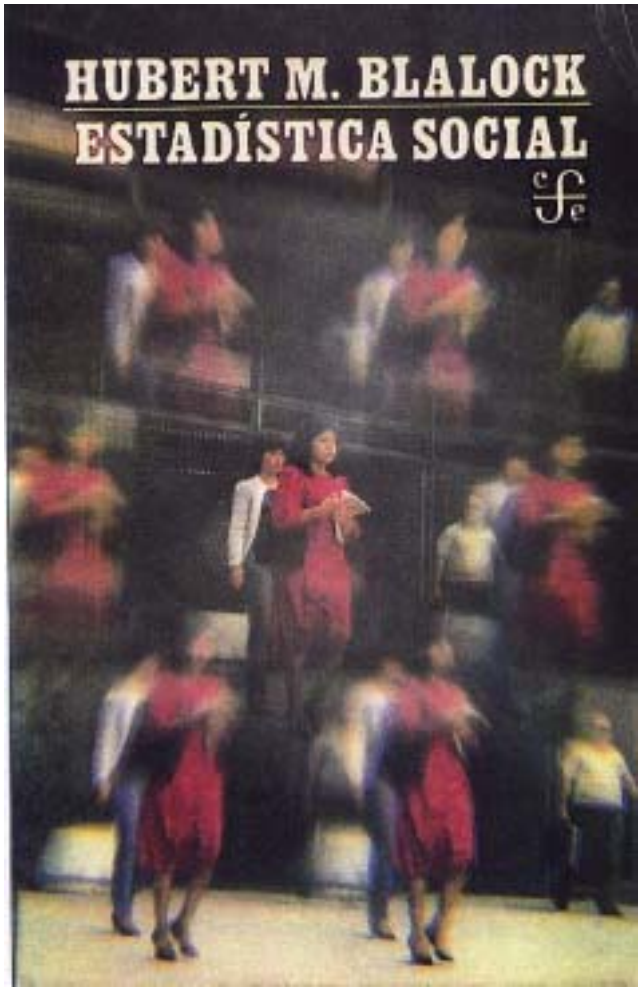
ESTADÍSTICA SOCIAL

HUBERT M. BLALOCK



FONDO DE CULTURA ECONÓMICA

México, 1986

Este material se utiliza con fines
exclusivamente didácticos

ÍNDICE

Prefacio 9

Primera Parte
INTRODUCCIÓN

1. Introducción: objetivos y límites de la estadística 15
1.1. Funciones de la estadística..... 16
1.2. El lugar de la estadística en el proceso de la investigación 19
1.3. Advertencia 20
Bibliografía 21

II. Teoría, medición y matemáticas 22
II. l. Teoría e hipótesis: definiciones operativas 22
II.2. El nivel de medición: escalas nominales, ordinales y de intervalo 26
II.3. Medición y estadística 32
II.4. Organización del libro 37
Bibliografía..... 40

Segunda Parte
ESTADÍSTICA DESCRIPTIVA UNIVARIADA

III. Escalas nominales: proporciones, porcentajes y razones 43
III. 1. Proporciones 43
III. 2. Porcentajes..... 45
III. 3. Razones..... 49
Bibliografía 52

IV. Escalas de intervalo: distribuciones de frecuencia y representación gráfica 53
IV.1. Distribuciones de frecuencia: agrupamiento de los datos 53
IV.2. Distribuciones de frecuencia cumulativa 60
IV.3. Presentación gráfica: histogramas, polígonos de frecuencia y ojivas 61
Bibliografía 66

V. Escalas de intervalo: medidas de tendencia central 67
V.1. La media aritmética 67
V.2. La mediana 71
V.3. Cálculo de la media y la mediana de datos agrupados 73
V 4. Comparación de la media y la mediana 81
V.5. Otras medidas de tendencia central 85
V.6. Deciles, cuartiles y percentiles 86
Bibliografía 88

VI. Escalas de intervalo: medidas de dispersión 90
VI. l. El recorrido 90
VI.2. La desviación cuartil 92
VI.3. La desviación media 92
VI.4. La desviación estándar 93
VI.5. El coeficiente de variabilidad 101
VI.6. Otras medidas resumidas 102
Bibliografía 103

VII. La distribución normal 104
VII. l. Distribuciones de frecuencias finitas versus infinitas 104
VII.2. Forma general de la curva normal..... 107
VII.3. Áreas bajo la curva normal 109
VII.4. Ilustraciones suplementarias del empleo de la tabla normal 113

Bibliografía 116

Tercera Parte
ESTADÍSTICA INDUCTIVA

VIII. Introducción a la estadística inductiva 119
VIII. l. Estadística y parámetros 119
VIII.2. Pasos en la verificación de una hipótesis 120
VIII.3. La falacia de afirmar el consecuente 123
VIII.4. La forma de las hipótesis estadísticas 124
Bibliografía 127

IX. Probabilidad 128
IX.1. Probabilidad a priori 129
IX.2. Propiedades matemáticas de las probabilidades 133
IX.3. Permutas 145
IX.4. Valores esperados 151
IX.5. Independencia y muestreo aleatorio 153
Bibliografía 159

X. Pruebas de hipótesis: la distribución binomial 160
X1. La distribución de muestreo binomial 160
X.2. Pasos en las pruebas estadísticas 164
X.3. Aplicaciones de la binomial 177
X.4. Extensiones del binomio 181
X.5. Sumario 183
Bibliografía 186

XI. Pruebas de muestras simples que implican medias y proporciones 187
XI. l. Distribución en muestreo de las medias 187
XI.2. Prueba para la media de la población, conociendo. 194
XI.3. La distribución t de Student 199
XI.4. Pruebas que comportan proporciones 204
Bibliografía 210

XII. Estimación de punto e intervalo 211
XII. l. Estimación del punto 212
XII.2. Estimación del intervalo 215
XII.3. Intervalos de confianza para otros tipos de problemas 221
XII.4. Determinación del tamaño de la muestra 224
Bibliografía 227

Cuarta Parte
ESTADISTICAS BIVARIADAS Y MULTIVARIADAS

XIII. Pruebas de dos muestras: diferencia de las medias y las proporciones 231
XIII. l. Prueba de la diferencia de las medias 231
XIII.2. Diferencia de proporciones 240
XIII.3. Intervalos de confianza 245
XIII.4. Muestras dependientes: pares asociados 246
XIII.5. Comentarios a propósito de los esquemas experimentales y pruebas de significación 248
Bibliografía 255

XIV. Escalas ordinales: pruebas no paramétricas de dos muestras. 256
XIV.1. Fuerza y eficiencia de la fuerza 257
XIV.2. La prueba de las secuencias (runs) de Wald-Wolfowitz 263
XIV.3. La prueba de Mann-Whitney o de Wilcoxon 269

XIV.4. La prueba de Kolmogorov-Smirnov 277
XIV.5. La prueba de Wilcoxon de pares asociados y órdenes provistos de signo 280
XIV.6. Resumen 284
Bibliografía 288

XV. Escalas nominales: problemas de contingencia 289
XV.1. La prueba de la X-cuadrada 289
XV.2. La prueba exacta de Fisher 301
XV.3. Medidas de la fuerza de la relación 306
XV.4. Control de otras variables 319
Bibliografía 330

XVI. Análisis de la variancia 332
XVI. l. Análisis simple de la variancia 332
XVI.2. Comparación de medias específicas 343
XVI.3. Análisis bimodal de la variancia 349
XVIA. Alternativas no paramétricas del análisis de variancia 365
XVI.5. Medidas de asociación: correlación intraclase 370
Bibliografía 376

XVII. Correlación y regresión 377
XVII. l. Regresión lineal y mínimos cuadrados 377
XVII.2. Correlación 393
Bibliografía 413

XVIII. Correlación y regresión [conclusión] 414
XVIII. 1. Prueba de significación e intervalos de confianza 414
XVIII. 2. Correlación no lineal y regresión 425
XVIII. 3. Efectos de los errores de medición 431
XVIII. 4. Escalas ordinales: correlación de rangos 433
Bibliografía 445

XIX. Correlación múltiple y parcial 447
XIX.1. Regresión múltiple y mínimos cuadrados- 447
XIX.2. Correlación parcial 451
XIX.3. Correlación parcial e interpretaciones causales 461
XIX.4. Mínimos cuadrados múltiples y los coeficientes beta 469
XIX.5. Correlación múltiple 473
XIX.6. Regresión múltiple y no linealidad 479
XIX.7. Pruebas de significación e intervalos de confianza 484
Bibliografía 489

XX. Análisis de covariancia y variables simuladas 491
XX.1. Relación de dos escalas de intervalo, control de la escala nominal 492
XX.2. Relación de una escala de intervalo y una escala nominal, control de la escala de intervalo 510
XX.3. Extensiones del análisis de covariancia 516
XX.4. Análisis de la variable simulada 517
XX.5. Observaciones finales 521
Bibliografía 526

Quinta Parte
MUESTREO

XXI. Muestreo 531
XXI.1. Muestreo aleatorio sencillo 532
XXI.2. Muestreo sistemático 537

XXI.3. Muestreo estratificado 539
XXIA. Muestreo por conglomerados 546
XXI.S. Muestreo sin probabilidad 552
XXI.6. Errores no de muestreo y tamaño de la muestra 553
Bibliografía 554

APÉNDICES

I. Resumen de operaciones algebraicas 559
Cuadros 565
Índice de figuras 599
Índice de cuadros 603

VI. ESCALAS DE INTERVALO: MEDIDAS DE DISPERSIÓN

En la investigación sociológica la atención se concentra en muchos casos en medidas de tendencia central. Por ejemplo, podemos querer comparar varios tipos de religión en relación con la asistencia media a la iglesia o el nivel medio de ingreso. Podemos también desear obtener, sin embargo, medidas de homogeneidad. Tal vez hayamos partido de la hipótesis que una de las religiones extraerá sus adeptos en mayor grado que las otras de una misma capa social. Sin embargo, aun si estamos interesados ante todo en comparar medidas de tendencia central, necesitamos, con todo, saber algo acerca de la dispersión en cada grupo. Nos damos cuenta intuitivamente de que, si cada religión fuera extremadamente heterogénea en cuanto al ingreso y a la asistencia a la iglesia, una diferencia determinada entre sus medias (digamos de $2000) no sería tan importante o indicativa como sería el caso si cada grupo fuera perfectamente homogéneo.

Cuando lleguemos a la estadística inductiva, estaremos en condiciones de justificar dicha intuición y de apreciar por qué las medidas de dispersión son tan importantes. En el presente capítulo vamos a concentrarnos en el mecanismo, en tanto que en el siguiente daremos una interpretación de la medida de dispersión más importante: la desviación estándar.

VI. 1. El recorrido

De las distintas medidas de dispersión que vamos a examinar en este capítulo, el recorrido es con mucho el más simple. El recorrido se define como o la diferencia entre la marca más alta y la más baja. Así, pues, en relación con los datos proporcionados en el capítulo anterior (72,81,86,69 y57), el recorrido sería la diferencia entre 86 y 57, o sea 29. Por lo regular solemos indicar el recorrido ya sea por medio de la diferencia real (29), o dando las dos marcas extremas, *v.gr*. 57 y 86. Si los datos se han agrupado, tomamos como recorrido la diferencia entre los *puntos medios* de las categorías extremas. Así, pues, si el punto medio del intervalo inferior es 2 450 y el del intervalo superior 7 450, el recorrido será de 5 000.

La simplicidad extrema del recorrido corno medida de dispersión presenta a la vez ventajas e inconvenientes. En efecto, el recorrido puede resultar muy útil si se trata de obtener unos cálculos muy rápidos que puedan proporcionar una indicación bruta de la dispersión, o si los cálculos ha de hacerlos alguna persona que no esté familiarizada con la estadística. Si los datos han de presentarse a una audiencia relativamente ingenua, el recorrido será tal vez la única medida de dispersión que aquélla esté en condiciones de interpretar fácilmente. Sin embargo, el nivel de preparación de los sociólogos está alcanzando rápidamente un punto tal, que podemos legítimamente suponer que entenderán también medidas algo más complicadas y satisfactorias. El inconveniente del recorrido es obvio: se basa exclusivamente en dos casos, qué son, además, los dos casos extremos. Y como quiera que los casos extremos suelen ser raros o poco comunes en la mayoría de los problemas empíricos, nos damos cuenta que por lo regular es una cuestión de azar que obtengamos uno o dos de ellos en nuestra muestra. Supóngase, por ejemplo, que en la localidad investigada hay un millonario. Si escogemos 10 personas al azar, es probable que aquél no esté incluido entre ellas. Pero, supóngase que sí está. En tal caso el recorrido de los ingresos será extraordinariamente amplio y muy engañador en cuanto medida de dispersión. Si nos servimos del recorrido como medida, nada sabemos acerca de la variación de las marcas entre los dos valores extremos, excepto que éstas se sitúan en algún lugar en el interior de dicho recorrido. Así, pues, como resulta del ejemplo anterior, el recorrido variará considerablemente de una muestra a otra. Por otra parte, el recorrido será por lo regular mayor en las muestras grandes que en las pequeñas, simplemente porque en los primeros tenemos más probabilidades de incluir a los casos individuales extremos. Ésta es la razón de que el recorrido no se emplee por lo regular en sociología, excepto al nivel de tipo más exploratorio.

Otra medida sumamente simple, la *razón de variación*, puede ser utilizada en el caso de los datos en grupo, lo que resulta especialmente adecuado en el caso de las escalas nominales. Consiste básicamente en una medida del grado en que se concentran los datos en la categoría modal, en lugar de que se, les encuentre distribuidos uniformemente a lo largo de todas las categorías. Se define así:

$$\text{V.R.} = 1 - f_{\text{modal}} /\text{N},$$

en donde f_{modal} se refiere al número de casos en la categoría modal, y N al número total de casos. Es evidente que esta medida resulta insensible a la distribución de casos en las categorías no modales, siendo por otra parte dependiente del proceso de categorización. Su ventaja radica en su sencillez extrema y en su atracción intuitiva, además del hecho de que en el caso de las escalas nominales no es posible hacer uso de una ordenación de categorías que permita habilitar medidas de un mayor refinamiento.

VI. 2. La desviación cuartil

Otra medida empleada algunas veces en los campos de la psicología y la enseñanza, pero que raramente aparece en la literatura sociológica, es la desviación cuartil o recorrido serni-intercuartil. La desviación cuartil Q es un tipo de recorrido, pero, en lugar de representar la diferencia entre los valores extremos, se define arbitrariamente como la mitad de la distancia entre el primero y el tercer cuartiles. O en forma simbólica:

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

en donde Q_1 y Q_3 representan respectivamente al primero y tercer cuartiles. Obsérvese que la derivación cuartil mide el recorrido ocupado por la mitad central de los casos. Como quiera que Q_1 y Q_3 variarán menos de una muestra a otra que los casos más extremos, la desviación cuartil representa una medida mucho más estable que el recorrido. Por otra parte, en cambio, no saca provecho del conjunto de la información. No estamos midiendo la variabilidad entre los casos centrales ni tomarnos en consideración lo que ocurre en los extremos de la distribución. De ahí, pues, que enderecemos nuestra atención a otras dos medidas que sí poseen esta propiedad deseable.

VI. 3. La desviación media

Si deseamos servirnos de todos los datos, el sentido común nos sugerirá que tomemos las desviaciones de cada dato con respecto a alguna medida de tendencia central y que calculemos luego alguna especie de promedio de dichas desviaciones, con objeto de controlar el número de casos comprendidas. Sería posible tomar como medida de tendencia central la mediana o el modo, pero por lo regular tomamos la media, ya que ésta es en la mayoría de los casos la medida particular más satisfactoria. Supóngase que sumáramos simplemente las desviaciones efectivas respecto de la media. Por desgracia, como sabemos, el resultado, sería siempre cero, ya que las diferencias positivas y negativas se compensan mutuamente. Esto sugiere que, para obtener una medida de dispersión alrededor de la media, hemos de deshacernos en una forma u otra de los signos negativos. Se nos ocurren inmediatamente dos métodos: *1*) ignorar los signos y tomar solo los valores absolutos de las diferencias, o 2) cuadrar las diferencias. Estos dos métodos conducen efectivamente a las dos medidas restantes de dispersión que hemos de examinar en este capítulo, a saber: la desviación media y la desviación estándar.

La desviación media se define como la media aritmética de las diferencias absolutas de cada marca con respecto a la media, o en símbolos:

$$\text{Desviación media} = \frac{\sum_{i=1}^{N} |X_i - \overline{X}|}{N} \qquad \text{(VI.2)}$$

La media de los números 72, 81, 86, 69 y 57 es 73.0. Si sustraemos 73.0 de cada uno de dichos números, ignorando los signos, y luego adicionamos los resultados y dividimos entre 5, obtenemos:

$$\frac{\sum_{i=1}^{N} |X_i - \overline{X}|}{N} = \frac{1 + 8 + 13 + 4 + 16}{5} = \frac{42}{5} = 8.4$$

Podemos por consiguiente decir que el promedio de los datos difiere de la media en 8.4.

Pese a que la desviación media presenta una interpretación intuitiva más directa que la desviación estándar, tiene, con todo, varios inconvenientes graves. Primero, los valores absolutos no se dejan manipular

algebraicamente con facilidad. Segundo y más importante, la desviación media no es de fácil interpretación teórica ni conduce a resultados matemáticos simples. Con fines puramente descriptivos, la desviación media puede ser adecuada, pese a que, según veremos, la desviación estándar se deja interpretar más fácilmente en términos de la curva normal. Cuando lleguemos a la estadística inductiva veremos que la desviación estándar se utiliza sobre todo a causa de su superioridad teórica. Ésta es la razón de que sólo raramente encontremos en la literatura sociológica referencias a la desviación media.

VI.4. La desviación estándar

Habiendo eliminado más o menos otras varias medidas de dispersión, podemos ahora dirigir nuestra atención a la más útil y frecuente de las medidas: la *desviación estándar*. Ésta se define como la raíz cuadrada de la media aritmética de las desviaciones cuadradas con respecto a la media, o en símbolos:

$$s = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N}} \qquad \text{(VI.3)}$$

en donde *s* se emplea para designar la desviación estándar.[1] O en palabras: tomamos la desviación de cada marca con respecto a la media, cuadramos cada diferencia, sumamos los resultados, dividimos entre el número de casos y extraemos la raíz cuadrada. Para conseguir una respuesta correcta, es indispensable que las operaciones se efectúen exactamente en el orden indicado. En nuestro ejemplo numérico la desviación estándar podría conseguirse como sigue:

X_i	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
72	− 1	1
81	8	64
86	13	169
69	− 4	16
57	− 16	256
$\overline{X} = 73.0$	0	506

$$s = \sqrt{506/5} = \sqrt{101.2} = 10.06$$

El significado intuitivo de la desviación estándar no nos aparecerá claramente hasta más adelante, cuando nos sirvamos de *s* para darnos las áreas bajo la curva normal. Por el momento la aceptamos simplemente como un número abstracto. Sin embargo, algunas propiedades de la desviación estándar son ya manifiestas desde ahora. Observamos, en efecto, que cuanto mayor es la dispersión alrededor de la media tanto mayor es la desviación estándar. Si todos los cinco valores hubieran sido cero, las desviaciones alrededor de la media habrían sido cero, y *s* también habría sido cero. Por otra parte, vemos que las desviaciones extremas con respecto a la media pesan más, con mucho, en cuanto a determinar el valor de la desviación estándar. En efecto, los valores 169 y 256 dominan las otras tres desviaciones cuadradas. Al cuadrar las desviaciones, pese a que después extraigamos la raíz cuadrada, estamos en realidad dando más

[1] Algunos textos definen *s* con *N - 1* en el denominador en vez de *N*. La razón de ello no resultará clara hasta en el capítulo XI

peso relativo a los valores extremos todavía de lo que era el caso al calcular la media. Esto sugiere que hemos de mitigar nuestro entusiasmo inicial a propósito de la desviación estándar en cuanto "la mejor" medida particular de dispersión. Ciertamente, si hay varios casos extremos, queremos que nuestra medida lo señale. Pero si la distribución presenta unos pocos casos muy extremos, la desviación normal puede conducir a resultados engañosos, en cuanto puede ser extraordinariamente grande. En tales casos nos serviríamos probablemente como medida de tendencia central de la mediana y, tal vez, de la desviación cuartil como medida de dispersión. Sin embargo, para la mayoría de los datos la desviación normal resultará adecuada.

Es razonable preguntar: "¿por qué molestarse en extraer la raíz cuadrada al calcular una medida de dispersión?" Una respuesta fácil, aunque poco satisfactoria, sería la de decir que así es como se define la desviación estándar. Podría justificarse la extracción de la raíz cuadrada señalando que, ya que hemos cuadrado cada desviación, lo que hacemos es compensar dicho paso anterior. Sin embargo, resulta más comprensible justificar la extracción de la raíz en términos de su carácter práctico. Como quiera que, en efecto, más adelante habremos de hacer un empleo considerable de la curva normal, la desviación estándar, tal como se la ha definido, resulta ser una medida muy útil. Para otros,, fines nos serviremos del cuadrado de la desviación normal o *variancia*, que se define como:

$$\text{Variancia} = s^2 = \frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N}$$

Los matemáticos han encontrado que el concepto de variancia poseía mayor valor teórico que la desviación estándar. A partir del capítulo XVI, haremos un uso creciente de? la variancia, pero de momento podemos limitar nuestra atención a la desviación estándar. Los dos conceptos son por lo demás tan fácilmente intercambiables, que podemos pasar sin dificultad del uno al otro. Que se defina la variancia como cuadrado de la desviación estándar o ésta como raíz cuadrada de la variancia, esto no reviste importancia alguna.

Cálculo de la desviación estándar de datos no agrupados. Si bien la desviación estándar puede calcularse siempre a partir de la fórmula básica que se acaba de dar, resulta a menudo más sencillo servirse de fórmulas de cálculo que no requieren la sustracción de la media de cada marca separada. En efecto, no sólo la media no será por lo regular un número entero, sino que usualmente se cometerán errores de redondeo al emplear la fórmula antes indicada. Con objeto de ver de qué modo podamos simplificar los cálculos, desarrollemos la expresión que está abajo del radical. Tenemos:

$$\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} = \frac{\sum_{i=1}^{N} (X_i^2 - 2X_i\overline{X} + \overline{X}^2)}{N}$$

$$= \frac{\sum_{i=1}^{N} X_i^2 - 2\overline{X} \sum_{i=1}^{N} X_i + N\overline{X}^2}{N}$$

Obsérvese que, como quiera que $\overline{X}$ es constante, pudimos tornarla frente al signo de sumación en el segundo término del numerador. En el tercer término, a su vez, nos hemos servido del hecho de que, para toda constante *k*, tenemos:

$$\sum_{i=1}^{N} k = Nk.$$

Pero, como quiera que $\overline{X} = \sum_{i=1}^{N} X_i/N$, el término central del numerador se reduce a $-2\overline{X}^2$, y podemos escribir:

$$\frac{\sum_{i=1}^{N} (X_i - \overline{X})^2}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - 2\overline{X}^2 + \overline{X}^2 = \frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2$$

Por lo tanto:

$$s = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \overline{X}^2} \qquad \text{(VI.4)}$$

Algunas otras fórmulas de cálculo alternativas son las siguientes:

$$s = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \left(\frac{\sum_{i=1}^{N} X_i}{N}\right)^2} \qquad \text{(VI.5)}$$

$$= \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}} \qquad \text{(VI.6)}^2$$

$$= \frac{1}{N}\sqrt{N \sum_{i=1}^{N} X_i^2 - \left(\sum_{i=1}^{N} X_i\right)^2} \qquad \text{(VI.7)}$$

[2]

Si bien cualquiera de las formas precedentes puede utilizarse como fórmula de cálculo, la ecuación (VI.7) es la que. comporta, con todo, menos errores de redondeo, por ello se la recomienda. Sirvámonos de una de dichas fórmulas de cálculo (ec. VI.7) en el problema anterior, en donde $N = 5$.

X_i	X_i^2
72	5184
81	6561
86	7396
69	4761
57	3249
365	27151

[2] La obtención. de las ecuaciones (VI.6) y (VI.7) a partir de la ecuación (VI-5) se deja como ejercicio.

En adición al número total de casos, las dos cantidades requeridas son

$$\sum_{i=1}^{N} X_i, \sum_{i=1}^{N} X_i^2$$

Ambas sumas pueden acumularse simultáneamente con las modernas calculadoras de oficina. Calculamos ahora s a partir de (VI.7):

$$S = 1/5 \sqrt{5(27151) - (365)^2} = 1/5 \sqrt{135755 - 133225} = 10.06$$

Nos hemos servido de este problema muy sencillo para ilustrar que la fórmula de cálculo da el mismo resultado numérico que la fórmula básica de la ecuación (VI.3). Como quiera que $\bar{X}$ resultó ser un entero, la fórmula de cálculo ha comportado en realidad más trabajo que la fórmula original. Pero normalmente, por supuesto, esto no será así.

* *Cálculo de la desviación estándar de datos agrupados.* Si los datos han sido agrupados, podemos simplificar nuestra labor considerablemente tratando cada caso como si se hallara en el punto medio de un intervalo y sirviéndose de una medida supuesta. Sin duda introducimos con ello alguna inexactitud, pero el ahorro de tiempo es sustancial. Siguiendo una convención corriente, supongamos que $x_i = X_i - \bar{X}$. En consecuencia, las *x* minúsculas representan desviaciones respecto de la media, y la fórmula básica de la desviación estándar se convierte en:

$$s = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}$$

Podemos modificar ahora la fórmula tomando en cuenta el hecho de que habrá un gran número de casos tratados todos como si tuvieran el mismo valor, esto es, uno de los puntos medios. Si multiplicamos el número de casos en cada clase por el punto medio propio y sumamos luego los productos, nos podemos ahorrar el trabajo de sumar todos los *N* casos. La fórmula de la desviación estándar se convierte así en:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{N}} \qquad \text{(VI.8)}$$

en donde f_i es el número de casos del intervalo i-ésimo y *k* el número de intervalos.[3]

Supongamos ahora que anticipamos una media y tomamos las desviaciones con respecto a ésta, en lugar de respecto de la media verdadera. Mostramos en el capítulo anterior que la suma de las desviaciones cuadradas de la media será menor –que cualquier otro valor– que la suma de las desviaciones cuadradas. En particular, la suma de las desviaciones cuadradas de la media anticipada será mayor que la cifra obtenida sirviéndonos de la media verdadera, a menos, por supuesto, que aquélla coincida con ésta. Puede pues, demostrarse que cuanto más cerca queda la media supuesta de la verdadera, tanto menor resulta la suma de las desviaciones cuadradas de la media supuesta. En otros términos: si nos servimos de una media supuesta, esperamos obtener una suma de cuadrados demasiado grande. Lo mismo que anteriormente, podemos servirnos de un factor de corrección, al que sustraemos luego del valor obtenido utilizando la media anticipada. La fórmula de la desviación estándar se convierte en tal caso en :

[3] Obsérvese que no se elevan al cuadrado las frecuencias fi que aparecen en el numerador de la expresión bajo el radical.

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{k} f_i d_i}{N}\right)^2} \qquad \text{(VI.9)}$$

en donde los d_i representan las diferencias entre cada marca y la media anticipada y son directamente análogos a los x_i de la ecuación (VI.8).

Antes de tomar un ejemplo numérico, examinemos la fórmula precedente con mayor atención. El segundo término debajo del radical representa el factor de corrección que ha de sustraerse de las desviaciones cuadradas de la media supuesta. Recordando la fórmula de la media expresada en términos de la media supuesta, o sea:

$$\overline{X} = \overline{X'} + \frac{\sum_{i=1}^{k} f_i d_i}{N}$$

vemos que

$$\frac{\sum_{i=1}^{k} f_i d_i}{N} = \overline{X} - \overline{X'}$$

y que, por lo tanto:

$$\left(\frac{\sum_{i=1}^{k} f_i d_i}{N}\right)^2 = (\overline{X} - \overline{X'})^2$$

De este modo, el factor de corrección resulta ser el cuadrado de la diferencia entre las medias verdadera y la supuesta. Vemos inmediatamente que, si hubiéramos anticipado la media exactamente, el factor de corrección habría sido cero. Por lo tanto, cuanto mayor sea la diferencia entre las medias verdadera y supuesta tanto mayor será el factor de corrección. Una suposición deficiente conducirá siempre al resultado correcto, pero comportará marcas numéricas mayores en ambos términos de la fórmula.

Ésta puede modificarse más todavía si preferimos pensar en términos de desviaciones graduales d_i. Lo mismo que en el capítulo v, ponemos en factor la amplitud del intervalo de cada d_i y multiplicamos el resultado final por i, una vez el proceso terminado. La fórmula se convierte así en:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i d_i^2}{N} - \left(\frac{\sum_{i=1}^{k} f_i d_i}{N}\right)^2} = \sqrt{\frac{i^2 \sum_{i=1}^{k} f_i d'^2_i}{N} - \left(\frac{i \sum_{i=1}^{k} f_i d'_i}{N}\right)^2}$$

Luego:

$$s = i\sqrt{\frac{\sum_{i=1}^{k} f_i d'^2_i}{N} - \left(\frac{\sum_{i=1}^{k} f_i d'_i}{N}\right)^2} \qquad \text{(VI.10)}$$

$$= i\sqrt{\frac{\sum_{i=1}^{k} f_i d'^2_i - \frac{\left(\sum_{i=1}^{k} f_i d'_i\right)^2}{N}}{N}} \qquad \text{(VI.11)}$$

$$= \frac{i}{N}\sqrt{N\sum_{i=1}^{k} f_i d'^2_i - \left(\sum_{i=1}^{k} f_i d'_i\right)^2} \qquad \text{(VI.12)}$$

Obsérvese que efectivamente no hemos hecho más que sacar la amplitud *i* del intervalo, de debajo del radical.

Al calcular la desviación estándar de datos agrupados, podemos ahora extender el procedimiento empleado para la media,

CUADRO VI.1. *Cálculo de la desviación estándar utilizando datos agrupados*

Límites verdaderos	*Puntos medios*	f_i	d'_i	$f_i d'_i$	$f_i d'^2_i$
$1 950–2 950	$2 450	17	−3	−51	153
2 950–3 950	3 450	26	−2	−52	104
3 950–4 950	4 450	38	−1	−38	38
4 950–5 950	5 450	51	0	0	0
5 950–6 950	6 450	36	1	36	36
6 950–7 950	7 450	21	2	42	84
Totales		189		−63	415

$$s = i\sqrt{\frac{\sum_{i=1}^{k} f_i d'^2_i}{N} - \left(\frac{\sum_{i=1}^{k} f_i d'_i}{N}\right)^2}$$

$$= 1\,000\sqrt{\frac{415}{189} - \left(\frac{-63}{189}\right)^2} = 1\,000\sqrt{2.196 - .111}$$

$$= 1\,444$$

añadiendo la columna $f_i d_i'^2$. Aunque en realidad podríamos obtener las desviaciones cuadradas $d_i'^2$ y multiplicar luego por f_i, resultará con todo mucho más simple multiplicar las dos últimas columnas empleadas en obtener la media (esto es: d_i' x $f_i d_i'$). En efecto, habiendo multiplicado d_i' por sí mismo, vemos

que todos los números negativos se hacen ahora positivos.[4] Calculemos ahora la desviación estándar de los datos agrupados utilizados en el capítulo precedente. Con fines de ilustración nos serviremos de la ecuación (VI.10), pese a que por lo regular la (VI.12) comportará menos errores de redondeo.

Obtuvimos en esta forma una media de $ 5 117 y una desviación estándar de $ 1444. Estos dos números pueden servir ahora para resumir los datos o para compararlos con datos de otra muestra. Según veremos más adelante que pueden emplearse también para verificar hipótesis o para apreciar medidas de población.

VI.5. El coeficiente de variabilidad

Es a veces conveniente. comparar varios grupos en relación con su homogeneidad relativa, en casos que dichos grupos tienen medias distintas. Podría, pues, resultar engañoso comparar las magnitudes absolutas de las desviaciones estándar. Cabría esperar que, con una media muy grande, podría encontrarse por lo menos una desviación estándar suficientemente grande. Así, pues, alguien podría interesarse en primer lugar por el tamaño de la desviación estándar *en relación con el de la media*. Esto sugiere que podemos obtener una medida de la variabilidad relativa dividiendo la desviación estándar entre la media. El resultado se ha llamado *coeficiente de variabilidad* y se designa con una V. Así, pues:

$$V = \frac{s}{\overline{X}}$$

Para ilustrar las ventajas del coeficiente de variabilidad con respecto a la desviación estándar, supóngase que un psicólogo social trata de demostrar que para todos los fines prácticos dos grupos son igualmente homogéneos en relación con la edad. En uno de los grupos la edad media es de 26, con una desviación estándar de 3. En el otro la edad media es de 38 años, con una desviación estándar de 5. Por lo tanto, los coeficientes de variabilidad son respectivamente 3/26 = .115 y 5/38= .132, o sea una diferencia mucho más pequeña que la que se da entre las dos desviaciones estándar. En vista del hecho de que por lo regular la edad exacta resulta menos importante, al determinar intereses, capacidades y posición social, a medida que aumenta la edad promedio de los miembros del grupo, la comparación de los dos coeficientes de variabilidad podría resultar muy bien, en este caso, mucho menos engañosa que si se emplearan las desviaciones estándar.

Si se desea, puede utilizarse también una variancia relativa. Por desgracia, estas medidas relativas de dispersión se hallan citadas con muy poca frecuencia en la literatura sociológica. Es mucho más frecuente, en efecto, encontrar las medias y las desviaciones estándar relacionadas en columnas adyacentes.

VI.6. Otras medidas resumidas

Sólo hemos examinado dos tipos de medidas resumidas: las de tendencia central y las de dispersión. Son posibles, además, otras medidas, aunque sólo se las utiliza raramente en la investigación sociológica. Sin duda, encontramos a menudo dada la distribución de frecuencia entera, pero esto no constituye una medida particular de resumen. Resulta a veces deseable indicar en una distribución el grado de asimetría. Una de las medidas de ésta saca provecho del hecho de que mayor es la asimetría tanto mayor resulta la diferencia entre la media y la mediana. Esta medida se halla dada por la fórmula:

$$\text{Asimetría} = \frac{3(\overline{X} - Md)}{s}$$

Si la distribución está desviada hacia la derecha (grandes marcas positivas), la media será mayor que la mediana, y el resultado será un número positivo. En tanto que la distribución desviada hacia la izquierda dará un resultado negativo.

[4] Obsérvese bien que la última columna del cuadro VI.1 no se obtiene elevando al cuadrado la columna $f_i d_i$, ya que al hacerlo traería consigo elevar también f_i al cuadrado.

Con muy poca frecuencia, también, hallamos en sociología referencias al carácter general de las cúspides de una distribución asimétrica. Utilízase el término de *picudez* en relación con dicha medida, que examinaremos brevemente una vez que hayamos visto la curva normal. Por lo regular, los textos de estadística escritos ante todo para los estudiantes de economía se ocupan más a fondo tanto de la desviación corno de la picudez. Tal vez cuando empecemos a alcanzar una mayor precisión en la descripción de las formas exactas de las distribuciones de las variables sociológicas hallaremos un mayor empleo para estas otras medidas descriptivas.

GLOSARIO

Coeficiente de variabilidad
Desviación media
Desviación cuartil
Recorrido
Desviación estándar
Variancia

Ejercicios

1. Calcúlense las desviaciones media y estándar de los datos indicados en el ejercicio 1, cap. V. Respuesta, 9.62; 11.59.

2. Calcúlense las desviaciones estándar y cuartil de los datos agrupados en el ejercicio 1, cap, IV. Hágase lo mismo con los del ejercicio 2, cap. IV.

3. Calcúlese la desviación estándar de los datos del ejercicio 4, cap. V. Contrólense los cálculos escogiendo una media anticipada y una fórmula de cálculo distintas. Respuesta, 10.83.

4. Indíquese en qué forma resultaría afectada la desviación normal por los cambios indicados en el ejercicio 5, cap. v.

BIBLIOGRAFÍA

Anderson, T. R., y M. Zelditeh: *A Basic Course in Statistics*, 2ª ed., Holt, Rinehart and Winston, Inc., Nueva York, 1968. pp. 76-84.

Downie, N. M., y R. W. Heath: *Basic Statistical Methods*, 2ª ed., Harper and Row, Publishers, Incorporated, Nueva York, 1952, cap. 9.

Hagood, M. J., y D. O. Price: *Statistics for Sociologists*, Henry Holt and Company, Inc., Nueva York, 1952, cap. 9.

McCollough, C., y L. van Atta: *Introduction to Descriptive Statistics and Correlation*, McGraw-Hill Book Company, Nueva York, 1965, cap. 3.

Mueller, J. H., K. Schuessler y H. L. Costner: Statistical Reasoning in Sociology, 2ª ed., Houghton Mifflin Company, Baston, 1970, cap. 6.

Weinberg, G. H., y J. A. Schumaker: *Statistics: An Intuitive Approach*, Wadsworth Publishing Company, Inc. Belmont, Cal, 1962, cap. 3.

Weiss, R. S.: *Statistics in Social Research*, John Wiley & Sons, Inc., Nueva York, 1968, cap. 7.