

RESTRINGIDA

E/CEPAL/R.250

9 de marzo de 1981

ORIGINAL: ESPAÑOL

C E P A L
Comisión Económica para América Latina

ANALISIS DE LA VARIACION DE UN FACTOR EN LAS
ENCUESTAS DE HOGARES

Carlos Cavallini */
Asesor Regional en Muestreo para
Estadísticas Demográficas

*/ Las opiniones expresadas en este documento son de exclusiva responsabilidad del autor y pueden no coincidir con las de la Organización.

81-2-280-170

ANALISIS DE LA VARIACION DE UN FACTOR EN LAS ENCUESTAS DE HOGARES

1. Las investigaciones en base a encuestas de hogares, que tratan sobre el estudio de variables basadas en modelos estadísticos, están sujetas a numerosas fuentes de error. Algunos de estos errores se pueden predecir, se pueden estimar y se pueden evitar, algunos se pueden predecir, se pueden estimar, pero no se pueden evitar y algunos no se pueden predecir ni evitar. Entre los primeros podemos mencionar, por ejemplo, el error o efecto propio de los entrevistadores que es introducido en la captación de la información, el cual podemos predecir y evitar. Entre los segundos podemos citar el uso de determinados marcos muestrales de población, de los cuales podemos predecir que están incompletos y podemos estimar la omisión, pero no podemos corregirlos dado que las unidades que no fueron censadas no pueden ser incluidas, salvo que se realice otro censo, lo cual daría otro marco. Entre los terceros, existen un cúmulo de errores ligados a componentes de propaganda, formas de pagos, medios de transporte, problemas personales, etapas de trabajos manuales, organización, etc. que se hace difícil poder predecir y más aún evitar.

Muchos de estos errores pueden introducir en los resultados sesgos grandes, macro sesgos, mientras que otros pueden producir sesgos pequeños, micro sesgos.

De allí que es importante, antes de comenzar una investigación, tener una idea de las fuentes productoras de sesgos y de la magnitud de los mismos, de manera tal de atacar, en función de los recursos disponibles, a aquellas fuentes cuyos sesgos se estiman más perjudiciales para los resultados que se persiguen. En las encuestas de hogares en base a muestras probabilísticas, los macro sesgos se originan, generalmente, si no se toman cuidados especiales, en los marcos muestrales y en la medición de las unidades de observación por parte de los entrevistadores.

Si un marco muestral no representa confiablemente a la población de estudio, ya sea por ser anacrónico o por no estar completo, los resultados obtenidos en base a una muestra seleccionada de dicho marco, serán buenos para el marco pero no para la población que se quiere estudiar. En cuanto

a la medición de las unidades, si los valores obtenidos no son los reales, y si no se han adoptado técnicas de evaluación el error, esta información errónea podrá invalidar, a la postre, la investigación realizada.

Se hace oportuno mencionar en este punto, que existe una necesidad a producir estadísticas más eficientes y a más bajo costo. Ello es debido a la demanda creciente que hay de las mismas en los países de América Latina pero que no lleva aparejado un similar aumento de recursos. Uno de los factores principales para bajar los costos por unidad de eficiencia es promover la aplicación de técnicas estadísticas más avanzadas. No se debe confundir "avance técnico" con "avance de dificultades", pues casi siempre, aquél facilita y aclara las distintas etapas de la investigación. Las herramientas estadísticas como la aleatorización, la estratificación y la replicación, ayudan a explicar mejor el fenómeno que se estudia, permiten optimizar la función costo-eficiencia y ayudan en la coordinación de trabajos y en la asociación de resultados.

A continuación se presentan algunas consideraciones a tener en cuenta cuando se conoce una fuente de variación o factor y se desea conocer el efecto del mismo a distintos niveles.

2. Consideremos a una super-población de M unidades dividida aleatoriamente en K poblaciones de N unidades cada población, siendo $M = N K$.

Simbolizando con Y_{ij} al valor real correspondiente a la j-unidad de la i-población y simbolizando con y_{ij} al valor observado de esa ij-unidad, establecemos los siguientes 2 modelos

$$Y_{ij} = u_i + e_{ij} \quad (1)$$

$$y_{ij} = u_i + g_i + e_{ij} \quad (2)$$

donde u_i es la media parámetro correspondiente a la i-población; g_i , es el efecto introducido en la medición de la unidad por un cierto factor cualitativo cuya fuente de origen se conoce; y e_{ij} , es el efecto residual

debido a factores fortuitos que se desconocen, siendo $E_j e_{ij} = 0$.

3. Ambas variables, Y_{ij} e y_{ij} , se distribuyen dentro de cada población en forma normal,

$$Y_{ij} \sim N(u_i; \sigma_o^2) \quad (3)$$

$$y_{ij} \sim N(u_i + g_i; \sigma_o^2) \quad (4)$$

Se observa que ambas σ_o^2 son iguales, por ser g_i constante dentro de cada nivel i de observación, aunque g_i puede ser distinta entre los distintos niveles.

4. Por ejemplo, si se necesita conocer qué tipo de entrevistador es más eficaz para medir el ingreso de los hogares en el estrato socioeconómico alto, tendremos que el factor es el entrevistador y su nivel puede ser, por caso, el sexo - edad - educación. La idea fundamental es comparar la variación ocasionada por la acción del factor con la variación residual. Si la diferencia entre ambas variaciones es significativa, ello indicaría que el factor ejerce una influencia considerable en las observaciones. En este caso, las medias de cada nivel se diferenciarán también de manera significativa. Si se establece que el factor influye considerablemente en las observaciones, luego habrá que determinar cuál de los niveles es el más realista, haciendo, por ejemplo, un control de la calidad de la información recogida.

5. Seleccionando una muestra de k poblaciones y dentro de cada población seleccionando una muestra de n unidades, de (1) y (2) obtenemos, sumando sobre $j = \overline{1;n}$ y dividiendo por n que

$$\bar{y}_i = \bar{Y}_i + g_i \quad (5)$$

6. La varianza estimada de \bar{y}_i , $v(\bar{y}_i)$, es

$$v(\bar{y}_i) = v(\bar{Y}_i) + v(g_i) \quad (6)$$

aceptando que la covarianza de $(\bar{Y}_i; g_i)$ es cero, por ser g_i independiente de \bar{Y}_i .

7. En (5) y (6) observamos que de no existir el efecto g_i , será $\bar{y}_i = \bar{Y}_i$ y $v(\bar{y}_i) = v(\bar{Y}_i)$, siendo $v(\bar{Y}_i) = \frac{\sigma_0^2}{n}$. Es decir, en este caso las diferencias entre las medias de los distintos niveles son atribuibles a factores aleatorios propio de las fluctuaciones muestrales. Pero es evidente, que de existir considerables diferencias entre las medias observadas de cada nivel, debido a la introducción del efecto g_i , será

$$v(\bar{y}_i) = \frac{\sigma_0^2}{n} + \sigma_1^2 \quad (7)$$

donde a $v(g_i)$ se la ha considerado como un estimador de σ_1^2 , componente aditivo de la varianza estimada de las medias observadas.

8. Es decir, a la varianza estimada de las medias observadas, $v(\bar{y}_i)$, se la puede considerar como un estimador de $\frac{\sigma_0^2}{n} + \sigma_1^2$, donde $\frac{\sigma_0^2}{n}$ es la

varianza residual, no explicable, y σ_1^2 es un componente de variación debido a la introducción, de un factor cualitativo conocido, en las poblaciones de estudio.

9. Para analizar un resultado, en nuestro caso $\bar{y} = \frac{1}{k} \sum_i^k \bar{y}_i$,

es necesario hacer un análisis de su varianza. Generalmente son muchas las fuentes de variación que afectan a una observación o a un resultado, y el problema consiste en separar y estimar estas fuentes de variación de manera tal de poder explicar el porqué de dicho resultado.

El ejemplo presentado en punto 4, es similar a seleccionar una muestra de hogares en el estrato alto y dividirla aleatoriamente en k submuestras, o replicaciones, donde cada una de ellas representa en igual grado a la población de hogares del estrato. Se asigna, luego, a cada replicación un entrevistador de distinto nivel. La media de la muestra, \bar{y} , se verá afectada, por tanto, i) por un componente de variación debido a las fluctuaciones propia de la muestra, $\frac{\sigma_0^2}{n}$ y ii) por otro componente de variación debido al factor entrevistador, σ_1^2 . Lo que deseamos averiguar es la significación de σ_1^2 , dado que si ella es importante estaría indicando

que algunos entrevistadores, o todos, podrían estar introduciendo en las observaciones algún efecto que altera el valor real de la observación. Si este es el caso, la media, \bar{y} , no estaría estimando acuradamente a la media poblacional y la distribución de la variable, en nuestro caso ingreso, no representaría a la real distribución del ingreso de la población.

Con todo esto se quiere resaltar la importancia que tiene la replicación en los diseños muestrales, en especial cuando se deben medir variables sensibles, como es el ingreso, y donde la calidad o manera de ser del entrevistador puede afectar la respuesta.

Por otro lado, el trabajo operativo que se lleve a cabo en áreas urbanas, en función de un diseño replicado, no ha de aumentar los costos, y en las áreas rurales estos costos estarán condicionados, principalmente, a la estratificación, a la definición de las unidades muestrales que se hagan, y a la organización de campo. Generalmente, la ganancia analítica que se obtiene, del fenómeno bajo estudio, aumenta por unidad de costo cuando se utiliza el método de las repeticiones.

10. Retomando el tema, para conocer cuál de los 2 sumandos de la (7) influye más en la $v(\bar{y}_i)$ tendríamos que comparar σ_1^2 con $\frac{\sigma_0^2}{n}$. A σ_1^2 no lo podemos estimar directamente, pero sí podemos calcular $v(\bar{y}_i)$ y estimar $\frac{\sigma_0^2}{n}$ con s_0^2 . El cociente $\frac{n v(\bar{y}_i)}{s_0^2}$ estima a $\frac{\sigma_0^2 + n \sigma_1^2}{\sigma_0^2}$

y se distribuye como una F. Si F es 1 significa que $\sigma_1^2 = 0$, por tanto no existiría el efecto del factor considerado. Solamente valores grandes de F supondrían que el factor ejerce un efecto significativo en las observaciones obtenidas.

11. La mejor estimación de σ_0^2 es

$$s_0^2 = \frac{1}{k(n-1)} \sum_i^k \sum_j^n (y_{ij} - \bar{y}_i)^2 \quad (8)$$

que es la varianza media observada en cada replicación. Se observa que se utilizan los desvíos observados, que se conocen, en lugar de los desvíos reales, que no se conocen, por ser ambos iguales,

$$(y_{ij} - \bar{y}_i) = (Y_{ij} - \bar{Y}_i) \quad (9)$$

dado que g_i es constante para el i - nivel.

12. La varianza estimada de \bar{y}_i es

$$v(\bar{y}_i) = \frac{1}{k-1} \sum_i^k (\bar{y}_i - \bar{y})^2 \quad (10)$$

13. Las sumas de cuadrados y los grados de libertad correspondiente a los distintos factores de variación se pueden tabular en el siguiente cuadro denominado Análisis de la Varianza.

Análisis de la Varianza

Factor de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Parámetros Estimados por los Cuadrados Medios
Cualitativo, entre replicaciones	$B = n \sum_i^k (\bar{y}_i - \bar{y})^2$	k-1	$\frac{B}{k-1}$	$\sigma_0^2 + n \sigma_1^2$
Residual, dentro de las replicaciones	$A = \sum_i^k \sum_j^n (y_{ij} - \bar{y}_i)^2$	k(n-1)	$\frac{A}{k(n-1)}$	σ_0^2
Total	A + B	nk-1	-	-

Esta forma de presentar los resultados, además de clarificar la propiedad aditiva de la suma de los cuadrados y de los grados de libertad, permite una útil y rápida comparación entre los cuadrados medios.

14. La hipótesis nula que se plantea puede ser formulada de distintas maneras

$$H_0 : \sigma_1^2 = 0 \quad (11)$$

$$H_0 : \sigma_i = 0 \quad i = \overline{1;K} \quad (12)$$

$$H_0 : E y_{ij} = u_i \quad (13)$$

$$H_0 : u_1 = u_2 = \dots = u_i = \dots = u_K \quad (14)$$

Generalmente, el test de la F, donde

$$F = \frac{\text{Cuadrados medios entre replicaciones}}{\text{Cuadrados medios dentro de las replicaciones}}$$

con $(k-1)$ y $k(n-1)$ grados de libertad, se hace al nivel de significación del .05 y del .01. Si el valor obtenido por F es mayor que el dado en la tabla, ello estaría indicando que la H_0 no debería aceptarse, y que por tanto el factor cualitativo está introduciendo un efecto significativo en algunos, o en todos, los niveles del factor considerado.

