Las Técnicas de la Investigación Social

Por Pauline V YOUNG

CAPITULO XI

Técnica y Conceptos Básicos de la Estadística

Por Calvin F. Schmid
Universidad de Washington

(Concluye)

CORRELACION

En el lenguaje popular la idea de la correlación se expresa frecuentemente, pero, por lo general, en una forma cualitativa e inarticulada. Se hacen comentarios referentes a la relación entre la criminalidad y la debilidad mental, el suicidio y las enfermedades mentales, los malos alojamientos y la morbosidad, la delincuencia juvenil y los hogares deshechos, el divorcio y el ciclo económico y muchos otros fenómenos. Como es natural que los estudiantes de investigaciones sociales se interesen en estos problemas e intenten estudiarlos, es probable que resulte útil cierto tipo de correlación estadística. Para medir la relación entre dos variables, únicamente debe seleccionarse cierto tipo de correlación simple. Para dicha correlación debe hacerse una selección de las siguientes fórmulas, que dependen de los datos del problema que se tiene a la mano: 1) momento-producto o pearsoniano; 2) diferencia u orden de rango; y 3) curvilínea. Si, por otra parte, la relación fuera entre diversas variables deben aplicarse técnicas de correlación parciales o múltiples, o análisis de los factores.

La discusión presente se dedicará principalmente al momento-producto o pearsoniano, puesto que es el tipo de correlación básico, y el más comúnmente usado. El coeficiente de correlación del momento-producto (r) es un número puro y alcanza en valor desde el uno positivo (+ 1.0), pasando por cero (0.0), hasta el uno negativo (— 1.0). Esto es, que la correlación puede ser directa o positiva, o inversa o negativa, de acuerdo con la dirección del cambio y el tamaño del coeficiente que indica el grado de relación. Cuando una de las variables aumenta (o disminuye) y la otra cambia en una cantidad constante o casi constante en la misma dirección, la relación de las dos series es positiva; pero si los cambios en las dos variables son en dirección opuesta, la correlación entre ambas series es negativa. Por ejemplo: la altura y el peso de los seres humanos están positivamente relacionados, puesto que es común que las personas más altas pesen más que las bajas. Por otra parte, en este país hay una relación negativa o inversa entre la posición socio-económica y la fertilidad. Las familias que tienen más ingresos tienen menos niños que las familias con poco dinero.

Con objeto de dilucidar más ampliamente el concepto de correlación, examinamos el problema de la medición de la relación entre el peso y la altura de los estudiantes. Los datos que serán analizados, se derivan del examen médico realizado a un grupo de 265 estudiantes varones de la Universidad de Washington.

1. El primer paso para computar un coeficiente de correlación es construir un diagrama de dispersión. Dicho diagrama representa, en forma gráfica, el grado y tipo de relación o covariación en dos series de datos. Además, si los datos se encuentran en un diagrama de dispersión, pueden transferirse fácilmente a una carta de correlación para su cómputo. Para hacer un diagrama de dispersión, el primer paso es seleccionar los intervalos de clase convenientes para las respectivas variables, de tal manera que cada una tenga, aproximadamente, de ocho a quince grupos. Este procedimiento es similar al que se sigue para construir una distribución de frecuencia. De hecho, el diagrama de dispersión es una distribución de frecuencia en dos sentidos.

En la figura 19 se observará que se han elegido los intervalos de diez, para series que representan peso, y los intervalos de uno, para los que representan la altura. Los intervalos de clase para la variable X se leen de izquierda a derecha y, al revés de lo que sucede con la distri-

bución de frecuencia convencional, los intervalos de la variable Y se leen de abajo a arriba.

2. Cada anotación en la carta representa siempre dos valores numéricos. En el problema ilustrativo un valor representa altura, y otro peso. Por ejemplo, el primer caso (estudiante) elegido en nuestra muestra,

ALTURA (pulgadas)

PESO (Libras)

Y \ X	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	*f*
200-209									•		•			•				3
190-199											•							1
180-189									•	•		•	•	•				11
170-179							•		•	•	•	•	•	•	•			15
160-169								•	•	•	•	•	•	•			•	39
150-159				•	•	•	•	•	•	•	•	•	•					47
140-149					•	•	•	•	•	•	•	•	•		•			66
130-139					•	•	•	•	•	•	•	•						42
120-129	•		•	•	•	•	•	•	•			•						30
110-119			•	•	•		•	•	•									9
100-109					•													1
90-99	•																	1
f	2	0	4	6	11	19	27	42	41	36	27	22	15	8	2	0	1	265

Fig. 19. *Diagrama de puntos esparcidos, que muestra la relación entre la altura y el peso de 265 estudiantes.*

mide 66 pulgadas de estatura y pesa 145 libras. Por supuesto que el valor que representa la altura se coloca en la columna correspondiente al intervalo de clase 66. La célula exacta se localiza seleccionando la hilera apropiada que incluye también el correspondiente valor de peso. Para

dicha ilustración un peso de 145 libras quedará incluído en la hilera correspondiente al intervalo 140 a 149. Los valores pares de cada caso se extraen de acuerdo con este procedimiento.

3. Se observará que los puntos muestran una tendencia a amontonarse en una amplia zona que va del extremo inferior izquierdo al superior derecho del diagrama, indicando que la correlación es positiva. Si los casos estuvieran distribuídos a lo largo de una banda que fuera del extremo superior izquierdo al inferior derecho, la correlación sería negativa.

Después que se haya completado el diagrama de dispersión y se juzgue que la distribución es rectilínea, los datos se transfieren a una carta de correlación. Es más barato, más fácil y más factible usar una carta de correlación impresa que hacerla a mano. La carta de correlación que se usará en el presente caso fué dividida por el profesor F. Stuart Chapin de la Universidad de Minnesota. [11]

Las siguientes instrucciones sintetizan los diferentes pasos que deben darse para computar un coeficiente de correlación en este tipo de carta:

1. Los intervalos de clase, tanto para las variables X como Y, deben escribirse en los espacios que están arriba y a la izquierda de la carta de correlación, y el número de casos debe quedar registrado en las células correspondientes. Las frecuencias de las dos variables también deben figurar en la carta. Esta operación consiste simplemente en transferir los datos esenciales del diagrama de dispersión a la carta de correlación. Al seleccionar los intervalos cero, para las dos variables, debe intentarse escoger intervalos en los cuales sea más fácil que se presenten las medias de las distribuciones respectivas. En este problema se escogió 69, para representar el intervalo cero, para la variable X y 140 a 149 para la variable Y.

11 Todas las formas impresas para computar el coeficiente de correlación son muy semejantes. Algunas de las cartas más conocidas son: 1) Thurstone (publicada por C. H. Stoelting Company, Chicago; 2) Otis (World Book Company, Nueva York); 3) Cureton and Dunlop (Psychological Corporation, Nueva York); 4) Tryon (Universidad de California); 5) Ruch-Stoddard (Universidad de Iowa); 6) Holzinger (Universidad de Chicago); 7) Dvorak (Longmans, Green, Nueva York); y 8) Kelley (World Book Company, Nueva York).

Fig. 20. *Carta de correlación.* Desarrollada por F. Stuart Chapin de la Universidad de Minnesota y reconstruida por Calvin F. Schmid

2. Para la variable X determinar los productos de (f) (d_x) y colocarlos en la columna fd_x. Multiplicar (f) (d_y) por la variable Y y después colocar el resultado en la columna fd_y. Determinar las sumas algebraicas de la columna fd_x y fd_y. En la figura 20 se observará que $fd_x = + 11$ y $fd_y = + 99$.

3. Los respectivos valores de fd_x^2 se obtienen en seguida, lo mismo que los de fd_y^2, y se registran en la carta. Agréguense las columnas fd_x^2 y fd_y^2. En el problema $fd_x^2 = 1{,}839$ y $fd_y^2 = 931$. Debemos repetir que el segundo y tercer pasos son idénticos a los que siguen para computar la desviación standard.

4. La cuarta operación es diferente de todo lo que se ha discutido aquí. Primero, nótense las pequeñas cifras que aparecen en el extremo superior izquierdo de cada célula. Segundo, obsérvense los signos de cada uno de los cuadrantes que se indican en el centro de la carta. El cuadrante inferior izquierdo y el superior derecho son positivos (+), y los otros dos negativos (—). Multiplíquese el número de casos de cada célula por la cifra que aparezca en ella, observando los signos. El producto se coloca en la columna fd_{xy} + o en la fd_{xy} —, según el signo. Ilustremos este paso llevando a cabo los cómputos en la hilera 130 a 139 de la figura 20. Multiplíquese cada una de las frecuencias en la hilera designada por el intervalo de clase 130 a 139 por las pequeñas cifras que se encuentran en cada una de las células correspondientes. Los productos para los números localizados en el cuadrado positivo son los siguientes: (4) (2) = 8; (3) (5) = 15; (2) (12) = 24; y (1) (7) = 7. La suma de estos productos, que es 54, se coloca en la columna fd_{xy} +. Los productos de los números para esta hilera, que se localizan en el cuadrado negativo son: (3) (1) = — 3; (2) (3) = — 6; y (1) (5) = — 5. Sumando estos productos tenemos — 14, cifra que se registra en la columna fd_{xy}—. Las cifras que hay en cada columna se suman y se colocan en la carta. Los pasos siguientes consisten en determinar la suma algebraica de + fd_{xy} y — fd_{xy}. En el problema las cifras son: 848 — 75 = 773.

5. Esto completa todos los cómputos preliminares de la carta. El último paso, consiste en substituir la fórmula del lado derecho de la carta

y proceder a los cálculos. Se observará en la figura 20 que han sido hechas las substituciones convenientes en la fórmula, y el coeficiente de correlación ha sido computado para este problema ilustrativo. El coeficiente de correlación entre el peso y la altura, para este ejemplo de 265 estudiantes varones, es $r = + .60$.

Cálculo Pearsoniano o coeficiente de correlación por el producto-momento (r) *para datos no agrupados.* Si no hay más de 50 ó 75 casos es generalmente más fácil calcular por medio del método no agrupado. Esta afirmación se hace siempre que se disponga de una tabla de cuadrados y de una buena máquina calculadora.

Como se observará en la Tabla VII, el procedimiento para calcular r a través de este método es muy simple. El problema en la Tabla VII es determinar la relación entre la mortalidad por cáncer y cierto índice cultural para los registros de mortalidad de los Estados Unidos.

1. La primera columna contiene los nombres de los Estados, la segunda las cifras de cáncer respectivas (x), y la tercera el índice de valores correspondientes (y).

2. La cuarta columna representa los productos de xy.

3. La quinta columna contiene los valores de x^2 y la sexta columna los valores de y^2.

4. En seguida se determinan las sumas de las cifras en cada una de las columnas.

5. Se hacen las substituciones convenientes en la fórmula y se calcula el coeficiente de correlación. En el problema ilustrativo $r = + 0.66$

TABLA VII

Cálculo del coeficiente de correlación por el momento-producto (r) para casos no agrupados. La relación es entre la mortalidad por cáncer y otros tumores malignos y la distribución en 152 items culturales. Para el Area de Registro de los Estados Unidos 1929 a 1931*

Estado	Cáncer Porcentaje x	Indice Cultural y	Porcentaje de frecuencia Indice xy	Porcentaje al cuadrado x^2	Indice al cuadrado y^2
Alabama	60.8	13	790.4	3,696.64	169
Arizona	65.0	41	2,665.0	4,225.00	1,681
Arkansas	46.9	10	469.0	2,199.61	100
California	91.1	93	8,472.3	8,299.21	8,649
Colorado	80.1	34	2,723.4	6,416.01	1,156
Connecticut	96.1	72	6,919.2	9,235.21	5,184
(Con el fin de ahorrar espacio solamente se han incluído en esta Tabla los datos detallados sobre los registros de la mortalidad en 1931 para 10 de los 47 Estados					
Washington	83.1	54	4,487.4	6,905.61	2,916
West Virginia	63.0	19	1,197.0	3,969.00	361
Wisconsin	89.6	38	3,404.8	8,028.16	1,444
Wyoming	62.9	50	3,145.0	3,956.41	2,500
Totales (Σ)	3,586.6	1,740	142,016.3	283,886.98	83,664

$$r = \frac{\frac{\Sigma xy}{N} - \left(\frac{\Sigma x}{N}\right)\left(\frac{\Sigma y}{N}\right)}{\sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}\ \sqrt{\frac{\Sigma y^2}{N} - \left(\frac{\Sigma y}{N}\right)^2}}$$

$$= \frac{\frac{142{,}016.3}{47} - \left(\frac{3586.6}{47}\right)\left(\frac{1740}{47}\right)}{\sqrt{\frac{283{,}886.98}{47} - \left(\frac{3586.6}{47}\right)^2}\ \sqrt{\frac{83{,}664}{47} - \left(\frac{1740}{47}\right)^2}}$$

$$= \frac{196.513}{\sqrt{216.779}\ \sqrt{409.531}} = \frac{196.513}{297.949} = +.65955 \text{ ó } +.66$$

* Para mayores datos relativos a la interpretación de la correlación entre la mortalidad por cáncer y el índice cultural en esta Tabla véase: Calvin F. Schmidt, *Mortality Trends in the State of Minnesota* (Tendencias de la Mortalidad en el Estado de Minnesota), pp. 191-193.

Interpretación del coeficiente de correlación. Al interpretar el coeficiente de correlación, el tamaño del mismo no es lo único que debe tomarse en consideración. Tan importante como el tamaño absoluto del coeficiente es el tamaño relativo del caso en general y la naturaleza de los datos. Decir que un coeficiente de correlación de 0.75 (más o menos) es "alto", puede producir grandes confusiones si el error standard o probable es relativamente grande. Decir también que un coeficiente de 0.40 debe considerarse siempre como "bajo", puede ser equivocado para ciertos tipos de datos.

Al juzgar el tamaño de un coeficiente particular deben tenerse en cuenta otros coeficientes que han sido derivados de la misma clase de datos. Puede encontrarse, por ejemplo, que los coeficientes de correlación para ciertas variables han sido constantemente inferiores a 0.20, de modo que un coeficiente de 0.40 sería considerado relativamente alto para este tipo particular de datos.

Sin embargo, con objeto de que el lector pueda disponer de hechos un poco más definitivos, para emplearlos en la interpretación del tamaño de los coeficientes de correlación, puede ser de alguna utilidad la siguiente clasificación general: 1) un coeficiente de 0.70 a 1.00 (más o menos), significa que hay un *alto* grado de asociación entre las series; 2) si el coeficiente es mayor de 0.40 pero menor de 0.70 hay una relación *substancial;* 3) si el coeficiente es mayor de 9.20 pero menor de 0. 40 hay una correlación *baja* y 4) si el coeficiente es menor de 0.20 hay una relación *insignificante.*

El hecho de que dos variables presenten una correlación "alta", no es una evidencia *per se* de una relación causal. Un coeficiente de correlación presenta únicamente el grado de asociación entre dos grupos de fenómenos. Si los fenómenos están o no causalmente relacionados, es cuestión de interpretación. Hay por lo menos tres interferencias posibles que pueden presentarse en un caso en que dos variables tengan un pronunciado grado de correlación; 1) uno puede ser la "causa" del otro; 2) ambos pueden estar relacionados con una o más "causas" comunes; y 3) la correlación puede haberse presentado por mera casualidad.

Método de correlación de diferencia de clasificación. El método de correlación por diferencia de clasificación puede resultar útil para ejemplos relativamente pequeños —generalmente menos de 30 casos—, o en donde

la clasificación de los items de las dos variables constituyen la única información de que se dispone. Desde un punto de vista algebraico el método de diferencia de clasificación es decididamente menos satisfactorio que el método de producto-momento. Por esta razón este último debe preferirse generalmente al primero.

Al describir el procedimiento para computar un coeficiente de correlación (ρ, letra griega rho) por el método de diferencia de clasificación, el problema de la Tabla VIII puede considerarse como una forma de ilustración.

1. El problema de la Tabla VIII, consiste en determinar la relación entre la incidencia del suicidio y la movilidad de población en las 25 ciudades más grandes de los Estados Unidos, en 1930. En la primera columna aparecen las 25 ciudades y en la segunda y tercera, respectivamente, el promedio de cifras de suicidio para el período de 2 años 1929 y 1930, y un índice de movilidad para 1930.

2. Las clasificaciones relativas de las diferentes ciudades de cada una de las variables, están registradas en la cuarta y quinta columnas. La clasificación significa simplemente, la numeración de los diferentes valores de acuerdo con las posiciones que ocupan cuando están ordenadas de acuerdo con la magnitud. El variante de valor más alto da la clasificación 1, el siguiente 2, y así sucesivamente. En caso de que haya dificultades en la clasificación, pueden emplearse uno o dos métodos; el de "paréntesis angulares", o el de "clasificación media", los items que tienen el mismo valor quedan en la misma clasificación, y al que sigue a la unión se le da la clasificación que hubiera tenido en caso de que no existiera ésta.

TABLA VIII

Cálculo del coeficiente de correlación de Diferencia de Clasificación (ϱ.) La relación se refiere al Suicidio y la Movilidad de Población para las veinticinco Ciudades Americanas más grandes: 1929 a 1930.

Ciudad	Variables		Grados		Rx — Ry	(Rx — Ry)²
	Suicidio %	Movilidad Indice	Suicidio	Movilidad		
	X	Y	Rx	Ry		
New York	19.3	54.3	13.5	11	+ 2.5	6.25
Chicago	17.0	51.5	19	8	+11	121.00
Philadelphia	17.5	64.6	16	17	— 1	1.00
Detroit	16.5	42.5	21	5	+16	256.00
Los Angeles	23.8	20.3	7.5	1	+ 6.5	42.25
Cleveland	20.1	52.2	12	10	+ 2	4.00
St. Louis	24.8	62.4	4	16	—12	144.00
Baltimore	18.0	72.0	15	23	— 8	64.00
Boston	14.8	59.4	23	14	+ 9	81.00
Pittsburg	14.9	70.0	22	22	0	0.00
San Francisco	40.0	43.8	1	6	— 5	25.00
Milwaukee	19.3	66.2	13.5	19	— 5.5	30.25
Buffalo	13.8	67.6	24	20	+ 4	16.00
Washington D. C..	22.5	37.1	9	4	+ 5	25.00
Minneapolis	23.8	56.3	7.5	12	+ 4.5	20.25
New Orleans	17.2	82.9	17.5	25	+ 7.5	56.25
Cincinnati	23.9	62.2	6	15	— 9	81.00
Newark	21.4	51.9	10	9	+ 1	1.00
Kansas City	24.5	49.4	5	7	— 2	4.00
Seattle	31.7	30.7	2	2	0	0.00
Indianapolis	21.0	66.1	11	18	— 7	49.00
Rochester	17.2	68.0	17.5	21	+ 3.5	12.25
Jersey City	10.1	56.5	25	13	+12	144.00
Louisville	16.6	78.7	20	24	— 4	16.00
Portland	29.3	33.2	3	3	0	0.00

$$\Sigma\,(Rx - Ry)^2 = 1199.50$$

$$\varrho = 1 - \frac{6\Sigma(Rx - Ry)^2}{N(N^2 - 1)}$$

$$= 1 - \frac{6(1199.50)}{25\ (625. - 1)}$$

$$= 1 - .461$$

$$= + .539$$

En el método de "clasificación media" se les da a los items unidos la misma clasificación, pero ésta representa la clasificación media de los mismos items unidos. El último método es generalmente preferible, y es el que se ha seguido en la Tabla VIII. Se observará que tanto en Nueva York como en Milwaukee el suicidio alcanza 13.5, Cleveland tiene 12 y Baltimore 15. Puesto que tanto Nueva York como Milwaukee tienen cifras de 19.3 su clasificación es propiamente $\frac{13 + 14}{2}$, sea 13.5. Con el método de "clasificación de paréntesis angulares", tanto Nueva York como Milwaukee tendrían clasificación 13, y las de Cleveland y Baltimore serían las mismas que con el método de "clasificación media".

3. El siguiente paso es encontrar las diferencias en las clasificaciones (Rx — Ry) para cada uno de los items.

4. En seguida se elevan al cuadrado las diferencias en las clasificaciones, y las cifras se registran en la séptima columna.

5. La suma de la séptima columna $(Rx - Ry)^2$, se obtiene y se hacen las substituciones convenientes en la fórmula de la Tabla VIII. Se observará que $\rho = +.54$.

El coeficiente de correlación en la diferencia de clasificación (ρ) y el coeficiente del producto-momento (r) no son idénticos, pero sí tienen un valor similar. Por ejemplo, en el problema ilustrativo $\rho = +.54$, que es el equivalente de $r = +.558$. Pearson ha inventado una fórmula que puede usarse para traducir ρ a r o viceversa. Las tablas que registran los valores de ρ y los correspondientes de r, se encuentran en muchas pruebas standard en las estadísticas. [12]

Correlación curvilínea. Como ya se indicó, el coeficiente de correlación pearsoniano o del momento-producto, se basa en la suposición de que la relación entre las dos variables es rectilínea. Cuando los datos no son lineales (curvilíneos) r no da la medida exacta de correlación entre las variables. La constancia de la proporción de cambio de las dos variables determina si la correlación es rectilínea o curvilínea. Si el conjunto de cambios entre las dos variables tiene una proporción constante, entonces la

12 E. g., Robert E. Chaddock, *Principles and Methods of Statistics* (Principios y Métodos Estadísticos), pp. 300-301-464.

correlación es rectilínea, pero si no, es curvilínea. Por ejemplo, la relación entre la fuerza y la edad de los seres humanos es curvilínea, puesto que la fuerza no está en proporción constante en relación con la edad a través de toda la vida. Los hechos de esta clase solamente pueden aclararse colocando los datos en un diagrama de dispersión. Hay también fórmulas que pueden aplicarse como pruebas para alineamientos de regresión. [13]

Correlación parcial y múltiple. Como el sociólogo se ocupa muy frecuentemente de un gran número de factores y de sus interrelaciones, puede suceder que la simple técnica de correlación resulte definitivamente inadecuada para ciertos problemas. La simple correlación no mide separadamente la relación entre dos variables, de tal manera que los efectos de otras variables relacionadas quedan eliminados, ni tampoco determina la relación entre cualquier variable y los efectos combinados de diversas variables relacionadas. Los dos problemas de correlación que se presentan en estos casos, pueden ser tratados más satisfactoriamente por métodos de correlación parcial o múltiple.

El análisis de la correlación parcial es una técnica estadística para medir el grado de relación entre dos variables, cuando los efectos de otras variables específicas, con las cuales están relacionados, se eliminan. La correlación parcial tiene ciertas características del método experimental, en el que es posible estudiar la relación entre dos factores cuando otros distintos se mantienen constantes. En las ciencias sociales es extraordinariamente difícil llevar a cabo experimentos controlados, como se hace en las ciencias físicas o químicas. Sin embargo, la correlación parcial ofrece al científico social uno de los substitutos más satisfactorios para la experimentación controlada.

El análisis de la correlación múltiple es una técnica estadística para medir la correlación entre una variable (dependiente), y el efecto combinado de cierto número de otras variables (independientes). [14]

13 Véase una de las pruebas registradas en la bibliografía correspondiente este capítulo, para una discusión más detallada de la correlación curvilínea.

14 Para una discusión más completa de la correlación parcial y múltiple, véase Henry E. Garrett, *Statistics in Psychology and Education* (Las Estadísticas de Psicología y Educación), pp. 221-265.

MUESTRAS

Uno de los problemas más importantes y más difíciles en la investigación social, es el problema de las muestras. En vez de estudiar todos los casos que pueden quedar lógicamente incluídos en la investigación, solamente se selecciona una pequeña parte de ellos para el análisis, y de allí se sacan las conclusiones. La mayoría de los estudios estadísticos se basan en muestras, y no en numeraciones completas de todos los dato importantes. Una muestra estadística es una pintura en miniatura de todo el grupo o conjunto del cual se tomó la muestra. El grupo del cual se haya elegido la muestra se conoce como el ',universo", "polación" o "refacción"

Desde un punto de vista ideal, probablemente se consideraría preferible una cuenta completa de todos los casos importantes a una simple muestra. Sin embargo, resultaría imposible o poco practicable incluir otra cosa que una pequeña parte del número total de casos. Los factores de tiempo y costos constituyen generalmente consideraciones importantes en la investigación social. Es más económico basar los estudios en muestras y, para propósitos más prácticos, las conclusiones extraídas de una muestra pueden tener tanto valor como las que hayan surgido del análisis de la universalidad de los casos.

Hay dos aspectos básicos en el problema de las muestras estadísticas: primero, la selección actual de los items que van a constituir la muestra y, segundo, el mensuramiento de la exactitud de la muestra. La presente sección se dedicará especialmente a los principios y procedimientos que se siguen para seleccionar una muestra.

Hablando en términos generales, las mismas leyes matemáticas de casualidal o probabilidad que gobiernan la fluctuación de las monedas, el rodamiento de un disco bien formado o el dibujo de las paredes simétricamente coloreadas de tamaño uniforme que forman una urna, deben regir las muestras estadísticas. Es obvio que la consideración más importante, al seleccionar una muestra, es que represente íntimamente al universo de que se trata. El tamaño de la muestra no nos da seguridad necesaria de que sea representativa. Muestras relativamente cortas, cuando han sido propiamente seleccionadas, pueden ser mucho más efectivas que muestras más grandes que no se han seleccionado con todo cuidado. La selección debe ser arreglada de tal manera, que todos los items que forman parte del universo que se considera tengan la misma oportunidad de ser incluídos en ella.

Algunos de los procedimientos mecánicos más comunmente usados para formar las muestras son: 1) una selección puramente casual, 2) selección a intervalos regulares, y 3) selección proporcional. [15]

1. *Selección puramente casual.* Si se conoce la extensión del universo, cada caso puede numerarse en una hoja de papel o en un disco colocado en un lugar conveniente, después de lo cual los items individuales que se van a incluir en la muestra se van sacando de allí. Cada vez que se saque un caso, las hojas o los discos deben mezclarse. De acuerdo con este procedimiento todos los items individuales tienen aproximadamente las mismas oportunidades para que se les incluya en la muestra.

2. *Selección a intervalos regulares.* Otro procedimiento es seleccionar los casos de las series a intervalos regulares, por orden alfabético o por medio de cualquier otro arreglo arbitrario. Por ejemplo, al seleccionar la muestra de 265 estudiantes que se han empleado para propósitos ilustrativos en el presente capítulo; cada vigésimoquinto caso ha sido extraído de las listas de estudiantes varones durante el año escolar de 1937-1938. Los 265 casos seleccionados de esta manera representan solamente una muestra de 5%. En vez de tomar un caso cada veinticinco, puede tomarse cada tres, cada cinco o cada diez, según el número de casos que se desee incluir en la muestra.

3. *Selección proporcional.* Si la composición del universo es conocida, es posible seleccionar una muestra tomando sub-muestras proporcionales al tamaño de los elementos significativos o sub-divisiones del universo. Después de que se ha determinado el tamaño relativo de cada sub-muestra, los items individuales se seleccionan, bien al azar o por intervalos regulares. Por ejemplo, si se conocen los datos de un cuerpo escolar relativos al sexo, año escolar y a filiación fraternal, las sub-muestras se seleccionan sobre la base de estas diferentes categorías, siendo determinado el tamaño de cada sub-muestra por el número de casos en cada clasificación.

No debe pensarse que en la práctica la selección de las muestras es tan mecánica y tan simple como estas reglas generales parecen indicar. Además, como ya lo hace notar Bowley, "no hay reglas formales que puedan reemplazar a la experiencia en la selección e interpretación de las

15 George A. Lundberg, *op. cit.*, pp. 99-105.

muestras" Muy frecuentemente factores inesperados e incontrolables pueden hacer que una muestra no resulte representativa.

Nunca debe perderse de vista el hecho de que la exactitud y validez de las propias conclusiones, en un estudio, dependen en alto grado de la validez de la muestra. Hay muchos estudios en las ciencias sociales que llevan las marcas más superficiales de erudición y autoridad, y que carecen de valor intrínseco porque se basan en muestras que no son representativas. [16]

CONFIABILIDAD

Una verdadera medida de grupo, tal como el promedio aritmético debe basarse en todos los casos que figuran en el universo. Si el verdadero promedio de peso de todos los estudiantes varones de la Universidad de Washington tuviera que ser computado, tendría que incluirse a todo estudiante que estuviera registrado. Debe hacerse hincapié en que el peso medio de 148.74 libras solamente es un valor probable. Si se escogiera otra muestra de 265 casos, el peso podría no ser igual al que resultó de la primera muestra, pero probablemente sería muy semejante. Además, es probable que ni el peso de la segunda muestra ni el de la primera, coincidieran exactamente con el peso medio verdadero basado en todo el universo. En otras palabras, las medidas basadas en muestras, generalmente son menores o mayores que las medidas verdaderas.

Como la mayoría de los estudios estadísticos se basan en muestras, es importante conocer hasta qué punto dichas medidas representan a las verdaderas muestras y también qué variación puede esperarse que ocurra si se analizan otras muestras. Las llamadas medidas de confiabilidad pueden resultar de gran utilidad para aclarar estos problemas.

Las medidas de confiabilidad se ocupan únicamente de las fluctuaciones ocurridas en las muestras que se toman al azar. Es obvio que no tienen ninguna relación con los errores de observación o de cómputo. Además, siempre que se computa una medida de confiabilidad se entiende que la muestra debe ser adecuada, y que se ha seleccionado de acuerdo con un procedimiento rígidamente científico.

16 Para una discusión práctica excelente de los principios y técnicas de las muestras, véase Frederick F. Stephan, "Practical Problems of Sampling Procedure" (Problemas Prácticos de Muestras de Procedimientos), *American Sociological Review,* I. (Agosto, 1936), 569-580.

Suponiendo que la muestra haya sido debidamente seleccionada, la medida de confiabilidad, lo mismo que el promedio aritmético, depende de: 1) el número de casos que hay en la muestra y 2) la variabilidad de los valores de la misma. El sentido común indica que la seguridad de una medida está relacionada con el tamaño de la muestra. Es natural que tengamos más confianza en la veracidad de una muestra que sea relativamente grande que en una más pequeña. Puede encontrarse matemáticamente que la seguridad de una muestra varía de acuerlo con la raíz cuadrada del número de items que comprende la muestra. Así, pór ejemplo, sería necesario cuadruplicar el número de casos en una muestra con objeto de doblar su confiabilidad. Si la muestra es muy reducida, digamos menor de 25 casos, hay muy poca justificación para sacar medidas de confiabilidad. Hay fórmulas especiales de confiabilidad para muestras pequeñas, pero debe tenerse mucho cuidado al usar muestras extremadamente reducidas en la investigación social.

El grado de variabilidad de los casos en una muestra, también tiene una influencia importante sobre la seguridad de las medidas que se han sacado de ella. Si los casos en la muestra están notablemente dispersos, es natural esperar una fluctuación mayor en las medidas. La relación entre el tamaño de la muestra y la variabilidad de los items está claramente indicada en la fórmula del promedio de error standard. El error standard es una medida de confiabilidad:

$$\sigma X = \frac{\sigma}{\sqrt{N}}$$

El error standard del peso medio para la muestra de 265 estudiantes de la Universidad de Washington es:

$$\sigma X = \frac{18.37}{\sqrt{265}}$$

$$= \frac{18.37}{16.28}$$

$$= 1.12$$

¿Qué significa la cifra 1.12? Puede decirse que hay aproximadamente 68 (68.26) probabilidades en 100 de que las medias de las muestras del

mismo tamaño, elegidas en forma similar, se encuentren entre el intervalo 148.7 + 1.12 y 148.7 — 1.12, o entre 149.82 y 147.88 libras. Puede también decirse que las probabilidades son de 68 entre 100 de que el promedio verdadero —es decir el promedio del universo— caiga entre ± 1 σ X ó 149.82 y 147.88.

Es muy frecuente que la confiabilidad de una medida se exprese como error probable más bien que como error standard. Básicamente el error probable es el mismo que el error standard. El error probable de cualquier medida es .6745 del error standard. Por lo tanto el error probable del promedio es:

$$\text{P. E. } X = .6745\frac{\sigma}{\sqrt{N}}$$

En el problema ilustrativo.

$$\text{P E. } = X\ (.6745)\left(\frac{18.4}{\sqrt{265}}\right)$$

$$= \quad (.6745)\ (1.12)$$

$$= \quad .76$$

El error probable tiene una interpretación un poco diferente puesto que incluye un tramo más pequeño que el error standard. Del error probable de la ilustración puede concluirse que si las muestras adicionales se escogen del mismo tamaño, las probabilidades son de (1/2 + 1/2) que en la generalidad de los promedios computados a través de muestras nuevas que caerían dentro de los límites de 148.7 + .76 y 148.7 — .76, o entre 149.46 y 147.94. Otra vez se representa la probabilidad de que el verdadero promedio se encuentre entre 149.46 y 174.94 libras. Es una práctica común colocar el error probable de una medida inmediatamente después de la misma, pero precedido por el signo ±. Para el problema ilustrativo el peso medio y su error probable se escribirían 148.1 ± .76. En la actualidad muchos estadísticos prefieren el error standard y su uso aumenta cada vez más. Sin embargo, siempre que se usa el error standard es conveniente indicarlo así, puesto que se entiende generalmente que cuando dos cifras están separadas por el signo ± la segunda es el error probable.

No debe perderse de vista el hecho de que cada medida estadística tiene su propia fórmula de confiabilidad. La fórmula para el X puede ampliarse hasta cualquier forma de distribución de frecuencia, pero la siguiente fórmula debe usarse solamente si la distribución de la muestra es bastante normal.

El error standard del coeficiente de correlación es:

$$\sigma r = \frac{1 - r^2}{\sqrt{N}}$$

En el problema ilustrativo en el que se ha computado el coeficiente de correlación para altura y peso es:

$$\sigma r = \frac{1 - (.59973)^2}{\sqrt{265}}$$

$$= \frac{1 - .359676}{16.28}$$

$$= .0393 \text{ ó } .04$$

La fórmula para el error standard de la desviación standard para ese promedio es:

$$\sigma\sigma = \frac{\sigma}{\sqrt{2N}}$$

$$\sigma \chi = 1.2533 \frac{\sigma}{\sqrt{N}}$$

SERIES DE TIEMPO

En muchos tipos de problemas sociales es necesario frecuentemente estudiar los cambios que se efectúan en un determinado período de tiempo. Para algunos de estos problemas pueden resultar muy satisfactorias las técnicas tabulares y gráficas relativamente sencillas. La carta de listas ver-

ticales, la gráfica de coordenadas rectangulares y la carta de semi-logaritmos constituyen las formas básicas para representar series de tiempo. En el siguiente capítulo se encontrará una discusión detallada de estas cartas.

Sin embargo, algunas veces es necesario usar técnicas matemáticas más refinadas para analizar adecuadamente ciertos tipos de series temporales. A causa de la limitación de espacio, solamente pueden incluirse en el presente estudio algunas de las técnicas más elementales, con la esperanza de que a través de ellas el estudioso pueda darse una idea de los problemas que comprende el análisis de esta clase de datos. En la práctica se verá que los movimientos en las series temporales asumen diversas formas. La longitud del período en el cual el movimiento completa su curso, las características generales o configuración de los movimientos y el grado de regularidad de los mismos, pueden presentar muchas variaciones. Con referencia a estos puntos, los movimientos pueden ser descritos como:[17]

1. *De tendencia secular o de mucho tiempo.* Los fenómenos sociales a menudo presentan una tendencia definitiva o persistente a aumentar o disminuir en un considerable período de tiempo.

2. *Fluctuaciones periódicas.* Quizás el tipo más común queda ilustrado en la variación estacional de los fenómenos sociales que con más o menos regularidad presentan una máxima y una mínima. También pueden observarse en los ciclos diurnos o semanales caracterizados por una marcada periodicidad.

3. *Movimientos ondulatorios o cíclicos.* Estos movimientos son de carácter ondulatorio pero no son definitivamente periódicos. El llamado ciclo de los negocios, con sus períodos alternantes de depresión y prosperidad, es un buen ejemplo de este tipo de movimiento.

4. *Variaciones irregulares.* Estos movimientos irregulares pueden ser episódicos, fortuitos o accidentales. Los cambios episódicos pueden ser causados por factores tales como huelgas, quiebras, conflagraciones, terremotos o cualquier otro tipo de cataclismo natural. Los cambios episódicos se traducen en quiebras profundas en la variable y no demuestran ninguna tendencia aparente a repetirse en intervalos establecidos. Los movimientos

17 Edmund E. Day, *Statiscal Analysis* (Análisis Estadístico), pp. 235 y ss.

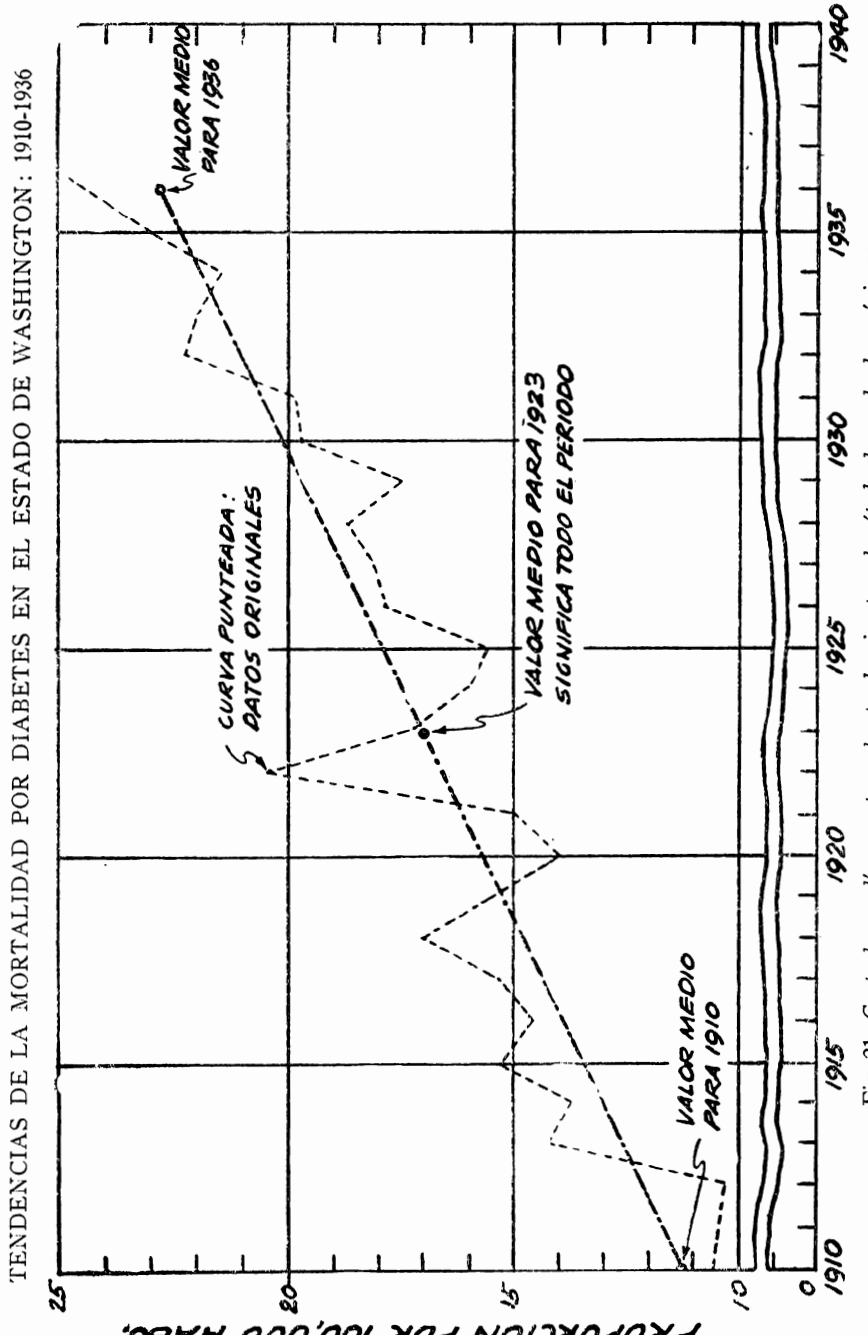

Fig. 21. *Corte de una línea recta en las tendencias por el método de cuadrados mínimos.*

fortuitos o accidentales son generalmente menos pronunciados que los cambios episódicos y casi nunca se deben a causas fácilmente determinables.

Con objeto de ilustrar algunas de las técnicas más elementales para analizar las series de tiempo, consideremos los datos de la Tabla IX. Dicho examen revela que ha habido una tendencia definitiva a que aumente la mortalidad por diabetes en el Estado de Washington, durante los últimos 27 años. Este hecho se pone de manifiesto al estudiar los datos de la carta. Pueden usarse tres procedimientos básicos de esta clase: 1) tirando una línea recta, 2) por el método de medias móviles, y 3) por el método de una curva matemática adecuada. El método de la línea libre es un procedimiento rudo y poco exacto para determinar los rasgos seculares de una serie, puesto que está basado simplemente en una visión general. Ordinariamente este procedimiento no debe usarse como no sea para un trabajo preliminar. El método de medias móviles es superior pero también posee limitaciones definidas. La Media Móvil se obtiene promediando grupos consecutivos en las series; omitiendo una vez un año y agregándolo en otra. Por ejemplo, si se va a calcular una media móvil de tres años para las series de la Tabla IX, el primer paso sería obtener la cifra media para los primeros tres años:

$$\frac{10.8 + 10.5 + 10.2}{3} = 10.5$$

La media obtenida se registraría para el año 1911, puesto que es el punto central del primer período de tres años. Con objeto de obtener valores similares para los años siguientes, una cifra se agrega y la otra se quita. Para obtener la segunda media móvil en la Tabla IX, se omite la cifra para 1910 y se agrega la de 1913

$$\frac{10.5 + 10.2 + 14.2}{3} = 11.6$$

El promedio 11.6 se coloca dentro del año 1912, que es el punto central del segundo período de tres años. Este procedimiento se continúa a través de toda la serie. Las medias móviles pueden basarse en intervalos de longitud diversa, según los datos originales. Sin embargo, no siempre es fácil determinar el mejor intervalo. Como las medias deben ser centradas, siempre que sea posible es conveniente usar un número de años impar para el intervalo.

TABLA IX

Determinación de la línea recta por el método de cuadrados. Datos que representan las cifras de mortalidad por 100,000 personas de Diabetes en el Estado de Washington: 1910 a 1936

AÑOS	PORCENTAJE Y	DESVIACION DEL PUNTO INTERMEDIO d	DESVIACION DE FRECUENCIA. PORCENTAJE dy	DESVIACION AL CUADRADO d^2
1910	10.8	—13	—140.4	169
1911	10.5	—12	—126.0	144
1912	10.2	—11	—112.2	121
1913	14.2	—10	—142.2	100
1914	13.8	— 9	—124.2	81
1915	15.3	— 8	—122.4	64
1916	14.6	— 7	—102.2	49
1917	15.4	— 6	— 92.4	36
1918	16.9	— 5	— 84.5	25
1919	15.4	— 4	— 61.6	16
1920	13.9	— 3	— 41.7	9
1921	15.1	— 2	— 30.2	4
1922	20.5	— 1	— 20.5	1
1923	17.4	0	0	0
1924	16.1	1	16.1	1
1925	15.5	2	31.0	4
1926	17.7	3	53.1	9
1927	18.0	4	72.0	16
1928	18.7	5	93.5	25
1929	17.5	6	105.0	36
1930	19.7	7	137.9	49
1931	19.9	8	159.2	64
1932	22.1	9	198.9	81
1933	22.0	10	220.0	100
1934	21.4	11	235.4	121
1935	23.1	12	277.2	144
1936	24.8	13	322.4	169

$$m = \frac{\Sigma\, dy}{\Sigma\, d^2}$$

$$= \frac{721.4}{1638}$$

$$= 0.4404$$

Tendencia de valores para:
(1) 1923 (punto intermedio)

$$= \frac{\Sigma\, y}{N}$$

$$= \frac{460.5}{27}$$

$$= 17.056$$

(2) 1910

$$= 17.056 + (-13)\ (m)$$
$$= 17.056 + (-13)\ (.4404)$$
$$= 17.056 + (-5.725)$$

$$= 11.331$$

(3) 1936

$$= 17.056 + (+13)\ (m)$$

$$= 17.056 + (+13)\ (.4404)$$

$$= 17.056 + (+5.725)$$
$$= 22.781$$

Al determinar matemáticamente los rasgos de una línea, debe elegirse la forma que le convenga. Puede ser recta o curva. Las líneas curvas pueden expresarse con una parábola, una hipérbola, una curva de interés compuesto con una forma más complicada. Para las series de la Tabla IX una línea recta parece que es la de tipo más conveniente. En la práctica se verá que el tipo de línea más común es el recto.

En la figura 21 se ha acomodado a las series de cifras de mortalidad, una línea recta a través de los cuadrados. Este tipo de línea es también conocido como una parábola de primer grado. Al computar una línea recta a través de los cuadrados, se pueden seguir uno o dos métodos comunes. El más sencillo, así como el más frecuentemente usado, queda ilustrado en la Tabla IX y en la figura 11. Se conoce como el método de los momentos.

1. En la Tabla IX se observará que los años y las cifras correspondientes en las series, están registrados en la primera y segunda columnas.

2. El punto medio de las series se localiza y las desviaciones se marcan negativamente en los primeros años, y positivamente en los últimos. En el problema de la Tabla IX el año de 1923 es el año intermedio o punto de origen.

3. Los valores y, se multiplican por las correspondientes desviaciones y los productos se colocan en la cuarta columna (dy).

4. Las desviaciones se elevan al cuadrado y se ponen en la quinta columna (d^2).

5. El valor del año intermedio (b interceptada) se obtiene sacando el promedio de valores de la columna y. En el problema.

$$\frac{\Sigma y}{N} = \frac{460.5}{27} = 17.056$$

6. Con objeto de determinar la inclinación (m) de la línea de cuadrados, es necesario substituir en la siguiente fórmula: $m = \frac{\Sigma dy}{d^2}$.

En el problema, $m = \frac{721.4}{1638} = .044$. Se verá que m puede ser positivo o negativo, según que la línea de los cuadrados muestre un movimiento hacia arriba o hacia abajo.

7. La ordenada de cualquier año en las series puede obtenerse por la fórmula $y = b + mx$. Debe hacerse notar que b representa la media de la columna Y, y es la ordenada que corresponde al año medio de las series. En el problema, $b = 17.056$. Al representar la línea recta de los cuadrados en una gráfica, no es necesario calcular las ordenadas de cada año en las series. Dos ordenadas, una de cada lado de la media, es todo lo que se necesita. Computemos, por lo tanto, las ordenadas para el primero y último año de las series. Para 1910, $Y = 17.056 + (.044)\ (-13) = 11.331$ y para 1936, $Y = 17.056 + (.044)\ (+ 13) = 22.781$. En la figura 11 las ordenadas para 1910 y 1936 fueron trazadas en la gráfica y después conectadas con la línea recta.

ERRORES ESTADISTICOS

Para ser realmente eficiente en el trabajo estadístico, se necesita desde luego un conocimiento profundo de las diversas técnicas, pero además es indispensable poseer cualidades tales como el sentido común, el buen juicio, un escepticismo saludable, objetividad, experiencia, y una amplia comprensión del tema que se estudia. Muchos de los errores más serios en el trabajo estadístico son de carácter no matemático. Este hecho a menudo es pasado por alto por el principiante que ha adquirido algunos conocimientos relacionados con las fórmulas y técnicas estadísticas. El análisis estadístico no es una simple función mecánica de aplicación de las fórmulas y funcionamiento de la máquina calculadora. El trabajo estadístico requiere un buen juicio, una actitud crítica y un pensamiento cuidadoso.

Con objeto de ponernos a cubierto de los errores estadísticos más comunes, sinteticemos algunas de las fuentes de error más características:

1) datos inadecuados y descuidados; 2) equivocaciones mecánicas; y 3) interpretaciones absurdas. [18]

Bajo el título de datos inadecuados y poco cuidadosos pueden mencionarse las siguientes fuentes de errores: 1) datos insuficientes; 2) muestras que no son representativas; 3) datos que no han sido deliberadamente falsificados por los informantes; 4) datos inadecuados que pueden resultar de una observación descuidada; y 5) standards y unidades de mensuramiento poco seguras.

Los errores mecánicos incluyen: 1) equivocaciones en los procesos matemáticos; 2) aplicación de fórmulas equivocadas; y 3) errores de copia.

Las falacias de interpretación más comunes son: 1) falta de consideración de todos los factores importantes; 2) ignorancia de la evidencia negativa; 3) correlación de causación equivocada; 4) comparación de datos no comparables; y 5) interpretaciones absurdas para servir a ideas preconcebidas y prejuicios.

Todo el que aspire a ejecutar una investigación estadística, debe familiarizarse con las cuatro sencillas reglas del procedimiento estadístico, formuladas por Adolph Quetelet: 1) No tener nunca ideas preconcebidas acerca de lo que las cifras van a probar; 2) No rechazar nunca un número que parece contrario a lo que se esperaba, simplemente porque se aparta del promedio aparente; 3) Pesar y registrar cuidadosamente todas las causas posibles que puedan intervenir en un evento y no atribuir a una lo que en realidad es el resultado de la combinación de varias; 4) No comparar datos que no sean completamente comparables.

PREGUNTAS Y SUGESTIONES PARA UN ESTUDIO POSTERIOR

1. Explicar las características principales de una unidad estadística.

2. ¿Qué es una variable? ¿Un atributo? ¿Una formación? ¿Una distribución de frecuencia?

18 Robert Emmet Chaddock, *op. cit.*, pp. 10-39; Manuel C. Elmer, *Social Statistics* (Estadísticas Sociales), pp. 234-244; I. S. Falk, *The Principles of Vital Statistics* (Los Fundamentos de las Partes Vitales de las Estadísticas), pp. 219-237; Arthur Newsholme, *The Elements of Vital Statistics* (Los Elementos de las Partes Vitales de las Estadísticas), pp. 527-541 y 598-606; Jerome B. Cohen, "The Misuse of Statistics" (El Abuso de las Estadísticas), *Journal of the American Statistical Association,* XXXIII, (diciembre de 1938), pp. 657-674.

3. ¿Cuáles son las diferencias entre las tablas de propósitos generales y las de propósitos especiales?

4. Desígnense las medias de los siguientes intervalos de clase: 30 a 34; 70 a 79; 16 a 19; 0 a 7; 0 a 2; 6 a 7.

5. Exprésese, en forma algebráica, la fórmula para computar la media de una distribución de frecuencia por el método abreviado.

6. Enumérense algunas de las características básicas de la media, la mediana y el modo.

7. ¿Qué entiende usted por desviación standard?

8. Defínase la asimetría positiva; la asimetría negativa.

9. ¿Qué es el coeficiente de variación? Póngase un ejemplo de esta aplicación.

10. Explíquese (a) el error standard y (b) el error probable.

11. ¿Qué es una muestra estadística? Explíquense tres procedimientos diferentes para seleccionar una muestra estadística.

12. Explíquese el concepto de correlación en el análisis estadístico

13. ¿Qué es correlación parcial? ¿Correlación múltiple?

14. Hágase una distribución de frecuencia de los grados recibidos por un gran número de estudiantes examinados, usando intervalos de clase de cinco.

15. Compútense porcentajes para la distribución y arréglense todos los datos en una tabla estadística con título, hileras, cabezas y otras características esenciales.

16. Constrúyase un histograma, un polígono de frecuencia y una curva uniforme de frecuencia.

17. Determínese la media, la mediana y el modo en una distribución de frecuencia.

18. Calcúlese la desviación standard.

19. Reúnanse los resultados de otro examen efectuado en la misma clase y compútese el coeficiente de correlación para las dos series de

grados. Primero constrúyase un diagrama de dispersión, y luego transfiéranse los datos a una carta de correlación standard.

20. Pónganse, en un diagrama de dispersión, dos series de datos que se consideren relacionados. Compútese el coeficiente de correlación.

21. De los datos del último ejemplo del *Statistical Abstract* (Resumen Estadístico), determínese la relación entre la cifra de mortalidad infantil y el por ciento de vehículos de motor, en la población de los 48 Estados. Explíquense los resultados.

22. Examínese el trabajo de Dennis H. Cooke, *Minimum Essentials of Statistics* (Mínimo Indispensable en las Estadísticas), e indíquese el valor del trabajo social y sociológico.

23. Examínese el trabajo de Elmer R. Enlow, *Statistics in Education and Psychology* (Estadísticas de educación y psicología), e indíquese hasta qué punto la estadística en este terreno puede tener algún valor para el sociólogo, el economista o el trabajador social.

24. Consúltese la obra de Albert E. Waugh, *Elements of Statistical Methods* (Elementos de los métodos estadísticos), y compárese con alguno de los textos primitivos. (Véase la Bibliografía para este capítulo.)