muestras de 10 o más, la distribución de muestreo de S bajo la hipótesis nula será aproximadamente normal, con media de cero y variancia dada por:

$$\sigma_s^2 = {}^1/_{18}N(N-1)(2N+5) \qquad \text{(XVIII.13)}$$

Hablando estrictamente, la fórmula anterior es aplicable sólo cuando no hay empates, pero puede ser usada cuando el número de éstos es relativamente pequeño. Si se da un gran número de empates, un factor de corrección bastante voluminoso habrá de ser aplicado.

Para probar la significancia de tau con los datos del campo de trabajo, comenzamos por computar σ_s^2 como sigue:

$$\sigma_s^2 = {}^1/_{18}(16)(15)(37) = 493.3$$

Obteniendo la raíz cuadrada tenemos:

$$\sigma_s = 22.21$$

valor que puede ser usado en el denominador de Z al probar la hipótesis nula de que A y B no están relacionados. Así

$$Z = \frac{S-0}{\sigma_s} = \frac{60.0}{22.21} = 2.70$$

y vemos que un valor de tau de .526 es significativo al nivel de .01.

Medidas ordinales para datos agrupados: tau$_c$, gamma, d$_{yx}$ y d$_{xy}$. Una de las ventajas de tau respecto de r_s es que aquélla puede utilizarse fácilmente cuando se da un número grande de empates. Pese a que el cálculo de rutina que se acaba de describir resultaría sumamente fastidioso en tales casos, podemos simplificar mucho el procedimiento cuando ambas categorías se han agrupado en categorías algo toscas. Así, por ejemplo, puede haberse colocado a personas en cinco clases sociales, considerándolas como empatadas con respecto a la posición. Si la segunda variable se ha categorizado en la misma forma, podemos servirnos de una fórmula de tau modificada, aprovechando con ello la información de que los datos han sido efectivamente ordenados, y no simplemente puestos en categorías.

Podemos calcular $S = C - D$ mediante un procedimiento que se describe más abajo. Sirviéndonos de las fórmulas que se acaban de dar, encontraremos que el límite superior de tau$_b$ sólo será la unidad cuando el número de hileras y de columnas sea el mismo. Con objeto de corregir para el caso en que $r \neq c$, formamos la razón:

$$\tau_c = \frac{S}{\tfrac{1}{2}N^2[(m-1)/m]} \qquad \text{(XVIII.14)}$$

donde $m = \text{Min}\ (r,c)$

Aquí seguimos a Kendall en el empleo del símbolo τ_c, con objeto de distinguir la ecuación (XVIII.14) de las fórmulas precedentes. Veamos ahora cómo se calcula τ_c.

Los datos del cuadro XVIII.4 representan los lugares asignados a 217 estudiantes de introducción a la sociología en la Universidad de Michigan. La variable B comporta el interés general del estudiante en cuanto a adoptar las formas "apropiadas" o "correctas" de comportamiento en los medios convencionales. En tanto que la variable A comporta el deseo de formar parte de organizaciones únicamente con objeto de mejorar la posición social.

Toda vez que la medición de ambas variables fue más bien tosca, se decidió dividir cada una de ellas en cuatro categorías: interés alto, moderadamente alto, moderadamente bajo y bajo. De este modo, si bien cada variable comporta una escala ordinal con un

CUADRO XVIII.4. *Datos comparados para el cálculo de la tau de Kendall a partir de datos agrupados*

Grado del deseo de formar parte de organizaciones (A)	*Interés en la conducta adecuada (B)*				*Total*
	Alto	*Moderadamente alto*	*Moderadamente bajo*	*Bajo*	
Alto	18	19	12	8	57
Moderadamente alto	16	16	12	10	54
Moderadamente bajo	11	14	18	16	59
Bajo	5	5	15	22	47
Total	50	54	57	56	217

gran número de empates, los resultados pueden con todo reunirse en forma de una tabla de contingencia.

Al calcular S será conveniente obtener separadamente C y D, ya que dichas cantidades serán utilizadas también para otras mediciones discutidas en esta sección. Observamos en primer término que las marcas de A se han ordenado nuevamente de altas a bajas, con la diferencia de que ahora tenemos 57 individuos "empatados" en cuanto a las marcas altas, 54 en cuanto a las moderadamente altas, 59 en cuanto a las moderadamente bajas y 47 en cuanto a las bajas. Considerando primero a los de marcas altas en cuanto a A, vemos que 18 las tienen también altas en B; 19 moderadamente altas, etcétera. Para obtener las contribuciones a C y D (y por lo tanto a S) observamos que, como quiera que todos los individuos de la categoría alta de A están empatados, ninguno de estos pares contribuirá a C o D. Y en forma análoga, ninguno de los pares de la misma columna contribuirá a C o D, debido al hecho de que todos ellos están empatados con respecto a B. Si nos fijamos en una casilla determinada cualquiera, todas las marcas que se hallan por debajo y a la derecha de la misma contribuirán al número de pares C concordantes, en tanto que todas las que se encuentran por debajo y a la izquierda contribuirán a D. Así, por ejemplo, cada uno de los 18 individuos de la casilla producirá pares concordantes con cada una de las marcas

$$16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22$$

que quedan por debajo y a la derecha de dicha casilla. En total, pues, la contribución de la casilla en cuestión a C será de:

$$18(16 + 14 + 5 + 12 + 18 + 15 + 10 + 16 + 22) = 18(128)$$

A continuación nos fijamos en los 16 casos inmediatamente debajo del ángulo izquierdo superior. Cada uno de estos individuos tiene también marcas altas de B. Con objeto de contar los pares de contribuciones a C, volvemos a adicionar las cantidades que figuran debajo y a la derecha. Multiplicando luego por el número de casos tenemos:

$$16(14 + 5 + 18 + 15 + 16 + 22) = 16(90)$$

Al pasar a las columnas segunda y siguientes, empezamos a encontrar contribuciones a C y D, ya que las columnas de la izquierda tienen marcas superiores de B. Así, para la primera casilla de la segunda columna obtenemos como contribución a C:

$$19(12 + 18 + 15 + 10 + 16 + 22) = 19(93)$$

y como contribución a D la cantidad $19(16 + 11 + 5) = 19(32)$. Recorriendo la tabla hacia abajo y hacia la derecha en forma semejante, podemos obtener S hasta cierto punto con facilidad, como sigue:

$$\begin{aligned} C &= 18(128) + 16(90) + 11(42) + 19(93) + 16(71) + 14(37) \\ &\qquad + 12(48) + 12(38) + 18(22) \\ &= 9055 \\ D &= 19(32) + 16(16) + 14(5) + 12(67) + 12(35) + 18(10) \\ &\qquad + 8(112) + 10(68) + 16(25) \\ &= 4314 \end{aligned}$$

Por tanto: $S = 9055 - 4314 = 4741$

Así pues:

$$\tau_c = \frac{4741}{\frac{1}{2}(217)^2[(4-1)/4]} = .268$$

Obsérvese que el denominador de τ_c depende sólo del número de hileras y columnas, y no de las distribuciones marginales, las que por supuesto determinan el número de empates. Esto hace que τ_c sea difícil de interpretar, y, en este sentido, menos satisfactoria que τ_b.[3] Hay también otras varias medidas que di-

[3] Puede demostrarse que en el caso $k \times k$, en el que *todos* los totales marginales son exactamente N/k, τ_b y τ_c serán iguales. De otra forma, en el caso $k \times k$, τ_c será generalmente menor que τ_b en valor numérico, aun cuando pueda ser mayor que τ_b en el caso $r \times c$.

fieren en relación con el manejo de los empates en el denominador. La más conocida de dichas medidas es gamma (γ), la que excluye por completo los empates en el denominador, y puede además ser aplicada a datos no agrupados. La fórmula para gamma es la siguiente:

$$\gamma = \frac{C - D}{C + D}$$

En el ejemplo que estamos considerando obtenemos:

$$\gamma = \frac{9055 - 4314}{9055 + 4314} = .354$$

Se indicó en el capítulo xv que la Q de Yule, igual a $(ad - bc)/(ad + bc)$ es un caso especial de gamma. Podemos por ello esperar que gamma se conduzca esencialmente igual en los casos en que las distribuciones marginales son muy desiguales, debiendo observarse las mismas precauciones que se aplicaron a Q. Como tanto gamma como τ_a y τ_b tienen todas los mismos numeradores y puesto que el denominador de gamma excluye todos los empates, puede verse fácilmente que $|\gamma| \geq |\tau_b| \geq |\tau_a|$. En general, hasta el grado en que los totales marginales para A y B son muy diferentes, gamma puede exceder a τ_b por una cantidad apreciable. Por ejemplo, en el caso del siguiente cuadro hipotético:

	B			
A	*Alta*	*Media*	*Baja*	*Total*
Alta	100	80	0	180
Media	0	20	80	100
Baja	0	0	20	20
Total	100	100	100	300

observamos que no hay pares discordantes, de modo que $\gamma = 1.0$. Sin embargo, $\tau_b = .77$ y $\tau_c = .68$. El que uno desee o no referirse a la anterior asociación considerándola "perfecta", dependerá de los supuestos en relación con la causa de que las distribuciones marginales no sean idénticas.

Además de las taus y gamma, tenemos dos medidas asimétricas, d_{yx} y d_{xy}, ideadas por Sommers [20] y definidas como sigue:

$$d_{yx} = \frac{C - D}{C + D + T_y}$$

y

$$d_{xy} = \frac{C - D}{C + D + T_x}$$

en donde T_x es el número de pares que están empatados en X pero no en Y, y T_y es el número de pares empatados en Y pero no en X. Si hacemos que T_{xy} se refiera al número de pares empatados tanto en X como en Y, y volviendo a la ecuación (XVIII. 12) para τ_b, veremos que $T = T_x + T_{xy}$, y $U = T_y + T_{xy}$, y por tanto, ya que el número total de pares $\frac{1}{2}N(N-1) = C + D + T_x + T_y + T_{xy}$, tendremos $C + D + T_y = \frac{1}{2}N(N-1) - (T_x + T_{xy}) = \frac{1}{2}N(N-1) - T$. De manera análoga, el denominador de d_{xy} es $C + D + T_x = \frac{1}{2}N(N-1) - U$. Así, el producto $d_{yx}d_{xy} = \tau_b^2$. En este sentido puede pensarse en las medidas asimétricas como *análogos declives*. Sin embargo, como su asimetría es función del número de empates, los que habitualmente dependen de los procedimientos de clasificación, la analogía con los declives b_{yx} y b_{xy} es, en el mejor de los casos, muy tenue.

Costner [5] ha señalado que puede darse a gamma una interpretación de reducción proporcional en el error semejante a la dada a las τ_b o λ_b de Goodman y Kruskal. Supongamos que deseamos predecir el orden de un par de casos con respecto a B. Si prescindimos de empates, nuestra probabilidad de incurrir en error, no conociendo nada más, sería de .5. Pero si conocemos el orden con respecto a A, resulta que el valor absoluto de gamma es igual al número de errores esperados conociendo A, menos el número esperado no conociendo A, dividido entre el número esperado no conociendo A.

Tenemos así disponible un número de medidas ordinales que difieren sólo en relación con el tratamiento de los empates en el denominador. Por desgracia, no tenemos de ordinario reglas claras de decisión para elegir entre dichas medidas, ya que las razones para los empates permanecen frecuentemente en la oscuridad. Wilson [23] ha demostrado que la propiedad de gamma, de reducción proporcional en el error, desaparece si se admite que los errores pueden cometerse cuando se predice un orden con respecto a B si, en realidad, el par está empatado con respecto a B.[4] Parece como si este problema del manejo de los empates no tuviera solución sencilla. Tal vez la mejor regla empírica consista en hacer uso de tantas categorías de cada variable como sea posible, reduciendo así el número de empates, a la vez que las diferencias entre las distintas medidas.

[4] Wilson [23] hace observar que tales empates no están excluidos del análisis de los modelos de regresión. Así, si dos casos se encuentran sumamente próximos en relación con sus puntuaciones de X, prediceríamos que sus puntuaciones de Y también lo estarían. En este sentido, si hay un par empatado con respecto a X, podemos esperar que lo esté también con respecto a Y, y cometeríamos un error si así *no* fuera. ¿Cuál es la importan-

Kruskal [17] ha demostrado que la medida de la r_s de Spearman puede ser interpretada en función de tríos de observaciones en lugar de pares, preguntándose cuál es la probabilidad de que, por lo menos, una de las tres observaciones sea concordante con las otras dos *a la vez*. Tal interpretación tiene una mucho menor atracción intuitiva que las interpretaciones mediante pares, aparte del hecho de que son mayores nuestros conocimientos acerca de los errores de muestreo de tau y gamma. Por estas razones prefiere Kruskal la tau a la r_s. Sin embargo, si la distribución básica de las dos variables es realmente bivariada normal, el valor absoluto de r_s será mayor que el de tau, y su comportamiento puede resultar mucho más semejante al de la correlación momento-producto. Trabajos previos no publicados muestran que el comportamiento de las r_s *parciales* (después de corregidos los empates) es muy singular al de las correlaciones parciales cuando las relaciones verdaderas son normales multivariadas (véase la definición en el próximo capítulo), por lo que sigue sin aclararse cuál de las medidas es preferible. Ante tal situación, el investigador deberá aplicar varias medidas diferentes para comprobar si se comportan de manera semejante al aplicarse a los datos que se examinan.

Finalmente, debemos tomar nota de un argumento de Wilson [22], quien afirma que *ninguna* medida ordinal que implique la idea de pares (o tríos) puede tener propiedades plenamente deseables. El punto básico de Wilson está en que el razonamiento teórico se funda normalmente en leyes que son apropiadas para un caso *único*, como cuando especificamos por ejemplo que el cambio de una unidad en X debe producir en Y el cambio de b_{yx} unidades. Con base en tales teorías, no tiene sentido pensar en función de pares ordenados, los que por necesidad nos fuerzan a realizar comparaciones a través de los casos. Si, por ejemplo, la propia teoría especifica que un cambio en el porcentaje de negros producirá un cambio en los niveles de discriminación, uno se está refiriendo tal vez a una "ley" que opera *en el interior* de una simple localidad (u otras unidades de observación). No se aplica directamente a comparaciones a través de pares de observaciones. Por supuesto que, en tanto uno defina su tarea como una simple generalización de poblaciones fijas, no se planteará este tipo de dificultad conceptual. El lector deberá consultar a Wilson si desea un análisis más completo. Está bien claro que el

cia de este "error" al predecir incorrectamente los empates, comparada con la del error de hacer predicciones equivocadas en los casos no empatados? Como puede verse, toda esta cuestión de la exclusión de empates, procedimiento que tiende a favorecer a gamma en relación con las demás medidas, no resulta cosa sencilla. Por ello, cuanto mayor sea el número de empates debidos a la crudeza de la medición, tanto más ambigua será la elección entre las medidas y mayor la sensibilidad de los resultados de tal elección.

empleo de medidas ordinales trae consigo cierto número de dificultades que hasta el momento no han sido resueltas adecuadamente.

Ejercicios

1. En los ejercicios 1 y 2 del capítulo XVII se calcularon tres coeficientes de correlación.

   a) Para cada uno de dichos coeficientes, empléese el análisis de variancia para verificar la hipótesis nula de que $\varrho = 0$. Respuesta, $F = .67$; $F = 7.09$; $F = 9.6$.
   b) Colóquense intervalos de confianza del 99.9 por ciento con respecto a las tres r.
   c) Verifíquese la relación entre la integración moral y la heterogeneidad en el caso de no linealidad.
   d) Conviértanse los mismos datos en órdenes y obténganse la tau de Kendall y la r_s de Spearman para las tres correlaciones.
   e) Verifíquese cada uno de estos coeficientes de rango ordenados en cuanto a significación.

2. En el ejercicio 3 del capítulo XVII se agruparon los índices de integración moral y de heterogeneidad. Calcúlense para estos datos agrupados la tau_c y la gamma de Kendall y compárese el resultado con el que se acaba de obtener antes en el ejercicio 1*d* de esta sección.

Bibliografía

1. Anderson, T. R., y M. Zelditch: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968, caps. 7 y 8.
2. Blalock, H. M.: "Estimating Measurement Error Using Multiple Indicators and Several Points in Time", *American Sociological Review*, vol. 35, pp. 101-111, 1970.
3. Bohrnstedt, G. W.: "Observations on the Measurement of Change", en Edgar Borgatta (ed.), *Sociological Methodology 1969*, Jossey-Bass Inc., Publishers, San Francisco, 1969, cap. 4.
4. Christ, Carl: *Econometric Models and Methods*, John Wiley & Sons, Inc., Nueva York, 1966, Parte III.
5. Costner, H. L.: "Criteria for Measures of Association", *American Sociological Review*, vol. 30, pp. 341-353, 1965.
6. Costner, H. L.: "Theory, Deduction and Rules of Correspondence", *American Journal of Sociology*, vol. 75, pp. 245-263, 1969.
7. Croxton, F. E., y D. J. Cowden: *Applied General Statistics* 3ª ed., Prentice-Hall, Inc., Englewood Cliffs, N. J., 1967, cap. 20.
8. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Clasifications", *Journal of the American Statistical Association*, vol. 49, pp. 732-764, 1954.
9. Goodman, L. A., y W. H. Kruskal: "Measures of Association for Cross Classifications, II: Further Discussion and References", *Journal of the American Statistical Association*, vol. 54, pp. 123-163, 1959.
10. Goodman, L. A., y W. H. Kruskal: "Measures of Association for

Cross Classifications, III: Aproximate Sampling Theory", *Journal of the American Statistical Association*, vol. 58, pp. 310-364, 1963.
11. Haggard, E. A.: *Intraclass Correlation and the Analysis of Variance*, The Dryden Press, Inc., Nueva York, 1958, pp. 22-26.
12. Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 23.
13. Hays, W. L.: *Statistics*, Holt, Rinehart and Winston, Inc., Nueva York, 1963, cap. 16.
14. Heise, D. R.: "Separating Reliability and Stability in Test-Retest Correlation", *American Sociological Review*, vol. 34, pp. 93-101, 1969.
15. Johnston, J.: *Econometric Methods*, McGraw-Hill Book Company, Nueva York, 1963, Parte II.
16. Kendall, M. G.: *Rank Correlation Methods*, Hafner Publishing Company, Inc., Nueva York, 1955, caps. 1, 3 y 4.
17. Kruskal, W. H.: "Ordinal Measures of Association", *Journal of the American Statistical Association*, vol. 53, pp. 814-861, 1958.
18. Mueller, J. H., K. Schuessler, y H. L. Costner: *Statistical Reasoning in Sociology*, 2ª ed., Houghton Mifflin Company, Boston, 1970, cap. 10.
19. Siegel, Sidney: *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company, Nueva York, 1956, cap. 9.
20. Somers, R. H.: "A New Asymmetric Measure of Association for Ordinal Variables", *American Sociological Review*, vol. 27, pp. 799-811, 1962.
21. Wallis, W. A., y H. V. Roberts: *Statistics: A New Approach*, The Free Press of Glencoe, Ill., Chicago, 1956, cap. 17.
22. Wilson, T. P.: "A Critique of Ordinal Variables", *Social Forces*, vol. 49, pp. 432-444, 1971.
23. Wilson, T. P.: "A Proportional-Reduction-in-Error Interpretation for Kendall's tau-b", *Social Forces*, vol. 47, pp. 340-342, 1969.

XIX. CORRELACIÓN MÚLTIPLE Y PARCIAL

En los dos últimos capítulos nos hemos ocupado de la relación entre dos escalas de intervalo, entre una variable dependiente y una sola variable independiente. Los análisis de correlación y regresión pueden extenderse fácilmente para comprender cualquier número de escalas de intervalo, una de las cuales puede tomarse como dependiente, y las demás como independientes. El problema se puede concebir como un problema de predicción en el que tratamos de predecir una variable dependiente Y a partir de las variables $X_1, X_2, \ldots, X_k$. Habremos de servirnos de nuevo de un modelo muy sencillo, que será directamente análogo a la regresión lineal, excepto en cuanto al hecho de que habrá más de dos dimensiones.

El concepto de correlación se generalizará en dos formas. Emplearemos el término de *correlación parcial* para designar la correlación entre dos variables cualesquiera cuando los efectos de otras variables se han controlado. El de *correlación múltiple*, en cambio, servirá para indicar qué tanto de la variación total de la variable dependiente puede explicarse por todas las variables independientes actuando conjuntamente. Veremos que los materiales examinados en el presente capítulo comportan en su mayor parte extensiones directas de razonamientos presentados anteriormente. Una vez que hayamos ampliado las nociones de correlación y regresión, estaremos en condiciones, en el capítulo siguiente, de emprender el análisis de covariancia, que comporta una combinación de las técnicas de regresión con el análisis de la variancia.

XIX.1. *Regresión múltiple y mínimos cuadrados*

En la regresión múltiple tratamos de predecir una sola variable dependiente a partir de cualquier número de variables independientes. Si se da un gran número de variables de escala de intervalo que deban relacionarse entre sí, será posible, por supuesto, predecir cualquier variable particular a partir de cualquier combinación de las demás. Por lo regular resultará claro del contexto cuáles variables han de considerarse como independientes y cuáles como dependientes.[1] Así, por ejemplo, puede querer predecirse el éxito en la universidad a partir de una serie de marcas de aptitud y del éxito en la escuela secundaria. O puede resultar posible predecir la tasa de crecimiento de una ciudad

[1] Cuando se crea que existe una causación recíproca, o retroalimentación, de la variable "dependiente" hacia alguna de las demás, deberán emplearse ecuaciones simultáneas en lugar de mínimos cuadrados. Véanse [4] y [12].