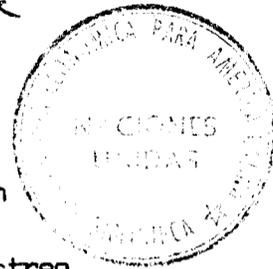


Biffista

CEPAL/Borrador/EST/149

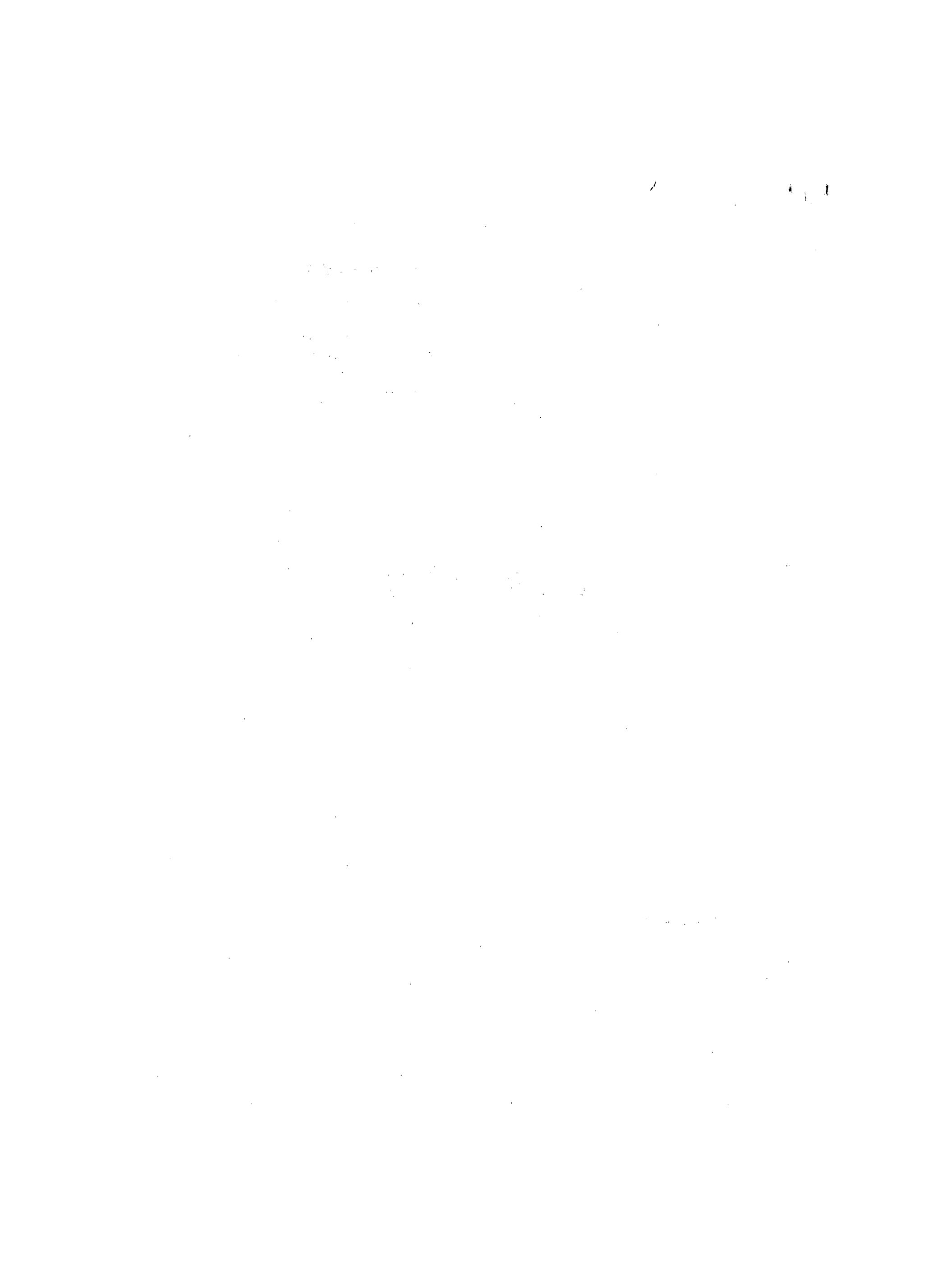


Borrador para discusión
Carlos Cavallini
Asesor Regional en Muestreo
para Estadísticas Demográficas
adscrito a la CEPAL
División de Estadística
Noviembre de 1976

DETERMINACION DEL LIMITE EN LA
CONSTRUCCION DE DOS ESTRATOS

76-11-2420-300

17 DEC 1976



I

1. La ciencia estadística es en sí misma un método científico.

Se encuentran estadísticos trabajando en todo campo donde exista evidencia cuantitativa.

En la demografía, en los problemas actuariales, en los campos de la probabilidad, en la astronomía, en la biología, en la agricultura, en la economía, en la química, y así a través de todo el alfabeto, los métodos estadísticos que hoy configuran la ciencia estadística, han intervenido en el desarrollo de las distintas materias. Si bien en un principio, estas líneas de desarrollo se mantuvieron en cierta forma independientes, actualmente, en este siglo, se ha visto que las mismas obedecen a conceptos comunes. Los trabajos de Francis Galton, Karl Pearson, Isaac Newton, Girolamo Cardano, Abraham de Moivre, Daniel Bernoulli, Siméon Poisson, Karl Gauss, Adolphe Quetelet, George Udny Yule, Ronald Fisher, Thomas Bayes, Jerzy Neyman, Maurice Kendall, A.A. Markoff, George Snedecor, P.C. Mahalanobis, entre muchos otros, sirvieron para plasmar el método estadístico. Sin embargo muchas veces la teoría estadística es confundida como parte de la matemática pura. Esto se debe a que muchos estadísticos de comienzos del presente siglo fueron matemáticos competentes y escribieron en la tradición matemática. El futuro de la estadística no se apoya solamente en el desarrollo de nuevas ideas matemáticas sino en el de hacer buen uso de las ya existentes. Es deplorable el hecho, por ejemplo, de que la mayoría de los trabajos estadísticos que aparecen en las revistas especializadas sean completamente inentendibles, salvo para personas que poseen una alta preparación en matemática. Esto no significa que la matemática no contribuya al progreso estadístico, todo lo contrario, sino que muchas personas que podrían aportar en el campo estadístico pueden llegar a pensar que este es un campo muy dificultoso.

2. Una rama importante de la estadística y que ha tenido un amplio desarrollo en los últimos cincuenta años es la que concierne al muestreo estadístico. Esto es, inferir de poco con respecto al todo. Este proceso de generalizar y de obtener conclusiones para toda una población examinando sólo una parte de ella es un razonamiento inductivo. Los resultados así obtenidos no necesariamente son exactos, ellos poseen márgenes de confiabilidad, o de error, confiabilidad que el investigador puede fijar antes de realizar el experimento. La teoría estadística provee las bases para la aplicación de este proceso inductivo.

En el diseño de una investigación por muestreo estadístico es esencial tener en cuenta, entre otros, los siguientes principios básicos: aleatorización, estratificación y replicación.

La aleatoridad tiene que ver con la teoría de la probabilidad. Sólo las muestras probabilísticas permiten medir la confiabilidad de los resultados obtenidos. La estratificación aumenta la eficiencia por unidad de costo y la replicación permite aumentar la sensibilidad de la investigación.

3. La estratificación de una población en subpoblaciones o estratos se adopta por diversas razones. En especial se utiliza para dar estimaciones por estrato, o para mejorar la representatividad de la muestra, o para aplicar distintos procedimientos muestrales en las diferentes subpoblaciones, o por conveniencias de tipo administrativo, pero singular beneficio reporta el hecho de que una adecuada estratificación permite reducir los costos de la investigación sin que por ello necesariamente disminuya la precisión de los resultados.

Como ejemplo se puede citar la investigación llevada a cabo en 1976 en la Oficina Técnica de Estudios de Mano de Obra (OTEMO) del Ministerio del Trabajo de la República del Perú y en la cual colaboró personal técnico de la División de Estadística de la CEPAL. En general, en esta investigación que abarcó a todo el Perú, las unidades primarias muestrales de selección, que eran conglomerados compuestos por unas 500 viviendas cada una, fueron estratificadas en unos 138 estratos. Estos estratos, relativamente homogéneos dentro de sí pero heterogéneos entre sí, fueron formados principalmente en base a la siguiente información: Regiones Geográficas, Zonas Geográficas, Concentración Urbana, Altitud sobre el Nivel del Mar, Comunicaciones, Servicios Educativos, Agua Potable y Servicios Eléctricos.

Luego se seleccionó una unidad por estrato, lo cual redujo significativamente el costo relativo (costo por unidad de eficiencia) de la investigación en comparación al costo relativo que hubiera representado si no se hubiese estratificado. Dado que se obtuvo una observación por estrato, para poder hallar la confiabilidad de las estimaciones se debieron aparear los estratos de a dos y utilizar la técnica estadística conocida como el método de los estratos conjugados o estratos contraídos (collapsed strata).

4. Los indicadores estadísticos, como ser los coeficientes de correlación intraclase que permiten medir la homogeneidad interna de las unidades muestrales, los costos de accesibilidad a las distintas unidades que componen la población de estudio, las estimaciones de resultados y errores por dominios de interés que pueden conocerse por experiencias pasadas, las tasas de crecimiento de la población por distintas áreas, los índices del desarrollo edilicio en las zonas urbanas, entre otros, y los indicadores económicos y sociales como son por ejemplo el conocimiento

de variables que permitan agrupar a la población en grupos que respondan a niveles de ingreso, hacinamiento, educación, actividad económica, así como el cruzamiento de estos indicadores con las características geológicas y telúricas del país, permiten al estadístico poder diseñar una investigación que aumente la acuracidad de los resultados sin aumentar los costos y manteniendo tamaños muestrales reducidos. Estos indicadores pueden ser logrados a través de las investigaciones, continuas o no, que realizan las Oficinas Nacionales de Estadística, aunque el objetivo básico de la investigación no los contemple.

Por otro lado, el hecho de diseñar muestras pequeñas permite tener un mejor control de la información recogida. La experiencia nos ha demostrado en Latinoamérica, que muchos de los problemas que entorpecen las investigaciones estadísticas se deben frecuentemente al hecho de usar muestras grandes, en las cuales interviene mucha gente que escapa, tanto ellos como la información que recogen, a los controles establecidos, creando posteriormente un serio entorpecimiento en el procesamiento y análisis de los datos.

5. Entre las reglas prácticas que generalmente se utilizan para la determinación de los distintos estratos podemos mencionar:

- i) Regla de Dalenius - Hodges. Acepta el supuesto que la distribución de la variable de estudio dentro de un estrato es constante, en especial si los estratos son numerosos y estrechos. En este caso se forman estratos igualando el acumulado de $\sqrt{f(y)}$, donde $f(y)$ es la distribución de frecuencias de la variable de estudio y .
- ii) Regla de Ekman. Los límites se construyen de tal manera que los valores $N_h R_h$ sean constantes en cada estrato. N_h es el total de unidades en el estrato h y R_h es la amplitud de intervalo del estrato h .

- iii) Regla de Dalenius - Gurney. Los estratos se forman haciendo constantes los valores de $N_h \sqrt{V_h}$ en los distintos estratos. V_h corresponde a la varianza en el estrato h.
- iv) Regla de Mahalanobis - Hansen - Hurwitz - Madow. Se trata de formar estratos haciendo constantes los totales por estrato de la variable de estudio. Esta regla es eficiente cuando los coeficientes de variación por estrato son aproximadamente iguales.

El problema que a continuación se formula en este borrador tiene que ver con la determinación del límite en la construcción de dos estratos, considerando una variable de estudio.

6. Presentación del Problema. Consideremos un conjunto finito de valores positivos c_i , con $i = 1, 2, \dots, n$, ordenados en forma no decreciente, cuyo intervalo de variación es

$$g \leq c_i \leq G$$

donde g es el minorante y G el mayorante del conjunto.

Dividimos al conjunto en dos subconjuntos I y II de acuerdo con un valor arbitrario K , con $g \leq K \leq G$.

Definimos

$$c_i = x_i + y_i \quad (1)$$

con

$$x_i = \begin{cases} c_i & \text{si } c_i \leq K \\ K & \text{si } c_i > K \end{cases}$$

e

$$y_i = \begin{cases} c_i - K & \text{si } c_i > K \\ 0 & \text{si } c_i \leq K \end{cases}$$

por tanto,

$$g \leq x_i \leq K \quad \text{y} \quad 0 \leq y_i \leq G - K.$$

Haciendo

$$\bar{c} = \bar{x} + \bar{y} \quad (2)$$

es

$$\bar{c} = \frac{1}{n} \sum_{i=1}^n c_i \quad \text{media del conjunto}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{media modificada del subconjunto I}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{media modificada del subconjunto II.}$$

Podemos establecer la siguiente función de K

$$f(K) = \text{Var } x + \text{Var } y \quad (3)$$

con

$$\text{Var } x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

y

$$\text{Var } y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

El problema consiste en determinar el valor K para el cual la función f (K) es un mínimo.

7. Para que la función $f(K)$ pase por un mínimo, debe ser

$$\frac{d}{dK} f(K) = 0 \quad (4)$$

o sea

$$\frac{d}{dK} \text{Var } x + \frac{d}{dK} \text{Var } y = 0 \quad (5)$$

8. Las derivadas parciales de la ecuación (5) son

$$\frac{d}{dK} \text{Var } x = \frac{1}{n} \left\{ 2KG - 2K^2 - 2\bar{x}(G - K) \right\} \quad (6)$$

(ver Apéndice punto 1)

y

$$\frac{d}{dK} \text{Var } y = \frac{1}{n} \left\{ 2KG - K^2 - G^2 - 2\bar{y}(K - G) \right\} \quad (7)$$

(ver Apéndice punto 2)

9. Reemplazando (6) y (7) en (5) y desarrollando el álgebra se obtiene la siguiente ecuación de segundo grado en K

$$K^2 - \frac{A}{3}K - \frac{B}{3} = 0 \quad (8)$$

donde

$$A = 2(2G + \bar{x} - \bar{y})$$

y

$$B = -2G(\bar{x} - \bar{y}) - G^2$$

(ver Apéndice punto 3)

10. Las dos raíces que satisfacen a la ecuación (8) son $K_1 = G$

$$\text{y } K_2 = \frac{G + 2(\bar{x} - \bar{y})}{3} \quad (9)$$

(ver Apéndice punto 4)

11. Resumiendo

$$f(K) = \text{Var } x + \text{Var } y \quad (10)$$

siendo la derivada primera (ver Apéndice punto 3)

$$n f'(K) = -3K^2 + AK + B \quad (11)$$

la cual se anula para $K_1 = G$ y para $K_2 = \frac{G + 2(\bar{x} - \bar{y})}{3}$

La derivada segunda es

$$n f''(K) = \frac{-4}{n} K^2 + \left(\frac{8G}{n} - 6\right) K + 4G - \frac{4G^2}{n} + 2(\bar{x} - \bar{y})$$

la cual para un n^2 grande podemos escribir

$$f''(K) = \frac{1}{n} (-6K + A) \quad (12)$$

Evaluándola para K_1 y K_2 (ver Apéndice punto 5),

$$\left. \begin{array}{l} f''(K) < 0 \\ K = G \end{array} \right\} \quad (13)$$

y

$$\left. \begin{array}{l} f''(K) > 0 \\ K = \frac{G + 2(\bar{x} - \bar{y})}{3} \end{array} \right\} \quad (14)$$

Por tanto la función $f(K)$ tiene un máximo en $K = G$ y tiene un mínimo en $K = \frac{G + 2(\bar{x} - \bar{y})}{3}$. En el punto $K = G$ y en el punto

$K = g$ es $f(K) = \text{Var } c$.

12. El valor de $K = \frac{G + 2(\bar{x} - \bar{y})}{3}$ que minimiza a $f(K)$ depende de un K_a de cálculo que determina a las dos medias modificadas. Un criterio para determinar K es hallar el valor en el cual K se iguala a K_a . Para ello escribimos

$$K = \frac{G}{3} + \frac{2}{3} (\bar{c} - 2\bar{y}) \quad (15)$$

dado que $\bar{x} = \bar{c} - \bar{y}$.

Además por definición

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{II} (c_i - K_a) \\ &= \frac{1}{n} \sum_{II} c_i - \frac{n_2}{n} K_a \end{aligned} \quad (16)$$

donde n_2 es el número de valores comprendidos a la derecha de K_a o sea pertenecientes al subconjunto II. Por tanto

$$\begin{aligned} K &= \frac{G}{3} + \frac{2}{3} \left(\bar{c} - \frac{2}{n} \sum_{II} c_i + \frac{2n_2}{n} K_a \right) \\ &= \frac{G}{3} + \frac{2}{3} \bar{c} - \frac{4}{3n} \sum_{II} c_i + \frac{4n_2}{3n} K_a \end{aligned} \quad (17)$$

Llamando

$$Q = \frac{G}{3} + \frac{2}{3} \bar{c}$$

y

$$q = \frac{4}{3n}$$

queda

$$K = Q + q \left(n_2 K_a - \sum_{II} c_i \right) \quad (18)$$

siendo Q y q constantes. Dándole valores a K_a se obtienen valores para n_2 para $\sum_{II} c_i$ y para K . El valor de K buscado será aquél que haga

un mínimo a la diferencia $|K_a - K|$.

13. Ejemplo hipotético. Consideremos el conjunto ordenado de cuatro valores: 2,3,4 y 7. Asignando valores a K_a construimos el siguiente cuadro:

K_a	n_2	$\sum_{II} c_i$	K	$ K_a - K $
2	3	14	$\frac{7}{3}$	$\frac{1}{3}$
3	2	11	$\frac{10}{3}$	$\frac{1}{3}$
4	1	7	4	0
7	0	0	5	2

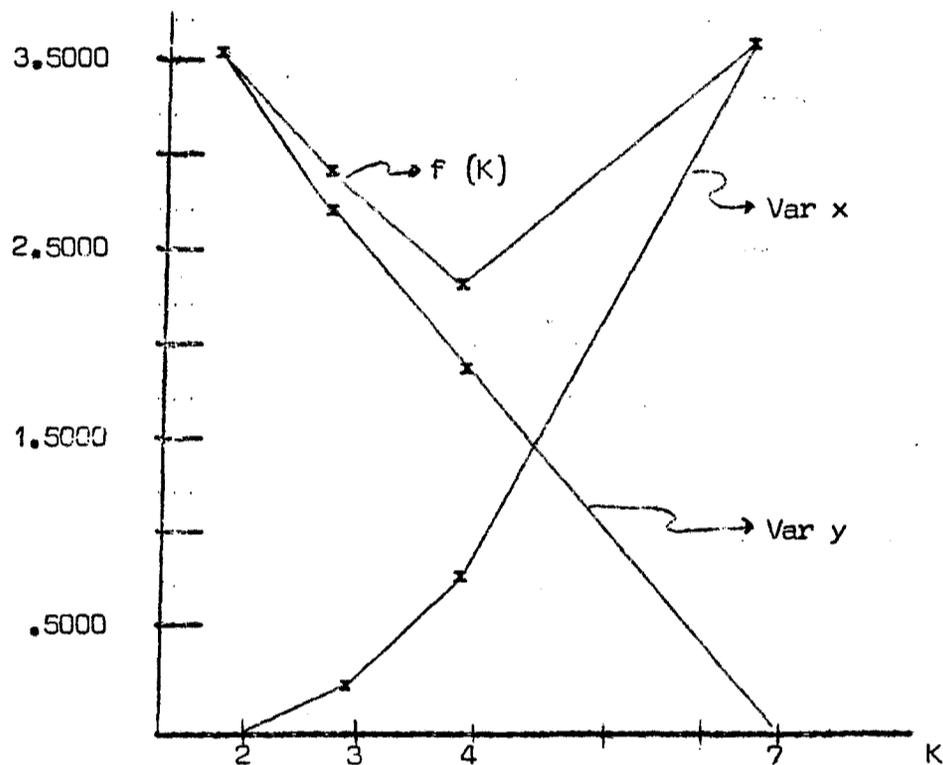
La diferencia $|K_a - K|$ es mínima para $k = 4$.

14. Comprobación.

K	$\text{Var } x$	$\text{Var } y$	$f(K) = \text{Var } x + \text{Var } y$
2	.0000	3.5000	3.5000
3	.1875	2.6875	2.8750
4	.6875	1.6875	2.3750
7	3.5000	0.0000	3.5000

donde se observa que para $K = 4$ la función $f(K)$ adquiere su valor mínimo.

15. Gráficamente



16. Se debe tener en cuenta que el criterio adoptado en párrafo 12 para determinar K no garantiza que $f(K)$ pase por un mínimo. Además la variable utilizada es discreta.

En el ejemplo que presenta Cochran, Cuadro 5 A. 12, el K que se obtiene aplicando este criterio corresponde al intervalo 40-45, cuando en rigor $f(K)$ se minimiza para un K que cae en el intervalo 20-25. La diferencia relativa entre ambos $f(K)$ es en este caso del 16 por ciento.

Apéndice

Punto 1. Demostrar que

$$\frac{d}{dK} \text{Var } x = \frac{1}{n} \left\{ 2KG - 2K^2 - 2\bar{x} (G - K) \right\} \quad (1.1)$$

1.1 Escribimos a $\text{Var } x$ como

$$\text{Var } x = \frac{1}{n} \sum_i^n x_i^2 - \frac{1}{n} \left(\sum_i^n x_i \right)^2 \quad (1.2)$$

siendo por definición

$$\sum_i^n x_i^2 = \sum_I c_i^2 + \sum_{II} K^2 \quad (1.3)$$

donde \sum_I indica la suma sobre el subconjunto I y \sum_{II} indica la

suma sobre el subconjunto II.

1.2 La derivada de $\sum_i^n x_i^2$ con respecto a K es

$$\frac{d}{dK} \sum_i^n x_i^2 = \frac{d}{dK} \sum_I c_i^2 + \frac{d}{dK} \sum_{II} K^2 \quad (1.4)$$

Si tomamos

$$\begin{aligned} \frac{d}{dK} \sum_I c_i^2 &= \frac{d}{dK} \int_g^K c^2 dc \\ &= \frac{d}{dK} \left[\frac{c^3}{3} \right]_g^K = K^2 \quad (1.5) \end{aligned}$$

y

$$\begin{aligned} \frac{d}{dK} \sum_{II} K^2 &= \frac{d}{dK} \int_K^G K^2 dc \\ &= \frac{d}{dK} \left\{ K^2 \left[c \right]_K^G \right\} = 2KG - 3K^2 \quad (1.6) \end{aligned}$$

se obtiene que

$$\frac{d}{dK} \sum_i^n x_i^2 = 2KG - 2K^2 \quad (1.7)$$

1.3 La derivada de $\left(\sum_i^n x_i \right)^2$ con respecto a K es

$$\frac{d}{dK} \left(\sum_i^n x_i \right)^2 = 2 \left(\sum_i^n x_i \right) \frac{d}{dK} \sum_i^n x_i \quad (1.8)$$

donde por definición,

$$\frac{d}{dK} \sum_i^n x_i = \frac{d}{dK} \left\{ \sum_I c_i + \sum_{II} K \right\} \quad (1.9)$$

Hallando las derivadas parciales y reemplazando en (1.8), obtenemos

$$\frac{d}{dK} \left(\sum_i^n x_i \right)^2 = 2 \left(\sum_i^n x_i \right) (G - K) \quad (1.10)$$

1.4 De (1.2) podemos escribir que

$$\frac{d}{dK} \text{Var } x = \frac{1}{n} \left\{ \frac{d}{dK} \sum_i^n x_i^2 - \frac{1}{n} \frac{d}{dK} \left(\sum_i^n x_i \right)^2 \right\} \quad (1.11)$$

Reemplazando (1.7) y (1.10) en (1.11) se obtiene la expresión (1.1) que se quería demostrar.

Punto 2. Demostrar que

$$\frac{d}{dK} \text{Var } y = \frac{1}{n} \left\{ 2KG - K^2 - 2\bar{y} (K - G) \right\} \quad (2.1)$$

2.1 Para determinar la derivada de Var y con respecto a K seguimos un procedimiento similar al utilizado en el punto 1 para determinar la derivada de Var x.

Teniendo presente que por definición es

$$\sum_i^n y_i^2 = \sum_{II} c_i^2 + \sum_{II} K^2 - 2K \sum_{II} c_i \quad (2.2)$$

y que

$$\sum_i^n y_i = \sum_{II} c_i - \sum_{II} K \quad (2.3)$$

podemos hacer

$$\frac{d}{dK} \sum_i^n y_i^2 = 2KG - K^2 - G^2 \quad (2.4)$$

y

$$\frac{d}{dK} \left(\sum_i^n y_i \right)^2 = 2 \left(\sum_i^n y_i \right) (K - G) \quad (2.5)$$

Por tanto

$$\frac{d}{dK} \text{Var } y = \frac{1}{n} \left\{ 2KG - K^2 - G^2 - 2\bar{y}(K - G) \right\} \quad (2.6)$$

que es lo que se quería demostrar.

Punto 3. La derivada primera de $f(K)$ se obtiene de (1.1) y (2.1)

$$n f'(K) = 2KG - 2K^2 - 2\bar{x}(G - K) + 2KG - K^2 - G^2 - 2\bar{y}(K - G) \quad (3.1)$$

Operando el álgebra e igualando a cero queda la ecuación de segundo grado en K

$$-3K^2 + K(2G + \bar{x} - \bar{y}) - 2G(\bar{x} - \bar{y}) - G^2 = 0 \quad (3.2)$$

y llamando

$$A = 2 (2 G + \bar{x} - \bar{y}) \quad (3.3)$$

$$B = - 2 G (\bar{x} - \bar{y}) - G^2 \quad (3.4)$$

se obtiene la expresión (B) que es

$$K^2 - \frac{A}{3} K - \frac{B}{3} = 0 \quad (3.5)$$

Punto 4. El discriminante de la ecuación $K^2 - \frac{A}{3} K - \frac{B}{3} = 0$

es $\frac{A^2}{9} + \frac{4}{3} B$.

Sustituyendo los valores de A y B y operando algebraicamente se obtiene que

$$\frac{A^2}{9} + \frac{4}{3} B = \frac{1}{9} \left[2 G - 2 (\bar{x} - \bar{y}) \right]^2 \quad (4.1)$$

La resolvente de la ecuación es

$$K = \frac{1}{2} \left\{ \frac{A}{3} \pm \left(\frac{A^2}{9} + \frac{4}{3} B \right)^{\frac{1}{2}} \right\} \quad (4.2)$$

Reemplazando los valores de (3.3) y de (4.1) en (4.2) se obtienen las dos raíces $K_1 = G$ y $K_2 = \frac{G + 2 (\bar{x} - \bar{y})}{3}$ que satisfacen a la ecuación.

Punto 5. La derivada de $f(K)$ es, ver (3.1),

$$n f'(K) = -3K^2 + K^2(2G + \bar{x} - \bar{y}) - 2G(\bar{x} - \bar{y}) - G^2 \quad (5.1)$$

5.1 Tomando $\frac{d}{dK} \bar{x} = \frac{G-K}{n}$ y $\frac{d}{dK} \bar{y} = \frac{K-G}{n}$

se puede escribir para un n^2 grande

$$n f''(K) = -6K + 2(2G + \bar{x} - \bar{y}) \quad (5.2)$$

Evaluando para $K = G$

$$f''(K) \Big|_{K=G} = -6G + 4G + 2(\bar{x} - \bar{y}) < 0 \quad (5.3)$$

por ser $G > \bar{x}$. Por tanto para $k = G$ la función $f(K)$ pasa por un máximo.

5.2 Se observa que cuando $K = G$ por definición es $x_i = c_i$ e $y_i = 0$, siendo por tanto $f(K = G) = \text{Var } c$.

5.3 Evaluando la derivada segunda para $K = \frac{G + 2(\bar{x} - \bar{y})}{3}$ se

obtiene

$$f''(K) \Big|_{K = \frac{G + 2(\bar{x} - \bar{y})}{3}} = 2(G + y - x) > 0 \quad (5.4)$$

es decir en este punto la función $f(K)$ pasa por un mínimo.

5.4 Si tomamos $K = g$ será $x_i = g$ e $y_i = c_i - g$, por definición de variable, obteniéndose que $f(K = g) = \text{Var } c$.

Bibliografía

1. Cavallini, C.M., Informe de la Misión de Asesoría Realizada en el Perú. CEPAL, Julio de 1976
2. Cochran, W.G., Técnicas de Muestreo. Primera Edición en Español. Compañía Editorial Continental, S.A. México, 1971
3. Kendall, M.G., The History and Future of Statistics. Statistical paper in honor of G.W. Snedecor. The Iowa State University Press, Ames, Iowa 1972
4. Ranjan Kumar Som, A Manual of Sampling Techniques. Heinemann Educational Books Ltd, London, 1973

1. The first part of the document is a list of names and addresses.

2. The second part is a list of names and addresses.

3. The third part is a list of names and addresses.

4. The fourth part is a list of names and addresses.

5. The fifth part is a list of names and addresses.

6. The sixth part is a list of names and addresses.

7. The seventh part is a list of names and addresses.

8. The eighth part is a list of names and addresses.

9. The ninth part is a list of names and addresses.

10. The tenth part is a list of names and addresses.

11. The eleventh part is a list of names and addresses.

12. The twelfth part is a list of names and addresses.

13. The thirteenth part is a list of names and addresses.

14. The fourteenth part is a list of names and addresses.

15. The fifteenth part is a list of names and addresses.