Las Técnicas de la Investigación Social

Por Pauline V. YOUNG.

CAPITULO XI

Técnica y Conceptos Básicos de la Estadística [1]

Por Calvin F. Schmid
Universidad de Washington

> "... la expresión cuantitativa del hecho social debe ser preferida para los propósitos científicos, siempre que pueda ser usada. Reduce la parcialidad individual al mínimo, permite la verificación por otros investigadores, reduce y, al mismo tiempo, hace evidente el margen de errores y reemplaza el significado menos exacto de las palabras descriptivas con la precisión de una anotación matemática." *Stuart A. Rice.*

A causa de la urgente necesidad de métodos más precisos y objetivos, los procedimientos estadísticos han estado ganando incesantemente mayor aceptación entre los sociólogos. Debe tenerse en cuenta que el desarrollo de toda ciencia se caracteriza por el grado en que los datos y procedimientos cuantitativos y exactos sobrepasan a la simple especulación y a las impresiones cualitativas. La astronomía, la física y la química, son

1 El autor debe a su colega Dr. Joseph Cohen el haber hecho una lectura crítica de este artículo y haberle ofrecido valiosas sugestiones.

citadas frecuentemente como ciencias "exactas", especialmente porque sus datos y su metodología son relativamente precisos y cuantitativos. No resulta aventurado decir que a medida que las ciencias sociales avanzan de los estados impresionistas y cualitativos, los métodos estadísticos se van haciendo cada vez más importantes. De hecho, en la actualidad, un conocimiento profundo, siquiera de los fundamentos de la estadística es parte indispensable del equipo del investigador en el campo de las ciencias sociales.

Un conocimiento de los conceptos estadísticos básicos, así como de la técnica empleada, es también necesario para una comprensión inteligente de la literatura corriente en las ciencias sociales. Generalmente los escritores suponen que el lector conoce el significado de términos tan elementales como media, mediana, graduación, desviación, tipo, correlación y confiabilidad. El provecho que se obtiene de leer estudios técnicos de esta clase aumenta en proporción de la comprensión de los métodos estadísticos.

El presente capítulo es un simple bosquejo de los principios y métodos más comunes del análisis estadístico. La selección y estudios de los tópicos se basa en el juicio del autor, sobre lo que él considera más significativo y fundamental para una discusión simple y concisa de este tema. Se ha esforzado por presentar el material de la manera más clara y simple que sea posible, sin tratar los procedimientos más avanzados y teóricos. No se necesita más que un conocimiento mínimo de las matemáticas para comprender los tópicos discutidos en este capítulo. Se presume que el lector conoce aritmética y rudimentos de álgebra. Debe observarse que la aplicación práctica de los procedimientos estadísticos para concretar los problemas, ha sido depurada y que todos los ejemplos ilustrativos se basan en datos reales. A causa del espacio limitado se ha hecho hincapié en la habilidad para el cómputo, pero sin excluir la interpretación. El principiante no debe tener la impresión de que las estadísticas son métodos mecánicos. Una visión y un juicio crítico, así como una imaginación fértil y constructiva son quizás más importantes en el trabajo estadístico, que la simple habilidad para computar.

UNIDADES ESTADISTICAS

Todo trabajo estadístico presupone la existencia de unidades. Las unidades se expresan en forma cuantitativa y pueden ser usadas para

contar o medir. Las unidades se totalizan, se multiplican, se dividen y se manejan en muchas otras formas y son de una importancia básica en la selección, análisis e interpretación de los datos estadísticos. Es indispensable que cada tipo de unidad sea definido cuidadosa y exactamente. Los resultados de un estudio pueden quedar completamente viciados si las unidades son vagas y ambiguas.

Las unidades estadísticas pueden ser clasificadas en cinco grupos principales: 1) seres naturales; 2) objetos producidos; 3) cualidades producidas; 4) unidades mensurables; y 5) unidades de valores pecuniarios. [2] La unidad "seres naturales" se usa para contar y se aplica a seres tales como una persona, una vaca, un árbol. Como ejemplos de "objetos producidos" podemos tomar una casa, un rancho, un automóvil y una máquina de escribir. Las "cualidades producidas" o atributos son especialmente comunes en las estadísticas sociales e incluyen unidades tales como un divorcio, un crimen, un dependiente, un desocupado y un ciudadano. Las "unidades mensurables" se usan para medir y no para contar. Ejemplos de unidades mensurables son el pie, la libra, el grado y el mes. Las unidades de "valores pecuniarios" representan una variedad de las unidades mensurables, pero en esta clasificación están limitadas a medir los valores financieros. El dólar, la libra y el franco son ejemplos de unidades de valor pecuniario.

Una unidad estadística satisfactoria debe poseer las cualidades siguientes: 1) propiedad; 2) claridad; 3) mensurabilidad; y 4) comparabilidad. [3] La propiedad de una unidad está determinada por el propósito del estudio. Una unidad apropiada para una clase de estudio, puede no serlo para otra. La claridad implica precisión y simplicidad en la definición. La definición de una unidad debe ser fácilmente comprensible y debe poseer siempre el mismo significado. Las unidades estadísticas satisfactorias deben llenar el requisito de mensurabilidad en el más amplio sentido del término. Las unidades deben expresarse en forma objetiva y cuantitativa puesto que se emplean para contar y medir. Al planear un estudio estadístico debe uno esforzarse en definir las unidades de tal manera que puedan compararse con unidades de otros estudios similares.

2 Robert Riegel, *Elements of Business Statistics* (Elementos de Estadística Comercial), pp. 139-159.

3 Harry Jerome, *Statistical Method* (El Método Estadístico); George A. Lundberg, *Social Research* (La Investigación Social), pp. 67-72.

TABULACION DE LOS DATOS ESTADISTICOS

En los capítulos precedentes han sido discutidos, con alguna extensión, diversos métodos para reunir datos estadísticos y de otra clase. Cuando se ha juntado una masa de datos, es necesario arreglar el material en un orden conciso y lógico. El procedimiento se conoce con el nombre de tabulación. Los datos reunidos pueden estar en cédulas usadas por enumeradores en el terreno estudiado, en cuestionarios que hayan sido contestados y enviados por los informantes o en formas tomadas de los archivos de alguna organización privada o de alguna oficina pública.

El primer paso para tabular los datos estadísticos es el elaborar un sistema detallado de clasificación. La clasificación es fundamental para cualquier clase de análisis científico. El esquema general de la clasificación casi siempre se determina antes de reunir los datos, pero rara vez se completa realmente antes de que se hayan juntado estos. La base de cualquier clasificación estadística se determina tanto por el problema mismo, como por los rasgos característicos de los datos. En la práctica se encontrará que, para la clasificación estadística, se siguen invariablemente alguno o algunos de los siguientes criterios: 1) geográfico, 2) cronológico o temporal, 3) cualitativo o atributivo y 4) cuantitativo. Si se elige el criterio geográfico como base de la clasificación, los datos deben ser organizados en términos de división geográfica como en estado, región, ciudad, país, etc. Algunas unidades de tiempo como el día, la semana, el mes o el año, pueden servir de base para la clasificación de ciertas clases de datos estadísticos. Algunas cualidades o atributos, tales como sexo, color, natalidad, ocupación y estado matrimonial, se usan también frecuentemente en las clasificaciones estadísticas. Cierto criterio cuantitativo, como el expresado por el tamaño o magnitud, es una base de clasificación particularmente común en los trabajos estadísticos. La distribución de la frecuencia que se discutirá con cierto detalle en la sección siguiente es un excelente ejemplo de este tipo de clasificación.

Después de que se haya terminado la clasificación, los casos o items individuales que se encuentran comprendidos en los datos reunidos, se distribuyen y se cuentan, de acuerdo con las diversas categorías de la clasificación. El proceso de distribución y recuento puede hacerse a mano o en máquina, según el número de casos o items que se necesite tabular,

así como según el dinero de que se disponga. Si el número de casos es relativamente pequeño, el trabajo de tabulación puede hacerse a mano. Si los datos están en tarjetas o en hojas pequeñas, pueden distribuirse en montones para contarlas directamente. Este sistema era el que se usaba para los proyectos en grande escala antes de que se introdujera la tabulación mecánica. Otro método muy común es el de poner cada caso en el compartimiento apropiado, en una hoja marcada con líneas, cuadros

PESO	TABULACION	FRECUENCIA
90-99	/	1
100-109	/	1
110-119	𝍸 ////	9
120-129	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	30
130-139	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	42
140-149	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	66
150-159	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 //	47
160-169	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	39
170-179	𝍸 𝍸 𝍸	15
180-189	𝍸 𝍸 /	11
199-199	/	1
200-209	///	3
	NUMERO TOTAL DE CASOS	265

Fig. 11. *Hoja para el trabajo de tabulación por distribución de la frecuencia.* Datos que representan los pesos de 265 estudiantes de primer año, de la Universidad de Washington.

u otros medios. La hoja marcada es un registro con columnas y espacios apropiados. La fig. 11 es un ejemplo de una hoja marcada para una distribución de frecuencia. Muestra la distribución, por peso, de 265 estudiantes nuevos de la Universidad de Washington. Si hay un gran número de items o casos, o si se trata de muchas formas tabulares detalladas, la tabulación a mano puede resultar tan laboriosa y tan cara, que se necesite

recurrir a la tabulación mecánica. En dicho procedimiento los datos de las formas originales se transfieren a tarjetas, haciendo una perforación en cada item. La perforación de las tarjetas se basa en un sistema mecánico. Después de que hayan sido perforadas, pasan primero por una máquina seleccionadora y después por otra tabuladora. Con un equipo tabulador tan moderno como el de las máquinas Hollerith y Powers, es posible tabular decenas de miles de items en un día de trabajo. [4]

DISTRIBUCION DE FRECUENCIA

Los datos estadísticos que están clasificados de acuerdo con la magnitud o tamaño, frecuentemente se arreglan en forma de una distribución de frecuencia. La figura 11 es una simple ilustración de una distribución de frecuencia. Se observará que los intervalos expresados en libras se indican a la izquierda y el número de casos o frecuencia de cada intervalo está señalado a la derecha.

A fin de aclarar la subsecuente discusión acerca de las distribuciones de frecuencia es necesario que el estudiante comprenda los siguientes conceptos:

1. Una *variable* es cualquier cantidad o característica que puede poseer diferentes valores numéricos. El tiempo, la edad, los salarios, y los puntos, en las pruebas de inteligencia, son ejemplos de variables.

2. Los *valores variantes* se refieren a los valores indicados por una variable. Cuando se consideran en conjunto constituyen *series.*

3. Generalmente se hace una distinción entre las variables que son *continuas* y las que son *discontinuas* o *distintas.* Una variable continua tiene un número ilimitado de valores posibles, que van desde lo más bajo hasta lo más alto, mientras que una variable discontinua o distinta tiene límites o fronteras en los valores variantes y por lo tanto no es capaz de tener una referencia indefinida. Todo valor de una variable distinta es diferente y separado, mientras que los valores de una variable continua se encadenan entre sí a través de una gradación paulatina. Edad,

4 Para un estudio completo del problema de la tabulación mecánica, véase G. W. Baehne, *Practical Application of the Punched Card Method in Colleges and Universities* (Aplicación Práctica del Método de las Tarjetas Perforadas en Colegios y Universidades).

peso y temperatura son ejemplos de variables continuas y gente, casas y automóviles son ejemplos de variables discontinuas.

Al construir una distribución de frecuencia es necesario determinar, desde el principio: 1) el número de intervalos de clase que se usarán; 2) el tamaño de los intervalos; y 3) la designación de los mismos.

1. Ordinariamente no debe haber menos de ocho o diez ni más de dieciocho o veinte intervalos de clase, según la naturaleza de los datos y el número de casos estudiados. Con objeto de entender claramente los datos originales, los items individuales se arreglan frecuentemente en orden de magnitud ascendente o descendente. Esta clasificación se conoce como una *formación.* Después de anotar el valor más alto y el más bajo, así como los rasgos característicos de los datos, el número de intervalos puede ser determinado más fácilmente.

2. Deben hacerse todos los esfuerzos para que los intervalos sean de tamaño uniforme. Los intervalos no deben ser tan reducidos que pierdan las ventajas de la sumarización, ni tan grandes que oculten las características más importantes de la distribución. Además, si los intervalos de clase son demasiado pequeños, pueden presentarse intervalos vacantes o en blanco. Si se desea hacer comparaciones entre datos similares, es recomendable que se seleccionen intervalos del mismo tamaño para todas las distribuciones. Siempre que sea posible, los intervalos de clase deben representar divisiones numéricas comunes y convenientes, tales como cinco o diez, mejor que divisiones difíciles como tres o siete.

3. Después de que se haya determinado el tamaño de los intervalos es muy importante que se designen claramente en la tabla de frecuencia. Cada intervalo debe tener límites superior e inferior definidos, y debe expresarse en tal forma que se excluya toda posibilidad de mala interpretación o confusión. Los intervalos de clase se indican generalmente por el primero y el último número del intervalo. Así, por ejemplo, 20 a 24 representa un intervalo de cinco; 10 a 19 un intervalo de diez; 2 a 3 un intervalo de dos; y 9 a 11 un intervalo de tres. Debe hacerse la distinción entre los límites expresados y los límites reales de un intervalo de clase. La separación real entre los intervalos depende de la precisión en la medida de los datos originales. Si las medidas se dividen hasta un grado centesimal los límites reales de un intervalo, digamos, de diez

serán de 9.99. Como se observará en la fig. 12, es una práctica comúnmente aceptada en las estadísticas expresar los límites superiores de los intervalos de clase como números integrales. En las series diferentes, los intervalos de clase se señalan por sí mismos, puesto que cada unidad determina los límites de su propio intervalo.

Todo intervalo de clase tiene un punto intermedio que está colocado entre el punto máximo y el mínimo. La fig. 12 ilustra los puntos intermedios en un intervalo regular y en uno fraccionado. Se observará que el punto intermedio del intervalo 10 a 19 es 15 y que el del intervalo 10 a 14 es 12.5. Después de que se hayan determinado los intervalos de clase, es relativamente fácil contar el número de casos que caben en cada intervalo.

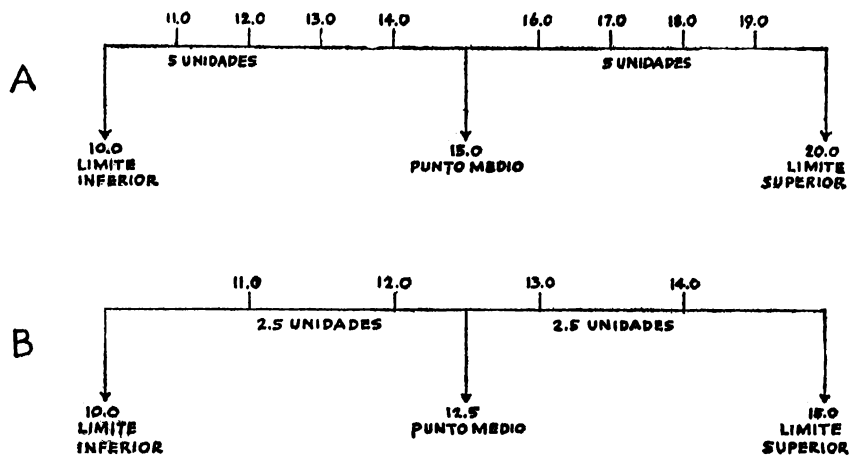

Fig. 12. *Clase de intervalos y puntos intermedios.* A) Clase de intervalos regulares. B) Clase de intervalos fraccionados.

TABLAS ESTADISTICAS

Después que los datos hayan sido tabulados, el paso siguiente es arreglar, por lo menos una parte, en tablas estadísticas. Las tablas estadísticas se consideran con la "síntesis de la estadística." No importa qué tipo de problema estadístico se esté investigando, invariablemente se necesitan las tablas. Por lo tanto, es de extrema importancia para el que estudia las investigaciones sociales, tener una idea clara de la construcción de las tablas. Las ventajas de presentar los datos estadísticos en

forma tabular son tan evidentes, que todo comentario resulta superfluo. Sin embargo, puede hacerse notar que: 1) las tablas estadísticas ahorran espacio y reducen las explicaciones y las descripciones a un mínimo; 2) la visualización de las relaciones y del proceso de comparación se facilita mucho; 3) los datos tabulados pueden recordarse más fácilmente que los que no lo están; 4) en una tabla es más fácil sumar los items y extraer los errores u omisiones; y 5) las tablas estadísticas sirven de base para las computaciones. [5]

En toda discusión de las reglas y prácticas que deben seguirse en la construcción de tablas estadísticas, es importante hacer una distinción entre la tabla de propósitos generales y la de propósitos especiales. El primer tipo de tablas es conocido también como tipo original, primario o de referencia y el último se ha designado como analítico, sumario, interpretativo, derivativo o secundario. La tabla de propósitos generales está destinada a incluir grandes cantidades de datos originales en forma conveniente y accesible, mientras que las tablas de propósitos especiales tienden a ilustrar o demostrar determinados puntos en un análisis estadístico, o hacer resaltar las relaciones significativas de los datos. La diferencia entre las tablas de propósitos generales y las de propósitos especiales está bastante bien ilustrada en los numerosos reportes publicados por el United States Bureau of the Census. Virtualmente todas las tablas referentes a población, estadísticas vitales, agricultura, manufactura y otros temas, son del tipo de propósito general. Representan modelos extensivos de información estadística. Los ejemplos de las tablas de propósitos especiales se encuentran en monografías y artículos en los que se ha dado particular importancia al estudio estadístico. En la discusión presente nuestro principal interés es para la tabla de propósitos especiales.

Las reglas y procedimientos para construir las tablas estadísticas, no están completamente estandarizadas, pero hay ciertos principios generales aceptados a los cuales debemos adherirnos íntimamente.

1. Toda tabla debe tener un título. Dicho título debe representar una descripción sucinta del contenido de la tabla y debe hacerla inteligible sin necesidad del texto. El título debe ser claro, conciso y adecuado y

5 Harry Jerome, *op. cit.*, pp. 28-36; Horace Secrist, *An Introduction to Statistical Methods* (Introducción a los Métodos Estadísticos), pp. 134-138.

debe responder a las preguntas ¿*qué*?, ¿*dónde*? y ¿*cuándo*? Debe colocarse siempre sobre el cuerpo de la tabla.

2. Toda tabla debe ser identificada con un número para facilitar la referencia. El número puede ser arábigo o romano y puede estar encima del título o colocado en la primera línea de él.

3. Las cabezas de columna y los talonarios deben ser claros y breves.

4. Todas las notas explicatorias que se refieren a la tabla, deben estar directamente debajo de ella y con objeto de evitar cualquier posible confusión con las notas del texto deben usarse símbolos especiales como el asterisco, la daga, la doble daga, la marca de sección, etc.

5. Si los datos de una serie de tablas han sido obtenidos de fuentes diferentes, es aconsejable que se indiquen dichas fuentes en un sitio debajo de la tabla.

6. Con objeto de hacer resaltar la significación relativa de ciertas categorías, pueden usarse diferentes clases de tipos, espacios y colocación.

7. Se usan generalmente marcas que guían al ojo a través de la columna, excepto cuando los datos están tan cerca que no son necesarias.

8. Es importante que todas las cifras de la columna estén propiamente alineadas. Los puntos decimales y los signos de mínimo deben estar también en un alineamiento perfecto.

9. Usualmente las columnas están separadas entre sí por medio de líneas. Dichas líneas ponen de manifiesto más claramente las relaciones de los datos y hacen que la tabla sea más legible y atractiva. Siempre se tiran líneas arriba y abajo de la tabla, lo mismo que entre las columnas. No es necesario tirar también líneas a los lados de la tabla.

10. Algunas veces las columnas se numeran para facilitar la referencia.

11. Los items de naturaleza diversa y excepcional se colocan generalmente en la última columna de la tabla.

12. Como resultaría muy confuso leer una tabla larga cuando todas las columnas y las líneas son de un espacio, es una práctica común agrupar los talones de acuerdo con las cabezas de columna. Generalmente la agrupación de cuatro o cinco es muy satisfactoria.

13. Deben evitarse las abreviaturas, siempre que sea posible, lo mismo que las notas abreviadas.

14. El arreglo práctico de las grandes clases de la tabla depende de los hechos y relaciones que deban considerarse como de mayor importancia. No debe olvidarse, sin embargo, que una tabla estadística debe hacerse tan lógica, clara, completa y simple como sea posible.

15. Las columnas y las hileras que deben compararse entre sí deben estar cerca una de otra.

16. Los totales pueden ponerse a la cabeza o al final de la tabla. Debe hacerse notar a este respecto, que la parte más notable de la tabla es la esquina superior de la izquierda.

17. El arreglo de las categorías en una tabla, puede ser cronológico, alfabético o estar de acuerdo con la magnitud. Las series cronológicas pueden leerse de arriba a abajo o de abajo para arriba, de derecha a izquierda o de izquierda a derecha, según la importancia de los datos que figuran en ellas. Con excepción de las series de tiempo, las categorías generalmente se arreglan de acuerdo con la magnitud. Si el orden de importancia no tiene una significación particular, puede emplearse el orden geográfico. El orden alfabético se usa mucho más frecuentemente en las tablas de propósitos generales que en las de propósitos especiales.

PRESENTACION GRAFICA DE LAS DISTRIBUCIONES DE FRECUENCIA

Al analizar o registrar los datos de distribuciones de frecuencia, es a menudo deseable, usar alguna forma de presentación gráfica. Hay tres clases de gráficas que se usan para presentar las distribuciones de frecuencia: 1) el histograma o diagrama en columna; 2) el polígono de frecuencia; y 3) la curva uniforme de frecuencia. Todas estas gráficas están dibujadas en coordenadas rectangulares. El eje X o eje de las abscisas

representa los intervalos de clase y el eje Y, o eje de las ordenadas, representa las frecuencias. El eje horizontal se lee de izquierda a derecha y el eje vertical de abajo a arriba.[6]

1. *Histogramas.* Al construir un histograma, las líneas verticales se elevan hasta los límites de los intervalos de clase, formando series de rectángulos continuos. La altura de cada rectángulo representa las respec-

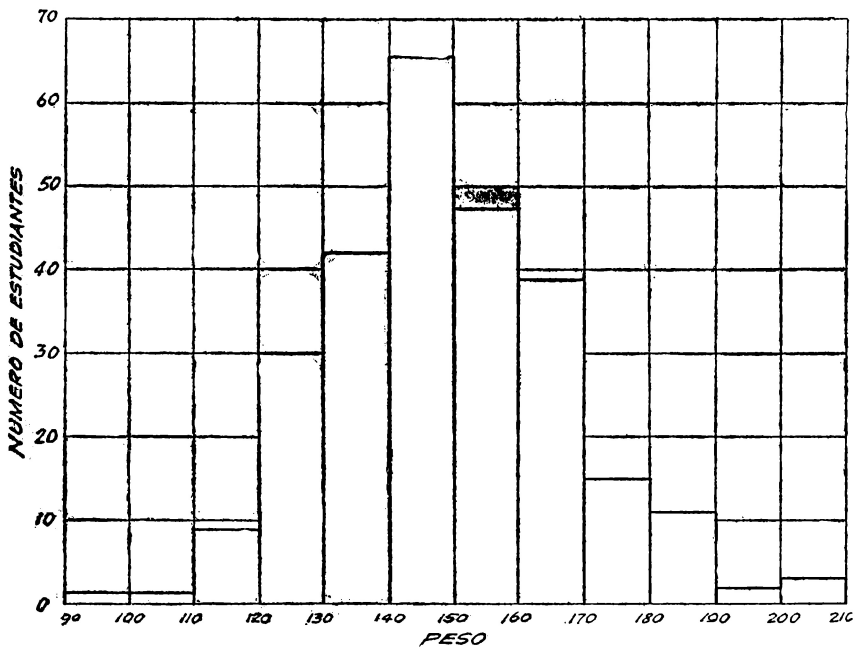

Fig. 13 *Histograma.* Basado en los datos de la Tabla I.

tivas frecuencias de clase. La fig. 13 es una ilustración de un histograma. Los histogramas son especialmente valiosos para representar series discretas.

2. *Polígonos de frecuencia.* Al construir un polígono de frecuencia, el término apropiado de cada clase se encuentra en el punto medio del intervalo y los puntos extremos quedan conectados por líneas rectas. Si en la misma gráfica se representan dos o más series, las curvas pueden

6 Para una discusión más detallada de los principios de la representación gráfica, véase el Capítulo XII.

trazarse de acuerdo con reglas diferentes. En la fig. 14 se observará que un polígono está representado con una línea continua y otro con una punteada. Si el número total de casos en dos series es diferente, las frecuencias a menudo se reducen a porcentajes. En este tipo de cartas, las líneas verticales representan más bien porcentajes que frecuencias absolutas. El polígono de frecuencia es particularmente apropiado para representar series continuas.

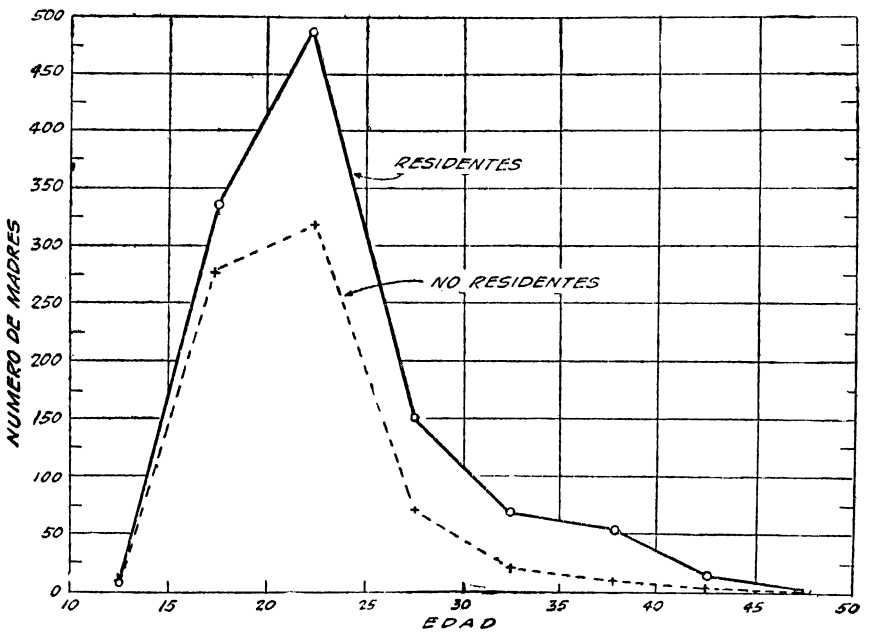

Fig. 14. *Polígonos de frecuencia.* Basados en los datos originales de Calvin F. Schmid, *Mortality. Trends in the State of Minnesota* (Curso de la Mortalidad en el Estado de Minnesota), pp. 273-275. Imprenta de la Universidad de Minnesota, 1937.

3. *Curvas uniformes de frecuencia.* Como muchas distribuciones de frecuencia se basan en muestras relativamente pequeñas, algunas veces es conveniente representar los datos por medio de una curva uniforme, más bien que con un polígono. Al uniformar una curva de frecuencia se presume que las fluctuaciones menores en la distribución se deben a un número relativamente escaso de ejemplos en observación.

Si aumenta el número de casos que se toman como muestra, las irregularidades tienden a desaparecer y la naturaleza real de la distribución

se hace más aparente. Teóricamente, de la misma manera se supone que la curva uniforme representa más verdaderamente las características generales de los datos originales. Sin embargo, nunca es suficiente la insistencia que debe hacerse en que el principiante sea muy cuidadoso al uniformar las curvas de distribuciones de frecuencia.

De las diferentes técnicas usadas para uniformar las curvas de frecuencia, el siguiente método gráfico es el más satisfactorio para propósitos prácticos: [7]

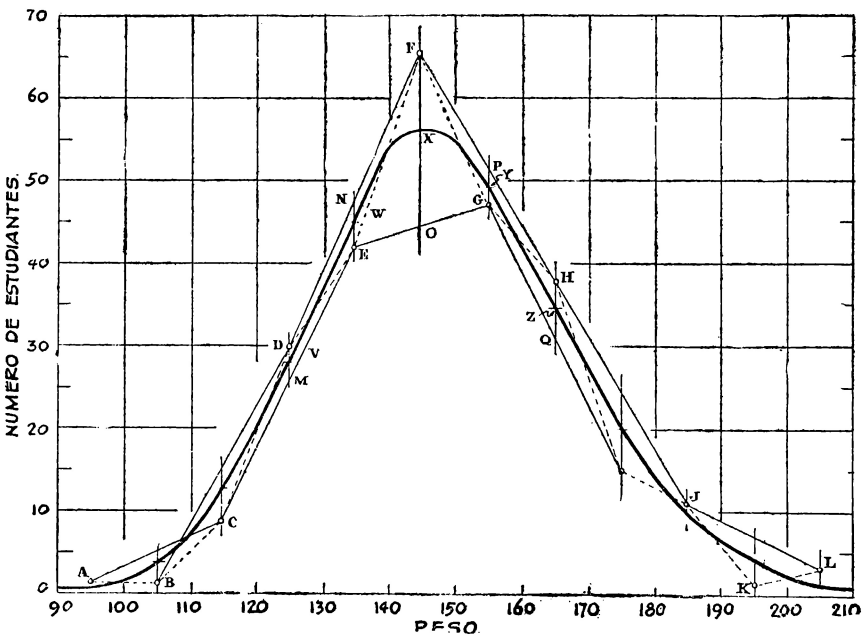

Fig. 15. *Uniformando una distribución de frecuencia.* Datos basados en la Tabla I.

1. El primer paso es dibujar un polígono de frecuencia de los datos. En la fig. 15 el polígono original de frecuencia se presenta por medio de una línea delgada. Los puntos de referencia están representados por pequeños círculos y designados por las letras de la A a la L inclusive.

7 L. L. Thurstone, *The Fundamentals of Statistics* (Los Fundamentos de las Estadísticas), pp. 39-44.

2. Todos los otros puntos de referencia están conectados por una línea recta. Esto es, los puntos A y C, B y D, C y E, D y F.

3. Hay líneas verticales que atraviesan cada uno de los puntos de referencia y que interceptan la conexión de las líneas AC, BD, CE, DF, etc.

4. Los puntos centrales entre los puntos de referencia y la conexión de las líneas AC, BD, CE, DF, se determinan por inspección. Como ilustración a este respecto obsérvense las líneas verticales DM, NE, FO, PG y HQ, en la fig. 15. Los puntos intermedios de las líneas respectivas se designan con las letras V, W, X, Y y Z.

5. El último paso es conectar los puntos intermedios consecutivos, tales como V, W, X, Y y Z de las líneas verticales por medio de una curva uniforme.

MEDIDAS DE DIMENSIONES REGULARES.—PROMEDIOS

Las distribuciones de frecuencia son extremadamente valiosas para sintetizar grandes masas de datos, pero el proceso de sintetización puede llevarse mucho más adelante reuniendo las características de series enteras en unas cuantas cifras significativas y si es posible, en una sola.

Estas cifras tienen el carácter de promedios y representan los valores típicos de una variable. Los promedios tienen un lugar muy importante en todos los tipos de trabajos estadísticos. De hecho, la estadística ha sido llamada "la ciencia de los promedios." Hay muchas clases de ellos, pero nosotros consideramos solamente tres, a saber: 1) el promedio aritmético o medio; 2) el mediano; y 3) el de grados.

Promedio aritmético. Quizás el más familiar de todos es el promedio aritmético. Es relativamente fácil de calcular y se emplea mucho en las investigaciones estadísticas. Si se conoce la medida de cada item en las series, el promedio puede obtenerse sumando las medidas y dividiendo el resultado entre el número de items. Si cinco estudiantes reciben calificaciones de 60, 75, 86, 88 y 96 respectivamente, el promedio será el resultado de la suma de las calificaciones, dividido entre cinco:

$$\frac{405}{5} = 81$$

El procedimiento que se emplea para calcular el promedio de datos que no han sido agrupados, puede expresarse en términos algebráicos a través de la siguiente fórmula:

$$X = \frac{\Sigma m}{N}$$

X es el promedio.

Σ (La S mayúscula en griego, o sigma) es el signo de suma convencional. No representa una cantidad separada, sino que indica "la suma de" lo que siga.

m representa la medida o tamaño de los casos o items separados.

N es el número total de items o casos.

Si hay relativamente pocos casos, digamos menos de 25 o 30, no es muy difícil tratar los items individualmente y computar el promedio de acuerdo con la fórmula anterior. Si hay muchos casos, más de 30, y si es necesario computar más tarde otras medidas de tendencia central o dispersión, generalmente es más fácil arreglar los datos de una distribución de frecuencia y calcular el promedio en esta forma. Hay dos métodos comunes para calcular el promedio de una distribución de frecuencia: 1) el método "largo" y 2) el método de "presunción" o "recortado." Como este último es muy superior al método "largo" y como se usa también para calcular la desviación standard, el coeficiente Pearsoniano de correlación, será el único que se considere en la presente discusión.

TABLA I

Cálculo del promedio (X) de una distribución de frecuencia. Datos que representan el peso de 265 estudiantes de la Universidad de Washington.

Intervalo de clase (Peso)	f	d	fd
90- 99	1	—5	— 5
100-109	1	—4	— 4
110-119	9	—3	—27
120-129	30	—2	—60
130-139	42	—1	—42
140-149	66	0	0
150-159	47	1	47
160-169	39	2	78
170-179	15	3	45
180-189	11	4	44
190-199	1	5	5
200-209	3	6	18
	N = 265		
	Σ fd = 237-138		
	= 99		

$$X = g + \frac{\Sigma fd}{N} \quad (i)$$

$$= 145 + \left(\frac{99}{265}\right) (10)$$

$$= 145 + (.3736)\ (10)$$

$$= 145 + 3.74$$

$$= 148.74$$

Los pasos que se siguen en el proceso de cálculo del promedio de una distribución de frecuencia, de acuerdo con el método recortado son los siguientes:

1. Arreglar los datos de una distribución de frecuencia y determinar el número de éstos. Del problema ilustrativo en la Tabla I se deduce que los intervalos de clase representan los pesos de 265 estudiantes de la Universidad de Washington, los cuales comprenden la primera columna y las frecuencias la segunda.

2. Por una simple inspección encuéntrese el intervalo de la distribución que tiene más probabilidades de contener el promedio. Por lo que se refiere a los resultados finales, cualquier intervalo podría servir, pero con objeto de mantener el tamaño de los números en el proceso de computación en un mínimo, debe tenerse cuidado de seleccionar un intervalo

de clase que esté lo más cerca posible del que contiene el promedio. El punto medio del intervalo elegido se designa en la fórmula con la letra g. En el problema de la Tabla I el intervalo 140 a 149 se supone que contiene el promedio y, por lo tanto, g = 145.

3. En la tercera columna las desviaciones (d) del promedio supuesto van siendo marcadas consecutivamente en pasos o intervalos. Las que están arriba del promedio elegido se designan con más (+) y las que están abajo como menos (—).

4. La siguiente operación consiste en multiplicar cada frecuencia por su correspondiente desviación (d). Los productos se ponen en la cuarta columna (fd). Naturalmente que debe tenerse cuidado de observar los signos.

5. Encuéntrese la suma algebraica (Σ) de fd. En la Tabla I, Σ fd neg = —138 y Σ fd pos = + 237.

Por lo tanto la suma algebraica de las figuras en la columna fd es: Σ fd = + 99.

6. Dividir Σ fd entre N. En el problema tenemos $\frac{99}{265} = .3736$ o quitando la cuarta cifra decimal por carecer de significación 0.374. El cociente obtenido por medio de esta operación da la corrección en términos de intervalos de clase. Algunas veces $\frac{\Sigma \text{ fd}}{N}$ puede ser positivo y otras negativo, según la posición que tenga el medio elegido.

7. Con objeto de obtener la corrección en términos de unidades reales en la distribución multiplíquese $\frac{\Sigma \text{ fd}}{N}$ por el tamaño del intervalo de clase (i). En el problema de la Tabla I, 0.374 está multiplicado por 10. El resultado es 3.74.

8. El último paso del proceso es agregar algebraicamente g y $\frac{\Sigma \text{ fd}}{N}$ (i)

En el problema $g = 145$ y $\frac{\Sigma fd}{N}(i) = +3.74$, por lo tanto la media de distribución es 148.74 libras.

Mediana. Esta es otra forma simple de emplear los promedios para medidas de tendencia central. Muchos estadísticos han definido el promedio mediano como el tamaño del item intermedio, cuando los items están ordenados en orden de magnitud. Esto significa, desde luego, que hay tantos items arriba, es decir mayores que el central, como abajo hay menores. Si el número de items es igual, el punto intermedio se toma como la cifra aritmética de los items centrales. Este concepto que estamos discutiendo no debe considerarse como una mediana genuina sino como el valor de un item medio en un caso medio. El concepto "mediana" debe reservarse para las distribuciones de frecuencia y no para las simples ordenaciones. Si se hace esta distinción, entonces el punto intermedio puede ser definido como el punto en una escala de la variable (el punto en la escala —X— en una gráfica de frecuencia) que divide la distribución en dos partes iguales.

En una distribución de frecuencia el punto intermedio se deriva de la interpolación de una de las clases de la distribución. Hay dos métodos de interpolación: 1) el aritmético y 2) el gráfico. El método aritmético es el más generalmente usado. Sin embargo, el método gráfico debe considerarse también en la discusión presente puesto que ayuda al principiante a entender más claramente la significación del concepto.

Sinteticemos primero los pasos que hay que dar para derivar el punto mediano (χ) de acuerdo con el método aritmético.

1. La primera y segunda columnas en la Tabla II son idénticas a las de la Tabla I.

2. La tercera columna representa una frecuencia de distribución acumulativa. Las frecuencias, en esta columna están acumuladas desde el principio de la distribución. Se encontrará que la frecuencia acumulativa de cualquier intervalo de clase particular representa la suma de las frecuencias de éste y todos los intervalos de clase precedentes.

3. Con objeto de determinar el intervalo en que se encuentra el punto intermedio, el próximo paso es dividir N por 2 $\left(\frac{N}{2}\right)$. En la Tabla

II, $\frac{N}{2} = \frac{265}{2} = 132.5$. Se observará que 132.5 cae dentro del intervalo de clase 140-149.

TABLA II

Cálculo de la mediana (χ). Datos que representan el peso de 265 estudiantes del sexo masculino, del primer año de la Universidad de Washington.

Intervalos de clase (Peso)	f	Acumulativo f "Menos que"
90- 99	1	1
100-109	1	2
110-119	9	11
120-129	30	41
130-139	42	83
140-149	66	149
150-159	47	196
160-169	39	235
170-179	15	250
180-189	11	261
190-199	1	262
200-209	3	265

$N = 265$

$\frac{N}{2} = \frac{265}{2} = 132.5$

$$\chi = l + \left(\frac{w}{f}\right)(i)$$

$$= 140 + \left(\frac{132.5\text{-}83}{66}\right)(10)$$

$$= 140 + \left(\frac{49.5}{66}\right)(10)$$

$$= 140 + (.750)\ (10)$$

$$= 140 + 7.50$$

$$= 147.5$$

4. Determinar cuántos casos se requieren en este intervalo particular para alcanzar $\frac{N}{2}$ o sea 132.5. Esto se consigue restando el número de casos que quedan abajo del punto intermedio de $\frac{N}{2}$. Haciendo la sustitución en la presente ilustración tenemos 132.5 — 83 = 49.5. Este número ha sido designado con la w en la fórmula.

5. Dividir w entre el número de casos (f) en el intervalo de clase mediano. En la ilustración de la frecuencia del intervalo de clase en la cual se encuentra, la mediana es 66. Por lo tanto $\frac{49.5}{66} = .750$

6. Multiplíquese este cociente por el tamaño del intervalo de clase (i). En el problema, i es 10 y 10 x .750 = 7.50

7. Agréguese este producto al límite mínimo (l) del intervalo mediano. Se observará que 140 es límite mínimo del intervalo mediano en el problema ilustrativo y 140 + 7.50 = 147.5.

Para derivar un punto mediano por interpolación gráfica es necesario construir bien un tipo "inferior" o "superior" (o ambos) de curva de suma u ojiva. Los datos del tipo "inferior" se acumulan desde el principio de la distribución de frecuencia y los del tipo "superior" desde el fin. La escala —X de la gráfica de frecuencia acumulativa es la misma que la de la gráfica de frecuencia simple, pero la escala —Y difiere en que los valores deben ser de calidad más amplia. Al dibujar las frecuencias acumulativas de las de las diferentes clases, los puntos intermedios de los intervalos no se usan nunca. Para el tipo "inferior" de curva, se emplean los límites superiores de los intervalos de clase y para el tipo "superior" los inferiores. Como se verá en la figura 16 la típica ojiva suave tiene las características generales de una S, alargada.

Para calcular el punto medio de cualquier distribución $\frac{N}{2}$ se coloca en el eje vertical de la gráfica en una línea de interpolación paralela al eje horizontal que intercepta las curvas. Se tira una línea perpendicular a la escala —X. El punto en que dicha línea corta a la escala X es la mediana. Se observará en la fig. 16 que las dos ojivas se interceptan en un punto que divide la distribución en dos partes iguales. Por supuesto que este punto es el mismo que se ve atravesado por la línea de interpolación.

Cuartiles. Las técnicas para calcular los cuartiles son muy semejantes a las que se usan para derivar las medianas. Sin embargo, debe entenderse que los cuartiles no son medidas de tendencia central. Los cuartiles,

deciles y otras medidas de esta clase deberían incluirse, más lógicamente, dentro del título de variabilidad.

Debe hacerse notar que el punto mediano divide a la distribución de frecuencia en dos partes iguales. Por otra parte, los cuartiles la dividen en cuatro partes. Lo mismo que la mediana, los cuartiles, representan puntos en la escala de variables. Uno de los tres puntos es la mediana,

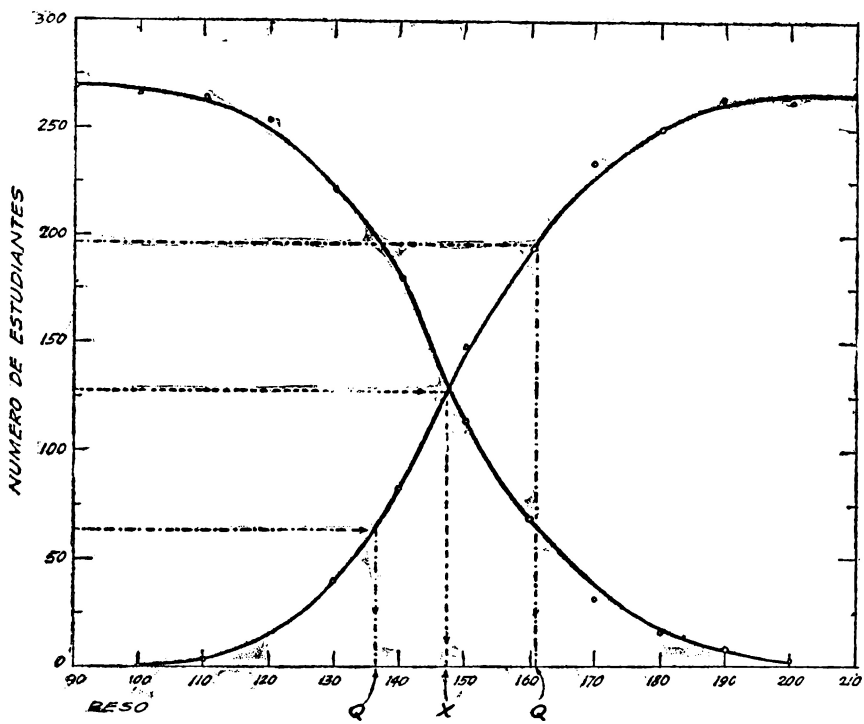

Fig. 16. *Curvas de frecuencia acumulativa u ojiva ilustran la técnica gráfica de interpolación para las medianas y cuartiles.* Basada en los datos de la Tabla I.

que puede designarse realmente con Q_2, o el segundo cuartil. Los cuartiles pueden computarse, tanto por procedimientos algebráicos como gráficos.

Los procedimientos que se siguen para calcular los cuartiles, de acuerdo con el método algebraico, son los siguientes:

1. El arreglo de los intervalos de clase, las frecuencias simples y las frecuencias acumulativas son iguales a las de la Tabla II.

2. Con objeto de localizar los intervalos que contienen los tres cuartiles, primero se divide N entre 4. $\frac{N}{4}$ localiza Q_1 (el primer cuartil); $\left(\frac{N}{4}\right)(2)$ localiza Q_2 (el cuartil segundo o medio); y $\left(\frac{N}{4}\right)(3)$ localiza Q_3 (el tercer cuartil).

3. Después de que los intervalos de clase que contienen los cuartiles hayan sido localizados, se aplica la misma fórmula de interpolación para el punto medio. A modo de ilustración, computamos Q_1 con los datos de la Tabla II: Primero $\frac{N}{4} = \frac{265}{4} = 66.25$. Por lo tanto, observamos que Q_1 se encuentra en el intervalo de 130 a 139.

Substituyendo en la fórmula $Q_1 = l + \left(\frac{w}{f}\right)(i)$ tenemos:

$$Q_1 = 130 + \left(\frac{25.25}{42}\right)(10) = 130 + 6.01 = 136.01.$$

La Fig. 16 ilustra el método gráfico de calcular los cuartiles a través de la suma de las curvas. Por ejemplo, con objeto de derivar Q_1, el valor $\frac{N}{4}$ se coloca en la escala vertical de la carta y, a través de la ojiva "inferior" se tira una línea horizontal de interpolación. Luego se traza otra línea perpendicular al punto de intersección de la línea de interpolación y la ojiva. Se verá, en la Fig. 16 que Q_1 corresponde aproximadamente a 136.0.

Modo. En las series simples el modo es el tamaño de la medida que se presenta más frecuentemente. En la siguiente serie de valores: 15, 17, 18, 22, 24, 25, 25, 25, 27, 28, 28, 30, y 31, el modo es 25 porque esta cifra se representa más que cualquiera otra. En una distribución el modo es el punto de la escala de variables en que la frecuencia es mayor. El

principiante no debe olvidar nunca que en una distribución de frecuencia, el punto medio, la mediana y el modo, son valores en la escala de variables. En una curva de frecuencia quedarán representados por puntos en la escala-X.

En la práctica se encontrará que las distribuciones de frecuencia a menudo tienen más de un modo. Las distribuciones que no tienen más que uno se denominan unimodales. Las que tienen dos, bimodales, las que tienen tres, trimodales y las de más, multimodales.

TABLA III

Cálculo del Modo (X). Datos que representan el peso de 265 estudiantes del primer año de la Universidad de Washington.

Clase de intervalo	f
90- 99	1
100-109	1
110-119	9
120-129	30
130-139	42
140-149	66
150-159	47
160-169	39
170-179	15
180-189	11
190-199	1
200-209	3

$$X = l + \left(\frac{f_2}{f_2 + f_1}\right)(i)$$

$$= 140 + \left(\frac{47}{47 + 42}\right)(10)$$

$$= 140 + \left(\frac{47}{89}\right)(10)$$

$$= 140 + 5.3$$

$$= 145.3$$

Hay varios métodos que pueden ser usados para derivar el modo en una distribución de frecuencia, pero ninguno es enteramente satisfactorio o universalmente aceptado. En el presente estudio mencionaremos únicamente dos métodos. Primero, el que se basa en la inspección, por medio de la cual se determina el modo-clase y el punto central del intervalo es considerado el modo. Este modo se designa frecuentemente como el modo *inspeccional* o *crudo*. El segundo método se basa en la interpolación. Después de que haya sido seleccionado el modo-clase, la localización exacta del modo dentro de la clase se determina por medio de la interpolación.

Este procedimiento es más factible, cuando se aplica a variables continuas en las cuales la distribución es casi normal. La fórmula se basa en la suposición de que la localización exacta del modo en la clase de modo está determinada por las frecuencias de las dos clases adyacentes. Si las respectivas frecuencias de las clases adyacentes son idénticas, entonces el modo es el punto central del intervalo de clase de modo. Si las frecuencias de las clases adyacentes son diferentes, entonces el modo es la dirección que lleva la frecuencia más grande, a partir del punto central de la clase de modo.

Los pasos que se dan para derivar el modo, de acuerdo con este método, son los siguientes:

1. Encontrar el modo de clase por inspección [8].

2. El segundo paso es establecer el límite inferior del modo de clase. En el problema de la Tabla III es 140.

3. Dividir la frecuencia de la clase adyacente abajo del modo de clase (f_2) entre la suma de las frecuencias de ambas clases adyacentes ($f_2 + f_1$). En el problema: $\dfrac{f_2}{f_2 + f_1} = \dfrac{47}{47 + 42} = \dfrac{47}{89} = 0.53.$

4. Multiplicar el cociente que se ha derivado de esa manera (0.53), por el tamaño del intervalo de clase (i). En el problema, 0.53 está multiplicado por 10 y el producto es 5.3.

8 Algunas veces es difícil localizar el modo de clase por la simple observación, puesto que las frecuencias de varias clases pueden ser muy parecidas. Entonces el modo de clase puede ser determinado por un proceso de agrupación y reagrupación de los intervalos de clase de la distribución. Los intervalos se agrupan primero en parejas, comenzando por la parte de arriba de la distribución. El paso siguiente es bajar un intervalo y repetir el proceso. Después se agrupan los intervalos en tres comenzando, como antes, por la parte de arriba de la distribución. Después de bajar un intervalo cada vez, se agrupan en cuatro y el proceso se repite. Por medio de este procedimiento de agrupación y reagrupación se encontrará que un intervalo tiende al modo de clase más frecuentemente que cualquiera de los otros. Para una discusión más amplia de este punto, véanse algunas de las pruebas uniformes mencionadas en la bibliografía al final de este capítulo.

5. El último paso es agregar (l) y la cifra derivada de

$$\left(\frac{f_2}{f_2 + f_1}\right)(i)$$

Por lo tanto, en el problema, $X = 140 + 5.3 = 145.3$ libras.

APLICACION Y RASGOS CARACTERISTICOS DE LA MEDIA, LA MEDIANA Y EL MODO

¿Cuáles son las cualidades y los efectos de los diferentes promedios y cuándo debe usarse cada uno de ellos? Un promedio satisfactorio debe ser 1) preciso y rígidamente definido, 2) simple y concreto, 3) fácilmente calculable, 4) rápidamente comprensible, y 5) susceptible de un tratamiento algebraico. Juzgando de acuerdo con este criterio, la media es sin duda superior tanto a la mediana como al modo, pero en último análisis, la naturaleza de los datos y el propósito inmediato son los que realmente deben determinar el promedio particular que se elija. Ningún promedio debe considerarse como el *mejor* bajo todas las circunstancias. Además, no debe concederse demasiada importancia a un solo valor. Una descripción adecuada de una distribución de frecuencia ordinariamente requiere el cómputo de dos o más promedios así como otras clases de medidas estadísticas.

Sinteticemos brevemente algunas de las más importantes características de la media, la mediana y el modo: [9].

Media:

1. El promedio aritmético es el más conocido y el más frecuentemente usado. A menudo es conveniente emplearlo simplemente porque se comprende fácilmente.

2. El promedio aritmético se ve afectado por el valor de cada caso de la serie. Por lo tanto, a veces se les concede un valor indebido a items extremos y equivocados. Por ejemplo, el salario medio de un grupo puede

9. G. Udny Yule, *An Introduction to the Theory of Statistics* (Introducción a la Teoría de las Estadísticas), pp. 106-132; Frederick Cecil Mills, *Statistical Methods* (Los Métodos Estadísticos), pp. 143-146.

resultar equivocado si hay items extremos y no característicos en cada extremo de la escala.

3. Desde el punto de vista algebraico el promedio aritmético es mejor que la mediana o el modo.

Mediana:

1. La mediana no está influenciada por el tamaño de los items extremos. Se basa en los valores que radican inmediatamente en ambos extremos. Este tipo de promedio puede usarse muy efectivamente cuando los items extremos pueden tener una influencia indebidamente desproporcionada sobre el conjunto.

2. La mediana se calcula fácilmente.

3. Generalmente este tipo de promedio es menos verificable que la media y no se adapta tan bien a la manipulación algebraica.

Modo:

1. El modo, como la mediana, es un promedio de posición que no resulta afectado por los items extremos. Por lo tanto es muy útil en los casos en que es deseable eliminar los efectos de las variaciones extremas.

2. Frecuentemente es muy difícil localizar el modo.

3. El modo tiene poca significación, a menos que haya una tendencia central distinta y a menos que se aplique a una muestra relativamente grande.

4. El modo no es susceptible de manipulación algebraica.

VARIABILIDAD

Otro concepto importante en la estadística es el de la variabilidad [10]. La media, la mediana y el modo dan solamente una característica esencial

10 Este concepto ha sido también designado como "dispersión", "esparcimiento", "difusión" o "derivación".

de la distribución de frecuencia —su tamaño típico o tendencia central. Decir, por ejemplo, que el valor medio o mediano de las residencias en una ciudad es de $3,000 no da una idea muy exacta del valor de la propiedad. Además, es esencial conocer cómo varían los valores arriba y abajo de la media o de la mediana. Algunas casas pueden valer solamente $500, mientras que otras pueden tener un valor de $50,000 o más. Por otra parte, es posible que varias distribuciones tengan el mismo promedio y sean marcadamente diferentes en cuanto a variabilidad. En algunas distribuciones los casos pueden agruparse estrechamente alrededor del promedio, mientras que en otras pueden estar extensamente diseminadas. Por lo tanto, es muy importante determinar el alcance de los valores individuales de ambos lados de la tendencia central.

En la discusión presente se tendrán en cuenta las siguientes medidas de variabilidad: 1) la oscilación, 2) la desviación media, 3) la desviación standard, y 4) la desviación semi-intercuartilar.

1. *Oscilación.* La oscilación de un conjunto no agrupado de medidas es simplemente la diferencia entre el tamaño del item más grande y del más pequeño. Por ejemplo, la oscilación de las series 10, 11, 13, 16, 17, y 19 es 9 porque 19-10 es igual a 9. La oscilación es una medida muy inestable de variabilidad, puesto que depende exclusivamente de los dos items extremos de las series y virtualmente no indica nada acerca de la forma general o del perfil de las mismas. Además, no puede usarse muy eficazmente en una distribución de frecuencia puesto que la oscilación exacta ordinariamente no puede determinarse. Si se desea únicamente la distribución total de los items de las series o si los datos son demasiado escasos para dificultar el cómputo de una medida de variabilidad más exacta, entonces puede usarse la oscilación.

2. *Desviación media.* La desviación media o desviación de promedio, como se le llama frecuentemente, es el promedio de la suma de las desviaciones (sin considerar el signo) de alguna medida de tendencia central. Generalmente la media se toma como standard aunque algunas veces se usan la mediana y el modo. Debe tenerse siempre cuidado de especificar el promedio particular que se haya elegido para computar la desviación media. Para ilustrar el cálculo de la desviación media, consideremos las siguientes series simples de valores:

m, series de valores: 2, 3, 5, 8, 10, 11, 12, 13, 14, 16, 18

d, la desviación del promedio: 8, 7, 5, 2, 0, 1, 2, 3, 4, 6, 8

$$Xm = \frac{110}{11} = 10$$

$$Xd = \frac{46}{11} = 4.18$$

Se observará que el punto medio de las series se derivó primero sumando los números y dividiéndolos entre 11. Las desviaciones de cada valor del coeficiente de las series (Xm = 10) se indican en la segunda línea. La desviación general es la suma de estas desviaciones (signos no considerados) del promedio dividido entre el número de casos de las series o $\frac{46}{11} = 4.18$.

Cuando la desviación del promedio se calcula en una distribución de frecuencia, puede emplearse el procedimiento siguiente:

1. Computar las medidas de tendencia central que deberán usarse como standards.

2. Indicar, en una columna separada, el punto medio de cada intervalo de clase en la distribución.

3. En la columna siguiente tabular las desviaciones (d) de acuerdo con el standard que se haya elegido en los puntos centrales de los diversos intervalos de clase.

4. En la última columna, tabular los productos de las frecuencias de cada intervalo de clase por la desviación (fd).

5. Sumar las cifras en la columna fd, sin tomar en cuenta los signos (Σfd).

6. Dividir fd por el número de casos en la distribución

$$\left(\frac{\Sigma fd}{N}\right). \text{A. D.} = \frac{\Sigma fd}{N}.$$

En la práctica se encuentra que la desviación media tiene poco valor y que raras veces se usa en la investigación social. La causa principal de que se le haya concedido tanto espacio, es ayudar al estudiante a obtener una comprensión más clara del concepto de variabilidad y especialmente de la desviación standard, que se usa más extensamente y es muy superior en todos aspectos a la desviación media.

Desviación standard. La desviación standard, lo mismo que la media, representa un promedio de las desviaciones de los items. Sin embargo, es diferente de la desviación media en cuanto a que las desviaciones se elevan al cuadrado antes de ser sumadas. La suma de las desviaciones al cuadrado se divide por el total de observaciones (casos) en la distribución y entonces se extrae la raíz cuadrada de este cociente. La desviación standard siempre se computa a través del promedio aritmético, mientras que la desviación media puede computarse por la media, la mediana y a veces, el modo. Algebraicamente estos pasos pueden representarse con la siguiente fórmula:

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}$$

La desviación standard constituye una medida más refinada, más válida y más importante desde el punto de vista estadístico que la desviación media.

TABLA IV

Cálculo de la desviación standard (σ). Datos que representan el peso de 265 estudiantes de primer año de la Universidad de Washington

Intervalo de Clase (Peso)	f	d	fd	fd^2
90- 99	1	—5	—5	25
100-109	1	—4	—4	16
110-119	9	—3	—27	81
120-129	30	—2	—60	120
130-139	42	—1	—42	42
140-149	66	0	0	0
150-159	47	1	47	47
160-169	39	2	78	156
170-179	15	3	45	135
180-189	11	4	44	176
190-199	1	5	5	25
200-209	3	6	18	108
$N = 265$	$\Sigma fd = 99$		$\Sigma fd^2 = 931$	

$$\sigma = \left(\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N} \right)^2} \right) (i)$$

$$\sigma = \left(\sqrt{\frac{931}{265} - \left(\frac{99}{265} \right)^2} \right) (10)$$

$$\sigma = \left(\sqrt{3.5132 - .1396} \right) (10)$$

$$\sigma = \left(\sqrt{3.3736} \right) (10)$$

$$\sigma = (1.8367) \quad (10)$$

$$\sigma = 18.37 \text{ ó } 18.4$$

La desviación standard queda simbolizada por la abreviación S. D., pero más frecuentemente por (σ) la letra griega minúscula que significa *s*. Al computar la desviación standard de una distribución de frecuencia, el procedimiento siguiente es el que se usa más frecuentemente. Se conoce como el método más "abreviado".

1. Se observará en el problema ilustrativo de la Tabla IV que la primera columna contiene los intervalos de clase y la segunda, las frecuencias. El número total de casos (N) es 265.

2. Se selecciona como origen arbitrario el intervalo en el cual es más fácil que se presente la media y se marcan las desviaciones en tér-

minos de intervalos superiores (más) e inferiores (menos). El intervalo 140 a 149 se eligió como intervalo cero.

3. Multiplíquese la frecuencia de cada clase por su correspondiente desviación (fd) y colóquense los productos en la cuarta columna. Encuéntrese Σfd. El procedimiento, hasta ahora, es idéntico al que se sigue para encontrar la media.

4. En la quinta columna, tabúlense los productos de (fd) (d), esto es, los valores de fd^2.

5. Obténgase Σfd^2. Se observará que nunca hay cantidades mínimas en la columna fd^2. En el problema Σfd^2 es igual a 931.

6. La corrección (c) que es $\frac{\Sigma fd}{N}$ se eleva al cuadrado. En el problema.

$$\frac{\Sigma fd}{N} = \frac{99}{265} = .374 \text{ y } (.374)^2 = .1399$$

Se observará que c queda expresada en términos de intervalos de clase y no en la unidad original de la escala.

7. El próximo paso es hacer las sustituciones adecuadas en la fórmula completa de la desviación standard.

$$\sigma = \sqrt{\frac{\Sigma fd_2}{N} - c^2} \quad (i)$$

Se verá que c^2 es siempre una cantidad mínima. Después de que se han hecho los cálculos bajo el signo radical, se extrae la raíz cuadrada y los resultados se multiplican por el tamaño del intervalo de clase (i). En la ilustración de la Tabla IV la desviación standard es 18.4. Con objeto de verificar la exactitud de los cálculos se puede elegir otro punto intermedio diferente y recomputar el problema de acuerdo con él.

Desviación semi-intercuartilar. La desviación semi-intercuartilar o desviación cuartilar es la mitad de la diferencia entre el tercer cuartil (Q_3) y el primero (Q_1). La desviación semi-intercuartilar comúnmente se simboliza por Q. La fórmula de la desviación semi-intercuartilar es:

$$Q = \frac{Q_3 - Q_1}{2}$$

La desviación semi-intercuartilar para la distribución de la Tabla I es, por lo tanto:

$$Q = \frac{160.71 - 136.01}{2}$$

$$= \frac{24.70}{2}$$

$$= 12.35$$

Al contrario de lo que sucede con la desviación media y con la standard, la desviación semi-intercuartilar no se mide a partir de un promedio central. Debe hacerse notar que los cuartiles dividen la distribución en cuatro partes iguales. Q_1 es un punto en la escala de variables, abajo del cual está el 25 por ciento y encima el 75 por ciento de los casos y Q_3 se localiza de modo que el 25 por ciento de los casos quedan arriba y el 75 abajo. Si la distribución de frecuencia es perfectamente simétrica Q_1 y Q_3 son equidistantes de χ. En una distribución asimétrica no se mantiene esta relación. La desviación semi-intercuartilar no es una medida de variabilidad tan satisfactoria como la desviación standard, puesto que sólo se toma en consideración una parte de la distribución, descuidándose el resto y, por ciertos tipos de distribuciones, los cuartiles resultan indeterminados y faltos de significación.

RELACION ENTRE LA DESVIACION MEDIA, LA STANDARD Y LA DESVIACION SEMI-INTERCUARTILAR

Debe hacerse notar que la media, la mediana y el modo representan puntos o valores en la escala X y que la desviación media, la standard y la desviación semi-intercuartilar representan distancias en la misma escala. En una distribución de frecuencia simétrica: 1) ¿qué proporción del área o de los casos queda incluída dentro de los límites de las respectivas medidas de variabilidad? y 2) ¿cuál es la relación entre la desviación media, la standard y la desviación-cuartilar? En una distribución perfectamente

simétrica el A. D. (X o $\chi \pm$ A. D.), define los límites del centro 57.5 por ciento de los casos el σ ($X \pm \sigma$) marca la distancia del centro 68.26% de los casos y Q ($\chi \pm Q$) indica el centro 50 por ciento de los items.

Se observará en la Fig. 17, que una distribución perfectamente simétrica, o en una que se aparte de este tipo muy poco, la desviación media corresponde aproximadamente a cuatro quintas partes de la desviación standard y la desviación semi-intercuartilar comprende a cerca de dos tercios de la desviación standard.

La tabla siguiente presenta más exactamente la relación entre estas tres medidas de variabilidad.

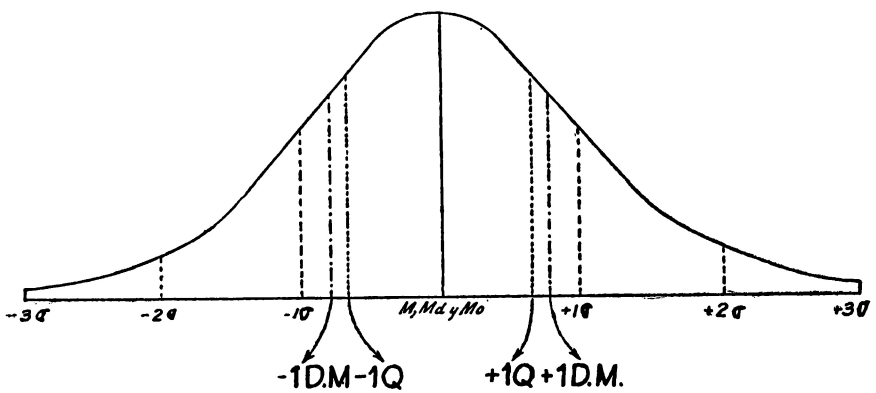

Fig. 17. *Curvas simétricas mostrando la relación de las medidas de variabilidad.*

TABLA V

Relación entre la desviación media standard y la desviación semi-intercuartilar, tal como se encuentra en las distribuciones de frecuencia simétricas

MEDIDAS		RELACION
σ	=	1.2533 A. D.
σ	=	1.4825 Q
A. D.	=	.7979 σ
A. D.	=	1.1843 Q
Q	=	.6745 σ
Q	=	.8453 A. D.

Como ilustración para interpretar una medida de variabilidad, tenemos la desviación standard de la distribución que aparece en la Tabla IV.

Se encontró que la desviación standard era 18.4. En una distribución perfectamente simétrica la desviación standard, cuando es medida abajo y arriba del punto medio incluye en esta distancia aproximadamente el 68.26 por ciento de los casos de la distribución. Este hecho también suele encontrarse en distribuciones que son moderadamente asimétricas.

En este problema, una desviación standard inferior a la media (148.7) sería 130.3 y una superior sería 167.1 en la escala; por lo tanto, se podría decir que aproximadamente dos terceras partes (68.26 por ciento si la distribución fuera perfectamente simétrica) de los estudiantes que figuran en esta Tabla, pesan entre 130.3 y 167.1 libras.

COEFICIENTE DE VARIACION

La desviación media, la standard, y la desviación semi-intercuartilar, representan medidas de variabilidad absoluta. También es necesario, frecuentemente, medir la variabilidad relativa de dos o más distribuciones de frecuencia.

En la Tabla VI figuran las respectivas edades medias, las desviaciones standard y los coeficientes de variación de cuatro grupos de mujeres que tuvieron uno o más hijos en Minneapolis durante el período de cinco años comprendido entre 1931 y 1935.

TABLA VI

Medias, desviaciones standard y coeficientes de variación de las distribuciones de edades de cuatro grupos de madres que tuvieron un hijo o más, en la ciudad de Minneapolis: 1931 a 1935 *

Clasificación	X	σ	c. v.
Casadas residentes	28.2	6.0	21.3
Casadas no residentes . . .	29.5	6.0	20.3
Residentes sin casar	23.4	5.8	24.8
No residentes sin casar. . .	21.7	3.7	17.1

* Datos tomados de Calvin F. Schmid, *Mortality Trends in the State of Minnesota* (Curso de la Mortalidad en el Estado de Minnesota), pp. 273-275.

¿Cuál grupo presenta el grado más alto, relativamente, de variabilidad y cuál el menor? Con un simple examen de las desviaciones standard es imposible decirlo. Sin embargo, cuando las desviaciones standard de varias distribuciones están relacionadas con sus correspondientes medias, es posible determinar la cantidad relativa de variabilidad de un número determinado de distribuciones de frecuencia. Karl Pearson elaboró una medida simple de variabilidad relativa que se conoce generalmente como el coeficiente de variación.

$$\text{C. V. o V} = \frac{\sigma}{X}\left(\frac{100}{1}\right)$$

En la Tabla VI se verá que las madres solteras no residentes presentan la variabilidad relativamente menor (17.1) en edad y las residentes solteras, la mayor (24.8). El coeficiente de variación para el problema de la Tabla IV es:

$$\text{C. V.} = \left(\frac{18.367}{148.736}\right)\left(\frac{100}{1}\right) = 12.3 \text{ por ciento}$$

ASIMETRIA

En la práctica, las distribuciones de frecuencia raras veces son simétricas; muestran varios grados de asimetría. En una distribución perfectamente simétrica la media, la mediana y el modo coinciden, lo cual no sucede en una distribución asimétrica u oblicua.

La figura 18 presenta ilustraciones de curvas que son notablemente oblicuas. La superior tiene una oblicuidad negativa y la inferior positiva. Se observará que en la curva que tiene asimetría negativa la media es menor que el modo, mientras que en la curva con asimetría positiva la media es mayor que el modo. El modo no resulta afectado ni por el tamaño ni por el número de los items extremos en una distribución de frecuencia, pero la media sí.

Los coeficientes de asimetría indican tanto la dirección como el grado de ésta, bien sea absoluta o relativamente. Una media comunmente usada, pero poco exacta es la diferencia que hay entre la media menos el modo ($X - X$). El signo indica la dirección de la asimetría y la dife-

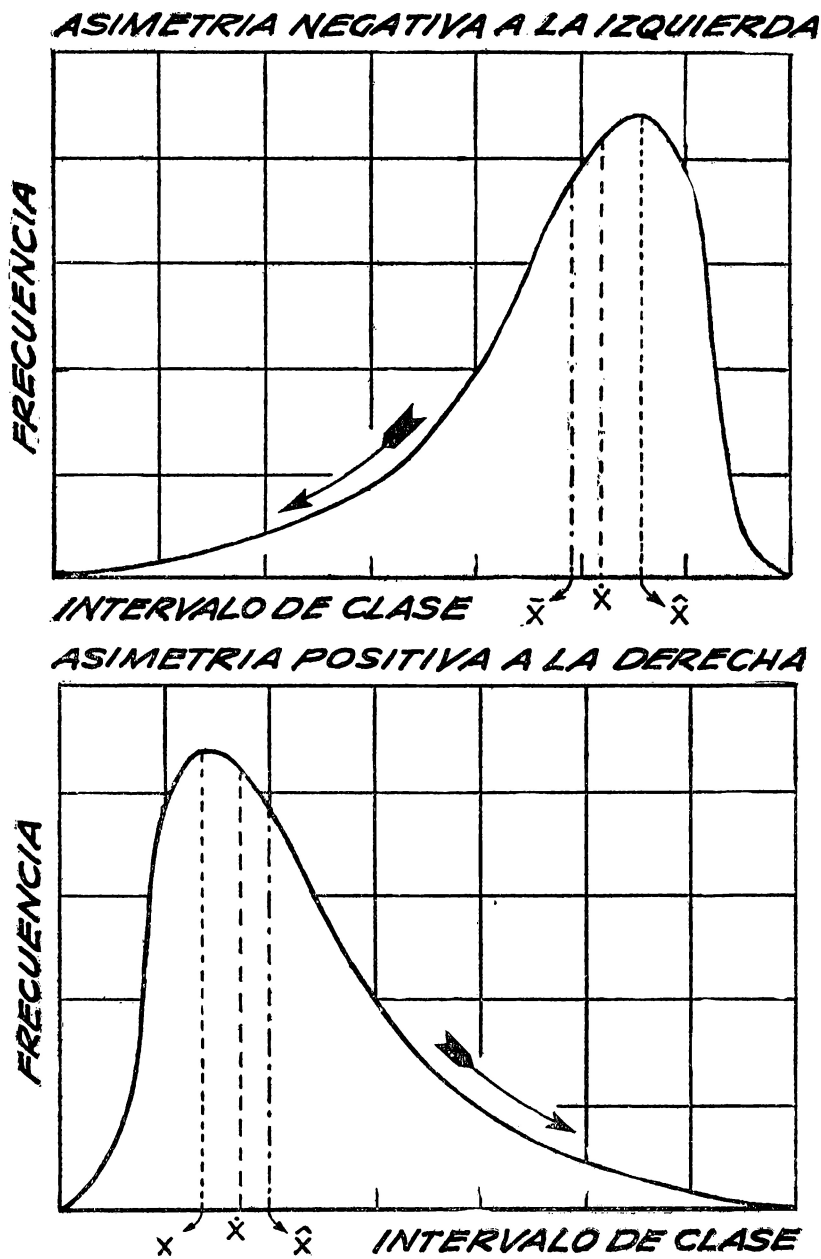

Fig. 18. *Curvas de frecuencia que muestran la inclinación positiva y negativa.*

rencia. La distribución por peso de 265 estudiantes demuestra una simetría ligera, pero positiva, pues

$$X = 148.74 \text{ y } X = 145.28 \text{ y } 148.74 - 145.28 = + 3.46.$$

Una medida de asimetría más satisfactoria en la cual se le da el valor debido al grado de variabilidad es:

$$Sk = \frac{X - x}{\sigma}$$

Puesto que el verdadero modo frecuentemente es difícil de determinar, puede aplicarse otra fórmula para encontrarlo, la cual ha sido desarrollada por Karl Pearson, especialmente si las series son moderadamente asimétricas. De acuerdo con esta fórmula:

$$Sk = \frac{3\ (X - \chi)}{\sigma}$$

Las distribuciones de frecuencia que han sido consideradas hasta ahora son del tipo común en el cual las frecuencias son relativamente bajas al final y altas hacia el centro. Además de este tipo hay curvas de frecuencia que muestran una tendencia constante a aumentar o a disminuir. Estas curvas poseen las características generales de una J mayúscula y se conocen como curvas en forma de J. Hay también otro tipo de distribución irregular en el cual las frecuencias son relativamente altas en los dos extremos y bajas hacia el centro. Este tipo se conoce con el nombre de tipo en forma de U. Estos dos tipos no se presentan tan frecuentemente como el tipo regular en forma de campana.

(Continuará)