Capítulo 3

Métodos de estimación de la varianza surgidos en el marco de población infinita

En este capítulo se presentan los métodos de cálculo de varianza que surgieron en el marco de población infinita, los cuales son: *linearización, jackknife y bootstrap.* La linealización por series de Taylor no se basa en el remuestreo pero ha sido una técnica muy difundida desde mediados de los setenta, ya que se puede aplicar a través de varios paquetes de cómputo. Por otra parte, las investigaciones sobre la aplicación del *jackknife* y del *bootstrap* en muestreos complejos se comienzan a dar a mediados de los ochenta; y no es hasta los noventa que algunas instituciones adoptan alguno de estos dos métodos formalmente. La referencia cronológica explica el por qué las investigaciones dirigidas a la comparación de métodos de cálculo de varianza (que se encontraron en la bibliografía), exploran las nuevas técnicas contrastándolas con la linearización. Por tal motivo, se consideró necesario presentar su desarrollo, junto con el *jackknife* y el *bootstrap*, los cuales son métodos de remuestreo que están cobrando auge en la solución de problemas de muestreos complejos, gracias al gran desarrollo de la computación.

3.1 Linealización

El método de linealización por series de Taylor es quizás el más utilizado en aplicaciones con encuestas de tamaño considerable, donde se hace necesario el uso de paquetes computarizados de difusión comercial. Por ejemplo, el Instituto Nacional de Estadística, Geografía e Informática (INEGI), reporta en la Encuesta Nacional de la Dinámica Demográfica, 1992 (ENAID) que obtuvo las varianzas mediante el programa *clusters*, el cual usa este método.

El método consiste en expresar un parámetro no-lineal, Θ, como una función de medias de otras variables. La expansión por series de Taylor provee una aproximación lineal de la estadística de interés. De tal forma, se busca la varianza de esa aproximación, lo que representa un estimador sesgado de la estadística no-lineal.

Entre las primeras referencias encontradas respecto al uso de series de Taylor para estimar la varianza de una estadística, se hallan dos trabajos: Woodruff (1971) y Woodruff y Causey (1976). En el primero se advierte la utilidad del método en el caso de muestras complejas, y en el segundo, se describe un programa en el que se ha desarrollado la aproximación por series de Taylor para calcular errores de muestreo.

3.1.1 Caracterización del estimador

Cuando el parámetro de interés es de la forma $\Theta = f(\mathbf{Y})$, donde $\boldsymbol{Y}$ es un vector p-variado[1] de parámetros de la población, $\boldsymbol{Y} = (\boldsymbol{Y_1}, \boldsymbol{Y_2}, ..., \boldsymbol{Y_p})$, resulta natural elegir el estimador $\hat{\Theta} = f(\hat{\mathbf{Y}})$. Sin embargo, la varianza de tal estimador no se obtiene directamente, en la mayoría de las ocasiones.

Bajo el supuesto de que $f(\hat{\mathbf{Y}})$ posee derivadas de segundo orden en una hiperesfera que contiene a $\hat{\mathbf{Y}}$ y a $\mathbf{Y}$, entonces, al aplicar la expansión de Taylor se tiene:

$$f(\hat{\mathbf{Y}}) = f(\mathbf{Y}) + \sum_{j=1}^{p} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j) + R_{n_\mathbf{S}}(\hat{\mathbf{Y}},\mathbf{Y}) \tag{3.1}$$

$R_{n_\mathbf{S}}(\widehat{\mathbf{Y}},\mathbf{Y})$ es el residuo[2], al desechar los términos de segundo orden y superiores, el cual depende del tamaño de muestra y está dado en función de los parámetros poblacionales y los estimadores.

El error cuadrático medio del estimador $f(\widehat{\mathbf{Y}})$ está dado por:

$$ECM(f(\hat{\mathbf{Y}})) = E\{(f(\widehat{\mathbf{Y}}) - f(\mathbf{Y}))^2\}$$

Ya se comentó en el Capítulo 1 que en ocasiones se considera una aproximación del ECM como la aproximación de la varianza, debido a que el $Sesgo^2(\hat{\Theta})$ es de orden menor a $Var(\hat{\Theta})$. Por lo tanto:

$$\begin{aligned} ECM(f(\hat{\mathbf{Y}})) &\approx Var\{f(\widehat{\mathbf{Y}})\} \\ &= Var\{\sum_{j=1}^{p} \frac{\partial f(\mathbf{Y})}{\partial y_j}(\hat{Y}_j - Y_j)\} \\ &= \sum_{j=1}^{p}\sum_{i=1}^{p} \frac{\partial f(\mathbf{Y})}{\partial y_j}\frac{\partial f(\mathbf{Y})}{\partial y_i} Cov(\hat{Y}_j,\hat{Y}_i)\ . \end{aligned}$$

Así pues, al considerar la matriz de varianza-covarianza de $\hat{\mathbf{Y}}$, $\xi_{n_{(\mathbf{S})}}$, y el vector $1 \times p$ de derivadas, d, cuyos elementos son de la forma $d_j = \frac{\partial f(\mathbf{Y})}{\partial y_j}$, se llega a:

[1] Nótese que $\boldsymbol{Y}$ es una función p-variada de $y = (y_1, ..., y_N)$, lo que de acuerdo a la notación del Capítulo 1 representa un vector de valores en la población de tamaño N. Pero ahora se considera $y_i = [y_{i1}, y_{i2}, ..., y_{ip}]$, es decir, la función de interés es función de varias variables en la población.

[2] $n_\mathbf{S}$ es el tamaño de la muestra completa, como se definió en el Capítulo 1.

$$ECM(f(\hat{\mathbf{Y}})) \approx d\xi_{n_{(\mathbf{S})}} d' \tag{3.2}$$

Por otra parte, el vector de derivadas se consigue evaluando en los valores muestrales de la siguiente forma:

$$\hat{d}_j = \frac{\partial f(\widehat{\mathbf{Y}})}{\partial y_j}$$

Sin embargo, Woodruff (1971) advirtió que si las estadísticas $\hat{Y}_j$ están dadas por una combinación lineal de observaciones,

$$\hat{Y}_j = \sum_{i=1}^{n_{\mathbf{S}}} w_i y_{ij} \text{ donde j=1,...,p.}$$

entonces el cálculo se simplifica a un problema univariado. El razonamiento de Woodruff se muestra a continuación:

$$\begin{aligned} ECM(\hat{\Theta}) &= Var[\sum_{j=1}^{p} \frac{\partial f(\mathbf{Y})}{\partial y_j} \hat{Y}_j] \\ &= Var[\frac{\partial f(\mathbf{Y})}{\partial y_j} \sum_{i=1}^{n_{\mathbf{S}}} w_i y_{ij}] \\ &= Var(\sum_{i=1}^{n_{\mathbf{S}}} w_i \sum_{j=1}^{p} \frac{\partial f(\mathbf{Y})}{\partial y_j} y_{ij}) \\ &= Var(\sum_{i=1}^{n_{\mathbf{S}}} w_i v_i) \text{ donde, } v_i = \frac{\partial f(\mathbf{Y})}{\partial y_j} y_{ij} \end{aligned}$$

La estimación de esas v_i se realiza como se mostró antes en el caso del vector d. Es decir,

$$\hat{v}_i = \frac{\partial f(\widehat{\mathbf{Y}})}{\partial y_j} y_{ij}$$

3.1.2 Consideraciones prácticas

Se observó que se puede llegar a la expresión de la varianza de una estadística que es combinación lineal de las observaciones, ponderadas por la derivada de la función original, evaluada en valores muestrales obtenidos de la misma muestra. Lo cual implica, que se debe requerir un estimador insesgado de $Var(y_{ij})$.

Muchas veces, se puede aplicar este método para calcular varianzas casi a ciegas, si se tiene acceso a programas que desarrollen esta técnica. Sin embargo, aún para estos programas pueden haber ciertas dificultades al buscar las derivadas en cuestión. Ejemplos de situaciones difíciles lo son los coeficientes de correlación múltiple y parcial. En tales casos pueden ser útiles los métodos descritos en Woodruff y Casey (1976).

Por otra parte, hay que tener presente que los términos en $\xi_{n(\mathbf{s})}$, o bien, la expresión para $Var(\sum_{i=1}^{n_\mathbf{s}} w_i v_i)$, dependerán del diseño de muestreo. Wolter (1985), advierte que es posible que para calcularla, sea necesario aplicar algún otro método para la estimación de la varianza.

Es útil mostrar un ejemplo de la aplicación de esta técnica. Sea $\hat{R}$ el estimador de razón entre dos variables, tal que $\hat{R} = \frac{\overline{y}}{\overline{x}}$. De acuerdo a (3.2), se debe determinar el vector de dos dimensiones d, y la matriz $\xi_{n(\mathbf{s})}$, los cuales son:

$$\hat{\xi}_{n(\mathbf{s})} = \begin{bmatrix} \widehat{Var}(\overline{x}) & \widehat{Cov}(\overline{x},\overline{y}) \\ \widehat{Cov}(\overline{x},\overline{y}) & \widehat{Var}(\overline{y}) \end{bmatrix}$$

$$d = \begin{bmatrix} -\frac{\overline{y}}{\overline{x}^2} \\ \\ \frac{1}{\overline{x}} \end{bmatrix}$$

Por lo tanto, el estimador de la varianza de una razón, de acuerdo al método de linealización, está dado por:

$$\widehat{Var}(\hat{R}) = \left(\frac{\widehat{Var}(\overline{x})}{\overline{x}^2}\hat{R}^2 + \frac{\widehat{Var}(\overline{y})}{\overline{x}^2} - 2\frac{\widehat{Cov}(\overline{x},\overline{y})}{\overline{x}^2}\hat{R} \right) \tag{3.3}$$

Una forma alternativa, mas simple, para su cálculo, es

$$\widehat{Var}(\hat{R}) = \frac{1}{\overline{x}^2}\widehat{Var}(\overline{y} - \hat{R}\overline{x}) \tag{3.4}$$

Se puede verificar en Cochran (1977), y otros autores que esta aproximación de la varianza obtenida por linealización, es el estimador que usualmente se proporciona por fórmula. Cuando se conoce $\overline{X}$, el valor poblacional, se usa $\overline{X}^2$ en lugar de $\overline{x}^2$ en (3.3). Es evidente que cada uno de los elementos de $\hat{\xi}_{n(\mathbf{s})}$ se tiene que estimar de acuerdo al diseño muestral, por lo que la evaluación de esta expresión podría hacerse bastante pesada. Cabe señalar que si se conoce el valor poblacional, $\overline{X}$, se usa $\overline{X}^2$ en lugar de $\overline{x}^2$ en (3.3).

3.2 Jackknife

Hasta hace unos años, el único método para cálculo de varianza en muestreos complejos que aplicaban muchas instituciones era el de grupos aleatorios dependientes, o bien la linealización. El creciente desarrollo de los equipos de cómputo ha facilitado la exploración de métodos de intenso trabajo computacional. Tal es el caso del *jackknife*, el cual, cada vez está siendo adoptado por más analistas de encuestas en el mundo, en parte porque es más simple de aplicar que el *bootstrap*, además de tener cierto apoyo teórico y de que los estudios empíricos revelan, en general, un buen comportamiento.

3.2.1 Caracterización del estimador

La técnica a la que se ha denominado por *jackknife* partió de un estimador del coeficiente de correlación serial, con corrección del sesgo, el cual fue creado por Quenouille (1949) Años más tarde, Tukey (1958), retomó el trabajo de Quenouille y propuso un estimador de la varianza y el sesgo de una estadística. De manera general, el *jackknife* es un método no-paramétrico basado en remuestreo. Sus inicios se dieron en el contexto de poblaciones infinitas aunque más tarde fue aplicado al caso de poblaciones finitas.

Es posible revisar la lógica tras la idea de Quenouille para estimar el sesgo, y posteriormente ver cómo fue aprovechada por Tukey para conseguir un estimador de la varianza. En primer lugar, Quenouille (1949) advierte que para una estadística t_n, basada en una muestra de tamaño n, que se pueda expandir a través de una serie de Taylor, que sea consistente, con cumulantes finitos, se cumple que:

$$\mathrm{E}(t_n - \Theta) = \frac{a_1}{n} + \frac{a_2}{n^2} + \frac{a_3}{n^3} +$$

Igualmente, si se calculara la misma estadística, pero con $n-1$ elementos, se tendría que:

$$\mathrm{E}(t_{n-1} - \Theta) = \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \frac{a_3}{(n-1)^3} + \ldots$$

Es claro que bajo los supuestos indicados, el sesgo disminuye, a medida que n crece. Efron (1982), da una explicación gráfica al planteamiento de Quenouille. (Para entenderla mejor, conviene ver la Figura 3.1). Sea E_n la esperanza del estimador t_n, basado en un tamaño de muestra n, entonces la gráfica de E_n vs $\frac{1}{n}$ estaría dada por una curva creciente y cóncava, que parte del punto $(0, \Theta)$, el cual correspondería al caso $n = \infty$. Aproximando tal curva con una línea, se tendría que:

$$\frac{E_n - E_\infty}{E_{n-1} - E_n} = \frac{1/n}{1/(n-1) - 1/n}$$

Se entiende por E_∞, el verdadero valor del parámetro, o sea Θ. De ahí que el sesgo del estimador basado en n elementos, está dado por $E_n - E_\infty$, lo cual, resolviendo la relación anterior, resulta en

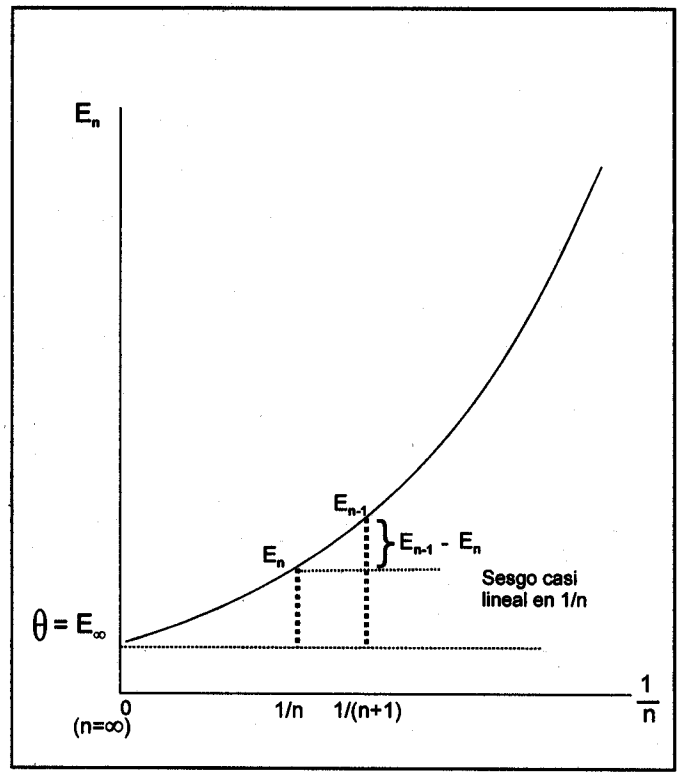
$$Sesgo = E_n - E_\infty = (n-1)(E_{(n-1)} - E_n)$$

De la misma relación se obtiene

$$\Theta = E_\infty = nE_n - (n-1)E_{(n-1)}$$

La propuesta de Quenouille es obtener, a partir de la muestra de tamaño n, un estimador θ, basado en los n elementos, junto con n estimadores basados en $(n-1)$

Figura 3.1: Gráfica tomada de Efron (1982), que explica la lógica tras el *Jackknife* original



Nota: E_n representa la esperanza del estimador basado en n datos.

elementos cada uno, denominados por, $\hat{\Theta}_{(i)}$. Así pues, se calcula el promedio de los últimos:

$$\hat{\Theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\Theta}_{(i)}. \tag{3.5}$$

El estimador del sesgo está dado por:

$$\widehat{Sesgo} \equiv (n-1)(\hat{\Theta}_{(.)} - \hat{\Theta}).$$

También se deduce un nuevo estimador con corrección del sesgo:

$$\tilde{\Theta} = \hat{\Theta} - \widehat{Sesgo} = n\hat{\Theta} - (n-1)\hat{\Theta}_{(.)}$$

Se puede verificar que si bien $\hat{\Theta}$ tiene un sesgo del orden de $1/n$, el sesgo del nuevo estimador es del orden $1/n^2$

Basado en el planteamiento de Quenouille, Tukey (1958), sugirió calcular n *pseudovalores*, denominados $\tilde{\Theta}_i$, los cuales se construyen a partir del estimador que se consigue al dejar el dato i fuera de la muestra. Como se ve a continuación el pseudovalor de Tukey corresponde directamente al estimador con corrección del sesgo de Quenouille.

$$\tilde{\Theta}_i = \hat{\Theta} + (n-1)(\hat{\Theta} - \hat{\Theta}_{(i)}) = n\hat{\Theta} - (n-1)\hat{\Theta}_{(i)} \tag{3.6}$$

El estimador *jackknife* de Θ, estaría dado por:

$$\tilde{\bar{\Theta}}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{\Theta} \tag{3.7}$$

La estimación de la varianza se basa en las desviaciones de los pseudovalores con respecto a su media, de la siguiente forma:

$$\widehat{VAR}_{J1} = \frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(\tilde{\Theta}_i - \tilde{\bar{\Theta}}_J)^2 \tag{3.8}$$

Es más común asociar el *jackknife* directamente a los estimadores $\hat{\Theta}_{(i)}$ (i=1,...,n), quizás porque es más conocida la aplicación del método a problemas de regresión, donde no se involucran pseudovalores como los que propone Tukey para un problema de muestreo. Así pues, se puede ver que (3.8) es equivalente a:

$$\widehat{VAR}_{J1} = \frac{(n-1)}{n}\sum_{i=1}^{n}(\hat{\Theta}_{(i)} - \hat{\Theta}_{(.)})^2 \tag{3.9}$$

Por lo tanto, es factible obtener el estimador de la varianza *jackknife*, sin necesidad de calcular los pseudovalores de Tukey. Por otra parte, en la práctica, $\widehat{VAR}_{J1}$ no sólo se utiliza para estimar la varianza de $\tilde{\bar{\Theta}}$ sino también la de $\hat{\Theta}$. De hecho, en ocasiones, también se usa el estimador:

$$\widehat{VAR}_{J2} = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\Theta}_{(i)} - \hat{\Theta})^2 \tag{3.10}$$

3.2.2 Aplicación del *jackknife* en muestreo

Se ha explicado la naturaleza del estimador *jackknife*, pero lo que nos ocupa es saber cómo se aplicaría este método en el caso de una encuesta. Hasta ahora, lo planteado se asocia a la situación de un muestreo aleatorio simple, donde se calculan n estimadores al eliminarse cada dato. La lista de pasos que se presenta a continuación se refiere a la eliminación de un *grupo*, en lugar de un dato. De esta forma, se entiende que la mecánica que se presenta, tomada de Wolter (1985), es el primer intento de plantear la aplicación del *jackknife* en una manera generalizada para un muestreo complejo. Lo primero que salta a la vista es que se supone que se cuenta con k grupos de m elementos cada uno; lo cual acarrería el problema de formar esos grupos, o de determinar el tamaño de k. En realidad, la mecánica que se expone a continuación no es la mejor para un muestreo complejo, pero se exhibe porque representa una versión que en ocasiones es de utilidad, incluso en problemas de regresión.

De manera general, la estimación por *jackknife* consiste en los siguientes pasos:

1. Se calcula un estimador $\hat{\Theta}$ de Θ basado en toda la muestra y respetando el diseño.
2. La muestra es particionada en k grupos con m observaciones cada uno. Es decir, se supone n múltiplo de k ($n = mk$).
3. Se calculan k estimadores, con la misma forma de $\hat{\Theta}$ pero omitiendo en cada ocasión uno de los grupos. Es decir, se obtienen k estimaciones con $m(k-1)$ unidades a las que se denominan mediante $\hat{\Theta}_{(i)}$ (donde $i = 1, ..., k$).
4. Se calcula, para cada grupo, su pseudovalor denominado por: $\tilde{\Theta}_i$, de la siguiente forma.
$$\tilde{\Theta}_i = k\hat{\Theta} - (k-1)\hat{\Theta}_{(i)}$$
5. Se obtiene el estimador *jackknife* promediando los pseudovalores, como en (3.7).
$$\tilde{\bar{\Theta}}_J = \frac{\sum_{i=1}^{k} \tilde{\Theta}_i}{k}$$

Así pues:

$$\begin{aligned}
\tilde{\bar{\Theta}}_J &= k\hat{\Theta} - (k-1)\frac{\sum_{(i)=1}^{k} \hat{\Theta}_{(i)}}{k} \\
&= \sum_{(i)=1}^{k} \hat{\Theta}_{(i)}\left(\frac{1}{k} - 1\right) + \sum^{k} \hat{\Theta} \\
&= \frac{1}{k}\sum_{(i)=1}^{k} \hat{\Theta}_i - \sum_{i=1}^{k}(\hat{\Theta}_{(i)} - \hat{\Theta})
\end{aligned}$$

Vemos que $\tilde{\bar{\Theta}}_J$ no es simplemente el promedio de las estimaciones de k *grupos*, sino que se considera un término que aporta información sobre la desviación simple de las estimaciones de cada grupo respecto al estimador general.

Wolter (1985), indica que cuando se tiene un diseño por conglomerados, se deben considerar como los grupos a eliminar, los conglomerados primarios. Es decir, aquellos conglomerados que están constituidos por unidades de segunda, tercera etapa, hasta las unidades elementales.

Resulta necesario mostrar el mecanismo a seguir para aplicar el *jackknife* bajo un diseño más complicado. Nuevamente, Wolter (1985), es la primera referencia explícita que considera un diseño más complicado. Cabe señalar que, según su planteamiento, los pseudovalores son ajustados para que el estimador *jackknife* sea igual al que resulta en el caso de una estadística lineal. Además, es importante añadir las siguientes advertencias, que son válidas para cualquier diseño estratificado, sea simple, o de varias etapas:

- Si se desea estimar la varianza dentro de cada estrato, entonces se debe conseguir el estimador *jackknife* del estrato, de acuerdo al diseño dentro de éste
- Si se desea estimar la varianza total entre todos los estratos, entonces se deben crear grupos aleatorios bajo la misma estratificación, que contengan elementos de cada uno de los estratos.

Ahora se brinda el esquema general para un diseño estratificado por conglomerados, en el que se incluyen L estratos con n_h (h=1,2,...,L) conglomerados en cada uno. La aplicación del *jackknife* consiste en ir eliminando conglomerados, como se aprecia en los siguientes pasos:

1. Se calcula $\hat{\Theta}_{(hi)}$ de la misma forma que $\hat{\Theta}$ pero no se incluye el conglomerado i del estrato h.
2. Se obtiene un estimador del estrato promediando sobre todos los $\hat{\Theta}_{(hi)}$. De forma explícita:

$$\hat{\Theta}_{(h.)} = \sum_{i=1}^{n_h} \frac{\hat{\Theta}_{(hi)}}{n_h} \tag{3.11}$$

3. Dado $n = \sum_{h=1}^{L} n_h$, se obtiene un estimador global dado por:

$$\hat{\Theta}_{(..)} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{\hat{\Theta}_{(hi)}}{n}$$

4. Se obtiene el promedio de las estimaciones de los estratos, al que se denomina por $\overline{\Theta}_{(..)}$.

$$\overline{\Theta}_{(..)} = \sum_{h=1}^{L} \frac{\hat{\Theta}_{(h.)}}{L} \tag{3.12}$$

5. Se procede a calcular los pseudovalores, que para este caso son como sigue:

$$\hat{\Theta}_{hi} = (LW_h + 1)\hat{\Theta} - LW_h\hat{\Theta}_{(hi)} \tag{3.13}$$

donde, $i = 1, ..., n_h$; $h = 1, ..., L$.
$W_h =$ $(n_h - 1)(1 - \frac{n_h}{N_h})$ si es muestreo sin reemplazo
$1)$ para muestreo con reemplazo.

6. Finalmente, el estimador *jackknife* de Θ se define como:

$$\begin{aligned}\hat{\Theta}' &= \sum_{h=1}^{L}\sum_{i=1}^{n_h}\frac{\hat{\Theta}_{hi}}{Ln_h} \\ &= (1 + \sum_{h=1}^{L} W_h)\hat{\Theta} - \sum_{h=1}^{L} W_h\hat{\Theta}_{(h.)}\end{aligned}$$

7. El estimador de la varianza para el muestreo estratificado es:

$$v(\hat{\Theta}) = \sum_{h=1}^{L}\frac{W_h}{n_h}\sum_{i=1}^{n_h}\left(\hat{\Theta}_{(hi)} - \hat{\Theta}_{(h.)}\right)^2 \tag{3.14}$$

W_h es como se definió en (3.13), del inciso 5.

Es evidente, que se requerirán tantas iteraciones *jackknife* como sea el total de conglomerados. Del desglose anterior, se desprende también que de esta forma se obtiene un estimador de la varianza global, mas no se obtienen los componentes de varianza. Por otra parte, se observa que se consideran factores de corrección por población finita para conglomerados en cada estrato.
Rao (1996), considera una forma mucho más expedita de llevar a cabo el *jackknife* en un muestreo estratificado bietápico o cualquier muestreo complejo. Su método se aplica a estimadores que se consiguen como función de los factores de expansión. Por ejemplo, en un muestreo estratificado bietápico por conglomerados, con L estratos, n_h conglomerados muestreados en el estrato h y m_{hi} unidades encuestadas en el conglomerado i-ésimo del estrato h y a su vez, w_{hij} representa el factor de expansión de la unidad j-ésima, del conglomerado i-ésimo del estrato h, se puede considerar el siguiente estimador de la media de X:

$$\hat{\bar{X}} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}x_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}} \tag{3.15}$$

De igual forma, la correlación lineal entre dos variables x y y, bajo el diseño mencionado, se estima como sigue:

$$\hat{\rho_{xy}} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}(x_{hij}-\hat{\bar{x}})(y_{hij}-\hat{\bar{y}})}{\sqrt{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}(x_{hij}-\hat{\bar{x}})^2}\sqrt{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij}-\hat{\bar{y}})^2}} \tag{3.16}$$

Rao al igual que Wolter, estipula que se consigue cada réplica del *jackknife* eliminando un conglomerado de un estrato, pero su metodología se sintetiza en un ajuste de los factores de expansión. En cada iteración, suponga que se elimina el conglomerado l del estrato k y sean w_{hij}^{kl} los factores de expansión ajustados como se indica a continuación:

$$w_{hij}^{kl} = \left\{ \begin{array}{rll} w_{hij} & (h \neq k) & \text{Para las unidades que no están en el estrato donde se está eliminando un conglomerado.} \\ w_{hij}\frac{n_h}{n_h-1} & (\text{i} \neq l) & \text{Para las unidades en el estrato } k \text{ pero que no son del conglomerado } l. \\ 0 & (\text{i} = l) & \text{Para las unidades del conglomerado que se elimina.} \end{array} \right\}$$

Se obtiene el estimador $\Theta_{(kl)}$ de la misma forma que Θ (quitando el conglomerado l-ésimo del estrato k), pero con los nuevos factores de expansión. Luego la varianza de $\hat{\Theta}$ se estima de forma similar a (3.10)

$$v_J(\hat{\Theta}) = \sum_{h=1}^{L}\frac{n_h-1}{n_h}\sum_{i=1}^{n_h}(\hat{\Theta}_{(hi)}-\hat{\Theta})^2 \tag{3.17}$$

o bien, de acuerdo a (3.9), si se obtienen las diferencias respecto a $\hat{\Theta}_{(..)} = \sum^{L}\sum^{n_h}\hat{\Theta}_{(hi)}/n$ $(n = \sum_{h=1}^{L} n_h)$:

$$v_J(\hat{\Theta}) = \sum_{h=1}^{L}\frac{n_h-1}{n_h}\sum_{i=1}^{n_h}(\hat{\Theta}_{(hi)}-\hat{\Theta}_{(..)})^2 \tag{3.18}$$

Esta mecánica, resulta mucho más directa y general, además de que sus estimadores son de forma similar a los que arroja el planteamiento original del *jackknife*. El estimador de la varianza dado por Wolter (1985), para un muestreo estratificado bietápico, y éste dado por Rao, no son iguales. Se observa que el estimador de Rao no incluye factores de corrección por población finita (fcpf). Al respecto, Canty y Davison (1998), toman la misma expresión de Rao, incorporando los fcpf de la siguiente forma:

$$v_J(\hat{\Theta}) = \sum_{h=1}^{L}\frac{n_h-1}{n_h}\left(1-\frac{n_h}{N_h}\right)\sum_{i=1}^{n_h}(\hat{\Theta}_{(hi)}-\hat{\Theta})^2 \tag{3.19}$$

En el Capítulo 6 se verá que esta última versión no fue calculada en la aplicación. Por otra parte, se vio que el estimador de Rao y el de Wolter son numéricamente muy cercanos, lo que implica también, probablemente, que las fracciones de muestreo eran pequeñas.

3.2.3 Inferencia basada en el estimador *jackknife*

Inicialmente, se estudiaron las propiedades del estimador *jackknife* en el contexto de poblaciones infinitas. En 1958, Tukey sugirió que los estimadores individuales de las submuestras se podían considerar independientes e idénticamente distribuidos. De tal suerte, propuso el estimador de la varianza de $\hat{\theta}$ que se presentó en la sección 3.2.1. como $\widehat{VAR}_{J1}$ (3.9). Pero no se conformó con hacer una simple estimación de la varianza. La idea de independencia lo llevó a considerar los pseudovalores, $\tilde{\Theta}_i$, como las x_i en el caso $\hat{\Theta} = \overline{x}$. De tal suerte, propuso basar pruebas de hipótesis e intervalos de confianza para Θ en la estadística $\tilde{\overline{\Theta}}_J$, dada en (3.7). Es decir, un intervalo para Θ con una confianza de $1 - \alpha$, estaría dado por:

$$\tilde{\overline{\Theta}} \pm t_{\alpha/2,n-1}\sqrt{\widehat{VAR}_{J1}} \tag{3.20}$$

donde $t_{\alpha/2,n-1}$ es el cuantil de la distribución t con $n-1$ grados de libertad y probabilidad de obtener una t mayor de $\alpha/2$. Finalmente, se comprobó que las sospechas de Tukey eran acertadas en el caso asintótico, aún en el caso de población finita. Esto se debe al teorema de Krewski y Rao (1981), que como se discutió en el Capítulo 1, permite la creación de intervalos de confianza basados en la Normal (0,1) y en el estimador de varianza *jackknife*, cuando la estadística es una función suave de las observaciones.

Debido a este resultado o a los mecanismos que siempre se utilizan en el área de muestreo, es usual ver que los intervalos de confianza que se construyen a partir de estadísticas que provienen de muestras grandes (como suele ser el caso de muestreos complejos), se basan en la normal estándar, debido a que se considera que el tamaño de muestra así lo permite. Sin embargo, Rao (1996) hace ver que los grados de libertad del estimador de varianza obtenido por un método de remuestreo dependen del número de iteraciones o repeticiones que se hacen, por lo que, en la mayoría de los casos, los intervalos deben basarse en la distribución *t de Student* y no en la normal. En el anexo C se ofrece información sobre el cálculo de grados de libertad.

Recientemente Yung y Rao (1996), establecieron la consistencia de la varianza *jackknife* bajo ajustes de post-estratificación y no-respuesta. Estos resultados son importantes porque avalan su aplicación en situaciones donde con frecuencia la obtención de estimadores es laboriosa.

3.2.4 El problema de la mediana en el *jackknife*

Como se ha dicho anteriormente, los resultados asintóticos que se tienen para el *jackknife* son para estadísticas suaves. Se ha visto que el estimador de varianza *jackknife* para los cuantiles, o la mediana, en particular, no es consistente. Lo que implica, que aún teniendo muestras grandes el estimador de la varianza se puede alejar mucho del verdadero valor. Esta inconsistencia se ha demostrado (Efron 1985, Capítulo 3) en el caso de un muestreo aleatorio simple, en el marco de población infinita. Aunque no existen estudios teóricos sobre el comportamiento del *jackknife* cuando se trata de un muestreo complejo o una población finita, se conserva la alarma y se recomienda no aplicar este estimador de remuestreo. No obstante, Rao (1996, 1997), plantea la posibilidad de que en un muestreo

complejo, donde el *jackknife* se realiza eliminando conglomerados, no exista tal inconsistencia. Por lo pronto, se consideró de interés exponer en qué consiste el problema de la estimación de la varianza de la mediana por *jackknife*.

La inconsistencia resulta más evidente cuando n es par. Sean $(x_{(1)}, ..., x_{(m)}, x_{(m+1)}, ..., x_{(n)})$ las estadísticas de orden, de una muestra de tamaño n, donde $n = 2m$. Entonces la mediana es el punto medio entre $x_{(m)}$ y $x_{(m+1)}$. Claro está que al ejecutar el *jackknife*, se realizan $2m$ iteraciones. Cuando se elimina algún dato, $x_{(i)}$, tal que $(i) \leq (m)$, entonces, el estimador de la mediana es $x_{(m+1)}$; mientras que si el valor que se elimina es tal que $(i) \geq (m+1)$, entonces, la mediana es estimada por $x_{(m)}$. Así pues, de las $2m$ estimaciones, m de ellas están dadas por $x_{(m)}$ y las otras m, por $x_{(m+1)}$. El promedio de las n iteraciones *jackknife* es igual al parámetro pues:

$$\frac{m(x_{(m)} + x_{(m+1)})}{n} = \frac{x_{(m)} + x_{(m+1)}}{2}$$

Pero la varianza *jackknife*, dada en (3.10), resulta ser:

$$\begin{aligned} var(mediana)_J &= \frac{n-1}{n}\{m(x_{(m+1)} - \frac{x_m + x_{(m+1)}}{2})^2 \\ &\quad + m(x_{(m)} - \frac{x_m + x_{(m+1)}}{2})^2\} \\ &= \frac{n-1}{2}\left[\left(\frac{x_{(m+1)} - x_{(m)}}{2}\right)^2 + \left(\frac{x_m - x_{(m+1)}}{2}\right)^2\right] \\ &= \frac{n-1}{4}\left(x_{(m+1)} - x_{(m)}\right)^2 \end{aligned}$$

De manera general, existe un teorema que dice[3] que si $X_1, ..., X_n$ son variables aleatorias independientes e idénticamente distribuidas, con densidad $f(.)$ y función de distribución $F(.)$, estrictamente monótona, y con solución única para el cuantil p; entonces, la estadística que estima dicho percentil p, $\hat{X}_{np}$, se distribuye como una normal con media μ_p y varianza $\frac{p(1-p)}{n\ f(\mu_p)^2}$. En particular, en el caso de la mediana $p = 0.5$ y lo que ocurre con la verdadera varianza del estimador de este cuantil es[4]:

$$Var(mediana) \to \frac{1}{4n\ f(\mu_p)^2},$$

lo cual significa que si la $f(.)$ es $N(\mu, \sigma^2)$, entonces la varianza del estimador de la mediana es igual a $\pi\sigma^2/2n$. Lo grave del *jackknife* es que su estimador de varianza, bajo un m.a.s., no se acerca en el límite a la verdadera varianza, y como se puede apreciar, tiende a incrementarse según n crece. Ya que se mencionó el poco conocimiento que se tiene de lo que ocurre en muestreos complejos sobre este respecto, será interesante ver en el Capítulo 6, los resultados que se obtuvieron de la varianza *jackknife* de la mediana en la ENAL'96, y su

[3]Ver Mood (1974), pág. 255

[4]Ver Efron (1985), Capítulo 3, pág. 16.

comparación con el estimador de repeticiones balanceadas, el cual no posee este problema para los cuantiles. (Claro está, no existe referencia sobre la verdadera varianza).

3.3 Bootstrap

3.3.1 Caracterización del estimador

Efron (1979), dio a conocer una técnica para estimar la distribución acumulada de probabilidad, F, de una variable aleatoria. Al presentar el *bootstrap*, como llamó al método, resaltó su relación con el ya conocido *jackknife*, además de argüir mayor generalidad. Demostró que el *jackknife* es una aproximación lineal del *bootstrap*, lo cual, sin la referencia cronológica, podría pensarse que ocurriera al contrario (que primero se desarrollara el *bootstrap* y luego se aproximara linealmente). Cabe resaltar que aunque el *bootstrap* se concibió en el marco de una población infinita, Efron (1979, 1982) consideró brevemente el caso de población finita. Posteriormente, autores como Rao y Wu (1985, 1988) y Kovar, Rao y Wu (1988), incursionaron en la aplicación del *bootstrap* en problemas de muestreo, para obtener estimadores de la varianza e intervalos de confianza.

El problema más elemental se basa en la observación de una muestra independiente de tamaño n de una variable aleatoria X, cuya distribución de probabilidad F es desconocida. Se tiene,

$$X_i \sim F \quad \text{donde, } X_i \text{ i.i.d., } \text{i=1,2,...,n.}$$

Por otra parte, se observa $x = (x_1,x_2,...,x_n)$, la realización de $X = (X_1,X_2,...,X_n)$. Ahora bien, si se tiene la variable aleatoria $R(X,F)$, que depende de X y F desconocida, se desearía estimar la distribución muestral de R con base en la observación x. Para tal efecto Efron (1979) propone la consecución de los siguientes pasos, que describen el *bootstrap* en su forma más simple:

1. Se construye la distribución empírica, $\hat{F}$, atribuyendo una masa de $1/n$ a cada observación $x_1,x_2,...,x_n$.

2. Se considera

$$X_i^* \sim \hat{F} \quad \text{donde, } X_i^* \text{ i.i.d., } \text{i=1,2,...,n.}$$

   A partir de $\hat{F}$ se obtiene la muestra "*bootstrap*" de tamaño n,$x^* = (x_1^*,x_2^*,...,x_n^*)$; lo que equivaldría a obtener dicha muestra, de $(x_1,x_2,...,x_n)$ con reposición. Con base en x^* se obtiene $R^*(X^*,\hat{F})$, el valor "*bootstrap*" de R.

3. Se lleva a cabo el inciso 2 repetidas veces dejando $\hat{F}$ fija, para así aproximar la distribución muestral de $R(X,F)$ mediante la distribución de las $R^*(X^*,\hat{F})$.

Es claro que si lo que se deseara fuera aproximar la distribución de X_i, el estimador máximo verosímil, no paramétrico, de F es $\hat{F}$, o sea, la distribución empírica; ante lo cual,

no tendría sentido intentar afinar $\hat{F}$. Es importante hacer énfasis en el objetivo del *bootstrap*, porque si no es bien comprendido, pudiera parecer que es una solución redundante a un problema que ya sabemos resolver.

Cuando se obtiene la muestra $(x_1,x_2,...,x_n)$, se tienen n observaciones de X, y como ya mencionamos, podemos estimar $F_X(X) = F$. Sin embargo, si nos interesa $R(X,F) = t(X)$, sólo contamos con una observación, es decir, se tiene una muestra de tamaño $n = 1$, de $t(X)$. Obviamente, nos vemos imposibilitados para estimar $F_{T(X)}(T(X))$, pues además, $F_X(X)$ es desconocida.

La solución de Efron valora a $\hat{F}$ como máximo verosimil, pues parte de la similitud que espera de esta distribución con F, para simular una muestra de $R(X,F) = t(X)$, representada por los valores $R^*(x,\hat{F}) = t^*(x)$; de lo que se sigue, estimar la distribución de R, con la de R^*, a la cual se denomina *distribución bootstrap*. Queda claro entonces que el paso clave radica en cómo se calcule la distribución *bootstrap*. Para resolver este problema fundamental Efron (1979), propone tres alternativas, que se describen a continuación:

1. Cálculo teórico directo. (Obviamente, este método no siempre es posible aplicarlo).

2. Aproximación Monte Carlo. Se generan repetidas observaciones de $X*$, conservando $\hat{F}$ fija, se calculan los valores correspondientes de $R^*(X_1^*,\hat{F})$, $R^*(X_2^*,\hat{F})$,..., $R^*(X_n^*,\hat{F})$, y el histograma de estos últimos, lo que se tomará como la aproximación de la distribución *bootstrap*.

3. Utilizar algún método de linealización por series de Taylor para obtener la media y varianza de la distribución *bootstrap* de R^*.

La segunda alternativa de las mencionadas anteriormente representa la forma en que se ha generalizado la solución de la distribución *bootstrap*. Con base en ella, se puede explicar la mecánica y la lógica en cuestión: A partir de los valores "*bootstrap*", $R^*(X*,\hat{F})$, se construye a su vez lo que sería la distribución empírica de una estadística, $t^*(x)$, basada en la empírica de X; la cual viene a ser la distribución *bootstrap* deseada.

Es obvio que en la medida en que $\hat{F}$ se parezca a F, la estimación de la distribución de R, será mejor; esto es válido en toda la estadística ya que si la muestra es representativa, la inferencia es mejor. Esta dependencia sobre la muestra se presenta al utilizar cualquier estimador no-paramétrico. Por otra parte, Efron (1979), puntualiza que el que la distribución *bootstrap* de R^* sea buen estimador de la distribución de $R = t(X)$, también dependerá de la forma de $R(X,F)$; aspecto que se relaciona con las cantidades pivotales. Un ejemplo claro es que se espera que la distribución de $R(X,F) = \frac{t(X)-E_F(t)}{\sqrt{Var_F(t)}}$ se aproxime mejor que la de $R(X,F) = t(X)$. Más adelante se discutirá cómo este principio se ha aplicado para mejorar los intervalos de confianza.

Se ha revisado ya el fundamento general del *bootstrap*, por lo que se proseguirá con la aplicación de este método al problema que nos ocupa: el cálculo de la varianza de un estimador en un muestreo complejo. Del planteamiento de Efron (1979), se deduce que utilizando el mismo mecanismo con el que se estima la distribución de la estadística de interés, se obtendrá igualmente, una estimación de los parámetros de dicha distribución.

Rao y Wu (1985) explican la versión más simple de la aplicación del *bootstrap* al caso de un muestreo de una población finita, a la cual denominan "el *bootstrap* ingenuo" (*naive bootstrap*).

Supongamos que se tiene una población finita de la que interesa conocer el parámetro Θ, que es una función del vector de medias de p variables medidas en cada elemento de la población, $\overline{\boldsymbol{X}} = (\overline{X}_1, \overline{X}_2, ..., \overline{X}_p)^T$. Se considera el estimador $\hat{\Theta} = g(\overline{\boldsymbol{X}})$[5], que aun siendo una función no-lineal de las observaciones, puede ser calculado, pero no así su varianza. En el caso de un muestreo aleatorio simple, de una población finita, el "*bootstrap* ingenuo" se lleva a cabo mediante los siguientes pasos:

1. Partiendo de la muestra observada, se selecciona una muestra aleatoria simple con reemplazo, cuyos valores se denominan $\{x_i\}_1^n$. Se calculan $\overline{x}^* = \frac{1}{n}\sum_1^n x_i^*$, y $\hat{\Theta}^* = g(\overline{x}^*)$.

2. Se repite el paso anterior una gran cantidad de veces, B, y cada estimación de Θ se denomina $\hat{\Theta}^{*1}, \hat{\Theta}^{*2}, ..., \hat{\Theta}^{*B}$. Se obtiene entonces la media de estas estimaciones:

$$\hat{\Theta}_a^* = \sum_{b=1}^{B} \hat{\Theta}^{*b}/B$$

3. El estimador de la varianza de $\hat{\Theta} = g(\overline{x})$ está dado por:

$$v_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\Theta}^{*b} - \hat{\Theta}_a^*)^2 \tag{3.21}$$

Se puede apreciar que el estimador de la varianza es la aproximación Monte Carlo a

$$Var(\hat{\Theta}^*) = E(\hat{\Theta}^* - E_*\hat{\Theta}^*)^2,$$

donde $E_*\hat{\Theta}^*$ es la esperanza bajo la distribución *bootstrap*.

Es de esperar que si Rao y Wu (1988), llamaron a la metodología anterior "ingenua", ellos hallan presentado un cambio en la técnica, con una justificación adecuada. A continuación se presentará la aplicación del *bootstrap* a muestreos más elaborados. En la discusión del muestreo estratificado se verá el método propuesto por Rao y Wu, que considera el cálculo de pseudovalores.

3.3.2 Aplicación del *bootstrap* en muestreo

Muestreo estratificado

De acuerdo a Rao y Wu (1988), la extensión al muestreo estratificado requiere de "*bootstraps*" independientes en cada estrato, por cada iteración que se realiza (b=1,...,B). Es decir, se obtiene una muestra aleatoria con reemplazo en cada estrato, de tamaño n_h, de

[5]Krewski y Rao (1981), demostraron que el estimador $\hat{\Theta} = g(\overline{\boldsymbol{X}})$ se puede expresar como $\hat{\Theta} = g(\overline{x})$.

entre las unidades que están contenidas en la muestra total, las cuales se pueden denotar por $\{x_{hi}\}$. Se calcula, para cada estrato,

$$\overline{x}_h^* = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}^*$$

Nótese que la nomenclatura que incluye un "*", se refiere a los valores obtenidos de la simulación *bootstrap*. Por ejemplo, x_{hi}^* significa el estimador del total de una unidad que ha sido seleccionada (en el *bootstrap*) y $\overline{x}_h^*$ es el promedio del estrato h obtenido en una iteración *bootstrap*. Así también se obtiene $\overline{x}^* = \sum W_h \overline{x}_h^*$ y $\hat{\Theta}^* = g(\overline{x}^*)$, donde W_h son los pesos de los estratos. Al igual que antes, se repite el muestreo (*bootstrap*) por estrato y la obtención de estimadores, que tambien se llaman $\hat{\Theta}^{*1},...,\hat{\Theta}^{*B}$, seguida del promedio de éstos, $\hat{\Theta}_a^* = \frac{\sum \hat{\Theta}^{*b}}{B}$. El estimador de la varianza está dado por (3.21) de la sección 3.

Al examinar el caso lineal, con $p = 1$, $\hat{\Theta}^* = \sum W_h \overline{x}_h^* = \overline{x}^*$, Rao y Wu (1988), compararon la varianza que se obtendría mediante el *bootstrap*, $var_b(\overline{x}^*)$ con $var(\overline{x})$ la varianza real según diseño, y advirtieron que $\frac{var_b(\overline{x}^*)}{var(\overline{x})}$ *no converge a 1 en probabilidad* a menos que el número de estratos sea fijo y $n_h \to \infty$ en cada estrato. Como consecuencia, $var_b(\overline{x}^*)$ no es un estimador consistente de la varianza de $\overline{x}$, ni tampoco lo es la varianza *bootstrap* de cualquier estadística no lineal

Para solucionar este problema, Rao y Wu (1988) propusieron otro método de realizar el *bootstrap*, que se diferencía del anterior en la obtención de pseudovalores antes de calcular el estimador *bootstrap*. La propuesta es como sigue:

1. Se obtiene una muestra aleatoria simple en cada estrato, de tamaño m_h, $\{x_{hi}^*\}_{i=1}^{m_h}$, a partir de la muestra observada $\{x_{hi}\}_{i=1}^{n_h}$. Teniendo en mente que $\overline{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$ es la media del estrato h que resulta de la muestra original (total), se calculan:

$$\begin{aligned}
\tilde{x}_{hi} &= \overline{x}_h + \frac{\sqrt{m_h}}{\sqrt{n_h - 1}}(x_{hi}^* - \overline{x}_h) \\
\tilde{x}_h &= \frac{1}{m_h}\sum_{i=1}^{m_h} \tilde{x}_{hi} \\
\tilde{x} &= \sum W_h \tilde{x}_h \\
\tilde{\Theta} &= g(\tilde{x}).
\end{aligned}$$

   Cabe advertir que si el muestreo es sin reemplazo, en lugar de la expresión anterior para $\tilde{x}_{hi}$, se deberá considerar (3.23) de la sección 3.

2. Se repite lo anterior un gran número de veces[6], B, y se obtienen, $\tilde{\Theta}^1,...,\tilde{\Theta}^B$, así como el promedio de éstos, $\tilde{\Theta}_a^* = \frac{\sum_{b=1}^{B} \tilde{\Theta}^b}{B}$.

[6]Se entiende que el paso 1 se repite B veces, por lo que cada una de las estadísticas que allí se mencionan llevarían un índice b ($b = 1, ..., B$).

3. El estimador de la varianza está dado por:

$$\tilde{v}_b(a) = \frac{1}{B-1}\sum_{b=1}^{B}(\tilde{\Theta}^b - \tilde{\Theta}_a^*)^2 \tag{3.22}$$

Obviamente el estimador representa la estimación Monte Carlo de $E(\tilde{\Theta} - E_*\tilde{\Theta})^2$.

Con la adecuación que se acaba de señalar, ocurre que en el caso lineal, (3.22) se reduce a $var(\overline{x})$, para cualquier tamaño elegido, m_h. En el caso no lineal, $\tilde{v}_b(a)$ es consistente para $Var(\hat{\Theta})$. Por otra parte, cabe advertir que la normalidad asintótica de $\frac{(\tilde{\Theta}-\hat{\Theta})}{\sqrt{v_b}}$ se cumple para m_h comparable a n_h, en cada estrato. Si se elige $m_h = n_h - 1$ el método propuesto por Rao y Wu se reduce al *bootstrap* ingenuo, en el que se toman muestras de tamaño $n_h - 1$ en cada estrato. En tal situación, los pseudovalores $\tilde{x}_{hi}$ son iguales a las observaciones simuladas en el *bootstrap* ingenuo, es decir, $\tilde{x}_{hi} = x_{hi}^*$. Si ocurre que $n_h = 2$, Rao y Wu recomiendan tomar $m_h = 3$ o 4, ya que para $m_h = 1$ la varianza obtenida por el método de "Muestreo por Mitades" (*Half-sampling*) o repeticiones balanceadas sería más estable.

Es importante advertir que si el muestreo es estratificado simple **sin reemplazo** entonces, los pseudovalores consideran un factor por corrección finita de la siguiente forma:

$$\tilde{x}_{hi} = \overline{x}_h + \frac{\sqrt{m_h}}{\sqrt{n_h - 1}}\sqrt{1 - f_h}(x_{hi}^* - \overline{x}_h) \tag{3.23}$$

donde, $f_h = n_h/N_h$. Además, aunque se elija $m_h = n_h - 1$, si el muestreo es sin reemplazo, ocurre que $\tilde{x}_{hi} \neq x_{hi}^*$. El objetivo de estos ajustes por pseudovalores es lograr que la varianza *bootstrap* de una estadística lineal sea igual a la conocida. Este proceso de reescalamiento, según Sitter (1992) tiene el inconveniente de que en muestreos complejos se necesitan varias estadísticas de resumen, además de que los ajustes varían para distintos diseños y puede llegar a ocurrir que algún $\hat{\Theta}^b$ sea negativo, aún cuando por definición, $\hat{\Theta} \geq 0$.

Muestreo por conglomerados en dos etapas

El caso particular que ahora se discutirá es la aplicación del *bootstrap* cuando se tiene una población con N conglomerados, cada uno con M_i unidades. Se seleccionan n conglomerados **sin reemplazo** y m_i unidades en el conglomerado i-ésimo de los seleccionados, a través de un muestreo aleatorio simple sin reemplazo, nuevamente. Usualmente se denota por M_0 el tamaño de la población, es decir, $M_0 = \sum_1^N M_i$; lo que muchas veces es desconocido.

Para aplicar un *bootstrap* ahora hay que llevar a cabo dos etapas de aleatorización. Primero se elige una muestra aleatoria simple de n conglomerados *con reemplazo*. Luego se selecciona una muestra aleatoria con reemplazo de m_i elementos, entre las m_i unidades de la muestra original, en el conglomerado correspondiente. Si un mismo conglomerado se selecciona varias veces, se realizan tantos muestreos independientes como veces haya sido elegido. La secuencia anterior produciría una iteración del *bootstrap*, por lo que habría que repetirla B veces, para llegar a obtener el estimador *bootstrap* de la varianza.

Con el objeto de ser más explícita, hay que entrar en detalles de notación. Se recordará que el estimador en cuestión, $\hat{\Theta}$, se puede expresar como $\hat{\Theta} = g(\hat{\overline{X}})$. Ahora bien, la expresión para el estimador insesgado de $\overline{X}$ es:

$$\hat{\overline{X}} = \frac{\hat{X}}{M_0} = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{X}_i}{\overline{M}_0} \tag{3.24}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i}\sum_{j=1}^{m_i}\frac{M_i X_{ij}}{\overline{M}_0}$$

$$\text{Puesto que, } \hat{X} = \frac{N}{n}\sum_{i=1}^{n} M_i \overline{x}_i,$$

$$\overline{x}_i = \frac{1}{m_i}\sum_{i=1}^{m_i} x_{ij}$$

$$\text{y, } \overline{M}_0 = \frac{M_0}{N}$$

Por otra parte, sea:

x_{ij}^{**} = el valor de x, del elemento simulado por el *bootstrap*.

M_i^* = el tamaño del conglomerado iésimo del *bootstrap*.

m_i^* = el tamaño de muestra en el *bootstrap*, en el conglomerado i-ésimo.

$\hat{X}_i^*$ = el estimador del total en el conglomerado i-ésimo del *bootstrap*.

$$\hat{X}_i^* = M_i^* \overline{x}_i^{**}$$

$$\overline{x}_i^{**} = \sum_j x_{ij}^{**}/m_i^*$$

$$\lambda_1^2 = \frac{n}{n-1}(1-\frac{n}{N})$$

$$\lambda_{2i}^{*2} = \frac{n}{N}(1-\frac{m_i}{M_i})\frac{m_i^*}{m_i^*-1}$$

Rao y Wu indican que se deben calcular los siguientes valores:

$$\tilde{x}_{ij} = \hat{\overline{X}} + \lambda_1\left[\frac{\hat{X}_i^*}{\overline{M}_0} - \hat{\overline{X}}\right] + \lambda_{2i}^*\left[\frac{M_i^* x_{ij}^{**}}{\overline{M}_0} - \frac{\hat{X}_i^*}{\overline{M}_0}\right] \tag{3.25}$$

y

$$\tilde{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i^*}\sum_{j=1}^{m_i^*}\tilde{x}_{ij}$$

$$= \hat{\overline{X}} + \frac{\lambda_1}{n}\sum_{i=1}^{n}\left[\frac{\hat{X}_i^*}{\overline{M}_0} - \hat{\overline{X}}\right] + \frac{1}{n}\sum_{i=1}^{n}\lambda_{2i}^*\left[\frac{\hat{X}_i^{**}}{\overline{M}_0} - \frac{\hat{X}_i^*}{\overline{M}_0}\right] \tag{3.26}$$

Por supuesto, para la ejecución del *bootstrap*, se deben repetir las aleatorizaciones y los cálculos arriba mencionados, una gran cantidad de veces, B. Al igual que en los otros casos vistos, se obtienen B estimaciones de $\hat{\Theta}$, y el promedio de ellas. Finalmente el estimador de la varianza de $\hat{\Theta}$ se consigue de manera equivalente a (3.22) de la sección 3.

Estratificado bietápico por conglomerados

En el caso del *jackknife* y de repeticiones balanceadas se mostró un procedimiento de llevarlos a cabo que simplifica muchos cálculos, ajustando los factores de expansión, y que se hace extensivo a cualquier diseño y estimador. Con la misma lógica, Rust y Rao (1996), describen lo que probablemente es la forma más corta de hacer un *bootstrap*, en un muestreo estratificado bietápico por conglomerados. Aquí la generalización a cualquier diseño no es directa porque el ajuste de los pesos no es tan evidente. Para cada réplica del *bootstrap*, en cada estrato h, se eligen por m.a.s. con reemplazo z_h ($z_h > 0$) conglomerados de los n_h que existen en la muestra del estrato. (Como se aprecia se simula la elección de conglomerados y no de unidades dentro de conglomerados.) En cada iteración, se obtiene, para cada conglomerado de todos los estratos, el número de veces que fue seleccionado, al que se denomina $r_{hi}^{(b)}$. Es obvio que $0 < r_{hi}^{(b)} \leq z_h$. El ajuste de los factores de expansión, en la iteración b, se obtiene de la siguiente forma:

$$w_{hij}^{b} = w_{hij}\left(\left[1 - \sqrt{\frac{z_h}{(n_h - 1)}}\right] + \sqrt{\frac{z_h}{(n_h - 1)}}\frac{n_h}{z_h} r_{hi}^{(b)}\right) \tag{3.27}$$

Anteriormente se vio que para un diseño estratificado, se recomendaba la obtención de unos pseudovalores y varias estadísticas por estrato, de tal suerte que el estimador resultaba el conocido si se trataba de una estadística lineal. Este ajuste de pesos simplifica el proceso que inicialmente recomendaba Rao, pero se aplica solamente al estratificado por conglomerados. Obviamente, esta metodología no proporciona elementos para encontrar los componentes de varianza, por lo que, de tener interés en ello, habría que replantear la solución, considerando la aleatorización dentro de conglomerados.

Aunque ya se han visto ejemplos de cómo se estiman estadísticas que se expresan en términos de los factores de expansión, se presentará el caso de la mediana estimada por *bootstrap*, bajo el diseño que aquí tocamos.

En el caso en que se quiere estimar un cuantil y su varianza[7], como puede ser la mediana, se requiere estimar parte de la función de distribución muestral de la variable de interés, $F_{n(\mathbf{S})}$. Sea $y = (y_1, ..., y_N)$ el vector de valores de dicha variable en la población. El cuantil "p" en la población se estima de:

$$\xi_p = F_{n(\mathbf{S})}^{-1}(p) = inf\left[y_i : \ F_{n(\mathbf{S})}(y) \ \geq \ p\right] \tag{3.28}$$

La estimación de $F_{n(\text{bf S})}$ se hace creando variables indicadoras, I_z, tales que

[7]Ver Kovar, Rao, Wu (1988) y Kish (1972).

$$I(z) = I(y_{hij} \leq z) = 1 \text{ Si } y_{hij} \leq z$$
$$= 0 \text{ si no se cumple lo anterior}$$

En la expresión anterior y_{hij} denotaría que el diseño es estratificado bietápico por conglomerados y que este valor representa la respuesta de la unidad j-ésima del conglomerado i-ésimo del estrato h. Kovar, Rao y Wu (1988), recomiendan basar la estimación de la función de distribución en el pseudovalor $\tilde{y}_{hi}$, que se consigue mediante (3.22) o (3.23)[8].

De tal suerte la función de distribución (considerando el mismo diseño), evaluada en z, $F_{n(\mathbf{s})}(z)$, se obtiene de [9]:

$$F_{n(\mathbf{s})}(z) = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}^{b} I(z)_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}^{b}}$$

Esta función se debe evaluar en varios puntos (z), hasta encontrar el valor del cuantil deseado, como indica (3.28). La varianza del cuantil se consigue repitiendo el proceso de aleatorización y cálculos B veces y utilizando el estimador dado en (3.22), donde es claro que $\tilde{\Theta}^b$ es el cuantil que se estima en la iteración número b.

3.3.3 Intervalos de confianza basados en el estimador *bootstrap*

Se vio que Krewski y Rao (1981), dieron elementos para basar los intervalos de confianza en la distribución normal estándar cuando el estimador de la varianza es del tipo *jackknife*, de linealización o de repeticiones balanceadas. En el caso del *bootstrap*, los intervalos sólo se basan en la propia distribución que éste genera. Chaudhuri y Stenger (1992), abordan brevemente el *bootstrap* y la forma de hacer intervalos, así también Kovar, Rao y Wu (1988).

La manera más sencilla empieza por la elaboración de un histograma de los estimadores que se obtuvieron en las B iteraciones (En realidad, podría tenerse solamente los datos ordenados y no la gráfica en sí). Se consigue el valor de $\hat{\Theta}$ que tiene un área a su izquierda de 100 $\alpha/2$ (el que se denomina $\hat{\Theta}_{\alpha/2,l}$) e igualmente, aquél valor del estimador que tiene la misma área a su derecha ($\hat{\Theta}_{\alpha/2,u}$). Siguiendo el *método del percentil*, como comúnmente se conoce lo que se acaba de describir, finalmente, se considera que:

$$\left(\hat{\Theta}_{\alpha/2,l},\ \hat{\Theta}_{\alpha/2,u} \right)$$

[8] Así pues (3.29) se obtiene de $I(z) = I(\tilde{y}_{hij} \leq z)$.

[9] De manera general, considerando que la muestra total es de $n(\mathbf{s})$ unidades, se estimaría mediante:

$$F_{n(\mathbf{s})}(z) = \frac{\sum_{i=1}^{n(\mathbf{s})} w_i^{(b)} I(z)_i}{\sum_{i=1}^{n(\mathbf{s})} w_i^{(b)}}$$

es un intervalo de confianza para Θ con un nivel de $100\ (1-\alpha)$

Existe otra alternativa para elaborar intervalos basados en un *bootstrap*. Consiste en obtener estadísticas de la forma de una *t de Student*, a partir de los estimadores de cada iteración y la varianza *bootstrap*, como sigue:

$$t^b = \frac{\left(\tilde{\Theta}^b - \tilde{\Theta}_a\right)}{\sqrt{\tilde{v}_J^b}}$$

donde $\tilde{v}_J^b$ es el estimador de varianza *jackknife* aplicado a los datos simulados en la iteración b del *bootstrap*. Posteriormente, se encuentran los valores de $t_{\alpha/2,l}$ y $t_{\alpha/2,u}$, de la misma forma que $\hat{\Theta}_{\alpha/2,l}$ y $\hat{\Theta}_{\alpha/2,u}$. O sea, observando el área que queda a la izquierda y derecha, respectivamente, de estos estimadores. El intervalo para Θ, con un nivel de confianza de $100(1-\alpha)$

$$\left(\Theta - t_{\alpha/2,u}\sqrt{\tilde{v}_J}\ ,\ \hat{\Theta} - t_{\alpha/2,l}\sqrt{\tilde{v}_J}\right) \tag{3.29}$$

Si se recuerda, al discutir los inicios del *bootstrap*, se comentó que era de esperar que se lograra una mejor o más rápida aproximación a la distribución de $R(X,F) = \frac{t(X)-E_F(t)}{\sqrt{Var_F(t)}}$ que a la de $R(X,F) = t(X)$; tal es la justificación del intervalo dado en (3.29). Kovar, Rao y Wu (1988), en un estudio empírico encontraron que para los casos de estimador de razón y coeficiente de correlación, esta lógica para hallar un intervalo sólo probó ser mejor en el caso de intervalos de un solo lado. La gran dificultad operativa de los intervalos *bootstrap*es que además de requerir de un gran número de aleatorizaciones se suma el hecho de que necesitan del cálculo de la varianza *jackknife* $B+1$ veces (Una vez para toda la muestra y B veces para las iteraciones)