Padilla Terán, Alberto Manuel

Working Paper

Estimación de razón-remuestreo en muestreo estratificado

Working Papers, No. 2016-02

Provided in Cooperation with:
Bank of Mexico, Mexico City

Suggested Citation: Padilla Terán, Alberto Manuel (2016) : Estimación de razón-remuestreo en muestreo estratificado, Working Papers, No. 2016-02, Banco de México, Ciudad de México

This Version is available at:
https://hdl.handle.net/10419/174432

Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2016-02

Estimación de razón-remuestreo en muestreo estratificado

Alberto Manuel Padilla Terán
Banco de México

Febrero 2016

Documento de Investigación
2016-02

Working Paper
2016-02

Estimación de razón-remuestreo en muestreo estratificado*

Alberto Manuel Padilla Terán†
Banco de México

Resumen: En el muestreo estratificado existen diseños en los que se extrae una unidad por estrato o se desean hacer estimaciones para dominios no planeados en los que se tiene una unidad en algunos estratos. En estos casos, la estimación de la varianza se efectúa en general con el método de estratos colapsados, el cual requiere la identificación de los estratos por colapsar previo al proceso de estimación. Esto puede ser complicado en encuestas con muchos estratos o variables por estimar. En este trabajo se propone una alternativa a esta problemática por medio de dos estimadores de razón basados en submuestras de promedios por estrato tipo jackknife, los cuales son fáciles de calcular sin necesidad de colapsar estratos. Los estimadores son sesgados y se construyen las expresiones para el sesgo, así como las estimaciones del sesgo con datos muestrales. Se presentan algunos ejemplos, entre ellos, estimaciones usando datos de empleo y de medidas de pobreza.
Palabras Clave: Estimador de razón, remuestreo, estimación de varianza, intervalo de confianza, subpoblaciones, estratos colapsados.

Abstract: In stratified sampling there are designs with one unit selected per stratum or when one is willing to make estimations on unplanned domains with one unit in some strata. In these cases the variance is generally estimated by the collapsed strata method, which requires identification of the strata to be collapsed previous to variance estimation. This can be quite complex in surveys with a lot of strata or variables to estimate. In this article we propose an alternative method by means of two ratio estimators based on strata means obtained from jackknife subsamples, which are easy to compute without collapsing strata. The estimators are biased and we build the expressions for them, together with their sample estimators. Some examples are given, among them, estimations with data from the Mexican employment survey and poverty measures.
Keywords: Ratio estimator, Resampling, Variance estimation, Confidence interval, Subpopulations, Collapsed strata.
JEL Classification: C80, C83

*El autor agradece a los participantes del seminario del Banco de México, así como a dos revisores del Banco de México por sus comentarios y sugerencias.

† Dirección General de Investigación Económica. Correo electrónico: ampadilla@banxico.org.mx.

1. INTRODUCCIÓN

En el muestreo aleatorio estratificado hay ocasiones en las que se seleccionan muestras que tienen un elemento por estrato, el cual muchas veces es un conglomerado grande del que a su vez se extraen submuestras, como en la Mini-Finland Health Survey (1977). Al tener un tamaño de muestra igual a uno en uno o más estratos, ya no se cuenta con un estimador insesgado de la varianza de una media o un total. Las fórmulas de los estimadores de varianza dentro de estratos tienen un componente igual al número de elementos en muestra menos uno en el denominador. Si se tiene un tamaño de muestra uno en un estrato, el denominador toma el valor cero y no puede calcularse el estimador de varianza. De esta manera, la información muestral de dicho estrato solamente proviene de un elemento y no se tiene idea de la variabilidad dentro de ese estrato. Cuando esto sucede, es necesario construir un estimador de varianza que permita, por ejemplo, agrupar los totales de estratos con otros semejantes. Este tipo de construcciones inducen sesgos en las estimaciones de varianza, a menos que la agrupación sea tal que el estrato con tamaño de muestra uno se asocie con otro que sea igual en variabilidad, lo cual es poco probable que suceda en la práctica. También hay diseños estratificados en los que se tienen algunos estratos con tamaños de muestra uno o con muy pocas unidades, dos o tres unidades en muestra, como en la Canadian Health Measures Surveys, CHMS, encuesta bianual iniciada en 2007. Por otra parte, hay ocasiones en las que en una muestra estratificada se desea obtener estimaciones en dominios no planeados que abarquen todos los estratos. En este caso es posible que se termine con tamaños de muestra uno o muy pequeños en algunos estratos y

no sea posible obtener un estimador de varianza o dicha estimación sea inestable a causa de tamaños de muestra muy pequeños en algunos estratos. En este caso, la inestabilidad se tendría en el estimador de varianza si dentro de algún estrato hay pocas observaciones, como dos o tres, y los valores presentan un coeficiente de variación grande, digamos mayor al 15% o 20%. Por supuesto, la situación se complica si la selección de la muestra dentro de cada estrato corresponde a un diseño complejo, es decir, que comprende conglomeración, otra estratificación y probabilidades desiguales de selección de conglomerados o elementos, véase Wolter (1985).

Para obtener estimaciones en el caso de los diseños que tienen una unidad por estrato, uno de los métodos más empleados es el de estratos colapsados, véase Cochran (1986) y Särndal et al. (1992). En la versión más sencilla del método, se colapsan por pares los estratos cuyos totales no difieran mucho entre sí y se trabaja con una fórmula de estimación muy sencilla. El problema con este método es que la decisión de los estratos por colapsar debe ser hecha previo a la selección de la muestra, de lo contrario se pueden inducir sesgos en la estimación. Este método se complica en caso de tener un número impar de estratos. Fuller (1970) propuso un método distinto en el que se recorren con una muestra sistemática circular los estratos y se realizan las estimaciones. En Mantel & Giroux (2009) se encuentra una metodología desarrollada por los autores para la estimación en la CHMS, que es una modificación del método de estratos colapsados. Además, comparan su método con algunos otros que se han desarrollado, incluidos el remuestreo, jackknife y bootstrap, en los estratos colapsados. Breidt et al. (2014) propusieron recientemente un estimador no paramétrico, en el cual el criterio para colapsar estratos se basa en las distancias entre los valores por estrato

de una variable auxiliar que esté correlacionada positivamente con la variable por estimar. Dichas distancias, divididas por un ancho de banda, se emplean como argumento en un kernel ponderado para cada estrato. En este método, se pueden tener estratos colapsados con una cantidad diferente de estratos y podría haber estratos colapsados con dos o más estratos según sea la distancia entre ellos. Por otra parte, tanto el estimador puntual del promedio o total, como el de la varianza, requieren de un esfuerzo medio de programación. Sin embargo, el aspecto que requiere atención especial es la elección o determinación del ancho de banda. Breidt et al. (2014) no mencionan la manera en que lo eligieron; empero, en el ejemplo de las estimaciones de porcentajes de pobreza con base en datos del Consejo Nacional de Evaluación de la Política de Desarrollo Social, CONEVAL, se mencionarán las dificultades encontradas al tratar de implementar este método. Al igual que con la estimación de estratos colapsados arriba mencionada, se requiere el uso de una variable auxiliar por estrato.

En relación con la estimación en dominios no planeados en diseños estratificados, un método que en general proporciona buenos resultados es el de la estimación de razón separada o combinada, dependiendo del tipo de información con que se cuente, véase Särndal et al. (1992) y anexo 3. Sin embargo, en el caso en el que se tengan una o muy pocas observaciones por dominio y estrato, la estimación de varianza puede ser inestable.

Es importante notar que, salvo el método de estratos colapsados, el resto de los métodos son difíciles de implementar o requieren más conocimiento del diseño usado dentro de cada estrato para estimar la varianza.

En este trabajo se presenta un estimador de razón basado en el remuestreo de promedios por estrato, el cual es relativamente fácil de usar y puede emplearse para el problema de estimación en estratos con una o muy pocas unidades por estratos, así como en la estimación de dominios no planeados en estratificación. También se propone una estimación de varianza basada en la estimación de razón. El método se ilustra con varios ejemplos, entre ellos, se emplean datos del último trimestre de 2012 de la ENOE para estimar el ingreso promedio de ocupación por hora en México, así como la varianza para dicha estimación. También se estima el porcentaje de pobreza en México, así como su varianza, con datos publicados por el CONEVAL, para el 2010 y 2012. Para este último ejemplo se realiza una comparación con el método de estratos colapsados y se mencionan algunos puntos encontrados al tratar de construir los estimadores no paramétricos propuestos por Breidt et al. (2014).

El artículo se encuentra organizado de la siguiente manera, en la sección 2 se introduce la notación, así como las expresiones para las estimaciones estratificadas y de razón. También se explica el método de remuestreo llamado jackknife, que se empleará como base de los estimadores propuestos. En la sección 3 se describen brevemente los métodos de estratos colapsados y de estimación no paramétrica de Breidt et al. (2014). En la sección 4 se presenta el estimador propuesto, así como la estimación de varianza y el sesgo del estimador. En la sección 5 se ilustra el método con varios ejemplos. En esta sección, con los datos de pobreza por entidad federativa publicados por el CONEVAL para 2010 y 2012 se efectúa una comparación con el método de estratos colapsados y se mencionan los problemas que se tuvieron al tratar de usar el método no paramétrico de Breidt et al. (2014).

Es importante mencionar que el estimador propuesto no se aplica a dominios no planeados en los que no se tiene información en algún estrato o estratos. Esto se debe a la manera en la que está construido el estimador, el cual requiere de información en todos los estratos. En caso de que no se tenga información del dominio no planeado en algún estrato, podría efectuarse una imputación y aplicarse el método; empero, esto es un tema para investigación futura.

2. DEFINICIONES Y NOTACIÓN

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}*. El tamaño de muestra se denotará con *n*. Como se tratarán los diseños estratificados y el estimador de razón, a continuación se presenta la notación empleada.

2.1 Notación, población y muestreo aleatorio estratificado.

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 escribirá como $\bar{y}_{st} = \sum_{h=1}^{H} W_h \bar{y}_h$, donde $W_h = N_h / N$ y $\bar{y}_h = \sum_{i=1}^{N_h} y_{hi} / N_h$; en tanto que el estimador del promedio es $\hat{\bar{y}}_{st} = \sum_{h=1}^{H} W_h \hat{\bar{y}}_h$, con $\hat{\bar{y}}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$. Las W_h son tales que $W_h \in (0,1)$. El total poblacional y

su estimador puntual por estrato se representará con y_h y $\hat{y}_h$ respectivamente. La varianza poblacional entre elementos dentro de estratos se escribirá como s_{hU}^2, la estimación muestral como s_h^2 y la varianza poblacional del estimador del promedio, usando muestreo aleatorio simple, *mas*, dentro de estratos se denotará como $v_{mae}(\hat{y}_{st}) = \sum_{h=1}^{H} W_h^2 (1 - n_h/N_h) s_{hU}^2 / n_h$ y la estimación muestral se escribirá como $\hat{v}_{mae}(\hat{y}_{st})$ con s_h^2 en lugar de s_{hU}^2 en la fórmula para v_{mae}. Aquí *mae* se refiere al muestreo aleatorio estratificado.

2.2 Notación, estimador de razón.

La notación expuesta en esta sección, se refiere a expresiones del estimador de razón para una población finita en general. Si se desea aplicar a un diseño estratificado con estimaciones separadas por estrato, sólo se añade el subíndice *h* a las expresiones.

Para algún diseño muestral en el que se tenga una variable auxiliar, este estimador se define como, véase Cochran (1986), $\hat{r} = \hat{y}/\hat{x}$, para $\hat{x} > 0$, los estimadores $\hat{y}$ y $\hat{x}$ son del tipo Horvitz-Thompson, véase Särndal et al. (1992), Horvitz & Thompson (1952) y Anexo 3. Por otra parte, la correspondiente cantidad poblacional se denotará como $r_u = y/x$, con $x > 0$, donde las cantidades del numerador y denominador son los totales poblacionales de las dos variables. Nótese que tanto el estimador como la cantidad poblacional pueden también calcularse usando los promedios muestrales y poblacionales respectivamente. Este estimador presenta una varianza menor que la del *mas* cuando la variable auxiliar está

correlacionada positivamente con la variable de interés y_i, por lo cual es de bastante utilidad en la práctica.

El estimador de razón es un estimador sesgado ya que es el cociente de dos variables aleatorias y en general, la esperanza del cociente de dos variables aleatorias no es igual que el cociente de las esperanzas, véase Cochran (1986) ó Särndal et al. (1992) y la magnitud del sesgo se mide con la relación sesgo a error estándar. Esta última cantidad es menor o igual que el coeficiente de variación de la variable auxiliar x, y será despreciable si dicho coeficiente de variación es menor o igual que 10%, véase la tabla 1.1 de la sección 1.8 de Cochran (1986). Dicho autor recomienda esto, ya que si la relación sesgo a desviación estándar es menor o igual que 0.10, en lugar de trabajar con un 95% de confianza para un intervalo, se estaría trabajando con un 94.89%. Por otro lado si dicha relación fuese igual a 0.40, se estaría trabajando al 93.15%.

La varianza del estimador de razón utilizando muestreo aleatorio simple es: $v_r(\hat{r}) = (1-f)s_r^2 / (n\bar{x}_u^2)$, con $f = n/N$, $\bar{x}_u$ es el promedio poblacional y $s_r^2 = \sum_{i=1}^{N}(y_i - r_u x_i)^2 / (N-1)$. A f se le conoce como la fracción de muestreo. Bajo *mas*, la estimación de la varianza se efectúa con la siguiente fórmula $\hat{v}_r(\hat{r}) = (1-f)\hat{s}_r^2 / (n\hat{\bar{x}}^2)$, en donde $\hat{\bar{x}}$ es la estimación muestral de la media poblacional de x y $\hat{s}_r^2 = \sum_{i=1}^{n}(y_i - \hat{r}x_i)^2 / (n-1)$.

2.3 Algunos puntos acerca de un método de remuestreo: el jackknife

El jackknife fue propuesto por Quenouille (1949, 1956), véase también Cochran (1986), en el contexto del problema de estimación de varianza en el muestreo sistemático y estratificado. El término jackknife fue acuñado por Tukey (1958) y se refiere a una navaja multiusos, fácil de portar. Tukey lo propuso como un procedimiento general para pruebas de hipótesis y el cálculo de intervalos de confianza. La versión más sencilla del jackknife se esboza a continuación y se trabajará con ella en el presente artículo. Para una muestra aleatoria de tamaño $n, (x_1, x_2, \cdots, x_n)$, las muestras jackknife se calculan dejando fuera un elemento x_i de los n a la vez, para $i \in \{1, \cdots, n\}$ y se denota como $x_{(-i)} = (x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n)$.

De esta manera, el estimador de un promedio se calcula de la siguiente manera, $\hat{x}_{(-i)} = \sum_{j=1}^{n} \frac{x_j}{n-1}$, con $j \neq i$ y el estimador jackknife del promedio, *jkf*, es:

$$\hat{x}_{jkf} = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_{(-i)}.$$

El estimador *jkf* del error estándar tiene la siguiente expresión:

$$\hat{e}_{jkf} = \left[\frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{x}_{(-i)} - \hat{x}_{jkf} \right)^2 \right]^{1/2}.$$

3. ESTIMADOR DE ESTRATOS COLAPSADOS Y NO PARAMÉTRICO

3.1 Estimador de estratos colapsados

En aquellos diseños estratificados en los que se extrae una unidad por estrato, no se puede estimar insesgadamente la varianza y, como se mencionó al principio, un método que se emplea con regularidad en estos casos es el de estratos colapsados, véase Cochran (1986) y Särndal et al. (1992). Los estratos se colapsan de manera previa a ver los resultados de la muestra y deberían agruparse estratos cuyos totales difieran poco entre sí. El estimador puntual de la media es el mencionado en la sección 2, $\hat{\bar{y}}_{st} = \sum_{h=1}^{H} W_h \hat{\bar{y}}_h$, siempre que se tenga un estimador insesgado por estrato. Para la estimación de la varianza en una población con un número par de estratos, se emplea la siguiente expresión, si se conoce el total de elementos en población, *N*, Cochran (1986),

$$\hat{v}_c(\hat{\bar{y}}_{st}) = \frac{1}{N^2} \sum_{j=1}^{H/2} \left(\hat{y}_{j1} - \hat{y}_{j2}\right)^2 . \qquad (1)$$

En esta fórmula, $\hat{y}_{j1}$ y $\hat{y}_{j2}$, se refieren a la estimación de totales, y_{j1} y y_{j2} en el j-ésimo estrato colapsado y el subíndice *c* en $\hat{v}_c(\hat{\bar{y}}_{st})$ al estimador de varianza para estratos colapsados. El valor esperado de esta cantidad es,

$$v_c(\hat{\bar{y}}_{st}) = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{1}{N_H}\right) s_{hu}^2 + \frac{1}{N^2} \sum_{j=1}^{H/2} \left(y_{j1} - y_{j2}\right)^2 . \quad (2)$$

El primer término de la derecha corresponde a la varianza poblacional del estimador del muestreo estratificado con tamaño de muestra igual a uno en cada estrato. El segundo término casi siempre será mayor que cero y puede hacer que la varianza sea grande si la elección de estratos colapsados no es adecuada. El estimador (1) es simple y fácil de calcular; empero, la dificultad radica en que los estratos deben agruparse antes de ver los datos. Esto puede ser difícil en la práctica si se trabaja en un diseño que tenga muchos estratos, como uno de los ejemplos que se mostrarán adelante, o si se desean hacer estimaciones de varias variables. También se presentan complicaciones en el evento de tener un número impar de estratos. Para un mayor detalle de métodos relacionados con los estratos colapsados, véase Wolter (1985).

3.2 Estimador no paramétrico propuesto por Breidt et al. (2014)

A continuación se presentará el estimador puntual de la varianza estimada por métodos no paramétricos propuestos por Breidt el al. (2014). No se construirá dicha expresión, ya que se encuentra en el mencionado artículo, sólo se comentarán los puntos en los que se encontraron algunos problemas al intentar implementarlo. La notación del mencionado artículo se adaptó en general a la usada en este documento.

El estimador del promedio es el usual en el muestreo probabilístico, es decir, el estimador de Horvitz-Thompson que se etiquetó como $\hat{y}_{st}$ en la sección 2.1 del presente documento.

Sea u_r, $r \in \{1,2,\cdots,H\}$ una variable conocida para todos los estratos, con correlación positiva alta con la variable de interés y que se emplea para colapsar los estratos. El estimador de varianza no paramétrico, $\hat{v}_{NP}(\hat{y}_{NP})$, se calcula como:

$$\hat{v}_{NP}(\hat{y}_{NP}) = \frac{1}{N^2 C_d} \sum_{r=1}^{H} \left(\hat{y}_r - \sum_{j=1}^{H} d_j(h) \hat{y}_j \right)^2 .$$

En esta fórmula, C_d es una constante de normalización y tiene la siguiente forma:

$$C_d = \frac{1}{H} \sum_{r=1}^{H} \left(1 - 2d_r(r) + \sum_{j=1}^{H} d_j^2(r) \right) .$$

Obsérvese que esta constante depende de las cantidades $d_j(r)$, las cuales son ponderadores que se determinan con base en el kernel o núcleo de la estimación no paramétrica y se calculan como:

$$d_j(r) = \frac{K\left(\frac{u_r - u_j}{h_b} \right)}{\sum_{k=1}^{H} K\left(\frac{u_r - u_k}{h_b} \right)} .$$

Aquí, $K(\cdot)$ es un núcleo o kernel acotado y simétrico, en tanto que h_b es un ancho de banda (en inglés se le conoce como bandwith). En el artículo de Breidt el al. (2014) se usa el kernel de Epanechnikov, el cual se define como $K(x) = 0.75(1 - x^2) 1_{\{|x| \le 1\}}$. En el contexto de esta estimación, para cada u_r, $r \in (1,2,\cdots,H)$, se encuentran aquellas u_j tales que $|x_j(r)| \le 1$, con $x_j(r) = (u_r - u_j)/h_b$. La suma en el denominador para $d_j(r)$ sólo se aplica sobre las u_k

que satisfacen $|x_k(r)| \leq 1$. Nótese que una vez dadas las u_r, el requerimiento $|x| \leq 1$ impone restricciones sobre los posibles valores para h_b.

En el contexto de la estimación no paramétrica la selección del valor h_b es crucial y, en el caso que nos ocupa, afecta al número de elementos que conforman cada estrato colapsado. Con esta metodología los estratos colapsados no necesariamente contienen dos estratos, pueden tener diferente número de estratos, inclusive uno solo. En esta última situación, obsérvese que se tiene $u_r - u_r = 0$ y, por lo tanto, $K(\cdot) = 0$; por lo que $d_r(r)$ no está definida y no puede usarse el estimador de varianza $\hat{v}_{NP}$.

Este caso fue el que se encontró al tratar de aplicar este método con los datos del ejemplo 3 de la sección 5, usando las W_h.

4. ESTIMADORES PROPUESTOS

En esta sección se presentan tres tipos de estimadores puntuales para la media, junto con estimadores de varianza. Dos de ellos, denominados tipo a y b, requieren del tamaño relativo de los estratos, W_h, en tanto que el tercero, llamado tipo c, no hará uso de los tamaños relativos, sino de los factores de expansión y es el que se propondrá para la estimación en dominios no planeados. El estimador de varianza tipo a, es muy fácil de usar, ya que sólo se requieren los tamaños relativos de los estratos de una población y las estimaciones puntuales de cada estrato provenientes de una muestra estratificada. Como se verá en los ejemplos, esta información aparece publicada como resultado de encuestas de

organismos gubernamentales como el INEGI o el CONEVAL. Para el estimador tipo b se necesita la misma información que el tipo a, pero se hacen más cálculos, lo cual se traduce en unas cuantas líneas de programación. No obstante lo anterior, continúa siendo un estimador fácil de construir para tener una idea de la variabilidad de la estimación en caso de que no haya sido publicada.

Por otra parte, al final de esta sección se encuentra un resultado en el que se establece una relación entre la varianza del estimador de razón y la del jackknife para una muestra aleatoria de elementos con igual ponderación. Este resultado no se usa en los ejemplos, pero en opinión del autor abre un camino de investigación que no ha sido explorado en la literatura.

4.1 Tipo a, estimador de razón-remuestreo con variable auxiliar W_i

Considérese una población U de N elementos particionada en H estratos en la que se desea estimar un promedio poblacional, así como una estimación de varianza de dicho promedio con base en una muestra extraída con algún diseño, posiblemente complejo, que produzca estimaciones insesgadas de promedios por estrato. El estimador propuesto, que se etiquetará como *erra*, se motiva con el procedimiento siguiente.

a. Calcúlense todas las medias por estrato, $\hat{y}_h$, con el estimador propio del diseño muestral. En caso de que el estrato tenga una observación, está constituirá la media estimada.

b. Para $i \in \{1,2,\ldots,H\}$, elimínese la información de la *i*-ésima observación, es decir, W_i y $\hat{y}_i$, y calcúlese un nuevo estimador del promedio sin el *i*-ésimo estrato de la siguiente manera: pondere las W_h, con $h \neq i$, para que sumen uno, $\partial_{h(-i)} = W_h/(1-W_i)$ y calcule $\hat{y}_{(-i)} = \sum_{h=1}^{H} \partial_{h(-i)} \hat{y}_{h(-i)}$. Al usar el símbolo *(-i)* se entiende que las operaciones se realizan sin considerar el *i*-ésimo estrato.

c. Estimador de razón-remuestreo para la media. Construya el estimador en estratos como un estimador de razón combinado, véase Särndal et al. (1992) y Anexo 3, usando los *H* valores $\hat{y}_{(-i)}$:

$$\hat{y}_{erra} = \sum_{i=1}^{H} \hat{y}_{(-i)} \Big/ \sum_{i=1}^{H} \sum_{h=1}^{H} \partial_{h(-i)} . \qquad (3)$$

Esto es un estimador de razón, pero en este caso, el denominador es igual a *H*, ya que $\sum_{h=1}^{H} \partial_{h(-i)} = 1$, por lo cual:

$$\hat{y}_{erra} = \sum_{i=1}^{H} \hat{y}_{(-i)} \Big/ H . \qquad (4)$$

Obsérvese que para calcular el estimador en (3) sólo se requiere contar con los promedios estimados por estrato con base en el diseño usado en la extracción de la muestra, así como los tamaños relativos de los estratos W_h.

Sesgo del estimador para la media.

El estimador $\hat{y}_{erra}$ es un estimador sesgado y el sesgo se obtiene al hacer,

$$E(\hat{y}_{erra}) - y_{st} = \frac{1}{H}\sum_{i=1}^{H}\frac{W_i}{1-W_i}(y_{st} - y_i) \ . \qquad (5)$$

La demostración se encuentra en el anexo 2. La esperanza corresponde a la densidad inducida por el diseño con el que se extrajo la muestra. Nótese que el sesgo puede estimarse con los datos de la muestra, siempre que las medias por estrato sean insesgadas. También se puede demostrar que si todos los estratos tienen el mismo tamaño, entonces el sesgo es cero. Esto es importante, porque en una población con muchos estratos que tengan tamaño parecido y con estimaciones insesgadas de media por estrato, el sesgo será despreciable. Por otra parte, puede tenerse un sesgo de magnitud considerable si se tienen pocos estratos y valores grandes de W_h y $y_{st} - y_h$.

Varianza del estimador de razón-remuestreo de la media.

A continuación se presentan la expresión para la varianza poblacional y su estimador, en caso que exista, del estimador del promedio dado en (3). Estas expresiones se incluyen solamente con fines ilustrativos, ya que el estimador de varianza que se empleará no es el que se mostrará a continuación.

La varianza del estimador de razón-remuestreo con base en una muestra aleatoria estratificada es:

$$v(\hat{y}_{erra}) = \frac{1}{H^2}\sum_{i=1}^{H}\left[\sum_{h=1,h\neq i}^{H}\left(\frac{1}{1-W_h}\right)^2 W_i^2 v(\hat{y}_i)\right]. \qquad (6)$$

Esta expresión se obtiene al aplicar la varianza al estimador $\hat{y}_{erra}$ en la fórmula (4) y notando que las selecciones de muestras entre estratos son independientes entre sí, por lo cual las covarianzas son cero. La estimación de la varianza en (6), siempre que exista un estimador insesgado de varianza por estrato, está dada por:

$$\hat{v}(\hat{y}_{erra}) = \frac{1}{H^2}\sum_{i=1}^{H}\left[\sum_{h=1,h\neq i}^{H}\left(\frac{1}{1-W_h}\right)^2 W_i^2 \hat{v}(\hat{y}_i)\right]. \qquad (7)$$

No se usará el estimador (7) ya que si en un estrato se extrajo una muestra de tamaño uno, no se tiene un estimador insesgado de $v(\hat{y}_i)$ y se estaría en el mismo caso que en la estimación de varianza con estratos colapsados. Las propiedades del estimador (7), en cuanto a eficiencia relativa con el muestreo aleatorio estratificado, se encuentran bajo estudio.

Estimador de varianza de razón-remuestreo.

Por lo expresado en el párrafo anterior, es necesario contar con un estimador de varianza que sea diferente de (7). Retomando la idea detrás de la construcción del estimador en (3), de que se está trabajando con un estimador de razón, un estimador de varianza para el estimador dado en (3) es similar al estimador de varianza de una razón, véase Cochran (1986), sin considerar la corrección por población finita ya que se tienen *H* réplicas independientes del estimador $\hat{\bar{y}}_{(-i)}$. El estimador es,

$$\hat{v}(\hat{y}_{erra}) = \frac{H-1}{H}\sum_{i=1}^{H}\left(\hat{y}_{(-i)} - \hat{y}_{erra}\right)^2. \qquad (8)$$

Es interesante notar que si en lugar de *H* estratos se tuviesen *n* observaciones como en el jackknife de la sección 2.3 y si se hace $W_i = 1/n$, los pesos ∂_i serían $\partial_i = 1/(n-1)$. Al construir las ∂_i y las $\hat{y}_{(-i)}$ para formar el estimador tipo erra, la expresión para la estimación de varianza dada en (8) es la misma que la del jackknife de la sección 2.3.

Con esto se aprecia que el estimador de varianza del estimador de razón-remuestreo para un promedio, es la estimación de varianza dada por el jacknife.

4.2 Tipo b, estimador de varianza usando promedio de varianzas dentro de réplicas

Con pocos estratos, se construye un estimador que promedia las varianzas dentro de promedios de las *H* poblaciones replicadas. La construcción se basa en el estimador tipo *erra* conforme al procedimiento siguiente.

a. Calcúlense todas las medias por estrato, $\hat{y}_h$, con el estimador propio del diseño muestral. En caso de que el estrato tenga una observación, está constituirá la media estimada.

b. Para $i \in \{1,2,\ldots,H\}$, elimínese la información de la *i*-ésima observación, es decir, W_i y $\hat{y}_i$, y calcúlese un nuevo estimador del promedio sin el *i*-ésimo estrato de la siguiente manera: pondere las W_h para que sumen uno, $\partial_{h(-i)} = W_h/(1-W_i)$ y calcule $\hat{y}_{(-i)} = \sum_{h=1}^{H} \partial_{h(-i)} \hat{y}_{h(-i)}$.

c. Estimador de razón-remuestreo para la varianza de la *i*-ésima réplica. Construya el estimador de varianza usando la varianza estimada del estimador de razón, véase Särndal et al. (1992), usando los *H-1* valores $\hat{y}_{h(-i)}$ y $\hat{y}_{(-i)}$:

$$\hat{v}\left(\hat{y}_{(-i)}\right)=\frac{H-1}{H}\sum_{h=1}^{H}\hat{\partial}_{h(-i)}^{2}\left(\hat{y}_{h(-i)}-\hat{y}_{(-i)}\right)^{2}. \qquad (9)$$

A continuación se construye un estimador de varianza promediando las varianzas de cada réplica:

$$\hat{v}\left(\hat{y}_{errb}\right)=\frac{1}{H}\sum_{i=1}^{H}\hat{v}\left(\hat{y}_{(-i)}\right). \qquad (10)$$

En resumen, para el cálculo del estimador *errb* se construyen *H* réplicas de población estratificada de medias y se hace uso de la estimación de varianza del estimador de razón al interior de cada población replicada.

La idea de promediar varianzas se encuentra en el trabajo de Nórlen & Waller (1979) en Särndal et al. (1992). Es importante mencionar que Nórlen & Waller (1979) propusieron un estimador de varianzas basado en un promedio de varianzas obtenidas con un método de remuestreo denominado grupos aleatorios. Este método de estimación es diferente al propuesto en el presente artículo y para una exposición a detalle, véase el capítulo 11 de Särndal et al. (1992).

4.3 Tipo c, estimador de razón-remuestreo para dominios no planeados

En el tema de la estimación de dominios no planeados, véase Särndal et al. (1992), para variables que se encuentren con mediciones en todos los estratos, casi siempre se cuenta con factores de expansión que permiten realizar estimaciones de la variable de interés usando un estimador de razón. El estimador resultante tiene la forma de un estimador de razón combinado en el cual, el numerador tiene la suma de los totales expandidos por estrato para la variable de interés y el denominador contiene la suma de los factores de expansión, que devuelve una estimación del número de elementos en la población que pertenecen al dominio de interés. Esta es una situación diferente a los dos estimadores vistos, ya que las expansiones de totales por estrato hacen uso de los factores de expansión y no necesariamente se cuenta con el número de elementos en población, N_h, en el *h*-ésimo estrato. De hecho, casi siempre esta cantidad se estima con la suma de los factores de expansión en el *h*-ésimo estrato para el dominio de interés.

Con el fin de trabajar con un estimador de razón combinado, se construye el estimador de razón-remuestreo en dominios no planeados, en lo sucesivo *errc*, de la siguiente manera.

a. Calcúlense todos los totales expandidos por estrato, $\hat{y}_h$, con el estimador propio del diseño muestral o usando los factores de expansión en caso de tenerlos. En el evento de que el estrato tenga una observación, calcúlese el total, esto constituirá el total estimado.
b. Para $i \in \{1,2,\ldots,H\}$, elimínese la información de la *i*-ésima observación, es decir, $\hat{y}_i$, y calcúlese un nuevo estimador del total simplemente sumando los totales expandidos

sin la observación del *i*-ésimo estrato; haga lo mismo para los factores de expansión del denominador. Así, calcule $\hat{y}_{(-i)} = \sum_{h=1}^{H} \hat{y}_{h(-i)}$ y $\hat{\gamma}_{(-i)} = \sum_{h=1}^{H} \omega_{h(-i)}$, donde ω_h se refiere a la suma de los factores de expansión en el *h*-ésimo estrato.

c. Estimador de razón-remuestreo, errc. Construya el estimador *errc* como un estimador de razón combinado, véase Särndal et al. (1992) o sección A3.2 del anexo 3, usando los H valores $\hat{y}_{(-i)}$:

$$\hat{y}_{errc} = \sum_{i=1}^{H} \hat{y}_{(-i)} \Big/ \sum_{i=1}^{H} \omega_{h(-i)} . \qquad (11)$$

Obsérvese que un total estimado se encuentra en *H-1* réplicas, por lo cual, la expresión (11) queda como:

$$\hat{y}_{errc} = \sum_{i=1}^{H} \hat{y}_h \Big/ \sum_{i=1}^{H} \hat{\gamma}_h = \hat{y} / \hat{\gamma} . \qquad (12)$$

Es importante hacer notar que la expresión del lado derecho en (12) corresponde al estimador usual de dominios no planeados en el que se calcula el cociente de la suma de totales expandidos en el numerador y denominador.

Como el estimador de la fórmula (12) es sesgado, ya que la cantidad en el denominador varía de réplica a réplica, se tiene que verificar que el sesgo sea despreciable. Hartley y Ross (1954) encontraron que, véase Cochran (1986), el límite superior de la relación sesgo a error estándar del estimador de razón está dado por el coeficiente de variación del denominador. De esta manera, con datos muestrales se calcula dicha cantidad y si el coeficiente es menor al 10%, véase Cochran (1986), el sesgo es despreciable.

De manera similar a los motivos expuestos para la construcción de la fórmula (8), una estimación de varianza del estimador errc (11), al tener *H* réplicas independientes del estimador $\hat{y}_{(-i)}$, es:

$$\hat{V}(\hat{y}_{errc}) = \frac{1}{H(H-1)\hat{\bar{y}}^2} \sum_{i=1}^{H} \left(\hat{y}_{(-i)} - \hat{y}_{errc}\hat{\bar{y}}_{(-i)}\right)^2, \quad (13)$$

donde $\hat{\bar{y}}$ es un estimador del tamaño promedio por estrato de la variable auxiliar, no es el promedio de los *H* totales estimados por estrato. Obsérvese que sí $\hat{\bar{y}} = 1/(H-1)$, entonces el término que multiplica a la suma en (13) se convierte en $(H-1)/H$ y coincide con el término que multiplica a la suma en (8).

5. EJEMPLOS

Ejemplo 1, estimador erra, razón-remuestreo con variable auxiliar W_i

Supóngase que se tiene una población de 120 elementos con *H*=5 estratos, de la cual se extrajo una muestra de tamaño 40 con el fin de estimar el promedio poblacional. Supóngase que el diseño muestral produce estimaciones insesgadas de promedios, como el muestreo aleatorio simple por estrato. En uno de los estratos se extrajo una muestra de tamaño uno. A continuación, tabla 1, se encuentra la información de los estratos, así como los promedios muestrales.

Tabla 1

Información del muestreo estratificado

Estrato	N_h	W_h	n_h	$\hat{y}_h$
1	13	11%	6	2.33
2	18	15%	1	4.02
3	26	22%	10	5.04
4	26	22%	16	7.01
5	37	30%	7	9.86
Total	120		40	

En la tabla 2 se encuentran los principales resultados de la muestra. Las casillas que se encuentran en blanco, se refieren al i-ésimo estrato que se omitió para calcular los valores de las réplicas, Rép1 a Rép5 en el primer renglón de la tabla 2.

Tabla 2

Resultados de las réplicas

Estrato	Rép1	Rép2	Rép3	Rép4	Rép5
1		0.302	0.329	0.329	0.366
2	0.678		0.773	0.773	0.861
3	1.246	1.304		1.422	1.584
4	1.733	1.814	1.977		2.203
5	3.324	3.480	3.792	3.792	
$\hat{y}_{(-i)}$	6.980	6.900	6.871	6.316	5.015

A continuación se muestran los valores del estimador de razón-remuestreo para el promedio, así como la estimación de varianza.

Tabla 3

Resultados de la estimación

Estimadores de razón-remuestreo	Valor
Promedio	6.416
Varianza	2.187
Error estándar	1.479
Coeficiente de variación	23%

El promedio estimado de razón-remuestreo $\hat{y}_{erra} = 6.416$ se compara con el estimador del promedio bajo muestreo aleatorio estratificado, $\hat{y}_{st} = 6.4683$, el cual es un estimador insesgado. El sesgo estimado es igual a -0.052, el cual es muy pequeño comparado con el valor $\hat{y}_{st} = 6.4683$. Este sesgo se calculó con la versión muestral de (5), $\frac{1}{H}\sum_{h=1}^{H}\frac{W_h}{1-W_h}(\hat{y}_{st} - \hat{y}_h)$. Los valores se obtienen de la tabla 1. Como se tiene un tamaño de muestra igual a uno en el estrato 2, no existe un estimador insesgado de la varianza bajo muestreo aleatorio estratificado.

El estimador del promedio de razón-remuestreo presenta un valor moderado de variabilidad, ya que se tiene un coeficiente de variación del 23%; sin embargo, solamente se usó la información de 5 promedios y de los tamaños relativos de los estratos.

Ejemplo 2, estimador errb, simulación con datos de la ENOE.

A continuación se muestra un ejercicio de simulación en el que se emplean datos de la Encuesta Nacional de Ocupación y Empleo, ENOE, correspondientes al cuarto trimestre de 2012, véase INEGI. Dicha encuesta es de tipo panel y se levanta trimestralmente. Se entrevista a personas en hogares y la intención es que un hogar permanezca 5 trimestres en muestra. Cada trimestre se renueva una quinta parte de la muestra y la encuesta se aplica en aproximadamente 120,000 viviendas. Se tienen varios criterios de estratificación, geográficos y sociodemográficos, y para el 2012 se tuvieron un total de 884 estratos.

Ejercicio de simulación. Con el fin de evaluar empíricamente el estimador *errb*, se consideró la muestra del cuarto trimestre del 2012 para personas con ingreso mayor que cero, como si fuera una población con H=884 estratos, de la cual se extraerán 20,000 muestras de tamaño $n_h=1$ en cada estrato, teniendo así $n=884$ elementos por muestra. La característica por estimar en esta población es el 'promedio de ingreso por hora trabajada de la población ocupada' para la población entre 14 y 98 años. Es importante mencionar que sólo se emplean los valores de ingreso mayores que cero, ya que aparecen personas que tuvieron actividad, pero sin remuneración.

En resumen, en esta población se tienen $N=119{,}296$ personas, en $H=884$ estratos, por lo cual $n=884$, y el promedio poblacional del ingreso por hora trabajada es 32.794 pesos, que es la cantidad por estimar. Nótese que, como este es un ejercicio de simulación, se conoce el valor poblacional, en este caso 32.794. En la tabla 4, se encuentran algunas estadísticas

descriptivas de los tamaños relativos de los estratos W_h, así como de las N_h, el número elementos por estrato.

Tabla 4
Algunas estadísticas descriptivas de tamaños relativos
y absolutos de los estratos de la ENOE 2012

Estadística	W_h	N_h
Mínimo	0.000017	2
Máximo	0.011367	1,356
Media	0.001131	135
Mediana	0.000570	68

Puede observarse en la tabla 4 que los tamaños relativos, W_h, sugieren una asimetría positiva y el valor máximo es casi 10 veces mayor que la media. Por otra parte, la varianza poblacional de un diseño estratificado extrayendo un elemento por estrato es $v_{mae}(\hat{y}_{st}) = 4.8678$; sin embargo, esta varianza no tiene un estimador insesgado y hay que emplear algún método para estimar la varianza, como el de estratos colapsados o el que se propone en este artículo y cuyos resultados de una simulación se muestran a continuación. Nótese que en este ejemplo no puede emplearse la fórmula (7) para estimar la varianza por el tamaño de muestra uno en los estratos, por lo cual se usará la expresión dada en (8).

Recordemos que el ejercicio de simulación consistió en extraer 20,000 muestras aleatorias estratificadas de tamaño *n=884*, con $n_h = 1$ para cada estrato. En cada estrato se registró el ingreso por hora trabajada y se calculó el estimador del promedio estratificado, el de razón-remuestreo, fórmula (4), y la varianza estimada de este último estimador, fórmula (8). Se

calculó la cobertura al 95% suponiendo normalidad del estimador y los resultados se muestran a continuación.

Tabla 5

Resultados de la simulación

Cantidad	Valor simulación	Valor poblacional
Estimador $\hat{y}_{mr}$	32.779	32.794
Varianza	4.966	4.868
Error estándar	2.206	2.228
Cobertura	92%	

Con el fin de tener una cantidad empírica con la cual medir la estabilidad de los estimadores de varianza, se calculó el coeficiente de variación entre los 20,000 estimadores de varianza del estimador de razón-remuestreo y resultó en 6.8%. Por lo expuesto en el cuarto párrafo de la sección 2.2, un coeficiente de variación menor al 10% sugiere un desempeño aceptable del estimador de razón-remuestreo.

La cobertura quedó 3 puntos debajo de la nominal, lo cual puede deberse a que sólo se tiene un elemento por estrato en muestra, aunque esto es un tema que requiere estudiarse a fondo. Los cálculos de las simulaciones se efectuaron en el sistema *R*.

Ejemplo 3, Estimación de varianza usando datos de pobreza de CONEVAL.

Ejemplo 3.1 Estimador errb, estimador de varianza usando promedio de varianzas dentro de réplicas

En este ejemplo se estimarán los porcentajes de pobreza nacional publicado por el CONEVAL para el 2010 y 2012, así como estimaciones de varianza de dichas cantidades usando el estimador tipo b. La información que se emplea se encuentra en la tabla A1 del Anexo 1 y se trata de las mediciones de pobreza realizadas por el CONEVAL para el 2010 y 2012 por entidad federativa. Con base en la información de pobreza por entidad, se estimará el porcentaje nacional de pobreza para 2010 y 2012, así como una estimación de varianza usando la estimación tipo b.

Se requieren los tamaños relativos W_h, H=32, los cuales se obtuvieron de información publicada por CONAPO para mediados de 2010 y 2012, véase la tabla A2 del Anexo 1. Estos tamaños relativos se calculan para 2010 y 2012 como el cociente de la población de cada entidad federativa entre el total de población del año en cuestión. Para cada año, se calculó el estimador del porcentaje de pobreza usando la fórmula (4), ya que los estimadores puntuales del promedio coinciden para el tipo a y b. Los estimadores puntuales del porcentaje de pobreza, error estándar, así como los límites inferior y superior de intervalos de confianza al 90% se muestran a continuación. Las cantidades entre paréntesis corresponden a las estimaciones de los Estados Unidos Mexicanos publicadas por CONEVAL para 2010 y 2012, véase tabla A1 del Anexo 1.

Tabla 6
Resultados estimación puntual y de varianza de pobreza
CONEVAL, 2010 y 2012
(Cantidades en %)

Cantidad	*Estimaciones*	
	2010	*2012*
Intervalo de confianza límite inferior	45.24	44.70
Porcentaje de pobreza	46.10 (46.11)	45.48 (45.48)
Intervalo de confianza límite superior	46.95	46.25
Error estándar	0.5162	0.4728

De esta tabla se aprecia que las estimaciones puntuales del porcentaje de pobreza son prácticamente iguales a las estimadas por el CONEVAL y el error estándar no refleja demasiada variación alrededor del estimador de porcentaje. Por otra parte, no es posible comparar el error estándar obtenido con la información del CONEVAL, ya que dicha institución no publicó, al menos en la tabla A1 del anexo 1, la información de los errores estándar obtenidos.

Ejemplo 3.2 Estimador de varianza usando estratos colapsados

En la sección 3.1 se mencionó que para la estimación de varianza usando estratos colapsados se requiere el uso de una variable a nivel estrato que esté relacionada con la variable por estimar, en este caso, el porcentaje de pobreza. Como en la sección 3.2, sea u_r, $r \in \{1,2,\cdots,H\}$ una variable conocida para todos los estratos y que se emplea para colapsar

los estratos en pares. Es importante notar que una vez definidos los *H/2* estratos colapsados con base en la variable u_r, ésta no se utiliza en la estimación, a menos que se trate de las W_h. Debido a que se podían tener muchas variables auxiliares para colapsar, no se usó alguna variable en particular, sino que se permutaron los *H=32* estratos y se colapsaron por pares como aparecieron en la permutación. Una vez permutados, se estimó la varianza por el método de estratos colapsados usando la formula (1). Este procedimiento de permutación se realizó 100,000 veces para cada año, obteniéndose los valores mínimo y máximo de las varianzas estimadas de las 100,000 permutaciones, los cuales se muestran más abajo.

Más abajo, también se encuentran dos diagramas de caja y brazos correspondientes a las raíces cuadradas de las estimaciones de varianza de las simulaciones.

Antes de mostrar los valores del resultado de las simulaciones se mostrarán unas cuantas permutaciones para aclarar el procedimiento.

Tabla 7
Ejemplo de tres posibles permutaciones y estratos colapsados que les corresponden

Estrato colapsado	**1**	**1**	**2**	**2**	**3**	**3**	**4**	**4**	**5**	**5**	**6**	**6**	**7**	**7**	**8**	**8**
Orden original	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Permutación 1	6	23	30	14	4	10	15	5	8	26	16	21	25	7	11	24
Permutación 2	4	19	29	5	13	11	20	31	9	28	6	17	24	30	1	21

Estrato colapsado	**9**	**9**	**10**	**10**	**11**	**11**	**12**	**12**	**13**	**13**	**14**	**14**	**15**	**15**	**16**	**16**
Orden original	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Permutación 1	12	9	29	18	28	22	1	32	3	2	19	17	27	31	13	20
Permutación 2	25	27	3	14	12	22	8	26	23	7	10	18	32	15	16	2

En esta tabla, que se partió en dos con fines de presentación, se aprecian tres permutaciones, entre muchas que hay, siendo el orden original de los estratos (alfabético de la entidad federativa) una de ellas. La permutación dos por ejemplo, colapsaría a la entidad federativa 4 y 19 en el estrato colapsado número 1, lo cual corresponde a los estados de Campeche y Nuevo León, según la tabla A1 del Anexo 1. El resto de los datos se lee de la misma manera. Nótese que si en la permutación dos el 19 apareciera antes que el 4, dejando fijo el resto de los números de la permutación, esto no alteraría la asignación de ambas entidades al estrato colapsado 1, ni modificaría el resultado de la varianza estimada.

El objetivo de estas permutaciones es el de aproximarse al valor más pequeño de la varianza estimada por el método de estratos colapsados, ya que cualquier selección de variable auxiliar u_r para colapsar estratos, se traduce en una permutación como las de la tabla 7.

Tabla 8
Resultados de estimación de varianza, estratos colapsados
con 100,000 permutaciones
Estimación de porcentaje de pobreza
CONEVAL, 2010 y 2012
(Cantidades en porcentaje, excepto varianza estimada)

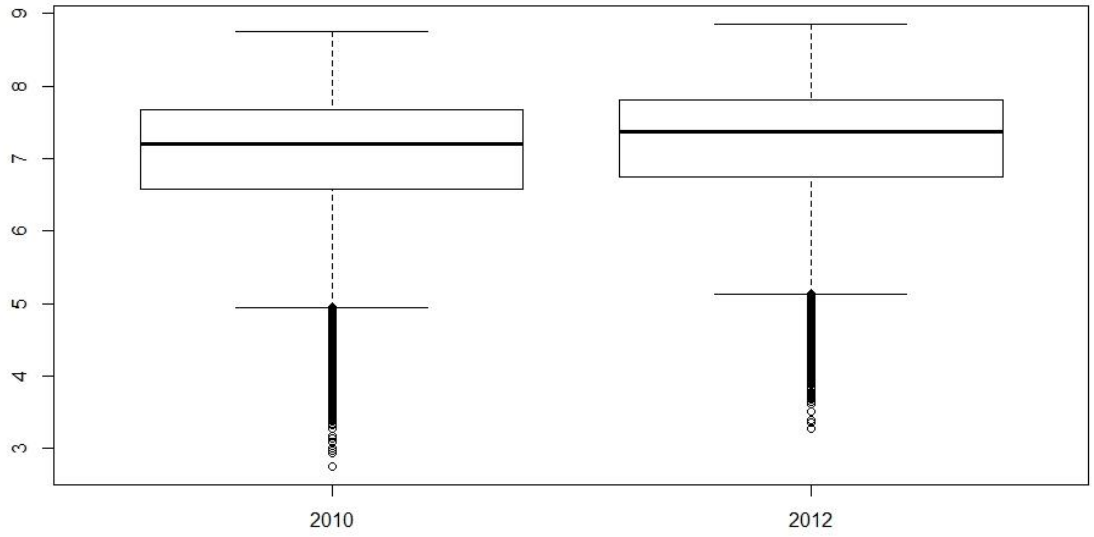
Cantidad	*Estimaciones*	
	2010	*2012*
Varianza mínima	7.58	10.70
Error estándar mínimo	2.75	3.27
Coeficiente de variación	16.4	26.5
Porcentaje de pobreza	46.11	45.48

Comparando los errores estándar mínimos de la tabla 8 con los de la tabla 6 para 2010 y 2012, se tiene que los valores mínimos observados en las permutaciones con el método de estratos colapsados son 5.3 y 6.9 veces más grandes que los obtenidos con el estimador de razón-remuestreo; por lo cual, en este caso, es preferible usar el estimador de varianza de razón-remuestreo.

Es importante notar que esto no es un resultado general y podría ser diferente con otro conjunto de datos; sin embargo, siempre es posible comparar el estimador de varianza de razón-remuestreo con el de estratos colapsados a través de las permutaciones para una muestra dada.

A continuación se muestran dos diagramas de caja y brazos en los que se aprecia, con base en las 100,000 permutaciones por año, una aproximación al rango de valores posibles del error estándar del método de estratos colapsados. Recordemos que en este tipo de diagramas, los límites inferior y superior de la caja corresponden a los percentiles 25 y 75 respectivamente. La línea superior, paralela al eje horizontal de la caja, se construye añadiendo 1.5 veces el rango intercuartílico al límite superior de la caja, en tanto que la línea inferior, paralela al eje horizontal de la caja, se construye de manera similar y solo se resta del límite inferior de la caja.

Gráfica 1
Aproximación al rango de valores posibles para el error estándar
con el método de estratos colapsados
(100,000 permutaciones de estratos por año)



En ambos diagramas se aprecia que los errores estándar pueden llegar a ser bastante grandes en relación con los valores estimados de porcentaje de pobreza, 46.11% y 45.48% para 2010 y 2012 respectivamente. Este tipo de cálculos nos da una guía para comparar la eficiencia relativa al estimar la varianza con el método de estratos colapsados al usar una variable auxiliar u_r, con el valor mínimo estimado de varianza obtenido con un número grande de simulaciones.

Recuérdese que esto puede hacerse así ya que, como se mencionó anteriormente, cualquier selección de variable auxiliar u_r para colapsar estratos, se traduce en una permutación como las de la tabla 7.

Ejemplo 3.3 El estimador no paramétrico propuesto por Breidt et al. (2014) y problemática encontrada

Al tratar de aplicar este método usando las W_h se obtuvieron algunos estratos colapsados de tamaño uno; por lo cual, algunas $K(\cdot) = 0$, impidiendo calcular el estimador no paramétrico de varianza. Esto sucedió por la asimetría positiva alta de las W_h debido a los valores grandes de tamaño relativo del Estado de México, 13.6% y 13.8% para 2010 y 2012, respectivamente. De hecho, en diversos cálculos, no se encontraron valores h_b que satisficieran simultáneamente la condición $|x_j(r)| \leq 1$ y que todos los estratos tuvieran un tamaño mayor o igual que dos. Es claro que pueden elegirse otras variables para las cuales se conjetura que existe una correlación positiva alta con la variable por estimar, en este caso, el porcentaje de pobreza. Sin embargo, no se abundará más sobre este tema, ya que no es el motivo de investigación del presente documento, aunque es necesario mencionar que Breidt el al. (2014) no hacen mención alguna a la determinación de h_b en los ejemplos que desarrollan en su artículo. Dichos autores tampoco mencionan los pasos por seguir para determinar el valor de h_b cuando se tengan variables auxiliares con una observación atípica.

Ejemplo 4, estimador errc, de razón-remuestreo para dominios no planeados, ENOE 2012

Se emplearán los mismos datos de la encuesta correspondiente al último trimestre del 2012 de la ENOE del ejemplo 2, pero el objetivo es estimar el 'promedio de ingreso por hora trabajada para la población ocupada' para la población entre 14 y 98 años, así como

construir intervalos de confianza al 90%. En el ejemplo 2 se había considerado a la muestra como una población, ahora se usará la muestra para estimar una variable en un dominio no planeado. Esto variable constituye un dominio no planeado, ya que se desconoce, en el momento de seleccionar la muestra, sí el entrevistado tuvo un ingreso mayor que cero y pertenece al rango de edad deseado. Es importante mencionar que el estimador dado en (12) reproduce los datos publicados por el INEGI para esta variable.

La estimación de la varianza que el INEGI propone en su metodología, véase INEGI, Diseño muestral y bases de datos (2010), es una aproximación de varianza usando totales de unidades primarias de muestreo dentro de estratos a la fórmula de varianza del estimador de razón combinado. Esta aproximación se emplea usualmente cuando se seleccionan conglomerados con un muestreo sistemático con probabilidad proporcional a alguna medida de tamaño, véase Särndal et al. (1992). Esto sucede porque en el muestreo sistemático no existe un estimador insesgado de la varianza. Infortunadamente, no puede emplearse la fórmula dada por el INEGI para la estimación de varianza de esta variable ya que hay algunos estratos que tienen un tamaño de muestra igual a uno para varios estratos. En este caso podría emplearse el estimador de varianza dado en (1), estratos colapsados, pero tendrían que formarse los estratos colapsados previo a ver los datos de los H=884 estratos, lo cual se ve complicado al tener una buena cantidad de estratos por colapsar. En este tipo de situaciones se aprecia la utilidad del estimador de varianza de razón-remuestreo dado en (13), ya que sólo se requiere de los promedios expandidos para cada uno de los H estratos, así como de la información de la variable del denominador por estrato. Se requiere programar los puntos a-c de la sección 4.3 para obtener las estimaciones; sin embargo,

dicha programación es sencilla y tiene un grado de dificultad menor al de establecer criterios para colapsar estratos o programar los estimadores puntual y de varianza de Breidt et al. (2014).

La variable de interés 'promedio de ingreso por hora trabajada' para la población ocupada entre 14 y 98 años, es mayor que cero y se encuentra en todos los estratos. Se denotará como $\hat{y}_h = \sum_{i=1}^{n_h} \omega_{hi} y_{hi}$ al total estimado de ingreso por hora trabajada en el *h*-ésimo estrato. Las variables y_{hi} y ω_{hi} se refieren al ingreso por hora trabajada y al factor de expansión, respectivamente, de la *i*-ésima persona en el *h*-ésimo estrato. La cantidad n_h se refiere al número de personas en muestra en el *h*-ésimo estrato. Esta cantidad es uno para algunas personas en algunos estratos al formar el dominio no planeado para la variable de interés.

En la tabla 9 se muestran algunos datos de la encuesta empleada en este ejemplo. El número de viviendas en muestra es de 119,286, en tanto que el total de estratos con una unidad primaria de muestreo es de 20, es decir, un $20/884 = 2.3\%$ de los estratos.

Tabla 9

Algunos datos de ponderadores y totales expandidos

Estadística	Factores ω_h	*Valores* $\hat{y}_h$
Mínimo	119	5,395
Máximo	805,620	20'860,474
Media	39,250	1'217,356
Mediana	20,900	676,610

El empleo del estimador de razón-remuestreo para dominios no planeados en este ejemplo se calculó con 884 medias usando la fórmula (12) y se tiene que:

$$\hat{y}_{errc} = \frac{1'076{,}142{,}579}{34'697{,}369} = 31.015 \cdot$$

El coeficiente de variación para las 884 réplicas de estimadores del total en el denominador fue de 0.19%, el cual es pequeño, y el sesgo es despreciable. Como se hizo notar en el tercer párrafo de la sección 2.2., el sesgo en un estimador de razón será despreciable si el coeficiente de variación de la variable auxiliar, en este caso el total estimado del denominador, es menor al 10%. Para la estimación de varianza se usó la fórmula (13) y se obtuvo un valor de $\hat{v}(\hat{y}_{errc}) = 0.441$. Los límites al 90% suponiendo normalidad son $(29.923,\ 32.108)$. Es importante recalcar que el promedio estimado de 31.015 coincide con la cifra publicada por el INEGI para dicha variable.

6. CONCLUSIONES

En este trabajo se presentó un estimador de razón-remuestreo que permite realizar estimaciones de varianza para diseños estratificados que tengan una o muy pocas unidades por estrato en muestra. El estimador es del tipo de razón, el cual es sesgado, y se trabaja con el estimador de varianza de este tipo de estimadores, usando réplicas de medias estimadas por estrato tipo jackknife. Es adecuado su uso en el caso de contar con los tamaños relativos de los estratos y estimadores insesgados de las medias en cada estrato. En este caso el sesgo

se estima con los datos muestrales. También se aplica para el caso de estimaciones en dominios no planeados en los que se tengan observaciones en todos los estratos. Se ilustró con ejemplos la facilidad relativa con la que se obtienen estimaciones, comparado con otros métodos, en el caso de encuestas complejas. Asimismo, se están investigando si existen condiciones en las que el estimador de varianza propio del método de razón-remuestreo, véase fórmula siete, es mejor que el estimador de varianza del muestreo aleatorio estratificado.

Bibliografía

Breidt, F. J., Opsomer, J.D. & Sánchez-Borrego, I. (2014), *Nonparametric variance estimation under fine stratification: an alternative to collapsed strata*. Journal of the American Statistical Association. Forthcoming.

Canadian Health Measures Surveys, 2007, Statistics Canada.

Cochran, W., *Técnicas de Muestreo*, Ed. CECSA, México, 1986.

Fuller, W. A., *Sampling with Random Stratum Boundaries*, Journal of the Royal Statistical Society, B 32, pp. 209-226, 1970.

Horvitz, D.G. & Thompson, D.J., *A generalization of sampling without replacement from a finite universe*, Journal of the American Statistical Association, Vol. 47, No. 260, (Dec. 1952), pp. 663-685.

INEGI, *Encuesta Nacional de Ocupación y Empleo 2010*. Diseño Muestral y bases de datos.

Mantel, H. & Giroux, S., *Variance estimation in complex surveys with one PSU per stratum. Joint Statistical Meetings*, Washington, D.C., USA, August 1-6, 2009.

Mini-Finland Health Examinations Survey 1977, Rehabilitation Research Centre & Research Institute for Social Security of the Social Insurance Institution.

Norlén, U. & Waller, T., *Estimation in a complex survey-experiences from a survey of buildings with regard to energy usage*, Statistik Tidskrift, 17, pp. 109-124, 1979.

Quenouille, H. H., *Problems in plane sampling*, Ann. Math. Stat., 20, pp. 355-375, 1949.

Quenouille, H. H., *Notes on bias in estimation*, Biometrika, 43, pp. 353-360, 1956.

R Development Core Team. 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, 2010.

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

Wolter, K. M., *Introduction to variance estimation*, Springer-Verlag, New York, 1985.

Anexo 1

Tabla A1
Resultados de mediciones de pobreza para México por entidad federativa
CONEVAL, 2010 y 2012

CONEVAL
Consejo Nacional de Evaluación de la Política de Desarrollo Social

Medición de la pobreza, Estados Unidos Mexicanos, 2012
Evolución de la pobreza y pobreza extrema nacional y en entidades federativas, 2010-2012

Entidad federativa	Pobreza						Pobreza extrema					
	Porcentaje		Miles de personas		Cambios en el número de personas		Porcentaje		Miles de personas		Cambios en el número de personas	
	2010	2012	2010	2012	Porcentual	Absoluto (Miles de personas)	2010	2012	2010	2012	Porcentual	Absoluto (Miles de personas)
Aguascalientes	38.1	37.8	456.8	467.6	2.3	10.7	3.8	3.4	45.1	42.0	-7.0	-3.1
Baja California	31.5	30.2	1,019.8	1,010.1	-0.9	-9.7	3.4	2.7	109.1	91.5	-16.1	-17.6
Baja California Sur	31.0	30.1	203.0	211.3	4.1	8.3	4.6	3.7	30.3	25.8	-14.8	-4.5
Campeche	50.5	44.7	425.3	387.9	-8.8	-37.4 *	13.8	10.4	116.1	90.7	-21.8 *	-25.3
Coahuila	27.8	27.9	775.9	799.3	3.0	23.3	2.9	3.2	81.9	92.7	13.1	10.8
Colima	34.7	34.4	230.3	237.2	3.0	6.9	2.5	4.0	16.7	27.4	63.9	10.7
Chiapas	78.5	74.7	3,866.3	3,782.3	-2.2 *	-84.0 *	38.3	32.2	1,885.4	1,629.2	-13.6	-256.2
Chihuahua	38.8	35.3	1,371.6	1,272.7	-7.2	-98.9	6.6	3.8	231.9	136.3	-41.2 *	-95.6
Distrito Federal	28.5	28.9	2,537.2	2,565.3	1.1	28.2	2.2	2.5	192.4	219.0	13.9	26.6
Durango	51.6	50.1	864.2	858.7	-0.6	-5.5	10.5	7.5	175.5	128.0	-27.1 *	-47.5
Guanajuato	48.5	44.5	2,703.7	2,525.8	-6.6	-177.9	8.4	6.9	469.5	391.9	-16.5	-77.6
Guerrero	67.6	69.7	2,330.0	2,442.9	4.8	112.9	31.8	31.7	1,097.6	1,111.5	1.3	14.0
Hidalgo	54.7	52.8	1,477.1	1,465.9	-0.8	-11.1	13.5	10.0	364.0	276.7	-24.0	-87.3
Jalisco	37.0	39.8	2,766.7	3,051.0	10.3	284.3	5.3	5.8	392.4	446.2	13.7	53.8
México	42.9	45.3	6,712.1	7,328.7	9.2	616.7	8.6	5.8	1,341.2	945.7	-29.5 *	-395.6
Michoacán	54.7	54.4	2,424.8	2,447.7	0.9	22.9	13.5	14.4	598.0	650.3	8.8	52.4
Morelos	43.2	45.5	782.2	843.5	7.8	61.3	6.9	6.3	125.4	117.2	-6.6	-8.3
Nayarit	41.4	47.6	461.2	553.5	20.0	92.3 *	8.3	11.9	92.7	138.7	49.6	46.0
Nuevo León	21.0	23.2	994.4	1,132.9	13.9	138.4	1.8	2.4	86.4	117.5	36.1	31.1
Oaxaca	67.0	61.9	2,596.3	2,434.6	-6.2	-161.7	29.2	23.3	1,133.5	916.6	-19.1	-216.9
Puebla	61.5	64.5	3,616.3	3,878.1	7.2	261.9	17.0	17.6	1,001.7	1,059.1	5.7	57.3
Querétaro	41.4	36.9	767.0	707.4	-7.8	-59.6 *	7.4	5.2	137.5	98.7	-28.2 *	-38.7
Quintana Roo	34.6	38.8	471.7	563.3	19.4	91.6 *	6.4	8.4	87.5	122.2	39.5	34.6
San Luis Potosí	52.4	50.5	1,375.3	1,354.2	-1.5	-21.1	15.3	12.8	402.6	342.9	-14.8	-59.6
Sinaloa	36.7	36.3	1,048.6	1,055.6	0.7	6.9	5.5	4.5	156.3	130.2	-16.7	-26.1
Sonora	33.1	29.1	905.2	821.3	-9.3	-83.9	5.1	5.0	140.1	139.8	-0.2	-0.3
Tabasco	57.1	49.7	1,291.6	1,149.4	-11.0	-142.2 *	13.6	14.3	306.9	330.8	7.8	23.9
Tamaulipas	39.0	38.4	1,301.7	1,315.6	1.1	13.9	5.5	4.7	183.4	160.2	-12.7	-23.2
Tlaxcala	60.3	57.9	719.0	711.9	-1.0	-7.1	9.9	9.1	118.2	112.2	-5.0	-5.9
Veracruz	57.6	52.6	4,448.0	4,141.8	-6.9	-306.2	18.8	14.3	1,449.0	1,122.0	-22.6 *	-327
Yucatán	48.3	48.9	958.5	996.9	4.0	38.3	11.7	9.8	232.5	200.6	-13.7	-31.9
Zacatecas	60.2	54.2	911.5	835.5	-8.3 *	-76.0 *	10.8	7.5	164.1	115.3	-29.7 *	-48.8
Estados Unidos Mexicanos	***46.1***	***45.5***	***52,813.0***	***53,349.9***	***1.0***	***536.9***	***11.3***	***9.8***	***12,964.7***	***11,529.0***	***-11.1 ****	***-1,435.7***

* El cambio en pobreza respecto de 2010 es estadísticamente significativo con un nivel de significancia de 0.05.

Fuente: estimaciones del CONEVAL con base en el MCS-ENIGH 2010 y 2012.

Tabla A2
Número de personas por entidad federativa en México
CONAPO, 2010 y 2012
Información a mediados de año
(Valores Wh calculados por el autor con los datos de CONAPO)

Entidad federativa	2010		2012	
	Número de personas	Tamaño relativo, Wh	Número de personas	Tamaño relativo, Wh
AGUASCALIENTES	1,195,787	0.010	1,233,921	0.011
BAJA CALIFORNIA	3,224,844	0.028	3,328,623	0.028
BAJA CALIFORNIA SUR	649,616	0.006	695,409	0.006
CAMPECHE	836,748	0.007	866,375	0.007
CHIAPAS	4,903,755	0.043	5,050,568	0.043
CHIHUAHUA	3,525,273	0.031	3,598,792	0.031
COAHUILA	2,782,013	0.024	2,854,334	0.024
COLIMA	658,910	0.006	685,394	0.006
DISTRITO FEDERAL	8,944,599	0.078	8,911,665	0.076
DURANGO	1,669,815	0.015	1,709,741	0.015
GUANAJUATO	5,558,502	0.049	5,668,181	0.048
GUERRERO	3,444,264	0.030	3,499,507	0.030
HIDALGO	2,690,086	0.024	2,768,973	0.024
JALISCO	7,442,625	0.065	7,644,152	0.065
MEXICO	15,571,679	0.136	16,106,485	0.138
MICHOACAN	4,420,271	0.039	4,494,730	0.038
MORELOS	1,803,340	0.016	1,850,812	0.016
NAYARIT	1,108,860	0.010	1,155,448	0.010
NUEVO LEON	4,723,273	0.041	4,868,844	0.042
OAXACA	3,868,109	0.034	3,930,833	0.034
PUEBLA	5,863,823	0.051	6,002,161	0.051
QUERETARO	1,848,191	0.016	1,912,803	0.016
QUINTANA ROO	1,350,945	0.012	1,440,115	0.012
SAN LUIS POTOSI	2,616,459	0.023	2,675,311	0.023
SINALOA	2,851,334	0.025	2,905,750	0.025
SONORA	2,727,032	0.024	2,809,806	0.024
TABASCO	2,252,641	0.020	2,309,071	0.020
TAMAULIPAS	3,334,664	0.029	3,419,338	0.029
TLAXCALA	1,186,143	0.010	1,224,637	0.010
VERACRUZ	7,712,247	0.067	7,858,604	0.067
YUCATAN	1,980,690	0.017	2,036,694	0.017
ZACATECAS	1,509,019	0.013	1,536,674	0.013
Total	114,255,557	1.000	117,053,750	1.000

Anexo 2

Demostración de la expresión (5), sesgo del estimador para la media.

Primero obtenemos el valor esperado del estimador $\hat{y}_{erra}$, usando el hecho de que el estimador de la media por estrato es insesgado:

$$E(\hat{y}_{erra}) = \frac{1}{H}\sum_{i=1}^{H}\sum_{h=1}^{H}\frac{W_h}{1-W_i}E\left(\hat{y}_{h(-i)}\right) = \frac{1}{H}\sum_{i=1}^{H}\sum_{h=1}^{H}\frac{W_h}{1-W_i}\bar{y}_{h(-i)} \ .$$

Notando que $\sum_{h=1}^{H} W_h \bar{y}_{h(-i)} = \bar{y}_{st} - W_i \bar{y}_i$, la expresión anterior puede escribirse de la siguiente manera:

$$E\left(\hat{y}_{erra}\right) = \frac{1}{H}\sum_{i=1}^{H}\frac{\bar{y}_{st}}{1-W_i} - \frac{1}{H}\sum_{i=1}^{H}\frac{W_i}{1-W_i}\bar{y}_i \ .$$

Al restar $\bar{y}_{st}$ a la fórmula anterior, simplificar y ordenar términos se obtiene el resultado (5).

Anexo 3

A3.1 Estimadores del tipo Horvitz-Thompson.

Este tipo de estimadores fueron desarrollados por Horvitz-Thompson (1952) para estimar totales de una población finita con base en una muestra probabilística, sin emplear información adicional. Se tienen *n* valores muestrales y_k y cada uno tiene una probabilidad de selección conocida π_k, con $\pi_k > 0$. Dicho estimador tiene la siguiente forma $\hat{y} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$.

Este estimador es insesgado y en caso de que se requiera una estimación insesgada del promedio poblacional, se divide $\hat{y}$ entre *N*, siempre que esta última cantidad sea conocida. Al recíproco de las π_k se les denomina factores de expansión y tienen la propiedad: $\sum_{k=1}^{n} \frac{1}{\pi_k} = N$.

A3.2 Estimación de razón en muestreo aleatorio estratificado

Al usar el estimador de razón en el muestreo aleatorio estratificado, en el que se tengan estimaciones insesgadas del promedio por estrato, se pueden construir dos expresiones para el estimador de razón de la población con base en una misma muestra estratificada dependiendo de la manera en la que se construyan las estimaciones:

a) Estimador de razón separado. Se obtiene un estimador de razón por estrato, $\hat{r}_h = \frac{\hat{y}_h}{\hat{x}_h}$, y después se emplea el siguiente estimador de la media poblacional: $\hat{r}_s = \sum_{h=1}^{H} W_h x_{Uh} \hat{r}_h$, donde x_{Uh} es el total de la variable auxiliar *x* en el h-ésimo

estrato. La estimación de varianza tiene la siguiente forma, Cochran (1986):

$$\hat{v}(\hat{r}_s) = \frac{1}{x_U^2}\sum_{h=1}^{H} W_h^2(1-f_h)\frac{1}{n_h}\frac{\sum_{i=1}^{n_h}(y_{hi}-\hat{r}_h x_{hi})^2}{n_h-1}.$$

En esta fórmula, $f_h = \frac{n_h}{N_h}$ y $x_U = \sum_{h=1}^{H}\frac{x_h}{N_h}$. Nótese que se requiere conocer el total de la variable auxiliar por estrato, situación que no siempre se puede garantizar en la práctica.

b) Estimador de razón combinado. Con las $\hat{y}_h$ y $\hat{x}_h$ se construye un estimador de la forma, $\hat{r}_c = \frac{\hat{y}_{st}}{\hat{x}_{st}}$, con $\hat{y}_{st} = \sum_{h=1}^{H} W_h \hat{y}_h$ y $\hat{x}_{st}$ calculado de manera análoga. La estimación de varianza tiene la siguiente forma, Cochran (1986):

$$\hat{v}(\hat{r}_c) = \frac{1}{x_U^2}\sum_{h=1}^{H} W_h^2(1-f_h)\frac{1}{n_h}\frac{\sum_{i=1}^{n_h}(y_{hi}-\hat{r}_c x_{hi})^2}{n_h-1}.$$

Obsérvese que en el estimador de razón combinado no se requiere conocer el total de la variable auxiliar por estrato, sólo poblacional.

En general, es preferible usar el estimador de razón combinado al separado cuando se tengan tamaños de muestra pequeños por estrato, menores que 30, ya que en el estimador de razón separado el sesgo se añade a través de los estratos, véase Cochran (1986) o Särndal et al. (1992).