Capítulo 1

Resultados de inferencia en muestreo

Ya que el cálculo de varianza en un muestreo complejo, representa un problema en torno a una estimación, resulta conveniente señalar, aunque sea brevemente, ciertos resultados de inferencia estadística en el caso particular del muestreo en una población finita. Básicamente, la inferencia estadística es una sola, sin embargo, en el caso del muestreo se tienen dos particularidades que impiden la aplicación directa de sus leyes: en primer lugar, el análisis se refiere a una población finita, en la que sus elementos pueden ser debidamente identificados (o etiquetados) y en segundo lugar, las observaciones no son independientes e idénticamente distribuidas; una situación particular que explica el último punto es la correlación entre datos de un mismo conglomerado. Además de estos aspectos, se añade el hecho de que en inferencia (general) no se habla sobre el concepto de *diseño*. Si se reflexiona un poco sobre este punto, quizás solamente se halle similitud con el manejo de condiciones experimentales en el área de diseño de experimentos. Sin embargo, en todo uso de estadística hay *diseño*, si se entiende por ello, la forma de obtener los elementos de estudio, los procedimientos de medición, etcétera. Lamentablemente, los textos de inferencia no lo mencionan, y todo se desarrolla en torno a variables aleatorias independientes e idénticamente distribuidas (iid), de donde se supone un muestreo aleatorio simple de poblaciones infinitas, o con reemplazo de poblaciones finitas. Pero no se dice que en realidad en poblaciones infinitas no se puede tomar una muestra aleatoria, sino que se toma de algún modo y se supone que es aleatoria y las variables iid. Por otra parte, en poblaciones finitas, se supone que la aleatoriedad la introduce el investigador al tomar muestras probabilísticas. En otras palabras, se dice que hay un proceso que produce la aleatoriedad.

Este capítulo pretende señalar, dentro del marco de la inferencia, qué propiedades se pueden pedir de un estimador de varianza en un muestreo complejo y encontrar los elementos que determinen la utilidad de estimarla. A continuación se abordan varios puntos que son fundamentales en cualquier problema de estimación en inferencia.

1.1 Verosimilitud *vs.* diseño

En el ámbito del muestreo, existen dos corrientes conceptuales, las cuales se catalogan como: *inferencia por aleatorización o por diseño* e *inferencia basada en modelo* (Smith, 1993) o bien

como: *acercamiento de población fija* y *acercamiento de superpoblación* (Cassel, Sarndal, Wretman, 1977) respectivamente. Aplicando el enfoque de *inferencia por aleatorización*, las respuestas de interés Y_i no son aleatorias. Se consideran constantes o parámetros, por lo que se cuenta un vector de N parámetros, $\mathbf{y} = (\mathbf{y_1}, ..., \mathbf{y_N})$, de los cuales, a través del muestreo, se conocerán n de ellos, bajo el supuesto de que no hay error en la medición.

Resulta conveniente hacer explícita cierta notación, la cual se apega a Cassel, Särdal y Wretman (1977, pág. 29):

$\mathbf{y} = (y_1, y_2, ..., y_N)$	=	Valores (fijos) de interés en la población de tamaño N.
$k \epsilon \mathbf{s}$		Notación que indica que k es un componente de la secuencia **s**.
$\mathbf{s} = (k_1, ..., k_{n(S)}) = (k : k \epsilon \mathbf{s})$	=	Secuencia de unidades no necesariamente distintas.
$n(S)$	=	Tamaño de la muestra completa.
$s = (k : k \epsilon s)$	=	Muestra de unidades distintas, sin orde
$\mathbf{S}$	=	Secuencia aleatoria que toma los valores $\mathbf{s} \epsilon \Omega *$.
S	=	Secuencia aleatoria que toma valores s $\epsilon \Omega$.
$\Omega* = \{ \mathbf{S} \}$	=	Conjunto de todas las muestras ordenadas, **s**
$\Omega = \{S\}$	=	Conjunto de todas las muestras no-ordenadas, s.
$\mathbf{D} = ((k, y_k) k \epsilon \mathbf{S})$	=	Var. aleatoria de datos ordenados.
$D = ((k, y_k); k \epsilon S)$	=	Var. aleatoria de datos donde no se repiten unidades.
$\mathbf{d} = ((k, y_k) k \epsilon \mathbf{s})$	=	Realización de **D**.
$d = ((k, y_k); k \epsilon s)$	=	Realización de D.

Siendo y fija, ocurre que la variable aleatoria es entonces,

$$I(\mathbf{s}) = [I(1, \mathbf{s}), ..., I(N, \mathbf{s})]$$

$$\text{donde, } I(i, \mathbf{s}) = \left\{ \begin{array}{ll} 1 & \text{si el elemento i } \epsilon \ \mathbf{s}, \\ 0 & \text{en otro caso.} \end{array} \right\}$$

para, i= 1,...,N.

Es decir, lo aleatorio es la selección de los individuos en lugar de la característica que se pretende medir, de donde se deriva el hecho de que la distribución muestral de un estimador

(bajo este enfoque) depende de la acción del estadístico, a través del diseño que elige. Ahora bien, recordando que se entiende por *verosimilitud* la probabilidad de observar una muestra, vista como función de los parámetros y denotando a la probabilidad de obtener una muestra en particular como P(s), se tiene que

$$L(y) = \left\{ \begin{array}{ll} P(\mathbf{s}) & \forall\, y_i \in \Omega \ \ (\mathrm{i}=1,\ldots,\mathrm{N}) \\ 0 & \text{en otro caso.} \end{array} \right\}$$

Sin embargo, bajo este esquema de pensamiento, resulta que la verosimilitud es constante en relación a los parámetros, y como consecuencia, la inferencia basada en la verosimilitud es imposible.

La inferencia *basada en modelo*, como lo cataloga Smith, sí se basa en la verosimilitud, pero ésta se refiere a un modelo que se supone para la población. De tal forma, basándose en los valores observados en la muestra, se pretende estimar los no muestreados. Las críticas a este enfoque radican, principalmente, en la suposición de un modelo en la población.

Por el contrario, la inferencia por aleatorización está desligada de la verosimilitud. Resulta interesante advertir que si se aplicara el principio de verosimilitud al caso de una encuesta, se tendría que para dos diseños muestrales, tales que las probabilidades de las observaciones $s(y_i,\ i \in S)$ en uno sean un múltiplo de las probabilidades en el segundo diseño, cualquier inferencia sobre la población, debería ser la misma en ambos casos.

Lo anterior significa que de aplicarse el principio de verosimilitud, las inferencias **no** deben depender del diseño de muestreo, ni siquiera a través de las probabilidades $P(\mathbf{s})$, de las muestras observadas. Quien conoce un poco los métodos generalizados en la práctica de muestreo se puede percatar que éstos se ubican en el enfoque de *aleatorización* o *población fija* y que son inconsistentes con el principio de verosimilitud. Los estimadores ampliamente utilizados dependen funcionalmente del diseño y se buscan criterios como insesgamiento y reducción de la varianza.

Existe otro planteamiento basado en la verosimilitud de los datos observados, y_s, mediante una distribución hipergeométrica multivariada (Royall, 1968), que permite, hasta cierto punto, encontrar estimadores máximo verosímiles del número de elementos de la población que poseen determinado valor de la variable de interés. Consecuentemente, se puede encontrar un estimador máximo verosímil para la media y otros estimadores. Sin embargo, esta solución difícilmente se da para otro diseño que no sea el aleatorio simple o el aleatorio simple estratificado.

Resumiendo lo que compete al interés de este trabajo, se tiene que los estimadores a los cuales se hará referencia no son *máximo verosímiles*, como no lo son los métodos de estimar su varianza, más importante aún, hay que ubicar que el enfoque del que parten los métodos discutidos en este trabajo es el de *población fija*. Pero, ¿qué justificación existe para usar estimadores que consideren el diseño? Como ya se mencionó, bajo este enfoque, lo aleatorio no son los valores de la variable observada, sino el proceso de selección. De tal forma, al introducir la dependencia hacia el diseño de muestreo en el estimador, se introduce información sobre la aleatorización. En consecuencia, no sólo se habla de determinado estimador, $\hat{\Theta}$, sino que se especifica lo que se llama una estrategia de muestreo $H(D, \hat{\Theta})$, donde D representa el diseño de muestreo. Se puede ser más explícito aún indicando la

estrategia como: $H(D(S,P),\hat{\Theta})$, pues el diseño está dado por las posibles muestras que se pueden obtener, S, y su probabilidad, P.

Por último, cabe señalar que el interés en una encuesta no está en conocer aquellos valores de la población que no se midieron en la muestra, lo que interesa es alguna función de todos los valores de la población. De ahí se deriva la lógica frecuentista en la que se basa el enfoque de *población fija*, de imaginar repetidas muestras del mismo tamaño y con el mismo diseño. Para hacer inferencia se define un estimador y se evalúa la distribución de este estimador en repetidas muestras. Más adelante, en la sección 1.4, se describen las bases establecidas por Hansen, Madow y Tepping para hacer inferencia en muestreo, basados en este enfoque.

1.2 Suficiencia y varianza mínima

En un problema de estimación en inferencia, los conceptos de *suficiencia* y *suficiencia minimal* son centrales. Es de interés saber si existen resultados sobre suficiencia en el caso de muestreo. La relación directa con el tema de este trabajo es saber si se pueden obtener estimadores de menor varianza a través de la aplicación del teorema de *Rao-Blackwell* y quizás un estimador uniformemente con varianza mínima (UMVUE). De existir esa posibilidad, tal debe ser el estimador a utilizarse, y se procedería con el cálculo de su varianza.

Siendo precisos, sí existen resultados sobre suficiencia en muestreo de poblaciones finitas. Existe un estimador suficiente minimal para el vector de parámetros de la población $y = (y_1, y_2, ...y_N)$, pero éste no es completo, por lo que es posible conseguir un estimador de menor varianza (aunque a veces no resulta un estimador práctico), pero no existe un UMVUE en la clase de todos los estimadores, dada por Λ. Más adelante se verá que al restringir la clase de estimadores, sí existe UMVUE, siendo un caso particular, la media muestral.

El estimador suficiente minimal para $y = (y_1, y_2, ...y_N)$ está dado por la variable aleatoria que representa el conjunto de los datos etiquetados, en el que no se repiten unidades. Esta variable está dada por $D = ((k, y_k);\ k \epsilon S)$ (incluida en la nomenclatura de la sección anterior). La definición de suficiencia en el caso de muestreo es equivalente al concepto en inferencia, pero conviene explicitarla aquí:

Una estadística $Z = u(\mathbf{D})$ es **suficiente** para el parámetro $y = (y_1, y_2, ...y_N)$, si y sólo si la distribución condicional de $\mathbf{D}$ dado $Z = z$, no depende de y, siendo que su distribución condicional está bien definida.

Una estadística $Z_1 = u_1(\mathbf{D})$ es **suficiente minimal** para $y = (y_1, y_2, ...y_N)$, si y sólo si para cada estadística suficiente $Z = u(\mathbf{D})$, cada partición de P_u es un subconjunto de la partición inducida por Z_1,P_{u1}.

Finalmente, ocurre que $D = u_0(\mathbf{D})$ es suficiente y suficiente minimal para y, cuya demostración se aprecia en Cassel, Särdal y Wretman (1977, págs. 36-38). Este resultado no es del todo satisfactorio, pues simplemente los datos, sin considerar orden ni unidades repetidas, es la estadística suficiente. Como se dijo antes, son posibles algunas aplicaciones del teorema de Rao-Blackwell, pero no ofrecen una solución práctica.

No existe UMVUE en la clase de *todos los estimadores insesgados*, ni en la clase de los *insesgados lineales*, y para casi todos los diseños tampoco lo existe para los *lineales insesgados homogéneos*. Casi todos estos resultados se deben a Godambe (1955) y Godambe y Joshi (1965). Sin embargo, Hege, Hanurav y Lanke[1] vieron que los resultados de Godambe se aplicaban al caso de un diseño que no fuera *de uni-conglomerado*; entendiéndose por un diseño *uni-conglomerado* aquel en el que cualesquiera dos muestras $\mathbf{s}_1$, $\mathbf{s}_2$, donde $p(\mathbf{s}_1) > 0$, $p(\mathbf{s}_2) > 0$, ocurre que las muestras son disjuntas o son equivalentes en el sentido de que ambas contienen el mismo conjunto de unidades distintas. Así pues el mismo Godambe (1965), corroboró que una condición necesaria y suficiente para que exista un UMVUE en la clase de estimadores insesgados lineales homogéneos, es que el diseño sea *uni-conglomerado*. Bajo estas restricciones, se encuentra que el estimador Horvitz-Thompson es UMVUE en la clase de estimadores insesgados para la media. También lo es la media muestral.

En Sukhatme y colaboradores (1984, págs. 38-40) se aprecia la demostración de que restringiendo la clase de estimadores insesgados para la media a los lineales, la media muestral es el estimador con menor varianza, por lo tanto es el mejor estimador lineal insesgado de la media poblacional. También demostraron que para la clase de estimadores del tipo Horvitz-Thompson, dada por

$$T_{HT} = \sum_{i \in s} \beta_i Y_i$$

donde, las β_i son el inverso de las probabilidades de inclusión, la media muestral es el único estimador insesgado y por lo tanto es el mejor estimador insesgado para la media poblacional[2], $\overline{Y}$. En Chaudhuri (1988), Chauduhuri y Vos (1988) y Cassel, Sardal y Wretman (1977), se discuten estos hallazgos, así como aspectos sobre admisibilidad de estimadores, pero para los propósitos de este trabajo no se considera necesario abundar más al respecto.

Con los antecedentes que se han planteado, se entiende que el problema a abordar no tiene salida por los conceptos de inferencia clásica. Si se tiene un diseño complejo y estimadores que no son lineales, no se encuentra un estimador de varianza mínima, que sería el lógico a escoger, ni siquiera se sabe si un estimador insesgado será el más adecuado.

1.3 Propiedades de los estimadores

Otros criterios, quizás menos exigentes, para elegir un estimador en el caso de una población finita, son los de *insesgamiento*, *consistencia* y *error cuádrático medio*. A continuación se explican los criterios mencionados.

Estimador insesgado: cuando el valor promedio del estimador, sobre todas las muestras posibles de tamaño n, es exactamente igual al valor poblacional, se dice que el estimador es insesgado. Es decir, para que sea $\hat{\Theta}$ un estimador insesgado de Θ, entonces:

[1]Citados en Chaudhuri A, *Optimality of Sampling Strategies*, Handbook of Statistics, Vol. 6, Elsevier Science Publishers B.V., 1988, págs. 47-96.

[2]Sukhatme y colaboradores (1984, pág. 81).

$$E(\hat{\Theta}) = \sum_{s \epsilon S} \hat{\Theta}_s P(s) = \Theta$$

donde $\hat{\Theta}_s$ es el valor de $\hat{\Theta}$ que corresponde a la muestra s.

Por otra parte, el sesgo de un estimador está dado por:

$$SESGO(\hat{\Theta}) = E(\hat{\Theta}) - \Theta$$

El sesgo resulta ser de especial interés pues los estimadores de las varianzas de muestreos complejos suelen ser sesgados. Ocurre además que el método *jackknife* se concibió como un estimador para reducir sesgo.

Estimador consistente: en el caso de poblaciones infinitas se dice que un estimador $\hat{\Theta}$ es consistente si:

$$lim_{n \to \infty} \left(P[\mid \hat{\Theta} - \Theta \mid > \epsilon] \right) \longrightarrow 0$$

En otras palabras, el estimador $\hat{\Theta}$ es consistente si asume el valor Θ con una probabilidad que converge a 1, cuando el tamaño muestral aumenta indefinidamente. Cabe advertir que un estimador consistente no necesariamente será insesgado, pero si es sesgado, el sesgo tenderá a cero mientras n tiende a infinito. Ahora bien, para el caso de poblaciones finitas, Sukhatme *et al.* (1984), define un estimador consistente como aquel que asume el valor de Θ cuando se toma la población total como muestra, es decir, cuando $n = N$.

Los estimadores usados con mayor frecuencia son consistentes pero no insesgados. Sin embargo, un estimador inconsistente puede ser útil, pues puede tener buena precisión cuando n es pequeña comparada con N.

Error cuadrático medio: para un estimador $\hat{\Theta}$, éste se define como:

$$ECM_{\Theta}\left[\hat{\Theta}\right] = E(\hat{\Theta} - \Theta)^2 = V(\hat{\Theta}) + SESGO^2$$

El error cuadrático medio es útil para comparar dos estimadores sesgados o un insesgado con un sesgado.

1.4 Bases de inferencia en muestreo

Los intervalos de confianza para datos que provienen de muestreos complejos, basados en la distribución *Normal(0,1)* o en la distribución t de *Student* son de uso generalizado. Sin embargo, el por qué se hacen de esta manera, a veces no es claro para los usuarios de estadística.

Hansen, Madow y Tepping (1983, págs. 776-793), describen los principios de inferencia desde el punto de vista de la aleatorización (enfoque de *población fija*). Definen un diseño de muestreo probabilístico como aquél que consiste de:

a) Un plan de muestreo tal que cada miembro de la población tiene una probabilidad conocida, mayor a cero, de ser incluido en la muestra.

b) Procedimientos de inferencia tales que para muestras razonablemente grandes, la exactitud de las inferencias no dependen de la suposición de un modelo.

De esta conceptualización se deriva el hecho de que las inferencias se basan en distribuciones límites que son inducidas por la aleatorización; por lo tanto, un intervalo de confianza, en este contexto, es válido si se interpreta en términos de la probabilidad de aleatorización. Es decir, la probabilidad de que un intervalo de confianza contenga el valor estimado, es igual o mayor al valor nominal del nivel de confianza, **bajo la probabilidad inducida** por la forma en que se hizo la selección.

Así pues, las inferencias usualmente se derivan de la cantidad pivotal:

$$t = \frac{\hat{\Theta} - \Theta}{\sqrt{v(\hat{\Theta})}}$$

En la expresión de arriba, $v(\hat{\Theta})$ es un estimador de la varianza de $\hat{\Theta}$, y se supone que t se distribuye, al menos aproximadamente, como una N(0,1). Su aproximación a la distribución t de *Student* es más correcta, pero como las muestras en encuestas son grandes, se tiende a utilizar automáticamente la distribución normal. Sin embargo, en el caso en que la varianza se calcula por métodos de remuestreo, los grados de libertad son mucho menores y se debe observar el uso de la distribución t de *Student*. Más adelante en este trabajo se muestra que estos intervalos son válidos para varios de los métodos discutidos. En particular, en la sección 1.6 se ven resultados para el *jackknife*, *repeticiones balanceadas* (*Half- sampling*) y *linealización*.

Si se tiene un intervalo de confianza, es lógico querer transformarlo para construir una prueba de hipótesis. Sin embargo, se ha visto que cuando las muestras son grandes, tales pruebas no son adecuadas. "Si la muestra tiene un tamaño suficientemente grande, la relación más débil y menos significativa aparecerá con significancia estadística"[3]. Este planteamiento ha sido por años uno de los puntos debatidos por los estadísticos bayesianos. Sin duda, lo razonable es expresar resultados en intervalos de confianza. Pero, con la reciente difusión de los métodos de estimación por remuestreo, se ha visto que éstos proveen una forma robusta de hacer pruebas de significancia[4]. Anteriormente, se señaló que al aplicar estas técnicas, los grados de libertad son mucho menores, lo que hace a un lado el problema que representa un tamaño de muestra tan grande, es decir, el hecho de que en todas las pruebas de significancia con gran tamaño de muestra todo sea significativo aunque no sea importante. El cálculo de grados de libertad bajo diferentes métodos de remuestreo se muestra en el Apéndice B, citando a Rust y Rao (1996), pues es la única referencia que se encontró al respecto.

[3]Palabras de Yates citadas por Kish en: Kish, L., Muestreo de Encuestas, Trillas, 1972, pág. 678.

[4]Rust K.F., Rao J.N.K., *Variance Estimation for Complex Surveys Using Replication Techniques*, Statistical Methods in Medical Research, 1996; 5, págs. 283-310.

1.5 Aspectos de interés sobre la varianza

En el marco de inferencia, la varianza es uno de los parámetros que define la distribución sobre la cual se quiere inferir y es en este aspecto, en el que radica la importancia de estimarla. Asimismo, en el caso de muestreo de poblaciones finitas, se vio, en la sección anterior, que el estimador de la varianza o la desviación estándar juega un papel importante para establecer intervalos de confianza y en algunos casos, hacer pruebas de hipótesis. Al respecto conviene citar a Rao (1997): "ignorar el diseño de muestreo y analizar los datos de forma convencional puede llevar a una seria sub-estimación de la desviación estándar, niveles de pruebas inflados y diagnósticos erróneos".

Es claro entonces, que bajo el esquema de aleatorización es fundamental la consideración del diseño en los estimadores de varianza. A su vez, la estimación de la varianza en muestreos complejos tiene importancia como medida de la precisión del diseño; esto permite establecer comparaciones entre distintos diseños. Obviamente, el interés de una encuesta no es estimar una varianza, pero sí lo es obtener varios estimadores de valores en la población, y éstos se quieren conocer con la mayor precisión posible; además, se quiere conocer qué tan precisos (cercanos a θ) son los posibles valores de $\hat{\theta}$. La comparación de diferentes diseños es útil, sobre todo, en situaciones en las que se requiere periódicamente levantar una encuesta similar, pues se puede optar por cambios en el diseño para un futuro proyecto. Por ejemplo se puede apreciar si cierta estratificación tuvo o no el efecto deseado en reducción de varianza.

Lo más común es la comparación con el muestreo aleatorio simple con reemplazo y sin reemplazo. Las medidas más difundidas en la práctica para este efecto son: el **DEFF**, cuando se compara contra el m.a.s. sin reemplazo (m.a.s.s.r.), y el $DEFT^2$, si se compara contra el m.a.s. con reemplazo (m.a.s.c.r.). Tales medidas son definidas por Kish (1989), de la siguiente manera:

$$DEFF(\hat{\Theta}) = \frac{Var(\hat{\Theta}_{\text{diseño}})}{Var(\hat{\Theta}_{\text{m.a.s.s.r.}})} \tag{1.1}$$

$$DEFT^2(\hat{\Theta}) = \frac{Var(\hat{\Theta}_{\text{diseño}})}{Var(\hat{\Theta}_{\text{m.a.s.c.r.}})} \tag{1.2}$$

Las verdaderas varianzas son desconocidas en la mayoría de los casos, por lo que normalmente, sólo se obtienen estimadores de dichas medidas, substituyendo en el numerador y denominador los estimadores de las varianzas correspondientes. En ocasiones esta medida sirve para simplificar la presentación de resultados y evitar la necesidad de cálculos extensos para estimar varianzas. Cuando hay elementos para pensar que algunos estimadores de la encuesta tienen un DEFF similar (usualmente se sabe por encuestas anteriores) sólo se exhiben estimadores basados en m.a.s. - los que calcula un paquete de cómputo fácilmente - y se indica el valor de DEFF para ese grupo de variables. De esta forma, quien está interesado en conocer una varianza o un intervalo, estima la varianza mediante el producto de la varianza de m.a.s. y el DEFF. Esta simplificación de resultados puede ser muy objetada, porque depende mucho de la debida inspección de encuestas anteriores y la correcta

agrupación de variables; sin embargo es una práctica muy difundida, y que se debe, en parte, a las dificultades que enfrenta el equipo de análisis de una encuesta. Kish (1989), discute ampliamente la utilidad de esta medida.

Se ha hablado de la necesidad de conocer la varianza, pero tras la breve exposición de resultados de inferencia en muestreo, resulta de interés establecer bajo un enfoque de *población fija*, qué propiedades se pueden buscar en un estimador de varianza cuando se usa un muestro complejo. Las propiedades que se revisan en los estimadores de varianza en general, son: insesgamiento y consistencia. En el caso de estimadores no lineales, en ocasiones se busca que el método de estimación, de ser aplicado a una estadística lineal (por ejemplo, la media), produzca el mismo estimador que se obtendría por la fórmula, según el diseño, y que la esperanza del estimador de varianza sea también la conocida para el caso. Obviamente, esto no garantiza que al aplicarse el método a una estadística no lineal, se obtengan las mismas propiedades.

Por otra parte, es conveniente aclarar que muchas veces se habla de varianza o *error cuadrático medio* (ECM) casi indistintamente. Se sabe que:

$$ECM(\hat{\Theta}) = Var(\hat{\Theta}) + Sesgo^2(\hat{\Theta}).$$

Cuando el $Sesgo^2(\hat{\Theta})$ es de un orden mucho menor que $Var(\hat{\Theta})$, una aproximación para $ECM(\hat{\Theta})$ de primer orden, también lo sería para $Var(\hat{\Theta})$. Por tal razón, muchas veces lo que se obtiene es una aproximación del ECM y se toma como estimación o expresión para la varianza. Así pues, en ocasiones, se elige entre dos posibles estimadores de acuerdo al que tenga menor error cuadrático medio. En el caso de estimadores de varianza, cabe la posibilidad de comparar $ECM(\hat{Var}(\hat{\Theta}))$, pero llega a ocurrir que un estimador con menor ECM, no da los mejores intervalos, es decir, la probabilidad de inclusión en el intervalo es menor al valor nominal. Lo que se verá al respecto en este trabajo (en la sección 4.3.) es que el comportamiento de los métodos de remuestreo, han sido inspeccionados a través de simulaciones por distintos autores, ya que no es posible hacerlo de forma analítica.

Es interesante mencionar que el mismo problema que representa el cálculo de varianza en un muestreo complejo, se da para calcular una matriz de varianza-covarianza. Estas matrices son necesarias, especialmente, cuando se desea hacer un análisis multivariado. Nathan (1988), considera que el método de *linealización*, el de *repeticiones balanceadas* y el *jackknife* producen buenos estimadores para estas matrices.

Como se advirtió en la introducción, son pocos los resultados de inferencia clásica que servirán de apoyo en la discusión de los métodos de cálculo de varianza. No existen resultados en los que se pueda basar la elección de un estimador, pero sí existen otros que permiten hacer inferencia sobre los estimadores $\hat{\Theta}$, basándose en $\hat{Var}(\hat{\Theta})$, obtenida por distintos métodos.

1.6 Resultados específicos sobre los métodos de remuestreo

Krewski y Rao (1981), sentaron las bases de la teoría asintótica del estimador de linealización, el *jackknife* y el de repeticiones balanceadas (*Half-Sampling*). Esta teoría se da en el marco de una secuencia infinita de poblaciones finitas, $\{\prod_L\}_{L=1}^{\infty}$ con L estratos en $\{\prod_L\}$. Los resultados son válidos en el contexto de un diseño estratificado multietápico en el que las unidades primarias de muestreo se seleccionan *con reemplazo* y en las que se toman muestras independientes, cuando una unidad resulta seleccionada más de una vez. Además, el análisis de Krewski y Rao se avoca al caso en el que se tienen muchos estratos con pocas unidades primarias en cada uno de éstos.

El interés radica en parámetros de la forma $\Theta = g(\overline{\boldsymbol{Y}})$, y el estimador correspondiente, $\hat{\Theta} = g(\overline{\boldsymbol{y}})$, siendo $g(.)$ una función real, $\overline{\boldsymbol{Y}}$ la media poblacional, ($\overline{\boldsymbol{Y}}= \sum_{h=1}^{L} W_h \overline{\boldsymbol{Y}}_{\boldsymbol{h}}$), $\overline{\boldsymbol{y}}$ un estimador insesgado de $\overline{\boldsymbol{Y}}$ ($\overline{\boldsymbol{y}}= \sum_{h=1}^{L} W_h \overline{\boldsymbol{y}}_{\boldsymbol{h}}$), y W_h el peso del estrato h, dado por la proporción de unidades en el estrato relativas al total de ellas en la población ($W_h = \frac{N_h}{N}$). Se denomina por $\boldsymbol{Y_{hi}}= (Y_{hi1},...,Y_{hip})'$, los valores poblacionales de la unidad i-ésima del estrato h, por $\boldsymbol{y_{hi}}= (y_{hi1},...,y_{hip})'$ los valores muestrales respectivos donde p es el número de variables; de manera similar, $\overline{\boldsymbol{Y}}_{\boldsymbol{h}}= (\overline{Y}_{h1},...,\overline{Y}_{hp})'$ y $\overline{\boldsymbol{y}}_{\boldsymbol{h}}= (\overline{y}_{h1},...,\overline{y}_{hp})'$, son los vectores de medias poblacionales y muestrales de los estratos. La matriz de varianza-covarianza de $\overline{\boldsymbol{y}}$ está dada por:

$$\Sigma = \sum_{h=1}^{L} W_h^2 n_h^{-1} \Sigma_h$$

donde Σ_h es una matriz $(p \times p)$ cuyo elemento típico es de la siguiente forma:

$$\sigma_{hj,hj'} = N_h^{-1} \sum_{i=1}^{N_h} (Y_{hij} - \overline{Y}_{hj})(Y_{hij'} - \overline{Y}_{hj'})$$

Por otra parte, se supone que la secuencia de poblaciones satisface las siguientes condiciones de regularidad:

i) $\sum_{h=1}^{L} W_h E\{|\, y_{hij} - \overline{Y}_h j|^{2+\delta}\} = 0(1)$, para algún $\delta > 0$, $j = (1,...,p)$.

ii) $max_{1 \le h \le L}\{n_h\} = 0(1)$

iii) $max_{1 \le h \le L}\{W_h\} = 0(L^{-1})$

iv) $n \sum_{h=1}^{L} W_h^2 n_h^{-1} \Sigma_h \rightarrow \Sigma_*$, donde Σ_* es p x p, definida positiva.

v) $\overline{Y}_k \rightarrow \mu_k$(finita) para $k=1,...,p$.

vi) Las primeras derivadas $g_j(.)$ de $g(.)$, son continuas en una vecindad de $\mu = (\mu_1,...,\mu_p)$.

Sea $v_L(\hat{\Theta})$, $v_J(\hat{\Theta})$, $v_{RB}(\hat{\Theta})$ los estimadores de varianza obtenidos por linealización, *jackknife* y repeticiones balanceadas respectivamente; así también, t_L, t_J y t_{RB} son estadísticas

que parten de los mismos métodos. Bajo las condiciones anteriores, Krewski y Rao (1981) demostraron los siguientes resultados:

a)

$$\begin{aligned} nv_L(\hat{\Theta}) &\xrightarrow{p} \sigma_\Theta^2 \\ t_L = \frac{\hat{\Theta}-\Theta}{\sqrt{v_L(\hat{\Theta})}} &\xrightarrow{d} N(0,1) \end{aligned}$$

b)

$$\begin{aligned} nv_J(\hat{\Theta}) &\xrightarrow{p} \sigma_\Theta^2 \\ t_J = \frac{\hat{\Theta}-\Theta}{\sqrt{v_J(\hat{\Theta})}} &\xrightarrow{d} N(0,1) \end{aligned}$$

c)

$$\begin{aligned} nv_{RB}(\hat{\Theta}) &\xrightarrow{p} \sigma_\Theta^2 \\ t_{RB} = \frac{\hat{\Theta}-\Theta}{\sqrt{v_{RB}(\hat{\Theta})}} &\xrightarrow{d} N(0,1) \end{aligned}$$

Al examinar las condiciones iniciales del teorema, se aprecia que para considerar válidos estos resultados, se requiere del supuesto de que el número total de unidades $n = \sum n_h$, sea grande, y que ningún factor de expansión sea desproporcionadamente mayor a los demás. Los resultados expresan que los tres estimadores de varianza son asintóticamente consistentes para la verdadera varianza y que los intervalos de confianza asociados a la teoría normal también son válidos asintóticamente.