UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

ANÁLISIS COMPARATIVO DE ESTIMACIÓN DE INTERVALOS DE CONFIANZA PARA PROPORCIONES.

TESIS

QUE PARA OBTENER EL GRADO ACADÉMICO DE MAESTRO EN CIENCIAS MATEMÁTICAS

PRESENTA

JOSÉ GUSTAVO RODRÍGUEZ JIMÉNEZ.

DIRECTORA DE TESIS: DRA. GUILLERMINA ESLAVA GÓMEZ

MÉXICO,D.F. ENERO, 2009.

INTRODUCCIÓN

La utilización de muestras y encuestas es un recurso muy usado en los diversos ámbitos de la ciencia, donde el objetivo es la obtención de estimaciones confiables sobre algún parámetro poblacional de interés. Para este fin es necesario contar con estimadores de alta precisión, es decir que cuenten con errores de estimación pequeños.
El trabajo que aquí se presenta centra su atención en una clase muy particular de estimadores, que tienen una estructura matemática de tipo fraccional, conocidos como estimadores de razón.
Se ilustra una aplicación de estos estimadores en el cálculo de proporciones, y de diferencias de proporciones; que es una situación de interés en el caso de desear aproximar las preferencias electorales de una población con la finalidad de identificar un ganador.
Con el fin de estimar de manera más precisa la varianza de los estimadores, y con ellos construir intervalos de confianza, en la aplicación se hace uso de técnicas de remuestreo.
Dichas técnicas se utilizan generalmente cuando,

- Es más fácil implementar los métodos de remuestreo, que programar las fórmulas correspondientes al estimador de varianza.
- No existe fórmula analítica para el estimador de varianza.

Mediante trabajo computacional, las técnicas de remuestreo nos permiten

aproximar el error de estimación incluso cuando la muestra ha sido seleccionada bajo un diseño muestral complejo, es decir cuando la muestra está constituida posiblemente en varias etapas, cuenta con conglomeramiento y quizá también con estratificación.

Antes del surgimiento de las técnicas de remuestreo se utilizaba una metodología que permitió dar una aproximación del error mediante el uso de la fórmula de Taylor, pero cuenta con sus propias limitaciones que más adelante presentaremos.

En la práctica cada técnica tiene sus ventajas y sus desventajas pero a fin de cuentas debe existir congruencia entre unos métodos y otros.
El objetivo del presente trabajo es corroborar la congruencia entre los resultados obtenidos mediante los métodos de remuestreo tomando como base los resultados del método aproximación por Linealización.

Para ese fin se compararon los intervalos de confianza generados con remuestreo en repetidas simulaciones y con diversos tamaños de muestra. Contrastándolos con el intervalo que brinda el método de Linealización en cada caso.

Este trabajo se divide en 6 capítulos; incluyendo esta introducción y las conclusiones. En el capítulo 1 iniciaremos el desarrollo de la teoría necesaria dando una visión general de la definición de un diseño muestral, se presentan las propiedades óptimas del estimador de razón, así como el concepto de muestreo estratificado. En el capítulo 2 se presentan a manera de resumen los estimadores para proporciones bajo la perspectiva de no contar con un muestreo complejo, se estudia el método de Linelización y se ilustra, con un ejemplo, cómo se utiliza la técnica con estimadores de razón.

El capítulo 3 se dedica al estudio de los estimadores de razón en diseños estratificados así también como sus propiedades. En el capítulo 4 se exponen dos técnicas de remuestreo, el *Jackknife* y el *Bootstrap*. También se presentan las herramientas con las que se mide la estabilidad de los estimadores, las cuales se usarán más adelante para establecer comparaciones, que es nuestro principal objetivo. Para terminar, en el capítulo 5 se presentan los resultados. Para que sea más ilustrativo el trabajo se incluyen algunas gráficas y tablas. En el capítulo 6 se presentan algunas conclusiones.

RESUMEN

En este trabajo se compara el desempeño de los intervalos de confianza producidos con los estimadores de razón combinado y de post-estratificación o separado, la comparación se hace tomando en cuenta estimadores de varianza que se usan en remuestreo. Para esto se establecen medidas como lo son el sesgo relativo de la varianza, estabilidad relativa de la varianza, longitud de los intervalos, etc. Estas medidas se basan en un número grande de simulaciones, las cuales consistieron en obtener muestras, calculando en cada una de ellas los estimadores de razón, estimadores de diferencias de razones, así como también sus respectivos estimadores de varianza. Las muestras obtenidas fueron extraídas mediante un diseño muestral estratificado.
Se consideraron tres tamaños de muestra: 600, 900 y 1500 unidades. Se compararon los intervalos de confianza producidos con la técnica de Linealización de Taylor como base, con dos técnicas de remuestreo, el *Jackknife* y el *Bootstrap*.

El objetivo principal de este trabajo es identificar cuáles métodos de estimación de varianza producen estimaciones más precisas en el sentido de cumplir con todos o la mayoría de los criterios de comparación.

Se pretende que a partir de este trabajo, se contemple el método que mejor resulte aquí como la opción inmediata a aplicarse en el futuro.

Los motivos para elegir los dos procedimientos de remuestreo que aquí se presentan son los siguientes.

- La programación resulta ser muy sencilla.
- La teoría sobre los métodos señala que son eficientes para aproximar varianzas.

En este trabajo se explica brevemente la teoría detrás de cada uno de los métodos y su implementación práctica. También se da una breve descripción de las ventajas y desventajas de cada uno de los métodos. Finalmente se hace una aplicación utilizando los datos generados en el PREP[1], de la elección para presidente de la República Mexicana del año 2000, se usaron a nivel sección electoral aún cuando aparecen a nivel casilla. Los datos del PREP se encuentran contenidos en el disco compacto, sistema de consulta, **Estadísticas de las elecciones Federales de México 2000**. Instituto Federal Electoral. Versión 1.0, México 2000.

[1]Para más información ver apéndice D

Índice general

Índice de figuras

Índice de Tablas

Capítulo 1

Preliminares

Al hablar de muestreo nos referimos al conjunto de técnicas estadísticas que estudian la forma de seleccionar una muestra probabilística de una población, cuya información permita inferir algunas características de interés de toda la población, cometiendo un error medible y acotable en términos probabilísticos. A partir de la muestra, seleccionada mediante un determinado método de muestreo, se estiman características poblacionales de cierta variable de interés, por ejemplo, media, total, proporción, etc., con un error cuantificable y acotable en términos probabilísticos. Las estimaciones se realizan a través de funciones matemáticas de la muestra denominadas estimadores, que se convierten en variables aleatorias al considerar la aleatoriedad de las muestras. Los errores debidos al muestreo generalmente se cuantifican mediante sesgos y varianzas, error estándar o errores cuadráticos medios de los estimadores, que miden la precisión de éstos.

Existen varios tipos de muestreo, dependiendo de que la población sea finita o infinita, materia sobre la que existe amplia literatura, pero en este trabajo solo se considera el muestreo en poblaciones finitas. La población finita que se desea investigar se denomina población objetivo.

Por otro lado, para seleccionar una muestra, es necesario un listado de unidades de muestreo denominado marco muestral, que teóricamente debiera coincidir con la población objetivo, éste no siempre existe. En este trabajo suponemos que ambos coinciden.

Por otra parte es muy importante tener en cuenta que para medir la precisión es necesario usar muestreo probabilístico. Recordemos que se dice que el muestreo es probabilístico cuando pueda establecerse la probabilidad de obtener cada una de las muestras que sea posible seleccionar, esto es, cuando todos y cada uno de los elementos de la población objetivo tienen una probabilidad conocida y positiva de ser seleccionados. A continuación se presentan elementos básicos en la teoría de muestreo cuando la forma de selección de la muestra es probabilística.

1.1. Diseño muestral

Consideremos los sucesos elementales asociados a un fenómeno aleatorio dado $s_1, \ldots, s_m$. Entendiendo por sucesos elementales los más simples posibles, es decir, aquellos que no pueden ser descompuestos en otros sucesos. El conjunto $\{s_1, \ldots s_m\}$ se denomina espacio muestral asociado al fenómeno.
Si consideramos como fenómeno la extracción aleatoria de muestras dentro de una población por un procedimiento o método de muestreo dado, podemos considerar como sucesos elementales las muestras obtenidas, constituyendo el conjunto de las mismas el espacio muestral.
Habitualmente en los métodos de muestreo comúnes, se consideran iguales las muestras con los mismos elementos, aunque estén colocados en orden diferente (el orden de colocación no interviene).
Una muestra aleatoria de tamaño n extraída de una población $U = \{u_1, \ldots, u_N\}$

de tamaño N mediante un método de muestreo dado, suele denotarse como

$$s = \{u_1, \ldots, u_n\}.$$

De esta forma, el conjunto de las $\binom{N}{n}$ muestras posibles sin reemplazo de tamaño n que se pueden formar con los N elementos de la población U es el espacio muestral s.

Evidentemente, para establecer la probabilidad de todas las muestras posibles derivadas de un procedimiento de muestreo dado, será necesario conocer ese conjunto de muestras; es decir, será necesario delimitar tanto el método de muestreo como el espacio muestral derivado del mismo.

Así, un diseño muestral es sencillamente una función mediante la que se seleccionan las muestras de modo que cada una tenga una determinada probabilidad de ser elegida. Por lo tanto, el diseño muestral empleado para seleccionar la muestra, define en el espacio muestral a L, una función de probabilidad $\mathbb{P}$ que representa:

$$L : \{s_1, s_2, \ldots, s_k\} \rightarrow \mathbb{P}(.) : \{\mathbb{P}(s_1), \mathbb{P}(s_2), \ldots, \mathbb{P}(s_k)\}$$

Donde k es el número de muestras posibles y además $\mathbb{P}$ satisface las siguientes propiedades:

1. $\mathbb{P}(s_i)\ 0$.
2. $\sum_{s_i \in L} \mathbb{P}(s_i) = 1$.

Esta función L juega un papel central porque determina propiedades estadísticas esenciales, como la distribución de muestreo, el valor esperado, la varianza y otras propiedades de los estimadores calculados a partir de la muestra.

Como ejemplo se puede considerar el muestreo aleatorio simple sin reemplazo. Consiste en seleccionar elementos del universo de manera secuencial como a continuación se describe:

1. Seleccionar con igual probabilidad, $\frac{1}{N}$, un primer elemento de N en la población.
2. Seleccionar con igual probabilidad, $\frac{1}{N-1}$, un segundo elemento de los $N-1$ restantes.

$\vdots$

n. Seleccionar con igual probabilidad, $\frac{1}{N-n+1}$, un n-ésimo elemento de los $N-n+1$ elementos restantes después de $N-1$ selecciones.

Un conjunto específico s con n elementos puede ser obtenido como resultado de $n!$ diferentes sucesiones (equiprobables).
El diseño muestral es entonces:

$$\mathbb{P}(s) = \begin{cases} 1/\binom{N}{n} & \text{Si } s \text{ tiene } n \text{ elementos.} \\ \\ 0 & \text{En otro caso.} \end{cases}$$

Visto como una función, lo podemos representar como sigue.

$$L : \{s_1, s_2, \ldots, s_k, \ldots, s_{\binom{N}{n}}\} \to \mathbb{P}(.) : \{1/\binom{N}{n}, 1/\binom{N}{n}, \ldots, 1/\binom{N}{n}, \ldots, 1/\binom{N}{n}\},$$

El aleatorio simple es uno de los diseños muestrales más sencillos que se pueden encontrar, y sirve como base de comparación con los otros diseños muestrales por lo que juega un papel muy importante en la teoría de muestreo.

Existen otros diseños usuales en los que no pondremos atención en este trabajo como son: muestreo Bernoulli, Poisson, sistemático, muestreo con probabilidad de selección proporcional al tamaño, etc.

1.2. Muestreo estratificado

Con frecuencia, tenemos información adicional de los elementos poblacionales que nos ayuda a diseñar una muestra. Si las variables de interés asumen valores promedio diferentes en diferentes subpoblaciones, es posible obtener estimaciones más precisas al tomar una muestra estratificada.
El fundamento es la partición de la población en H subpoblaciones que llamaremos estratos, dentro de los cuales exista más homogeneidad que en la población completa en relación a la característica que se estudia. Así, matemáticamente la estructura de estratos no es otra cosa que una partición de la población, $\{U_1, \ldots, U_H\}$ que satisface.

1. $U = \bigcup_{h=1}^{H} U_h$.

2. $U_h \cap U_l = \emptyset \ \ \forall h \neq l \ \ h, l : 1, \ldots, H$,

la reducción de varianza en las estimaciones se conseguirá siempre que se procure que haya homogeneidad dentro de $U_1, U_2, \ldots U_H$. Como cada subpoblación tiene N_h elementos, el tamaño de la población es entonces

$$N = \sum_{h=1}^{H} N_h$$

En su aspecto más general, el muestreo estratificado se realizará aplicando, de forma independiente, y en cada estrato, un diseño muestral, para obtener la muestra s_h de tamaño n_h, la unión de todas estas muestras dará lugar a la muestra total s.
Matemáticamente $\{s_1, \ldots, s_H\}$ satisface.

1. $s = \bigcup_{h=1}^{H} s_h$.

2. $s_h \cap s_l = \emptyset \ \ \forall h \neq l \ \ h, l : 1, \ldots, H$.

Esto nos lleva a que la muestra global s consta de $n = \sum_{h=1}^{H} n_h$ elementos. Posteriormente, reunimos la información de cada estrato para obtener las estimaciones globales de la población.

Cuando se usa el mismo diseño en cada estrato entonces el diseño general lo podemos llamar de manera específica como por ejemplo:

- **STPPS**: Estratificación con probabilidad de selección proporcional al tamaño del estrato y con reemplazo
- **STΠPS**: Estratificación con probabilidad de selección proporcional al tamaño del estrato y sin reemplazo
- **STSY**: Estratificación con selección sistemática en cada estrato.
- **STBE**: Estratificación con selección Bernoulli
- **STSI**: Estratificación con selección aleatoria simple en cada estrato

El diseño muestral STSI es el que se utiliza en este trabajo.

Finalmente podemos mencionar que el muestreo estratificado tiene las siguientes ventajas:

- Si las mediciones dentro de cada estrato son homogéneas, la estratificación producirá un valor más pequeño (δ) para el error de estimación.
- Se puede reducir el costo por observación al estratificar la población en grupos convenientes.
- Puede facilitar el trabajo de campo.
- Es más fácil controlar los errores muestrales y no muestrales.

Ahora vamos a hablar un poco de las propiedades de los estimadores.

1.3. Estimadores y sus propiedades

Consideremos a Y como una variable aleatoria con una distribución de probabilidad, que depende de un parámetro desconocido θ, consideremos a $Y_1, Y_2, \ldots, Y_n$ como una muestra de Y en donde para cada unidad en la muestra se determina una medición, i.e. $Y_1(u_1) = y_1, \ldots, Y_n(u_n) = y_n$. A estos valores muestrales los denotaremos simplemente con $y_1, y_2, \ldots, y_n$.

Los estimadores son funciones de la muestra que permiten la estimación del parámetro desconocido θ. A tal efecto entenderemos como estimador a cualquier variable aleatoria $g(Y_1, Y_2, \ldots, Y_n)$, o simplemente g una función de la muestra, que se defina a partir de la sucesión de variables aleatorias $Y_1, Y_2, \ldots, Y_n$, que integran una muestra de tamaño n extraída de manera probabilística de la población, es decir, toma un valor para cada n observaciones.

Deberemos valorar en un estimador su capacidad de extraer al máximo la información contenida en la muestra, ya que redundará en la calidad y precisión de las estimaciones.

El valor que toma g, es decir $g(y_1, y_2, \ldots, y_n)$, se conoce como una estimación de θ o valor estimado, y habitualmente se escribe,

$$\widehat{\theta} = g(y_1, y_2, \ldots, y_n).$$

Desde este punto de vista podemos decir que un estimador $\widehat{\theta}$ es una variable aleatoria de la que nos interesarán sus características, principalmente de centralización y dispersión, particularmente su esperanza, su varianza y sus momentos, así como otras medidas relativas a su precisión.

Para analizar la precisión de un estimador suelen usarse varios conceptos como son los siguientes:

1. Sesgo.
2. Consistencia.
3. Consistencia en error cuadrático medio.

A continuación se explican estos y otros conceptos.

Sea $\widehat{\theta}$ un estimador del parámetro desconocido θ, entonces, $\widehat{\theta}$ es un estimador insesgado para θ, si

$$\forall\theta \in \Theta, \quad E(\widehat{\theta}) = \theta.$$

Θ es el espacio paramétrico de θ.

Un estimador que no satisface la propiedad anterior se dice que es sesgado, y entonces existe una cantidad llamada sesgo del estimador que está dada por:

$$B(\widehat{\theta}) = E(\widehat{\theta}) - \theta.$$

Frecuentemente se relaja un poco la condición de insesgamiento y se reemplaza por la siguiente.
Considere $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_n$ una sucesión de estimadores de θ, en donde
$\widehat{\theta}_1 = \widehat{\theta}(X_1)$, $\widehat{\theta}_2 = \widehat{\theta}(X_1, X_2)$,..., $\widehat{\theta}_n = \widehat{\theta}(X_1, \ldots, X_n)$. Entonces, $\widehat{\theta}_n$ es un estimador asintóticamente insesgado para θ si

$$\forall\theta \in \Theta, \quad \lim_{n\to\infty} E(\widehat{\theta}_n) = \theta.$$

Es deseable que un estimador sea insesgado o asintóticamente insesgado, aunque puede haber ocasiones en las cuales podríamos preferir estimadores sesgados; veremos este caso más adelante. A fin de hacer una buena elección

en tales casos presentamos el siguiente concepto.

Se dice que $\widehat{\theta}_n$ es un estimador consistente de θ si

$$\forall \epsilon > 0 \quad \lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > \epsilon\right) = 0.$$

El concepto de consistencia será muy importante más adelante en este trabajo por la siguiente razón:
Suponga que los estimadores $\widehat{\theta}_1, \ldots \widehat{\theta}_r$ son consistentes para $\theta_1, \ldots, \theta_r$. Entonces existen algunas funciones f, $f(\widehat{\theta}_1, \ldots \widehat{\theta}_r)$ consistentes para $f(\theta_1, \ldots, \theta_r)$. En otras palabras funciones de estimadores consistentes son consistentes, Särndal *et al.* (1991, p.168).

Por ejemplo, si $\widehat{\bar{y}}_U$ y $\widehat{\bar{z}}_U$ son estimadores consistentes para las medias poblacionales $\bar{y}_U$ y $\bar{z}_U$ respectivamente, entonces $\frac{\widehat{\bar{y}}_U}{\widehat{\bar{z}}_U}$ será consistente para $\frac{\bar{y}_U}{\bar{z}_U}$.

Para terminar esta sección se define al error cuadrático medio (ECM) del estimador $\widehat{\theta}_n$ de la siguiente forma

$$ECM(\widehat{\theta}_n) = E(\widehat{\theta}_n - \theta)^2 = V(\widehat{\theta}_n) + B^2(\widehat{\theta}_n).$$

Así podemos definir el siguiente concepto:
Se dice que $\widehat{\theta}_n$ es un estimador consistente en error cuadrático medio de θ

$$si\ \forall \theta \in \Theta, \quad \lim_{n\to\infty} ECM(\widehat{\theta}_n) = 0.$$

Capítulo 2

Estimación de proporciones y de sus varianzas

2.1. Estimación de proporciones en muestreo aleatorio simple

Un estadístico utilizado para variables dicotómicas es la proporción de casos en una categoría, que representa la frecuencia relativa de una categoría con respecto al total.
En el caso de que se esté estudiando una única variable y y que ésta sea dicotómica se codifica dicha variable en ceros y unos. El valor uno se suele reservar para el código con el que se quiere designar la ocurrencia de la característica de interés, y cero si no la tiene, así.

$$y_i = \begin{cases} 1 & \text{si la } i \text{ ésima unidad tiene la característica de interés.} \\ 0 & \text{si no la tiene.} \end{cases}$$

Así, la proporción de unidades que tienen cierta característica es sólo un caso especial de la media.

$$p = \frac{\text{(No. de unidades con la característica en la población)}}{N} = \frac{1}{N}\sum_{y_i \in U} y_i.$$

Es natural usar la proporción que se ha calculado a partir de una muestra para estimar p.

$$\widehat{p} = \frac{1}{n}\sum_{y_i \in s} y_i.$$

Esta proporción claramente es un estimador lineal de la proporción poblacional bajo muestreo aleatorio simple.

Una manera de estimar la proporción poblacional p no es dando un único valor, sino un rango de valores posibles para p, determinando entonces un intervalo alrededor de $\widehat{p}$. Este intervalo contendrá o no al verdadero parámetro p, pero los procedimientos estadísticos vistos en la sección precedente aseguran que el intervalo lo incluya con un nivel de confianza del $(1-\alpha)100\,\%$.

Es claro que bajo muestreo aleatorio simple y si $n \to \infty$ entonces $\widehat{p} \dot{\backsim} N(p, \widehat{V}(\widehat{p}))$, donde $\widehat{p}$ es el estimador de máxima verosimilitud de p.

Utilizando el principio de invarianza de estos estimadores, tenemos que

$$\widehat{V}(\widehat{p}) = \frac{\widehat{p}(1-\widehat{p})}{n-1}.$$

y en consecuencia,

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}} \dot{\backsim} N(0,1).$$

Esta cantidad pivotal da origen a un intervalo que recibe el nombre de intervalo de confianza de *Wald*. Es derivado de la aproximación Normal y se calcula con la siguiente fórmula.

$$\widehat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n-1}}.$$

En la literatura, es conocido que dicho intervalo tiene serios problemas en dos circunstancias.

1. Cuando p está cercano a 0 ó 1.

2. Cuando el tamaño de muestra n es pequeño.

Una regla empírica útil para poder usar el intervalo de *Wald* consiste en satisfacer

$$n\widehat{p} > 5 \;\; y \;\; n(1-\widehat{p}) > 5$$

simultáneamente.

Alternativamente bajo muestreo aleatorio simple existen otros intervalos planteados más adelante por Lawrence D.B., T. T.Cai. y A. DasGupta (2002).
A continuación presentamos estos intervalos bajo muestreo aleatorio simple.
Consideremos primero las cantidades siguientes.

$$\tilde{Y} = \sum_{y_i \in s} y_i + \frac{z^2_{(1-\frac{\alpha}{2})}}{2}, \;\; \tilde{n} = n + z^2_{(1-\frac{\alpha}{2})}, \;\; \tilde{p} = \frac{\tilde{Y}}{\tilde{n}}, \;\; \tilde{q} = 1 - \tilde{p}.$$

En donde $z_{(1-\frac{\alpha}{2})}$ es el cuantil de una Normal estándar.
Entonces,

1. El intervalo de confianza de *Wilson* se define mediante.

$$\tilde{p} \pm \frac{z_{(1-\frac{\alpha}{2})}\sqrt{n}}{n + z^2_{(1-\frac{\alpha}{2})}} \sqrt{\widehat{pq} + \frac{z^2_{(1-\frac{\alpha}{2})}}{4n}}.$$

2. El intervalo de confianza de *Agresti-Coull*[1] es el siguiente.

$$\tilde{p} \pm z_{(1-\frac{\alpha}{2})} \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}.$$

3. El intervalo de confianza de Razón de Verosimilitud.
   Este intervalo es construido por inversión del cociente de verosimilitud,

[1]Referencia [1]

bajo la hipótesis nula $H_0 : p = p_0$, $-2\log(\Lambda_n) \dot{\backsim} \chi^2_{(1)}$, donde Λ_n es el cociente de verosimilitud

$$\Lambda_n = \frac{p_0^{\sum y_i}(1-p_0)^{n-\sum y_i}}{\widehat{p}^{\sum y_i}(1-\widehat{p})^{n-\sum y_i}}.$$

4. El intervalo de confianza de *Jeffreys*.
   Se construye mediante el teorema de Bayes, considerando una distribución inicial para p, Beta, $\beta(\frac{1}{2}, \frac{1}{2})$ que es considerada no informativa, obteniendo.

$$\left(B(\frac{\alpha}{2}, \sum_{y_i \in s} y_i + \frac{1}{2}, n - \sum_{y_i \in s} y_i + \frac{1}{2}), B(1 - \frac{\alpha}{2}, \sum_{y_i \in s} y_i + \frac{1}{2}, n - \sum_{y_i \in s} y_i + \frac{1}{2}) \right)$$

   Donde $B(\alpha, m_1, m_2)$ denota el α cuantil de una distribución $\beta(m_1, m_2)$, como Lawrence D.B.*et al.* refieren[2].

El estudio completo de los intervalos lo podemos encontrar en el trabajo de A.DasGupta *et al.* (2002) que acabamos de mencionar.
Adicionalmente, podemos mencionar el trabajo de Böhning y Viwatwongkasen (2005) que presenta otros intervalos, en el caso que $\widehat{p}$ es una función lineal de los datos muestrales analiza y estudia sus propiedades.

Hasta aquí dejamos los estimadores lineales de p bajo muestreo aleatorio simple, pues éste no es el objeto de estudio de este trabajo, que se centra en un caso en que el estimador $\widehat{p}$ no es una función lineal de la muestra.

[2]Referencia [37]

2.2. Estimación de varianza

A continuación presentamos un método analítico de uso muy común para aproximar varianzas.

2.2.1. Método de Linealización por series de Taylor para la estimación de varianzas

Tepping (1968)[3], fue el primero en sugerir el uso de Linealización por series de Taylor para la estimación de varianza en muestras complejas.
La estimación de medias y coeficientes de regresión lineal en muestras complejas fue presentada por Kish y Frankel (1974) y Folson (1974).
Woodruff (1971) presentó la aplicación general del método de Linealización a funciones explícitas de los datos observados.
Binder (1985) brindó un tratamiento más formal y riguroso a la Linealización en muestras complejas indicando cómo usarla cuando el parámetro de interés no se pueda expresar como una función explícita de los datos, además de probar la normalidad asintótica para estas estimaciones.
Básicamente la metodología es la siguiente. Supóngase que el parámetro poblacional de interés θ es de la forma $\theta = f(\mathbf{t})$, donde $\mathbf{t} = (t_1, \ldots, t_p)$ es un vector de parámetros poblacionales.
Se considera entonces $\widehat{\theta} = f(\widehat{\mathbf{t}})$ en donde $\widehat{\mathbf{t}} = (\widehat{t_1}, \ldots, \widehat{t_p})$ y $\widehat{t_i}$ es un valor estimado en la muestra. Se supone adicionalmente que $f(\widehat{\mathbf{t}})$ tiene derivadas de segundo orden en una vecindad de $\widehat{\mathbf{t}}$ que contiene a $\mathbf{t}$, entonces θ puede ser expandido en serie de Taylor alrededor de $\widehat{\mathbf{t}} = \mathbf{t}$ de la siguiente forma.

$$\widehat{\theta} = f(\mathbf{t}) + \sum_{j=1}^{p} \frac{\partial f}{\partial \widehat{t_j}}(\mathbf{t})(\widehat{t_j} - t_j) + Res(\mathbf{t}, \widehat{\mathbf{t}}). \tag{2.1}$$

Donde, $Res(\mathbf{t}, \widehat{\mathbf{t}})$ es el residuo, el cual se desecha por contener términos de

[3]Referencia [46]

orden superior al segundo.
Por lo tanto

$$ECM(\widehat{\theta}) \doteq V(\widehat{\theta}) = V\left(\sum_{j=1}^{p} \frac{\partial f}{\partial \widehat{t}_j}(\mathbf{t})(\widehat{t}_j - t_j)\right) = \dots$$

$$\dots = \sum_{j=1}^{p}\sum_{i=1}^{p} \frac{\partial f}{\partial \widehat{t}_j}(\mathbf{t})\frac{\partial f}{\partial \widehat{t}_i}(\mathbf{t})C(\widehat{t}_j, \widehat{t}_i). \tag{2.2}$$

En donde C es la covarianza entre estimadores.

A continuación se presenta un ejemplo de la utilización del método.

Ejemplo 1, Linealización:
Supóngase que se requiere encontrar la varianza por Linealización para el estimador de un cociente de totales poblacionales. En donde t_1 y t_2 son totales de dos variables de interés.

$$R = \frac{t_1}{t_2},$$

Se propone como su estimador a,

$$\widehat{R} = \frac{\widehat{t}_1}{\widehat{t}_2} = f(\widehat{t}_1, \widehat{t}_2).$$

Primero se calculan las derivadas parciales de la función,

$$\frac{\partial \widehat{R}}{\partial \widehat{t}_1} = \frac{1}{\widehat{t}_2},\ \frac{\partial \widehat{R}}{\partial \widehat{t}_2} = -\frac{\widehat{t}_1}{\widehat{t}_2^2}$$

después evaluamos las parciales en el punto $\mathbf{t} = (t_1, t_2)$,

$$\frac{\partial \widehat{R}}{\partial \widehat{t}_1}(\mathbf{t}) = \frac{1}{t_2},\ \frac{\partial \widehat{R}}{\partial \widehat{t}_2}(\mathbf{t}) = -\frac{t_1}{t_2^2} = -\frac{R}{t_2}.$$

con lo que podemos obtener la expansión de Taylor del Estimador $\widehat{R}$ según la ecuación (2.1).

$$\widehat{R} \doteq R + \frac{(\widehat{t_1} - R\widehat{t_2})}{t_2}. \tag{2.3}$$

y la varianza por Linealización según la ecuación (2.2) es,

$$V(\widehat{R}) \doteq \frac{1}{t_2^2}V(\widehat{t_1}) - \frac{R}{t_2^2}C(\widehat{t_1}, \widehat{t_2}) - \frac{R}{t_2^2}C(\widehat{t_2}, \widehat{t_1}) + \frac{R^2}{t_2^2}V(\widehat{t_2}).$$

Que nos lleva a una expresión muy conocida en la literatura,

$$V(\widehat{R}) = V(\frac{\widehat{t_1} - R\widehat{t_2}}{t_2}) \doteq \frac{1}{t_2^2}\left[V(\widehat{t_1}) + R^2V(\widehat{t_2}) - 2RC(\widehat{t_1}, \widehat{t_2})\right]. \tag{2.4}$$

Como ventajas del uso de este método podemos mencionar las siguientes:

1. Puede ser aplicado en cualquier diseño muestral.
2. La teoría está bien desarrollada.
3. Existen muchos programas de software que calculan los estadísticos por Linealización, por ejemplo.
   - **AM Software** de *American Institutes for Research.*
   - **Bascula** de *Statistics Netherlands.*
   - **CLUSTERS** de *University of Essex.*
   - **Epi Info** de *Centers for Disease Control.*
   - **Generalized Estimation System (GES)** de *Statistics Canada.*
   - **IVEware** de *University of Michigan.*
   - **PCCARP** de *Iowa State University.*
   - **R survey package** de *R Project.*

- **SAS** de *SAS Institute.*
- **SPSS Complex Samples** de *SPSS Inc.*
- **Stata** de *Stata Corporation.*
- **SUDAAN** de *Research Triangle Institute.*
- **VPLX** de *U.S. Bureau of the Census.*
- **WesVar** de *Westat, Inc.*

Desventajas:

1. Encontrar las derivadas parciales puede ser difícil.
2. Se necesita una fórmula diferente para cada estadística.
3. La exactitud de la aproximación por Linealización depende del tamaño de muestra, y la estimación de varianza con frecuencia tiene un sesgo hacia abajo, se subestima si la muestra no es suficientemente grande, Särndal *et al.* (1991, p.176).

En esta sección presentamos un método de estimación de varianza en el cual el parámetro de interés debe ser una función de totales poblacionales de las variables utilizadas.
En la siguiente sección se presenta una forma muy general de estimar totales poblacionales, también se presenta la estimación de la varianza de estos estimadores de totales.

2.2.2. El estimador Horvitz-Thompson

En 1952 los investigadores D.G. Horvitz y D.J. Thompson plantearon un estimador para el cálculo de totales con una forma funcional muy sencilla, que puede utilizarse cuando se hace uso de esquemas de muestreo con y sin reemplazo en poblaciones finitas. Este estimador es importante porque es un estimador lineal que es insesgado y consistente.

Para definir el estimador considere a π_j como la probabilidad de que una unidad u_j de la población U pertenezca a la muestra, $\pi_j = \mathbb{P}(u_j \in s)$. Los valores π_j son conocidos también como probabilidades de inclusión de primer orden. También considere a π_{ij} como la probabilidad de que las unidades u_i y u_j pertenezcan a la muestra, $\pi_{ij} = \mathbb{P}(u_i, u_j \in s)$. Los valores π_{ij} son conocidos también como probabilidades de inclusión de segundo orden.

Así, se define a

$$\widehat{t}_{y\pi} = \sum_{y_k \in s} \frac{y_k}{\pi_k}.$$

como el estimador del total

$$t_y = \sum_{y_k \in U} y_k.$$

Llamado estimador de total de Horvitz y Thompson. Es común que este estimador se presente en los libros con la siguiente forma.

$$\widehat{t}_{y\pi} = \sum_{y_k \in s} w_k y_k.$$

en donde las cantidades $w_k = \frac{1}{\pi_k}$ son llamadas pesos muestrales o factores de expansión.

La varianza del estimador de Horvitz y Thompson es

$$V(\widehat{t}_{y\pi}) = \sum_{y_k \in U} \sum_{y_l \in U} (\pi_{kl} - \pi_k \pi_l) \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Un estimador consistente de la varianza bajo el supuesto de que $\pi_{ij} \neq 0$, $\forall i, j : 1, \ldots, N$. es el siguiente.

$$\widehat{V}_1(\widehat{t}_{y\pi}) = \sum_{y_k \in s} \sum_{y_l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

Alternativamente presentamos el estimador de varianza debido a Yates-Grundy (1953) y Sen (1953), para diseños de tamaño de muestra fijo:

$$\widehat{V}_2(\widehat{t}_{y\pi}) = -\frac{1}{2} \sum_{y_k \in s} \sum_{y_l \in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2.$$

Cochran (1977, p.261) mencionó que ambos estimadores pueden asumir valores negativos para algunos diseños muestrales y configuraciones específicas de los valores muestrales, además mencionó también que $\widehat{V}_2$ es más estable que $\widehat{V}_1$ cuando el diseño muestral tiene tamaño de muestra fijo. $\widehat{V}_2$ es más estable en el sentido de que presenta valores más pequeños en el coeficiente de variación con respecto a los que presenta el coeficiente de variación de $\widehat{V}_1$.

Como ventaja principal del estimador de Horvitz y Thompson podemos decir que se puede implementar en cualquier diseño muestral que satisfaga la hipótesis sobre las probabilidades de inclusión de segundo orden. En el muestreo sistemático, por ejemplo, no se puede aplicar porque $\pi_{ij} = 0$ para algunos valores de i, j.
Sin embargo, hay que observar que en las fórmulas anteriores es necesario que $\pi_{kl} > 0 \ \forall u_k, u_l \in U \ k, l : 1, \ldots, N$ y para cualquier muestra que se seleccione.

En general resulta muy complicado el cálculo de varianza de este estimador, salvo que el diseño muestral sea muy sencillo, el muestreo aleatorio simple es un ejemplo clasico de un diseño muestral sencillo. En la práctica el estimador de Horvitz y Thompson es muy socorrido solo para obtener el estimador puntual y el cálculo de su varianza se hace por aproximación mediante Linealización.

El estimador de Horvitz y Thompson en la estimación de proporciones

En lo sucesivo del presente trabajo una situación de especial interés se tiene cuando se cuenta con una variable y que únicamente toma los valores 0 y 1. En dicha situación $\bar{y}_U = \frac{t_y}{N}$ se convierte automáticamente en la proporción de ocurrencias de y. Pero como es usual no se conoce muchas veces el numerador, en otras no se conoce el denominador y en algunas otras no se conocen ambas. Sin embargo solo son interesantes dos situaciones:

1. El numerador se desconoce pero el denominador, N, es una cantidad conocida con lo que obtenemos un estimador lineal.
$$\widehat{p} = \frac{\widehat{t}_{y\pi}}{N} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k}.$$

2. El numerador se desconoce y el denominador también, entonces obtenemos un estimador no lineal.
$$\widehat{R} = \frac{\widehat{t}_{y\pi}}{\widehat{N}} = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}}.$$

Como podemos observar en ambas situaciones se utilizan para aproximar los totales de las variables a los estimadores de Horvitz y Thompson.

En el caso de un muestreo aleatorio simple ambos estimadores se reducen a calcular las medias muestrales de la variable y, como podremos observar a continuación. En un muestreo aleatorio simple $\pi_i = \frac{n}{N}$ y $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$, al sustituir estos valores obtenemos.

$$\widehat{p} = \frac{\widehat{t}_{y\pi}}{N} = \frac{1}{N}\sum_{k\in s}\frac{y_k}{\pi_k} = \frac{1}{N}\frac{\sum_{k\in s} y_k}{\frac{n}{N}} = \frac{1}{n}\sum_{y_k\in s} y_k.$$

$$\widehat{R} = \frac{\widehat{t}_{y\pi}}{\widehat{N}} = \frac{\sum_{k\in s}\frac{y_k}{\pi_k}}{\sum_{k\in s}\frac{1}{\pi_k}} = \frac{\frac{N}{n}\sum_{i=1}^{n} y_k}{\sum_{i=1}^{n}\frac{N}{n}} = \frac{1}{n}\sum_{y_k\in s} y_k.$$

Särndal *et al.* (1991,p.182) señalan que muchas veces aún cuando N es conocida, el estimador $\widehat{R}$ es mejor estimador que $\widehat{p}$, es decir, usar $\frac{\widehat{t}_{y\pi}}{\widehat{N}}$ es preferible a usar $\frac{\widehat{t}_{y\pi}}{N}$.

Ya vimos las posibilidades existentes en el caso de diseño aleatorio simple y un estimador lineal para la estimación de proporciones, también vimos el cálculo de estimadores y sus varianzas. En el siguiente capítulo veremos una alternativa en el caso de que el estimador de proporción no es lineal y considerando un diseño muestral complejo como lo es el estratificado aleatorio simple.

Capítulo 3

Estimadores de razón para proporciones

3.1. Definición del estimador

Hasta el momento se ha presentado a la proporción p en su forma habitual,

$$p = \frac{1}{N} \sum_{y_k \in U} y_k$$

haciendo que y_k tome valores 0 y 1.
La proporción p se puede presentar de manera alternativa como un cociente de totales poblacionales,

$$R = \frac{t_y}{t_z} = \frac{\sum_{y_k \in U} y_k}{\sum_{z_k \in U} z_k}.$$

En donde $z_k = 1$ para $k = 1, \ldots, N$.

Cuando deseamos estimar p, podemos estimar ambos totales mediante el estimador de Horvitz y Thompson visto en el capítulo anterior, obteniendo el siguiente estadístico.

$$\widehat{R} = \frac{\widehat{t}_{y\pi}}{\widehat{t}_{z\pi}} = \frac{\sum_{y_k \in s} \frac{y_k}{\pi_k}}{\sum_{z_k \in s} \frac{z_k}{\pi_k}}.$$

A este estadístico se le conoce comúnmente como estimador de la razón.

3.2. Propiedades del estimador de razón

Esta clase de estimadores resulta ser sesgado pero consistente, sin embargo, el valor de $\widehat{R}$ es bastante cercano a R cuando n es suficientemente grande. Para eso Raj (1968) demuestra la siguiente desigualdad.

$$\frac{\left(E(\widehat{R}) - R\right)^2}{V(\widehat{R})} \leq \frac{V(\widehat{t}_{z\pi})}{{t_z}^2}.$$

Es decir, el sesgo relativo está acotado por el coeficiente de variación del total de la variable en el denominador de la razón. Dicha desigualdad también puede presentarse así.

$$\frac{\left|E(\widehat{R}) - R\right|}{\sqrt{V(\widehat{R})}} \leq CV(\widehat{t}_{z\pi}) = \frac{\sqrt{V(\widehat{t}_{z\pi})}}{E(\widehat{t}_{z\pi})}. \tag{3.1}$$

Sugiere que en la práctica se debe tomar la muestra de modo que el $CV(\widehat{t}_{z\pi})$ sea pequeño. En consecuencia para considerar ignorable el sesgo de los estimadores de razón es recomendable que esté acotado superiormente. Cochran ((1977), p.153) sugiere que

$$CV(\widehat{t}_{z\pi}) < \frac{1}{10}.$$

3.3. Estimación de varianza del estimador de razón

Usando el método de Linealización podemos encontrar fácilmente una fórmula para calcular la varianza del estimador de razón.
En el ejemplo 1, Linealización se encuentra el desarrollo para encontrar la fórmula, de la ecuación (2.4) se sigue que

$$V\left(\widehat{R}\right) \doteq \frac{1}{{t_z}^2}\left[V\left(\widehat{t}_{y\pi}\right) + R^2 V\left(\widehat{t}_{z\pi}\right) - 2RC\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right)\right].$$

Que es la expresión que Särndal *et al.* (1991,p.179) presentan en su libro. Más aún afirman que se puede estimar mediante:

$$\widehat{V}\left(\widehat{R}\right) \doteq \frac{1}{\widehat{t}_{z\pi}^2}\left[\widehat{V}\left(\widehat{t}_{y\pi}\right) + \widehat{R}^2\widehat{V}\left(\widehat{t}_{z\pi}\right) - 2\widehat{R}\widehat{C}\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right)\right].$$

Señala que $\widehat{V}(\widehat{R})$ tiende a subestimar $V(\widehat{R})$ si el tamaño de muestra es pequeño. Este es un problema heredado del método de Linealización.

3.4. Los estimadores de razón en diseños estratificados

Supóngase que tenemos una muestra estratificada s de tamaño n con H estratos. En esta situación existen dos tipos de estimadores de razón que son los siguientes

3.4.1. El estimador de razón combinado

El estimador combinado supone que la razón a estimar en cada estrato es más o menos parecida. Es decir $R_h \approx R$, para $h = 1, \ldots, H$. Tiene la siguiente

expresión.

$$\widehat{\mathcal{R}} = \frac{\sum_{h=1}^{H} \sum_{y_k \in s_h} w_{hk} y_k}{\sum_{h=1}^{H} \sum_{z_k \in s_h} w_{hk} z_k} = \frac{\sum_{h=1}^{H} \widehat{t}_{y\pi h}}{\sum_{h=1}^{H} \widehat{t}_{z\pi h}}.$$

En donde $w_{hk} = \frac{1}{\pi_{hk}}$ son los pesos correspondientes a cada diseño probabilístico utilizado en cada uno de los estratos.

La estimación estratificada por razón combinada presenta como principal ventaja la no acumulación de los sesgos de las estimaciones en los estratos, lo que reduce el sesgo del estimador final.
El principal inconveniente de este método es la imposibilidad de obtención de estimaciones separadas por estratos, lo que no permite disponer de información de la población al subnivel de estratos.
En la práctica suele usarse cuando las muestras en los estratos son de tamaño pequeño y cuando la cantidad de estratos es grande.

Bajo un esquema de muestreo aleatorio simple en cada uno de los estratos, el estimador puede reescribirse de la forma siguiente.

$$\widehat{\mathcal{R}} = \frac{\sum_{h=1}^{H} \sum_{y_k \in s_h} \frac{N_h}{n_h} y_k}{\sum_{h=1}^{H} \sum_{z_k \in s_h} \frac{N_h}{n_h} z_k} = \frac{\sum_{h=1}^{H} \frac{N_h}{N} \overline{y}_{s_h}}{\sum_{h=1}^{H} \frac{N_h}{N} \overline{z}_{s_h}}. \tag{3.2}$$

Se puede apreciar que el estimador es una combinación lineal de medias muestrales en cada estrato tanto en el numerador como en el denominador.
Cochran (1977, p.155) demuestra que mediante Linealización la varianza se puede aproximar mediante.

$$V(\widehat{\mathcal{R}}) \doteq \frac{1}{{t_z}^2} \sum_{h=1}^{H} {N_h}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S^2_{yU_h} + R^2 S^2_{zU_h} - 2RS_{yzU_h}\right], \tag{3.3}$$

en donde

$$S^2_{yU_h} = \frac{1}{N_h - 1} \sum_{y_k \in U_h} \left(y_k - \bar{y}_{U_h}\right)^2,$$

$$S_{zU_h}^2 = \frac{1}{N_h - 1} \sum_{z_k \in U_h} \left(z_k - \bar{z}_{U_h}\right)^2$$

y

$$S_{yzU_h} = \frac{1}{N_h - 1} \sum_{y_k, z_k \in U_h} \left(y_k - \bar{y}_{U_h}\right) \left(z_k - \bar{z}_{U_h}\right).$$

Särndal *et al.* (1991, p.253) aseguran que podemos aproximar al estimador de la varianza mediante:

$$\widehat{V}(\widehat{\mathcal{R}}) \doteq \frac{1}{\widehat{t}_{z\pi}^2} \sum_{h=1}^{H} {N_h}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S_{ys_h}^2 + (\widehat{\mathcal{R}})^2 S_{zs_h}^2 - 2\widehat{\mathcal{R}} S_{yzs_h}\right], \qquad (3.4)$$

en donde

$$S_{ys_h}^2 = \frac{1}{n_h - 1} \sum_{y_k \in s_h} \left(y_k - \bar{y}_{s_h}\right)^2,$$

$$S_{zs_h}^2 = \frac{1}{n_h - 1} \sum_{z_k \in s_h} \left(z_k - \bar{z}_{s_h}\right)^2,$$

y

$$S_{yzs_h} = \frac{1}{n_h - 1} \sum_{y_k, z_k \in s_h} \left(y_k - \bar{y}_{s_h}\right) \left(z_k - \bar{z}_{s_h}\right).$$

Finalmente, el sesgo estandarizado del estimador de razón combinado tiene como límite superior (Cochran 1977, p.165) el mencionado en la ecuación (3.1) el cual no debe sobrepasar 0.1 para que el sesgo sea despreciable.

$$\frac{\left|E(\widehat{\mathcal{R}}) - R\right|}{\sqrt{V(\widehat{\mathcal{R}})}} \leq CV(\widehat{t}_{z\pi}).$$

3.4.2. El estimador de razón separado

Éste se utiliza cuando sospechamos que la razón en cada estrato es diferente. Es decir $R_h \neq R_{h'}$, para $h \neq h'$ y $h, h' = 1, \ldots, H$. También es conocido como estimador de razón de post-estratificación. Cochran (1977, p.165) señala que el tamaño de muestra en cada estrato debe ser grande y que en consecuencia

la aproximación de varianza usando las fórmulas correspondientes está limitada en las aplicaciones prácticas.
Särndal *et al.* (1991, p.270) recomiendan que el tamaño de muestra por estrato debe ser de cuando menos de veinte unidades, $n_h \geq 20$ para $h = 1, \ldots, H$.

Este método presenta como principal inconveniente la acumulación de los sesgos de las estimaciones en los estratos, aumentando el sesgo del estimador final.
El estimador de razón separado es el siguiente.

$$\widehat{\mathcal{B}} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\sum_{y_k \in s_h} w_{hk} y_k}{\sum_{z_k \in s_h} w_{hk} z_k} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\widehat{t}_{\pi yh}}{\widehat{t}_{\pi zh}} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \widehat{\mathcal{R}}_{s_h}.$$

El estimador es una suma de los estimadores de razón combinados de cada estrato, ponderada por el tamaño relativo de cada estrato.
Destaca que este estimador requiere el conocimiento previo de los totales t_{zUh} en cada estrato Cochran (1977, p.164).
Presenta como principal ventaja la obtención de estimaciones separadas por estratos, lo que permite ofrecer información de la población al subnivel de estratos.
Cochran (1977, p.167) afirma que a menos que R_h sea constante en cada uno de los estratos, usar estimaciones de razón por separado en cada uno de los estratos es más preciso siempre que el tamaño de muestra en cada uno de ellos sea grande, con lo que la fórmula de estimación de la varianza por Linealización del estimador separado es una aproximación válida a la varianza de este estimador, y la acumulación de los sesgos por estrato que pueden afectar al estimador de razón separado pueden considerarse ignorables.
De igual forma que en el estimador combinado, vemos que bajo un esquema de muestreo aleatorio simple en cada estrato, el estimador puede reescribirse

de la siguiente forma.

$$\widehat{\mathcal{B}} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\sum_{y_k \in s_h} \frac{N_h}{n_h} y_k}{\sum_{z_k \in s_h} \frac{N_h}{n_h} z_k}. \tag{3.5}$$

Cochran (1977, p.164) demuestra en el teorema 6.4, que mediante Linealización la varianza se puede aproximar mediante.

$$V(\widehat{\mathcal{B}}) \doteq \frac{1}{{t_z}^2} \sum_{h=1}^{H} {N_h}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S^2_{yU_h} + {R_h}^2 S^2_{zU_h} - 2R_h S_{yzU_h}\right]. \tag{3.6}$$

Särndal *et al.* (1991, p.271) aseguran que podemos aproximar al estimador de la varianza mediante.

$$\widehat{V}(\widehat{\mathcal{B}}) \doteq \frac{1}{\widehat{t}^2_{z\pi}} \sum_{h=1}^{H} \left(\frac{\bar{z}_{U_h}}{\bar{z}_{s_h}}\right)^2 {N_h}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S^2_{ys_h} + \widehat{\mathcal{R}}^2_{s_h} S^2_{zs_h} - 2\widehat{\mathcal{R}}_{s_h} S_{yzs_h}\right]. \tag{3.7}$$

Finalmente el sesgo estandarizado del estimador de razón separado debe satisfacer la desigualdad (Cochran (1977),p.166) para que el sesgo del estimador sea despreciable.

$$\frac{\left|E(\widehat{\mathcal{B}}) - R\right|}{\sqrt{V(\widehat{\mathcal{B}})}} \leq \sqrt{H} \frac{\sum_{h=1}^{H} CV(\widehat{t}_{z\pi h})}{H} < \frac{1}{10}.$$

3.5. Estimadores de diferencias de razones

En las encuestas con frecuencia es necesario estimar la diferencia o el cociente entre dos estimadores de razón y en consecuencia también es necesario conocer la varianza del nuevo estadístico, con la finalidad de establecer comparaciones. En este trabajo se analiza la diferencia de estimadores de razón que estén correlacionados de alguna manera.

Para calcular la diferencia se suele usar una combinación lineal de estimadores de razón.

$$\widehat{R_i - R_j} = \widehat{R}_i - \widehat{R}_j, \tag{3.8}$$

aunque los estimadores de razón no lo sean,

$$\widehat{R}_i = \frac{\widehat{t}_{yi}}{\widehat{t}_{zi}}, \quad \widehat{R}_j = \frac{\widehat{t}_{yj}}{\widehat{t}_{zj}}.$$

Utilizando las propiedades de la varianza obtenemos que.

$$V(\widehat{R_i - R_j}) = V(\widehat{R}_i) + V(\widehat{R}_j) - 2C(\widehat{R}_i, \widehat{R}_j). \tag{3.9}$$

Los dos primeros términos de esta última expresión pueden calcularse con las fórmulas expuestas en secciones anteriores.

Es fácil ver que el elemento de covarianza puede aproximarse por
Linealización mediante la siguiente relación[1], Lethonen *et al.* (1994,p.175).

$$C(\widehat{R}_i, \widehat{R}_j) \doteq \frac{1}{t_{z_i} t_{z_j}} [C(\widehat{t}_{y_i}, \widehat{t}_{y_j}) + R_j R_i C(\widehat{t}_{z_j}, \widehat{t}_{z_i}) -$$

$$R_i C(\widehat{t}_{y_j}, \widehat{t}_{z_i}) - R_j C(\widehat{t}_{y_i}, \widehat{t}_{z_j})]. \tag{3.10}$$

Como podemos ver las fórmulas anteriores son generales y solo dependen de los totales de las variables de interés.

La fórmula (3.10), sirve como auxiliar para encontrar la covarianza entre dos estimadores de razón combinados o separados considerando muestreo estratificado aleatorio simple.

Para dos estimadores de razón combinados [2] tenemos que.

[1] Ver demostración en apéndice A
[2] Ver demostración en apéndice B

$$C(\widehat{\mathcal{R}}_j, \widehat{\mathcal{R}}_i) \doteq \frac{1}{t_{z_j} t_{z_i}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} [S_{y_j y_i U_h} + R_j R_i S_{z_j z_i U_h} - R_i S_{y_j z_i U_h} - R_j S_{y_i z_j U_h}]. \tag{3.11}$$

Y para dos estimadores de razón separados [3] tenemos que.

$$C(\widehat{\mathcal{B}}_j, \widehat{\mathcal{B}}_i) \doteq \frac{1}{t_{z_j} t_{z_i}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} [S_{y_j y_i U_h} + R_{hj} R_{hi} S_{z_j z_i U_h} - R_{hi} S_{y_j z_i U_h} - R_{hj} S_{y_i z_j U_h}]. \tag{3.12}$$

Obsérvese que para calcular $V(\widehat{R}_i - \widehat{R}_j)$ es necesario el cálculo del término de covarianza, esto puede resultar ser muy complicado. Para aproximar $\widehat{V}(\widehat{R}_i - \widehat{R}_j)$, también requiere de una buena aproximación de la covarianza entre estimadores. Cochran (1977, p.180) presenta una manera de aproximar $\widehat{V}(\widehat{R}_i - \widehat{R}_j)$ sin usar directamente la covarianza entre estimadores.
Considere nuevamente $\widehat{R}_i = \frac{\widehat{t}_{yi}}{\widehat{t}_{zi}}$ y $\widehat{R}_j = \frac{\widehat{t}_{yj}}{\widehat{t}_{zj}}$ estimadores de razón que tienen el mismo denominador, $\widehat{t}_{zi} = \widehat{t}_{zj} = \widehat{t}_z$ entonces se construye,

$$\widehat{R}_i - \widehat{R}_j := \frac{\widehat{t}_{yi} - \widehat{t}_{yj}}{\widehat{t}_z}, \tag{3.13}$$

un estimador de razón de la diferencia de razones, al cual se le puede aproximar la varianza utilizando la fórmula (3.4) en el caso de tener estimadores combinados, resultando la siguiente expresión.

$$\widehat{V}(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j) \doteq$$

$$\frac{1}{\widehat{t}_{z\pi}^2} \sum_{h=1}^{H} {N_h}^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S^2_{(y_i - y_j)s_h} + (\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j)^2 S^2_{zs_h} - 2(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j) S_{(y_i - y_j)zs_h}\right]. \tag{3.14}$$

Para el estimador separado, (3.13) se puede usar a nivel de estrato resultando,

$$\widehat{V}(\widehat{\mathcal{B}}_i - \widehat{\mathcal{B}}_j) \doteq \frac{1}{\widehat{t}_{z\pi}^2} \ldots$$

[3] Ver demostración en apéndice C

$$\sum_{h=1}^{H}\left(\frac{\bar{z}_{U_h}}{\bar{z}_{s_h}}\right)^2 {N_h}^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}\left[S^2_{(y_i-y_j)s_h}+(\widehat{\mathcal{R}}_{is_h}-\widehat{\mathcal{R}}_{js_h})^2 S^2_{zs_h}-2(\widehat{\mathcal{R}}_{is_h}-\widehat{\mathcal{R}}_{js_h})S_{(y_i-y_j)zs_h}\right]. \tag{3.15}$$

En este trabajo los estimadores de razón combinado y separado comparten el denominador por lo que las fórmulas (3.14) y (3.15) son usadas en los cálculos que mas adelante veremos.

La fase de aplicación de estas fórmulas en este trabajo consiste en lo siguiente. Para la población a considerar:

1. Cálculo de razones y diferencias de razones poblacionales.

2. Obtención de las varianzas por Linealización para los estimadores de razón combinado con (3.3) y con (3.6) para el separado.

3. Cálculo de varianza de los estimadores de la diferencia, considerando el término de covarianza. Usando (3.11) para el estimador combinado y (3.12) para el separado.

En la fase de simulación, en cada muestra se calcula lo siguiente:

1. Los estimadores de razón usando (3.2) para el estimador combinado y (3.5) para el estimador separado.

2. El estimador de la diferencia de razones con (3.8).

3. Los estimadores de varianza de los estimadores, usando (3.4) para el combinado y (3.7) para el separado.

4. Los estimadores de varianza de las diferencias usando las relaciones (3.14) y (3.15).

El procedimiento anterior lo podemos resumir en los siguientes esquemas.
Se calcula $V_L(\widehat{\mathcal{R}})$ y $V_L(\widehat{\mathcal{B}})$.

$$\begin{cases} V_L(\widehat{\mathcal{R}}) & \text{Varianza del estimador combinado para la proporción mediante Linealización.} \\ V_L(\widehat{\mathcal{B}}) & \text{Varianza del estimador separado para la proporción mediante Linealización.} \end{cases}$$

De igual manera se calculan $V_L(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j)$ y $V_L(\widehat{\mathcal{B}}_i - \widehat{\mathcal{B}}_j)$.

$$\begin{cases} V_L(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j) & \text{Varianza para la diferencia de estimadores combinados mediante Linealización.} \\ V_L(\widehat{\mathcal{B}}_i - \widehat{\mathcal{B}}_j) & \text{Varianza para la diferencia de estimadores separados mediante Linealización.} \end{cases}$$

En cada muestra se calculan $\widehat{\mathcal{R}}$ y $\widehat{\mathcal{B}}$.

$$\begin{cases} \widehat{\mathcal{R}} & \text{estimador combinado para la proporción} \\ \widehat{\mathcal{B}} & \text{estimador separado para la proporción} \end{cases}$$

Se calculan $\widehat{V}_L(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j)$ y $\widehat{V}_L(\widehat{\mathcal{B}}_i - \widehat{\mathcal{B}}_j)$.

$$\begin{cases} \widehat{V}_L(\widehat{\mathcal{R}}_i - \widehat{\mathcal{R}}_j) & \text{Estimador de varianza para la diferencia de combinados mediante Linealización.} \\ \widehat{V}_L(\widehat{\mathcal{B}}_i - \widehat{\mathcal{B}}_j) & \text{Estimador de varianza para la diferencia de separados mediante Linealización.} \end{cases}$$

Con la finalidad de simplificar la notación, las diferencias $R_i - R_j$ las denotaremos con $\mathcal{D}$.

En este capítulo se presentó el método analítico de Linealización que sirve como punto de referencia para comparar las técnicas de remuestreo que presentaremos a continuación.

Capítulo 4

Métodos de remuestreo

4.1. El estimador *Jackknife*

El uso ordinario de la palabra *Jackknife* describe una navaja de bolsillo con una gran variedad de herramientas disponibles, como el desarmador y las tijeras. El valor de estos dispositivos es su facilidad para cargarlos y su capacidad de ser usados para una gran variedad de tareas. Con esta analogía, la palabra *Jackknife* fue propuesta por Tukey (1958) para uso en estadística, para describir una aproximación general para pruebas de hipótesis y para calcular intervalos de confianza en situaciones donde no hay métodos analíticos disponibles. Esta técnica partió de un estimador del coeficiente de correlación serial, con corrección del sesgo el cual fue creado por Quenouille (1949).
Este método es computacionalmente intensivo aún para los estándares de la capacidad actual de las computadoras.
Una forma de justificar la idea del método *Jackknife* es pensar en términos de lo que usualmente se hace cuando estimamos un valor promedio, pero desde un inusual punto de vista.

Ejemplo, *Jackknife*

Supóngase que en una muestra aleatoria simple de n valores $y_1, y_2, \ldots, y_n$, calculamos la media muestral.

$$\bar{y}_s = \frac{1}{n} \sum_{y_i \in s} y_i.$$

Esta es usada para estimar la media de la población.

A continuación supóngase que la media muestral es calculada con la j-ésima observación omitida, es decir.

$$\bar{y}_{(j)} = \frac{1}{n-1} \left(\sum_{y_i \in s} y_i - y_j \right).$$

Las dos últimas ecuaciones pueden ser resueltas para y_j resultando.

$$y_j = n\bar{y}_s - (n-1)\bar{y}_{(j)}. \tag{4.1}$$

Esto muestra que es posible determinar el valor muestral y_j de la media muestral completa y también es posible determinar la media muestral con el valor y_j eliminado. Además podemos estimar la media promediando los valores.

$$\bar{y}_s = \frac{1}{n} \sum_{y_i \in s} y_i = \frac{1}{n} \sum_{y_i \in s} \left[n\bar{y}_s - (n-1)\bar{y}_{(i)} \right] = \widehat{\bar{y}}_J. \tag{4.2}$$

Se sigue de las ecuaciones (4.1) y (4.2) lo siguiente:

$$\widehat{V}_J \left(\widehat{\bar{y}}_J \right) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\left[n\bar{y}_s - (n-1)\bar{y}_{(i)} \right] - \widehat{\bar{y}}_J \right)^2 =$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(y_i - \bar{y}_s \right)^2 = \frac{S_{ys}^2}{n}.$$

Como se observa en este ejemplo , el método *Jackknife* y la fórmula matemática proporcionan una estimación idéntica de la varianza muestral.

El método es potencialmente útil en situaciones donde el parámetro que está siendo estimado es una función analítica de la muestra diferente a una media muestral. Inclusive se puede usar en estadísticas que no sean lineales. En una situación general, supóngase que un parámetro θ es estimado para alguna función de los n valores muestrales. Dicha estimación puede ser denotada por $\widehat{\theta}\,(y_1, y_2, \ldots, y_n)$, escrita brevemente $\widehat{\theta}$ y la estimación con el valor y_j removido (estimación parcial) será $\widehat{\theta}_{(j)}$.
Por analogía a la ecuación (4.1) existen entonces el conjunto de pseudovalores para $j = 1, \ldots, n$.

$$\widehat{\theta}_j = n\widehat{\theta} - (n-1)\,\widehat{\theta}_{(j)}. \tag{4.3}$$

Estos pseudovalores juegan el mismo papel que los valores y_j en una estimación de la media, y el promedio de los pseudovalores, es lo que se conoce como estimador *Jackknife* de θ. También es conocido como el estimador de Quenouille.

$$\widehat{\theta}_J = \sum_{j=1}^{n} \frac{\widehat{\theta}_j}{n}.$$

Tratando a los pseudovalores como una muestra aleatoria simple de estimaciones independientes entonces esto sugiere que la varianza de las estimaciones *Jackknife* puede ser estimada por $\frac{S^2}{n}$, donde S^2 es la varianza muestral de los pseudovalores. Definimos.

$$\widehat{V}_J\left(\widehat{\theta}_J\right) = \frac{S^2}{n} = \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\widehat{\theta}_j - \widehat{\theta}_J\right)^2.$$

el estimador de varianza del estimador de Quenouille.
El método *Jackknife* es importante no solo porque genera una estimación de varianza para el estimador de Quenouille, sino porque se puede generar una estimación de varianza para cualquier estimador. La fórmula para dicha estimación es análoga a la de la estimación de varianza del estimador de

Quenouille simplemente haciendo la sustitución de $\widehat{\theta}_J$ por $\widehat{\theta}$,

$$\widehat{V}_J\left(\widehat{\theta}\right) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \left(\widehat{\theta}_j - \widehat{\theta}\right)^2.$$

Es el estimador *Jackknife* de varianza del estimador $\widehat{\theta}$.
Särndal *et al.* (1991,p.438) hacen la observación de que en el caso general

$$\widehat{V}_J\left(\widehat{\theta}_J\right) \leq \widehat{V}_J\left(\widehat{\theta}\right).$$

Para el estimador de Quenouille un intervalo aproximado de $100(1-\alpha)\,\%$ de confianza para θ está dado por.

$$\widehat{\theta}_J \pm z_{(1-\frac{\alpha}{2})} \sqrt{\widehat{V}_J\left(\widehat{\theta}_J\right)}.$$

Y para cualquier otro estimador está dado por.

$$\widehat{\theta} \pm z_{(1-\frac{\alpha}{2})} \sqrt{\widehat{V}_J\left(\widehat{\theta}\right)}.$$

Donde $z_{(1-\frac{\alpha}{2})}$ es el cuantil de una Normal estándar.
Con todo este tratamiento, un problema de estimación se puede reducir a un problema de estimación de una media muestral.

El método tiene como ventajas principales:

1. Es fácil de implementar en cualquier diseño muestral
2. Es un método muy general que permite estimar cualquier función continua de los datos
3. La fórmula para el cálculo de varianza es prácticamente la misma sin importar el diseño muestral
4. Existen varios paquetes estadísticos que calculan varianzas por este método, por ejemplo:

- *AM Software.*
- *Generalized Estimation System.*
- *IVEware.*
- *R survey package.*
- *Stata.*
- *SUDAAN.*
- *VPLX.*
- *WesVar.*

Desventajas:

1. En general los pseudovalores están correlacionados en algún grado, así que la varianza estimada del estimador *Jackknife* está sesgada hacia arriba o hacia abajo. Es difícil predecirlo teóricamente. Por lo tanto, la varianza *Jackknife* necesita ser justificada por estudios empíricos o analíticos antes de ser usada.

2. El estimador del parámetro de interés debe ser una función explícita de los n valores muestrales.

3. Si el estimador no es una función continua el estimador de varianza *Jackknife*, es inconsistente.

Por ejemplo, los cuantiles son funciones que no son continuas. Efron(1982) encontró que en este caso, el estimador de varianza *Jackknife* es inconsistente. Tomando la mediana muestral él demostró que,

$$\frac{\widehat{V}_J}{V_n} \xrightarrow{d} \left(\frac{\chi_2^2}{2}\right)^2$$

donde V_n es la varianza asintótica de la mediana muestral y χ_2^2 es una variable aleatoria ji-cuadrada con 2 grados de libertad.

Antes de que el *Jackknife* fuera sugerido como una herramienta general para inferencia por Tukey(1958), Quenouille(1956) había ya tenido la idea de reemplazar un estimador por su respectiva versión *Jackknife* teniendo en mente reducir el sesgo en un orden de $\frac{1}{n}$.
Para esto, suponga que el valor esperado (la media) del estimador $\widehat{\theta}$ de θ basado en todo el conjunto de n observaciones es $\theta\left(1+\frac{A}{n}\right)$ con el factor A de sesgo . Entonces el valor esperado del estimador parcial $\widehat{\theta}_{(j)}$ con la j-ésima observación eliminada es $\theta\left\{1+\frac{A}{n-1}\right\}$. Se sigue que el valor esperado del j-ésimo pseudovalor definido en la ecuación (4.3) es

$$E\left[\widehat{\theta}_j\right] = n\left[\theta\left(1+\frac{A}{n}\right)\right] - (n-1)\left[\theta\left\{1+\frac{A}{n-1}\right\}\right] = \theta$$

con el factor A de sesgo cancelado.
El trabajo de Tukey fue muy importante en la década de los sesentas.
Más recientemente una revisión y nuevos descubrimientos fueron hechos por Gray y Schucany (1972) y Miller (1974). Una bibliografía de 162 referencias fue producida por Parr y Schucany (1980), ésta fue actualizada por Frangos en (1987). Una revisión más reciente fue proporcionada por Hinkley (1983), y en muestreo complejo podemos referirnos a Shao y Tu (1995).
Otras generalizaciones del método son las siguientes:

1. *Jackknife* en muestreo multietápico , el cual se aplica solamente a nivel de unidad primaria.

2. *Jackknife* que remueve más de una unidad muestral por replicación, conocido como d-*Jackknife*.

3. *Jackknife* en diseños muestrales estratificados. En donde se aplica en cada uno de los estratos.

En este trabajo sólo explicaremos y aplicaremos la última generalización. La cual detallamos a continuación.

4.1.1. Estimador *Jackknife* en muestreo estratificado

Las primeras referencias señalan que el *Jackknife* en diseños estratificados fue planteado por Wolter (1985, p.176). Más recientemente Shao y Tu (1995) analizaron otras formas de la metodología *Jackknife* en el caso de tener una muestra estratificada. Consideraron muestreo con y sin reemplazo,y además dieron a conocer algunos resultados teóricos.
A continuación se describe el procedimiento de aplicación del *Jackknife* en una muestra estratificada aleatoria simple sin reemplazo, considerando además los estimadores de razón descritos en capítulos anteriores.
Se utiliza la metodología *Jackknife* descrita por Rao y Rust (1996) que es diferente a la que usó Wolter en 1985. El *Jackknife* de Rao y Rust elimina una unidad primaria, ajusta los pesos muestrales (w_h) en cada una de las iteraciones y no hace uso de pseudovalores. El de Wolter elimina una unidad primaria sin ajustar los pesos muestrales y usa pseudovalores.
En este trabajo se aplica el procedimiento de Rao y Rust (1996) porque el procedimiento produce buenas estimaciones de varianza.

4.1.2. El método de estimación *Jackknife* de Rao y Rust (1996)

Supóngase que se tiene una muestra que tiene H estratos y n_h unidades en cada estrato. Además no especificaremos el tipo de estimador de razón (separado o combinado) en vista de que el proceder es el mismo para ambos. De manera general hay que seguir los siguientes pasos para calcular una estimación *Jackknife*.

1. Para el parámetro poblacional de interés R, se calcula un estimador $\widehat{R}$ utilizando todos los datos muestrales y respetando el diseño muestral, en este caso estratificado aleatorio simple.

2. Para cada estrato, $1 \leq h \leq H$, $1 \leq i \leq n_h$, calculamos $\widehat{R}_{(-hi)}$ un estimador de R con la misma forma funcional que tiene $\widehat{R}$ pero eliminando la observación i del estrato h, tomando en cuenta nuevos pesos muestrales, descritos de la siguiente forma.
   Para $1 \leq h \leq H$, $1 \leq i \leq n_h$:

$$w_{hi} = \begin{cases} \frac{N_h}{n_h} & \text{Para unidades fuera del estrato donde se elimina la observación } i. \\ \frac{N_h}{n_h - 1} & \text{Para unidades en el estrato } h. \\ 0 & \text{Para la unidad que se elimina.} \end{cases}$$

   Obtenemos así las estimaciones.

$$\widehat{R}_{(-11)}, \ldots, \widehat{R}_{(-1n_1)}, \ldots, \widehat{R}_{(-H1)}, \ldots, \widehat{R}_{(-Hn_H)}.$$

3. Se construye el estimador de varianza de Rao y Rust para el estimador de razón en muestreo estratificado.

$$\widehat{V}_J\left(\widehat{R}\right) = \sum_{h=1}^{H} \frac{n_h - 1}{n_h} \sum_{i=1}^{n_h} \left(\widehat{R}_{(-hi)} - \widehat{R}\right)^2. \tag{4.4}$$

Podemos basar nuestros intervalos de confianza para R en la estadística $\widehat{R}$. Es decir, un intervalo para R con una confianza de $(1-\alpha)100\,\%$, estaría dado por:

$$\widehat{R} \pm t_{(1-\frac{\alpha}{2}, n-H)} \sqrt{\widehat{V}_J\left(\widehat{R}\right)}.$$

donde $t_{(1-\frac{\alpha}{2}, n-H)}$ es el cuantil de la distribución t con n-H grados de libertad. Rao y Rust (1996) hacen ver que los grados de libertad dependen del número de iteraciones que se hacen (tantas como unidades primarias de muestreo), por lo que, en la mayoría de los casos, los intervalos deben basarse en la distribución t de Student y no en la Normal. Aunque en este trabajo ese caso

no se presenta, debido a que en este trabajo la diferencia $n-H$ vale al menos 300 en el peor de los casos.

Existen expresiones alternativas para calcular la varianza *Jackknife*, presentamos algunas de ellas a continuación:

1. La expresión que brinda Wolter (1985).

$$\widehat{V}_J\left(\widehat{R}\right) = \sum_{h=1}^{H} \frac{N_h^2}{n_h(n_h-1)} \sum_{i=1}^{n_h} \left(\widehat{R}_{(-hi)} - \widehat{R}\right)^2.$$

2. La expresión de Canty y Davison (1998) que incorpora un factor de corrección finita, ver Palmer, Eslava,*et al.* (2001, p.45).

$$\widehat{V}_J\left(\widehat{R}\right) = \sum_{h=1}^{H} \frac{n_h-1}{n_h}\left(1-\frac{n_h}{N_h}\right) \sum_{i=1}^{n_h} \left(\widehat{R}_{(-hi)} - \widehat{R}\right)^2.$$

Ninguna de estas dos expresiones se calcularon en este trabajo.

4.1.3. Estimadores calculados con el método *Jackknife*

La fase de aplicación consiste en la obtención de los estimadores de varianza para ambos estimadores de razón combinado, separado y sus diferencias, mediante *Jackknife* considerando tres tamaños de muestra diferentes.
Lo podemos resumir en los siguientes esquemas.
Se calculan los estimadores $\widehat{V}_J(\widehat{\mathcal{R}})$ y $\widehat{V}_J(\widehat{\mathcal{B}})$ con la expresión (4.4).

$$\begin{cases} \widehat{V}_J(\widehat{\mathcal{R}}) & \text{Estimador de varianza } \textit{Jackknife} \\ & \text{del estimador combinado para la proporción.} \\ \widehat{V}_J(\widehat{\mathcal{B}}) & \text{Estimador de varianza } \textit{Jackknife} \\ & \text{del estimador separado para la proporción} \end{cases}$$

De igual manera para las diferencias D, $\widehat{V}(\widehat{\mathcal{D}})$ es aproximado con el cálculo de $\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$ y $\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$ usando (4.4).

$$\begin{cases} \widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}}) & \text{Varianza estimada } \textit{Jackknife} \text{ para la diferencia} \\ & \text{de estimadores combinados.} \\ \widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}}) & \text{Varianza estimada } \textit{Jackknife} \text{ para la diferencia} \\ & \text{de estimadores separados.} \end{cases}$$

4.2. El método *Bootstrap*

Un conjunto de datos de tamaño n tiene $2^n - 1$ subconjuntos no vacíos; el método *Jackknife* solo utiliza n de estos.
El *Jackknife* puede ser mejorado usando una estadística basada en más de n subconjuntos, inclusive usar los $2^n - 1$ subconjuntos. Esta idea fue discutida por Hartigan (1969), pero para su época las computadoras no tenían la capacidad de trabajar con tantos elementos. El *Bootstrap* es un método computacionalmente intenso por lo que su desarrollo e implementación comenzó apenas hace 2 décadas.
El método *Bootstrap* simple fue ideado por Efron (1979), pero este no captura la dependencia impuesta por el muestreo sin reemplazo. En consecuencia muchas aproximaciones Bootstrap fueron ideadas para poblaciones finitas. Una de las primeras se le debe a Gross (1980) que propone el *Bootstrap* corto sin reemplazo (BWO) en el caso de muestreo aleatorio simple. Más adelante veremos que para diseños estratificados la teoría al respecto ha crecido bastante, y para diseños complejos y análisis estadísticos existen algunos trabajos como los siguientes:

- La idea de remuestreo *Bootstrap* de datos fue aplicado para calcular la distribución final en análisis Bayesiano. Llamado inicialmente *Bootstrap* Bayesiano, elaborado por Rubin (1981).

- Kuk (1989) propuso un método *Bootstrap* para un muestreo sistemático con probabilidad proporcional al tamaño y sin reemplazo.

- Sitter (1992) adaptó el método Rao-Hartley-Cochran (RHC) para muestreo con probabilidad proporcional al tamaño con reemplazo.

Las propiedades de los métodos han sido estudiadas para el caso en que tenemos una estimación puntual de la varianza para varias combinaciones de diseños y estimadores puntuales.

Para explicar la forma de proceder del método expondremos un ejemplo muy sencillo. Supongamos que realizamos 3 medidas de una determinada característica de un objeto o fenómeno con el objetivo de hallar un intervalo de confianza al 95 % para el valor poblacional de la característica. Si las medidas realizadas son 2, 3 y 5, los métodos de la estadística tradicional basados en hipótesis de normalidad asintótica para la población de la cual proceden los datos, establecen que el intervalo de confianza al 95 % para el verdadero valor de la característica es $\bar{y}_s \pm t_{(0,975,2)} \frac{S_{ys}}{\sqrt{3}}$ donde $\bar{y}_s$ es la media de estos tres datos, S_{ys} es la desviación estándar y $t_{(0,975,2)}$ es el percentil 97.5 de una distribución t de Student con 2 grados de libertad. Para este caso concreto el intervalo de confianza al 95 % es $[0.758, 5.909]$. La metodología *Bootstrap* procede de la siguiente manera para hallar el intervalo al 95 %. En primer lugar no hace ningún tipo de suposición acerca de la distribución de la cual proceden los datos. Lo que hace es obtener nuevas muestras de tamaño 3 con reemplazo de la muestra original, llamadas muestras *Bootstrap*, y calcula la media para cada una de estas nuevas muestras. Una de estas nuevas muestras pudiera ser 2, 3 y 3; la media para esta muestra es 8/3. El objetivo es obtener la función de distribución para la media de una supuesta población formada por los elementos 2, 3 y 5. Para hallar dicha distribución podemos proceder de dos maneras, una aproximada y otra exacta.

1. **Aproximada**. Para hallarla de esta manera podríamos lanzar un dado, por ejemplo 18 veces, con el siguiente convenio: Si sale 1 ó 2 elegimos el primer valor de la muestra original (es decir, el 2), si sale un 3 ó 4 elegimos el segundo (es decir, el 3) y si sale un 5 ó 6 elegimos el tercero (es decir, el 5). Después de los tres primeros lanzamientos del dado habremos obtenido una nueva muestra de tamaño 3 a la cual le calcularemos la media. Seguiremos así hasta obtener 6 muestras posibles de tamaño 3, las cuales nos producirán 6 valores para

la media. Para terminar realizaremos con estos 6 valores de la media una tabla de frecuencias relativas y frecuencias acumuladas. En realidad lo que estamos haciendo de esta manera es obtener la función de distribución de la media por simulación *Monte Carlo*. Desde luego en la era de las computadoras no vamos hacerlo como en la época de "Student"(lanzando dados) sino que se realizará mediante una computadora. Si repetimos el procedimiento hasta obtener 1000 valores de la media, se tiene que el intervalo de confianza *Bootstrap* al 90 % para la característica es $[7/3, 13/3] = [2.333, 4.333]$. Este método (aproximado) presenta la desventaja que añade, en la mayoría de los casos, una incertidumbre debida a la simulación que el método exacto no hace. Sin embargo en la mayoría de los casos es lo único con que el investigador cuenta para aproximar la varianza.

2. **Exacta**: Esta forma de proceder es conocida como *Bootstrap* total o ideal. Para el ejemplo que nos ocupa procederíamos de la siguiente manera. En primer lugar se obtendrían todas las muestras *Bootstrap* que dan valores diferentes de la media. Posteriormente se calculan las probabilidades para cada una de estas muestras diferentes y problema resuelto. En el caso de una muestra original de 3 elementos y_1, y_2 y y_3 como la que tenemos, las diferentes muestras *Bootstrap* son.
$\{y_1, y_1, y_1\}$, $\{y_1, y_1, y_2\}$, $\{y_1, y_1, y_3\}$, $\{y_1, y_2, y_2\}$, $\{y_1, y_2, y_3\}$,
$\{y_1, y_3, y_3\}$, $\{y_2, y_2, y_2\}$, $\{y_2, y_2, y_3\}$, $\{y_2, y_3, y_3\}$ y $\{y_3, y_3, y_3\}$.
Obsérvese que los subíndices en la muestras anteriores parten del 111 hasta 333 en orden creciente, es decir con el subíndice siguiente siempre mayor o igual que el anterior. La probabilidad de la muestra $\{y_2, y_2, y_3\}$ (probabilidad multinomial) es $\frac{3!}{(0!2!1!)}\frac{1}{3^3} = \frac{1}{9}$; la de la muestra $\{y_2, y_2, y_2\}$ es $\frac{3!}{(0!3!0!)}\frac{1}{3^3} = \frac{1}{27}$ y así sucesivamente. La función de distribución para la media se presenta en la siguiente tabla. A partir de la tabla se deduce que el intervalo de confianza exacto *Bootstrap* al 90 % para el verdadero

valor de la característica es también $[7/3, 13/3] = [2.333, 4.333]$ igual al aproximado.

$\bar{y}_s$	2	7/3	8/3	3	10/3	11/3	4	13/3	5
$F(\bar{y}_s)$	0.04	0.11	0.26	0.41	0.63	0.74	0.85	0.96	1

Tabla 4.1: Función de distribución exacta de la media

4.2.1. *Bootstrap* en muestreo estratificado

En diseños estratificados podemos encontrar varias modificaciones al método original de Efron (1979). A continuación mencionamos algunas de estas variaciones.

1. El *Bootstrap* ingenuo o *Naive Bootstrap* llamado así por Efron (1982).

2. Bickel y Freedman (1984) discuten el BWO (*Bootstrap sin reemplazo*) de Gross que mencionamos en la sección anterior, y lo extienden a una versión en un diseño estratificado aleatorio simple sin reemplazo.

3. El *Bootstrap* con reemplazo BWR (*Bootstrap-With-Replacement*) fue hecho en base a una modificación del *Bootstrap* ingenuo.
   McCarthy y Snowden (1985) mejoran el BWO-en diseños estratificados con selección aleatoria simple.

4. Rao y Wu (1988) proponen el *Bootstrap* con reescalamiento BRS (*Rescaling-Bootstrap*).

5. Sitter (1992) propuso *The Mirror-Match-Bootstrap* BMM y también hizo una extensión del BWO-en diseños estratificados con selección aleatoria simple.

En este trabajo se aplicó el *Bootstrap* con rescalamiento de Rao y Wu (1988), que es una variante del método *Bootstrap* tradicional, pues *Bootstrap* independientes se hacen en cada estrato, además esta variante utiliza muestreo con reemplazo.
Este *Bootstrap* se escoge para hacer este trabajo porque es fácil de programar, emplea poco tiempo para ejecutarse en la computadora, la velocidad en que converge es la misma que tiene el *Jackknife* y porque es una modificación del *Bootstrap* ingenuo que corrige la inconsistencia del estimador de varianza que éste presenta.

A continuación describiremos en qué consisten el método *Bootstrap* ingenuo y *Bootstrap* con rescalamiento.
Consideremos una población finita de N unidades, particionada en H estratos con $N_1, \ldots, N_H$ unidades respectivamente. Al seleccionar una muestra, un muestreo aleatorio simple sin reemplazo es utilizado independientemente en cada uno de los estratos. Los tamaños de muestra al interior de cada uno de los estratos son $n_1, \ldots, n_H$ respectivamente y el tamaño de muestra total es $n = \sum_{h=1}^{H} n_h$. Un vector de características para cada unidad se representa con $\mathbf{x}_{hi} = (x_{1hi}, \ldots, x_{\tau hi})^T$, donde $h = 1, \ldots, H$; $i = 1, \ldots, n_h$.
El vector de medias poblacionales es denotado con $\bar{\mathbf{x}}_U = (\bar{\mathbf{x}}_{1U}, \ldots, \bar{\mathbf{x}}_{\tau U})^T$, el vector de medias muestrales con $\bar{\mathbf{x}}_s = (\bar{\mathbf{x}}_{1s}, \ldots, \bar{\mathbf{x}}_{\tau s})^T = \sum_{h=1}^{H} W_h \bar{\mathbf{x}}_h$, donde $\bar{\mathbf{x}}_h = \frac{1}{n_h} \sum_{\mathbf{x}_{hi} \in s_h} \mathbf{x}_{hi} = (\bar{x}_{1s_h}, \ldots, \bar{x}_{\tau s_h})^T$ y $W_h = \frac{N_h}{N}$.

4.2.2. *Bootstrap* ingenuo

Muy comúnmente en estadística se consideran estimadores de la forma

$$\widehat{\theta} = g(\bar{\mathbf{x}}_s),$$

que es una función de promedios. Para estimar la varianza de dicho estimador, el *Bootstrap* ingenuo hace lo siguiente.

1. Se obtiene una muestra aleatoria simple con reemplazo que denominaremos $s_h^\star$ en cada uno de los estratos, con tamaño n_h y que denotaremos con $\{\mathbf{x}_{hi}^\star\}_{i=1}^{n_h}$ a partir de la muestra observada s_h en cada estrato, que denotaremos $\{\mathbf{x}_{hi}\}_{i=1}^{n_h}$.
   A continuación se calcula lo siguiente.

$$\bar{\mathbf{x}}_h^\star = \frac{1}{n_h} \sum_{\mathbf{x}_{hi}^\star \in s_h^\star} \mathbf{x}_{hi}^\star = \left(\bar{\mathbf{x}}_{1s_h^\star}^\star, \ldots, \bar{\mathbf{x}}_{\tau s_h^\star}^\star\right)^T,$$

   en donde para $t = 1, \ldots, \tau$ tenemos que,

$$\bar{\mathbf{x}}_{ts_h^\star}^\star = \frac{1}{n_h} \sum_{x_{thi}^\star \in s_h^\star} x_{thi}^\star.$$

   Después se calcula,

$$\bar{\mathbf{x}}_s^\star = \sum_{h=1}^{H} W_h \bar{\mathbf{x}}_h^\star$$

   con lo que se evalúa,

$$\widehat{\theta}^\star = g(\bar{\mathbf{x}}_s^\star). \tag{4.5}$$

2. Repetir el paso uno un número grande de veces, B. Para obtener,

$$\widehat{\theta}_1^\star, \ldots, \widehat{\theta}_B^\star.$$

   las replicaciones calculadas a partir de (4.5). A continuación se toma como *el estimador Bootstrap* de θ al promedio de estas replicaciones,

$$\widehat{\theta}(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\theta}_b^\star.$$

$$\widehat{V}_\star(\widehat{\theta}(\cdot)) = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\theta}_b^\star - \widehat{\theta}(\cdot))^2$$

   es tomado como el estimador de varianza para el estimador $\widehat{\theta}(\cdot)$.

Si escribimos $E_\star$ y $V_\star$ los operadores de esperanza y varianza con respecto al muestreo *Bootstrap* y continuamos este muestreo indefinidamente entonces $\widehat{\theta}(\cdot) \approx E_\star(\widehat{\theta}^\star)$ y $\widehat{V}_\star(\widehat{\theta}(\cdot)) \approx V_\star(\widehat{\theta}^\star)$. Sin embargo este procedimiento no siempre es consistente.

Ejemplo:

Considere el caso de que $\tau = 1$ y $\widehat{\theta} = \bar{x}_{1s}$ entonces,

$$E\left[\widehat{V}_\star(\bar{x}^\star_{1s})\right] \neq E\left[\widehat{V}(\bar{x}_{1s})\right]$$

A menos que n_h sea grande en cada estrato.
Esto es debido a que por una parte,

$$E\left[\widehat{V}(\bar{x}_{1s})\right] = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S^2_{xU_h}.$$

Por la otra tenemos,

$$E\left[\widehat{V}_\star(\bar{x}^\star_{1s})\right] = E\left[E\left(\widehat{V}_\star(\bar{x}^\star_{1s}) \mid s\right)\right] = \sum_{h=1}^{H} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{n_h - 1}{n_h} S^2_{xU_h}.$$

El factor de escala $\frac{n_h - 1}{n_h}$ no tiene un efecto apreciable cuando el tamaño de muestra por estrato es grande.
El caso contrario es tener por ejemplo $n_h = 2$ unidades por estrato, situación que se presenta más adelante en este trabajo, $\frac{n_h - 1}{n_h} = \frac{1}{2}$ en cada uno de los estratos y por lo tanto existe un sesgo relativo[1] de $\widehat{V}_\star$ de al menos 50 %.
Efron (1982) dice que es un problema de escalamiento para este procedimiento *Bootstrap*. Sugiere que una manera de remediar el problema es tomar $n_h - 1$ unidades en cada muestra *Bootstrap* en lugar de n_h.

[1]Mas adelante se aclarará el significado de este término

4.2.3. *Bootstrap* con rescalamiento (RS)

Para estimar la varianza de $\widehat{\theta} = g(\bar{\mathbf{x}}_s)$, Rao y Wu (1988) proponen el siguiente procedimiento.

Un vector es remuestreado con reemplazo, de la muestra original, después cada vector es rescalado y el estimador original es aplicado al vector rescalado. Los factores de rescalamiento son escogidos de tal manera que la varianza bajo las remuestras sea la misma que la producida por el estimador de varianza en el caso lineal. En un muestreo estratificado aleatorio simple, el método de rescalamiento es como sigue.

1. Seleccionar una muestra aleatoria simple con reemplazo s_h^* en cada estrato, de tamaño m_h con $1 \leq m_h \leq n_h$, que denotaremos $\{\mathbf{x}_{hi}^*\}_{i=1}^{m_h}$ a partir de la muestra observada en cada estrato s_h que denominaremos $\{\mathbf{x}_{hi}\}_{i=1}^{n_h}$.

   Luego se construyen los factores de rescalamiento, Rao y Wu (1988) utilizan

   $$c_h = \sqrt{\frac{m_h(1-\frac{n_h}{N_h})}{n_h - 1}}. \tag{4.6}$$

   Los factores de rescalamiento (4.6) se sustituyen en la siguiente ecuación.

   $$\tilde{\mathbf{x}}_h^* = \bar{\mathbf{x}}_h + c_h(\bar{\mathbf{x}}_h^* - \bar{\mathbf{x}}_h).$$

   Estos cambios se hacen solo con la finalidad de reducir el sesgo del estimador de varianza.

   $$\bar{\mathbf{x}}_h^* = \frac{1}{m_h} \sum_{\mathbf{x}_{hi}^* \in s_h^*} \mathbf{x}_{hi}^* = \left(\bar{\mathbf{x}}_{1s_h^*}^*, \ldots, \bar{\mathbf{x}}_{\tau s_h^*}^*\right)^T,$$

   y para $t = 1, \ldots, \tau$ tenemos que,

   $$\bar{\mathbf{x}}_{ts_h^*}^* = \frac{1}{m_h} \sum_{x_{thi}^* \in s_h^*} x_{thi}^*.$$

A continuación, se calcula

$$\tilde{\mathbf{x}}^* = \sum_{h=1}^{H} W_h \tilde{\mathbf{x}}_h^*$$

con $W_h = \frac{N_h}{N}$.
Se construye el estimador con misma forma funcional que el estimador original.

$$\widehat{\theta}_b^* = g(\tilde{\mathbf{x}}^*).$$

2. Se repite el paso uno un gran número de veces, digamos B y con esto se obtienen las replicaciones,

$$\widehat{\theta}_1^*, \ldots, \widehat{\theta}_B^*$$

Se toma como el *estimador Bootstrap* de θ al promedio de estos,

$$\widehat{\theta}_{RS} = \frac{\sum_{b=1}^{B} \widehat{\theta}_b^*}{B} \tag{4.7}$$

y la varianza se estima mediante la aproximación de *Monte Carlo*.

$$\widehat{V}(\widehat{\theta}^*) \approx E_* \left(\widehat{\theta}^* - E_* \widehat{\theta}^* \right)^2,$$

que se aproxima con,

$$\widehat{V}_{RS} \left(\widehat{\theta}_{RS} \right) = \frac{1}{B-1} \sum_{b=1}^{B} (\widehat{\theta}_b^* - \widehat{\theta}_{RS})^2. \tag{4.8}$$

Para calcular los intervalos de confianza de $\widehat{\mathcal{R}}$ y $\widehat{\mathcal{B}}$ se usó el método de intervalo de confianza estándar *Bootstrap*, que más adelante describiremos.

El cambio propuesto por Rao y Wu (1988) al *Bootstrap* ingenuo, se justifica porque de manera heurística demostraron que en el caso de que $\widehat{\theta} = g(\bar{\mathbf{x}}_s)$ con una función g no lineal, la distribución de los estimadores del *Bootstrap* con rescalamiento es consistente.

Este procedimiento elimina el problema de escalamiento del *Bootstrap* ingenuo que mencionamos en el último ejemplo citado. Es decir, si $\tau = 1$ y $\widehat{\theta} = \bar{x}_{1s}$,

$$\widehat{V}_* \left(\bar{x}_{1s}\right) = \sum_{h=1}^{H} W_h^2 \frac{m_h}{n_h - 1} \left[\frac{1}{m_h} \frac{n_h - 1}{n_h} \left(1 - \frac{n_h}{N_h}\right)\right] S^2_{x1s_h} = \sum_{h=1}^{H} \frac{W_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S^2_{x1s_h}$$

Entonces

$$E\left[\widehat{V}_*(\bar{x}^*_{1s})\right] = E\left[\widehat{V}(\bar{x}_{1s})\right].$$

Como ventaja podemos señalar que la técnica de *Bootstrap* servirá, a diferencia del *Jackknife*, para funciones no suaves, como es el caso de los cuantiles en diseños generales de muestreo.
Da como desventaja principal que requiere más cálculos que el *Jackknife* y que cualquier otra técnica, porque B es una cantidad en general muy grande.

Según Sitter (1992) el *Bootstrap* con rescalamiento 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 para algunas elecciones de los valores de m_h, algún $\widehat{\theta}^*_b$ sea negativo, aún cuando por definición, $\widehat{\theta} \geq 0$.
Actualmente pocos paquetes estadísticos cuentan con algunas rutinas programadas, pero ninguno tiene el *Bootstrap* con rescalamiento.

- *AM Software.*

- *R survey package.*
- *Stata,* aunque no considera diseño muestral.
- *SUDAAN*, aunque necesita de pesos de replicación.
- *WesVar*, también necesita de pesos de replicación.

4.2.4. Aplicación del método de rescalamiento

Antes de tratar de aplicar el método debemos preguntarnos lo siguiente: ¿Los dos tipos de estimadores de razón pueden ser expresados como una función g de medias poblacionales?. Esta pregunta surge debido a que el método de rescalamiento requiere que el estimador sea una función de medias. A continuación veremos que sólo el estimador de razón combinado satisface la condición, pues si pensamos en $\tau = 2$,

$$\widehat{\mathcal{R}} = \frac{\sum_{h=1}^{H} \sum_{y_k \in s_h} w_{hk} y_k}{\sum_{h=1}^{H} \sum_{z_k \in s_h} w_{hk} z_k} = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{y_k \in s_h} y_k}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{z_k \in s_h} z_k} = \frac{\sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{s_h}}{\sum_{h=1}^{H} \frac{N_h}{N} \bar{z}_{s_h}} = \frac{\sum_{h=1}^{H} W_h \bar{y}_{s_h}}{\sum_{h=1}^{H} W_h \bar{z}_{s_h}} =$$

$$g\left(\sum_{h=1}^{H} W_h \bar{\mathbf{x}}_h\right) = g\left(\left(\bar{\mathbf{y}}_s, \bar{\mathbf{z}}_s\right)^T\right) = g\left(\bar{\mathbf{x}}_s\right).$$

Pero el estimador de razón separada no la cumple,

$$\widehat{\mathcal{B}} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\sum_{y_k \in s_h} w_{hk} y_k}{\sum_{z_k \in s_h} w_{hk} z_k} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\sum_{y_k \in s_h} \frac{N_h}{n_h} y_k}{\sum_{z_k \in s_h} \frac{N_h}{n_h} z_k} = \sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\frac{N_h}{N} \bar{y}_{s_h}}{\frac{N_h}{N} \bar{z}_{s_h}} =$$

$$\sum_{h=1}^{H} \frac{t_{zU_h}}{t_z} \frac{\bar{y}_{s_h}}{\bar{z}_{s_h}} = g\left(\bar{y}_{s_1}, \ldots, \bar{y}_{s_H}, \bar{z}_{s_1}, \ldots, \bar{z}_{s_H}\right) \neq g\left(\bar{\mathbf{x}}_{\mathbf{s}}\right).$$

En este trabajo se pretende apreciar si la falta de esta hipótesis afecta de alguna manera el resultado de las estimaciones.

Para la aplicación debemos seguir los siguientes pasos.

1. Obtener una muestra aleatoria simple con reemplazo
   $s_h^* = \{\mathbf{x}_{hi}^*\}_{i=1}^{m_h} = \{(y_{hi}^*, z_{hi}^*)^T\}_{i=1}^{m_h}$ en cada estrato, de tamaño m_h, a partir de la muestra observada $s_h = \{\mathbf{x}_{hi}\}_{i=1}^{n_h} = \{(y_{hi}, z_{hi})^T\}_{i=1}^{n_h}$. Se calculan

$$\tilde{\mathbf{x}}_h^* = \bar{\mathbf{x}}_h + \sqrt{\frac{m_h(1-\frac{n_h}{N_h})}{n_h-1}}\,(\bar{\mathbf{x}}_h^* - \bar{\mathbf{x}}_h) = (\tilde{y}_h^*, \tilde{z}_h^*)^T =$$

$$\left(\bar{y}_{s_h} + \sqrt{\frac{m_h(1-\frac{n_h}{N_h})}{n_h-1}}\left(\bar{y}_{s_h}^* - \bar{y}_{s_h}\right), \bar{z}_{s_h} + \sqrt{\frac{m_h(1-\frac{n_h}{N_h})}{n_h-1}}\left(\bar{z}_{s_h}^* - \bar{z}_{s_h}\right)\right)^T.$$

   y

$$\tilde{\mathbf{x}}^* = \sum_{h=1}^{H} W_h \tilde{\mathbf{x}}_h^* = (\tilde{\mathbf{y}}_s^*, \tilde{\mathbf{z}}_s^*)^T.$$

   Con lo que el estimador de razón combinado en una iteración *Bootstrap* es el siguiente.

$$\widehat{\mathcal{R}}^* = \frac{\tilde{\mathbf{y}}_s^*}{\tilde{\mathbf{z}}_s^*}.$$

   El estimador de razón separada no se puede llevar hasta la última sustitución por lo que nos quedamos un paso antes,

$$\widehat{\mathcal{B}}^* = \sum_{h=1}^{H}\left(\frac{t_{z_{U_h}}}{t_z}\right)\frac{\tilde{y}_h^*}{\tilde{z}_h^*}.$$

2. Se repite el paso uno B veces, con esto se obtienen para ambos estimadores.

$$\widehat{\mathcal{R}}_1^*, \widehat{\mathcal{R}}_2^*, \ldots, \widehat{\mathcal{R}}_B^*$$

   y

$$\widehat{\mathcal{B}}_1^*, \widehat{\mathcal{B}}_2^*, \ldots, \widehat{\mathcal{B}}_B^*$$

Se toma como el *estimador Bootstrap* de R al promedio de estos en cada tipo de estimador,

$$\widehat{\mathcal{R}}_{RS} = \frac{\sum_{b=1}^{B} \widehat{\mathcal{R}}_b^*}{B}. \tag{4.9}$$

$$\widehat{\mathcal{B}}_{RS} = \frac{\sum_{b=1}^{B} \widehat{\mathcal{B}}_b^*}{B}. \tag{4.10}$$

Se estima la varianza mediante lo siguiente.

$$\widehat{V}_{RS}\left(\widehat{\mathcal{R}}_{RS}\right) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\mathcal{R}}_b^* - \widehat{\mathcal{R}}_{RS})^2. \tag{4.11}$$

$$\widehat{V}_{RS}\left(\widehat{\mathcal{B}}_{RS}\right) = \frac{1}{B-1}\sum_{b=1}^{B}(\widehat{\mathcal{B}}_b - \widehat{\mathcal{B}}_{RS})^2. \tag{4.12}$$

Y finalmente se utiliza para construir un intervalo de confianza del $(1-\alpha)100\,\%$,

$$\widehat{\mathcal{R}}_{RS} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{V}_{RS}\left(\widehat{\mathcal{R}}_{RS}\right)}$$

y

$$\widehat{\mathcal{B}}_{RS} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{V}_{RS}\left(\widehat{\mathcal{B}}_{RS}\right)}.$$

A continuación presentamos brevemente algunas de las diversas formas en las que se pueden encontrar intervalos de confianza para el estimador *Bootstrap*.

4.2.5. Intervalos de confianza *Bootstrap*

Los intervalos de confianza se basan en la distribución que se genera, llamada distribución *Bootstrap*. Hay varias formas de calcular estos intervalos, podemos mencionar algunas a continuación.

El método del intervalo de confianza estándar, el método del percentil simple de Efron (1979), el método de los percentiles corregidos, el método t de Efron

(1981), DiCiccio y Romano (1988,1990) propusieron el método del percentil automático, Owen (1988) propuso métodos no paramétricos de verosimilitud. Más recientemente el método del percentil de los desvíos de Hall (1992). Chaudhuri y Stenger (1992) encontraron algunas otras formas.

Los cuatros procedimientos más comunes para construir intervalos de confianza, son los siguientes.

1. El método del intervalo de confianza *Bootstrap* estándar.
2. El método del percentil simple de Efron.
3. Método del intervalo de confianza *Bootstrap* por *Bootstrap*-t.
4. Método del intervalo de confianza *Bootstrap* por percentil de los desvíos.

La opción que en este trabajo se utilizó es la del método de rescalamiento y es el intervalo de confianza *Bootstrap* estándar que a continuación describimos.

El método del intervalo de confianza *Bootstrap* estándar

Es el que más se asemeja a la manera de construir intervalos de confianza paramétricos, se supone que el estimador *Bootstrap* $\widehat{\theta}$ se distribuye en forma Normal asintótica, con media θ y desviación estándar σ y por eso existe una probabilidad $1-\alpha$ tal que

$$\widehat{\theta} - 2\sigma < \theta < \widehat{\theta} + 2\sigma$$

aplique para una muestra aleatoria de la cual provenga $\widehat{\theta}$.
Para el caso de una distribución simétrica respecto a cero, se cumplirá que el negativo del percentil $100\left(\frac{\alpha}{2}\right)\%$ es igual al percentil $100\left(1-\frac{\alpha}{2}\right)\%$. En términos de un intervalo de confianza se expresa como.

$$\mathbb{P}\left[\widehat{\theta} - z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{V}(\widehat{\theta})} < \theta < \widehat{\theta} + z_{(1-\frac{\alpha}{2})}\sqrt{\widehat{V}(\widehat{\theta})}\right] = 1-\alpha.$$

4.2.6. Tamaño de las remuestras

Surge inmediatamente una pregunta al usar cualquiera de los métodos *Bootstrap*. ¿De qué tamaño deben elegirse los valores de m_h, el tamaño de las remuestras en cada estrato?.
Antes que nada debemos de recordar que se trabaja bajo el supuesto de que $n_h \geq 2$, $h = 1, \ldots, H$ y de que las unidades se seleccionan con reemplazo al momento de aplicar el *Bootstrap*. A continuación se presentan algunas propuestas de lo que en el pasado se ha aplicado.

1. Efron (1982) señala que para que el método sea eficiente debemos tener al menos kn ($k \geq 1$) estimaciones en cada iteración *Bootstrap*.

2. Rao y Wu (1988) recomiendan cuando $n_h = 2$ y por lo tanto $m_h = 1$, tomar $m_h = 3$ o $m_h = 4$ ya que la varianza obtenida por otro método de muestreo llamado por mitades o repeticiones balanceadas (BRR) sería más estable. De hecho cualquier estimador de varianza BRR puede ser visto como una aproximación al estimador de varianza *Bootstrap*.

3. En el *Bootstrap* con reemplazo (BWR) McCarthy y Snowden (1985) proponen para $h = 1, \ldots, H$ tomar
$$m_h \approx \frac{n_h - 1}{1 - \frac{n_h}{N_h}}$$

4. Rao y Wu (1988) sugieren tomar para $h = 1, \ldots, H$ cuando $n_h \geq 3$.
$$m_h \approx \frac{(1 - \frac{n_h}{N_h})(n_h - 2)^2}{(1 - 2\frac{n_h}{N_h})^2(n_h - 1)}$$

5. La simplificación de la relación anterior cuando la fracción de muestreo es pequeña.
$$m_h \approx \frac{(n_h - 2)^2}{(n_h - 1)} \tag{4.13}$$

Lenka *et al.* (2005) dicen que haciendo estudios empíricos, el *Bootstrap* con rescalamiento tiene un buen desempeño para estimar varianzas para funciones suaves cuando $m_h = n_h - 1$.
Siguiendo la observación de Lenka M. *et al.* (2005) se usó $m_h = n_h - 1$. Dando como resultado la tabla 4.2.

n_h	$m_h = n_h - 1$
2	1
3	2
5	4

n_h =Tamaño de muestra por estrato,
m_h =Tamaño de remuestra por estrato.

Tabla 4.2: Número de remuestras

4.2.7. Número de iteraciones *Bootstrap*

Surge otra pregunta: ¿Cuántas iteraciones *Bootstrap* deben considerarse para tener un *buen* intervalo de confianza?
El problema sigue abierto y en general varios autores señalan que no es un problema sencillo. A continuación presentamos algunas propuestas usadas anteriormente y reportadas en la literatura.

1. Calcular el método del percentil requiere de muchas iteraciones *Bootstrap*, así lo señala Efron (1982), quien sugiere que $50 \leq B \leq 200$ para construir una buena estimación de varianza *Bootstrap*. Y también sugiere que $1000 \leq B \leq 2000$ para construir un buen intervalo de confianza. Las afirmaciones se basan en la fórmula $B = \frac{1}{2}\epsilon_0^{-2}$, donde ϵ_0 es un valor predeterminado por el usuario, correspondiente a un cierto

margen de error permisible. Por ejemplo si $\epsilon_0 = 0{,}05$ entonces $B = 200$. Estas afirmaciones son respaldadas por Buckland (1984) el cual construyó una tabla en este contexto.

2. Babu y Singh (1983) muestran que B es una función de n que satisface

$$B \geq n \log n$$

3. Shi, Chen y Wu (1990) sugieren en cambio que

$$B = n^2 \log \log n$$

si pensamos que el error permisible es menor o igual a $\frac{1}{n}$.

En este trabajo se tomaron las siguientes iteraciones *Bootstrap* que se presentan a continuación en la tabla 4.3.
Se tomaron estos valores tratando de seguir la sugerencia de Efron (1982) para construir intervalos de confianza, después de hacer varios ensayos con el tamaño de muestra fijo, se observó que usar cantidades mayores a las reportadas en la tabla 4.3, prácticamente, ya no generaban cambios en los valores estimados de varianza.

n	B
600	700
900	1000
1500	1600

n =Tamaño de muestra,
B =Número de iteraciones *Bootstrap*.

Tabla 4.3: Número de iteraciones

Más recientemente Andrew y Buchinsky (1998) y Davidson y Mackinon (2000) proponen procedimientos para seleccionar el tamaño de iteraciones *Bootstrap* en la estimación de errores estándar, intervalos de confianza, regiones de confianza y contraste de hipótesis.
Como hemos visto, el objetivo del *Jackknife* y del *Bootstrap* es mas o menos el mismo. La estimación *Bootstrap* para la varianza es mejor que la obtenida con el *Jackknife*, al menos teóricamente, siempre que tengamos un número suficientemente grande de iteraciones.
En los intervalos de confianza para cualquier parámetro, el *Jackknife* supone que la distribución del muestreo es aproximadamente Normal; el método *Bootstrap* no depende necesariamente de este supuesto.

4.2.8. Estimadores calculados con el método *Bootstrap*

La fase de aplicación consiste en la obtención de los estimadores *Bootstrap* de proporción así como de sus varianzas y diferencias para tres tamaños de muestra diferentes.
Lo podemos resumir en los siguientes esquemas.
Se calculan $\widehat{\mathcal{R}}_{RS}$ usando (4.9) y $\widehat{\mathcal{B}}_{RS}$ usando (4.10).

$$\begin{cases} \widehat{\mathcal{R}}_{RS} & \text{Estimador combinado } \textit{Bootstrap} \text{ para la proporción.} \\ \widehat{\mathcal{B}}_{RS} & \text{Estimador separado } \textit{Bootstrap} \text{ para la proporción.} \end{cases}$$

Se calculan $\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ y $\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ usando (4.11) y (4.12) respectivamente.

$$\begin{cases} \widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS}) & \text{Estimador de varianza } \textit{Bootstrap} \\ & \text{del estimador combinado } \textit{Bootstrap} \text{ para la proporción.} \\ \widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS}) & \text{Estimador de varianza } \textit{Bootstrap} \\ & \text{del estimador separado } \textit{Bootstrap} \text{ para la proporción} \end{cases}$$

De igual manera para las diferencias, se calculan los estimadores $\widehat{\mathcal{D}}^{\mathcal{R}}_{RS}$ y $\widehat{\mathcal{D}}^{\mathcal{B}}_{RS}$ usando (4.9) y (4.10) respectivamente, para $\widehat{V}(\widehat{\mathcal{D}})$ se calcula con $\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$

y $\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$ con (4.11) y (4.12) respectivamente.

$$\begin{cases} \widehat{\mathcal{D}}^{\mathcal{R}}_{RS} & \text{Estimador combinado } \textit{Bootstrap} \\ & \text{de la diferencia de estimadores combinados } \textit{Bootstrap}. \\ \widehat{\mathcal{D}}^{\mathcal{B}}_{RS} & \text{Estimador separado } \textit{Bootstrap} \\ & \text{de la diferencia de estimadores separados } \textit{Bootstrap}. \end{cases}$$

$$\begin{cases} \widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS}) & \text{Varianza estimada } \textit{Bootstrap} \text{ para la diferencia} \\ & \text{de estimadores combinados } \textit{Bootstrap}. \\ \widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS}) & \text{Varianza estimada } \textit{Bootstrap} \text{ para la diferencia} \\ & \text{de estimadores separados } \textit{Bootstrap}. \end{cases}$$

Para terminar este capítulo presentamos una comparación de los métodos de remuestreo basada en la velocidad de convergencia.
Stanislav Kolenikov[2] (2002), tomando como base el método de Linealización presenta los siguientes resultados.
Diferentes estimadores *Jackknife* son $O_p(n^{-2})$ equivalentes,

- En el caso general
$$\frac{\widehat{V}_J(\widehat{\theta})}{\widehat{V}_L(\widehat{\theta})} = 1 + O_p(n^{-1});$$
- Cuando $n_h = 2$ para toda $h \in H$
$$\frac{\widehat{V}_J(\widehat{\theta})}{\widehat{V}_L(\widehat{\theta})} = 1 + O_p(n^{-2});$$
- Para estadísticas lineales o cuando $n_h = 2$ para toda $h \in H$ y $g(*)$ es una función cuadrática
$$\widehat{V}_J(\widehat{\theta}) = \widehat{V}_L(\widehat{\theta})$$

Con el método Bootstrap con rescalamiento existe evidencia empírica que señala que la varianza estimada no es muy estable.

[2]ver referencia[52]

- El histograma Bootstrap $\xrightarrow{P}$ a la distribución objetivo bajo la norma sup, y además
$$\frac{\widehat{V}_{RS}(\widehat{\theta})}{\widehat{V}_L(\widehat{\theta})} = 1 + O_p(n^{-1}).$$

Sin embargo Rao y Wu (1988)[3] señalan que agregando la siguiente condición.

$$\max_{h=1,\ldots,H} \frac{W_h}{n_h} = O(n^{-1}), \quad 0 < \delta_1 \leq \frac{m_h}{n_h - 1} \leq \delta_2 < \infty.$$

- En el caso de tener una función $g(*)$ no lineal
$$\frac{\widehat{V}_{RS}(\widehat{\theta})}{\widehat{V}_L(\widehat{\theta})} = 1 + O_p(n^{-2});$$

- En el caso de que $n_h = 2$ para toda $h \in H$, y se use el estimador de varianza *Jackknife* de Rao y Rust,
$$\frac{\widehat{V}_J(\widehat{\theta})}{\widehat{V}_L(\widehat{\theta})} = 1 + O_p(n^{-3})$$

En esta situación $\widehat{V}_J$ es asintoticamente mas cercano a $\widehat{V}_L$ que $\widehat{V}_{RS}$.

[3] Ver referencia[38]

Capítulo 5

Aplicación y resultados

En este capítulo se presenta la aplicación de los métodos, Linealización, *Jackknife* y *Bootstrap*, para calcular la varianza de varios estimadores y los intervalos de confianza asociados a ellos en un proceso de simulación con muestras estratificadas aleatorias simples, emulando conteos rápidos. Las estadísticas de interés, como se mencionó en el capítulo anterior son estimadores de proporciones y diferencias de proporciones usando estimadores de razón. Se comienza con la exposición del proceso de comparación de las simulaciones, los estimadores utilizados y posteriormente se muestran e interpretan los resultados.

5.1. Comparación de los métodos

En esta sección se presentan medidas para establecer la calidad de un estimador y de su estimador de varianza.

5.1.1. Medidas de estabilidad usando simulación

Las técnicas de estimación de varianza a analizar ya fueron presentadas: Linealización,*Bootstrap* y *Jackknife*. Para medir la calidad de los estimadores puntuales y de varianza se calculan las siguientes cantidades: el error cuadrático medio de los estimadores puntuales a lo largo de las simulaciones, el sesgo relativo al error cuadrático medio, la inestabilidad relativa de los estimadores de varianza y otros indicadores que describiremos con detalle mas adelante en este capítulo. Describimos la comparación de los métodos de la siguiente forma.

Al interior de cada estrato sólo se consideran tamaños de muestra fijos e iguales, i.e. $n_h = n_0 \ \forall h = 1, \ldots, H$. Con lo que el tamaño de la muestra global S es $n = n_0 H$. Como ya hemos mencionado n_0 toma los valores 2, 3, y 5 en esta aplicación.

Sitter (1992) utiliza el valor del error cuadrático medio de los estimadores puntuales aproximado mediante *Monte Carlo* como punto de referencia de la comparación. A diferencia de la propuesta original, en este trabajo los métodos de estimación de varianza son comparados con el valor de la varianza poblacional calculado bajo el método de Linealización de series de Taylor, V_L. El valor aproximado de V_L fue calculado mediante las fórmulas (3.3) y (3.6) para ambos estimadores de razón, adicionalmente se utilizó para calcular la varianza de las diferencias las fórmulas (3.11) y (3.12).

La simulación consiste en seleccionar $M = 1000$ muestras estratificadas con selección aleatoria simple al interior de cada uno de los estratos. Para cada una de las muestras se calculan los estimadores puntuales de razones y sus

respectivos estimadores de varianza.
La varianza de los estimadores es comparada en términos del sesgo relativo ($Bias$) a un estimador particular de varianza v y la inestabilidad relativa ($Instab$) con dicho estimador particular de varianza v.
A continuación definimos estas cantidades.

$$Bias := \frac{\bar{v} - V_L}{V_L} \tag{5.1}$$

$$Instab := \frac{\sigma_v}{V_L} \tag{5.2}$$

en donde,
V_L= Es la varianza poblacional del estimador calculada mediante
Linealización de Taylor,
v_m= Es la varianza estimada del estimador utilizando alguno de los tres métodos: Linealización, *Jackknife* o *Bootstrap* en la simulación m,

$$\bar{v} := \frac{1}{M}\sum_{m=1}^{M} v_m,$$

$$\sigma_v^2 := \frac{1}{M}\sum_{m=1}^{M} \left(v_m - V_L\right)^2,$$

Para comparar varios intervalos de confianza estimados, se utilizan los porcentajes de error en la cola superior e inferior, L y U. En este caso fueron comparadas con una tasa de error del 5 % en ambas colas.
A continuación definimos ambos porcentajes.

1. Porcentaje inferior de error en la cola.

$$L := \frac{100\ \#S\left(R < \widehat{R}_{Lm}\right)}{M} \tag{5.3}$$

En donde,
R = Es el valor poblacional de la proporción de interés,

$\widehat{R}_{Lm}$ = Es el valor estimado del extremo inferior del intervalo de confianza de la proporción en la m-ésima simulación,
$\#S\left(R < \widehat{R}_{Lm}\right)$ = es la cantidad de muestras que satisfacen la desigualdad $R < \widehat{R}_{Lm}$.

2. Porcentaje superior de error en la cola.

$$U := \frac{100\ \#S\left(R > \widehat{R}_{Um}\right)}{M} \tag{5.4}$$

En donde,
$\widehat{R}_{Um}$ = Es el valor estimado del extremo superior del intervalo de confianza de la proporción en la m-ésima simulación,
$\#S\left(R > \widehat{R}_{Um}\right)$ = es la cantidad de muestras que satisfacen la desigualdad $R > \widehat{R}_{Um}$.

Finalmente para comparar las longitudes de los intervalos estimados y los teóricos se utiliza lo siguiente.
Longitud estandarizada del intervalo.

$$Length = \frac{\frac{1}{M}\sum_{m=1}^{M}\left(\widehat{R}_{Um} - \widehat{R}_{Lm}\right)}{2z_{(1-\frac{\alpha}{2})}\sqrt{V_L}} \tag{5.5}$$

donde $\left(\widehat{R}_{Lm}, \widehat{R}_{Um}\right)$ es el intervalo de confianza estimado con la muestra m y $z_{(1-\frac{\alpha}{2})}$ es el cuantil de una Normal estándar.
En el mejor de los casos se deben obtener valores cercanos a los siguientes.

1. $Bias \approx 0$, $instab \approx 0$, $Length \approx 1$.

2. $L = 2.5$ y $U = 2.5$, si $\alpha = 0.05$.

Como ya se mencionó, estas medidas fueron definidas en el trabajo de Sitter (1992) pero usando ECM en lugar de V_L. En dicho trabajo se hizo un estudio de simulación donde compara los resultados del *Jackknife*, Linealización

y varias versiones del *Bootstrap*. Consideró ocho poblaciones bajo diseños estratificados simples sin reemplazo, con distintas cantidades de estratos y unidades. Concluyó que el estimador de razón *Jackknife* se comporta mejor que el estimador *Bootstrap*. En su trabajo se analizaron también estimadores de regresión y estimaciones de cuantiles como la mediana, resultando obtener errores grandes en las colas.

5.2. Desarrollo

La población que se está utilizando es la base de datos del PREP del año 2000. Esta población está subdividida mediante la partición que proporcionan los distritos electorales con lo que se cuentan 300 estratos.
Las unidades de esta población son las secciones electorales, originalmente suman 63445. Pero dos secciones electorales presentaban datos perdidos por lo que fueron eliminadas de la población, así se trabajó con 63443.

Como ya se ha mencionado, se trabaja con un esquema de extracción de muestras estratificadas y la selección de los elementos al interior de los estratos es aleatoria simple.
Se hicieron 1000 simulaciones, las cuales consistieron en la extracción repetida de este tipo de muestras que podemos considerar como conteos rápidos.

Las simulaciones se hacen considerando muestras de tres tamaños como podemos ver en la tabla 5.1 siguiente, al interior de cada estrato se toma una cantidad constante de elementos.
Al eliminar dos secciones electorales, los valores de los totales de las variables cambian, y en consecuencia los valores poblacionales de diversas proporciones también. Con la finalidad de referirnos a esos valores después, presentamos en la tabla 5.2 los totales poblacionales, en las variables que en este caso

Tabla 5.1: Tamaños muestrales utilizados en este trabajo

n	600	900	1500
n_h	2	3	5

n =Tamaño de muestra.
n_h =Tamaño de muestra por estrato. Para $h = 1, \ldots, 300$.

son los totales de votos a favor de cada partido y también se presentan las proporciones que tienen estas variables con respecto a los votos válidos totales de la elección.

Variable	Total de votos	$R = \frac{Partido}{Validos}$
PRI	13 579 718	**0.3688**
AC	15 989 636	**0.4343**
AM	6 256 780	**0.1699**
OTROS	955 866	**0.0259**
VALIDOS	36 813 461	1

Total de votos=Cantidad total de votos válidos del PREP.
R =Proporción de votos a favor de la Fuerza política.
PRI=Partido Revolucionario Institucional.
AC=Alianza por el cambio, integrado por PAN y PVEM.
AM=Alianza por México, integrado por PRD, CD, PSN y PAS.
OTROS=Partidos politicos restantes, PCD,PARM y DSPPN.

Tabla 5.2: Totales y razones poblacionales de las variables de la aplicación

De la misma forma que en la Tabla 5.2. Se calcularon las diferencias entre los porcentajes de los partidos. En la tabla 5.3 se presentan los valores de

estas diferencias, con el fin de compararlos posteriormente con los valores estimados en las simulaciones.

Variable	$R_i - R_j$
AC-PRI	**0.0654**
PRI-AM	0.1989
PRI-OTROS	0.3429
AC-AM	0.2643
AC-OTROS	0.4083
AM-OTROS	0.1439

Tabla 5.3: Diferencias entre las proporciones poblacionales

Como se puede observar en la Tabla 5.3, los partidos que se encuentran más próximos en el valor de proporción de votos válidos a favor son el PRI y la AC. En consecuencia para la estimación de proporciones y de las diferencias el único caso interesante se presenta en la diferencia del PRI con AC. Para no hacer extensivo este trabajo, solo presentamos resultados sobre los estimadores de AC, PRI y de las diferencias AC-PRI.

La eficiencia del estimador de razón depende en gran medida de la relación existente entre las variables numerador y denominador, siendo el caso más favorable aquel en el que existe una relación aproximada de proporcionalidad entre las variables. Gráficamente ello significa una nube de puntos concentrada en las proximidades de una línea recta que pasa por el origen.

En la figura 5.1 podemos observar que la relación que guardan las variables PRI y AC con la variable que cuenta la cantidad total de votos válidos en la elección, es de buena proporcionalidad. En mucho menor grado la proporcionalidad aproximada está presente para una nueva variable diferencia AC-PRI, que se presenta en la figura 5.2.

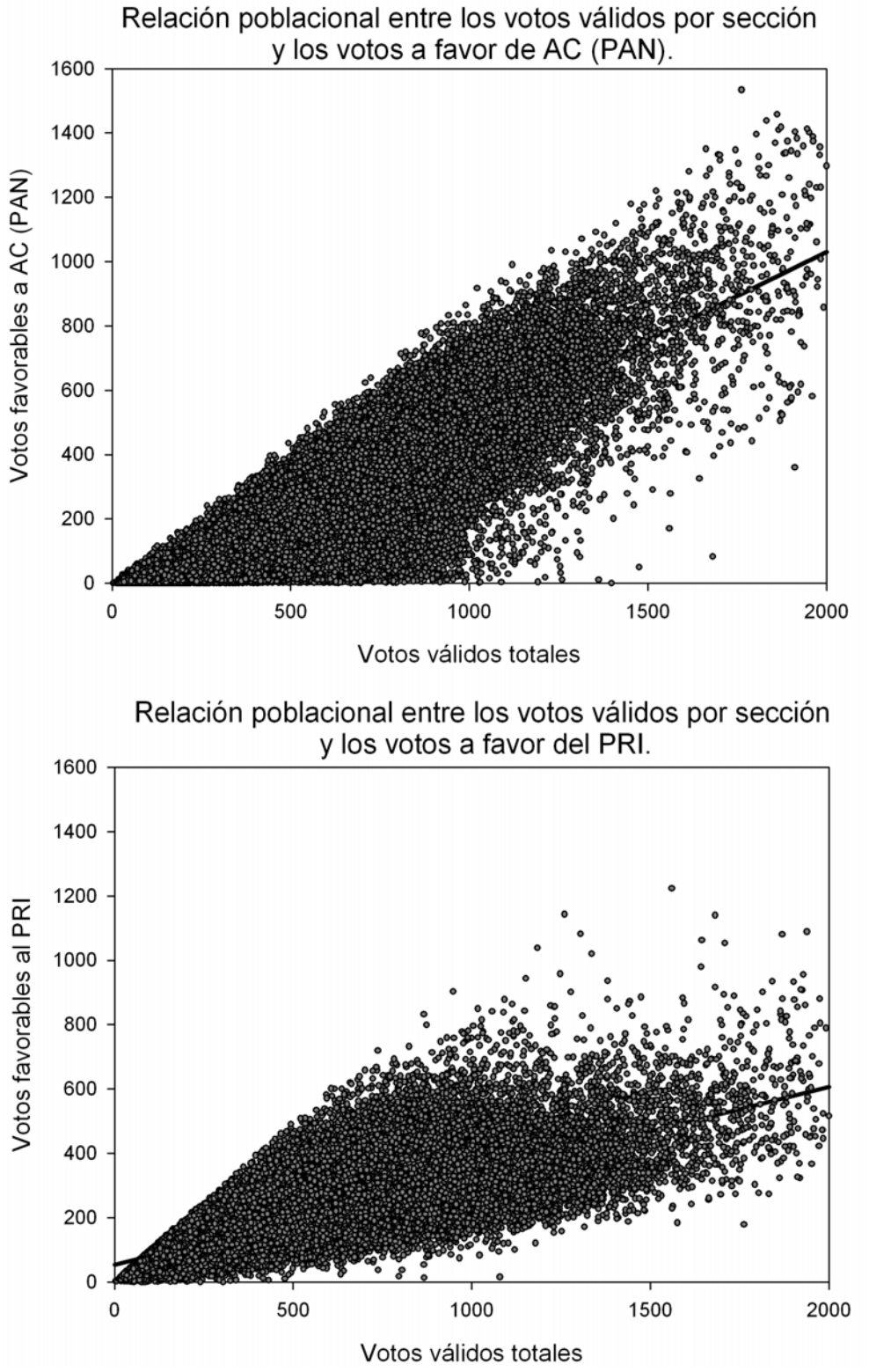


Figura 5.1: Proporcionalidad aproximada entre votos favorables a AC, PRI y válidos totales

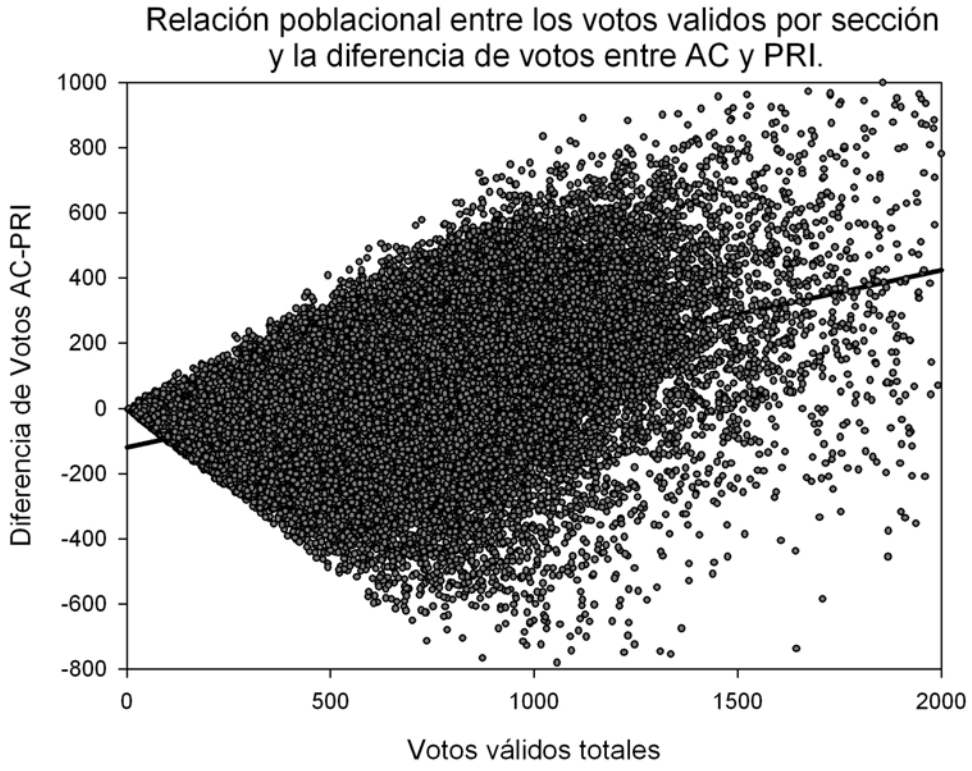


Figura 5.2: Proporcionalidad aproximada entre AC-PRI y válidos totales

5.2.1. Cálculo de varianzas de poblacionales

Como paso inicial se construyen los valores que deben tener las varianzas para los estimadores de razón combinado y separado de AC y PRI usando las fórmulas (3.3) y (3.6) respectivamente mediante Linealización. Después se calcula la varianza para la diferencia AC-PRI usando las fórmulas (3.9) y (3.11), para el estimador combinado (3.9) y (3.12) para el estimador separado nuevamente usando Linealización.
Esto se hace para los tres tamaños de muestra mencionados 600, 900 y 1500. Podemos resumir en el siguiente esquema el trabajo realizado. Se calculan los elementos encerrados en las llaves.

$$\begin{cases} V_L(\widehat{\mathcal{R}}) & \text{Varianza del estimador combinado para la proporción mediante Linealización.} \\ V_L(\widehat{\mathcal{B}}) & \text{Varianza del estimador separado para la proporción mediante Linealización.} \end{cases}$$

$$\begin{cases} V_L(\widehat{\mathcal{D}}^{\mathcal{R}}) & \text{Varianza para la diferencia de estimadores combinados mediante Linealización.} \\ V_L(\widehat{\mathcal{D}}^{\mathcal{B}}) & \text{Varianza para la diferencia de estimadores separados mediante Linealización.} \end{cases}$$

Los cálculos se realizaron empleando el programa 2 del apéndice E.
Obteniendo estos datos, presentamos en las tablas (5.4), (5.5) y (5.6) los valores correspondientes a los errores de muestreo.
Como resultado de estas tablas, a continuación se presentan los intervalos de confianza para los estimadores de razón y sus diferencias en la tabla (5.7).

Tabla 5.4: Error de estimación AC, $EE = 100 * 1.96\sqrt{V_L(\widehat{R})}$.

Partido	**AC**	
	EE	
n	Combinado	Separado
600	1.119	1.037
900	0.912	0.845
1500	0.703	0.651

Tabla 5.5: Error de estimación PRI, $EE = 100 * 1.96\sqrt{V_L(\widehat{R})}$.

Partido	**PRI**	
	EE	
n	Combinado	Separado
600	0.917	0.892
900	0.747	0.726
1500	0.576	0.560

Tabla 5.6: Error de estimación proporción AC-PRI, $EE = 100 * 1.96\sqrt{V_L(\widehat{d})}$.

Partido	**AC-PRI**	
	EE	
n	Combinado	Separado
600	1.887	1.808
900	1.537	1.473
1500	1.185	1.135

Tabla 5.7: Intervalos de C. del porcentaje de votos por Linealización de Taylor

Fuerza Política	Tipo de Estimador	$n = 600$		$n = 900$		$n = 1500$	
		I^-	I^+	I^-	I^+	I^-	I^+
PRI	Combinado	35.97	37.80	36.14	37.63	36.31	37.46
PRI	Separado	35.99	37.77	36.16	37.61	36.32	37.44
AC	Combinado	42.31	44.55	42.52	44.34	42.73	44.13
AC	Separado	42.39	44.47	42.58	44.27	42.78	44.08
Diferencia	Tipo de Estimador	inf	sup	inf	sup	inf	sup
AC-PRI	Combinado	4.65	8.43	5.00	8.08	5.36	7.73
AC-PRI	Separado	4.73	8.35	5.07	8.01	5.41	7.68

R =Proporción de votos a favor de la Fuerza política,
$I^- = 100(R - 1{,}96\sqrt{V_L(\widehat{R})})$,
$I^+ = 100(R + 1{,}96\sqrt{V_L(\widehat{R})})$,
Combinado=Estimador combinado,
Separado=Estimador separado.

En la tabla 5.7 se puede observar que los intervalos generados con estimadores combinados siempre son más amplios que los generados con estimadores separados. Cochran(1977, p.167) ya había observado este hecho.

5.2.2. Cálculo de varianzas en simulaciones

Para la fase de simulaciones, para cada muestra se siguen los siguientes pasos.

1. Se extrae una muestra estratificada aleatoria simple de secciones electorales de la población. Para la obtención de estas muestras se utilizó el programa 1 que se encuentra en el apéndice E.

2. Se calculan los estimadores de razón usando (3.2) para el estimador combinado $\widehat{\mathcal{R}}$ y (3.5) para el estimador separado $\widehat{\mathcal{B}}$ con los partidos AC y PRI.

   Se calculan los estimadores de varianza de estos estimadores, usando (3.4) para el combinado $\widehat{V}_L(\widehat{\mathcal{R}})$ y (3.7) para el separado $\widehat{V}_L(\widehat{\mathcal{B}})$ con los partidos AC y PRI.

   Se calcula el estimador de la diferencia de razones de AC-PRI con (3.8) usando los estimadores de razón combinado y separado.

   Se calculan los estimadores de varianza de la diferencia AC-PRI para la estimación combinada $\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$ y separada $\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$ utilizando(3.14) y (3.15).

   En un esquema las estimaciones son las siguientes. Se calculó.

   $$\begin{cases} \widehat{\mathcal{R}} & \text{Estimador combinado para las proporciones AC y PRI.} \\ \widehat{\mathcal{B}} & \text{Estimador separado para las proporciones AC y PRI.} \end{cases}$$

   $$\begin{cases} \widehat{V}_L(\widehat{\mathcal{R}}) & \text{Estimador de varianza por Linealización del} \\ & \text{estimador combinado para las proporciones AC y PRI.} \\ \widehat{V}_L(\widehat{\mathcal{B}}) & \text{Estimador de varianza por Linealización del} \\ & \text{estimador separado para la proporciones AC y PRI} \end{cases}$$

   $$\begin{cases} \widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}}) & \text{Varianza estimada por Linealización para la diferencia} \\ & \text{del estimador combinado de AC-PRI.} \\ \widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}}) & \text{Varianza estimada por Linealización para la diferencia} \\ & \text{del estimador separado de AC-PRI.} \end{cases}$$

3. Se calculan los estimadores de varianza de los estimadores bajo la metodología *Jackknife*, usando (4.4) para ambos tipos de estimadores, $\widehat{V}_J(\widehat{\mathcal{R}})$ y $\widehat{V}_J\widehat{\mathcal{B}})$ con los partidos AC y PRI.

   Se calculan los estimadores de varianza de la diferencia AC-PRI bajo la metodología *Jackknife*, usando nuevamente (4.4) para ambos tipos

de estimadores, $\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$ y $\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$.

Esquemáticamente se traduce en calcular lo siguiente.

$$\begin{cases} \widehat{V}_J(\widehat{\mathcal{R}}) & \text{Estimador de varianza } \textit{Jackknife} \text{ del} \\ & \text{estimador combinado para las proporciones AC y PRI.} \\ \widehat{V}_J(\widehat{\mathcal{B}}) & \text{Estimador de varianza } \textit{Jackknife} \text{ del} \\ & \text{estimador separado para la proporciones AC y PRI} \end{cases}$$

$$\begin{cases} \widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}}) & \text{Varianza estimada } \textit{Jackknife} \text{ de la diferencia} \\ & \text{de estimadores combinados de AC-PRI.} \\ \widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}}) & \text{Varianza estimada } \textit{Jackknife} \text{ de la diferencia} \\ & \text{de estimadores separados de AC-PRI.} \end{cases}$$

4. Se calculan los estimadores de razón *Bootstrap* usando (4.9) para el estimador combinado $\widehat{\mathcal{R}}_{RS}$ y (4.10) para el estimador separado $\widehat{\mathcal{B}}_{RS}$ con los partidos AC y PRI.

   Se calculan los estimadores de varianza de los estimadores, usando (4.11) para el combinado $\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ y (4.12) para el separado $\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ con los partidos AC y PRI.

   Se calculan los estimadores de la diferencia de razones de AC-PRI, $\widehat{\mathcal{D}}^{\mathcal{R}}_{RS}$ y $\widehat{\mathcal{D}}^{\mathcal{B}}_{RS}$ con las mismas fórmulas para los estimadores de razón combinado y separado.

   Se calculan los estimadores de varianza de la diferencia AC-PRI para la estimación combinada $\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$ y separada $\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$.

   Para las muestras de tamaño $n = 600$ se toman $B = 700$ submuestras de tamaño 300 cada una, $n_h = 2$ y $m_h = 1$.

   Para las muestras de tamaño $n = 900$ se toman $B = 1000$ submuestras de tamaño 600 cada una, $n_h = 3$ y $m_h = 2$.

   Para las muestras de tamaño $n = 1500$ se toman $B = 1600$ submuestras de tamaño 1200 cada una, $n_h = 5$ y $m_h = 4$.

   Los intervalos de confianza *Bootstrap* fueron construidos por medio de

los *ICB estándar* descritos anteriormente.
Los valores estimados son lo siguientes.

$$\widehat{R}\begin{cases}\widehat{\mathcal{R}}_{RS} & \text{Estimador combinado } \textit{Bootstrap} \text{ para las proporciones AC y PRI.}\\ \widehat{\mathcal{B}}_{RS} & \text{Estimador separado } \textit{Bootstrap} \text{ para las proporciones AC y PRI.}\end{cases}$$

$$\widehat{V}(\widehat{R})\begin{cases}\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS}) & \text{Estimador de varianza } \textit{Bootstrap} \text{ del estimador}\\ & \text{combinado } \textit{Bootstrap} \text{ para las proporciones AC y PRI.}\\ \widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS}) & \text{Estimador de varianza } \textit{Bootstrap} \text{ del estimador}\\ & \text{separado } \textit{Bootstrap} \text{ para la proporciones AC y PRI}\end{cases}$$

$$\widehat{\mathcal{D}}\begin{cases}\widehat{\mathcal{D}}^{\mathcal{R}}_{RS} & \text{Estimador combinado } \textit{Bootstrap} \text{ para la diferencia AC-PRI.}\\ \widehat{\mathcal{D}}^{\mathcal{B}}_{RS} & \text{Estimador separado } \textit{Bootstrap} \text{ para la diferencia AC-PRI.}\end{cases}$$

$$\widehat{V}(\widehat{\mathcal{D}})\begin{cases}\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS}) & \text{Varianza estimada } \textit{Bootstrap} \text{ de la diferencia}\\ & \text{de estimadores combinados } \textit{Bootstrap} \text{ de AC-PRI.}\\ \widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS}) & \text{Varianza estimada } \textit{Bootstrap} \text{ de la diferencia}\\ & \text{de estimadores separados } \textit{Bootstrap} \text{ de AC-PRI.}\end{cases}$$

Se ha mencionado repetidamente que se usaron 3 tamaños de muestra distintos, 600, 900, y 1500. Para cada tamaño de muestra se generaron 1000 muestras haciendo un total de 3000.
Terminadas las simulaciones se construyeron los indicadores de estabilidad presentados en el inicio de este capítulo.
Bias con (5.1), $Instab$ con (5.2), L con (5.3), U con (5.4) y *Length* con (5.5).
Estas medidas se presentan en tablas para cada tamaño de muestra, con cada fuerza política AC, PRI y para la diferencia y AC-PRI.

5.3. Ejemplo: Interpretación de las medidas de estabilidad

Para facilitar la interpretación de las medidas de estabilidad se proporcionan como ejemplo las gráficas correspondientes a la tabla 5.18, que son los resultados sobre las medidas de estabilidad para los estimadores de las diferencias AC-PRI y $n = 1500$ secciones electorales, en las figuras 5.3 y 5.8.

En la figura 5.3 se presentan los porcentajes de error en ambas colas que llamamos L y U, el estimador de razón combinado presenta valores cercanos a 2.5 en cada porcentaje como se espera que suceda cuando los intervalos de confianza se construyen con una confianza del 95 %. También se observa que los métodos de estimación de varianza Linealización y *Jackknife* presentan valores muy cercanos a los esperados y el *Bootstrap* valores más alejados.

El estimador de razón separado presenta valores muy diferentes a los esperados pues en todos los métodos presenta el valor de $L = 0$. Lo cual quiere decir que a lo largo de las simulaciones los intervalos de confianza estimados no dejaron escapar al valor poblacional de la proporción diferencia a través del límite inferior.
Presenta en contraste este estimador valores muy elevados para U, siendo más crítico en el método de estimación *Bootstrap*. Lo cual significa que el valor poblacional de la proporción AC-PRI no se encuentra en muchos intervalos de confianza estimados, porque se encuentra fuera de este y se encuentra muy por encima de los intervalos de confianza.

En la figura 5.4 se presentan los valores de sesgo e inestabilidad relativa de los estimadores de varianza.
A lo largo de los distintos métodos, los estimadores de varianza con menor

magnitud en el sesgo son los correspondientes a los estimadores de razón combinados. Debido a que estos valores están multiplicados por 100 se les puede interpretar como porcentajes.
Los estimadores de varianza asociados al estimador de razón separado presentan más frecuentemente sesgo hacia abajo y los asociados a los estimadores de razón combinados presentan sesgo hacia arriba.
En el método de Linealización el sesgo es negativo en ambos tipos de estimadores, indicando esto que al usar este método se corre el riesgo de subestimar las varianzas.
En términos de inestabilidad los valores que se presentan en general son pequeños para todos los métodos de estimación de varianza, resalta en la figura que bajo los métodos de Linealización y *Jackknife* el estimador de razón combinado para la diferencia AC-PRI se comporta más estable que el estimador de razón separado. Solamente en el método *Bootstrap* el comportamiento es al revés.
En la figura 5.5 observamos la longitud estandarizada de los intervalos de confianza. Se espera que esta medida tome valores cercanos a 1. En la figura este valor se marca con una línea horizontal punteada.
En este ejemplo el método de Linealización produce los intervalos más estrechos, el *Jackknife* produce los más amplios y el *Bootstrap* produce los más correctos en este contexto.
También podemos ver que los intervalos producidos con estimadores de varianza que usan el estimador de razón combinado son mejores que los del estimador separado.

Todos estos indicadores señalan que usar el estimador de razón combinado es preferible que usar el estimador de razón separado porque se presentan menores sesgo, mayor estabilidad y los intervalos de confianza tienen la longitud esperada.

Para complementar el ejemplo se presentan gráficamente los intervalos de confianza a lo largo de las mil simulaciones.
La figura 5.6 compara los intervalos producidos mediante Linealización para el estimador combinado contra el separado.
Las figuras 5.7 y 5.8 hacen la misma comparación bajo los métodos *Jackknife* y *Bootstrap*.
En cada caso el intervalo de confianza producido con la varianza poblacional calculada por Linealización se presenta en todas las figuras 5.6 a 5.8 como una franja con dos líneas punteadas como fronteras las cuales son el extremo superior e inferior de intervalo. La franja encierra una línea continua y constante que representa el valor poblacional de la proporción que en este ejemplo es una diferencia.

En las figuras 5.6 a 5.8 se aprecia que los estimadores puntuales de razón separada para la diferencia en la mayoría de las simulaciones no se encuentran en medio de la franja mencionada lo cual es indicativo de que estos estimadores subestiman el valor poblacional. Esto genera que los intervalos de confianza estén desplazados hacia abajo como ya lo habíamos mencionado y también es indicado con el valor de $L = 0$. En contraste los estimadores de razón combinada se mantienen en su mayoría y a lo largo de todos los métodos, dentro de la franja.

En la figura 5.9 se presentan los histogramas de los estimadores de razón combinados y separados, con una línea vertical se ubican los valores poblacionales.

Para terminar este ejemplo presentamos las figuras 5.10 a 5.12 en donde se presentan los histogramas correspondientes a los errores de muestreo de cada

uno de los métodos.El error de muestreo esta multiplicado por 100, calculados con $100*(1.96\sqrt{\widehat{V}(\widehat{\mathcal{D}})})$. Nuevamente los valores poblacionales están marcados con una línea vertical.

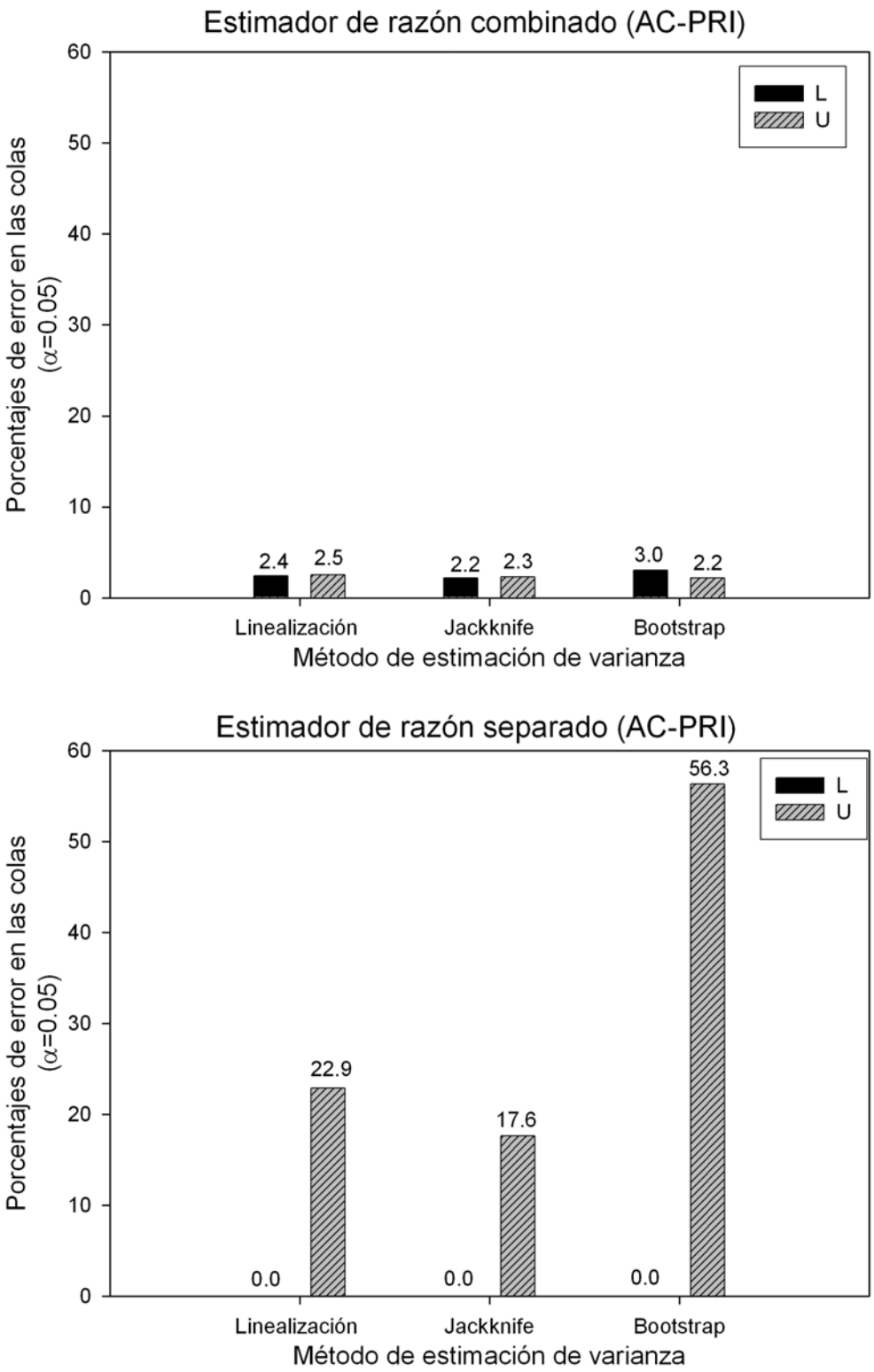
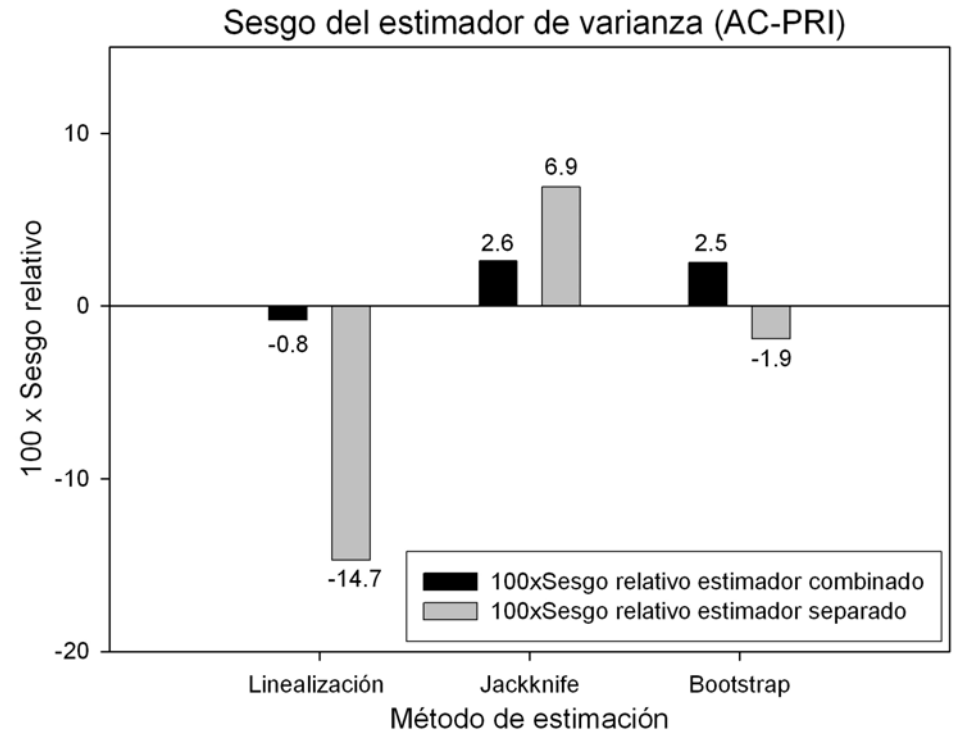


Figura 5.3: Error en las colas de los estimadores combinado y separado de diferencias AC-PRI.



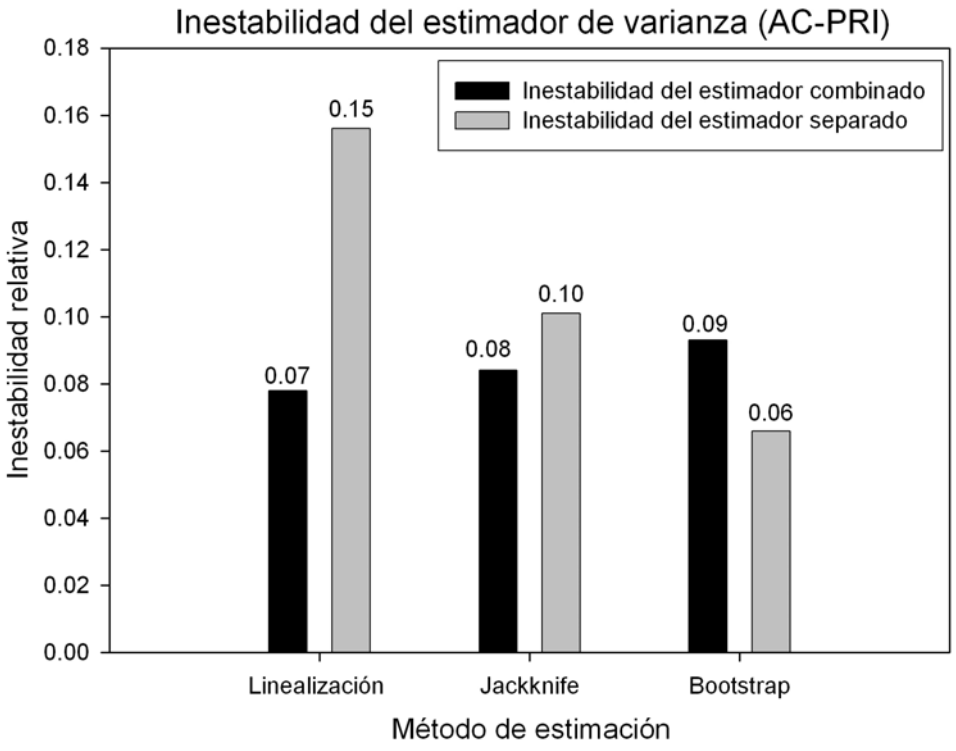


Figura 5.4: Sesgo e inestabilidad de los estimadores de diferencias AC-PRI.

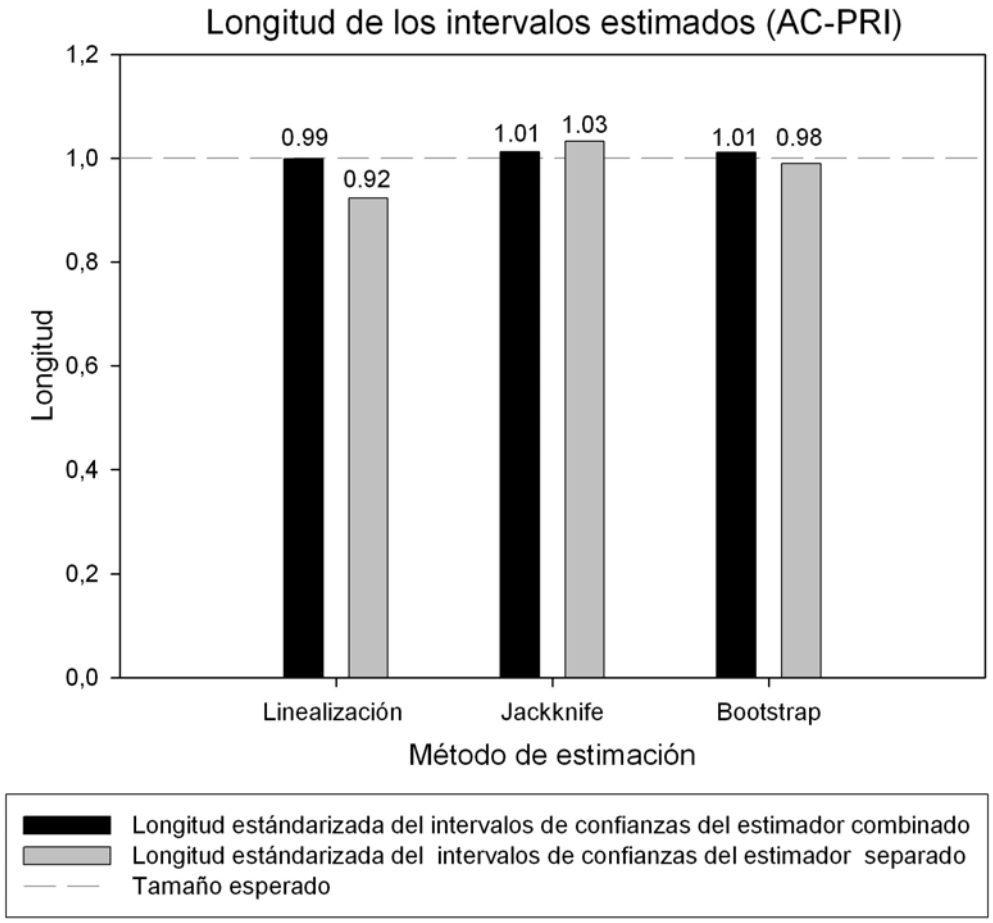


Figura 5.5: Longitud estandarizada de intervalos para las diferencias AC-PRI.

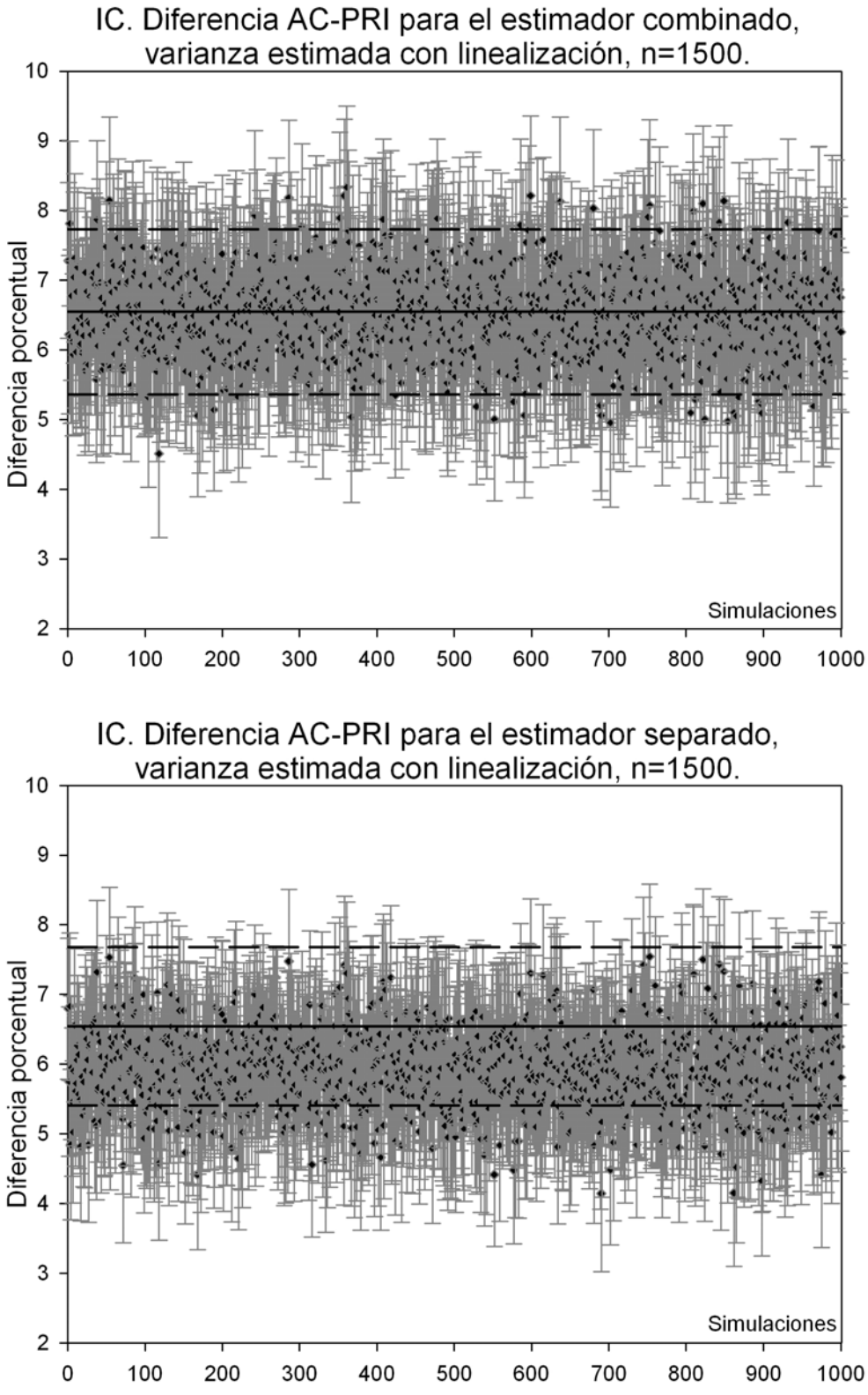


Figura 5.6: Intervalos calculados con Linealización, $100 * (\widehat{\mathcal{D}} \pm 1.96\sqrt{\widehat{V}_L(\widehat{\mathcal{D}})})$.

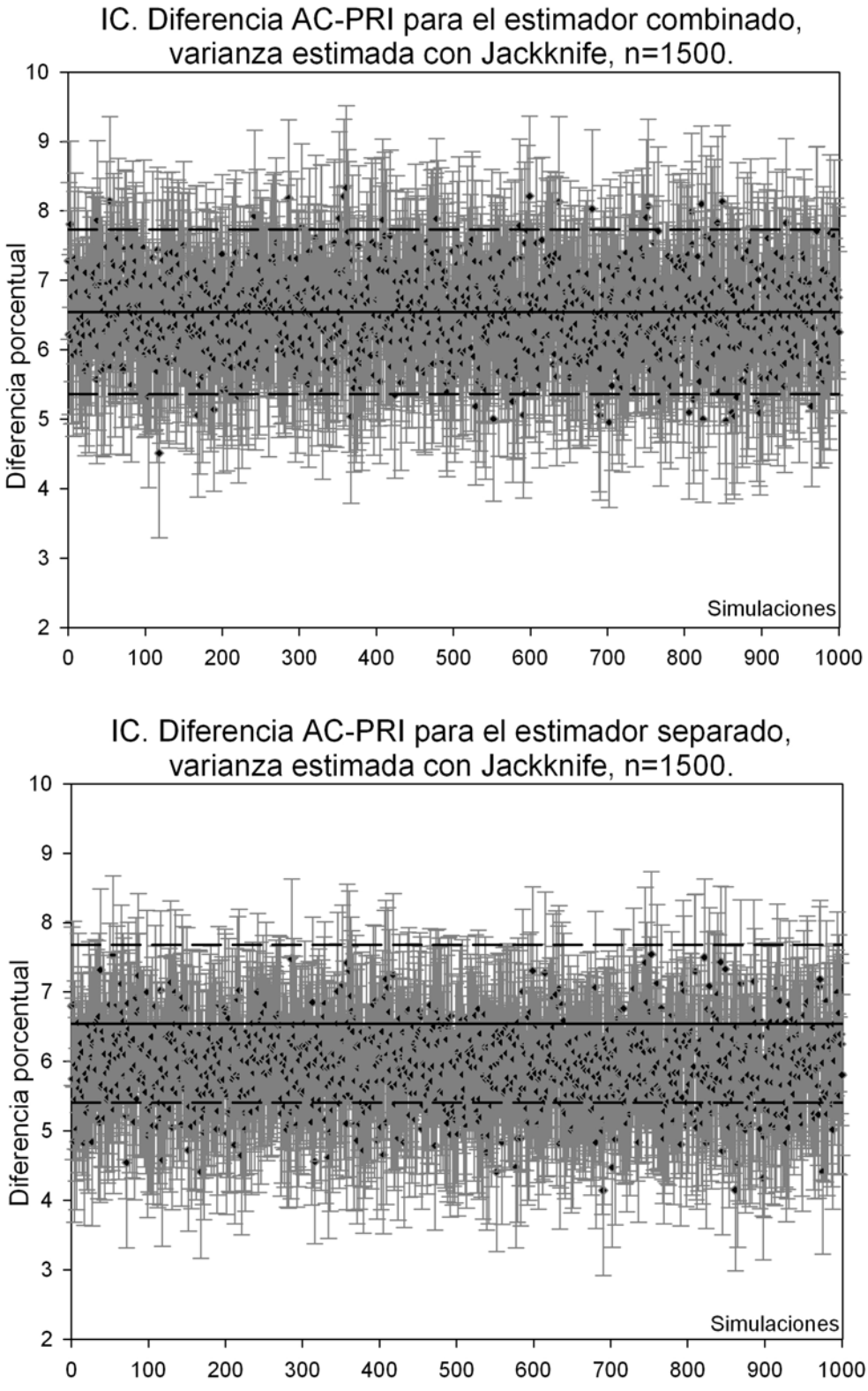


Figura 5.7: Intervalos calculados con *Jackknife*, $100 * (\widehat{\mathcal{D}} \pm 1.96\sqrt{\widehat{V}_J(\widehat{\mathcal{D}})})$.

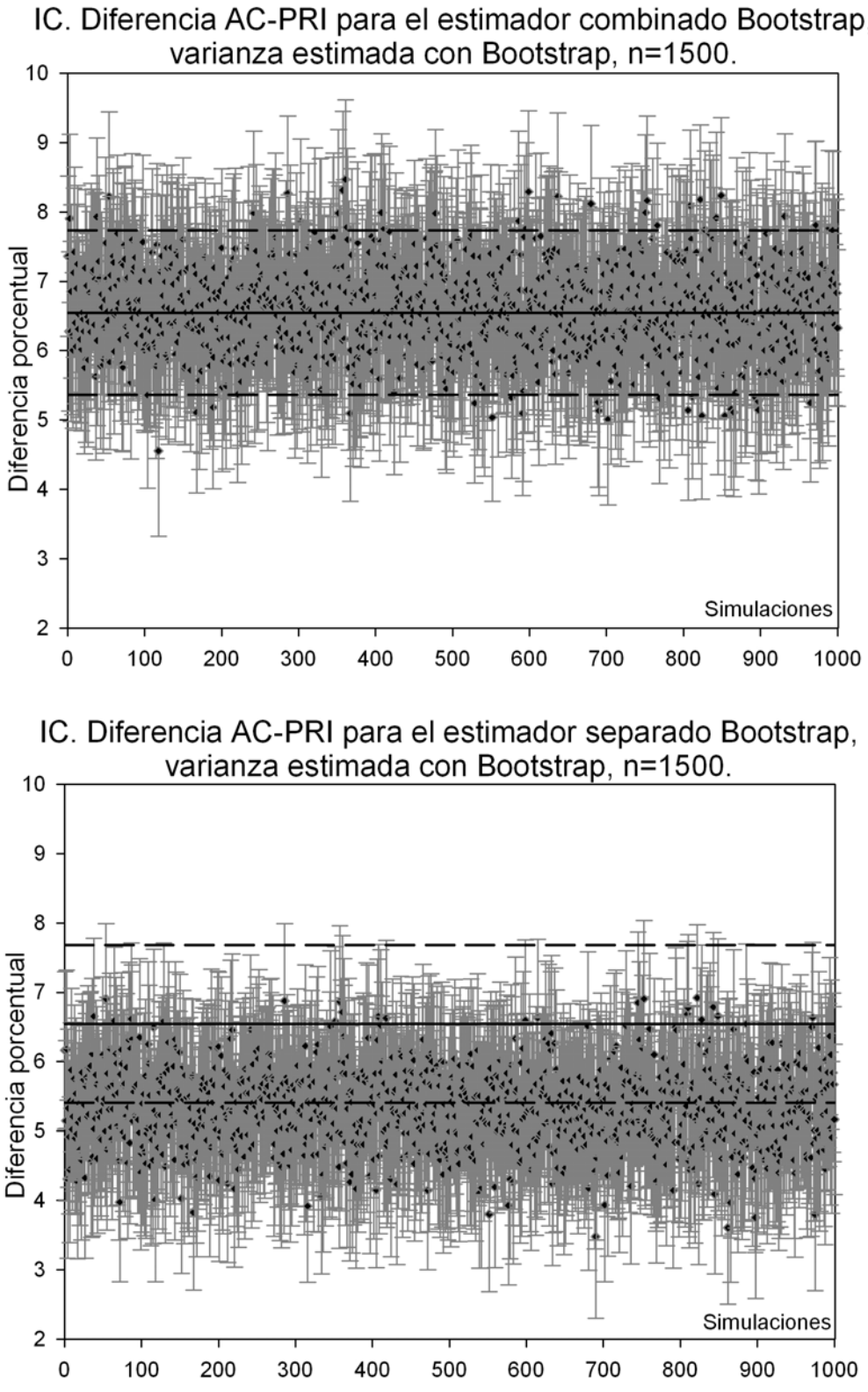


Figura 5.8: Intervalos calculados con *Bootstrap*, $100*(\widehat{\mathcal{D}}_{RS} \pm 1.96\sqrt{\widehat{V}_{RS}(\widehat{\mathcal{D}}_{RS})})$

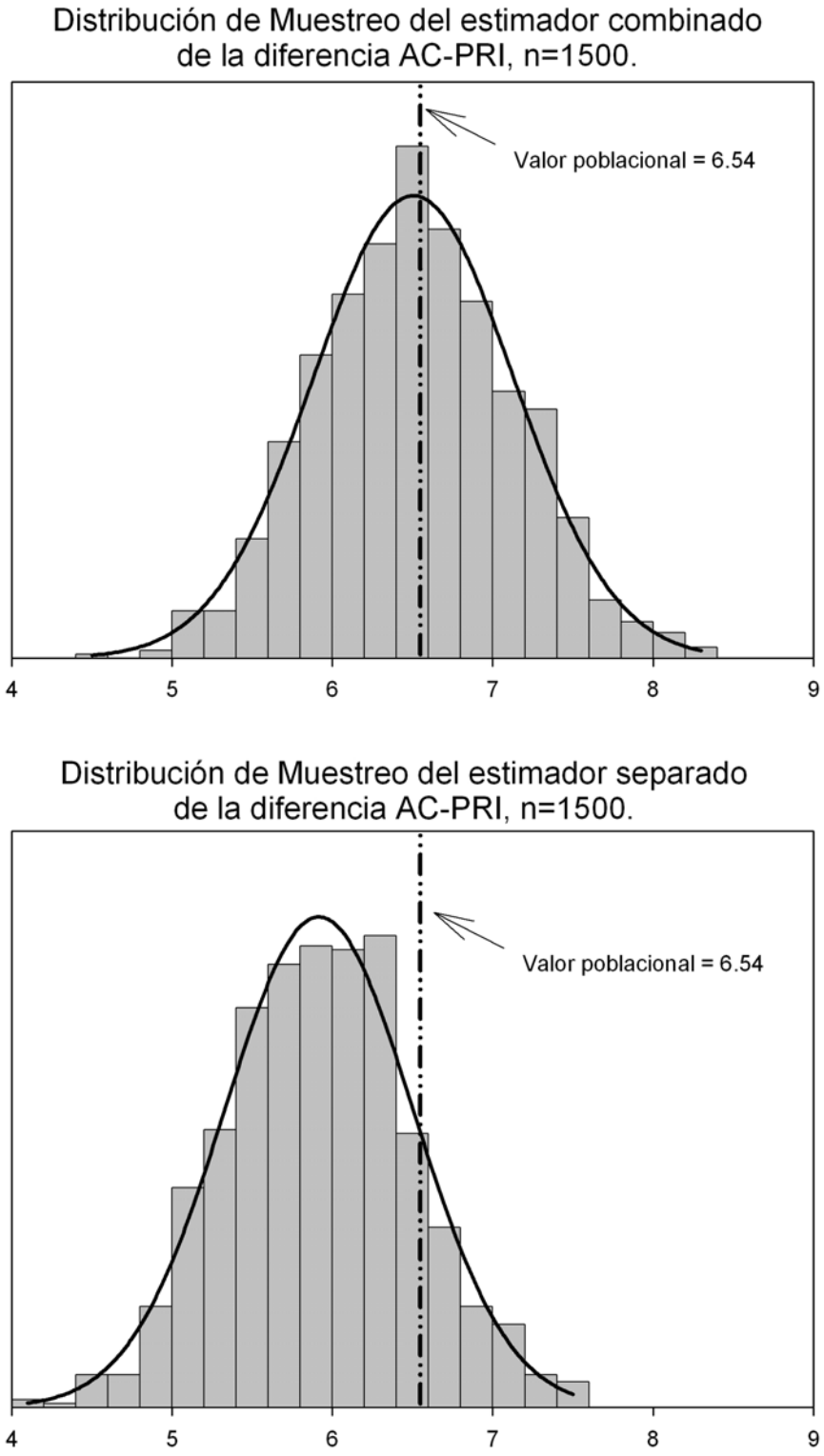


Figura 5.9: Distribución de muestreo de los Estimadores de diferencias AC-PRI.

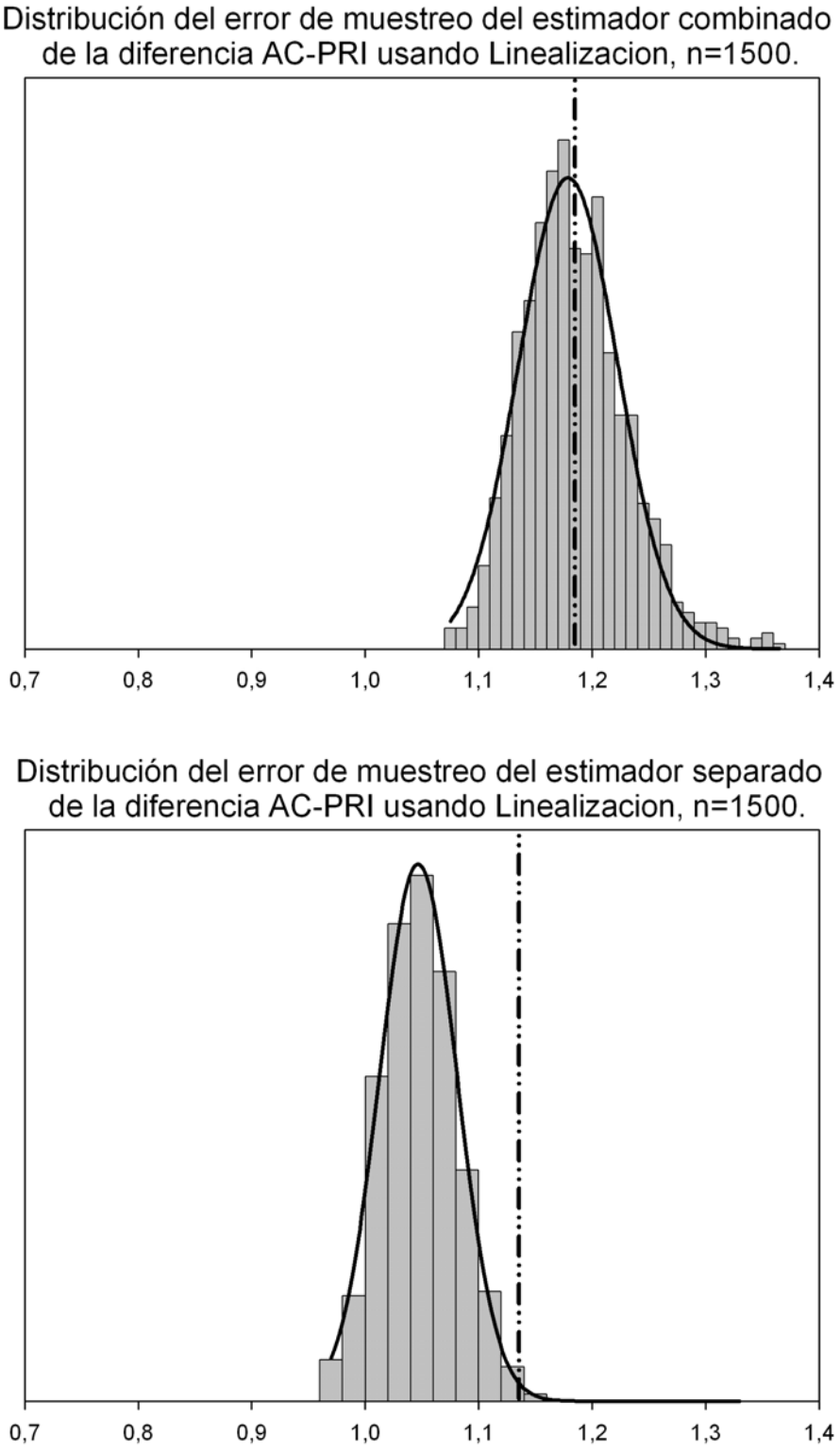


Figura 5.10: Errores de muestreo con Linealización, $100 * (1.96\sqrt{\widehat{V}_L(\widehat{\mathcal{D}})})$

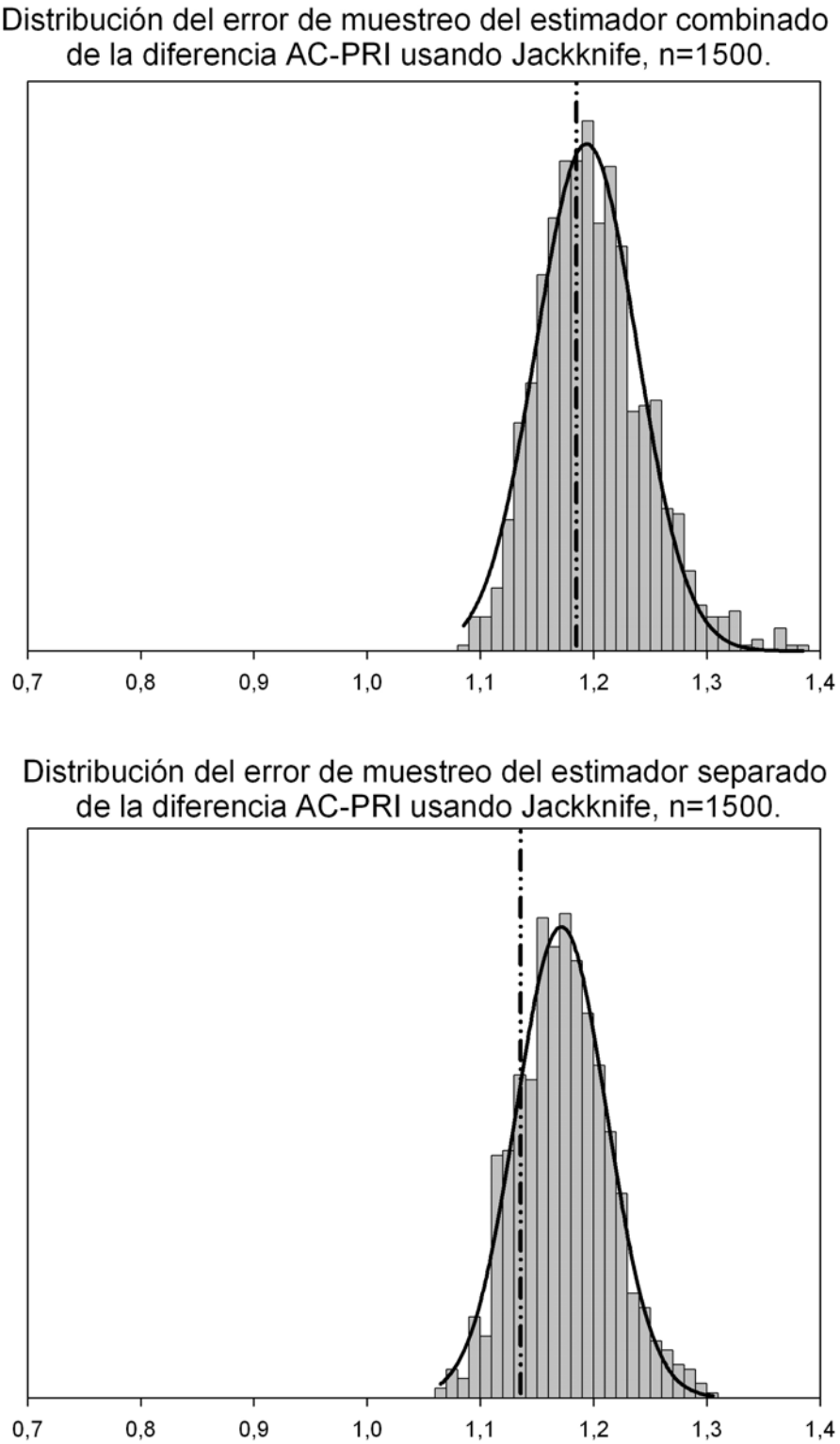


Figura 5.11: Errores de muestreo con *Jackknife*, $100 * (1.96\sqrt{\widehat{V}_J(\widehat{\mathcal{D}})})$

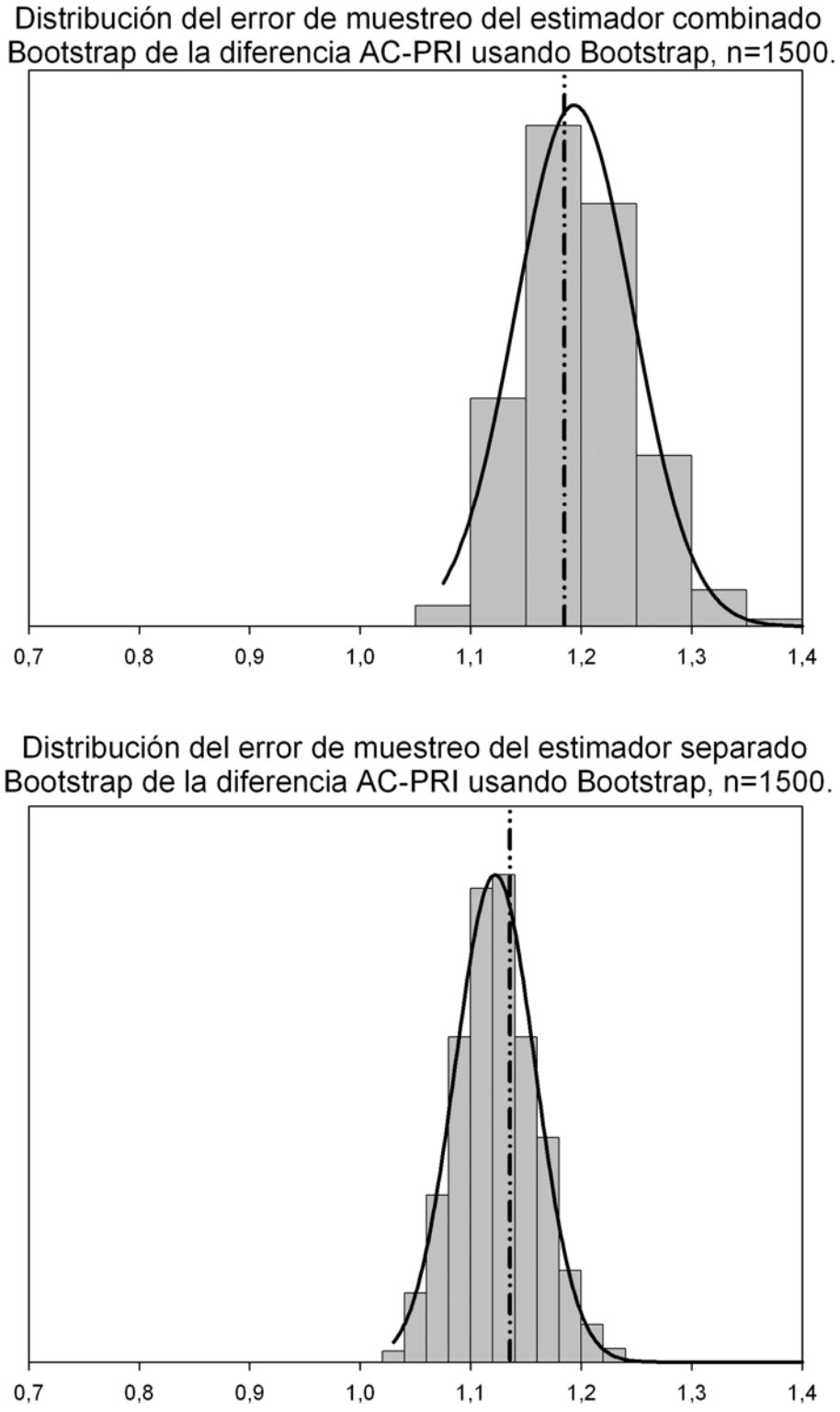


Figura 5.12: Errores de muestreo con *Bootstrap*, $100 * (1.96\sqrt{\widehat{V}_{RS}(\widehat{\mathcal{D}}_{RS})})$

5.4. Resultados

(Cochran 1967, p.165) señala que el estimador de razón separado necesita que el número de elementos en muestra de cada estrato sea grande. Särndal *et al.* (1991) han señalado que $n_h \geq 20$. En este trabajo se tomaron para n_h valores de 2, 3 y 5, lo cual generó estimadores puntuales sesgados y estimadores de varianza que están lejanos a ser los óptimos en términos de las medidas de estabilidad que se presentaron en el capítulo anterior (sesgo relativo, inestabilidad, etc.).

En las tablas 5.11, 5.15 y 5.19 se presentan los resultados correspondientes a los valores promedio de los estimadores puntuales y de varianza a lo largo de las simulaciones, variando el tamaño de muestra para las proporciones de votos a favor de los partidos AC, PRI y la proporción diferencia AC-PRI. Sobre el estimador de razón **separado** se puede observar lo siguiente.

- En promedio, los estimadores puntuales tanto los que se construyeron como cocientes de Horvitz-Thompson como los que se construyeron mediante Bootstrap con rescalamiento presentan subestimaciones fuertes con respecto a los valores poblacionales, considerando las proporciones AC y AC-PRI. Los estimadores de la proporción PRI presentan un comportamiento contrario al de los otros estimadores pues estos presentan sobre estimaciones grandes.

- Como era de esperarse por el bajo tamaño de muestra por estrato los estimadores de varianza bajo el método de Linealización presentaron subestimaciones muy fuertes.
  Los estimadores de varianza obtenidos con el método Bootstrap con rescalamiento son los que más cercanos a los poblacionales si consideramos solamente los valores promedio.

Sobre el estimador de razón **combinado** se puede observar lo siguiente.

- En promedio, los estimadores puntuales cocientes de Horvitz-Thompson de la mayoría de las proporciones presentan ligeras sobre estimaciones. Estas son un poco más grandes en los estimadores puntuales del *Bootstrap* con rescalamiento.

- Considerando a AC, los estimadores de varianza obtenidos mediante Linealización y *Bootstrap* con rescalamiento toman prácticamente los mismo valores en promedio. Ambas técnicas se acercan mucho a los valores poblacionales.

- En el caso de PRI el mejor método de estimación de varianza es el de Linealización solamente considerando sus valores promedio.

- En el caso de la diferencia AC-PRI el método que la mayoría de las veces estimó mejor en promedio fue el *Bootstrap* con rescalamiento.

En las tablas 5.8-5.10, 5.12-5.14 y 5.16-5.18 se encuentran los resultados correspondientes a las medidas de estabilidad de los estimadores de varianza. Para el estimador de razón **separado**.

- En las tablas correspondientes a la estimación de las proporciones AC y AC-PRI se observa en todos los métodos de estimación de varianza que el porcentaje inferior de error en la cola L es prácticamente cero y el porcentaje superior de error en la cola U toma valores grandes en magnitud y alejados de los esperados. Es decir, los intervalos de confianza para AC y AC-PRI quedan por debajo del valor poblacional.

- En las tablas correspondientes a la estimación de la proporciones PRI se observa en todos los métodos de estimación de varianza que el porcentaje inferior de error en la cola L es grande y el porcentaje superior de error en la cola U toma valores muy pequeños. La situación describe

un comportamiento contrario al que presentan las proporciones AC y AC-PRI. En este caso los intervalos de confianza para PRI quedan muy por encima del valor poblacional.

- Aunque en ningún momento es satisfactorio el desempeño del estimador de razón separado, se observa que en la mayoría de las tablas que en términos de sesgo relativo, inestabilidad relativa y longitud, el método de estimación de varianza que es más preciso, es el *Bootstrap* con rescalamiento. Como mencionamos en el capítulo anterior se espera que el sesgo relativo, e inestabilidad relativa deben estar muy próximas a cero y la longitud debe estar cercana a 1.

Para el estimador de razón **combinado**.

- En las tablas 5.8-5.10, 5.12-5.14 y 5.16-5.18 se observa que los métodos de estimación de varianza Linealización y *Jackknife* son equivalentes en términos de presentar valores muy cercanos en los porcentajes inferior y superior de error en las colas, además de ser próximos a los esperados.

- El método *Bootstrap* con rescalamiento tiende a tener porcentajes de error en ambas colas mayores a los niveles deseados.

- Se observa en la mayoría de las tablas, que en términos de sesgo relativo, inestabilidad relativa y longitud, el método de estimación de varianza que es más preciso, es el de Linealización.

- Si se quisiera comparar a las técnicas de remuestreo entonces debemos por un momento dejar de considerar al método de Linealización y pensar que solo se usaron *Jackknife* y *Bootstrap* con rescalamiento. Bajo este supuesto el método de estimación de varianza más preciso en términos de sesgo relativo y longitud es el *Bootstrap* con rescalamiento. Aunque si solo pensamos en inestabilidad el mejor es el *Jackknife*.

Tabla 5.8: Comparación de medidas de estabilidad de AC, $n = 600$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**AC**				
n	600				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.5	2.8	-0.5	0.139	0.995
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.5	2.8	0.4	0.140	1.000
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	4.6	1.6	0.0	0.150	0.997
$\widehat{V}_L(\widehat{\mathcal{B}})$	0.1	48.2	-33.5	0.342	0.814
$\widehat{V}_J(\widehat{\mathcal{B}})$	0.0	32.9	2.8	0.106	1.013
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	0.0	87.7	-11.7	0.151	0.9380

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L=Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.9: Comparación de medidas de estabilidad de AC, $n = 900$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**AC**				
n	900				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.5	3.0	0.0	0.102	0.999
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.4	3.0	1.6	0.106	1.006
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	6.6	1.5	0.5	0.108	1.001
$\widehat{V}_L(\widehat{\mathcal{B}})$	0.1	30.3	-23.9	0.247	0.871
$\widehat{V}_J(\widehat{\mathcal{B}})$	0.1	19.7	5.8	0.109	1.028
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	0.0	57.9	-4.9	0.091	0.974

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.10: Comparación de medidas de estabilidad de AC, $n = 1500$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**AC**				
n	1500				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.3	2.6	0.0	0.076	0.999
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.1	2.6	2.5	0.082	1.012
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	13.8	0.1	0.7	0.084	1.003
$\widehat{V}_L(\widehat{\mathcal{B}})$	0.0	20.4	-16.0	0.168	0.915
$\widehat{V}_J(\widehat{\mathcal{B}})$	0.0	15.5	5.1	0.086	1.025
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	0.0	18.2	-6.8	0.089	0.964

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.11: Comparación de promedios de 1000 estimaciones referentes a AC. Se presentan valores numéricos multiplicados por (100).

Partido	**AC**		
Parámetros y estimadores	$n = 600$	$n = 900$	$n = 1500$
R	43.434210	43.434210	43.434210
$\overline{\mathcal{R}}$	43.448782	43.444465	43.418954
$\overline{\mathcal{R}}_{RS}$	43.570538	43.631408	43.732133
$\overline{\mathcal{B}}$	42.617991	42.898851	43.101645
$\overline{\mathcal{B}}_{RS}$	41.900863	42.523119	43.093774
$V_L(\widehat{\mathcal{R}})$	0.003261	0.002163	0.001285
$\overline{V}_L(\widehat{\mathcal{R}})$	0.003244	0.002163	0.001285
$\overline{V}_J(\widehat{\mathcal{R}})$	0.003276	0.002198	0.001318
$\overline{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	0.003261	0.002174	0.001285
$V_L(\widehat{\mathcal{B}})$	0.002801	0.001858	0.001104
$\overline{V}_L(\widehat{\mathcal{B}})$	0.001862	0.001413	0.009267
$\overline{V}_J(\widehat{\mathcal{B}})$	0.002882	0.001967	0.001161
$\overline{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	0.002471	0.001767	0.001104

R =Proporción Poblacional,
$\overline{R}$ =Proporción promedio= $\sum_{i=1}^{1000} \frac{\widehat{R}_i}{1000}$,
$V_L(\widehat{R})$ =Varianza poblacional por Linealización,
$\overline{V}_{método}(\widehat{R})$ =Varianza promedio estimada. e.g.
$\overline{V}_{método}(\widehat{R}) = \sum_{i=1}^{1000} \frac{\widehat{V}_{método}(\widehat{R})_i}{1000}$.

Tabla 5.12: Comparación de medidas de estabilidad de PRI, $n = 600$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**PRI**				
n	600				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.1	2.6	-0.1	0.151	0.996
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.1	2.6	0.8	0.153	1.001
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	4.1	1.7	0.7	0.161	1.000
$\widehat{V}_L(\widehat{\mathcal{B}})$	4.6	6.10	-23.8	0.255	0.870
$\widehat{V}_J(\widehat{\mathcal{B}})$	28.4	0.0	24.8	0.285	1.115
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	94.6	0.0	6.1	0.139	1.028

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
Bias=Sesgo relativo, ver ecuación (5.1),
Instab=inestabilidad relativa, ver ecuación (5.2),
Length=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.13: Comparación de medidas de estabilidad de PRI, $n = 900$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**PRI**				
n	900				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.6	1.8	0.4	0.117	1.000
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.5	1.8	1.9	0.121	1.008
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	6.4	0.7	1.8	0.124	1.007
$\widehat{V}_L(\widehat{\mathcal{B}})$	4.2	3.9	-16.8	0.184	0.910
$\widehat{V}_J(\widehat{\mathcal{B}})$	19.6	0.10	16.6	0.200	1.078
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	89.8	0.0	12.1	0.154	1.058

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.14: Comparación de medidas de estabilidad de PRI, $n = 1500$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Partido	**PRI**				
n	1500				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{R}})$	2.8	2.7	0.2	0.080	1.000
$\widehat{V}_J(\widehat{\mathcal{R}})$	2.5	2.3	2.7	0.087	1.012
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	19.5	0.5	2.3	0.092	1.010
$\widehat{V}_L(\widehat{\mathcal{B}})$	4.6	4.3	-11.4	0.128	0.940
$\widehat{V}_J(\widehat{\mathcal{B}})$	15.2	0.4	9.9	0.127	1.047
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	87.5	0.0	3.0	0.070	1.014

$\widehat{V}_L(\widehat{\mathcal{R}})$ =Varianza estimada por Linealización de la proporción combinada,
$\widehat{V}_L(\widehat{\mathcal{B}})$ =Varianza estimada por Linealización de la proporción separada,
$\widehat{V}_J(\widehat{\mathcal{R}})$ =Varianza estimada *Jackknife* de la proporción combinada,
$\widehat{V}_J(\widehat{\mathcal{B}})$ =Varianza estimada *Jackknife* de la proporción separada,
$\widehat{V}_{RS}(\widehat{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la proporción separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.15: Comparación de promedios de 1000 estimaciones referentes a PRI. Se presentan valores numéricos multiplicados por (100).

Partido	**PRI**		
Parámetros y estimadores	$n = 600$	$n = 900$	$n = 1500$
R	36.887914	36.887914	36.887914
$\overline{\mathcal{R}}$	36.879219	36.891073	36.900007
$\overline{\mathcal{R}}_{RS}$	37.002803	37.075729	37.209090
$\overline{\mathcal{B}}$	37.626860	37.373235	37.182590
$\overline{\mathcal{B}}_{RS}$	38.574571	38.091269	37.769000
$V_L(\widehat{\mathcal{R}})$	0.002189	0.001452	0.000863
$\overline{V}_L(\widehat{\mathcal{R}})$	0.002186	0.001459	0.000865
$\overline{V}_J(\widehat{\mathcal{R}})$	0.002207	0.001481	0.000886
$\overline{V}_{RS}(\widehat{\mathcal{R}}_{RS})$	0.002205	0.001478	0.000883
$V_L(\widehat{\mathcal{B}})$	0.002070	0.001373	0.000816
$\overline{V}_L(\widehat{\mathcal{B}})$	0.001576	0.001141	0.000722
$\overline{V}_J(\widehat{\mathcal{B}})$	0.002584	0.001601	0.000897
$\overline{V}_{RS}(\widehat{\mathcal{B}}_{RS})$	0.002197	0.001540	0.000841

R =Proporción Poblacional,
$\overline{R}$ =Proporción promedio= $\sum_{i=1}^{1000} \frac{\widehat{R}_i}{1000}$,
$V_L(\widehat{R})$ =Varianza poblacional por Linealización,
$\overline{V}_{método}(\widehat{R})$ =Varianza promedio estimada. e.g.
$\overline{V}_{método}(\widehat{R}) = \sum_{i=1}^{1000} \frac{\widehat{V}_{método}(\widehat{R})_i}{1000}$.

Tabla 5.16: Comparación de medidas de estabilidad de AC-PRI, $n = 600$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Diferencia	**AC-PRI**				
n	600				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$	3.0	2.3	-0.4	0.145	0.995
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$	3.0	2.3	0.4	0.146	0.999
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$	2.9	2.1	0.6	0.156	1.000
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	51.5	-30.4	0.314	0.832
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	33.6	11.8	0.168	1.055
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$	0.0	95.3	-4.3	0.115	0.970

$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada por Linealización de la diferencia combinada,
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada por Linealización de la diferencia separada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada *Jackknife* de la diferencia combinada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada *Jackknife* de la diferencia separada,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.17: Comparación de medidas de estabilidad de AC-PRI, $n = 900$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Diferencia	**AC-PRI**				
n	900				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$	2.6	3.2	0.3	0.107	1.000
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$	2.5	3.2	1.8	0.111	1.007
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$	2.9	2.8	1.8	0.119	1.007
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	33.8	-21.7	0.227	0.883
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	22.1	10.1	0.143	1.048
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$	0.0	81.2	2.9	0.093	1.013

$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada por Linealización de la diferencia combinada,
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada por Linealización de la diferencia separada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada *Jackknife* de la diferencia combinada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada *Jackknife* de la diferencia separada,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.18: Comparación de medidas de estabilidad de AC-PRI, $n = 1500$. Los valores numéricos correspondientes a *Bias* están multiplicados por (100).

Diferencia	**AC-PRI**				
n	1500				
Estimador	L	U	$Bias$	$Instab$	$Length$
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$	2.4	2.5	-0.8	0.078	0.999
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$	2.2	2.3	2.6	0.084	1.012
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$	3.0	2.2	2.5	0.093	1.011
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	22.9	-14.7	0.156	0.923
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.0	17.6	6.9	0.101	1.033
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$	0.0	56.3	-1.9	0.066	0.989

$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada por Linealización de la diferencia combinada,
$\widehat{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada por Linealización de la diferencia separada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$ =Varianza estimada *Jackknife* de la diferencia combinada,
$\widehat{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$ =Varianza estimada *Jackknife* de la diferencia separada,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia combinada *Bootstrap*,
$\widehat{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$ =Varianza estimada *Bootstrap* de la diferencia separada *Bootstrap*,
L =Porcentaje de error inferior en la cola, ver ecuación (5.3),
U =Porcentaje de error superior en la cola, ver ecuación (5.4),
$Bias$=Sesgo relativo, ver ecuación (5.1),
$Instab$=inestabilidad relativa, ver ecuación (5.2),
$Length$=Longitud estandarizada, ver ecuación (5.5).

Tabla 5.19: Comparación de promedios de 1000 estimaciones referentes a AC-PRI. Se presentan valores numéricos multiplicados por (100).

Diferencia	**AC-PRI**		
Parámetros y estimadores	$n = 600$	$n = 900$	$n = 1500$
D	6.546295	6.546295	6.546295
$\overline{\mathcal{D}}^{\mathcal{R}}$	6.569563	6.553392	6.518887
$\overline{\mathcal{D}}^{\mathcal{R}}_{RS}$	6.597354	6.596747	6.595460
$\overline{\mathcal{D}}^{\mathcal{B}}$	4.991130	5.525616	5.919052
$\overline{\mathcal{D}}^{\mathcal{B}}_{RS}$	3.308828	4.421455	5.333410
$V_L(\widehat{\mathcal{D}}^{\mathcal{R}})$	0.009272	0.006151	0.003654
$\overline{V}_L(\widehat{\mathcal{D}}^{\mathcal{R}})$	0.005430	0.003624	0.002150
$\overline{V}_J(\widehat{\mathcal{D}}^{\mathcal{R}})$	0.009317	0.006265	0.003750
$\overline{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{R}}_{RS})$	0.009334	0.006263	0.003746
$V_L(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.008513	0.005648	0.003355
$\overline{V}_L(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.003438	0.002555	0.001649
$\overline{V}_J(\widehat{\mathcal{D}}^{\mathcal{B}})$	0.009518	0.006223	0.003589
$\overline{V}_{RS}(\widehat{\mathcal{D}}^{\mathcal{B}}_{RS})$	0.008144	0.005814	0.003290

D =diferencia Poblacional,

$\overline{\mathcal{D}}$ =diferencia promedio= $\sum_{i=1}^{1000} \frac{\widehat{\mathcal{D}}_i}{1000}$,

$V_L(\widehat{\mathcal{D}})$ =Varianza poblacional por Linealización,

$\overline{V}_{método}(\widehat{\mathcal{D}})$ =Varianza promedio estimada. e.g.

$\overline{V}_{método}(\widehat{\mathcal{D}}) = \sum_{i=1}^{1000} \frac{\widehat{V}_{método}(\widehat{\mathcal{D}})_i}{1000}$.

Capítulo 6

Comentarios y conclusiones

En este trabajo se comparó la estabilidad de estimadores de razones y de diferencias de razones en términos de los valores de varianza estimada de los estimadores y de los intervalos de confianza que se producen. Usando: los estimadores de razón combinado y separado, así como tres métodos de estimación de su varianza: aproximación por series de Taylor, *Jackknife* y *Bootstrap*. Se ilustró con un ejercicio en el contexto de los conteos rápidos, haciendo una aplicación con los datos del PREP[1] de la elección presidencial del año 2000. A continuación se exponen las conclusiones que se derivan de este trabajo.

- No hay un método de estimación de varianza que produzca estimaciones más precisas en el sentido de cumplir todos los criterios de comparación. Depende del estimador de proporción que se quiera usar.

- Considerando al estimador de razón separado para la estimación de proporciones y diferencias de proporciones, los intervalos de confianza producidos con el método *Bootstrap* con rescalamiento son los que satisfacen una mayor cantidad de criterios de comparación, pues en términos

[1] ver anexo E.

de sesgo relativo, inestabilidad relativa y longitud de los intervalos son los más precisos.

- Al usar el estimador de razón combinado para estimar proporciones y diferencias de proporciones, los intervalos de confianza producidos por Linealización o *Jackknife* son igualmente efectivos si solo pensamos en términos de los porcentajes de error en las colas L y U. Si se pretende comparar los intervalos en términos del sesgo, la inestabilidad relativa y la longitud; el mejor es el método de Linealización. Si solo nos restringimos a comparar únicamente las técnicas de remuestreo, el método *Jackknife* es el que presenta mejor estabilidad, pero si se quieren obtener buena longitud y menor sesgo relativo el mejor es el método *Bootstrap* con rescalamiento.

- Utilizando los criterios de comparación se observa que el estimador de razón combinado tiene mejor desempeño para estimar puntualmente proporciones y diferencias de proporciones así como sus varianzas, que el estimador separado aplicados a esta población y con estos tamaños de muestra.

- Al usar estimadores de razón separada es necesario tener un número grande de unidades en cada estrato, si no, los estimadores puntuales y de varianza subestiman o sobrestiman a los valores poblacionales. En este trabajo 5 unidades por estrato no fue suficiente.

- Aún en el caso más simple, aplicar el método *Bootstrap* en poblaciones estratificadas y con diseño aleatorio simple, no hay una guía teórica de la elección para la cantidad de remuestras B, ni para el tamaño de ellas m_h, $h = 1, \ldots H$. Haciendo pruebas antes de efectuar las simulaciones, observé que los valores de las estimaciones son muy sensibles a la elección de m_h y relativamente poco sensibles a la elección de B.

- Finalmente queda la interrogante de cual método se debe elegir para construir intervalos de confianza que regularmente estimen bien lo que se desea. En vista de los resultados obtenidos no es posible determinar un mejor método. La elección depende de cuáles medidas de estabilidad queremos que tengan buenos valores y por supuesto del estimador.
  A continuación se intenta dar una guía para tomar una decisión en la elección de un método.

  Primeramente, si el estimador es una función de promedios como es el caso del estimador de razón combinado usar Linealización es la opción que satisface más medidas de estabilidad en valores óptimos. En segundo lugar se recomienda usar el método *Bootstrap* con rescalamiento pues con este método se tienen valores pequeños de sesgo relativo y la longitud de los intervalos es en promedio la mejor. En tercer lugar se recomienda usar *Jackknife* pues la estabilidad de los intervalos es buena.

  Si el estimador no es una función de promedios como es el caso del estimador de razón separado se debe cuidar que el tamaño de muestra sea relativamente grande y la mejor opción es usar el *Bootstrap* con rescalamiento.

- Este trabajo tuvo como finalidad ilustrar el uso de diversos métodos de estimación de varianza, y su implementación en la computadora. En el ejercicio profesional ya no lo haría programando los métodos, porque ya existe mucha paquetería en la que ya están programadas la mayoría de estas rutinas.

Apéndice A

Demostración de la fórmula de covarianzas presentada por Lethonen

Consideremos dos estimadores de razón

$$\widehat{R}_i = \frac{\widehat{t}_{y_i}}{\widehat{t}_{z_i}},\ \widehat{R}_j = \frac{\widehat{t}_{y_j}}{\widehat{t}_{z_j}}.$$

Después expresamos ambos estimadores mediante su serie de Taylor de primer orden .

$$\widehat{R}_i \doteq R_i + \frac{1}{t_{z_i}}(\widehat{t}_{y_i} - R_i\widehat{t}_{z_i}),\ \ \widehat{R}_j \doteq R_j + \frac{1}{t_{z_j}}(\widehat{t}_{y_j} - R_j\widehat{t}_{z_j}).$$

Para continuar se calcula el valor del término de covarianza que existe entre ambos estimadores, haciendo el desarrollo mediante las propiedades de la covarianza.

$$C\left(\widehat{R}_i, \widehat{R}_j\right) \doteq C\left(R_i + \frac{1}{t_{z_i}}(\widehat{t}_{y_i} - R_i\widehat{t}_{z_i}), R_j + \frac{1}{t_{z_j}}(\widehat{t}_{y_j} - R_j\widehat{t}_{z_j})\right) =$$

$$C\left(\widehat{R}_i, \widehat{R}_j\right) = C\left(R_i, R_j\right) + C\left(R_i, \frac{1}{t_{z_j}}(\widehat{t}_{y_j} - R_j\widehat{t}_{z_j})\right) +$$

$$C\left(\frac{1}{t_{z_i}}(\widehat{t}_{y_i} - R_i\widehat{t}_{z_i}, R_j)\right) + C\left(\frac{1}{t_{z_i}}(\widehat{t}_{y_i} - R_i\widehat{t}_{z_i}), \frac{1}{t_{z_j}}(\widehat{t}_{y_j} - R_j\widehat{t}_{z_j})\right) =$$

$$C\left(\frac{1}{t_{z_i}}(\widehat{t}_{y_i} - R_i\widehat{t}_{z_i}), \frac{1}{t_{z_j}}(\widehat{t}_{y_j} - R_j\widehat{t}_{z_j})\right) = \ldots$$

Reagrupando términos obtenemos la expresión.

$$\ldots = \frac{1}{t_{z_i}t_{z_j}}\left\{C\left(\widehat{t}_{y_i}, \widehat{t}_{y_j}\right) + R_iR_jC\left(\widehat{t}_{z_i}, \widehat{t}_{z_j}\right) - R_jC\left(\widehat{t}_{y_i}, \widehat{t}_{z_j}\right) - R_iC\left(\widehat{t}_{y_j}, \widehat{t}_{z_i}\right)\right\}.$$

Apéndice B

Varianza entre estimadores de razón combinada

Para hacer dicha demostración es necesario el siguiente lema.

B.1. Lema B

Considerando un diseño estratificado aleatorio simple tenemos que.

$$C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{yzU_h}.$$

Demostración del lema.
Hansen M.H.*et al.* (1953, pag:190) referencia [21] afirma en la ecuación (4.4) del libro, lo siguiente.

$$\rho_{\bar{y}_U \bar{z}_U} = \frac{1}{\sigma_{\bar{y}_U} \sigma_{\bar{z}_U}} \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \sum_{y_{h_k}, z_{h_k} \in U_h} \frac{\left(y_{h_k} - \bar{y}_{U_h}\right)\left(z_{h_k} - \bar{z}_{U_h}\right)}{N_h - 1} \tag{B.1}$$

Es la expresión para calcular el coeficiente de correlación lineal bajo un diseño estratificado aleatorio simple.
Por otra parte es un hecho conocido que.

$$C\left(\bar{y}_U, \bar{z}_U\right) = \rho_{\bar{y}_U \bar{z}_U} \sigma_{\bar{y}_U} \sigma_{\bar{z}_U} \tag{B.2}$$

se deduce que si sustituyo (B.1) en (B.2) obtenemos lo siguiente.

$$C\left(\bar{y}_U, \bar{z}_U\right) = \frac{1}{N^2} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{yzU_h}. \tag{B.3}$$

Que nos da una expresión para calcular la covarianza entre estimadores de medias.
Por otra parte la covarianza entre medias se puede reescribir en términos de totales de la siguiente forma.

$$C\left(\bar{y}_U, \bar{z}_U\right) = C\left(\frac{\widehat{t}_{y\pi}}{N}, \frac{\widehat{t}_{z\pi}}{N}\right) = \frac{1}{N^2} C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right). \tag{B.4}$$

De la ecuación (B.4) obtenemos lo siguiente.

$$C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right) = N^2 C\left(\bar{y}_U, \bar{z}_U\right). \tag{B.5}$$

Sustituimos (B.3) en (B.5) con lo que obtenemos.

$$C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{yzU_h}.$$

B.2. Demostración de la fórmula de covarianza para una Razón Combinada en un diseño estratificado aleatorio simple

Sustituimos la expresión del lema B anterior en la fórmula del apéndice A.

$$C\left(\widehat{\mathcal{R}}_i, \widehat{\mathcal{R}}_j\right) = \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{y_i y_j U_h} +$$

$$R_i R_j \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{z_i z_j U_h} -$$

$$R_j \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{y_i z_j U_h} - R_i \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} S_{y_j z_i U_h} = \ldots$$

Reagrupando los términos obtenemos lo siguiente.

$$\ldots = \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} \left[S_{y_i y_j U_h} + R_i R_j S_{z_i z_j U_h} - R_j S_{y_i z_j U_h} - R_i S_{y_j z_i U_h}\right].$$

Apéndice C

Covarianza entre estimadores de razón separada

Para hacer dicha demostración es necesario el siguiente lema.

C.1. Lema C

Consideremos una muestra aleatoria simple tenemos que.

$$C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}S_{yzU_h}.$$

Demostración del lema.

Cochran (1977, p.25) demuestra lo siguiente.

$$C(\bar{y}_U, \bar{z}_U) = \frac{N-n}{nN}\frac{1}{N-1}\sum_{y_i, z_i \in U}\left(y_i - \bar{y}_U\right)\left(z_i - \bar{z}_U\right).$$

Haciendo un poco de álgebra esta expresión se puede reescribir como sigue.

$$C(\bar{y}_U, \bar{z}_U) = \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{y_i, z_i \in U}\frac{\left(y_i - \bar{y}_U\right)\left(z_i - \bar{z}_U\right)}{N-1}. \tag{C.1}$$

Se sigue de sustituir en (B.5) la ecuación (C.1) el resultado.

$$C\left(\widehat{t}_{y\pi}, \widehat{t}_{z\pi}\right) = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n}S_{yzU}.$$

C.2. Demostración de la fórmula de covarianza para una razón separada en un diseño estratificado aleatorio simple

Primero sustituimos la fórmula del estimador de razón separada en la expresión de covarianza entre estimadores.

$$C\left(\widehat{\mathcal{B}}_i, \widehat{\mathcal{B}}_j\right) = C\left(\sum_{h=1}^{H} \frac{t_{z_i U_h}}{t_{z_i}} \widehat{\mathcal{R}}_{s_{hi}}, \sum_{k=1}^{H} \frac{t_{z_j U_k}}{t_{z_j}} \widehat{\mathcal{R}}_{s_{kj}}\right) = \ldots$$

A continuación desarrollamos la fórmula anterior usando las propiedades de la covarianza.

$$\ldots = \sum_{h=1}^{H}\sum_{k=1}^{H} \frac{t_{z_i U_h} t_{z_j U_k}}{t_{z_i} t_{z_j}} C\left(\widehat{\mathcal{R}}_{s_{hi}}, \widehat{\mathcal{R}}_{s_{kj}}\right) = \sum_{h=1}^{H} \frac{t_{z_i U_h} t_{z_j U_h}}{t_{z_i} t_{z_j}} C\left(\widehat{\mathcal{R}}_{s_{hi}}, \widehat{\mathcal{R}}_{s_{hj}}\right) = \ldots$$

En vista de que la fórmula está en términos de estimadores de razón no estratificado podemos aplicar el apéndice A.

$$\ldots = \sum_{h=1}^{H} \frac{t_{z_i U_h} t_{z_j U_h}}{t_{z_i} t_{z_j}} \frac{1}{t_{z_i U_h} t_{z_j U_h}} [C\left(\widehat{t}_{y_i \pi h}, \widehat{t}_{y_j \pi h}\right) + R_{hi} R_{hj} C\left(\widehat{t}_{z_i \pi h}, \widehat{t}_{z_j \pi h}\right) -$$

$$R_{hj} C\left(\widehat{t}_{y_i \pi h}, \widehat{t}_{z_j \pi h}\right) - R_{hi} C\left(\widehat{t}_{y_j \pi h}, \widehat{t}_{z_i \pi h}\right)] = \ldots$$

Para terminar aplicamos el lema C a cada estrato.

$$\ldots = \frac{1}{t_{z_i} t_{z_j}} \sum_{h=1}^{H} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \left[S_{y_i y_j U_h} + R_{hi} R_{hj} S_{z_i z_j U_h} - R_{hj} S_{y_i z_j U_h} - R_{hi} S_{y_j z_i U_h}\right].$$

Apéndice D

PREP 2000

D.1. Programa de Resultados Electorales Preliminares (PREP)

El programa de Resultados Electorales Preliminares es un proceso para la captura y publicación en línea de las actas de escrutinio que contienen la cantidad de votos finales de cada una de las casillas y mesas receptoras instaladas por los organismos electorales el día de la elección.
El objetivo primordial de este programa es generar transparencia de la elección difundiendo de manera fehaciente y oportuna los resultados de la votación de los diferentes tipos de elección que estén en juego, garantizándole así al elector, la seguridad de que su voto ha sido contabilizado y registrado con transparencia y confiabilidad.

D.1.1. Ventajas y Desventajas del PREP

El PREP no es un cálculo de los resultados sobre la base de estimaciones estadísticas a partir de una muestra, los resultados que arroja el PREP son los resultados actuales de la votación, tal y como son asentados en la totalidad

de las actas de escrutinio y cómputo que los funcionarios electorales de cada una de las casillas elaboran al finalizar el proceso de votación.

Sus principales desventajas radican en que al ser un mecanismo de transparencia que busca recabar la totalidad de las actas de escrutinio, el tiempo que toma en terminar de transmitir y capturar la información es demasiado largo (de 6 a 12 horas en promedio después de haber cerrado oficialmente las casillas). Aunando a esto, por lo general el PREP nunca logra computar el 100 % de las actas, debido a diversas razones como lo son problemas con el procedimiento de cerrado y marcado de los paquetes electorales por parte de los funcionarios de casilla, impugnaciones por parte de los representantes de los partidos y/o candidatos o por la logística o complicación geográfica que tenga la sección electoral o mesa receptora para hacer llegar la información de la votación al centro de acopio.Es así, el PREP puede caer en un riesgo importante: el intercambio de ganador durante el proceso de captura de la información sobre todo en contiendas muy cerradas donde la diferencia es de quizás 2 o 3 puntos porcentuales. En México este fenómeno se ha presentado en diversos estados.

Apéndice E

Algunos programas

A continuación presentamos algunos de los programas utilizados al elaborar este trabajo. Descripción breve de los programas.

Programa 1: Obtiene las muestras Estratificadas aleatorias simples. Escrito en el paquete Visual Fox Pro 8.0.
Programa 2: Obtiene la varianza de los estimadores puntuales de razón de tipo combinado y separado mediante Linealización. Elaborado en el paquete SPSS 15.
Programa 3: Obtiene los estimadores puntuales y de varianza de los estimadores combinados y separados mediante Linealización. Elaborado en el paquete SPSS 15.
Programa 4: Obtiene los estimadores puntuales y de varianza de los estimadores combinados mediante *Jackknife*. Elaborado en Lenguaje R-2.5.0.
Programa 5: Obtiene los estimadores puntuales y de varianza de los estimadores separados mediante *Jackknife*. Elaborado en Lenguaje R-2.5.0.
Programa 6: Obtiene los estimadores puntuales y de varianza de los estimadores combinados y separados mediante *Bootstrap*. Elaborado en Lenguaje R-2.5.0.

E.1. Programa 1: Selector de muestras

```
** Programa para obtener muestras STSI
** De cada estrato se muestrean 2, 3, o 5 elementos.
** Rutina crear variables auxiliares N_acum, azar ,orden
** y prepara la tabla para generar muestras
&& Abre la tabla de datos
USE POB_1 INDEX ON dina TO indicebasura unique
COPY TO Pob_1_porborrar FIELDS dINA, id
ALTER TABLE Pob_1_porborrar RENAME COLUMN id TO N_ACUM
UPDATE Pob_1_porborrar SET n_acum=n_acum-1
CLOSE TABLES DELETE FILE indicebasura.IDX
&& Genera la tabla de donde se obtendrán las muestras
SELECT a.*, b.n_acum, 0.0000001 as azar, 99999 as orden,
9999.99999 as w;
FROM pob_1 a LEFT OUTER JOIN pob_1_porborrar b ON a.dina=b.dina ;
INTO CURSOR Cur_pob_1 READWRITE
ALTER TABLE Cur_pob_1 ALTER COLUMN id N(6,0);
&& Aumenta el ancho del campo id  x_hen una unidad
ALTER COLUMN X_h N(7,0)
** Proceso recursivo para obtener las muestras
&& Controla la cantidad de registros seleccionados por estrato
FOR j=2 TO 5 IF j<>4
&& Controla la cantidad de muestras aleatorias
FOR i=1 TO 1000
&& Nombre del archivo txt con la muestra
*arch_salida='S'+STR(j,1,0)+'_'+PADL(ALLTRIM(STR(i,4,0)),4,'0')+'.txt'
```

```
&& Nombre del archivo txt con la muestra
arch_salida='S'+STR(j,1,0)+'_'+ALLTRIM(STR(i,4,0))+'.txt'
&& Se asigna un número aleatorio a cada registro
UPDATE Cur_pob_1 SET azar=RAND()
&& Se ordenan los registros por estrato
&& y después por el número aleatorio
SELECT * from cur_pob_1 ORDER BY Dina,azar INTO
CURSOR Cur_pob_1_temporal READWRITE
&& Al interior de cada estrato se numeran los registros,
&& empezando siempre del 1
UPDATE Cur_pob_1_temporal SET ORDEN=RECNO()-N_acum, w=N_h/j
&& Se genera un archivo solamente con los j primeros
&& registros de cda estrato
SELECT * FROM Cur_pob_1_temporal WHERE orden<=j ORDER BY Dina,
id INTO CURSOR Cur_pob_1_temporal2
COPY TO borrar.txt FIELDS EXCEPT N_acum,Azar,Orden TYPE SDF
&& Pega el título de las columnas
RUN copy label.txt+borrar.txt &arch_salida
&& Borrar el archivo temporal de los resultados
DELETE FILE borrar.txt ENDFOR ENDIF ENDFOR
CLOSE TABLES DELETE FILE Pob_1_porborrar.DBF
```

E.2. Programa 2: Varianza de estimadores por Linealización

```
*Calcula la varianza por Linealización. SET PRINTBACK=OFF.
*Población solo con PRI PAN y DIFERENCIA.
GET FILE='C:\GUS\tesis\Data\Pob_1.sav'.
```

```
*totales poblacionales. COMPUTE uno = 1 . EXECUTE.
AGGREGATE /OUTFILE=*  MODE=ADDVARIABLES  /BREAK=uno
  /tot_pan=sum(pan)/tot_pri = sum(pri)/tot_val=sum(validos).
*total poblacionales por estrato.
AGGREGATE  /OUTFILE=*  MODE=ADDVARIABLES  /BREAK=dina
  /tot_pan_h=sum(pan)/tot_pri_h = sum(pri)/tot_val_h=sum(validos).
*Razones poblacionales.
COMPUTE R_pan=tot_pan/tot_val. COMPUTE R_pri=tot_pri/tot_val.
*Razones poblacionales por estrato.
COMPUTE R_pan_h=tot_pan_h/tot_val_h.
COMPUTE R_pri_h=tot_pri_h/tot_val_h.
*Tamaños de muestra por estrato.
COMPUTE m_h_2=2.COMPUTE m_h_3=3.COMPUTE m_h_5=5.
*Constantes.
COMPUTE cons_h_2=N_h*N_h*(1/m_h_2-1/N_h).
COMPUTE cons_h_3=N_h*N_h*(1/m_h_3-1/N_h).
COMPUTE cons_h_5=N_h*N_h*(1/m_h_5-1/N_h).EXECUTE .
*Promedios por estrato. AGGREGATE
  /OUTFILE=*  MODE=ADDVARIABLES  /BREAK=dina
  /pan_mean = MEAN(pan)  /pri_mean = MEAN(pri)
  /val_mean = MEAN(validos). *Valores S.
COMPUTE sum_Sy_pan=(pan-pan_mean)*(pan-pan_mean)/(N_h-1).
COMPUTE sum_Sy_pri=(pri-pri_mean)*(pri-pri_mean)/(N_h-1).
COMPUTE sum_Sxy_pan=(pan-pan_mean)*(validos-val_mean)/(N_h-1).
COMPUTE sum_Sxy_pri=(pri-pri_mean)*(validos-val_mean)/(N_h-1).
COMPUTE sum_Sx=(validos-val_mean)*(validos-val_mean)/(N_h-1).
COMPUTE sum_Syy=(pan-pan_mean)*(pri-pri_mean)/(N_h-1). EXE.
AGGREGATE /OUTFILE='C:\GUS\tesis\temp\auxiliar_U.sav'
/BREAK=dina /t=MEAN(tot_val)
```

```
/cons_h_2 = MEAN(cons_h_2) /cons_h_3 = MEAN(cons_h_3)
/cons_h_5 = MEAN(cons_h_5) /R_pan = MEAN(R_pan)
/R_pri = MEAN(R_pri) /R_pan_h = MEAN(R_pan_h)
/R_pri_h = MEAN(R_pri_h) /Sy_pan = SUM(sum_Sy_pan)
/Sy_pri = SUM(sum_Sy_pri) /Sxy_pan = SUM(sum_Sxy_pan)
/Sxy_pri = SUM(sum_Sxy_pri) /Sx = SUM(sum_Sx)
/Syy = SUM(sum_Syy).
GET FILE='C:\GUS\tesis\temp\auxiliar_U.sav'.
COMPUTE uno=1. COMPUTE pan_2c=(cons_h_2/(t*t))*
(Sy_pan+R_pan*R_pan*Sx-2*R_pan*Sxy_pan).
COMPUTE pri_2c=(cons_h_2/(t*t))*
(Sy_pri+R_pri*R_pri*Sx-2*R_pri*Sxy_pri).
COMPUTE pan_3c=(cons_h_3/(t*t))*
(Sy_pan+R_pan*R_pan*Sx-2*R_pan*Sxy_pan).
COMPUTE pri_3c=(cons_h_3/(t*t))*
(Sy_pri+R_pri*R_pri*Sx-2*R_pri*Sxy_pri).
COMPUTE pan_5c=(cons_h_5/(t*t))*
(Sy_pan+R_pan*R_pan*Sx-2*R_pan*Sxy_pan).
COMPUTE pri_5c=(cons_h_5/(t*t))*
(Sy_pri+R_pri*R_pri*Sx-2*R_pri*Sxy_pri).
COMPUTE pan_2s=(cons_h_2/(t*t))*
(Sy_pan+R_pan_h*R_pan_h*Sx-2*R_pan_h*Sxy_pan).
COMPUTE pri_2s=(cons_h_2/(t*t))*
(Sy_pri+R_pri_h*R_pri_h*Sx-2*R_pri_h*Sxy_pri).
COMPUTE pan_3s=(cons_h_3/(t*t))*
(Sy_pan+R_pan_h*R_pan_h*Sx-2*R_pan_h*Sxy_pan).
COMPUTE pri_3s=(cons_h_3/(t*t))*
(Sy_pri+R_pri_h*R_pri_h*Sx-2*R_pri_h*Sxy_pri).
COMPUTE pan_5s=(cons_h_5/(t*t))*
```

```
(Sy_pan+R_pan_h*R_pan_h*Sx-2*R_pan_h*Sxy_pan).
COMPUTE pri_5s=(cons_h_5/(t*t))*
(Sy_pri+R_pri_h*R_pri_h*Sx-2*R_pri_h*Sxy_pri).
COMPUTE C_2c=(cons_h_2/(t*t))*
(Syy+R_pan*R_pri*Sx-R_pri*Sxy_pan-R_pan*Sxy_pri).
COMPUTE C_3c=(cons_h_3/(t*t))*
(Syy+R_pan*R_pri*Sx-R_pri*Sxy_pan-R_pan*Sxy_pri).
COMPUTE C_5c=(cons_h_5/(t*t))*
(Syy+R_pan*R_pri*Sx-R_pri*Sxy_pan-R_pan*Sxy_pri).
COMPUTE C_2s=(cons_h_2/(t*t))*
(Syy+R_pan_h*R_pri_h*Sx-R_pri_h*Sxy_pan-R_pan_h*Sxy_pri).
COMPUTE C_3s=(cons_h_3/(t*t))*
(Syy+R_pan_h*R_pri_h*Sx-R_pri_h*Sxy_pan-R_pan_h*Sxy_pri).
COMPUTE C_5s=(cons_h_5/(t*t))*
(Syy+R_pan_h*R_pri_h*Sx-R_pri_h*Sxy_pan-R_pan_h*Sxy_pri). exe.
AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES /BREAK=uno /vpan_2c = SUM(pan_2c)
/vpri_2c = SUM(pri_2c)
/vpan_3c = SUM(pan_3c) /vpri_3c = SUM(pri_3c)
/vpan_5c = SUM(pan_5c) /vpri_5c = SUM(pri_5c)
/vpan_2s = SUM(pan_2s) /vpri_2s = SUM(pri_2s)
/vpan_3s = SUM(pan_3s) /vpri_3s = SUM(pri_3s)
/vpan_5s = SUM(pan_5s) /vpri_5s = SUM(pri_5s)
/co_2c = SUM(C_2c)  /co_3c = SUM(C_3c)
/co_5c = SUM(C_5c)  /co_2s = SUM(C_2s)
/co_3s = SUM(C_3s)  /co_5s = SUM(C_5s).
format vpan_2c to co_5s (f10.9).
*varianza de las diferencias.
compute vdif2_c=vpan_2c+vpri_2c-2*co_2c.
```

```
compute vdif3_c=vpan_3c+vpri_3c-2*co_3c.
compute vdif5_c=vpan_5c+vpri_5c-2*co_5c.
compute vdif2_s=vpan_2s+vpri_2s-2*co_2s.
compute vdif3_s=vpan_3s+vpri_3s-2*co_3s.
compute vdif5_s=vpan_5s+vpri_5s-2*co_5s. exe.
AGGREGATE /OUTFILE='C:\GUS\tesis\Data\var_U.sav'
/BREAK=uno
/vpan_2c =MEAN(vpan_2c)  /vpri_2c =MEAN(vpri_2c)
/vpan_3c =MEAN(vpan_3c)  /vpri_3c =MEAN(vpri_3c)
/vpan_5c =MEAN(vpan_5c)  /vpri_5c =MEAN(vpri_5c)
/vpan_2s =MEAN(vpan_2s)  /vpri_2s =MEAN(vpri_2s)
/vpan_3s =MEAN(vpan_3s)  /vpri_3s =MEAN(vpri_3s)
/vpan_5s =MEAN(vpan_5s)  /vpri_5s =MEAN(vpri_5s)
/vdif2_c =MEAN(vdif2_c)  /vdif3_c =MEAN(vdif3_c)
/vdif5_c =MEAN(vdif5_c)  /vdif2_s =MEAN(vdif2_s)
/vdif3_s =MEAN(vdif3_s)  /vdif5_s =MEAN(vdif5_s).
GET FILE='C:\GUS\tesis\Data\var_U.sav'.
FLIP VARIABLES= vpan_2c  vpri_2c  vpan_3c
  vpri_3c  vpan_5c  vpri_5c  vpan_2s  vpri_2s
  vpan_3s  vpri_3s  vpan_5s  vpri_5s  vdif2_c
  vdif3_c  vdif5_c  vdif2_s  vdif3_s  vdif5_s.
rename variables (var001=varianza).
compute se=sqrt(varianza).
formats varianza se (F10.9).execute.
sort cases by case_lbl (a).
SAVE TRANSLATE OUTFILE='C:\GUS\tesis\Data\var_U.dat'
/TYPE=TAB /MAP /REPLACE /FIELDNAMES
/CELLS=VALUES.SET PRINTBACK=ON.
```

E.3. Programa 3: Estimación de varianza con Linealización

```
*Estima la varianza por Linealización.
SET PRINTBACK=OFF.DEFINE TL (ssize=!charend('/')/
muestra=!charend('/')) .
*Conversión de muestras a SPSS.
GET DATA  /TYPE = TXT
/FILE=
!quote( !concat('C:\GUS\tesis\muestras\S',!ssize,'_',!muestra,'.TXT'))
/DELCASE = LINE
/DELIMITERS = " " /ARRANGEMENT = DELIMITED /FIRSTCASE = 2
/IMPORTCASE = ALL /VARIABLES = dina F3.0
id F5.0 x_h F6.0 N_h F3.0 validos F4.0
pan F4.0 pri F4.0 dif F5.0 .
CACHE.EXECUTE.
*Filtración de los casos validos.
FILTER OFF.USE ALL.
SELECT IF(NOT(dina =0)).EXECUTE .
*Pegado de faltante para el estimador separado.
MATCH FILES /FILE=*
 /TABLE='C:\GUS\tesis\Data\X_U_h.sav'
 /BY dina.EXECUTE.
compute w=N_h/!ssize.EXECUTE.
compute w_pan=w*pan.compute w_pri=w*pri.
compute w_val=w*validos.
COMPUTE uno = 1 .EXECUTE.
*Estimadores de totales.
AGGREGATE  /OUTFILE=*  MODE=ADDVARIABLES
```

```
  /BREAK=uno
  /t_pan = sum(w_pan)  /t_pri = sum(w_pri)
  /t_val = sum(w_val).
*Estimadores de totales por estrato.
AGGREGATE  /OUTFILE=*  MODE=ADDVARIABLES
  /BREAK=dina
  /t_pan_h = sum(w_pan)  /t_pri_h = sum(w_pri)
  /t_val_h = sum(w_val) /x_mean_s_h=mean(validos).
*tamaño de muestra por estrato.
COMPUTE m_h = !ssize.
*Estimador combinado.
COMPUTE R_pan=t_pan/t_val.
COMPUTE R_pri=t_pri/t_val.
*Estimador separado.
COMPUTE R_pan_h=t_pan_h/t_val_h.
COMPUTE R_pri_h=t_pri_h/t_val_h.
*Constantes por estrato.
COMPUTE cons_h_c=N_h*N_h*(1/m_h-1/N_h).
COMPUTE cons_h_s=
((X_mean_h*X_mean_h)/(x_mean_s_h*x_mean_s_h))*N_h*N_h*(1/m_h-1/N_h).
EXECUTE .
*promedios por estrato. AGGREGATE
/OUTFILE=*  MODE=ADDVARIABLES
/BREAK=dina  /pan_mean = MEAN(pan)
/pri_mean = MEAN(pri) /val_mean = MEAN(validos). *Valores S.
COMPUTE sum_Sy_pan=(pan-pan_mean)*(pan-pan_mean)/(m_h-1).
COMPUTE sum_Sy_pri=(pri-pri_mean)*(pri-pri_mean)/(m_h-1).
COMPUTE sum_Sx =(validos-val_mean)*(validos-val_mean)/(m_h-1).
COMPUTE sum_Sxy_pan=(pan-pan_mean)*(validos-val_mean)/(m_h-1).
```

```
COMPUTE sum_Sxy_pri=(pri-pri_mean)*(validos-val_mean)/(m_h-1).
EXECUTE. AGGREGATE
  /OUTFILE='C:\GUS\tesis\temp\auxiliar_S_A.sav'
  /BREAK=dina  /t=MEAN(t_val)
  /x_h=MEAN(x_h)  /cons_h_c = MEAN(cons_h_c)
  /cons_h_s = MEAN(cons_h_s)
  /R_pan = MEAN(R_pan)  /R_pri = MEAN(R_pri)
  /R_pan_h = MEAN(R_pan_h)  /R_pri_h = MEAN(R_pri_h)
  /Sy_pan = SUM(sum_Sy_pan)  /Sy_pri = SUM(sum_Sy_pri)
  /Sx = SUM(sum_Sx)  /Sxy_pan = SUM(sum_Sxy_pan)
  /Sxy_pri = SUM(sum_Sxy_pri).
GET FILE='C:\GUS\tesis\temp\auxiliar_S_A.sav'.
COMPUTE uno=1. COMPUTE ppanc=(cons_h_c/(t*t))*
(Sy_pan+R_pan*R_pan*Sx-2*R_pan*Sxy_pan).
COMPUTE ppric=(cons_h_c/(t*t))*
(Sy_pri+R_pri*R_pri*Sx-2*R_pri*Sxy_pri).
COMPUTE ppans=(cons_h_s/(t*t))*
(Sy_pan+R_pan_h*R_pan_h*Sx-2*R_pan_h*Sxy_pan).
COMPUTE ppris=(cons_h_s/(t*t))*
(Sy_pri+R_pri_h*R_pri_h*Sx-2*R_pri_h*Sxy_pri). exe.
AGGREGATE
/OUTFILE=* MODE=ADDVARIABLES /BREAK=uno
/xsum=SUM(x_h) /R_pan_2= MEAN(R_pan) /R_pri_2= MEAN(R_pri)
/vpanc = SUM(ppanc)  /vpric = SUM(ppric)
/vpans = SUM(ppans)  /vpris = SUM(ppris).
compute frac=x_h/xsum. compute pansep=frac*R_pan_h.
compute prisep=frac*R_pri_h.
execute. format vpanc to prisep (f10.9).
AGGREGATE
```

```
/OUTFILE='C:\GUS\tesis\temp\auxiliar_S_B.sav'
/BREAK=uno  /Rc_pan=MEAN(R_pan_2)
/vpanc = MEAN(vpanc) /Rc_pri=MEAN(R_pri_2)
/vpric = MEAN(vpric) /Rs_pan=SUM(pansep)
/vpans = MEAN(vpans) /Rs_pri=SUM(prisep)
/vpris = MEAN(vpris).
GET FILE='C:\GUS\tesis\temp\auxiliar_S_B.sav'.
COMPUTE sepanc=sqrt(vpanc).
COMPUTE sepric=sqrt(vpric).
COMPUTE sepans=sqrt(vpans).
COMPUTE sepris=sqrt(vpris). EXECUTE.
Delete variables vpanc vpric vpans vpris.
format Rc_pan to sepris (f10.9).
SAVE OUTFILE=
!quote(!concat('C:\GUS\tesis\temp\SE_S_',!ssize,'_',!muestra,'.sav'))
  /keep=Rc_pan sepanc Rc_pri sepric Rs_pan sepans Rs_pri sepris
   /COMPRESSED.
!ENDDEFINE.
*****.
DEFINE repetidor (ancial= !TOKENS(1)
/eneh = !TOKENS(1) /rept = !TOKENS(1))
!DO !i = !ancial !TO !rept.
*eneh es el tamaño de muestra 2 3 o 5.
*rept  repeticiones (1000).
TL  ssize=!eneh / muestra=!i.
!DOEND !ENDDEFINE.
*****.
DEFINE pegador (eneh=!TOKENS(1)
/ini=!TOKENS(1)/rept=!TOKENS(1))
```

```
GET FILE=
!quote(!concat('C:\GUS\tesis\temp\SE_S_',!eneh,'_',!ini,'.sav')).
!DO !i=!ini !to !rept.ADD FILES /FILE=*
 /FILE=
!quote(!concat('C:\GUS\tesis\temp\SE_S_',!eneh,'_',!i,'.sav')).
EXECUTE.!DOEND
Compute difc=Rc_pan-Rc_pri.
Compute sedifc=sqrt(sepanc*sepanc+sepric*sepric).
Compute difs=Rs_pan-Rs_pri.
Compute sedifs=sqrt(sepans*sepans+sepris*sepris).
format difc to sedifs (f10.9).
EXECUTE. SAVE OUTFILE=
!quote(!concat('C:\GUS\tesis\Data\final_L_SE_S_',!eneh,'.sav'))
 /COMPRESSED. !ENDDEFINE.
******. *Evaluación.
repetidor ancial=1 eneh = 5  rept =1000.
pegador eneh=5  ini=1  rept=1000. SET PRINTBACK=ON.
```

E.4. Programa 4: Estimación de varianza con *Jackknife*, estimador combinado

```
#Calcula el estimador de razón combinado y
#el error estándar usando \textit{Jackknife} para
#el partido DIFERENCIA para nh=2
#Función para calcular el estimador de razón combinado
comr<-function(peso,numera,denomina)
{denopes<-crossprod(peso,denomina)
numpes<-crossprod(peso,numera)
```

```
cratio<-numpes/denopes; return(cratio)}
jkdif<-function(fun,datas,enesubh)
{H<-length(tapply(datas$N_h,datas$dina,mean))
strata<-as.character(rep(1,enesubh))
for(h in 2:H){
strata<-as.character(c(strata,rep(h,enesubh)))}
upm<-as.character(datas$id)
temp<-paste(strata,upm,sep="**")
consecupm<-match(temp,unique(temp))
numupm<-length(unique(temp))
upmcol<-1:numupm
strcol<-strata[!duplicated(consecupm)]
temp<-rle(strcol)$lengths
nsubh<-rep(temp,temp)
#Construcción de pesos de replicación
nhdnhml<-nsubh/(nsubh-1)
wtmat<-matrix(datas$w,ncol=numupm,nrow=length(datas$w))
samestr<-strata[row(wtmat)]==strcol[col(wtmat)]
wtmat[samestr]<-wtmat[samestr]*matrix(nhdnhml,ncol=
numupm,nrow=length(datas$w),byrow=T)[samestr]
wtmat[consecupm[row(wtmat)]==upmcol[col(wtmat)]]<-0
#Calculo del estimador de razón combinado
combined<-fun(datas$w,datas$dif,datas$validos)
#Calculo de los estimadores de razón con las replicaciones
combinedrep<-fun(wtmat,datas$dif,datas$validos)
#Calculo de la estimación de la varianza
secombined<-sqrt(sum((combinedrep-combined[1])^2/nhdnhml))
exit<-c(combined,secombined)
return(exit)}
```

```
setwd("C:/GUS/tesis/muestras")
cant<-1000;pre<-paste("S2_",1:cant,sep="")
file<-paste(pre,"TXT",sep=".")
resul<-matrix(0,nr=length(file),nc=2)
for (i in 1:length(file))
{muestra<-read.table(file[i],header=T)
resul[i,]<-jkdif(comr,muestra,2)}
names<-c("dif2c","sejdif2c")
write.table(resul,file="C:/GUS/tesis/Data/difcomb2exit.dat",
sep="  ",row.names=F,col.names=names)
```

E.5. Programa 5: Estimación de varianza con *Jackknife*, estimador separado

```
#Calcula el estimador de razón separado y
#el error estándar usando \textit{Jackknife} para
#el partido DIFERENCIA para nh=5
jk2dif<-function(datas,me)
{#factor del estimador
Xh<-unique(datas$X_h);fac<-Xh/sum(Xh)
H<-length(tapply(datas$N_h,datas$dina,mean));n<-me*H
strata<-as.character(rep(1,me))
for(h in 2:H)
{strata<-as.character(c(strata,rep(h,me)))}
upm<-as.character(datas$id)
temp<-paste(strata,upm,sep="**")
consecupm<-match(temp,unique(temp))
```

```
numupm<-length(unique(temp))
upmcol<-1:numupm
strcol<-strata[!duplicated(consecupm)]
temp<-rle(strcol)$lengths
nsubh<-rep(temp,temp)
nhdnhml<-nsubh/(nsubh-1)
wtmat<-matrix(datas$w,ncol=numupm,nrow=length(datas$w))
samestr<-strata[row(wtmat)]==strcol[col(wtmat)]
wtmat[samestr]<-wtmat[samestr]*matrix(nhdnhml,ncol=
numupm,nrow=length(datas$w),byrow=T)[samestr]
wtmat[consecupm[row(wtmat)]==upmcol[col(wtmat)]]<-0
#Separated estimator
datas$totnum_x<-datas$dif*datas$w
datas$totdenom_x<-datas$validos*datas$w
num_x<-tapply(datas$totnum_x,datas$dina,sum)
deno_x<-tapply(datas$totdenom_x,datas$dina,sum)
rati<-num_x/deno_x
separate<-crossprod(fac,rati)
#Replicates ratio
datas$totnum<-datas$dif*wtmat
datas$totdenom<-datas$validos*wtmat
num_h<-matrix(0,H,n)
denom_h<-matrix(0,H,n)
for(k in 1:n){
num_h[,k]<-tapply(datas$totnum[,k],datas$dina,sum)
denom_h[,k]<-tapply(datas$totdenom[,k],datas$dina,sum)}
rat<-num_h/denom_h
seprat<-crossprod(fac,rat)
#Variance \textit{Jackknife} sep ratio
```

```
seR<-sqrt(sum((seprat-separate[1])^2/nhdnhml))
exit<-c(separate,seR)
return(exit)}

setwd("C:/GUS/tesis/muestras")
cant<-1000;pre<-paste("S5_",1:cant,sep="")
file<-paste(pre,"TXT",sep=".")
resul<-matrix(0,nr=length(file),nc=2)
for (i in 1:length(file))
{muestra<-read.table(file[i],header=T)
resul[i,]<-jk2dif(muestra,5)}
names<-c("dif5s","sejdif5s")
write.table(resul,file="C:/GUS/tesis/Data/difsep5exit.dat",
sep=",",row.names=F,col.names=names)
```

E.6. Programa 6: Estimación de varianza con *Bootstrap*

```
#Calcula el estimador de razón combinado, separado y
#errores estándar usando \textit{Bootstrap} para
#el partido DIFERENCIA para nh=3, mh=2, remuestras=1000
#muestras=1000;btdif<-function(datas,nh,m,repete)
{H<-length(tapply(datas$N_h,datas$dina,mean))
datas$n_h<-nh;datas$m_h<-m
datas$C_h<-sqrt((datas$m_h*(1-datas$n_h/datas$N_h))/(datas$n_h-1))
#valores muestrales para rescalamiento
Nh<-tapply(datas$N_h,datas$dina,mean);N<-sum(Nh)
W<-Nh/N;Xh<-tapply(datas$X_h,datas$dina,mean)
```

```
X<-sum(Xh);XX<-Xh/X
ymean<-tapply(datas$dif,datas$dina,mean)
xmean<-tapply(datas$validos,datas$dina,mean)
Ch<-tapply(datas$C_h,datas$dina,mean)
rescala<-function(zrar,zmen,cons)
{ zmen+cons*(zrar-zmen)}
#selector de submuestras
tota<-as.vector(table(datas$dina))
totcum<-cumsum(tota)
bce<-matrix(0,nr=repete,nc=1)
bse<-matrix(0,nr=repete,nc=1)
for(k in 1:repete)
{strsubs<-sample(tota[1],datas$m_h[1],replace=T)
if(H>1);for(h in 2:H)
{strsubs<-c(strsubs,sample(tota[h],datas$m_h[h],replace=T)+
totcum[h-1])}
strsubs<-sort(strsubs)
#create subsample
dina_s<-datas$dina[strsubs]
X_h_s<-datas$X_h[strsubs]
N_h_s<-datas$N_h[strsubs]
validos_s<-datas$validos[strsubs]
dif_s<-datas$dif[strsubs]
subsample<-data.frame(dina_s,X_h_s,N_h_s,validos_s,dif_s)
y_smean<-tapply(subsample$dif_s,subsample$dina_s,mean)
x_smean<-tapply(subsample$validos_s,subsample$dina_s,mean)
#rescalamiento
y<-rescala(y_smean,ymean,Ch)
x<-rescala(x_smean,ymean,Ch)
```

```
#Estimador combinado de la remuestra \textit{Bootstrap}
bce[k]<-crossprod(W,y)/crossprod(W,x)
#Estimador separado de la remuestra \textit{Bootstrap}
bse[k]<-crossprod(XX,(y/x))}
exit<-c(mean(bce),sqrt(var(bce)),mean(bse),sqrt(var(bse)))
return(exit)}
setwd("C:/GUS/tesis/muestras")
cant<-1000;pre<-paste("S3_",1:cant,sep="")
file<-paste(pre,"TXT",sep=".")
resul<-matrix(0,nr=length(file),nc=4)
for(i in 1:length(file))
{muestra<-read.table(file[i],header=T)
 resul[i,]<-btdif(muestra,3,2,1000)}
names<-c("bdif3c","sebdif3c","bdif3s","sebdif3s")
write.table(resul,file="C:/GUS/tesis/Data/difboot3exit.dat",
            sep=",",row.names=F,col.names=names)
```

Bibliografía

[1] **Agresti, A. y Coull, B. A.** (1998) Aproximate is better than exact for interval estimation of binomial proportions, *American Statistician*, **52**, 119-126.

[2] **Andrew, D.W.K. y Buchinsky, M.** (1998) On the number of *Bootstrap* repetitions for *Bootstrap* standard errors, confidence intervals, confidence regions, and tests, *Cowles Foundation Discussion paper* **1141 R**.

[3] **Babu, G.J. y Singh, K.** (1983) Inference on means using the *Bootstrap*, *Annals of Statistics*, **11**, 999-1003.

[4] **Bickel, P. J. y Freedman, D. A.** (1984) Asymptotic normality and the *Bootstrap* in stratified sampling, *Annals of Statistics*, **12**, 470-482.

[5] **Binder, D. A.** (1985) On the variances of asymptotically normal estimators from complex surveys., *International Statistical Review.* **51**, 37.

[6] **Böhning, D. y Viwatwongkasen CH.** (2005) Revisiting proportion estimators., *Statistical Methods in Medical Research.* **14**, 147-169.

[7] **Buckland, S. T.** (1984) Monte Carlo Confidence intervals., *Biometrics.* **40**, 811-17.

[8] **Canty, A. J., y Davison A.C.** (1998) *Variance estimation functions for two complex surveys in Switzerland*, Documento interno de Swiss fed-

eral institute of technology, bajo el contrato de Swiss federal statistical office.

[9] **Chaudhuri, A. y Stenger H.** (1992) *Survey Sampling*, Marcel Dekker Inc., E.U.A.

[10] **Cochran W.G.** (1977) *Sampling Techniques*,Third edition, John Wiley and Sons.

[11] **Davidson R. y Mackinon J.G.** (2000) *Bootstrap* tests: How many *Bootstrap*s?, *Econometrics Reviews*, **19**, 15-68.

[12] **DiCiccio, T. J. y Romano, J. P.** (1990) Nonparametric confidence limits by resampling methods and least favourable families, *International Statistical Review*, **58**, 56-76.

[13] **Efron B.** (1979) *Bootstrap* methods: Another look at the *Jackknife*, *The Annals of Statistics*,**7**, No.1,1-26.

[14] **Efron B.** (1981) Nonparametric standard errors and confidence intervals., *Canadian Journal of Statistics*, **9**,139-72.

[15] **Efron B.** (1982) The *Jackknife*, the *Bootstrap* and other resampling plans., *Philadelphia: SIAM monograph* no.**38**.

[16] **Eslava G. G., Méndez R.I. y Castrejon J.L.** (2003) *The 2000 Mexican Presidential election: Sampling Designs Evaluation* ,I.I.M.A.S.-U.N.A.M., No.118

[17] **Frangos, C. C.** (1987) An updated bibliography on the *Jackknife* method, *Communications in Statistics-Theory and Methods*, **16**, 1543-84.

[18] **Gray, H. L. y Schucany, W.R.** (1972) *The Generalized Jackknife Statistic*, Marcel Dekker, New York.

[19] **Gross, S. T.** (1980) Median estimation in sample surveys. Proceedings of the Section on survey Research. *Journal of American Statistical Association*, 181-184.

[20] **Hall, P.** (1992) *The Bootstrap and edge worth Expansion*. Springer-Verlag, New York.

[21] **Hansen M.H., Hurwitz W.N., Madow W.G.** (1953) *Survey methods and theory*. Vol I y II, New York, John Wiley & Sons.Inc.

[22] **Hartigan, J. A.** (1969) Using subsample values as typical value. *Journal of American Statistical Association*, **64**, 1303-1317.

[23] **Hinkley, D.V.** (1983) *Jackknife* methods. *Encyclopedia of Statistical Sciences*, **4**, 280-7.

[24] **Kish L.** (1987) *Statistical Design for Research* , New York; John Wiley & Sons.

[25] **Krewski, D. y Rao J.N.K.** (1981) Inference from stratified samples: Properties of the Linearization, *Jackknife* and balanced repeated replication methods, *Annals of Statistics*, **9**, 1010-1019.

[26] **Kuk, A. Y. C.** (1989) Double *Bootstrap* estimation of variance under systematic sampling with probability proportional to size, *Journal of Statistical Computation and Simulation*, **31**, 73-82.

[27] **Lawrence D.B., T. T.Cai. y A. DasGupta.** (2002) Confidence intervals for a binomial proportion and asymptotic expansions, *Annals of Statistics*, **30**, 160-201.

[28] **Lehtonen E. y Pahkinen** (1994) *Practical Methods for design and Analysis of Complex Surveys.*

[29] **Lenka M., Jean D. y Laurianes R.** (2005) *A Study of the properties of a Bootstrap Variance Estimator Under Sampling Without Replacement.* Statistics Canada,

`http://www.fcsm.gov/05papers/Mach_Dumais_Robidou_VA.pdf`

[30] **McCarthy, P.J., y Snowden, C. B.** (1985) *The Bootstrap and finite population sampling*, Vital and Health Statistics, Series 2. No.95.DHHS Púb. No.(PHS)85-1369. Washington, DC: Public Health Service, US Government Printing Office.

[31] **Miller, R.G.** (1974) The *Jackknife*-a review, *Biometrika*, **61**,1-15.

[32] **Owen, A. B.** (1988) Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, **75**, 237-49.

[33] **Palmer A.C., Eslava G. G. y Méndez R.I.** (2001) *Método de remuestreo para el cálculo de varianzas en muestreos complejos a la ENAL '96* .

[34] **Parr, W.C. y Shucany, W.R.**(1980) The *Jackknife*: a bibliography. *International Statistical Review*, **48**,73-8.

[35] **Quenouille, M.H.** (1949) Approximate tests of correlation in time series, *Journal of the Royal Statistical Society*, B, **11** 68-84.

[36] **Quenouille, M.H.** (1956) Notes on bias in estimation, *Biometrika*, **43** 353-60.

[37] **Rao, C. R.** (1973) *Linear Statistical Inference and Its Aplications*, 2nd ed.Wiley New York.

[38] **Rao, J.N.K. y C.F.J.Wu** (1988) Resampling inference with complex survey data, *Journal of American Statistical Association*, **83**, 231-241.

[39] **Rao, J.N.K. y K.F.Rust** (1996) Variance Estimation for Complex Surveys using Replication Techniques, *Statistical Methods in Medical Research*, **5**, 283-310.

[40] **Rubin, D. B.** (1981) The Bayesian *Bootstrap*, *Annals of Statistics*, **9**, 130-134.

[41] **Särndal, C.E., Swensson, B., y Wretman J.H.** (1991) *Model Assisted Survey Sampling*, Ed. Springer-Verlag.

[42] **Sen, A.R.** (1953) On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* **5**, 119-127.

[43] **Shao J. y Tu D.** (1995) *The Jackknife and Bootstrap* , New York; Springer-Verlag.

[44] **Shi, X., Chen, J. y Wu, C.F.J.** (1990) Weak and strong representations for quantile processes from finite populations with applications to simulation size in resampling methods, *Canadian Journal of Statistics*,**18**, 141-148.

[45] **Sitter, R.R.** (1992) Comparing three *Bootstrap* methods for survey data. *Proceedings of the Social Statistics Section. American Statistical Association*, 11-18.

[46] **Tepping, B. J.** (1968) Variance estimation in complex surveys, *Canadian Journal of Statistics*, **20**, No.2, 135-154.

[47] **Tukey, J.W** (1958) Bias and confidence in not quite large samples, *Annals of Mathematical Statistics*, **29**, 614.

[48] **Wolter K.M.** (1985) *Introduction to Variance Estimation* , New York; Springer-Verlag.

[49] **Woodruff, R. S.** (1971) A simple method for approximating the variance of a complicated estimate, *Journal of the American Statistical Association*, **66**. 411-414.

[50] **Yates, F., y Grundy, P.M.** (1953) Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, **B15**, 235-261.

[51] http://www.consulta.com.mx/interiores/99_pdfs/17_articulosinteres_pdf/ai_AR200501_EncSalidaCR.pdf.

[52] http://www.unc.edu/ skolenik/talks/survey-resampling-by2.pdf.