El efecto de diseño

Andrés Gutiérrez

Asesor Regional en Estadísticas Sociales - Comisión Económica para América Latina y el Caribe (CEPAL) - andres.gutierrez@cepal.org

Tabla de contenidos I

El efecto de diseño

Introducción

- En diseños de muestreo complejo, la falta de independencia entre las observaciones es común.
- La distribución de la variable de interés varía entre individuos debido a la complejidad del muestreo en encuestas de hogares.
- La inferencia correcta debe considerar las desviaciones significativas respecto al análisis estadístico clásico, que asume muestras aleatorias simples.
- A menudo, se requiere aumentar el tamaño de la muestra para lograr la precisión deseada.

Definición del Efecto de Diseño

- El efecto de diseño fue definido por Kish (1965, 258) como *la relación entre la varianza real de una muestra y la varianza real de una muestra aleatoria simple del mismo número de elementos*.
- Se expresa mediante la fórmula:

$$DEFF = \frac{Var(\hat{\theta})}{Var_{MAS}(\hat{\theta})}$$

En donde $Var(\hat{\theta})$ denota la varianza de un estimador $\hat{\theta}$ bajo un diseño de muestreo complejo $p(s)$ y $Var_{MAS}(\hat{\theta})$ denota la varianza del este estimador $\hat{\theta}$ bajo un diseño de muestreo aleatorio simple MAS

Interpretaciones del Efecto de Diseño

UN (2008, 49) ofrece varias formas de interpretar el efecto del diseño:

1. Como el factor por el cual la varianza del diseño de muestreo complejo es mayor que la de una muestra aleatoria simple del mismo tamaño.
2. Como la medida de cuánto peor es el plan de muestreo real que la muestra aleatoria simple en términos de precisión.
3. Como un reflejo de cuántos casos de muestra más tendrían que seleccionarse en el diseño de muestra planificado en comparación con una muestra aleatoria simple para lograr el mismo nivel de varianza de muestreo.

Estimación del efecto de diseño

Estimación del efecto de diseño

- El DEFF (Efecto de Diseño) depende del diseño muestral $p(s)$ y del estimador del parámetro θ.
- No es correcto describir al DEFF únicamente como una medida de eficiencia del diseño muestral, ya que puede variar según el parámetro que se esté estimando.
- Ninguno de los componentes del efecto de diseño es conocido y debe ser estimado.
- Un estimador insesgado de la varianza poblacional $S^2_{y_U}$ es la varianza muestral ponderada, expresada como

$$\hat{S}^2_{y_U} = \left(\frac{n}{n-1}\right) \frac{\sum_s w_k (y_k - \hat{\theta})^2}{\sum_s w_k - 1}$$

Estimación del Efecto de Diseño

- Para el caso de un promedio poblacional θ, la estimación de la varianza $Var_{MAS}(\hat{\theta})$ bajo muestreo aleatorio simple se calcula con la expresión

$$\widehat{Var}_{MAS}(\hat{\theta}) = \frac{1}{n}\left(1 - \frac{n}{\hat{N}}\right)\hat{S}^2_{y_U}.$$

En donde $\hat{N} = \sum_s w_k$

- La estimación del DEFF se obtiene a partir de la relación

$$\widehat{DEFF} = \frac{\widehat{Var}(\hat{\theta})}{\widehat{Var}_{MAS}(\hat{\theta})}$$

.

Evaluación del Estimador bajo Diferentes Escenarios

- La idea central del efecto de diseño consiste en evaluar el mismo estimador bajo distintos escenarios de muestreo.
- Al estar el estimador ponderado por factores de expansión de la encuesta, es conveniente utilizar el mismo criterio para evaluar ambas estrategias de muestreo.
- Es posible encontrar una discusión más profunda sobre el efecto de diseño en Gambino (2009, sec. 4.), Särndal, Swensson, y Wretman (2003, 188) y Gutiérrez, Zhang, y Montaño (2016, 101).

Descomposición del efecto de diseño en las encuestas de hogares

Descomposición del Efecto de Diseño (I):

Park (2003) propone que el efecto de diseño de una encuesta se puede descomponer en tres partes multiplicativas:

- $DEFF^W$ se debe a la ponderación desigual y puede ser estimado utilizando la expresión

$$DEFF^W = 1 + cv^2(w_k)$$

  En donde $cv(w_k)$ representa el coeficiente de variación de los pesos de muestreo w_k de las unidades en la encuesta
- $DEFF^S$ se refiere al efecto debido a la estratificación, evaluado por la relación

$$DEFF^S \cong \frac{\text{Varianza dentro de los estratos}}{\text{Varianza Total}}$$

Descomposición del Efecto de Diseño (II):

- $DEFF^C$ es el efecto debido a la conglomeración y está definido por

$$DEFF^C = 1 + (\bar{n}_{II} - 1)\rho_y$$

  En donde, $\bar{n}_{II}$ es el número de hogares promedio que se seleccionan en cada UPM, y ρ_y es el coeficiente de correlación intraclase.

- La fórmula general es

$$DEFF = DEFF^W \times DEFF^S \times DEFF^C$$

.

Control sobre el Efecto de Diseño

- El control sobre el valor de ρ_y es limitado una vez definido el marco de muestreo.
- Estrategias sugeridas para mitigar el efecto de conglomeración incluyen seleccionar más UPM, definir UPM más pequeñas y seleccionar un número fijo de viviendas dentro de las UPM.
- UN (2008) propone estrategias como seleccionar tantas UPM como sea posible, definir UPM pequeñas, seleccionar un número fijo de viviendas y usar muestreo sistemático en UPM.
- La segunda componente $DEFF^S$ debe ser mínima para los indicadores estudiados.

Control sobre el Efecto de Diseño

- El efecto de conglomeración y el uso del muestreo en varias etapas $DEFF^C$ debe ser mínimo.
- Se sugiere minimizar $DEFF^W$ al decidir qué variables de control utilizar en la calibración de estimadores, considerando la correlación entre indicadores de encuestas de hogares.
- La estrategia tripartita busca minimizar el efecto de diseño general de las encuestas asegurando que cada componente sea pequeña.
- Se destaca la importancia de decidir la relación entre UPM y hogares, así como la cantidad de UPM y hogares seleccionados.

Formas comunes del efecto de diseño

Estimación de la Varianza para Diferentes Parámetros

- Para la media poblacional ($\bar{y}$), la varianza del estimador bajo diseño de muestreo complejo se expresa como

$$Var(\hat{\bar{y}}) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right) S^2_{y_U}$$

- Para la proporción poblacional (P), la varianza del estimador se expresa como

$$Var(\hat{P}) = \frac{DEFF}{n}\left(1 - \frac{n}{N}\right) P(1-P)$$

Coeficiente de Correlación Intraclase (ρ_y)

- En diseños muestrales multietápicos, se utiliza el coeficiente de correlación intraclase para medir la homogeneidad dentro de los conglomerados.
- Se define como

$$\rho_y = 1 - \frac{N_I}{N_I - 1}\frac{SCD}{SCT}$$

donde,

- $SCT = \sum_U (y_k - \bar{y}_U)^2$ hace referencia a la suma de cuadrados total,
- $SCE = \sum_{U_I} N_I(\bar{y}_{U_I} - \bar{y}_U)^2$ es la suma de cuadrados entre, y
- $SCD = SCT - SCE$ es la suma de cuadrados entre.

Factores que Afectan el Efecto de Diseño

- El efecto de diseño ($DEFF$) aumenta cuando el coeficiente de correlación (ρ_y) crece, lo cual depende de la distribución de la variable de interés entre las UPM.
- También aumenta si el promedio de hogares seleccionados por UPM aumenta, un factor que se controla en la etapa de diseño de la encuesta.
- El coeficiente de correlación (ρ_y) es afectado por la distribución de la variable de interés entre las UPM, especialmente en casos de segregación entre hogares ricos y pobres.
- El promedio de hogares seleccionados por UPM es un factor controlable durante la etapa de diseño de la encuesta, lo que influye en el efecto de diseño.

Otras consideraciones

Efecto de Diseño en Subpoblaciones como Estratos

- Para subpoblaciones que son estratos, donde el tamaño poblacional es conocido, se utiliza la fórmula

$$DEFF_h = \frac{Var(\hat{\theta}_h)}{Var^h_{MAS}(\hat{\theta}_h)}$$

- El denominador $Var^h_{MAS}(\hat{\theta}_h)$ varía según el tipo de subpoblación y se calcula considerando el tamaño poblacional del estrato, el tamaño de la muestra y la varianza poblacional de la variable de interés en el estrato.

Varianza del Estimador del Promedio Poblacional

En el caso en el que $\hat{\theta}_h$ corresponda al estimador del promedio poblacional en el estrato h, su valor es el siguiente:

$$Var^h_{MAS}(\hat{\theta}_h) = \frac{1}{n_h}\left(1 - \frac{n_h}{N_h}\right) S^2_{y_{U_h}}$$

Siendo n_h el tamaño de la muestra en el estrato h, N_h el tamaño poblacional del estrato h y $S^2_{y_{U_h}}$ es la varianza poblacional de la variable de interés restringida al subgrupo h. Por lo tanto, los efectos de diseño para las medias muestrales en un diseño aleatorio estratificado, serán por definición iguales a uno.

Efecto de Diseño en Subpoblaciones como Subgrupos Aleatorios

- Para subpoblaciones que no son estratos, sino subgrupos aleatorios, con tamaños de muestra aleatorios, se utiliza la fórmula

$$DEFF_g = \frac{Var(\hat{\theta}_g)}{Var^U_{MAS}(\hat{\theta}_g)}$$

donde $Var^U_{MAS}(\hat{\theta}_g)$ es la varianza del estimador de interés.

- El denominador $Var^U_{MAS}(\hat{\theta}_g)$ varía según el tipo de subpoblación y se calcula considerando el tamaño de la muestra y la varianza poblacional de una nueva variable que toma valores según la pertenencia al subgrupo.

Varianza del Estimador del Promedio en Subpoblaciones

En el caso en el que $\hat{\theta}_g$ corresponda al estimador del promedio poblacional en el subgrupo g, entonces su varianza estaría dada por la siguiente expresión:

$$Var^U_{MAS}(\hat{\theta}_g) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{y_{gU}}$$

En donde $S^2_{y_{gU}}$ es la varianza poblacional de una nueva variable calculada en toda la población, que toma el valor de y_k, cuando la unidad k pertence al subgrupo g, y toma el valor de cero, en cualquier otro caso.

Consideraciones sobre la Estimación de Varianza

- Aunque la estimación de la varianza del diseño de muestreo complejo ($Var(\hat{\theta}_h)$ o $Var(\hat{\theta}_g)$) puede ser la misma, el denominador cambia según si el subgrupo es un estrato o no.
- Es crucial percatarse de las opciones de los software para calcular correctamente la cifra apropiada.
- Se destaca que el efecto del diseño compara la varianza de una media o total con la varianza de un estudio del mismo tamaño utilizando un muestreo aleatorio simple sin reemplazo.
- Es esencial utilizar el cálculo correcto del efecto de diseño, especialmente si los pesos de muestreo se han re-escalado o no son recíprocos a las probabilidades de inclusión.

El efecto de diseño general

En un diseño estratificado con H estratos, la varianza del estimador para un total poblacional t_y se descompone en la suma de las varianzas de cada estrato, ponderadas por el efecto de diseño del estrato, así

$$Var\left(\widehat{t_{y,\pi}}\right) = \sum_{h=1}^{H} Var_h\left(\widehat{t_{y,\pi}}\right)$$

donde

- $Var_h\left(\widehat{t_{y,\pi}}\right) = DEFF_h \times Var_{MAS,h}\left(\widehat{t_{y,\pi}}\right)$

Por otro lado,

$$Var\left(\widehat{t_{y,\pi}}\right) = DEFF \times Var_{MAS}\left(\widehat{t_{y,\pi}}\right)$$

Fórmula del Efecto de Diseño General (DEFF)

La fórmula del efecto de diseño general ($DEFF$) es una combinación lineal de los efectos de diseño de los H estratos, donde cada estrato contribuye con un peso w_h.

$$DEFF = \frac{\sum_{h=1}^{H} DEFF_h Var_{MAS,h}\left(\widehat{t_{y,\pi}}\right)}{Var_{MAS}\left(\widehat{t_{y,\pi}}\right)} = \frac{\sum_{h=1}^{H} DEFF_h \frac{N_h^2}{n_h}(1-\frac{n_h}{N_h})S_{y,U_h}^2}{\frac{N^2}{n}(1-\frac{n}{N})S_{y,U}^2}$$

Es decir, el efecto de diseño puede ser visto como una combinación lineal de los efectos de diseño de los H estratos

$$DEFF = \sum_{h=1}^{H} DEFF_h \; w_h$$

.

Peso de Estrato w_h

- El peso w_h se define como la proporción de la suma de cuadrados dentro del estrato sobre la suma de cuadrados totales de la variable de interés.

$$w_h = \frac{\frac{N_h^2}{n_h}(1 - \frac{n_h}{N_h})S_{y,U_h}^2}{\frac{N^2}{n}(1 - \frac{n}{N})S_{y,U}^2}$$

Peso de Estrato w_h en Muestras Autoponderadas

- En el caso de muestras autoponderadas en todos los estratos, w_h es proporcional a la variabilidad interna del estrato respecto a la variabilidad total de la variable de interés.
- En este caso, se tiene que $\frac{n_h}{N_h} = \frac{n}{N}$ para todo $h = 1, \ldots, H$, y se puede ver que

$$w_h = \frac{N_h S^2_{y,U_h}}{N S^2_{y,U}} = \frac{\sum_{U_h} (y_k - \bar{y}_{U_h})^2}{\sum_U (y_k - \bar{y}_U)^2}$$

Interpretación de los Pesos w_h

- Si los estratos están bien construidos (homogeneidad intraestrato y heterogeneidad interestrato), los pesos w_h son pequeños, y el $DEFF$ general es mucho menor que los $DEFF$ de los estratos.
- Si los estratos no consideran la variabilidad de la característica de interés, los pesos w_h pueden ser mayores, y el $DEFF$ general es un promedio ponderado de los efectos de los estratos.
- Algunos pesos w_h pueden ser mayores a 1 si los estratos no están bien construidos y no cumplen la relación $\frac{n_h}{N_h} = \frac{n}{N}$.

El efecto de diseño en las encuestas de hogares de la región

- Los esquemas de estratificación y selección desigual en encuestas de hogares afectan la varianza del muestreo.
- El efecto de estratificación tiende a reducir la varianza, mientras que el efecto de selección desigual tiende a aumentarla. Ambos efectos pueden anularse entre sí.
- En encuestas de hogares, el efecto de aglomeración, especialmente el coeficiente de correlación intraclase (ρ_y), juega un papel crucial en el efecto de diseño general ($DEFF$).
- La expresión generalizada del $DEFF$ destaca la importancia del número promedio de hogares seleccionados por UPM y el coeficiente de correlación intraclase.

Variación del Efecto de Diseño en Encuestas de Hogares

- El $DEFF$ varía según la subpoblación y el parámetro que se esté estimando.
- En UN (2005, cap. 7) presenta el comportamiento de esta medida a lo largo de tres encuestas de hogares en Brasil: la *Pesquisa Nacional por Amostra de Domicílios* (PNAD), la *Pesquisa Mensal de Emprego* (PME) y la *Pesquisa de Padrões de Vida* (PPV).
- Ejemplo: En la *Pesquisa Nacional por Amostra de Domicílios* (PNAD) de Brasil, el $DEFF$ para la proporción de hogares con electricidad difiere en distintas zonas[1](nacional, áreas metropolitanas, ciudades grandes, áreas rurales), indicando mayor heterogeneidad en algunas zonas.

[1]Se estimó que el efecto de diseño para este parámetro fue de 7.92 a nivel nacional, de 1.03 en las áreas metropolitanas, de 4.43 en las ciudades grandes y de 7.27 en las áreas rurales.

Impacto del Coeficiente de Correlación Intraclase (ρ_y)

- Ejemplo: Para la encuesta PNAD, la variación en el número promedio de cuartos utilizados como dormitorios también muestra diferencias en el $DEFF$ y el ρ_y entre las zonas[2], indicando mayor homogeneidad en algunas áreas.
- Conocer el valor del $DEFF$ permite realizar simulaciones y establecer el tamaño de muestra en la planificación o rediseño de encuestas de hogares después de los censos decenales.

[2]La estimación del $DEFF$ es de 2.14 a nivel nacional, de 2.37 en las áreas metropolitanas, de 1.72 en las ciudades grandes y de 2.09 en las áreas rurales.

¡Gracias!

Email: andres.gutierrez@cepal.org

Referencias

Gambino, Jack G. 2009. «Design Effect Caveats». *The American Statistician* 63 (2): 141-46. https://doi.org/10.1198/tast.2009.0028.

Gutiérrez, Andrés, Hanwen Zhang, y Cristian Montaño. 2016. «Calculo del tamaño de muestra para la estimación de una varianza en poblaciones finitas con funciones en R». *Comunicaciones en Estadística* 9 (1): 109. https://doi.org/10.15332/s2027-3355.2016.0001.06.

Kish, Leslie. 1965. *Survey Sampling*. John Wiley; Sons.

Park, Inho. 2003. «Design Effects and Survey Planning», 8.

Särndal, Carl-Erik, Bengt Swensson, y Jan Wretman. 2003. *Model Assisted Survey Sampling*. Springer Science; Business Media.

UN. 2005. *Household surveys in developing and transition countries*. Studies en methods / United Nations, Department of Economic y Social Affairs, Statistics Division Series F.

———. 2008. *Designing household survey samples: practical guidelines*. Studies en methods / United Nations, Department of Economic y Social Affairs, Statistics Division Series F. United Nations.