Capítulo 2

Métodos de remuestreo surgidos en la práctica

Por lo general, los problemas que se dan en la práctica se resuelven con base en una inteligencia intuitiva que dé soluciones rápidas. Posteriormente, esas soluciones se pueden afinar, incorporando elementos que incorporan aspectos teórico-empíricos. Este es el caso de los dos métodos de cálculo de varianza que se exponen en este capítulo. Ambos, el de *grupos aleatorios* y el de *repeticiones balanceadas* se iniciaron con lo que en un momento dado parecía una solución lógica y creativa. Posteriormente, se observaron ciertas deficiencias y los métodos fueron mejorados. A pesar de que datan de mediados de este siglo, los dos siguen siendo muy útiles; el de *repeticiones balanceadas* da solución a la estimación de varianzas con un diseño común en encuestas de gran cobertura; por otra parte, el de *grupos aleatorios* fue durante muchos años el único procedimiento factible de ser aplicado, y varios organismos internacionales reportan que no es hasta fechas recientes que han considerado la aplicación de otro método, si no es el caso de que continúan aplicándolo.

2.1 Grupos aleatorios

2.1.1 Caracterización del estimador

El método de grupos aleatorios, como ya se ha mencionado en la introducción, surgió en el trabajo de campo. Mahalanobis (1946), quien trabajaba en la India en problemas aplicados a la agricultura (en particular, la estimación de cosechas de diferentes productos) se vio en la necesidad de reducir las fuentes de error en sus encuestas. Es decir, estaba consciente de que el error total no consistía nada más de la variación del muestreo sino también de fallas humanas. Por tal razón, comenzó a diseñar muestras a las que llamó *interpenetrantes*, o *muestras a la mitad* ("half-sampling")[1] . Básicamente su técnica consistía en numerar

[1]Curiosamente Mahalanobis comenta que quiso llamarle *muestreo duplicado* (*Double sampling*), pero observó que este nombre producía en sus clientes la sensación de que la encuesta les costaba el doble, por lo que prefirió el término *Half-sampling*.

consecutivamente las unidades de muestreo dentro de cada estrato o unidad primaria; posteriormente, separaba la muestra de acuerdo a dicho número en nones y pares, formando dos conjuntos de datos que daban dos estimaciones. De tal suerte, con las dos muestras estaba también en posibilidades de calcular la varianza muestral. Aunque Mahalanobis procedía en su análisis suponiendo la independencia entre esas dos muestras, es evidente que los estimadores de cada una de ellas estaban correlacionados porque se construían obteniendo grupos mutuamente exclusivos; por lo tanto, no existía tal independencia y en realidad, su estimación incurría en un sesgo.[2]

El término *interpenetrante* alude a la aplicación del método al evaluar márgenes de error diferentes a la variabilidad del fenómeno que se observa. Sin embargo, la idea de Mahalanobis fue considerada por autores como Hansen, Hurwitz y Madow (1953), llamándole la *técnica del conglomerado último* (*Ultimate cluster*) o del *grupo aleatorio*, dándole un carácter más general. Deming (1956), la utilizó con el simple propósito de obtener un estimador de varianza sin importar lo complicado que fuera el estimador o el diseño muestral y le llamó *Muestreo replicado.*

A pesar de que las muestras de Mahalanobis no eran precisamente independientes, el concepto general de la metodología tuvo aceptación porque ayudaba a resolver varios problemas. De ahí que se desarrollaran dos formas de llevar a cabo esta técnica, basadas en: *muestras independientes* o *muestras dependientes.* En el primer caso, se supone una población finita P, de la cual, bajo cierto diseño, se extraen k muestras, s_1,s_2, ..., s_k; siendo que cada una de ellas se repone en la población antes de elegir la siguiente. De tal suerte, se obtienen k *muestras independientes.* Si en lugar de esta serie de muestras independientes se cuenta con una muestra total, la que es dividida aleatoriamente en k grupos, entonces se estaría hablando de k *muestras dependientes.* El estimador de la varianza en ambos casos es el mismo; la única diferencia es que al ser las muestras dependientes, se introduce un sesgo. Más adelante se mostrarán las características del estimador en un caso u otro. La única referencia explícita sobre este método es Wolter (1985), a excepción de la versión llamada *grupos aleatorios repetidos*, que se discute en este capítulo.

De manera general, sean $\hat{\Theta}_1$,$\hat{\Theta}_2$,...,$\hat{\Theta}_k$ los estimadores de Θ evaluados en los k grupos aleatorios. Se considera también a $\overline{\hat{\Theta}}$ como estimador del parámetro de interés, Θ, que se consigue de la siguiente forma:

$$\overline{\hat{\Theta}} = \sum_{i=1}^{k} \frac{\hat{\Theta}_i}{k}$$

Bajo el esquema de grupos dependientes, la esperanza del estimador en el grupo i-ésimo no necesariamente corresponde al valor esperado de la población. Es decir, $E(\hat{\Theta}_i) = \mu_i$, donde μ_i no necesariamente es igual a $\Theta = \mu$. De ahí que la esperanza del estimador de grupos dependientes es la media de las esperanzas, o sea, $E(\overline{\hat{\Theta}}) = \sum_{i=1}^{k} \frac{\mu_i}{k}$. Por otra parte, si los grupos son independientes entonces el estimador es insesgado.

[2]El señalar un error de alguien que dejó una obra tan reconocida no es agradable, pero resulta un hecho didáctico en el aspecto estadístico y de la vida en general.

Muestras independientes	Muestras dependientes
$\hat{\Theta}_1,...,\hat{\Theta}_k$ independientes.	$\hat{\Theta}_1,...,\hat{\Theta}_k$ correlacionadas.
$E(\hat{\Theta}_i) = \mu$ (i=1,...,k)	$E(\hat{\Theta}_i) = \mu_i$ (i=1,...,k)
$E(\overline{\hat{\Theta}}) = \mu = \Theta$.	$E(\overline{\hat{\Theta}}) = \overline{\mu}$.
$E(v(\overline{\hat{\Theta}})) = Var(\overline{\hat{\Theta}})$.	$E(v(\overline{\hat{\Theta}})) = Var(\overline{\hat{\Theta}})$ $+\frac{1}{k(k-1)}\sum_{i=1}^{k}(\mu_i - \overline{\mu})^2$ $-2\frac{\sum_{i=1}^{k}\sum_{j>i}^{k} Cov\{\hat{\Theta}_i,\hat{\Theta}_j\}}{k(k-1)}$.

Tabla 2.1: Propiedades de estimadores de grupos aleatorios

El estimador de la varianza de $\overline{\hat{\Theta}}$, tantò para el caso de grupos dependientes o independientes, está dado por:

$$v(\overline{\hat{\Theta}}) = \frac{\sum_{i=1}^{k}(\hat{\Theta}_i - \overline{\hat{\Theta}})^2}{k(k-1)} \tag{2.1}$$

Este estimador es insesgado si los grupos son independientes pero incurre en un sesgo si existe dependencia entre los grupos. En la tabla 2.1 se esbozan ciertas características de estos estimadores, de acuerdo al tipo de las muestras.

Como se advierte en la tabla anterior, no existe, en ninguno de los casos, el supuesto de que las varianzas de los k estimadores, $\hat{\Theta}_1$, ..., $\hat{\Theta}_k$, sean iguales. Esto implica que los k grupos, aunque sean independientes, no tienen que extraerse bajo un mismo diseño muestral. La verdad es que el supuesto de igualdad de varianza no se requiere para que $v(\overline{\hat{\Theta}})$ sea insesgado, pero si se desea hacer inferencias sobre Θ, dicho supuesto es necesario. A continuación se exhibe la esperanza de $v(\overline{\hat{\Theta}})$, bajo el supuesto de que los grupos son independientes.

Facilita el proceso, el considerar que $v(\overline{\hat{\Theta}})$ (definido en (2.1)) se puede expresar como sigue:

$$v(\overline{\hat{\Theta}}) = \frac{1}{k(k-1)}\left[\sum_{i=1}^{k}\hat{\Theta}_i^{\,2} - k\overline{\hat{\Theta}}^2\right]$$

Así pues,

$$\begin{aligned} E\left[v(\overline{\hat{\Theta}})\right] &= \frac{\sum_{i=1}^{k}\left(Var(\hat{\Theta}_i)+\mu^2\right)-k\left(Var(\overline{\hat{\Theta}})+\mu^2\right)}{k(k-1)} \\ &= \frac{k^2 Var(\overline{\hat{\Theta}})+k\mu^2-kVar(\overline{\hat{\Theta}})-k\mu^2}{k(k-1)} \\ &= \frac{(k^2-k)Var(\overline{\hat{\Theta}})}{k(k-1)} \\ &= Var(\overline{\hat{\Theta}}). \end{aligned}$$

Es posible que el lector haya advertido la posibilidad de utilizar otro estimador de Θ, diferente de $\overline{\hat{\Theta}}$. Éste sería el que se obtendría con la muestra total, compuesta de los k grupos. Es decir, una muestra de tamaño n, donde, $n = \sum_{i=1}^{k} n_i$, y n_i es obviamente, el tamaño de la muestra i-ésima. Si el estimador es lineal en las observaciones, este nuevo estimador, que se denomina $\hat{\Theta}$, será igual a $\overline{\hat{\Theta}}$. Pero generalmente:

$$\overline{\hat{\Theta}} \neq \hat{\Theta}$$

Si se utiliza $\hat{\Theta}$, entonces es lógico calcular su varianza como sigue:

$$v(\hat{\Theta}) = \frac{\sum_{i=1}^{k}(\hat{\Theta}_i - \hat{\Theta})^2}{k(k-1)}$$

Se ha visto que $v(\hat{\Theta}) \geq v(\overline{\hat{\Theta}})$. Pero al respecto, Wolter (1985, Capítulo 2), señala el hecho de que en muestreos complejos y de tamaño de muestra grande, los dos estimadores de varianza mencionados, no difieren mucho.

Grupos aleatorios repetidos

Otra versión de la técnica de grupos aleatorios, según Kovar, Rao y Wu (1988), es la de *grupos aleatorios repetidos* (GAR), la cual se avoca al caso en que se muestrean igual número de unidades, U, en cada estrato. La muestra se divide en U grupos, tal que cada una de los conglomerados o unidades en cada estrato, cae en sólo uno de esas submuestras; y se extraen los U estimadores, $\hat{\Theta}_u$, $(u = 1, ..., U)$ que les corresponden. Este proceso se repite un número arbitrario de veces, R, y la varianza se estima respecto al estimador global $\hat{\Theta}$, o respecto a la media del estimador en la repetición u-ésima dada por: $\overline{\Theta}_r = \sum_{u=1}^{U} \hat{\Theta}_{ru}/U$, $(r = 1, ..., R)$, como se ve a continuación:

$$v1_{GAR} = \frac{1}{R}\frac{\sum_{r=1}^{R}\sum_{u=1}^{U}(\hat{\Theta}_{ru} - \hat{\Theta})^2}{U(U-1)} \tag{2.2}$$

$$v2_{GAR} = \frac{1}{R}\frac{\sum_{r=1}^{R}\sum_{u=1}^{U}(\hat{\Theta}_{ru} - \bar{\Theta}_r)^2}{U(U-1)} \tag{2.3}$$

Esta forma de estimar la varianza consiste en obtener grupos dependientes, para una misma muestra total, un número repetido de veces. Entonces, dentro de cada réplica se tienen grupos dependientes, aunque dichas repeticiones son independientes entre sí. Con la información de la tabla 2.1 se da uno cuenta que la esperanza del estimador de varianza de GAR es la misma que la de grupos aleatorios Dependientes, ya que:

$$\begin{aligned} E\left[v1_{GAR}\right] &= \frac{1}{R}\sum_{r=1}^{R} E\left[v(\overline{\hat{\Theta}})\right] \\ &= E\left[v(\overline{\hat{\Theta}})\right] \end{aligned}$$

donde, $v(\overline{\hat{\Theta}})$ es el estimador de grupos aleatorios Dependientes.

Como un adelanto al siguiente método que se discute en este capítulo, se hace la notación de que cuando $U = 2$, el proceso resulta parecido al de *Muestreo por Mitades* o *repeticiones balanceadas* (RB) (Sección 2.2, y el inciso 2 de 4.4).

2.1.2 Consideraciones prácticas

En este trabajo se aborda el método de grupos aleatorios como una técnica para calcular la varianza, aunque se ha hecho mención de que en ocasiones se puede aplicar para comparar varianzas por encuestador u otro tipo de control. El método de muestras independientes, en particular, es apto para cumplir ambos propósitos.

Al construir grupos independientes, se recomienda que los encuestadores, así como otros trabajadores en la operación de la encuesta, se asignen a unidades que pertenezcan a un mismo grupo. De esta manera, se evita la existencia de correlación en los datos de diferentes grupos, debida a la fuente de error que puede introducir una persona al desempeñar su trabajo. Finalmente, la independencia entre las muestras sería lo más certero que se puede construir. Por otra parte, si un encuestador introduce cierto sesgo, éste se reflejará en los resultados que arroje su grupo, y se esperaría, que estas cifras contrastaran con las de los otros grupos. De ahí, la doble utilidad del método.

Sobre la construcción de muestras dependientes, es importante advertir que los grupos deben conservar el mismo diseño de muestreo que la muestra total. Igualmente, es recomendable, en el caso de muestras independientes pues así se garantizan las bases de inferencia, lo que se explica en la última sección de este capítulo. En el Apéndice A se exhiben las reglas dadas por Wolter (1985), para formar grupos conservando el diseño. A continuación se describe cómo se formarían grupos para un caso en particular. La tabla 3.2 muestra la estratificación que se estipuló para el estado de Jalisco en la ENAL96[3].

En esta encuesta las localidades dentro de los estratos representan conglomerados y las viviendas encuestadas en la localidad vienen a ser las unidades elementales de muestreo.

[3]L. Barragán, R. Hernández, A. Ávila y M.A. Ávila, Cartografía, Encuesta Nacional de Alimentación en el Medio Rural 1996, Instituto Nacional de Nutrición "Salvador Zubirán", México, D.F., 1997.

(Mayores detalles sobre el diseño y otros aspectos de esta encuesta se aprecian en el Capítulo 5). Bajo un diseño estratificado bietápico, como el que se utilizó en la encuesta, si se deseara formar tres grupos independientes, se tendrían que seleccionar tres muestras con el mismo diseño, donde los conglomerados son seleccionados con reemplazo, así como las viviendas encuestadas en cada conglomerado. No está por demás decir que el número de conglomerados por estrato debe ser igual en las tres muestras. Además debe quedar claro que la decisión de aplicar este método para cálculo de varianzas se toma en la etapa de planeación y diseño, pues se trata de contar con varias muestras independientes. Como consecuencia resulta evidente que el costo del levantamiento de la encuesta se puede incrementar. Sin embargo, puede ocurrir que esas tres muestras no se levanten a un mismo tiempo. Pudiera ser que, con la finalidad de evaluar cambios en la condición nutricional de un estado, éstas se levanten en diferentes periodos; en tal situación, las muestras son independientes y es posible el análisis comparativo entre ellas.

No.	Estrato	Localidades en estrato	Localidades encuestadas
1	HUEJUCAR	32	3
2	AMECA	35	3
3	TOMATLÁN	34	3
4	HUERTALA	45	3
5	CHIQUILISTAN	36	3
6	TAMAZULA DE GORDIANO	41	3
7	TLAJOMULCO	34	3
8	IXTLAHUACAN DEL RÍO	35	2
9	BARCALA	33	3
10	ATOTONILCO EL ALTO	38	3
11	TEPENTITLAN DE MORELOS	32	3
12	LAGOS DE MORENO	39	3

Tabla 2.2: Estratificación del Estado de Jalisco en la ENAL'96

Una situación muy distinta se daría si se cuenta con la muestra total, como aparece en la Tabla 2.2, y se decide formar tres grupos dependientes. En todos los estratos se encuestaron tres conglomerados, excepto en Ixtlahuacan del Río (8), donde sólo se levantó la encuesta en dos localidades. Si simplemente se piensa en asignar una localidad de cada estrato por grupo, ocurre que uno de los grupos tendría un estrato sin datos, lo cual no es deseable. Se puede pensar en hacer dos grupos, y según la regla (iv) del Apéndice A, existe la opción de

eliminar un conglomerado en cada uno de los 11 estratos restantes; o bien considerar que en cada grupo cerca de la mitad de los estratos tenga dos conglomerados y en el otro grupo, se cuente con una sola localidad. Posiblemente, ninguna de las alternativas sean satisfactorias, pues con cualquiera de ellas los diseños difieren bastante. Una opción para este caso (que no considera Wolter) es el colapsar desde un inicio el estrato no. 8, con el más parecido entre, estratos vecinos. De esta manera, es factible formar tres grupos dependientes, basados en 11 estratos, de los cuales 10 de ellos tienen una localidad encuestada en cada grupo y el restante (que correspondería al no. 8 junto con el que se colapsó), incluiría en dos grupos dos localidades, y en el tercer grupo, una localidad. Esta alternativa parece dar los grupos más parecidos pero es obvio que no permite calcular la varianza dentro de estratos por el hecho de que de manera general, sólo hay una localidad por estrato en los grupos. Aún que las varianzas dentro de estratos usualmente no interesen.

El ejemplo anterior sirvió para evidenciar que las reglas dadas por Wolter son de mucha ayuda, pero siempre acompañadas del criterio del estadístico, pues no resuelven todas las posibles situaciones que se pueden dar en la realidad.

2.1.3 Teoría entorno a grupos aleatorios

Las inferencias sobre Θ, cuando se calcula la varianza por el método de muestras independientes, se basan en el Teorema Central del Límite y en la distribución *t de Student*. Esto implica que se suponen $\Theta_1,...,\Theta_k$, no sólo independientes, sino también que provienen de una distribución normal (Θ, σ^2). Aunque la normalidad no se cumple exactamente, la teoría asintótica en muestreo, puede justificar el supuesto. Llama la atención igualmente, el que ahora se requiere suponer que las Θ_i se distribuyen idénticamente, con misma media y varianza, aun cuando la caracterización del estimador no considera igualdad de varianza. Por lo tanto, si se opta por grupos aleatorios independientes, pero con distinto diseño, no sería factible realizar inferencias sobre el parámetro Θ.

De manera general, cuando se estima la varianza del estimador de Θ mediante grupos aleatorios independientes, construidos bajo el mismo diseño, se considera el siguiente intervalo para Θ:

$$\left(\overline{\hat{\Theta}} - t_{k-1,\alpha/2}\sqrt{v(\overline{\hat{\Theta}})}\ ,\ \overline{\hat{\Theta}} + t_{k-1,\alpha/2}\sqrt{v(\overline{\hat{\Theta}})}\right),$$

donde $t_{k-1,\alpha/2}$ es el cuantil de la distribución t, con $k-1$ grados de libertad, y una probabilidad de obtener un valor mayor de $\alpha/2$. Con la misma base teórica se pueden realizar pruebas de hipótesis sobre Θ.

Claro está que cuando las muestras son dependientes, no se puede hacer inferencia sobre el parámetro Θ, ya que la misma construcción de las muestras invalida el supuesto más elemental. Por otra parte, si se cumple el supuesto de que los k estimadores son idénticamente distribuidos, pues se conserva el mismo diseño en todos los grupos.

Cabría la pregunta de cuán robusto puede ser el intervalo dado por la distribución t, pero al respecto no se encontró nada en la literatura. Sin embargo, se tienen referencias por comunicaciones personales, de que este método ha sido ampliamente utilizado por organismos internacionales durante décadas.

2.2 Repeticiones balanceadas

2.2.1 Caracterización del estimador

La técnica de repeticiones balanceadas ("Balanced Repeated Replication") también se conoce como *muestreo por mitades* ("Half-sampling"), *muestreo por mitades balanceadas* ("Balanced Half-samples"), *muestreo fraccional balanceado*, *Pseudorepetición* ("pseudoreplication") o método de semimuestras reiteradas (Mirás, 1985). Los inicios de esta metodología se dan en trabajos aplicados de muestreo, no en el ámbito teórico como ocurrió con el *jackknife* o el *bootstrap*. Aún así, está muy relacionado con estos métodos y con el de grupos aleatorios; como se mencionó en la introducción, las congruencias entre estos métodos no se deben a una evolución de las técnicas sobre una misma línea, sino que se han encontrado tras análisis conjunto.

El lector que haya revisado el tema de grupos aleatorios, puede imaginar que la técnica de repeticiones balanceadas se debe también a Mahalanobis, debido a que él llamó a sus primeros diseños que consideraban dos sub-muestras *"half-sampling"*. La coincidencia de nomenclatura parece deberse a la adopción del término por otros autores una vez que idearon una técnica a la que en realidad, le iba mejor dicho nombre. Sin embargo, aunque los diseños de Mahalanobis bien pudieron ser el preámbulo para la metodología que plantearon, no se menciona en la literatura revisada ninguna referencia de los iniciadores del muestreo por mitades, ninguna referencia a Mahalanobis.

Autores como Wolter (1985) y Judkins (1987), señalan que el método tuvo sus orígenes en la década de los cincuenta, en el Departamento del Censo de Estados Unidos de América, en trabajos de W.N. Hurwitz, M. Gurney y colaboradores. Posteriormente la técnica fue formalizada por McCarthy en 1966. Luego, estudios basados en simulaciones (Kovar, 1985; Hansen y Tepping, 1985) reportaron que el estimador de la varianza *jackknife* se comportaba mejor que el de repeticiones balanceadas en cuanto a sesgo y estabilidad. De ahí que Robert Fay (Dippo, Fay y Morganstein, 1984), del Departamento del Censo de Estados Unidos, planteara una modificación a la técnica de repeticiones balanceadas que mejora su estabilidad.

Se abordará ahora la técnica repeticiones balanceadas en su forma simple y posteriormente se comentará la adecuación de Fay. El método se aplica al caso en que se tienen L estratos con dos unidades primarias de muestreo en cada uno. Es decir, el tamaño de muestra en cada estrato es 2, $n_h = 2$, y el número total de unidades muestreadas[4] es $n = \sum_{i=1}^{L} n_h = 2L$. Bajo este diseño de muestreo, si se fuera a desarrollar la metodología de

[4]Cabe la posibilidad de que estas unidades se refieran a conglomerados, que de hecho es lo más usual. La sección 3.2.6 se refiere a esta situación.

grupos aleatorios **dependientes**, ocurriría que sólo se podrían formar dos grupos ($k = 2$), que conservaran la misma estructura de diseño, los cuales se podrían denominar por: $(x_{11}, x_{21}, \ldots, x_{L1})$ y $(x_{12}, x_{22}, \ldots, x_{L2})$.

Interesa estimar la varianza de un estimador $\hat{\Theta}$, que es de la forma $\hat{\Theta} = g(\overline{X})$, por lo que al abordar el caso de $\overline{x}$, se sigue la generalización para otro $\hat{\Theta}$. Como referencia, se debe recordar que la media global bajo un diseño estratificado, considerando $n_h = 2$ (h=1,2,...,L), N_h es el tamaño del estrato h y $N = \sum_{h=1}^{L} N_h$, se obtiene de:

$$\overline{x}_{est} = \sum_{h=1}^{L} W_h \overline{x}_h \tag{2.4}$$

$$\text{donde, } \overline{x}_h = \frac{(x_{h1} + x_{h2})}{2} \tag{2.5}$$

$$\text{y } W_h \text{ son los pesos de los estratos, } W_h = \frac{N_h}{N} \tag{2.6}$$

También conviene recordar que el estimador de la varianza de $\overline{x}_{est}$, en un muestreo estratificado simple, si la fracción de muestreo en cada estrato, n_h/N_h, es despreciable, se expresa como:

$$\begin{aligned} v(\overline{x}_{est}) &= \sum_{h=1}^{L} W_h^2 \frac{s_h^2}{2} \\ &= \sum_{h=1}^{L} \frac{W_h^2}{2} \left\{ \left[x_{h1} - \frac{(x_{h1} + x_{h2})}{2} \right]^2 + \left[x_{h2} - \frac{(x_{h1} + x_{h2})}{2} \right]^2 \right\} \\ &= \sum_{h=1}^{L} W_h^2 \frac{(x_{h1} - x_{h2})^2}{4} \\ &= \sum_{h=1}^{L} W_h^2 \frac{d_h^2}{4} \qquad (2.7) \\ \text{donde, } d_h &= x_{h1} - x_{h2} \end{aligned}$$

Bajo el método de grupos aleatorios, se obtienen dos estimadores, $\hat{\Theta}_1$ y $\hat{\Theta}_2$, obviamente considerando los pesos W_h $(h = 1, 2, \ldots L)$ de los L estratos, de la siguiente manera:

$$\overline{x}_1 = \sum_{h=1}^{L} W_h x_{h1} \text{ , de donde se obtiene, } \hat{\Theta}_1 = g(\overline{x}_1).$$

De igual forma se calcula el estimador del segundo grupo, obteniendo $\overline{x}_2$ y $\hat{\Theta}_2$. Aplicando (2.1) es inmediato que el estimador global sólo se basaría en un grado de libertad, lo que le daría poca estabilidad. La idea tras el método de muestreo por mitades, es considerar, *medias-muestras* compuestas por una unidad por estrato. Esto significa que se permite que las muestras se traslapen, pues se observan muestras con elementos comunes. De tal suerte,

existen 2^L posibles *medias-muestras*, las cuales representan una forma de pseudo-repetición o pseudo-remuestreo.

De cada media-muestra se deriva un estimador de la media, $\overline{x}_i$, el cual se exhibe a continuación:

$$\overline{x}_i = \sum_{i=1}^{L} W_h(\delta_{h1i}x_{h1} + \delta_{h2i}x_{h2}) \tag{2.8}$$

$$\text{Donde } \delta_{h1i} = \left\{ \begin{array}{ll} 1 & \text{si la unidad (h,1)} \in \text{la muestra i-ésima,} \\ 0 & \text{si la unidad (h,1)} \notin \text{la muestra i-ésima.} \end{array} \right\}$$

$$\delta_{h2i} = 1 - \delta_{h1i}. \tag{2.9}$$

En el caso de un estimador lineal, si se consideran las 2^L estimaciones, se tiene que el promedio de ellas es igual al estimador de un diseño estratificado; lo cual se deriva del hecho de que cada unidad se encuentra en la mitad de las 2^L muestras. En ocasiones el número de estratos, L, puede ser muy grande. Por lo tanto, resulta lógico que se considere sólo una fracción de esas 2^L posibles medias-muestras; digamos k de ellas. Intuitivamente el lector ya puede darse cuenta de que es posible construir un estimador de la varianza de $\overline{x}_{est}$, mediante el promedio del cuadrado de las diferencias de los $\overline{x}_i$ obtenidos (i=1,2,...,k), con respecto a $\overline{x}_{est}$. Tal suposición es cierta, sin embargo, al inspeccionar un poco el estimador se aprecia que no se deben considerar cualesquiera k medias-muestras. Aquí es donde entra la colaboración de McCarthy (1966), puntualizando que se deben escoger k muestras *balanceadas*. A partir de este argumento, se entiende por el estimador de muestreo por mitades ("Half-Sampling"), el que se basa en todas las 2^L muestras y sus respectivas estimaciones y por el estimador de repeticiones balanceadas ("Balanced Repeated Replication"), el que se basa en k medias-muestras balanceadas.

Antes de explicar qué son muestras balanceadas, es necesario examinar un poco el estimador en cuestión. De acuerdo a lo dicho en el párrafo anterior, se propone como estimador de la varianza de la media, basado en k muestras:

$$v_k(\overline{x}_{est}) = \sum_{i=1}^{k} \frac{(\overline{x}_i - \overline{x}_{est})^2}{k} \tag{2.10}$$

Y análogamente, para cualquier $\hat{\Theta}$ en general:

$$v_k(\hat{\Theta}) = \sum_{i=1}^{k} \frac{(\hat{\Theta}_i - \hat{\Theta})^2}{k} \tag{2.11}$$

$\hat{\Theta}_i$ es el estimador obtenido de la muestra i-ésima, aplicando (2.8). Por $\hat{\Theta}$ se entiende el estimador que se obtiene por las fórmulas usuales para diseño estratificado (Considerando $2L$ observaciones), o bien, $\hat{\Theta} = g(\overline{x}_{est})$. Cabe señalar que en ocasiones se construye (2.11) con base en las desviaciones hacia:

$$\hat{\Theta}_{(.)} = \sum_{j=1}^{k} \frac{\hat{\Theta}_j}{k} \tag{2.12}$$

y no respecto a $\hat{\Theta}$.

Al examinar el caso de la varianza de una estadística lineal, $\hat{\Theta}_{lin}$, se encuentra el argumento de McCarthy para definir y requerir *muestras balanceadas*; por cuanto lo que sigue de esta discusión, se enfoca hacia este tipo de estimador. De manera general, una estadística lineal en la media muestral se expresa como[5]:

$$\begin{aligned} \hat{\Theta}_{lin} &= \sum_{h=1}^{L} \left\{ \mu_h + \frac{1}{n_h} \sum_{j=1}^{n_h} W_h x_{hj} \right\} \\ &= \sum_{i=1}^{L} \left\{ \mu_h + W_h \overline{x}_h \right\} \end{aligned} \tag{2.13}$$

A partir de las variables indicadoras δ_{h1i} y δ_{h2i} definidas en 2.2.1, se puede construir otra variable indicadora, $\delta_h^{(i)}$, que corresponde a la muestra i-ésima, de la siguiente forma:

$$\delta_h^{(i)} = 2\delta_{h1i} - 1. \tag{2.14}$$

Así pues,

$$\delta_h^{(i)} = \left\{ \begin{array}{cl} 1 & \text{si la unidad (h,1) está en la muestra iésima} \\ -1 & \text{si la unidad (h,2) está en la muestra iésima} \end{array} \right\}$$

Ahora bien, cuando se considera la media-muestra i-ésima, la media del estrato h es igual al valor de la unidad incluida en ésta, es decir,

$$\overline{x}_h^i = \frac{x_{h1}(1 + \delta_h^{(i)}) + x_{h2}(1 - \delta_h^{(i)})}{2} \tag{2.15}$$

Si se hace la sustitución de (2.15) en (2.13), y se desarrolla un poco el álgebra, se ve que el estimador lineal de cualquiera de las medias-muestras ($i = 1, 2, ..., 2^L$) es igual a (2.13) más otro término que depende de las $\delta_h^{(i)}$; lo que se desglosa aquí:

$$\begin{aligned} \hat{\Theta}_{i_{lin}} &= \sum_{h=1}^{L} \left\{ \mu_h + \frac{W_h}{2} \left[x_{h1}(1 + \delta_h^{(i)}) + x_{h2}(1 - \delta_h^{(i)}) \right] \right\} \\ &= \sum_{h=1}^{L} \left\{ \mu_h + \frac{\sum_{j=1}^{2} W_h x_{hj}}{2} + \delta_h^{(i)} \frac{x_{h1} - x_{h2}}{2} \right\} \\ &= \hat{\Theta}_{lin} + \sum_{h=1}^{L} \delta_h^{(i)} W_h \frac{x_{h1} - x_{h2}}{2} \end{aligned} \tag{2.16}$$

[5] Ver Efron (1982), pág. 61.

Con estos elementos, (2.11) y (2.16), ya se puede desarrollar la varianza de $\hat{\Theta}_{lin}$, a partir de k muestras.

$$\begin{aligned} v\{\hat{\Theta}_{lin}\}_k\} &= \frac{1}{k}\sum_{j=1}^{k}(\hat{\Theta}_{i_{lin}} - \hat{\Theta}_{lin})^2 \\ &= \frac{1}{k}\sum_{j=1}^{k}\left[\sum_{h=1}^{L}\delta_h^{(i)}W_h\frac{(x_{h1}-x_{h2})}{2}\right]^2 \\ &= \frac{1}{k}\sum_{j=1}^{k}\sum_{h=1}^{L}W_h^2\frac{(x_{h1}-x_{h2})^2}{4} \\ &\quad +\frac{1}{k}\sum_{j=1}^{k}\sum_{h=1}^{L}\sum_{h'\neq h}\delta_h^{(j)}\delta_{h'}^{(j)}W_hW_{h'}\frac{(x_{h1}-x_{h2})(x_{h'1}-x_{h'2})}{4} \end{aligned} \tag{2.17}$$

El primer término en (2.17) se reduce a

$$\sum_{h=1}^{L}W_h^2\frac{(x_{h1}-x_{h2})^2}{4},$$

lo que, si se compara con (2.7), se aprecia que es igual a la varianza que se obtendría mediante las fórmulas conocidas para muestreo estratificado. Por otra parte, el segundo término puede ser muy grande para ciertos conjuntos de medias-muestras, pero se hace cero si ocurre que

$$\sum_{j=1}^{k}\delta_h^{(j)}\delta_{h'}^{(j)} = 0 \quad (\text{para } 1 \leq h < h' \leq L). \tag{2.18}$$

El método de McCarthy busca la manera de obtener para una estadística lineal, la misma varianza que se obtendría a través de la fórmula (2.7), pero con k muestras, donde $k < 2^L$. Para tal efecto, el método de *repeticiones balanceadas* se basa en cualquiera de las k muestras de las posibles 2^L *medias-muestras*, que cumplan (2.18). Por supuesto, el conjunto de 2^L medias-muestras satisface tal requerimiento, ya que cada unidad se encuentra en 2^{L-1} medias-muestras[6], pero la ventaja del método consiste en necesitar un número pequeño de pseudo-repeticiones.

Antes se mencionó la posibilidad de calcular la varianza considerando las desviaciones repecto a $\hat{\Theta}_{(.)}$, dado en (2.12). Si así se hiciera y la estadística en cuestión fuera lineal, se observaría que

[6]Es decir, $\sum_{j=1}^{2^L}\delta_h^{(j)}\delta_{h'}^{(j)} = \underbrace{\sum_{j\ni\delta_h^{(j)}=1}\delta_{h'}^{(j)}}_{} - \underbrace{\sum_{j\ni\delta_h^{(j)}=-1}\delta_{h'}^{(j)}}_{} = 0$

$$\hat{\Theta}_{lin_{(.)}} = \sum_{j=1}^{k} \frac{\hat{\Theta}_{jlin}}{k} \tag{2.19}$$

$$= \hat{\Theta}_{lin} \tag{2.20}$$

si se satisface el siguiente requisito por las medias-muestras elegidas:

$$\sum_{j=1}^{k} \delta_h^{(j)} = 0. \text{ (h=1,2,...,L).} \tag{2.21}$$

Lo cual se deduce claramente al expandir $\hat{\Theta}_{lin_{(.)}}$;

$$\hat{\Theta}_{lin_{(.)}} = \hat{\Theta}_{lin} + \frac{1}{k} \sum_{h=1}^{L} W_h \frac{x_{h1} - x_{h2}}{2} \sum_{j=1}^{k} \delta_h^{(j)}$$

Se dice que las medias-muestras poseen *balance ortogonal completo* si además de satisfacer (2.18), cumplen también con (2.21). No hay que olvidar que todas estas definiciones y características se atribuyen al caso de una estadística lineal en la media. Es concebible que se busquen estimadores para estadísticas no lineales que no se alteren, al aplicarlos a casos lineales. Pero lo cierto es que aún utilizando medias-muestras balanceadas, para el caso no lineal no se garantiza ninguna propiedad de las mencionadas; en particular, el promedio de las estimaciones $\hat{\Theta}^{(i)}$ $(i = 1, 2, ..., k)$, no será igual al estimador global obtenido por reglas de muestreo estratificado, basado en $2L$ observaciones; sobre la varianza, resulta obvio el equivalente del argumento anterior, ya que si se pudiera calcular el estimador de la varianza mediante las reglas generales de muestreo, en el caso no lineal, no tendría sentido ni éste ni ningún método. Sin embargo, sí existe una buena razón para considerar medias-muestras balanceadas, pues Krewski y Rao (1981), demostraron resultados asintóticos que permiten hacer inferencia basada en la varianza obtenida a partir de ellas.

2.2.2 Modificación Fay

El estimador de la varianza de repeticiones balanceadas, probó ser muy inestable al aplicarse a estadísticas no lineales. En particular, se detectó que el estimador es problemático al calcular la varianza de una razón, ya que el denominador puede ser cero o cercano a cero para algunas medias-muestras. Si se inspecciona (2.15), se aprecia que tácitamente, se atribuye a las unidades dentro de cada estrato un peso de 0 ó 2. Robert Fay, (Dippo, Fay y Morganstein, 1984) sugirió cambiar tales pesos para evitar los problemas mencionados. Su idea original era usar ponderaciones de 0.5 y 1.5, pero hizo el planteamiento general. En vez de los pesos $(1 + \delta_h^{(i)})$ y $(1 - \delta_h^{(i)})$, como se observan en 2.15, las unidades de los estratos se ponderan por $(1 + \delta_h^{(i)}(1 - \lambda))$ y $(1 - \delta_h^{(i)}(1 - \lambda))$; de tal suerte, se considera el siguiente estimador de la media de un estrato h $(h = 1, 2, ..., L)$:

$$\overline{x}_{hFay}^{(i)} = \frac{x_{h1}\left[1 + \delta_h^{(i)}(1 - \lambda)\right] + x_{h2}\left[1 - \delta_h^{(i)}(1 - \lambda)\right]}{2} \tag{2.22}$$

El estimador de la varianza consiste en las desviaciones del estimador *Fay* respecto al estimador de muestreo estratificado. En otras palabras, a partir de (2.22) se consigue $\overline{x}_{Fay} = \sum_{h=1}^{L} W_h \overline{x}_{hFay}$; luego, se busca $\hat{\Theta}_{Fay}^{(j)} = g(\overline{x}_{Fay})$ $(j = 1, 2, ..., k)$, para finalmente calcular:

$$v_{Fay}(\hat{\Theta}) = \sum_{j=1}^{k} \frac{(\hat{\Theta}_{Fay}^{(i)} - \hat{\Theta})^2}{k(1-\lambda)^2} \tag{2.23}$$

Nótese que si $\lambda = 0$, entonces la estimación de Fay coincide con la de *repeticiones balanceadas*, presentada anteriormente, tanto en la estimación de la media, como en la varianza. Igualmente debe llamar la atención la incorporación del término $\frac{1}{(1-\lambda)^2}$ en el estimador de la varianza. Al respecto, Judkins (1987, pág. 492) explica que el error cuadrático medio disminuye en la medida en que λ se acerca a la unidad, pero se transforma en un estimador razonable de varianza al multiplicarse por dicho factor. En el mismo artículo, Judkins hace ver que se puede pensar que $100(1-\lambda)$ es un factor de *perturbación*. O sea, en una aplicación de repeticiones balanceadas sin la adecuación de Fay, la perturbación es del 100%, pero si se ejecuta la modificación con $\lambda = 0.9$, la perturbación es del 10%. La determinación óptima de λ no ha sido dada por ningún autor. Los resultados de simulaciones de repeticiones balanceadas, adecuación de Fay y *jackknife* para razones de variables normales, de Judkins (1987, pág. 493), mostraron que cuando los errores estándares de los estratos son pequeños, todos los métodos son esencialmente equivalentes, pero si la variabilidad dentro de los estratos es grande y λ se acerca a 0, el estimador de la varianza es muy malo. Además ocurre que, sin importar la varianza de los estratos, la modificación de Fay y el *jackknife* convergen en la medida en que λ se acerca a la unidad. Por otra parte, dichas simulaciones de Judkins (1987), revelan que la modificación Fay conserva las características del caso lineal y reduce el sesgo y la varianza del estimador de la varianza de una razón.

2.2.3 Otros estimadores para estadísticas no lineales

Una peculiaridad de esta técnica es que existen varios estimadores para el caso no lineal. Para empezar, es necesario señalar que por cada conjunto de medias-muestras balanceadas, se tiene directamente otro que también es balanceado; éste es aquél cuyos indicadores $\delta_h^{(j)}$ $(j = 1, 2, ..., k$ y $h = 1, 2, ..., L)$ son el negativo del primero, razón por la cual se dice que es su *complemento*. Si el estimador es lineal, cualquier conjunto de medias-muestras balanceadas va a dar el mismo estimador, por lo que no tiene caso analizar dos conjuntos balanceados; pero cuando el estimador no es lineal, se obtienen resultados diferentes.

Sea $\hat{\Theta}_{(j)}^{c}$ el complemento de la media-muestra i-ésima. En seguida se deducen otras formas de calcular la varianza, tal como se aprecia aquí:

$$v_k^c(\hat{\Theta}) = \sum_{j=1}^{k} \frac{(\hat{\Theta}_{(j)}^{c} - \hat{\Theta})^2}{k} \tag{2.24}$$

$$\overline{v}_k(\hat{\Theta}) = \frac{v_k^c(\hat{\Theta}) + v_k(\hat{\Theta})}{2} \tag{2.25}$$

Wolter (1985), es quien presenta éstos y otros estimadores para el caso no lineal. Su libro no abarca la modificación de Fay que se discutió en la sección anterior. Sin embargo, es claro que si se utiliza el estimador Fay, se pueden calcular $v_k{}^c(\hat{\Theta})$ y $\overline{v}_k(\hat{\Theta})$, de la siguiente forma:

$$v_{k_{Fay}}^{c}(\hat{\Theta}) = \sum_{j=1}^{k} \frac{(\hat{\Theta}_{(j)_{Fay}}^{c} - \hat{\Theta})}{k(1-\lambda)} \tag{2.26}$$

$$\overline{v}_{k_{Fay}}(\hat{\Theta}) = \frac{v_{k_{Fay}}^{c}(\hat{\Theta}) + v_{k_{Fay}}(\hat{\Theta})}{2} \tag{2.27}$$

2.2.4 Consideraciones prácticas

Elección de las muestras balanceadas

Dada la descripción del estimador de repeticiones balanceadas, el primer problema práctico al que se enfrenta el estadístico que quiere aplicar esta técnica, es la selección de un conjunto de medias-muestras balanceadas. En primer lugar se pregunta qué valor de k es factible considerar, y luego se enfrenta a los requerimientos (2.18) y (2.21).

Esos cuestionamientos encontraron una solución en el trabajo de Plackett y Burman de 1946 (citado por Wolter, 1985), que aunque se trata de diseños de experimentos factoriales 2^n, fue igualmente útil para este problema de muestreo. Ellos construyeron matrices ortogonales de dimensión m X m, donde m es múltiplo de 4, y sus entradas son 1 ó -1. (Éstas son conocidas en matemáticas como matrices *Hadamard*).

Las columnas de estas matrices satisfacen (2.18), por lo que se identifican los estratos con las columnas y los renglones con las repeticiones (medias-muestras). A continuación se ofrecen advertencias en cuanto a la determinación de k, dado un diseño basado en L estratos:

a) Se tiene un **balance ortogonal completo** si

$$k > L \ , \ k \text{ múltiplo de } 4$$

b) Se tiene **balance**, pero no balance ortogonal completo si

$$k = L, \ \text{k y L múltiplos de } 4$$

c) No existe balance alguno si

$$k < L$$

La recomendación más evidente es pues, escoger k como el valor más pequeño que satisface el inciso (a). De esta manera se consigue además minimizar los cálculos. Por otra parte, es necesario basar la selección de las medias-muestras en la matriz Hadamard de dimensión k x k, una vez que se haya elegido el número de repeticiones.

Muestreo estratificado bietápico

Un aspecto importante a considerar es la generalización de esta técnica al caso en que las dos unidades muestreadas en cada estrato son conglomerados (Unidades Primarias de Muestreo), a lo que sigue una segunda etapa de muestreo. Se puede revisar en Cochran (1977) o Sukhatme (1984), el estimador insesgado para la media bajo este diseño está dado por:

$$\overline{x}_{\text{BI}} = \sum_{h=1}^{L} W_h \frac{1}{\overline{M}_h} \frac{1}{2} \sum_{i=1}^{2} M_{hi} \overline{x}_{hi} \tag{2.28}$$

Ahora el estimador lleva el subíndice "BI" para recordar que se aplica al muestreo bietápico. Por otra parte, a continuación se aclara la nomenclatura de (2.28).

- $W_h = \frac{N_h}{N}$, donde N_h es el número de conglomerados del estrato h ($h = 1, 2, ..., L$), N es el total de conglomerados, es decir, $N = \sum_{h=1}^{L} N_h$.
- M_{hi} es el tamaño del conglomerado i-ésimo ($i = 1, 2$) en el estrato h.
- $\overline{M}_h$ es el tamaño promedio de los conglomerados del estrato h. Sea $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$, entonces,

$$\overline{M}_h = \frac{M_{h0}}{N_h}.$$

- $\overline{x}_{hi}$ es la media muestral del conglomerado i-ésimo ($i = 1, 2$) en el estrato h. Es decir, si en el conglomerado hi se muestrean m_{hi} elementos,

$$\overline{x}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} x_{hij}.$$

Para aplicar la forma general para calcular la varianza de repeticiones balanceadas, (2.10) y (2.11), se debe comenzar por revisar el estimador que se basa en k medias-muestras balanceadas. Se supone que δ_h^j se define como antes y que se cumple (2.18). La media estimada por la media-muestra número j se expresa como:

$$\begin{aligned} \overline{x}_{BI}^{(j)} &= \sum_{h=1}^{L} W_h \frac{1}{\overline{M}_h} \frac{M_{h1}\overline{x}_{h1}(1+\delta_h^{(j)}) + M_{h2}\overline{x}_{h2}(1-\delta_h^{(j)})}{2} \\ &= \overline{x}_{BI} + \sum_{h=1}^{L} \frac{W_h}{\overline{M}_h} \delta_h^{(j)} \frac{M_{h1}\overline{x}_{h1} - M_{h2}\overline{x}_{h2}}{2} \end{aligned} \tag{2.29}$$

Con el estimador de la varianza ocurre algo similar a lo visto en el caso de muestreo estratificado simple.

$$v_{\text{BI}}^{k} = \sum_{j=1}^{k} \frac{(\overline{x}_{\text{BI}}^{(j)} - \overline{x}_{\text{BI}})^2}{k} = \frac{1}{k} \sum_{j=1}^{k} \sum_{h=1}^{L} \frac{W_h^2}{\overline{M}_h^2} \frac{M_{h1}\overline{x}_{h1} - M_{h2}\overline{x}_{h2}}{4} \tag{2.30}$$

Más adelante, en la sección 2.2.5. se explica cómo se puede llevar a cabo la aplicación de repeticiones balanceadas para estimar la varianza, de una forma directa, sin importar lo complicado del diseño o el estimador. También se hacen algunos comentarios sobre la posibilidad de aplicar esta técnica en diseños que no se ajustan exactamente al planteado, el cual se basa en dos conglomerados por estrato.

2.2.5 Generalizaciones

Interesa saber, cuán complicada es la aplicación de repeticiones balanceadas para obtener estimadores no lineales en muestreo complejo. Rust y Rao (1996), dan un procedimiento con el que se resuelve cualquier estimador bajo cualquier diseño. Se consideran los factores de expansión para cada unidad elemental muestreada; por ejemplo, en un muestreo bietápico estratificado por conglomerados, se puede hablar de w_{hij}, el factor de expansión de la unidad j-ésima, del conglomerado i-ésimo del estrato h. Cada uno de estos pesos se altera de la siguiente forma[7]:

$$w_{hij}^{(k)} = \left\{ \begin{array}{ll} (2-\lambda)w_{hij} & \text{Si el conglomerado i-ésimo del estrato } h \\ & \text{está en la media-muestra } k \\ (\lambda)w_{hij} & \text{en otro caso.} \end{array} \right\}$$

Posteriormente se obtiene el estimador $\hat{\Theta}_i$, para cada una de las medias-muestras (i=1,...,k), con los pesos ajustados. Por ejemplo, si la estadística de interés es un vector de coeficientes de regresión, β, en el modelo aplicado a los datos de una muestra:

$$y_{\mathbf{S}} = X_{\mathbf{S}}\beta + \epsilon_{\mathbf{S}}$$

donde $y_{\mathbf{S}}$ es el vector de respuestas en la muestra, $X_{\mathbf{S}}$ es la matriz de variables independientes o auxiliares y $W_{\mathbf{S}}$ es una matriz diagonal que contiene los factores de expansión de cada unidad elemental muestreada; entonces el estimador de β es[8]:

$$\hat{\beta} = (X'_{\mathbf{S}}W_{\mathbf{S}}X_{\mathbf{S}})^{-1}X'_{\mathbf{S}}W_{\mathbf{S}}y_{\mathbf{S}}$$

Quiere decir, que en cada media-muestra se obtendría un $\beta^{(k)}$ ajustando la matriz de pesos, $W_{\mathbf{S}}$ a $W_{\mathbf{S}}^{(k)}$, siguiendo la misma regla. Luego se obtiene la varianza de $\hat{\beta}$ mediante la expresión de la varianza que considera la modificación Fay, (expresión 2.23). Lo mismo se haría para cualquier otro estimador. Se debe apreciar que resulta mucho más fácil calcular un mismo estimador muchas veces, que encontrar y proporcionar las fórmulas adecuadas al personal que se encarga del cómputo en el análisis de una encuesta.

Se puede decir que una debilidad de esta técnica es el hecho de que se enfoque a casos donde el tamaño de muestra en el estrato es $n_h = 2$, pues esta situación limita sus posibilidades de aplicación. Sin embargo, la revisión bibliográfica revela que este procedimiento es muy estudiado y utilizado en el Departamento del Censo de Estados Unidos. L.R. Ernst

[7]Se sigue la notación de la sección 3.2.2.

[8]Ver a Kott (1994).

y T.R. Williams (1987), presentan un estudio muy singular, donde se colapsan estratos, de manera que pueden aplicar la técnica de repeticiones balanceadas.

El colapso de estratos consiste en unir o *colapsar* dos o más estratos en uno, de manera que se crea una nueva partición del conjunto de datos. Usualmente esta práctica se lleva a cabo cuando se tienen estratos con una sola unidad primaria de muestreo, pues en tal situación no se suele conseguir un estimador insesgado ni consistente de la varianza, ni siquiera para estadísticas lineales.

En este trabajo no se quiere entrar en detalles sobre el estimador de estratos colapsados, sino solamente mostrar que existen alternativas de acción ante un problema práctico, como el que presentan Ernst y Williams (1987). Para mayor información sobre este estimador se puede consultar a dichos autores y Wolter (1985, págs. 47- 54), o Cochran (1977).