Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2012-18

El Efecto del Diseño: Sesgo y Estimación de Varianza

Alberto Padilla
Banco de México

Diciembre 2012

Documento de Investigación
2012-18

Working Paper
2012-18

El Efecto del Diseño: Sesgo y Estimación de Varianza*

Alberto Padilla†
Banco de México

Resumen: El cálculo del tamaño de muestra es una parte fundamental en el proceso de planeación de una encuesta y puede hacerse de diferentes maneras, algunas de ellas requieren información que en ocasiones no se tiene o es costoso obtener. Una forma de realizar dicho cálculo hace uso del denominado estimador del efecto del diseño propuesto por Kish. Este estimador sirve también como una medida de eficiencia de un esquema de muestreo probabilístico, así como para la construcción de intervalos de confianza. A pesar del uso extendido del estimador del efecto del diseño en la práctica, se conocen poco sus propiedades estadísticas y no se cuenta con estimadores de varianza. En este trabajo se muestra que dicho estimador es sesgado, se construye una cota superior para la relación sesgo a error estándar y se propone un método para estimar la varianza. Con estos elementos es posible mejorar la precisión de los estimadores durante el proceso de planeación de una encuesta, así como en la etapa de estimación. Esto se traduce en una mejor asignación de recursos en la etapa de planeación de una encuesta.
Palabras Clave: Estimador de razón; Efecto del diseño; Varianza de varianzas; Tamaño de muestra; Coeficiente de variación; Método de remuestreo; Intervalo de confianza.

Abstract: The estimation of the sample size is a crucial part of the planning process of a survey and it can be accomplished in different ways, some of them require information not available or that may be obtained with a substantial cost. The estimation of the sample size can be done by using the design effect estimator proposed by Kish. This estimator is also used as an efficiency measure for a probability sampling plan and to build confidence intervals. Even though the design effect estimator is widely used in practice, little is known about its statistical properties and there are no variance estimators available. In this paper we show that the design effect estimator is biased, we give an expression for an upper bound to the ratio of the bias to the standard error and a method to estimate the variance. With these elements it is possible to improve the precision of the estimators during the planning and estimation stage of a survey. This also results in a better resource allocation during the planning stage of a survey.
Keywords: Ratio estimator; Design effect; Variance of variances; Sample size; Coefficient of variation; Resampling method; Confidence interval.
JEL Classification: C80, C83.

*El autor agradece a los participantes del seminario del Banco de México, a José Antonio Murillo, así como a dos revisores del Banco de México por sus comentarios y sugerencias.
† Dirección General de Investigación Económica. Email: ampadilla@banxico.org.mx.

1. Introducción

En el muestreo probabilístico, el problema básico consiste en estimar una variable de interés de una población finita, como podría ser estimar el gasto medio en alimentos por hogar en una ciudad. Si se tuviesen recursos suficientes para levantar un censo de todos los hogares de la ciudad en cuestión, se podría calcular dicho gasto y no habría necesidad de recurrir al muestreo. En este ejemplo, el gasto es lo que se conoce como una cantidad poblacional. En muchas situaciones no es factible levantar un censo, entonces se recurre a la extracción de una muestra para estimar la cantidad poblacional. La forma de seleccionar la muestra se conoce como diseño muestral y entre los principales diseños se encuentran los siguientes: el muestreo aleatorio simple, el muestreo aleatorio estratificado, el muestreo por conglomerados, el muestreo sistemático, el muestreo con probabilidades proporcionales a alguna medida de tamaño, entre otros. Para más detalle de estos y otros diseños muestrales usados en la práctica, véase Särndal et al. (1992). Por otra parte, para cada diseño muestral se tiene una expresión matemática particular del estimador de la cantidad poblacional de interés, por ejemplo, en el caso del muestreo aleatorio simple se emplea el promedio aritmético muestral como un estimador del correspondiente promedio poblacional y, como se está empleando un estimador, se construye una

fórmula para la varianza de dicho estimador. La varianza de un estimador es una cantidad poblacional, es decir, depende de cantidades que pueden calcularse al medir todos los elementos de la población de interés. Por este motivo, al trabajar con datos provenientes de una muestra, para cada diseño muestral se construye un estimador de la varianza y es el que se emplea para evaluar la precisión del estimador.

En el muestreo probabilístico, el efecto del diseño propuesto por Kish (1965) se define como el cociente de la varianza de un estimador, bajo un diseño muestral diferente del muestreo aleatorio simple, y la varianza de dicho estimador bajo muestreo aleatorio simple. El cálculo de efecto del diseño requiere del conocimiento de dos varianzas, es decir, de dos cantidades poblacionales. Esta cantidad poblacional se emplea con frecuencia, por parte de institutos de estadística y agencias gubernamentales que levantan encuestas, para el cálculo del tamaño de muestra, siempre que se tenga una estimación anticipada del efecto del diseño en cuestión, al obtener el tamaño de muestra del parámetro de interés bajo muestreo aleatorio simple y, posteriormente, multiplicando dicho tamaño por el efecto del diseño. Por otra parte, el efecto del diseño sirve como referencia para evaluar la pérdida o ganancia en eficiencia del estimador del diseño muestral diferente del muestreo aleatorio simple comparado con el muestreo aleatorio simple. Otro uso del efecto del diseño se tiene en la construcción de intervalos de

confianza en encuestas con diseños muestrales diferentes del muestreo aleatorio simple: la desviación estándar de algún estimador se obtiene al multiplicar la desviación estándar bajo muestreo aleatorio simple por la raíz cuadrada del efecto del diseño, véase Kish (1965).

Cabe mencionar que las comparaciones en el muestreo probabilístico se efectúan con el muestreo aleatorio simple, porque es el diseño más sencillo de analizar y, si el tamaño de muestra en relación con el tamaño de población, llamada fracción de muestreo, es despreciable, la varianza del estimador del promedio bajo muestreo aleatorio simple sería casi igual a la varianza de un estimador del promedio con base en una muestra aleatoria, es decir, variables aleatorias independientes e idénticamente distribuidas, y suponiendo que el segundo momento sea finito, véase Mood et al. (1985).

Es importante mencionar que en el cálculo del estimador del efecto del diseño, la varianza bajo muestreo aleatorio simple se calcula con los datos muestrales obtenidos por el diseño muestral diferente del muestreo aleatorio simple, haciendo caso omiso de las características del diseño como: estratificación, conglomeración, probabilidades desiguales de selección de elementos en muestra, etc. Esta forma de cálculo no garantiza una estimación insesgada de la varianza poblacional bajo muestreo aleatorio simple, véase Cochran (1977), y por este motivo en la literatura se han

propuesto formas de resolver este problema con Rao (1962), quien consideró estimadores insesgados bajo tres diseños muestrales, Cochran (1977), ejemplifica el método de Rao para estratificación y más recientemente Gambino (2009), construyó un estimador insesgado en términos del estimador Horvitz-Thompson (1952) del parámetro de interés de la población finita.

El estimador del efecto del diseño es una cantidad acerca de la cual se conocen poco sus propiedades y se emplea en la práctica sin que se cuestione su forma de cálculo o falta de estimadores de varianza. Con el fin de estudiar las propiedades del estimador del efecto del diseño, éste puede representarse como un estimador de razón. Un estimador de razón es un cociente de estimadores y se sabe que este tipo de estimadores son sesgados, véase Cochran (1977) o Mood et al. (1985). Esto se debe a que la esperanza de un cociente de variables aleatorias no es necesariamente igual al cociente de las esperanzas de variables aleatorias, suponiendo que las esperanzas existan y sean finitas. Una forma de verificar que el sesgo de un estimador de razón es despreciable, consiste en calcular el coeficiente de variación de la variable del denominador y si dicha cantidad es pequeña, el sesgo puede considerarse despreciable. Este resultado para estimadores de razón de un promedio o un total bajo muestreo aleatorio simple se debe a Hartley y Ross (1954), véase Cochran (1977), y es el límite superior de la

relación sesgo a error estándar del estimador de razón. En el presente artículo se construye dicho límite para el estimador del efecto del diseño, el cual a la fecha no ha sido publicado, con base en la literatura revisada.

Por otra parte, y con base en la revisión realizada a la fecha, no se ha encontrado una expresión matemática para la varianza del estimador del efecto del diseño, ni para el estimador de dicha varianza. Por esto se propone un método de remuestreo para la estimación de la varianza del estimador del efecto del diseño, el cual es una versión del bootstrap para diseños muestrales diferentes del muestreo aleatorio simple, véase Sitter (1992). Los resultados de la simulación sugieren que es una vía para estimar la varianza del estimador del efecto del diseño propuesto por Gambino (2009).

El artículo se encuentra organizado de la siguiente manera. En la sección 2 se proporcionan las definiciones, notación y las expresiones de varianzas para los diseños que se usarán en el presente artículo. El cálculo del efecto del diseño, estimación y sesgo, se encuentran en la sección 3, un límite superior de la relación sesgo a error estándar del estimador del efecto del diseño, cota, se muestra en la sección 4 y un par de ejemplos de cálculo de la cota se ilustra en la sección 5. En la sección 6 se trata el tema de la estimación de varianza del estimador del efecto del diseño, así como un ejemplo con un diseño aleatorio estratificado.

2. Definiciones y notación

Muestreo de poblaciones finitas. Algunos autores denominan a las encuestas de dos maneras: descriptivas y analíticas. La primera se refiere a la estimación de cantidades como: totales, medias, proporciones y razones, en tanto que la segunda se refiere al uso de modelos con base en los datos de una encuesta. En Chambers & Skinner (2003) se explica con mayor detalle los alcances de estos dos tipos de enfoque. Las fórmulas desarrolladas en este artículo son para estimaciones en encuestas descriptivas. Hay diversos enfoques para tratar el problema de estimación de medias o totales en el muestreo de poblaciones finitas, uno de los más usados en la práctica es la inferencia basada en el muestreo probabilístico o diseño, este último término proviene del término en inglés *design-based sampling*. Otro esquema que puede emplearse para la estimación es la inferencia basada en modelos, véase Valliant et al. (2000) y el término proviene del inglés *model-based*. En el presente artículo, todo se desarrolla con base en el muestreo probabilístico o basado en el diseño y debido a que no se trata de un artículo de divulgación del muestreo probabilístico, no se hace una explicación de esta teoría. El lector interesado en los supuestos y particularidades de este enfoque puede consultar el libro de Särndal et al. (1992).

Notación: sea *U* una población finita de *N* elementos etiquetados como *k=1,…,N*, *1<N*. Es usual representar a la población finita por sus etiquetas *k* como *U={1,2,…,k,…,N}.* Como se tratarán los diseños estratificados y conglomerados en dos etapas, a continuación se presenta la notación para estos diseños.

Muestreo aleatorio estratificado, *mae*: la variable bajo estudio se representará con y_{hi}, en donde *i* se refiere al *i*-ésimo elemento de la población en el *h*-ésimo estrato, con $i \in \{1,2,\cdots,N_h\}$. N_h y n_h denotarán el total de elementos, así como el tamaño de muestra en el *h*-ésimo estrato, $N = \sum_{h=1}^{H} N_h$ y $n = \sum_{h=1}^{H} n_h$, donde *H* es el total de estratos en la población. El promedio poblacional se denotará como $y_{st} = \sum_{h=1}^{H} W_h y_h$, donde $W_h = N_h / N$ y $y_h = \sum_{i=1}^{N_h} y_{hi} / N_h$. El estimador insesgado del promedio poblacional se calculará como $\hat{y}_{st} = \sum_{h=1}^{H} W_h \hat{y}_h$, en el que $\hat{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$. La varianza poblacional entre elementos dentro de estratos se escribirá como s_{hU}^2, la estimación muestral como $\hat{s}_h^2$ y la varianza del estimador del promedio, usando muestreo aleatorio simple, *mas,* dentro de estratos, se denotará como $v_{mae} = \sum_{h=1}^{H} W_h^2 (1 - n_h / N_h) s_{hU}^2 / n_h$. La estimación muestral se escribirá como $\hat{v}_{mae}$ con $\hat{s}_h^2$ en lugar de s_{hU}^2 en la fórmula para v_{mae}.

Muestreo por conglomerados en dos etapas: los conglomerados se denotarán como *UPM*, unidades primarias de muestreo y a los elementos dentro de conglomerados como *USM*, unidades secundarias de muestreo. *A* y *B* representarán al número de *UPM* en la población y al número de *USM* dentro de cada *UPM* respectivamente; en tanto que *a* y *b* representarán las respectivas cantidades muestrales. Se supondrá que *A*, *B*, *a* y *b* son mayores que uno y *a*<*A* y *b*<*B*. El total de elementos en población y muestra se denotarán como *N*=*AB* y *n*=*ab*, respectivamente. La variable bajo estudio se representará con y_{ij}, en donde *i* se refiere a la *UPM* y *j* a la *USM*. El promedio de y_{ij} en la i-ésima *UPM* es $y_i = \sum_{j=1}^{B} y_{ij} / B$ y el promedio poblacional $y_U = \sum_{i=1}^{A} y_i / A$. La varianza entre *UPM* se identificará con $s_1^2 = \sum_{j=1}^{A} (y_i - y_U)^2 / (A-1)$ y dentro de la i-ésima *UPM* se denotará como $s_i^2 = \sum_{j=1}^{B} (y_{ij} - y_i)^2 / (B-1)$. La varianza poblacional del muestreo por conglomerados en dos etapas (*mas*,*mas*), v_{mc2e} está dada por: $v_{mc2e}(\hat{y}) = (1 - a/A) s_1^2 / a + (1 - b/B) s_2^2 / (ab)$, donde $s_2^2 = \sum_{j=1}^{A} s_i^2 / A$. Esta expresión se encuentra en diversos textos de muestreo, por ejemplo, en la página 277 de Cochran (1977).

Observación 1: la notación *(mas,mas)* indica que tanto las *UPM*, como las *USM* en muestra, fueron seleccionadas por muestreo aleatorio simple.

Los ejemplos con los que se ilustrarán aspectos de los resultados en este artículo, se harán con los diseños arriba mencionados, estratificación y conglomeración. Los cálculos del límite superior para la relación sesgo a error estándar solo requieren que el diseño empleado use estimadores insesgados para la característica de interés. Así, el resultado no se refiere solo a diseños estratificados o conglomerados como los descritos. Estos diseños, que son bastante usados en la práctica, solo se emplean para ilustrar los resultados.

3. Ejemplo de usos del efecto del diseño, estimación y sesgo

Como se mencionó en la introducción, el efecto del diseño, efd_K, Kish (1965), se define como el cociente de la varianza de un estimador, $\hat{\theta}_{alt}$, bajo un diseño específico, $v_{alt}\left(\hat{\theta}_{alt}\right)$, y la varianza de dicho estimador bajo muestreo aleatorio simple, *mas*, $v_{mas}\left(\hat{\theta}_{mas}\right)$. De esta manera, el efecto del diseño poblacional se calcula como $efd_K\left(\hat{\theta}\right) = v_{alt}\left(\hat{\theta}_{alt}\right) / v_{mas}\left(\hat{\theta}_{mas}\right)$, con $v_{mas}\left(\hat{\theta}_{mas}\right) > 0$. Uno de los principales usos del efecto del diseño en la práctica se tiene en el cálculo del tamaño de muestra, la cual se realiza en la etapa del diseño de

una encuesta o plan de muestreo. En esta sección se mostrará con un ejemplo la forma de cálculo del tamaño de muestra empleando el efecto del diseño, también se mencionarán diversas encuestas o estudios en los que se emplea dicha cantidad y, por último, se desarrollará un ejemplo en el que se aprecia la manera en la que se estima el efecto del diseño en la práctica, el sesgo, así como las fuentes que contribuyen al sesgo y la forma de eliminar una de las fuentes o contribuciones al sesgo del estimador del efecto del diseño.

Antes de ejemplificar el uso del efecto del diseño para el cálculo del tamaño de muestra, en la siguiente observación se describe someramente la estrategia del cálculo del tamaño de muestra en el muestreo probabilístico.

Observación 2: en el muestreo probabilístico, el tamaño de muestra se calcula en general encontrando una ecuación que relacione el tamaño de muestra con el error de estimación esperado, absoluto o relativo. Esto se hace al suponer que el estimador del promedio o total de interés de la población finita, se distribuye como una normal con media igual al total o promedio de la población finita, y varianza igual a la del estimador del total o promedio bajo el esquema muestral considerado; es decir, muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados, etc. Así, se resuelve la siguiente ecuación para el tamaño de

muestra, $e_a = z_\alpha \sqrt{v(\hat{y})}$. En esta ecuación, e_a se refiere al error de estimación absoluto que se desea, en tanto que z_α se refiere al valor asentado en las tablas estadísticas de la distribución normal estándar para una confianza prefijada

En la ecuación, $e_a = z_\alpha \sqrt{v(\hat{y})}$, la varianza puede tener una expresión sencilla, en términos de solución para el tamaño de muestra, como en el caso del *mas* o más compleja como la expresión para el diseño *(mas,mas)* o mae. Una manera de encontrar el tamaño de muestra, *n*, para un diseño distinto del *mas* consiste en encontrar primero el valor de *n* bajo *mas* y después multiplicarlo por el efd_K, siempre que se tenga este valor. Si se desea conocer más acerca del cálculo del tamaño de muestra en el muestreo probabilístico, véase Särndal et al. (1992) ó Cochran (1977).

3.1 Ejemplo de uso del efecto del diseño en la práctica

Ejemplo 1: uso del efecto del diseño para el cálculo del tamaño de muestra. El INEGI hace uso del *efd* para el cálculo del tamaño de muestra de diferentes encuestas, como es el caso para la Encuesta Nacional de Ingreso Gasto de los Hogares 2008.

El INEGI empleó como variable de referencia el promedio del ingreso corriente total por hogar y la expresión utilizada fue:

$$n = \frac{z_{\alpha}^{2} s^{2} \mathrm{DEFF}}{r^{2} \bar{X}^{2} (1 - \mathrm{tnr}) \mathrm{PHV}} \qquad (1)$$

En la ecuación anterior:

- n = tamaño de la muestra.
- z_{α} = valor asentado en las tablas estadísticas de la distribución normal estándar para una confianza prefijada.
- s^2 = estimación de la varianza poblacional de la variable de interés.
- $\bar{X}$ = estimación del promedio de la variable de interés.
- DEFF = siglas en inglés del efecto de diseño, definido como el cociente de la varianza en la estimación del diseño utilizado, entre la varianza obtenida considerando un muestreo aleatorio simple para un mismo tamaño de muestra.
- r = error relativo máximo aceptable.
- tnr = tasa de no respuesta máxima esperada.
- PHV = promedio de hogares por vivienda.

- Fijando un nivel de confianza de 90%, un efecto de diseño de 3.3, una varianza poblacional de 1’767,586,177.77 un error relativo máximo aceptable de 4%, un promedio de ingreso corriente total por hogar de 34,127, una tasa de no respuesta máxima esperada de 15% y un promedio de hogares por vivienda de 1.02, se determinó una muestra a nivel nacional de 9,711 viviendas, la cual se ajustó a 10,000. Los valores del efecto del diseño, varianza poblacional e ingreso corriente total por hogar fueron obtenidos de la ENIGH-2006. Con el objeto de satisfacer los requerimientos adicionales de las instituciones externas que aportaron recursos para el levantamiento, la muestra final de la ENIGH-2008 fue de 35,146 viviendas.

La fórmula para el cálculo del tamaño de muestra que presenta el INEGI en la metodología de esta encuesta puede escribirse como:

$$n = \frac{z_\alpha^2 s^2 \mathrm{DEFF}}{r^2 \bar{X}^2 (1-\mathrm{tnr})\mathrm{PHV}} = \frac{z_\alpha^2 s^2}{r^2 \bar{X}^2} \mathrm{DEFF} \frac{1}{(1-\mathrm{tnr})\mathrm{PHV}} = n_{mas} \mathrm{DEFF} \frac{1}{(1-\mathrm{tnr})\mathrm{PHV}} \qquad (2)$$

En este caso, n_{mas} se refiere al tamaño de muestra obtenido por muestreo aleatorio simple, el cual es fácil de calcular, véase Cochran (1977). Nótese como el efecto del diseño aumenta, disminuye o no modifica el tamaño de muestra dependiendo del valor de la varianza del diseño distinto del *mas* con el cual se extraerá la muestra.

3.2 Ejemplo de estimación y cálculo del efecto del diseño en la práctica

Ejemplo 2: estimación sesgada de $v_{mas}(\hat{\theta})$*, cálculo del* efd_K *y sesgo*. En la introducción se mencionó que en el estimador del efd_K, la varianza bajo *mas* se calcula con los datos muestrales obtenidos por el diseño específico y que esta forma de cálculo no garantizaba una estimación insesgada de la varianza poblacional bajo *mas*. Un ejemplo con una población pequeña ilustrará este aspecto.

Considérese una población con $H = 2$ estratos y 8 elementos. Supóngase que de cada estrato se extrae una *mas* de 2 elementos y la característica de interés por estimar es el promedio poblacional.

Tabla 1

Ejemplo 2, valores poblacionales relevantes

Estrato	**Valores** y_{hi}	$\bar{y}_h$	s^2_{hU}
1	{2,3,4.5,2.5,3.4}	3.08	0.91
2	{11,14,18}	14.33	12.33
Población		**7.30**	**37.96**

Estimación sesgada de $v_{mas}(\hat{y})$. Bajo *mae*, el total de muestras aleatorias estratificadas es de 30, en tanto que el total de *mas* es 70. Al generar todas las posibles *mae* y calcular para cada una de ellas la varianza del *mas*, $\hat{v}_{sesg,mas}(\hat{y}_i) = (1-4/8)\sum_{j=1}^{4}(y_{hj}-\hat{y}_i)^2 \big/ (4-1)$, con $\hat{y}_i = \sum_{j=1}^{4} y_{hj} \big/ 4$ e $i = 1,\ldots,30$, se tiene que: $\sum_{i=1}^{30} v_{sesg,mas}(\hat{y}_i) \big/ 30 = 5.93$. Para este caso, la varianza poblacional bajo *mas* $v_{mas}(\hat{y}) = 4.75$, por lo que la estimación de la varianza del *mas* con este método tiene un sesgo de 5.93 - 4.75 = 1.18. El subíndice del estimador de varianza "*sesg, mas*", se refiere a la estimación sesgada de la varianza del *mas*.

Cálculo del efd_K. En este ejemplo $v_{mae}(\hat{y}_{st}) = 0.395$ y $efd_K(\hat{y}) = v_{mae}(\hat{y}_{st}) / v_{mas}(\hat{y}) = 0.083$. Nótese que esta es una cantidad poblacional y la varianza bajo *mas* se calculó usando la fórmula $v_{mas} = (1-n/N)\, s_U^2 / n$, con $s_U^2 = \sum_{i=1}^{N}(y_i - y_U)^2 / (N-1)$ y $y_U = \sum_{i=1}^{N} y_i / N$.

Estimación del efd_K. Usando la definición de Kish (1965) y como se mencionó en la introducción, la estimación del efd_K para el ejemplo se efectúa con la siguiente fórmula: $\hat{efd}_K(\hat{y}_i) = \hat{v}_{mae}(\hat{y}_{st,i}) / \hat{v}_{sesg,mas}(\hat{y}_i)$, donde $i = 1,\ldots,30$; es decir, una de las posibles *mae*. Al generar las 30 posibles *mae* y calcular para

cada una de ellas el estimador $\hat{efd}_K(\hat{y}_i)$, se tiene que $\sum_{i=1}^{30} \hat{efd}_K(\hat{y}_i)/30 = 0.067 \neq efd_K(\hat{y}) = 0.083$.

Observación 3: la diferencia entre los promedios de las estimaciones del estimador efd_K y el efd_K poblacional era de esperarse, debido a que el estimador del efd_K poblacional es un estimador de razón y es sesgado. Sin embargo, los estimadores de la varianza bajo *mas*, $v_{sesg,\ mas}$, empleados en el estimador del efd_K son sesgados, como se ha visto en el presente ejemplo, por lo cual contribuyen con un efecto más al sesgo. Por este motivo, es necesario corregir el estimador del denominador con la fórmula propuesta por Gambino (2009) y que se muestra a continuación:

$$\hat{v}_{ins,\ mas}(\hat{y}) = (1-\frac{n}{N})\frac{1}{n(N-1)}(\hat{y}_{cua} - \frac{\hat{y}^2 - \hat{v}_{alt}(\hat{y})}{N}) \qquad (3)$$

En esta fórmula, $\hat{y}_{cua,}$ $\hat{y}^2$ y $\hat{v}_{alt}(\hat{y})$son estimadores de las siguientes cantidades poblacionales: suma de cuadrados, total al cuadrado y varianza del total bajo el diseño diferente del *mas*. Con la fórmula (3), se tiene un estimador insesgado de la varianza poblacional bajo *mas*. Este estimador ya pertenece al tipo de los estimadores de razón descritos en el capítulo 6 de Cochran (1977). Un aspecto importante de estos estimadores, es que se tienen estimadores insesgados y positivos tanto del numerador, como del

denominador y la importancia del sesgo se encuentra relacionada con el coeficiente de variación de los estimadores del denominador.

Estimación del efd_K con estimación insesgada del denominador. Usando la fórmula 3 de Gambino (2009) para la varianza del promedio bajo *mas*, se calcularon las estimaciones de $v_{mas}(\hat{y})$ y efd_K, obteniéndose: $\sum_{i=1}^{30} \hat{efd}_G(\hat{y}_i)/30 = 0.084 \neq efd(\hat{y}) = 0.083$, donde $\hat{efd}_G$ se refiere al estimador del efecto del diseño usando la corrección de la varianza del denominador.

Observación 4: el estimador del efecto del diseño continúa siendo sesgado, pero el sesgo solo se debe a que se está usando un estimador de razón.

4. Cotas para la relación sesgo a error estándar del efd_K

A continuación se enuncia uno de los resultados principales del presente artículo. Primero, se encuentra la expresión general para el sesgo del estimador del *efd* y después se muestra tanto el sesgo del estimador efd_G como un límite superior de la relación sesgo a error estándar para el efd_G. Sean $\hat{y}_{alt}$ un estimador insesgado del promedio poblacional, con $v_{alt}(\hat{y}_{alt}) > 0$ y $\hat{ecm}$ un estimador de varianza tal que $E(\hat{ecm}) = v + sesgo$ si el

estimador es sesgado y $E(ec\hat{m}) = v$ si el estimador es insesgado. Aquí *v* se refiere a la varianza del estimador.

Teorema: para un diseño muestral, *alt*, distinto del *mas*, y estimadores de los errores cuadráticos medios, $ec\hat{m}_{alt}$ y $ec\hat{m}_{mas}$ con $ec\hat{m}_{mas} > 0$ tenemos:

$$E(ef\hat{d}_K) = E\left(\frac{ec\hat{m}_{alt}}{ec\hat{m}_{mas}}\right) = \frac{v_{alt} + sesgo_{alt}}{v_{mas} + sesgo_{mas}} - \frac{\text{cov}(ec\hat{m}_{mas}, ef\hat{d}_K)}{v_{mas} + sesgo_{mas}} \quad (4)$$

Corolario 1: para un diseño muestral *alt*, distinto del *mas*, con estimador $\hat{y}_{alt}$ del promedio de la población, estimador insesgado de varianza para el diseño *alt* y estimación insesgada de la varianza poblacional bajo *mas*, el sesgo del $ef\hat{d}_G$ se encuentra dado por: $-\text{cov}(ef\hat{d}_G, \hat{v}_{ins,mas}) / v_{mas}$.

Observación 5: en el corolario 1, $\hat{v}_{mas}$ corresponde a la fórmula de Gambino (2009) para la estimación insesgada de la varianza poblacional bajo *mas* y el $ef\hat{d}_K$ se reduce al $ef\hat{d}_G$.

Corolario 2: bajo las condiciones del teorema y corolario 1, un límite superior de la relación sesgo a error estándar del $ef\hat{d}_G$, se encuentra dada por $cv(\hat{v}_{ins,mas})$, es decir, el coeficiente de variación de las estimaciones insesgadas de la varianza poblacional bajo *mas*.

Corolario 3: bajo las condiciones del teorema, con $sesgo_{alt}=0$, el estimador propuesto por Gambino, $\hat{efd}_G$, tiene un sesgo relativo más pequeño que el estimador propuesto por Kish, $\hat{efd}_K$, siempre que:

$$1-\frac{sesgo(\hat{efd}_G)}{efd}<\left[1-\frac{sesgo(\hat{efd}_K)}{efd}\right]\left[\frac{1}{1+sesgo_{mas}/v_{mas}}\right] \qquad (5)$$

Las demostraciones del teorema y los corolarios se encuentran en el apéndice.

Observación 6: nótese que en los corolarios se está trabajando con la expresión para el $\hat{efd}_G$, la cual es diferente a $\hat{efd}_K$ que es la expresión empleada en la práctica y en libros de texto. El estimador del efd_K usado en la práctica no corrige el sesgo del estimador de la varianza poblacional bajo *mas*.

Es necesario mencionar que, salvo el caso del *mas,* véase Thompson (1997), no hay expresiones cerradas para la varianza de varianzas estimadas para diferentes diseños muestrales en poblaciones finitas. Por lo anterior, se trabajará con simulaciones de los dos diseños mencionados en la sección 2.

Como se comentó en la introducción, la magnitud del sesgo en el contexto de los estimadores de razón puede ser evaluada por medio del coeficiente de variación del estimador que se encuentra en el denominador del estimador de razón, si se tiene un estimador insesgado en el numerador. El sesgo será despreciable si el coeficiente de variación es pequeño, es decir, los valores del estimador que se encuentran en el denominador casi no cambian de muestra a muestra. En el estimador del efecto del diseño propuesto por Gambino, el sesgo será despreciable si el coeficiente de variación de los estimadores insesgados de la varianza bajo *mas* es pequeño, como es el caso del ejemplo del muestreo aleatorio en la siguiente sección. También puede tenerse un sesgo considerable del estimador del efecto del diseño, efd_G, si el coeficiente de variación de los estimadores insesgados de varianza no es pequeño, como se verá en la siguiente sección para algún casos del muestreo por conglomerados en dos etapas.

Por otra parte, todavía no se cuenta con un refinamiento del límite superior de la relación sesgo a error estándar para el efd_G y es un tema que se seguirá investigando. Los refinamientos de cotas o desigualdades requieren, en ocasiones, de supuestos o conocimiento adicional de la variable bajo estudio. Por ejemplo, en relación con desigualdades, la desigualdad de Chebyshev, véase Mood et al. (1985), requiere de pocos supuestos; sin embargo, resulta conservadora para su uso en la práctica, no así en su empleo en

demostraciones de algunos resultados en probabilidad y estadística. Si se imponen condiciones adicionales a la desigualdad de Chebyshev, como unimodalidad de la distribución de probabilidades, se refina dicha desigualdad obteniéndose lo que se conoce como desigualdad de Vysochanskii-Petunin, véase Hernández (2003).

5. Ejemplos de cálculo del efd y cota para la relación sesgo a error estándar

De acuerdo con lo mencionado en el último párrafo de la sección anterior, no se cuenta, en general, con una expresión cerrada para la varianza de varianzas estimadas, por lo cual no es posible tener una expresión exacta para la relación sesgo a error estándar. Por este motivo, es necesario simular extracciones de muestras con algún diseño, distinto del muestreo aleatorio simple, para calcular la cota para la relación sesgo a error estándar del efecto del diseño. Esto se ilustra a continuación con un par de ejemplos: en uno de ellos se simula la extracción de muestras de un diseño aleatorio estratificado, que típicamente tiene una varianza menor que la del muestreo aleatorio simple, véase el capítulo 5 de Cochran (1977), en tanto que en el otro

ejemplo se simula la extracción de muestras de un muestreo por conglomerados en dos etapas con tamaños iguales, que típicamente tiene una varianza mayor que la muestreo aleatorio simple, véase el capítulo 10 de Cochran (1977).

Ejemplo 3: población estratificada. Se simuló una población pequeña con $H = 3$ estratos y 57 elementos, tomando como base el ejemplo de la página 137 de Cochran (1977). Para cada estrato se simularon variables aleatorias uniformes con medias $\{2.20, 1.64, 5.00\}$ y varianzas entre elementos $\{1.82, 0.07, 1.00\}$ para los estratos 1, 2 y 3, respectivamente. En la tabla 2 se muestran los datos relativos a la población estratificada.

Tabla 2

Ejemplo 3, valores poblacionales relevantes

Estrato	N_h	n_h	W_h	y_h	s^2_{hU}
1	13	9	0.22	2.33	1.62
2	18	7	0.32	1.61	0.08
3	26	6	0.46	5.04	1.18
Población	**57**	**22**			**3.44**

Las varianzas poblacionales bajo *mas*, y *mae*, así como el efecto del diseño poblacional, efd_K, tienen los valores mostrados a continuación:

Tabla 3

Varianzas bajo *mas*, *mae* y efecto del diseño

Cantidad poblacional	Valor
v_{mas}	0.096
v_{mae}	0.035
efd_K	0.364

Se simuló la extracción de 5,000 muestras de tamaño 22 con el diseño de la tabla 2, cada muestra se extrajo por *mas*, y para cada muestra se calcularon los siguientes estimadores:

- varianza del *mae*, $\hat{v}_{mae}$,
- estimador sesgado de la varianza del *mas*, usando la definición de Kish, $\hat{v}_{sesg,mas}$,
- estimador insesgado de la varianza del *mas*, usando la corrección de Gambino, $\hat{v}_{ins,mas}$,
- estimador del efecto del diseño usando la fórmula de Kish, $\hat{efd}_K$,
- estimador del efecto del diseño usando la corrección de Gambino, $\hat{efd}_G$.

Los resultados del promedio de las 5,000 extracciones para cada uno de los estimadores, se encuentran en la siguiente tabla.

Tabla 4

Resultado de las simulaciones

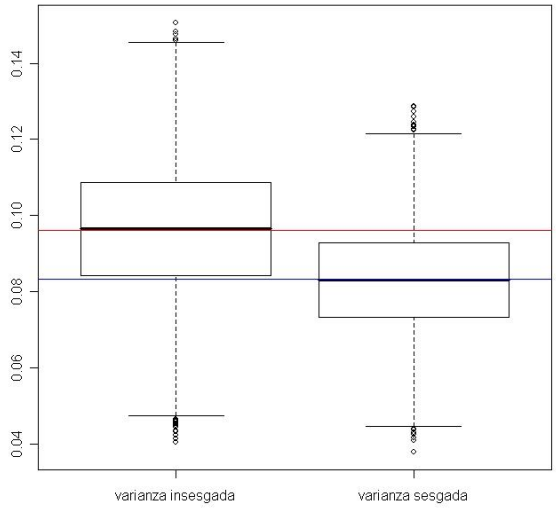
Estimador	Promedio de estimadores (A)	Valor poblacional (B)	Diferencia (%) = (A-B)/B
$\hat{v}_{ins,mas}$	0.0964	0.096	---
$\hat{v}_{sesg,mas}$	0.0832	0.096	-13.4%
$\hat{v}_{mae}$	0.0352	0.035	---
$\hat{efd}_K$	0.4318	0.364	18.7%
$\hat{efd}_G$	0.3724	0.364	2.4%

Aplicando el corolario 3 del teorema de la sección 4 a los datos de la tabla 4, se desprende que no es conveniente emplear el estimador efd_K por el tamaño del sesgo, casi 19%, en tanto que el $\hat{efd}_G$ presenta un sesgo pequeño. La cota para el sesgo relativo, $cv(\hat{v}_{mas})$, del corolario 2 de la sección 4, calculada con las $\hat{v}_{ins,mas}$ de la simulación, tuvo un valor de 18.4% lo cual sugiere que no hay un problema de inestabilidad en la estimación de la varianza del *mas*.

La cota para el sesgo relativo, $cv(\hat{v}_{mas})$, calculada con las $\hat{v}_{sesg,mas}$ de la simulación tuvo un valor de 17.1%; empero, el sesgo de la estimación de la varianza poblacional del *mas* fue de -13.4%. Lo anterior se aprecia en el siguiente diagrama de caja y brazos.

Gráfica 1

Estimadores insesgados y sesgados de la varianza bajo *mas*



En esta gráfica, la línea roja representa el valor de $v_{mas}(\hat{\theta})$ y la azul el promedio de las 5,000 simulaciones usando el estimador sesgado de la varianza poblacional bajo *mas*.

Ejemplo 4: población conglomerada con selección por mas en dos etapas. Se tiene una población con *A=8 UPM* y cada uno de los conglomerados tiene *B=8 USM* o elementos. El tamaño de muestra es de *a=3 UPM* y *b=4 USM*, con lo que se tiene un tamaño de muestra total, *n=ab=12* elementos. Los valores dentro de cada conglomerado se simularon con variables aleatorias uniformes, usando los mínimos, *mín*, y máximos, *máx*, de la cuarta columna de la siguiente tabla.

Tabla 5

Valores poblacionales relevante

UPM	y_i	s_i^2	*mín* y *máx*
1	0.332	0.0058	0.2 y 0.5
2	0.444	0.0009	0.4 y 0.5
3	1.237	0.0094	1.1 y 1.4
4	0.919	0.0037	0.8 y 1.0
5	0.223	0.0064	0.1 y 0.35
6	0.610	0.0030	0.5 y 0.7
7	0.970	0.0044	0.9 y 1.1
8	0.461	0.0077	0.3 y 0.6
Población	**0.650**	**0.1166**	

La correlación intraclase para esta población es de 0.96 y se calculó usando el resultado de la página 291 de Cochran (1977).

Usando una notación similar a la del ejemplo 3 y con los datos de la tabla 5, las varianzas poblacionales bajo *mas* y *mc2e*, así como el efecto del diseño poblacional, efd_K, tienen los valores mostrados a continuación:

Tabla 6

Varianzas bajo *mas*, *mc2e* y efecto del diseño

Cantidad poblacional	**Valor**
v_{mas}	0.0079
v_{mc2e}	0.0265
efd_K	3.3528

Se simuló la extracción de 3,500 muestras de tamaño 12 con el diseño (*mas, mas*) de la población generada con los datos de la tabla 5. Para cada muestra se calcularon los siguientes estimadores:

- varianza del *mc2e*, $\hat{v}_{mc2e}$,
- estimador sesgado de la varianza del *mas*, usando la definición de Kish, $\hat{v}_{sesg,mas}$,
- estimador insesgado de la varianza del *mas*, usando la corrección de Gambino, $\hat{v}_{ins,mas}$,
- estimador del efecto del diseño usando la fórmula de Kish, $\hat{efd}_K$,
- estimador del efecto del diseño usando la corrección de Gambino, $\hat{efd}_G$.

La extracción de muestras por conglomerados en dos etapas usando *mas* se programó en *R* y el cálculo de los estimadores del promedio y la varianza bajo *mae* se efectuó con la librería *survey* de *R*. Los resultados del promedio de las 3,500 extracciones para cada uno de los estimadores, se encuentran en la siguiente tabla.

Tabla 7

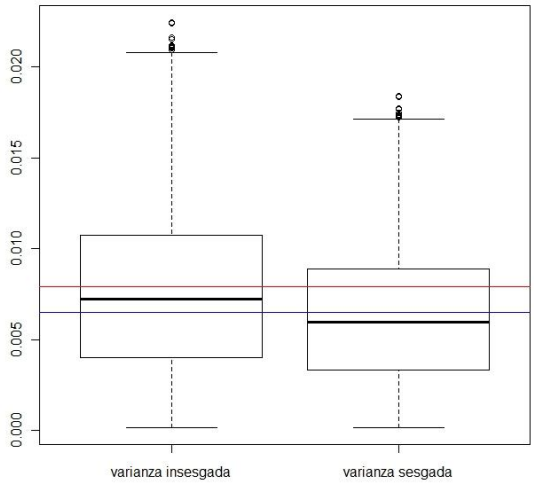
Resultado de las simulaciones

Estimador	Promedio de estimadores (A)	Valor poblacional (B)	Diferencia (%) = (A-B)/B
$\hat{v}_{ins,mas}$	0.0080	0.0079	---
$\hat{v}_{sesg,mas}$	0.0066	0.0079	-16.13%
$\hat{v}_{mc2e}$	0.0268	0.0265	---
$\hat{efd}_K$	3.912	3.3528	16.70%
$\hat{efd}_G$	3.253	3.3528	-3.00%

Aplicando el corolario 3 del teorema de la sección 4 a los datos de la tabla 7, se desprende que no es conveniente emplear el estimador del efecto del diseño propuesto por Kish. La cota para el sesgo relativo, $cv(\hat{v}_{mas})$, del corolario 2 de la sección 4 calculada con las $\hat{v}_{ins,mas}$ de la simulación, tuvo un valor de 63.50%, lo cual sugiere que hay inestabilidad en la estimación de la varianza del *mas*.

Gráfica 2

Estimadores insesgados y sesgados de la varianza bajo *mas*



En esta gráfica, la línea roja representa el valor de $v_{mas}(\hat{\theta})$ y la azul el promedio de las 3,500 simulaciones usando el estimador sesgado de la varianza poblacional bajo *mas*.

Extensión del ejemplo 4: con el fin de analizar el comportamiento del límite superior de la relación sesgo del efecto del diseño a error estándar, se generaron configuraciones que tienen, tanto el mismo promedio como la misma varianza poblacional entre elementos. Dichas configuraciones fueron generadas por medio de permutaciones de las *USM* entre *UPM* de la población de este ejemplo. Así, la varianza bajo *mas* no se altera al efectuar

estas permutaciones de las *USM*; sin embargo, el coeficiente de correlación intraclase y la varianza bajo *(mas, mas)* sí cambian y, por lo tanto, el valor del efecto del diseño. Para cada permutación de *USM* entre *UPM* se calculó el valor del coeficiente de correlación intraclase, la varianza bajo *(mas,mas)*, el efecto del diseño poblacional, así como el límite superior de la relación sesgo a error estándar. Para calcular este último valor, se extrajeron 3,500 muestras como en el ejemplo 3 para cada permutación, obteniéndose los resultados siguientes:

Tabla 8

Límite superior de la relación sesgo del efecto del diseño a error estándar según el valor de la correlación intraclase roh

rho	efd_K	Límite superior sesgo a error estándar	efd≈[1+rho(b-1)]
-0.14	0.70	0.25	0.58
-0.05	0.92	0.34	0.85
0.01	1.07	0.34	1.03
0.14	1.38	0.36	1.42
0.26	1.66	0.37	1.78
0.38	1.97	0.43	2.15
0.50	2.25	0.45	2.50
0.63	2.56	0.54	2.88
0.76	2.87	0.61	3.27
0.86	3.12	0.62	3.57
0.96	3.35	0.64	3.87

La fórmula mostrada en la cuarta columna de la tabla 8, es una aproximación que se usa en la práctica, al emplear un diseño *(mas,mas)*. *[1+rho(b-1)]* es una buena aproximación al efd_K (valor poblacional) siempre que los valores *(A-1)/A* y *(N-1)/N* sean casi uno, véase Kish (1965), capítulo 5, lo cual no sucede en el presente ejemplo por tratarse de una población pequeña. En cuanto al límite superior de la relación sesgo del efecto del diseño a error estándar, se aprecia en la tabla 8, el cual crece al hacerlo *rho*. Esto significa que conforme haya más homogeneidad en la población, valores de *rho>0*, el sesgo del estimador del efecto del diseño es grande comparado con el error estándar del efecto del diseño. Debido a que el efecto del diseño es mayor que cero, *[1+rho(b-1)]>0*, por lo que *rho>-1/(b-1)*. De la tabla se observa que en el caso de que se tenga heterogeneidad para la variable de interés en la población, *rho<0*, el efd_K es menor que uno, con lo que el muestreo por conglomerados en dos etapas es mejor que el *mas*. Para valores de *rho* alrededor del cero, el efd_K con lo que en términos de varianza el muestreo por conglomerados que nos ocupa es equivalente al *mas*. Por último, cuando se presenta homogeneidad en la población conglomerada, *rho>0*, el muestreo por conglomerados presenta una varianza que puede exceder por mucho a la del *mas*, como el cado de *rho=0.86* con un efd_K *=3.12*.

6. Estimación de la varianza del efecto del diseño

Como se mencionó en la introducción, no se encontraron referencias de cálculo de la varianza del efd_K. La manera en que se procede en el muestreo probabilístico para construir un estimador de varianza es la siguiente. Primero se encuentra una expresión para la varianza del estimador del efd_K y posteriormente, se construye un estimador de dicha varianza. No se cuenta con dichas expresiones, pero es posible estimar la varianza usando un método de remuestreo. En este caso, se hicieron pruebas con un tipo de bootstrap desarrollado para muestras provenientes de diseños complejos. En Chaudhuri & Stenger (2005) se encuentran ocho tipos de bootstrap para muestras de poblaciones finitas, entre ellos, algunos métodos propuestos por Sitter (1992), así como el empleado en el presente artículo.

Sitter (1992) propuso estimadores bootstrap para los siguientes diseños muestrales.

a) Muestreo aleatorio estratificado usando muestreo aleatorio simple en cada estrato.

b) Muestreo por conglomerados en dos etapas con tamaños iguales o desiguales.

c) Método de Rao-Hartley-Cochran, véase Rao et al. (1962), para el muestreo con probabilidad proporcional al tamaño.

En dicho artículo, Sitter propone tres métodos para la construcción de intervalos de confianza. Uno de estos métodos es el de percentiles, utilizado en el presente trabajo por su sencillez. Otro de los métodos, el cual es intensivo en términos de cálculos, se aplica un doble bootstrap a cada muestra. El tercero de los métodos, usado por el autor en su artículo, se trata de un estimador jackknife de la varianza aplicado a la muestra y la muestra replicada. Es importante mencionar que estos dos últimos métodos merecen un estudio aparte, en el que se comparen con el método de percentiles y se exploren experimentalmente algunas cuestiones como el número de réplicas necesarias para acercarse a la cobertura nominal de las estimaciones del efecto del diseño. Por otra parte, está documentado que el jackknife presenta complicaciones al aplicarse a poblaciones estratificadas, véase Särndal et al. (1992).

En este artículo se usó el bootstrap extendido de Sitter (1992), véase Chaudhuri & Stenger (2005), para muestras aleatorias estratificadas, el cual se describe a continuación.

Ignorando la parte entera de $n_h' = n_h - (1 - f_h)$ y $k_h = \frac{N_h}{n_h}(1 - \frac{1 - f_h}{n_h})$, con $f_h = n_h / N_h$, los siguientes son los pasos del método:

a) Replicar $(y_{h1}, \ldots, y_{hn_h})$ k_h veces de manera separada e independiente, *h=1,…,H*, para crear *H* seudo-estratos diferentes.

b) Extráigase una *mas* de tamaño n_h' del *h*-ésimo seudo-estrato y repita esto de manera independiente para cada *h=1,…,H*, generando así las observaciones bootstrap muestrales, $s^* = \{(y_{h1}^*, \ldots, y_{hn_h'}^*),\ h = 1, \ldots, H\}$y sea $\hat{\theta}^* = \hat{\theta}(s^*)$.

c) Repita el paso (b) un gran número de veces, digamos B, y calcule para cada b-ésima muestra bootstrap $\hat{\theta}_b^*,\ b = 1, \ldots, B$. Una vez que se tienen las B estimadores $\hat{\theta}_b^*$, se calculan las siguientes cantidades:

$$\begin{aligned} \hat{\theta}_B^* &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* \quad \text{y} \\ \hat{\mathrm{V}}_{\mathrm{BWO}} &= \frac{1}{B-1}\sum_{b=1}^{B} (\hat{\theta}_b^* - \hat{\theta}_B^*)^2 \end{aligned} \tag{5}$$

Con esto se tienen los estimadores del promedio o total de cada muestra y de la varianza. El estimador de varianza BWO también puede emplearse

como estimador de varianza para el estimador de la muestra original. Las siglas *BWO* se refieren al bootstrap para el muestreo sin reemplazo.

En este artículo, $\hat{\theta}_b^*$ puede ser un estimador del promedio estratificado, de razón, como el estimador del efecto del diseño, o de varianza, como la varianza bajo *mae* o *mas.*

Se usó el bootstrap extendido de Sitter, ya que es sencillo de implementar, comparado con los otros métodos mencionados por Chaudhuri & Stenger (2005). Sin embargo, no puede concluirse que se tengan mejores resultados que con cualesquiera de los otros bootstrap para la estimación de la varianza del efecto del diseño. De hecho, es un tema que será estudiado en el futuro.

Debido a que no se cuenta con una expresión para el estimador de la varianza del efd_G, se ilustrará a continuación la mecánica de cálculo del estimador bootstrap de la varianza con un diseño estratificado en una población pequeña.

Ejemplo 5: población estratificada de elementos con selección por mas. Se extendió la población del ejemplo 2 a cinco estratos, con un total de 120 elementos y un tamaño de muestra total de 40. La información de la población se encuentra resumida en la tabla que se muestra a continuación.

Tabla 9

Valores poblacionales relevantes

Estrato	N_h	n_h	W_h	y_h	s_{hU}^2
1	13	9	0.11	2.33	1.62
2	18	7	0.15	1.61	0.08
3	26	6	0.22	5.04	1.18
4	26	10	0.22	7.01	3.06
5	37	8	0.31	9.86	0.31
Población	**120**				**3.44**

En la siguiente tabla se encuentran los valores poblacionales para el cálculo del efecto del diseño.

Tabla 10

Varianzas bajo *mas*, *mae* y efecto del diseño

Cantidad poblacional	**Valor**
v_{mas}	0.1762
v_{mae}	0.0196
efd_K	0.1114

Los pasos que se siguieron en la simulación del bootstrap de Sitter, arriba mencionado son:

a) De la población estratificada, se simuló la extracción de 5,000 muestras de tamaño 40 con *mae.*
b) Para cada muestra, *mae*, se simularon B=2,000 muestras bootstrap con el método de Sitter ya mencionado. Se usó este valor con base en la recomendación de Stuart et al. (1999) para estimación de varianzas en el caso de variables independientes.
c) El estimador del efd_K se calculó tanto con el estimador efd_G usado en el ejemplo 2, como con el bootstrap. La varianza del estimador efd_G, para cada *mae*, se obtiene de $\hat{V}_{BWO}$, véase Chaudhuri & Stenger (2005) o Sitter (1992).

Se hizo un programa en *R*, veáse R Development Core Team (2010), y la extracción de las muestras *mae* se efectuó con el paquete *pps* de *R*, véase Gambino (2005). El método de Sitter se programó en *R* y los intervalos para la cobertura del efecto del diseño poblacional al 95%, se obtuvieron con el histograma bootstrap de los *efd* estimados. Con este método, se encuentran los percentiles correspondientes al 2.5% y al 97.5% de cada histograma y se determina si el efecto del diseño poblacional se encuentra entre el percentil 2.5% y el 97.5%; se cuenta el número de veces que esto ocurre y se divide

entre el total de réplicas, *B* réplicas, obteniéndose así la cobertura del estimador efd_G. Los resultados se muestran a continuación.

Tabla 11

Resultados de la simulación para la estimación de la varianza del efd_K

Estimador	**Promedio de estimadores (A)**	**Valor poblacional (B)**	**Diferencia (%) = (A-B)/B**
promedio *mae BWO*	6.144	6.143	0.02
V*mae BWO*	0.018	0.020	-7.60
V*mas BWO*	0.176	0.176	-0.10
efd *BWO*	0.103	0.111	-7.80
efd_G *mae*	**0.112**	0.111	0.30
desv efd BWO	**0.021**		
desv efd G *mae*	0.018		
Cobertura efd_G mae=	89.5%		

En la tabla 11, los estimadores que terminan con las siglas BWO fueron obtenidos con base en las simulaciones del bootstrap. Obsérvese que los estimadores bootstrap del promedio poblacional, promedio *mae BWO*, y la varianza bajo *mas*, V*mas BWO*, tuvieron una diferencia menor al 1% comparado con el valor poblacional. Por otra parte, les estimadores bootstrap de la varianza bajo *mae*, V*mae BWO*, y del efd_K, efd *BWO*, subestimaron el correspondiente valor poblacional casi en un 8%. A

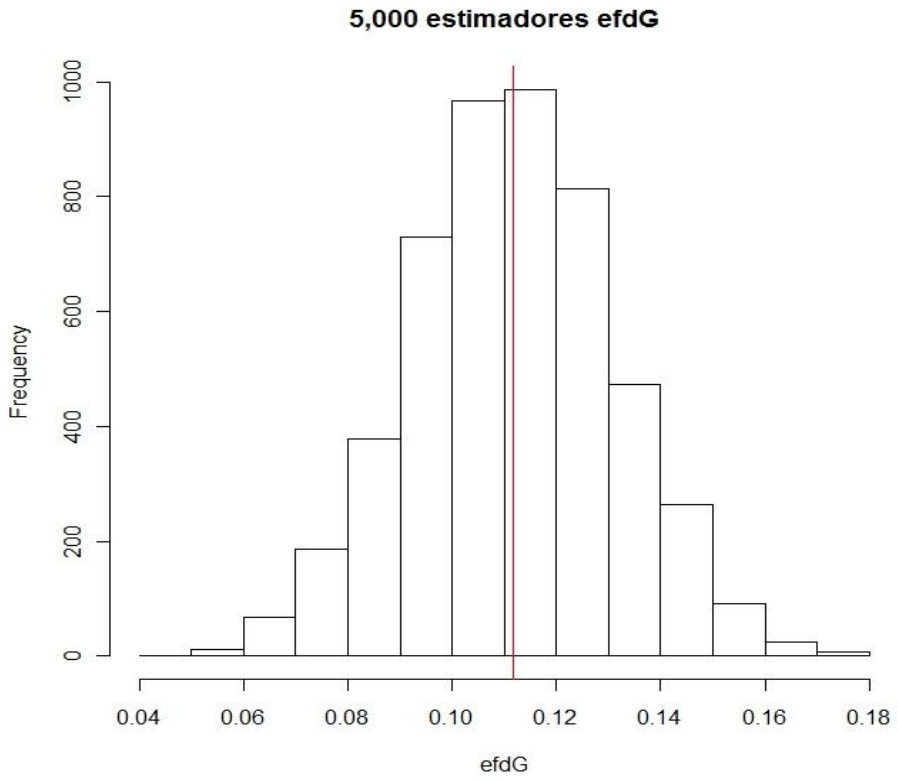
continuación se muestra un histograma con el resultado de las 5,000 muestras *mae* para el estimador efd_G. El estimador efd_G tuvo un sesgo muy pequeño ya que la diferencia con respecto al efd_K fue de 0.3%.

Por lo que concierne al estimador bootstrap de la varianza del efd_K, la raíz cuadrada del promedio de las 5,000 estimaciones de varianza *BWO* fue de 0.021. Es importante recordar que cada estimador de varianza *BWO* se obtiene con B=2,000 réplicas de una *mae*. Por otra parte y con base en las 5,000 muestras *mae*, se construyó también una estimación de la varianza del efd_K, empleando el estimador efd_G. La raíz cuadrada de la varianza entre los 5,000 estimadores efd_G fue de 0.018. Al comparar estas dos estimaciones, la estimación bootstrap fue un 17% más alta que la obtenida con los 5,000 estimadores efd_G. Obsérvese que la cobertura que se encuentra al final de la tabla 11 fue aproximadamente del 90%, lo cual está 5% debajo de la cobertura nominal del 95%.

Según Sitter (1992), la cobertura puede mejorarse con un método diferente al de los percentiles; sin embargo, se encontró durante las simulaciones que el incremento del número de réplicas mejoraba sustancialmente la cobertura. En un principio se usaron valores de B similares a los empleados por Sitter (1992), pero al consultar este tema en Stuart et al. (1999), se encontró que estos autores recomendaban al menos 2,000 réplicas para

estimaciones de varianza en el caso del bootstrap para muestras aleatorias. Debido a que no se tienen resultados de este tipo para el caso del muestreo probabilístico, se empleó dicho valor para el número de réplicas lo cual se tradujo en una mejora en la cobertura, comparado con los valores usados por Sitter (1992).

Gráfica 3

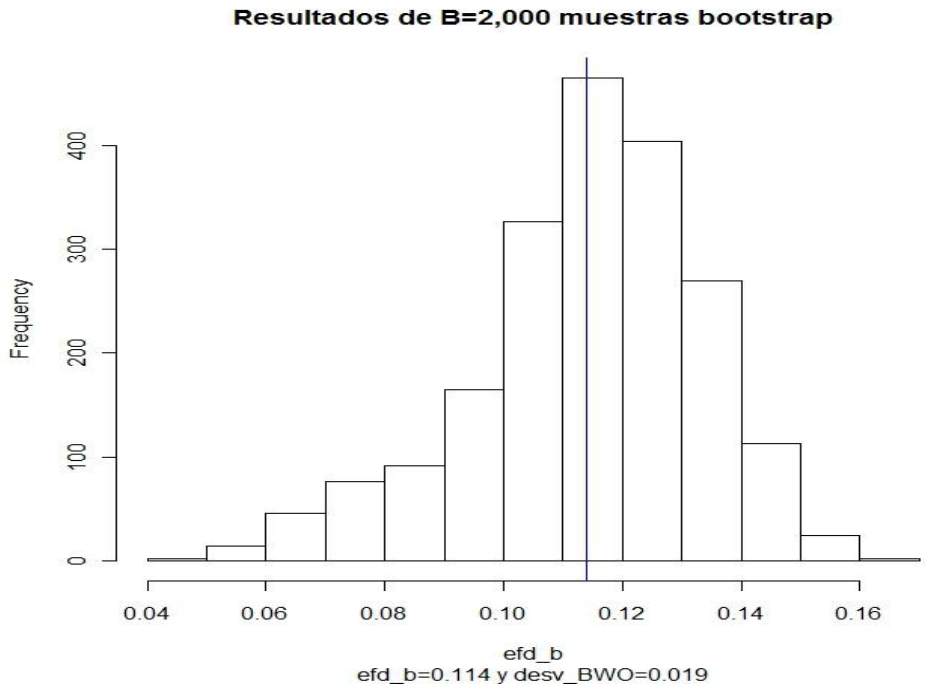


La gráfica 3 muestra el histograma de las 5,000 estimaciones efd_G, la línea roja corresponde al valor poblacional del efecto del diseño que es 0.1114. En

este ejemplo, la ligera asimetría del histograma podría obedecer a un tamaño de muestra pequeño, ya que según Cochran (1977), la distribución normal se aplica a los estimadores de razón, efd_G en este caso, cuando el tamaño de muestra excede de 30 unidades, el coeficiente de variación de las estimaciones insesgadas de la varianza no supera el 10% y la medida de asimetría de Fisher es cercana a cero, para este último punto véase Sugden et al. (2000).

En la práctica se tendría una sola muestra *mae* y las *B* muestras replicadas a partir de ella. Con la muestra *mae* se calcularía el estimador efd_G y con las muestras bootstrap se generaría un histograma del cual se obtendrían los límites inferior y superior de un intervalo calculado por el método de los percentiles. A continuación, gráfica 4, se muestra un histograma de los estimadores efd construido con *B=2,000* réplicas usando el bootstrap empleado en este trabajo, obtenido a partir de una de la última muestra *mae* de tamaño 40 usada en la simulación. Esta muestra no se eligió por algún motivo en particular, simplemente se escogió la última para mostrar un histograma.

Gráfica 4



La línea azul corresponde al valor poblacional del efecto del diseño. En este caso se obtuvo una varianza estimada de 0.000361 o una desviación estándar con un valor de 0.019. Es importante mencionar que no se ha estudiado el motivo por el cual en la gráfica 4 se aprecia una asimetría a la izquierda. No se puede afirmar que sea algo que suceda en todas las iteraciones del bootstrap y es un tema de estudio futuro.

Del histograma de la gráfica 4 se obtiene la información necesaria para construir un intervalo, con un porcentaje deseado, del estimador del efecto del diseño. Con los límites inferior y superior del intervalo se puede evaluar la

precisión del tamaño de muestra calculado con el estimador del efecto del diseño. Usando la última expresión del lado derecho de la ecuación 2, haciendo caso omiso de los valores *tnr* y *PHV* y con la notación del presente ejemplo, $n=n_{mas}efd_G$, se tendrían los límites inferior y superior para el tamaño de muestra. Es decir, se calcularían como $n_{inf}=n_{mas}inf\_efd_G$ y $n_{sup}=n_{mas}sup\_efd_G$, donde inf_efd_G y sup_efd_G son los límites inferior y superior del intervalo construido con base en el histograma bootstrap de los estimadores del efecto del diseño.

7. Conclusiones y recomendaciones

Se construyó una expresión exacta para el sesgo del estimador del efecto del diseño, así como un límite superior de la relación sesgo a error estándar para dicho estimador, al usar un estimador insesgado de la varianza bajo muestreo aleatorio simple. La cota para la relación sesgo a error estándar está dada por el coeficiente de variación de los estimadores insesgados de la varianza bajo muestreo aleatorio simple. Con base en los resultados de las simulaciones y el uso generalizado del efecto del diseño en la práctica, es recomendable analizar la estabilidad del estimador de la varianza bajo muestreo aleatorio simple, *siempre que sea posible*. A la luz de los resultados, es recomendable

emplear un estimador insesgado de la varianza bajo muestreo aleatorio simple.

Con el fin de estimar la varianza del estimador del efecto del diseño, se propuso la aplicación del bootstrap para diseños muestrales diferentes del muestreo aleatorio simple, empleando uno de los métodos de Sitter (1992). Esto se ejemplificó con una población pequeña y con un diseño aleatorio estratificado. Los resultados sugieren que es factible estimar la varianza con este esquema. La cobertura para el estimador del efecto del diseño requiere mejorar, lo cual podría hacerse con las recomendaciones de Sitter (1992) al emplear otra variante del bootstrap mencionadas en esta presentación.

Como se vio en el ejemplo de la sección 6, al calcular el tamaño de muestra con el estimador del efecto del diseño, es factible obtener límites inferior y superior para dicho tamaño de muestra. Esto permitirá una mejor asignación de recursos en la etapa de planeación de una encuesta.

Bibliografía

Chambers, R.L. & Skinner, C.J. (eds.) (2003) *Analysis of Survey Data*, Wiley Series in Survey Methodology.

Chaudhuri, A. & Stenger, H. (2005) *Survey Sampling: theory and methods*, 2nd edn., Chapman & Hall/CRC.

Cochran, W.G. (1977) *Sampling Techniques*, 3rd edn. New York: Wiley.

Gambino, J.G. (2009) *Design effects caveat*, The American Statistician, pp. 141-145.

Gambino, J.G. (2005) *pps: Functions for PPS sampling*. R package version 0.94.

Hernández, F.M. (2003) *Cálculo de Probabilidades.* Aportaciones matemáticas. Serie Textos Nivel Elemental No. 25. Sociedad Matemática Mexicana.

Horvitz, D.G. & Thompson, D. J. (1952) *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association 47, pp. 663-685.

INEGI, *Encuesta Nacional de Ingresos y Gastos de los Hogares 2008*. *Diseño Muestral*.

Kish, L. (1965) *Survey Sampling*, New York: Wiley & Sons.

Lumley, T. (2010) *survey: analysis of complex survey samples*. R package version 3.23-3.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1985) Introduction to the Theory of Statistics, McGraw Hill.

Padilla, A.M., *Una cota para el sesgo relativo del efecto del diseño*, Memorias electrónicas en extenso de la 4ª Semana Internacional de la Estadística y la Probabilidad. Julio 2011, CD ISBN: 978-607-487-324-5.

Padilla, A.M., *A bound for the relative bias of the design effect*. ICES IV, Fourth International Conference on Establishment Surveys, June 11-14, 2012, Montréal, Québec, Canada.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J.N.K. (1962) *On the estimation of the relative efficiency of sampling procedures*, Annals of the Institute of Statistical Mathematics, pp. 143-150.

Rao, J.N.K., Hartley, H.O. & Cochran, W.G. (1962) *On a simple procedure of unequal probability sampling without replacement*, Journal of the Royal Statistical Society B 24, pp. 482-491.

Särndal, C.E., Swensson, B. & Wretman, J.H. (1992) *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.

Sitter, R.R., (1992) *A resampling procedure for complex survey data*, Journal of the American Statistical Association, Vol. 87, pp. 755-765.

Stuart, A., Ord, K. & Arnold (1999) S. Kendall's Advanced Theory of Statistics (sixth edn). Volume 2ª, Classical Inference and the Linear Model. Edward Arnold, London.

Sugden, R.A., Smith, T.M.F. & Jones, R.P. (2000). *Cochran´s rule for simple random sampling*. Journal of the Royal Statistical Society, Series B (Statistical Methodology), Vol. 62, No.4, pp. 787-793.

Thompson, M.E. (1997). *Theory of Sample Surveys*. Chapman & Hall, London.

Valliant, R., Dorfman, A. and Royall, R.(2000) *Finite Population Sampling and Inference: a prediction approach*, John Wiley and Sons, New York.

Apéndice

Prueba del teorema, sección 4:

Sean $E(ecm\hat{n}_{alt}) = v_{alt} + sesgo_{alt}$ y $E(ecm\hat{n}_{mas}) = v_{mas} + sesgo_{mas}$ y

$efd\hat{a}_K = \frac{ecm\hat{n}_{alt}}{ecm\hat{n}_{mas}}$, entonces evaluemos la siguiente expresión:

$$\operatorname{cov}(ecm\hat{n}_{mas}, efd\hat{a}_K) = E(ecm\hat{n}_{mas} efd\hat{a}_K) - E(ecm\hat{n}_{mas})E(efd\hat{a}_K)$$

Despejamos $E(efd\hat{a}_K)$ y como $ecm\hat{n}_{mas} efd\hat{a}_K = ecm\hat{n}_{alt}$, se tiene que

$$E(efd\hat{a}_K) = \frac{E(ecm\hat{n}_{alt})}{E(ecm\hat{n}_{mas})} - \frac{\operatorname{cov}(ecm\hat{n}_{mas}, efd\hat{a}_K)}{E(ecm\hat{n}_{mas})}$$ y usando las expresiones

$E(ecm\hat{n}_{alt}) = v_{alt} + sesgo_{alt}$ y $E(ecm\hat{n}_{mas}) = v_{mas} + sesgo_{mas}$ se tiene el resultado del teorema.

Prueba del corolario 1, sección 4:

Como los estimadores del numerador y denominador son insesgados, entonces $efd\hat{a}_K = efd\hat{a}_G$, $E(ecm\hat{n}_{alt}) = v_{alt}$ y $E(ecm\hat{n}_{mas}) = v_{mas}$ y se tiene el resultado del corolario 1.

Prueba del corolario 2, sección 4:

Como $\left|sesgo(\hat{efd}_G)\right| = \frac{\left|\text{cov}(\hat{v}_{insesg,mas}, \hat{efd}_G)\right|}{v_{mas}}$ y por la definición de correlación

$\left|\text{cov}(\hat{v}_{insesg,mas}, \hat{efd}_G)\right| = \left|\rho(\hat{v}_{insesg,mas})\right|\sqrt{v(\hat{v}_{insesg,mas})}\sqrt{v(\hat{efd}_G)}$, se tiene que:

$$\frac{\left|sesgo(\hat{efd}_G)\right|}{\sqrt{v(\hat{efd}_G)}} = \frac{\left|\rho(\hat{v}_{insesg,mas}, \hat{efd}_G)\right|\sqrt{v(\hat{v}_{insesg,mas})}}{v_{mas}}.$$

En particular , ρ≤1, por lo que $\frac{\left|sesgo(\hat{efd}_G)\right|}{\sqrt{v(\hat{efd}_G)}} = \frac{\sqrt{v(\hat{v}_{insesg,mas})}}{v_{mas}} = cv(\hat{v}_{insesg,mas})$,

con lo cual queda demostrado el resultado.

Prueba del corolario 3, sección 4:

Al aplicar la ecuación (4) del teorema 4, con $sesgo_{alt}=0$, para $\hat{efd}_K$ se tiene

que: $E(\hat{efd}_K) = \frac{v_{alt}}{v_{mas} + sesgo_{mas}} - \frac{\text{cov}(\hat{ecm}_{mas}, \hat{efd}_K)}{v_{mas} + sesgo_{mas}}$

Y $E(\hat{efd}_K) = (efd - sesgo_{mas})(1 + \frac{sesgo_{mas}}{v_{mas}})^{-1}$

Al hacer lo mismo para el $\hat{efd}_G$ se tiene que: $E(\hat{efd}_G) = efd - sesgo_{mas}$

Con esto ya se pueden comparar los sesgos relativos para los dos estimadores del efecto del diseño, $\hat{efd}_K$ y $\hat{efd}_G$. El estimador de Gambino tendrá un sesgo relativo más pequeño sí:

$$1 - \frac{sesgo(\hat{efd}_G)}{efd} < (1 - \frac{sesgo(\hat{efd}_K)}{efd})(1 + \frac{sesgo_{mas}}{v_{mas}})^{-1}$$

.