

Un estimador de la varianza del total estimado bajo muestreo sistemático de una población específica

Mónica Tinajero Bravo *

Guillermina Eslava Gómez **

Luis Cruz Orive ***

Enero de 2010

*Candidato a Doctor en Ciencias Matemáticas, Posgrado en Ciencias Matemáticas, UNAM. Ciudad Universitaria, 04510, D.F., México, tinajerobm@gmail.com. Este reporte de investigación es parte del trabajo de investigación del programa de Doctorado en Ciencias Matemáticas.

**Depto. de Matemáticas, Facultad de Ciencias, UNAM. Ciudad Universitaria, 04510, D.F., México, eslava@matematicas.unam.mx.

***Depto. de Matemáticas, Estadística y Computación, Facultad de Ciencias, Universidad de Cantabria, Av. Los Castros s/n, E-39005 Santander, España, lcruz@matesco.unican.es.

Índice

1. Introducción	4
2. Varianza del estimador del total poblacional bajo muestreo sistemático	6
3. Algunos estimadores de la varianza	10
4. Varianza del estimador bajo el modelo específico “modelo de cola de ballena”, WTM	12
4.1. WTM población continua	13
4.1.1. Estimador del total	14
4.1.2. Varianza del estimador del total	15
4.1.3. Varianza del estimador del total considerando un error en el modelo . . .	17
4.1.4. Estimador de la varianza	18
4.2. WTM población finita	18
4.2.1. Estimador del total	21
4.2.2. Varianza del estimador del total	21
4.2.3. Varianza del estimador del total considerando un error en el modelo . . .	22
4.2.4. Estimador de la varianza	22
5. Experimento empírico y simulaciones	23
6. Anexos	25
6.1. Relación entre \widehat{V}_{Sdi} y \widehat{V}_{GC}	25
6.2. Derivación de $V_{WTM}(\hat{t}_{\pi}^*)$, población continua	26
6.3. Estimador de \hat{t}_{π}^* , WTM población finita	33

Resumen

En general no existen estimadores insesgados para la varianza del estimador del total poblacional bajo muestreo sistemático. Se han obtenido varias aproximaciones suponiendo que la población de estudio es una realización de una superpoblación que sigue un modelo específico, o bien mediante métodos no paramétricos basados en la teoría transitiva de G. Matheron.

En esta trabajo se presenta una aproximación, a partir de una sola muestra, de la varianza del estimador del total poblacional bajo muestreo sistemático sobre una población normal, reordenada de una manera especial que tiende a reducir dicha varianza. El gráfico de la función de medida resultante, f , es una versión ‘simetrizada’ de la inversa de la función de distribución de una variable aleatoria normal con las colas truncadas; debido a la forma de f le llamamos “Modelo de cola de ballena” (*Whale Tail Model*). Se demuestra que si el tamaño de muestra es un número par y la distribución normal se trunca de forma simétrica, la varianza del estimador del total es cero. Adicionalmente, se proponen dos estimadores de la varianza cuando hay un error en el modelo, $\tilde{f} = f + \epsilon$, donde ϵ es una variable aleatoria, es decir, cuando puede suponerse que la población de interés, antes de ordenarse, se aproxima a una población normal. Generando poblaciones a partir de una $N(0, 1)$ truncada y ordenándolas de la manera propuesta, se observa que uno de los estimadores sugeridos es el de menor sesgo y mayor estabilidad, en comparación con tres estimadores señalados en la literatura.

Palabras clave: Muestreo sistemático, estimador de Horvitz-Thompson, varianza de un estimador, población finita, población continua, superpoblación.

1. Introducción

El muestreo sistemático es usado ampliamente en la práctica debido a su simplicidad operativa, y a que, para cierta clase de poblaciones, genera estimadores más eficientes que el muestreo aleatorio simple. Su eficiencia depende de las propiedades de la distribución de los valores de la variable en la población, es decir, del orden de la misma. Algunas veces el sistemático es más preciso, pero en raras ocasiones se está seguro que será el más eficiente, por lo que es necesario tener información sobre la estructura de la población para usarlo de manera efectiva.

La desventaja principal del muestreo sistemático es que no existe una expresión analítica, basada en una sola muestra y sin algún supuesto sobre la población de la cual se extrajo la muestra, para estimar la varianza del estimador del total, \hat{t}_π , de manera insesgada (Cochran [2], Iachan [9]). No obstante, en la literatura se señalan algunas aproximaciones para el estimador de la varianza, que se han generado en al menos dos contextos: en el de las poblaciones finitas y en el de la Estereología, que trata principalmente con poblaciones continuas. En el primer enfoque se han propuesto estimadores bajo el supuesto que la población finita es una realización de una superpoblación que sigue un modelo determinado, estimadores que usan información auxiliar como el de regresión, estimadores que usan métodos de remuestreo y por último, modificar el diseño al elegir más de un arranque aleatorio. En el área de la Estereología, ciencia que se encarga de la inferencia estadística de parámetros cuantitativos de estructuras espaciales basados en muestras geométricas como un plano o secciones en línea a través de la estructura (Kiêu [12]), una herramienta importante la constituyen los métodos transitivos debidos a Matheron, los cuales son libres de distribución y consideran las transiciones o saltos de la primera derivada no continua de la función de medida f .

En el contexto de muestreo de poblaciones finitas, uno de los primeros trabajos se debe a Madow y Madow [13], quienes sentaron las bases para la teoría del muestreo sistemático desarrollando fórmulas para la varianza del estimador de la media en términos de la varianza y las autocorrelaciones poblacionales. Por su parte, Cochran [2] señala que no hay un estimador insesgado de la varianza a partir de una sola muestra sistemática, y cualquier estimación dependerá de los supuestos que se hagan sobre la forma de la población muestreada. Cochran [2] deriva expresiones para la varianza de tres diseños: sistemático, aleatorio simple y estratificado, suponiendo que la población finita es generada a partir de una población infinita (superpoblación), en la cual la autocorrelación entre dos unidades separadas u unidades, ρ_u , es una función monótona decreciente de u . Bajo este modelo, las varianzas son funciones lineales de la correlación y demuestra que no se puede establecer un resultado general para todas las poblaciones con ρ_u monótonamente decreciente. Sin embargo, si se hacen supuestos adicionales como que ρ_u es una función lineal de u , entonces el muestreo sistemático es el más eficiente, en

términos del valor esperado de la varianza, siguiéndole el estratificado y por último el aleatorio simple; si ρ_u sigue una función exponencial, entonces el muestreo más eficiente corresponde al sistemático, y además establece un estimador consistente para la varianza. Yates [21] propone usar lo que denomina “muestras sistemáticas parciales” las cuales consisten en dividir la muestra en segmentos para obtener comparaciones independientes. Quenouille [16] parte del trabajo realizado por Cochran [2] y señala que las expresiones para la varianza derivadas por éste último se pueden obtener bajo condiciones más generales.

Diferentes supuestos sobre la distribución de la población dan lugar a diferentes estimadores de la varianza, que son en general insesgados bajo el modelo de superpoblación que se asuma, Iachan [9]. Wolter [19], [20] compara ocho estimadores, analizando algunas de sus propiedades teóricas suponiendo cinco modelos para super poblaciones (aleatorio, tendencia lineal, efectos de estratificación, autocorrelacionado de orden uno y efectos periódicos), así como sus propiedades empíricas. De acuerdo con Wolter, estos estimadores son representativos de las diferentes soluciones que resultan útiles en la práctica: a) tratar la muestra sistemática como una muestra aleatoria simple sin reemplazo (SI), b) un estimador basado en diferencias sucesivas entre los valores muestrales, c) aproximación mediante un muestreo aleatorio estratificado con dos unidades seleccionadas de cada estrato de $2T$ unidades, d) considerar la segunda diferencia entre los datos en la muestra e) suponer que la correlación entre dos elementos separados u unidades es de tipo exponencial, $\rho_u = e^{-\lambda u}$, f) considerar diferencias de observaciones sucesivas hasta de orden 4, g) diferencias hasta de orden 8 y h) hacer una partición de la muestra en k submuestras de tamaño n/k . Si no se puede suponer algún modelo para la población, sugiere usar los estimadores señalados en b) y c), los cuales parecen ser buenas aproximaciones para diferentes clases de poblaciones.

Otra aproximación consiste en seleccionar dos o más arranques aleatorios de manera independiente, es decir, tomar réplicas de muestras sistemáticas. Esta idea se remonta a los trabajos de Madow y Madow [13], quienes describen este método de selección, así como a los de Mahalanobis [14], quien introdujo el concepto de muestras interpenetrantes para referirse a la réplicas. Este método genera un estimador insesgado de la varianza, pero tiene la desventaja de que se pierde precisión (Iachan [9]) si el número de submuestras es pequeño, o si el número de submuestras se aumenta dejando el tamaño total de muestra fijo. Gautschi [7] demuestra que si la superpoblación cumple con: a) la autocorrelación ρ_u es decreciente y b) el correlograma es cóncavo hacia arriba, es decir, $\rho_{u+1} - 2\rho_u + \rho_{u-1} \geq 0$, entonces el valor esperado de la varianza para un muestreo con una sola muestra sistemática es menor que el de un muestreo con varias muestras sistemáticas.

Otro enfoque es usar información auxiliar, es decir, involucrar información de otras variables

adicionales a la variable de interés que tienen relación con esta última. Berger [1], propone un estimador de varianza para un muestreo sistemático con probabilidades desiguales de selección para cada unidad en la población basado en un modelo de regresión.

Por otra parte, en el contexto de la Estereología, un total t puede verse como la aproximación de la integral de la función de medida $f = y$ a partir de datos discretos. La varianza del estimador se obtiene a través de los métodos transitivos, debidos a Matheron [15], los cuales se basan en el comportamiento del covariograma de f cerca del origen y han sido adaptados a este contexto por diversos autores. En Kiêu [10] y Kiêu et al. [12] la varianza del estimador, $V_{SY}(\hat{t}_\pi)$, se aproxima usando la fórmula de Euler-MacLaurin con varias modificaciones, mediante lo que se denomina el término de extensión, bajo los supuestos de que la función f es suave a trozos (m, p) y el intervalo muestral T es lo suficientemente pequeño, con $m \in \mathbb{Z}^+ \cup 0$. García-Fiñana y Cruz-Orive [6] proponen una generalización del estimador anterior donde m ya no necesariamente es un entero, misma que se obtiene a través de un refinamiento de la fórmula de Euler-MacLaurin usando herramientas de cálculo fraccional.

Tomando la experiencia que se maneja en esa área referente al grado de suavidad de la población y al orden de la misma [8], derivamos la varianza del estimador bajo muestreo sistemático asumiendo lo que denominamos el “Modelo de cola de ballena” (WTM por sus siglas en inglés *Whale Tail Model*). Posteriormente, proponemos dos estimadores de la varianza en base a los resultados obtenidos.

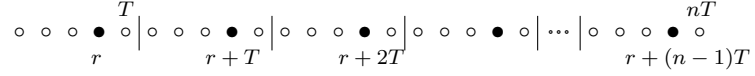
Este trabajo está organizado de la siguiente manera. En la sección 2 se introduce la notación básica, el estimador del área o del total y su varianza considerando dos tipos de poblaciones: continuas y discretas. En la sección 3 se reportan tres de las aproximaciones encontradas en la literatura para la varianza del estimador. En la sección 4, se desarrolla la varianza del estimador bajo el WTM y se proponen dos aproximaciones para la misma cuando se toma en cuenta un error en el modelo. Finalmente, en la sección 5 mediante un ejercicio de simulación, se muestra el desempeño de las aproximaciones propuestas.

2. Varianza del estimador del total poblacional bajo muestreo sistemático

Sea y_i el valor de la variable de interés para el elemento i de la población finita de tamaño N , cuyos elementos se enumerarán como $U = \{1, 2, \dots, N\}$. Se selecciona una muestra de tamaño n mediante muestreo sistemático, s_r , la cual consta de las unidades $s_r = \{r, r+T, r+2T, \dots, r+(n-1)T\}$, donde $T = N/n$ representa el salto o intervalo muestral y r denota el arranque que

es una realización de $U_r \sim U\{1, \dots, T\}$. Para simplificar el problema, se supondrá que $T \in \mathbb{Z}^+$ en el caso de una población finita y para simplificar notación, las unidades en la muestra se enumerarán serialmente como $s_r = \{1, \dots, n\}$. En este trabajo se usará indistintamente $i \in s_r$ o $i \in \{1, \dots, n\}$ para denotar que la unidad i pertenece a la muestra.

Esquemáticamente, el método de selección se puede representar como sigue



Considérese el caso del total poblacional t , el cual puede expresarse mediante

$$t = \sum_{i=1}^N y_i = \sum_{r=1}^T t_{s_r} \quad (1)$$

donde $t_{s_r} = \sum_{i \in s_r} y_i$.

En general, si s_r es una muestra probabilística, un estimador insesgado del total poblacional es el denominado estimador- π o de Horvitz-Thompson:

$$\hat{t}_\pi = \hat{t}_{s_r} = \sum_{i \in s_r} \frac{y_i}{\pi_i}, \quad (2)$$

equivalentemente, el estimador puede expresarse como una función lineal de las variables indicadoras I_i ,

$$\hat{t}_\pi = \sum_{i \in U} I_i \frac{y_i}{\pi_i},$$

donde

$$I_i = \begin{cases} 1 & \text{si } i \in s_r \\ 0 & \text{en otro caso} \end{cases} \quad y$$

$\pi_i = Pr(i \in s_r) = Pr(I_i = 1)$ es la denominada probabilidad de inclusión de primer orden.

La varianza de (2) está dada por (Särndal [17], resultado 2.8.1)

$$V(\hat{t}_\pi) = \sum_U \sum_U \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j = \sum_U \sum_U \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - \left(\sum_U y_i \right)^2, \quad (3)$$

donde $\pi_{ij} = Pr(i, j \in s_r) = Pr(I_i = 1, I_j = 1)$ corresponde a la probabilidad de inclusión de segundo orden. Si se cumple que $\pi_{ij} > 0$ para toda $i, j \in U$, un estimador insesgado de $V(\hat{t}_\pi)$ está dado por

$$\widehat{V}(\hat{t}_\pi) = \sum_{s_r} \sum_{s_r} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} \right) y_i y_j. \quad (4)$$

En el caso de un muestreo sistemático de unidades con igual probabilidad, las probabilidades de inclusión de primer y segundo orden, π_i y π_{ij} son, respectivamente:

$$\pi_i = \frac{n}{N} \text{ con } i = 1, \dots, N$$

$$\pi_{ij} = \begin{cases} \frac{n}{N} & \text{si } i, j \in s_r, r = 1, \dots, T \\ 0 & \text{si } i \in s_r \text{ y } j \in s_l \text{ con } r \neq l. \end{cases}$$

En consecuencia, el estimador del total (2) y su varianza (3) se reducen a las expresiones siguientes (Särndal [17], resultado 3.4.1)

$$\hat{t}_\pi = \hat{t}_{s_r} = T \sum_{i \in s_r} y_i, \quad (5)$$

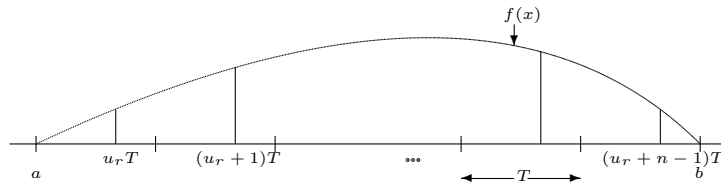
$$V_{SY}(\hat{t}_\pi) = T \sum_{r=1}^T \left(t_{s_r} - \frac{t}{T} \right)^2 = \frac{\sum_{r=1}^T (t_{s_r} - t)^2}{T}. \quad (6)$$

Desde el enfoque de la Estereología, el parámetro de interés t puede ser escrito como la integral de alguna función de medida f sobre un espacio unidimensional, es decir,

$$t = \int_{\mathbb{R}} f(x) dx \quad (7)$$

donde $f : \mathbb{R} \mapsto \mathbb{R}$, es integrable, con soporte acotado en \mathbb{R} .

Los puntos donde se observa f son tomados de una muestra sistemática, y por lo tanto, la información con la que se dispone es $\{(x, f(x)) : x \in s_r\}$, donde $s_r = \{u_r T + kT, k \in \mathbb{Z}\}$ y u_r es una realización de $U_r \sim U(0, 1)$. Gráficamente se tiene que



Análogamente a (5), un estimador insesgado de t es:

$$\hat{t}_\pi = T \sum_{x \in s_r} f(x). \quad (8)$$

Nótese que \hat{t}_π puede verse como la aproximación numérica de la integral a partir de datos discretos, por lo que la varianza del estimador se obtiene a partir de modificaciones a la fórmula de Euler-MacLaurin propuestas en esta área, mediante lo que se denomina el término de extensión, bajo los supuestos de que la función f es suave a trozos (m, p) y el intervalo muestral T es lo suficientemente pequeño.

Dada una función $f : \mathbb{R} \mapsto \mathbb{R}$, la amplitud del salto o transición de su m -ésima derivada $f^{(m)}$, se define mediante

$$Sf^{(m)}(x) := \lim_{y \rightarrow x^+} f^{(m)}(y) - \lim_{y \rightarrow x^-} f^{(m)}(y), \quad x \in \mathbb{R}, \quad m = 0, 1, 2, \dots$$

siempre que el límite exista, su soporte es el conjunto

$$Df^{(m)} = \{x : Sf^{(m)}(x) \neq 0\}.$$

Se dice que la función f es suave a trozos (m, p) , con $m, p \in \mathbb{N}$ si:

- a) $Df^{(k)} = \phi$, $k = 0, 1, \dots, m - 1$ y $Df^{(m)} \neq \phi$ y
- b) $f^{(k)}$ tiene un número finito de saltos y estos son finitos, $k = m, m + 1, \dots, m + p$.

El orden de la primera derivada de f no continua es m , es decir, es el orden de la derivada en la cual ocurren saltos o transiciones.

Originalmente, Kiêu [10] supone que $m \in \mathbb{Z}^+ \cup 0$ y obtiene que la varianza del estimador de t , bajo un muestreo sistemático, si f es de orden (m, p) , está dado por la expresión siguiente

$$\begin{aligned} V_{SY}(\hat{t}_\pi) &= (-1)^m T^{2m+2} \sum_{s, t \in Df^{(m)}} P_{2m+2, T}(t-s) Sf^{(m)}(s) Sf^{(m)}(t) + o(T^{2m+2}) \\ &= (-1)^m T^{2m+2} P_{2m+2, T}(0) \sum_{t \in Df^{(m)}} (Sf^{(m)}(t))^2 \\ &\quad + (-1)^m T^{2m+2} \sum_{s, t \in Df^{(m)}, t-s \neq 0} P_{2m+2, T}(t-s) Sf^{(m)}(s) Sf^{(m)}(t) + o(T^{2m+2}), \end{aligned} \tag{9}$$

donde $P_{l, T}(x) = P_l\left(\frac{x}{T} - \left[\frac{x}{T}\right]\right)$ es una función con período T , $[x]$ que denota la parte entera de x y $P_l(x)$ un polinomio de Bernoulli, $l \geq 1$.

La ecuación (9) se representa también de la siguiente manera

$$V_{SY}(\hat{t}_\pi) = V_E(\hat{t}_\pi) + Z(T) + o(T^{2m+2}).$$

donde

- i. $V_E(\hat{t}_\pi)$ es el llamado término de extensión y corresponde al primer sumando de la última igualdad en (9), sólo depende de las amplitudes de las transiciones de $f^{(m)}$.

- ii. $Z(T)$ es el denominado Zitterbewegung y corresponde a los sumandos con $t - s \neq 0$, lo cual depende tanto de las amplitudes de las transiciones como de su distribución en el eje muestral. Es una función oscilante de T .
- iii. $o(T^{2m+2})$ es una función tal que $\lim_{T \rightarrow 0} \frac{o(T^{2m+2})}{T^{2m+2}} = 0$.

Cuando T es suficiente pequeño $V_E(\hat{t}_\pi)$ constituye una buena aproximación de $V_{SY}(\hat{t}_\pi)$ ya que representa la tendencia central de la varianza. Este término puede escribirse como

$$V_E(\hat{t}_\pi) = -\frac{B_{2m+2}}{(2m+1)!} \frac{T^{2m+2}}{m+1} g^{(2m+1)}(0^+) \quad (10)$$

donde

$$g^{(2m+1)}(0^+) = \frac{1}{2} S g^{(2m+1)}(0) = \frac{-(-1)^m}{2} \sum_{t \in Df^{(m)}} S f^{(m)}(t),$$

g es el covariograma alrededor del origen de la función de medida f .

B_{2m+1} es un número de Bernoulli. Existe una relación entre los polinomios y los números de Bernoulli, $B_{2m+2} = (2m+2)! P_{2m+2}(0)$. Los primeros números de Bernoulli son: $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_4 = -1/30$ y $B_3 = B_5 = B_7 = \dots = 0$.

García-Fiñana y Cruz-Orive [6] proponen una generalización de la varianza anterior, la cual se deriva relajando el supuesto de que los saltos de la función de medida son finitos, en este caso m ya no necesariamente es un entero, denotándolo por q . La aproximación se obtiene a través de un refinamiento de la fórmula de Euler-MacLaurin usando herramientas de cálculo fraccional. La expresión analítica depende de un parámetro de suavización $q \in [0, 1]$, que generalmente se desconoce y habrá que estimarlo o bien suponerlo.

3. Algunos estimadores de la varianza

Como se señaló, la condición $\pi_{ij} > 0$ no se cumple para toda $i, j \in U$ en el caso del muestreo sistemático y por ello la ecuación (4) no se puede usar para estimar la varianza. No obstante, se han propuesto varias aproximaciones, a continuación se presentan tres de ellas por ser las que algunos autores recomiendan, y porque en una comparación empírica efectuada por los autores de este trabajo, basada tanto en datos reales como poblaciones simuladas, resultaron ser las de menor sesgo.

- i. Una aproximación se basa en diferencias sucesivas de los valores muestrales. Esta fue sugerida por Yates [22] y señalada también en Sukhatme, Sukhatme y Asok [18]

$$\widehat{V}_{Fdi}(\hat{t}_\pi) = N^2 (1-f) \frac{1}{n} \sum_{i=2}^n \frac{(y_i - y_{i-1})^2}{2(n-1)}, \quad (11)$$

donde $f = n/N = 1/T$ es la fracción de muestreo. Fuller [4] (resultado 5.3.5) señala que este estimador resulta de suponer un modelo en el que dos observaciones adyacentes tienen la misma media, es decir, $y_i = \mu + e_i$ con $e_i \sim ind(0, \sigma_i^2)$, $i = 1, \dots, n$.

ii. Otra considera la segunda diferencia de los datos muestrales (Wolter [19])

$$\widehat{V}_{Sdi}(\hat{t}_\pi) = N^2 (1 - f) \frac{1}{n} \sum_{i=3}^n \frac{(y_i - 2y_{i-1} + y_{i-2})^2}{6(n-2)}. \quad (12)$$

Cochran [3] señala que este estimador resulta de suponer un modelo lineal para la población, es decir, $y_i = \beta_0 + \beta_1 x_i + e_i$ con $e_i \sim ind(0, \sigma_i^2)$. Fuller [4] (resultado 5.3.7) también propone un estimador muy similar al anterior, la diferencia radica en el primer y último sumandos de la expresión.

iii. En el área de la Estereología, Kiêu [10] obtiene una aproximación del covariograma cerca del origen, $g^{(2m+1)}(0^+)$ para estimar $V_E(\hat{t}_\pi)$. Tal aproximación la hace considerando que: a) las estimaciones del covariograma están disponibles en un conjunto discreto de valores, b) usando la fórmula de Taylor y c) aplicando el método de mínimos cuadrados para estimar los coeficientes involucrados en la expansión de Taylor. El estimador resultante tiene la forma

$$\widehat{V}_E(\hat{t}_\pi) = -\frac{B_{2m+2}}{(m+1)} T \sum_{k=1}^l \lambda_k C_k \quad (13)$$

donde $C_k = \sum_{i=1}^{n-k} y_i y_{i+k}$ y l es un entero tal que $lT \in (0, \delta)$, intervalo en el que $g(lT)$ es $q + 1$ veces continuamente diferenciable, con $2m + 1 \leq q \leq 2m + 2p - 1$.

Si por ejemplo $m = 0$, $q = 2$ y $l = 2$, entonces

$$\widehat{V}_E(\hat{t}_\pi) = \frac{T^2}{12} (3C_0 - 4C_1 + C_2),$$

en cambio, si $m = 1$, $q = 3$ y $l = 2$, el estimador que se obtiene es

$$\widehat{V}_E(\hat{t}_\pi) = \frac{T^2}{240} (3C_0 - 4C_1 + C_2).$$

El estimador de varianza cuando $m = 0$ es veinte veces el estimador cuando $m = 1$. Posteriormente, García-Fiñana y Cruz-Orive [6] desarrollan la generalización siguiente a la ecuación (13) para el caso en que $m = q \in [0, 1]$

$$\widehat{V}_{GC}(\hat{t}_\pi) = \alpha(q) [3C_0 - 4C_1 + C_2] T^2,$$

donde $\alpha(q) = \frac{\Gamma(2q+2)\zeta(2q+2)\cos(q\pi)}{(2\pi)^{2q+2}(1-2^{2q-1})}$, $\zeta(w) = \sum_{k=1}^{\infty} \frac{1}{k^w}$ denota la función zeta de Riemann y q es una constante o parámetro de suavización que se puede estimar mediante $\hat{q} = \frac{1}{2\log k} \log \left(\frac{3C_0 - 4C_k + C_{2k}}{3C_0 - 4C_1 + C_2} \right) - \frac{1}{2}$. No existe una guía para el valor apropiado de $k = 2, 3, \dots$, pero los autores recomiendan $k = 2, 4$.

Si se considera la corrección por finitud, $(1 - f)$, el estimador anterior toma la forma siguiente

$$\widehat{V}_{GC}(\hat{t}_\pi) = N^2(1 - f) \frac{1}{n^2} \alpha(q) [3C_0 - 4C_1 + C_2]. \quad (14)$$

García-Fiñana y Cruz-Orive [6] muestran que para datos sintéticos, usando una función de medida con $q = 1/2$, y datos reales de cerebros humanos, su estimador de varianza con q estimado se comporta mejor que con $q = 0$ ó $q = 1$.

Por otra parte, haciendo un poco de álgebra, se puede demostrar que (ver Anexo 6.1)

$$\begin{aligned} \widehat{V}_{Sdi}(\hat{t}_\pi) &= N^2(1 - f) \frac{1}{3n(n-2)} \left[3 \sum_{i=1}^n y_i^2 - 4 \sum_{i=2}^n y_i y_{i-1} + \sum_{i=3}^n y_i y_{i-2} \right] \\ &\quad + N^2(1 - f) \frac{(-5y_1^2 - y_2^2 - y_{n-1}^2 - 5y_n^2 + 4y_2 y_1 + 4y_n y_{n-1})}{6n(n-2)}, \end{aligned}$$

si $n \approx n - 2$ y $\alpha(q = -1/2) = 1/3$, entonces

$$\widehat{V}_{Sdi}(\hat{t}_\pi) \approx \widehat{V}_{GC}(\hat{t}_\pi) - N^2(1 - f) \frac{1}{6n^2} [5y_1^2 + y_2^2 + y_{n-1}^2 + 5y_n^2 - 4y_2 y_1 - 4y_n y_{n-1}],$$

es decir, el estimador (12) puede expresarse como un caso particular de (14) con parámetro $\alpha(-1/2) = 1/3$ menos una cantidad que depende de y_1, y_2, y_{n-1} y y_n .

Si $2\sqrt{5} \approx 4$, \widehat{V}_{Sdi} será menor que \widehat{V}_{GC} con $q = -1/2$, ya que la expresión anterior se puede escribir como

$$\widehat{V}_{Sdi}(\hat{t}_\pi) \approx \widehat{V}_{GC}(\hat{t}_\pi) - N^2(1 - f) \frac{1}{6n^2} [(\sqrt{5}y_1 - y_2)^2 + (\sqrt{5}y_n - y_{n-1})^2].$$

4. Varianza del estimador bajo el modelo específico “modelo de cola de ballena”, WTM

La eficiencia del muestreo sistemático depende del orden de los elementos de la población, la varianza es más pequeña si los totales muestrales \hat{t}_{sr} son similares, como puede observarse en la expresión (6). Los estimadores $\widehat{V}_{Fdi}(\hat{t}_\pi)$ y $\widehat{V}_{Sdi}(\hat{t}_\pi)$ se derivan al suponer para la población un modelo de media constante entre dos elementos adyacentes y lineal, respectivamente. En la

derivación de $\widehat{V}_{GC}(\hat{t}_\pi)$ se encuentra que $V_{SY}(\hat{t}_\pi)$ depende del grado de suavidad de la función de medida para la cual se quiere estimar el total o la media, a través del parámetro de suavización q . Es importante hacer notar que al suponer un modelo, implícitamente se está induciendo un orden.

En esta sección la varianza del estimador se deriva bajo un modelo específico para la población que se ha denominado por su forma “modelo de cola de ballena” (WTM por sus siglas en inglés *Whale Tail Model*), mismo que involucra una población normal reordenada de cierta manera para reducir varianza. Primero se hará la derivación en el caso de una población continua y posteriormente se planteará un modelo análogo en el caso de una población finita.

4.1. WTM población continua

En el caso de que la población de interés sea continua y fija, se define como la función de medida, $f^*(p)$, a la función (figura 1)

$$f^*(p) = y^* = \begin{cases} \Phi^{-1}(p) + |a| & \text{si } \Phi_a \leq p \leq \Phi_b \\ \Phi^{-1}(2 - p) + |a| & \text{si } 2 - \Phi_b \leq p \leq 2 - \Phi_a \end{cases} \quad (15)$$

donde

$\Phi^{-1}(p)$ denota la función inversa de una $N(0, 1)$ para el cuantil p , es decir, si $y = \Phi^{-1}(p)$ entonces $p = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$,

$$\Phi(a) = \Phi_a \text{ y } \Phi(b) = \Phi_b,$$

$a = \Phi^{-1}(\Phi_a)$, $b = \Phi^{-1}(\Phi_b)$. Por definición de $f^*(p)$ se está trabajando con una normal truncada,

$y^* = y + |a|$ es la función y trasladada a unidades hacia arriba, con

$$f(p) = y = \begin{cases} \Phi^{-1}(p) & \text{si } \Phi_a \leq p \leq \Phi_b \\ \Phi^{-1}(2 - p) & \text{si } 2 - \Phi_b \leq p \leq 2 - \Phi_a \end{cases} . \quad (16)$$

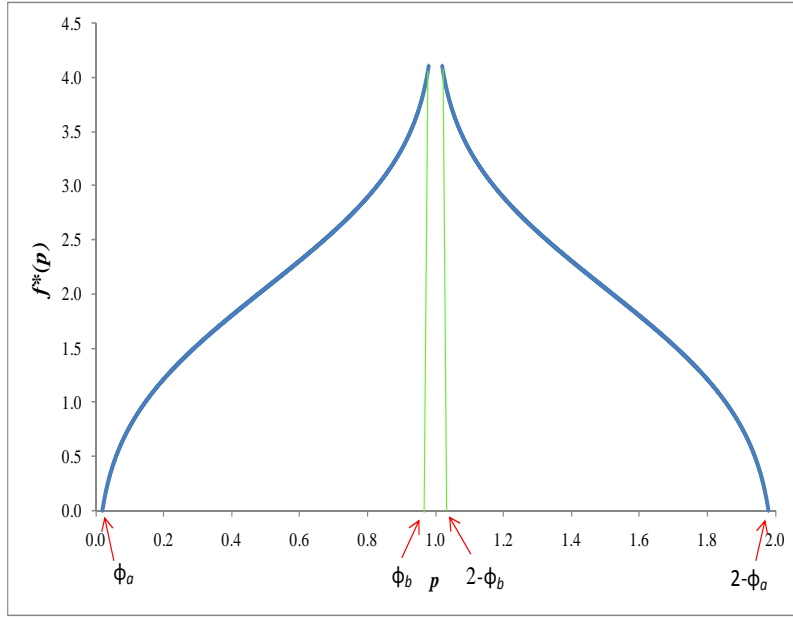


Figura 1: Modelo población continua, $y^* = f^*(p)$

4.1.1. Estimador del total

En esta situación, un parámetro objetivo t^* puede ser el área bajo la curva de la función de medida $f^*(p)$, es decir,

$$\begin{aligned}
 t^* &= \int_{\Phi_a}^{\Phi_b} f^*(p) dp + \int_{2-\Phi_b}^{2-\Phi_a} f^*(p) dp \\
 &= 2 \int_{\Phi_a}^{\Phi_b} f^*(p) dp \\
 &= 2 \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(p) dp + 2|a|(\Phi_b - \Phi_a).
 \end{aligned}$$

Un estimador insesgado de t^* bajo un muestreo sistemático, está dado por

$$\begin{aligned}
 \hat{t}_\pi^* &= T \sum_{i \in s_r} y_i^* \\
 &= T \sum_{i \in s_r} (\Phi^{-1}(p_i) + |a|)
 \end{aligned}$$

donde

$$T = \frac{2(\Phi_b - \Phi_a)}{n} \text{ es el intervalo muestral,}$$

$$p_i = \begin{cases} \Phi_a + [u + i - 1]T & \text{si } \Phi_a \leq p_i \leq \Phi_b \\ 2[1 - \Phi_b] + \Phi_a + [u_r + i - 1]T & \text{si } 2 - \Phi_b \leq p_i \leq 2 - \Phi_a, \end{cases}$$

u_r generado de una $U(0, 1)$.

4.1.2. Varianza del estimador del total

A continuación se derivará una expresión para $V_{WTM}(\hat{t}_\pi^*)$ considerando que el tamaño de muestra n es par, y se demostrará que cuando los límites del intervalo de integración se eligen tales que $\Phi_b = 1 - \Phi_a$, entonces la varianza del estimador es cero (ver Anexo 6.2).

Por definición, la varianza del estimador es igual a

$$\begin{aligned} V_{WTM}(\hat{t}_\pi^*) &= E((\hat{t}_\pi^*)^2) - t^2 \\ &= T^2 E\left(\sum_{i=1}^n y_i^*\right)^2 - t^2 \\ &= T^2 \left\{ \sum_{i=1}^n E(y_i^2) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n E(y_i y_j) + 2n|a| \sum_{i=1}^n E(y_i) + n^2 a^2 \right\} - t^2. \end{aligned} \quad (17)$$

donde los valores esperados de la última igualdad están dados por

$$\sum_{i=1}^n E(y_i) = \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(v) dv. \quad (18)$$

$$\sum_{i=1}^n E(y_i^2) = \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \{\Phi^{-1}(v)\}^2 dv. \quad (19)$$

$$\begin{aligned} \sum_{i=1}^{n-1} \sum_{j=i+1}^n E(y_i y_j) &= \frac{1}{T} \left\{ 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b - kT} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \right. \\ &\quad + \sum_{k=0}^{n/2-1} \int_{\Phi_a + kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \\ &\quad \left. + \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b - kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right\}. \end{aligned} \quad (20)$$

Sustituyendo los valores esperados (18), (19), (20) en la ecuación (17) y simplificando se obtiene

$$\begin{aligned} V_{WTM}(\hat{t}_\pi^*) &= 2T \left\{ \int_{\Phi_a}^{\Phi_b} \{\Phi^{-1}(v)\}^2 dv \right. \\ &\quad + 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b - kT} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \\ &\quad + \sum_{k=0}^{n/2-1} \int_{\Phi_a + kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \\ &\quad \left. - \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b - kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right\} \\ &\quad - \{t - 2|a|(\Phi_b - \Phi_a)\}^2. \end{aligned}$$

Un caso interesante se presenta cuando $\Phi_b = 1 - \Phi_a$, ya que en tal situación:

$$\begin{aligned}
V_{WTM}(\hat{t}_\pi^*) &= 2T \left\{ \int_{\Phi_a}^{1-\Phi_a} \{\Phi^{-1}(v)\}^2 dv \right. \\
&\quad + 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&\quad - 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&\quad \left. - \int_{\Phi_a}^{1-\Phi_a} \Phi^{-1}(v) \Phi^{-1}(v) dv \right\} \\
&\quad - \{t - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= - \{t - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= - \{2|a|(\Phi_b - \Phi_a) - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= 0.
\end{aligned} \tag{21}$$

Por lo tanto, bajo el modelo propuesto, $V_{WTM}(\hat{t}_\pi^*) = 0$ si se cumplen tres condiciones: *i*) que el tamaño de muestra n sea par, condición que se puede controlar, *ii*) que $\Phi_b = 1 - \Phi_a$, lo cual implica que se están truncando los valores extremos (cola izquierda y derecha) de la distribución subyacente con igual probabilidad y *iii*) $\Phi^{-1}(v) = -\Phi^{-1}(1-v)$, la cual se cumplirá si la variable subyacente a la inversa de la función de medida es simétrica, como lo es el caso de la distribución normal, la t -de student, la uniforme y otras.

Geoméricamente, el area bajo la función de medida $f^*(p)$ es igual al área del rectángulo de base $1 - 2\Phi_a$ y altura $|2a|$, lo cual se puede ver fácilmente si se corta la figura 1 en sus dos mitades simétricas y se da vuelta a una de ellas, ambas se pueden acoplar perfectamente formando un rectángulo con las dimensiones anteriores. Cuando se efectua un muestreo sistemático, la altura del primer punto seleccionado más la altura del $\frac{n}{2} + 1$ será igual a la altura del rectángulo, $|2a|$; lo mismo sucede con el segundo punto y el punto $\frac{n}{2} + 2$, y así sucesivamente, hasta los puntos $\frac{n}{2} - 1$ y n . Por lo tanto, el estimador \hat{t}_π^* siempre será igual a $(1 - 2\Phi_a)|2a|$.

Cuando se considera la función de medida $f(p) = y = y^* - |a|$, es fácil ver que la varianza del estimador del total $\hat{t}_\pi = T \sum_{i \in s_r} y_i$ es

$$V_{WTM}(\hat{t}_\pi) = V_{WTM}(\hat{t}_\pi^*),$$

y en caso de que $Y \sim N(\mu, \sigma^2)$ en lugar de que $Y \sim N(0, 1)$, se tiene que

$$V_{WTM}(\hat{t}_{\pi, \mu, \sigma^2}) = \sigma^2 V_{WTM}(\hat{t}_\pi) = 0.$$

4.1.3. Varianza del estimador del total considerando un error en el modelo

Un problema más general consiste en suponer que hay un error en el modelo propuesto, es decir, que la variable de interés es

$$\tilde{Y} = y^* + \epsilon = f^*(p) + \epsilon(p) \quad (22)$$

donde ϵ es una variable aleatoria que cumple:

- i. $E_M [\epsilon(p)] = 0$
- ii. $V_M [\epsilon(p)] = \sigma_\epsilon^2(p)$.

Supóngase que el objetivo es estimar el total $t = \sum_{i=1}^N \tilde{y}_i$, entonces el estimador de Horvitz-Thompson es

$$\begin{aligned} \tilde{t}_\pi &= T \sum_{i=1}^n \tilde{Y}_i \\ &= T \sum_{i=1}^n (y_i^* + \epsilon_i) = \hat{t}_\pi^* + T \sum_{i=1}^n \epsilon_i. \end{aligned}$$

Por otro lado, la esperanza condicional dada la muestra es

$$\begin{aligned} E_M [\tilde{t}_\pi | u_r] &= E_M \left[\hat{t}_\pi^* + T \sum_{i=1}^n \epsilon_i | u_r \right] \\ &= E_M [\hat{t}_\pi^* | u_r] + T \sum_{i=1}^n E_M [\epsilon_i | u_r] \\ &= \hat{t}_\pi^* + 0 \\ &= \hat{t}_\pi^*, \end{aligned}$$

$$\begin{aligned} V_M [\tilde{t}_\pi | u_r] &= V_M \left[\hat{t}_\pi^* + T \sum_{i=1}^n \epsilon_i | u_r \right] \\ &= V_M [\hat{t}_\pi^* | u_r] + V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] + 2Cov \left[\hat{t}_\pi^*, T \sum_{i=1}^n \epsilon_i | u_r \right] \\ &= 0 + V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] + 0 \\ &= V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right]. \end{aligned}$$

Sustituyendo estas dos ecuaciones en el resultado

$$V_{SY} (\tilde{t}_\pi) = V_{SY} (E_M [\tilde{t}_\pi | u_r]) + E_{SY} (V_M [\tilde{t}_\pi | u_r])$$

se obtiene

$$V_{SY}(\tilde{t}_\pi) = V_{SY}(\hat{t}_\pi^*) + E_{SY} \left(V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] \right).$$

Bajo el WTM, con n par y $\Phi_b = 1 - \Phi_a$ se llega finalmente a

$$\begin{aligned} V_{WTM}(\tilde{t}_\pi) &= 0 + E_{SY} \left(V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] \right) \\ &= E_{SY} \left(V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] \right). \end{aligned} \tag{23}$$

4.1.4. Estimador de la varianza

Dependiendo de los supuestos que se hagan acerca de $V_M[\epsilon(p)] = \sigma_\epsilon^2(p)$, se pueden proponer diferentes aproximaciones para estimar la varianza dada por la ecuación (23). Por ejemplo:

- Si se supone que los errores son independientes y $V_M[\epsilon(p)] = \sigma_\epsilon^2$, entonces

$$\hat{V}_{WTM,SI}(\tilde{t}_\pi) = T^2 (1 - f) n s_\epsilon^2$$

donde

$\tilde{y}_1, \tilde{y}_1, \dots, \tilde{y}_N$ son realizaciones de \tilde{Y} ,

$$\begin{aligned} \hat{\epsilon}_i &= e_i = \tilde{y}_i - y_i^* \\ \bar{e} &= \frac{\sum_{i=1}^n e_i}{n} \\ s_\epsilon^2 &= \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n - 1}. \end{aligned}$$

- Si se supone que los errores no son independientes, se propone usar alguna aproximación de las señaladas en la sección 3. En virtud de los resultados obtenidos para diversos ejercicios que hemos realizado, se sugiere estimar la varianza mediante

$$\hat{V}_{WTM,Sdi}(\tilde{t}_\pi) = T^2 (1 - f) n \sum_{i=3}^n \frac{(e_i - 2e_{i-1} + e_{i-2})^2}{6(n - 2)}.$$

4.2. WTM población finita

En muchas situaciones la población de interés no es de naturaleza continua, sino que se trata de una población finita y por lo tanto discreta, por ejemplo: número de habitantes, matrícula

en un nivel escolar, número de personas con determinado padecimiento, etc. En este caso se propone el modelo siguiente (figura 2) cuando el tamaño de la población N es par:

$$\begin{aligned}
 y_i^* &= E(y^* | p \in I_i) \\
 &= |a| + E(y | p \in I_i) \\
 &= |a| + \frac{1}{\Delta} \int_{I_i} f(p) dp, \quad i = 1, 2, \dots, N
 \end{aligned} \tag{24}$$

donde

$$\begin{aligned}
 y^* &= f^*(p) = f(p) + |a|, \\
 f(p) &= \begin{cases} \Phi^{-1}(p) & \text{si } \Phi_a \leq p \leq 1 - \Phi_a \\ \Phi^{-1}(2 - p) & \text{si } 1 + \Phi_a \leq p \leq 2 - \Phi_a \end{cases},
 \end{aligned}$$

$\Phi^{-1}(p)$ denota la función inversa de una $N(0, 1)$ para el cuantil p , es decir, si $y = \Phi^{-1}(p)$ entonces $p = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$,

$$\Phi(a) = \Phi_a,$$

$$a = \Phi^{-1}(\Phi_a),$$

$\Delta = \frac{2(1-2\Phi_a)}{N}$ corresponde a la longitud del intervalo I_i ,

$$I_i = \begin{cases} [\Phi_a + (i - 1)\Delta, \Phi_a + i\Delta] & \text{si } 1 \leq i \leq N/2 \\ [3\Phi_a + (i - 1)\Delta, 3\Phi_a + i\Delta] & \text{si } N/2 + 1 \leq i \leq N \end{cases}.$$

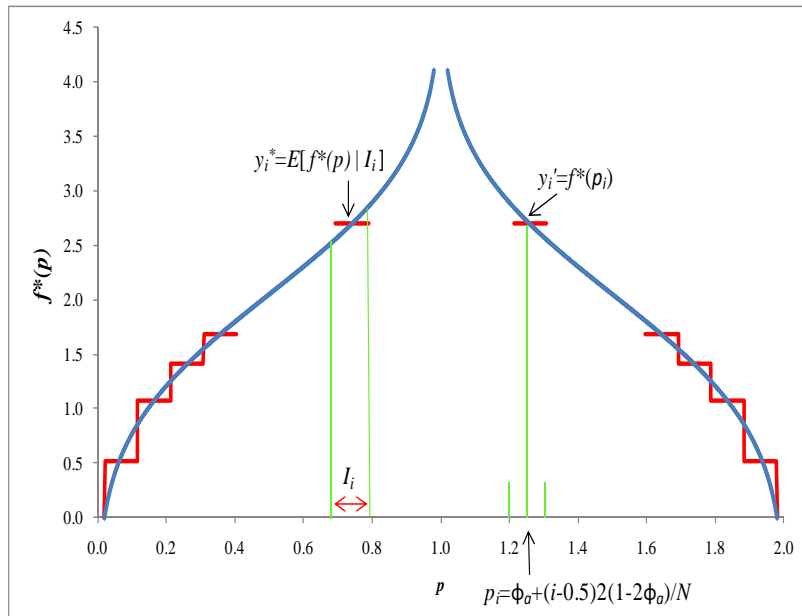


Figura 2: Modelo población finita, y_i^*

Se puede demostrar que en general se cumple

$$\int \Phi^{-1}(p)dp = -\frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}[\Phi^{-1}(p)]^2 \right\}.$$

Si $\Phi_a \leq p \leq 1 - \Phi_a$, o equivalentemente si $1 \leq i \leq N/2$, entonces

$$\begin{aligned} y_i &= \frac{1}{\Delta} \int_{I_i} f(p)dp \\ &= \frac{1}{\Delta} \int_{I_i} \Phi^{-1}dp \\ &= -\frac{1}{\Delta\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}[\Phi^{-1}(p)]^2 \right\} \Big|_{\Phi_a+(i-1)\Delta}^{\Phi_a+i\Delta} \\ &= \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2}[\Phi^{-1}(\Phi_a + (i-1)\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2}[\Phi^{-1}(\Phi_a + i\Delta)]^2 \right\} \right]. \end{aligned}$$

Análogamente, si $1 + \Phi_a \leq p \leq 2 - \Phi_a$, o bien si $N/2 + 1 \leq i \leq N$

$$\begin{aligned} y_i &= \frac{1}{\Delta} \int_{I_i} f(p)dp \\ &= \frac{1}{\Delta\sqrt{2\pi}} \exp \left\{ -\frac{1}{2}[\Phi^{-1}(2-p)]^2 \right\} \Big|_{3\Phi_a+(i-1)\Delta}^{3\Phi_a+i\Delta} \\ &= \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2}[\Phi^{-1}(2 - 3\Phi_a - i\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2}[\Phi^{-1}(2 - 3\Phi_a - (i-1)\Delta)]^2 \right\} \right]. \end{aligned}$$

Por lo tanto,

$$y_i = \begin{cases} \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2}[\Phi^{-1}(\Phi_a + (i-1)\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2}[\Phi^{-1}(\Phi_a + i\Delta)]^2 \right\} \right] & \text{si } 1 \leq i \leq \frac{N}{2} \\ \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2}[\Phi^{-1}(2 - 3\Phi_a - i\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2}[\Phi^{-1}(2 - 3\Phi_a - (i-1)\Delta)]^2 \right\} \right] & \text{si } \frac{N}{2} + 1 \leq i \leq N \end{cases} \quad (25)$$

Una alternativa al modelo (24) es que el valor de p asociado a la unidad i sea el punto medio del intervalo I_i , es decir,

$$y'_i = f^*(\bar{p}_i)$$

donde

$$\bar{p}_i = \begin{cases} \Phi_a + (i - \frac{1}{2})\Delta & \text{si } 1 \leq i \leq \frac{N}{2} \\ \bar{p}_i = 3\Phi_a + (i - \frac{1}{2})\Delta & \text{si } \frac{N}{2} + 1 \leq i \leq N. \end{cases}$$

Los valores y_i^* y y'_i serán muy similares (figura 2), las diferencias serán un poco más grandes en los extremos, cercanos a Φ_a ó $2 - \Phi_a$, y en la parte central, cercanos a $1 - \Phi_a$ ó $1 + \Phi_a$. Por ejemplo, si $N = 20$ y $\Phi_a = 0.02$, entonces $y_5^* = 1.9328$, $y'_5 = 1.9331$, $y_8^* = 2.6995$ y $y'_8 = 2.6971$, mientras que $y_{10}^* = 3.5838$ y $y'_{10} = 3.5446$.

4.2.1. Estimador del total

El parámetro objetivo t^* corresponde al total de y_i^* , es decir,

$$\begin{aligned}
t^* &= \sum_{i=1}^N y_i^* \\
&= N|a| + \sum_{i=1}^{N/2} \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + (i-1)\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + i\Delta)]^2 \right\} \right] \\
&\quad + \sum_{i=N/2+1}^N \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - i\Delta)]^2 \right\} - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - (i-1)\Delta)]^2 \right\} \right] \\
&= N|a| + \frac{1}{\Delta\sqrt{2\pi}} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a)]^2 \right\} - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + \frac{N}{2}\Delta)]^2 \right\} \right. \\
&\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - \frac{N}{2}\Delta)]^2 \right\} + \exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - N\Delta)]^2 \right\} \right].
\end{aligned}$$

Ya que $\Phi_a + \frac{N}{2}\Delta = 1 - \Phi_a$, $2 - 3\Phi_a - \frac{N}{2}\Delta = 1 - \Phi_a$ y $2 - 3\Phi_a - N\Delta = \Phi_a$, finalmente se llega a

$$t^* = N|a|.$$

Un estimador insesgado de t^* bajo un muestreo sistemático, está dado mediante

$$\begin{aligned}
\hat{t}_\pi^* &= T \sum_{i \in s_r} y_i^* \\
&= T \sum_{i \in s_r} (|a| + y_i)
\end{aligned}$$

donde

$$\begin{aligned}
T &= \frac{N}{n} \\
s_r &= r, r + T, r + 2T, \dots, r + (n-1)T.
\end{aligned}$$

Sustituyendo los valores de la expresión (25) y simplificando (ver Anexo 6.3. Estimador de \hat{t}_π^* , WTM población finita) se obtiene

$$\hat{t}_\pi^* = N|a|. \tag{26}$$

4.2.2. Varianza del estimador del total

Ya que el total estimado es igual a una constante entonces

$$V_{WTM}(\hat{t}_\pi^*) = 0.$$

Cuando se consideran los valores poblaciones sin trasladar, y_i , es fácil ver que

$$V_{WTM}(\hat{t}_\pi) = V_{WTM}(\hat{t}_\pi^*) = 0,$$

y en caso de que $Y \sim N(\mu, \sigma^2)$ en lugar de que $Y \sim N(0, 1)$, se tiene que

$$V_{WTM}(\hat{t}_{\pi, \mu, \sigma^2}) = \sigma^2 V_{WTM}(\hat{t}_{\pi}) = 0.$$

4.2.3. Varianza del estimador del total considerando un error en el modelo

Al igual que en el caso continuo, un problema más general consiste en suponer que hay un error en el modelo, es decir,

$$\tilde{Y}_i = y_i^* + \epsilon_i \quad (27)$$

donde ϵ es una variable aleatoria que cumple:

- i. $E_M[\epsilon_i] = 0$
- ii. $V_M[\epsilon_i] = \sigma_{\epsilon_i}^2$.

En este caso

$$\begin{aligned} \tilde{t}_{\pi} &= T \sum_{i=1}^n \tilde{Y}_i \\ &= \hat{t}_{\pi}^* + T \sum_{i=1}^n \epsilon_i. \end{aligned}$$

Análogamente al caso de una población continua, bajo el WTM, con n par y $\Phi_b = 1 - \Phi_a$ se llega a

$$\begin{aligned} V_{WTM}(\tilde{t}_{\pi}) &= 0 + E_{SY} \left(V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] \right) \\ &= E_{SY} \left(V_M \left[T \sum_{i=1}^n \epsilon_i | u_r \right] \right). \end{aligned} \quad (28)$$

4.2.4. Estimador de la varianza

Para aproximar la ecuación (28) se pueden hacer supuestos acerca de $V_M[\epsilon_i]$. Por ejemplo:

- Si los errores son independientes y $V_M[\epsilon_i] = \sigma_{\epsilon}^2$, entonces

$$\hat{V}_{WTM, SI}(\tilde{t}_{\pi}) = T^2 (1 - f) n s_{\epsilon}^2 \quad (29)$$

- Si los errores no son independientes, se propone

$$\hat{V}_{WTM, Sdi}(\tilde{t}_{\pi}) = T^2 (1 - f) n \sum_{i=3}^n \frac{(e_i - 2e_{i-1} + e_{i-2})^2}{6(n-2)}. \quad (30)$$

donde $\tilde{y}_1, \tilde{y}_1, \dots, \tilde{y}_N$ son realizaciones de \tilde{Y} , e_i y s_{ϵ}^2 se definen de manera análoga al caso continuo.

5. Experimento empírico y simulaciones

A continuación se analizará de manera empírica el comportamiento de los estimadores de la varianza (29) y (30), para el caso en que la población finita proviene de una población infinita con distribución normal.

Sea z_1, z_2, \dots, z_N una población finita generada a partir de una $N(0, 1)$ con $z \in [-a, a]$, donde $\Phi^{-1}(\Phi_a) = a$. Esta se ordena en forma de “cola de ballena” colocando a las unidades de la siguiente manera: $\tilde{y}_1 = z_{(1)}$, $\tilde{y}_2 = z_{(3)}$, $\tilde{y}_3 = z_{(5)}, \dots, \tilde{y}_{N/2} = z_{(N-1)}$, $\tilde{y}_{N/2+1} = z_{(N)}, \dots, \tilde{y}_{N-2} = z_{(6)}$, $\tilde{y}_{N-1} = z_{(4)}$, $\tilde{y}_N = z_{(2)}$, donde $z_{(1)}, \dots, z_{(N)}$ denotan las observaciones ordenadas de menor a mayor.

Bajo el WTM para poblaciones finitas el error observado estará dado por $e_i = \tilde{y}_i - y_i$ con y_i dada por la expresión (25).

En el experimento se simularon $K = 100$ poblaciones independientes y finitas de $N = 2, 200$ unidades bajo una distribución normal estándar truncada con $\Phi_a = 0.01$. Para cada población y cada una de las T posibles muestras de tamaño n se calcularon:

- La varianza del estimador, parámetro $V_{SY}(\hat{t}_\pi)$.
- Los estimadores tres señalados en la sección 3 y los dos propuestos en la sección anterior: $\hat{V}_{Fdi}(\hat{t}_\pi)$, $\hat{V}_{Sdi}(\hat{t}_\pi)$, $\hat{V}_{GC}(\hat{t}_\pi)$, $\hat{V}_{WTM,SI}(\hat{t}_\pi)$ y $\hat{V}_{WTM,Sdi}(\hat{t}_\pi)$.

La simulación tiene dos objetivos: i) comparar el desempeño de cada uno de los cinco estimadores con respecto a la varianza del estimador (parámetro), es decir, comparar el sesgo y ii) comparar los cinco estimadores en términos de su estabilidad.

En la figura 3 se presenta el promedio de las varianzas sobre las 100 poblaciones y las T muestras, $\frac{\sum \left(\frac{\sum_{r=1}^T \sqrt{\hat{V}_{sr}}(\hat{t}_\pi) / T}{K} \right)$, según el tamaño de muestra. Adicionalmente se grafica el promedio de la desviación estándar del estimador (parámetro), $\frac{\sum (\sqrt{V_{SY}}(\hat{t}_\pi))}{K} = \frac{\sum \left(\sqrt{\sum_{r=1}^T (\hat{t}_{sr} - \hat{t})^2 / T} \right)}{K}$ sobre las 100 poblaciones. En términos del sesgo, el mejor estimador corresponde a $\hat{V}_{WTM,Sdi}(\hat{t}_\pi)$, el cual decrece conforme el tamaño de muestra aumenta.

Como una medida de la estabilidad de las estimaciones, en la figura 4 se muestra el promedio de la desviación estándar de la varianza estimada, $\frac{\sum (\sqrt{V(\hat{V}(\hat{t}_\pi))})}{K} = \frac{\sum \left(\sqrt{\sum_{r=1}^T \left(\hat{V}_{SY} - \frac{\sum_{r=1}^T \hat{V}_{SY}}{T} \right)^2 / T} \right)}{K}$. Se observa que el estimador con una menor dispersión también es $\hat{V}_{WTM,Sdi}(\hat{t}_\pi)$.

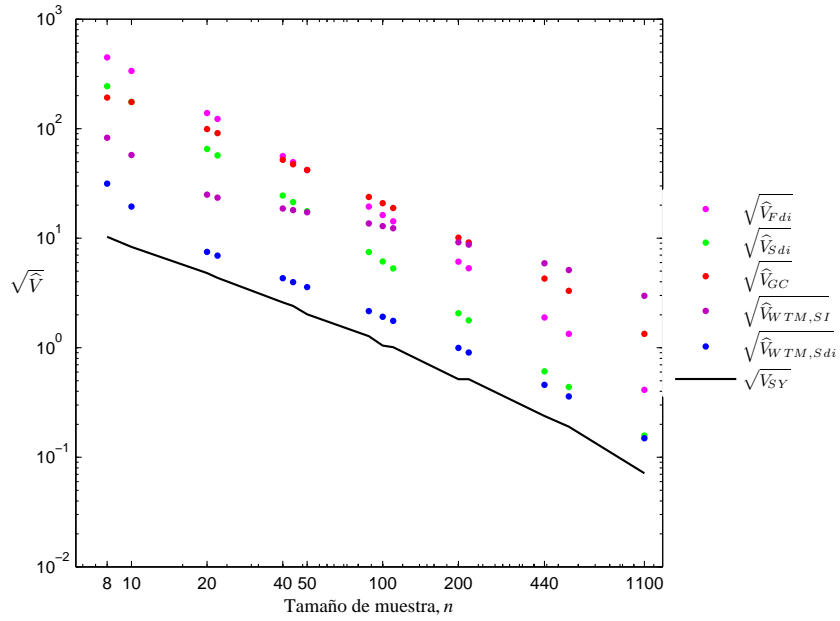


Figura 3: Poblaciones generadas $\sim N(0,1)$ y ordenadas en forma de “cola de ballena”. Desviación estándar poblacional y estimada promedio (escala logarítmica). La línea continua muestra el promedio de la desviación estándar poblacional sobre las 100 poblaciones generadas. Las líneas punteadas representan el promedio de la desviación estándar estimada, según tamaño de muestra, sobre las 100 poblaciones generadas y las T muestras; la varianza para cada muestra se estima mediante: \widehat{V}_{Fdi} , \widehat{V}_{Sdi} , \widehat{V}_{GC} , $\widehat{V}_{WTM,SI}$ y $\widehat{V}_{WTM,Sdi}$.

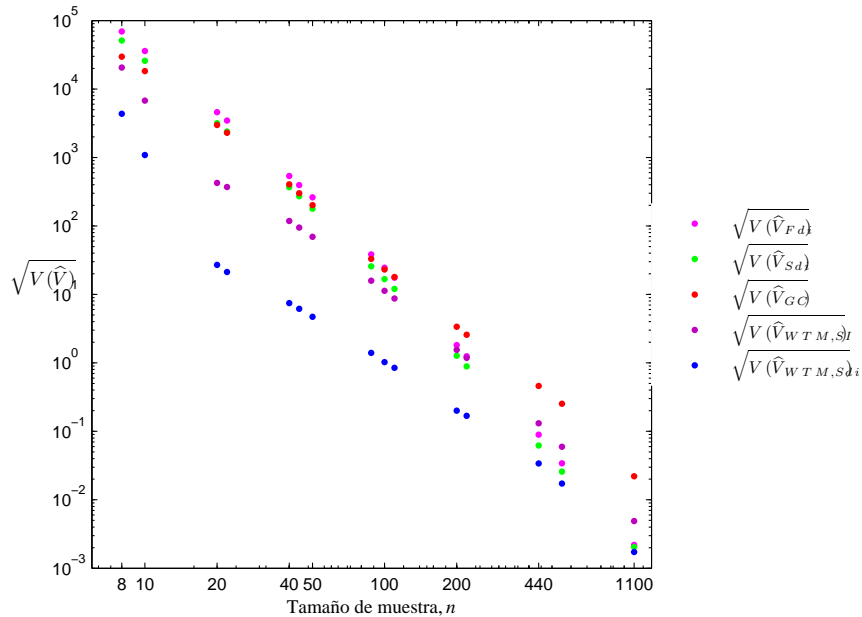


Figura 4: Poblaciones generadas $\sim N(0,1)$ y ordenadas en forma de “cola de ballena”. Promedio de la desviación estándar de la varianza, según tamaño de muestra y estimador, sobre las 100 poblaciones generadas (escala logarítmica). La varianza se estima usando: \widehat{V}_{Fdi} , \widehat{V}_{Sdi} , \widehat{V}_{GC} , $\widehat{V}_{WTM,SI}$ y $\widehat{V}_{WTM,Sdi}$.

Adicionalmente al trabajo presentado, es de interés explorar el desempeño de los estimadores propuestos para funciones de medida resultantes cuando la población subyacente no es la normal, así como en el caso de datos reales en lugar de simulados. En ambos casos es importante considerar alguna medida de distancia o diferencia de alejamiento de la función de medida analizada respecto al WTM.

6. Anexos

6.1. Relación entre \widehat{V}_{Sdi} y \widehat{V}_{GC}

Se había visto que

$$\begin{aligned}\widehat{V}_{GC}(\hat{t}_\pi) &= N^2(1-f)\frac{1}{n^2}\alpha(q)[3C_0 - 4C_1 + C_2] \\ &= N^2(1-f)\frac{1}{n^2}\alpha(q)\left[3\sum_{i=1}^n y_i^2 - 4\sum_{i=2}^n y_i y_{i-1} + \sum_{i=3}^n y_i y_{i-2}\right] \\ \widehat{V}_{Sdi}(\hat{t}_\pi) &= N^2(1-f)\frac{1}{n}\sum_{i=3}^n \frac{(y_i - 2y_{i-1} + y_{i-2})^2}{6(n-2)}.\end{aligned}\tag{31}$$

Desarrollando el trinomio al cuadrado en (31) se tiene que

$$\begin{aligned}\sum_{i=3}^n (y_i - 2y_{i-1} + y_{i-2})^2 &= \sum_{i=3}^n (y_i^2 + 4y_{i-1}^2 + y_{i-2}^2 - 4y_i y_{i-1} - 4y_{i-1} y_{i-2} + 2y_i y_{i-2}) \\ &= \sum_{i=1}^n 6y_i^2 - y_1^2 - y_2^2 - 4y_1^2 - 4y_n^2 - y_{n-1}^2 - y_n^2 \\ &\quad - \sum_{i=2}^n 8y_i y_{i-1} + 4y_2 y_1 + 4y_n y_{n-1} + \sum_{i=3}^n 2y_i y_{i-2} \\ &= 6\sum_{i=1}^n y_i^2 - 8\sum_{i=2}^n y_i y_{i-1} + 2\sum_{i=3}^n y_i y_{i-2} \\ &\quad - 5y_1^2 - y_2^2 - y_{n-1}^2 - 5y_n^2 + 4y_2 y_1 + 4y_n y_{n-1}.\end{aligned}$$

Haciendo $c = -5y_1^2 - y_2^2 - y_{n-1}^2 - 5y_n^2 + 4y_2 y_1 + 4y_n y_{n-1}$, y sustituyendo en (31), se tiene

$$\begin{aligned}\widehat{V}_{Sdi}(\hat{t}_\pi) &= N^2(1-f)\frac{1}{6n(n-2)}\left[6\sum_{i=1}^n y_i^2 - 8\sum_{i=2}^n y_i y_{i-1} + 2\sum_{i=3}^n y_i y_{i-2} + c\right] \\ &= N^2(1-f)\frac{2}{6n(n-2)}\left[3\sum_{i=1}^n y_i^2 - 4\sum_{i=2}^n y_i y_{i-1} + \sum_{i=3}^n y_i y_{i-2} + c/2\right] \\ &= N^2(1-f)\frac{1}{3n(n-2)}\left[3\sum_{i=1}^n y_i^2 - 4\sum_{i=2}^n y_i y_{i-1} + \sum_{i=3}^n y_i y_{i-2}\right] + N^2(1-f)\frac{c}{6n(n-2)},\end{aligned}$$

si $n \approx n - 2$ y $\alpha(q = -1/2) = 1/3$, entonces

$$\widehat{V}_{Sdi}(\hat{t}_\pi) \approx \widehat{V}_{GC}(\hat{t}_\pi) - N^2(1-f) \frac{1}{6n^2} [5y_1^2 + y_2^2 + y_{n-1}^2 + 5y_n^2 - 4y_2y_1 - 4y_ny_{n-1}].$$

Finalmente si $2\sqrt{5} \approx 4$, la última expresión se puede escribir como

$$\widehat{V}_{Sdi}(\hat{t}_\pi) \approx \widehat{V}_{GC}(\hat{t}_\pi) - N^2(1-f) \frac{1}{6n^2} [(\sqrt{5}y_1 - y_2)^2 + (\sqrt{5}y_n - y_{n-1})^2]. \quad (32)$$

6.2. Derivación de $V_{WTM}(\hat{t}_\pi^*)$, población continua

A continuación se derivará una expresión para $V_{WTM}(\hat{t}_\pi^*)$ considerando que el tamaño de muestra n es par, y se demostrará que cuando $\Phi_b = 1 - \Phi_a$, entonces la varianza del estimador es cero.

Por definición, la varianza del estimador es igual a

$$\begin{aligned} V_{WTM}(\hat{t}_\pi^*) &= E((\hat{t}_\pi^*)^2) - t^2 \\ &= T^2 E\left(\sum_{i=1}^n y_i^*\right)^2 - t^2 \\ &= T^2 E\left(\sum_{i=1}^n (y_i + |a|)\right)^2 - t^2 \\ &= T^2 E\left(\sum_{i=1}^n y_i + n|a|\right)^2 - t^2 \\ &= T^2 \left\{ E\left(\sum_{i=1}^n y_i\right)^2 + 2n|a| \sum_{i=1}^n E(y_i) + n^2 a^2 \right\} - t^2 \\ &= T^2 \left\{ \sum_{i=1}^n E(y_i^2) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n E(y_i y_j) + 2n|a| \sum_{i=1}^n E(y_i) + n^2 a^2 \right\} - t^2. \end{aligned} \quad (33)$$

Enseguida se desarrollarán los valores esperados de la última igualdad en la ecuación (33), considerando que n es par.

$$\sum_{i=1}^n E(y_i) = \sum_{i=1}^{n/2} E(y_i) + \sum_{i=n/2+1}^n E(y_i).$$

Obteniendo el valor esperado del primer sumando

$$\begin{aligned}
\sum_{i=1}^{n/2} E(y_i) &= \sum_{i=1}^{n/2} \int_0^1 \Phi^{-1}(\Phi_a + [u + i - 1]T) du \\
&= \sum_{i=1}^{n/2} \frac{1}{T} \int_{\Phi_a + [i-1]T}^{\Phi_a + iT} \Phi^{-1}(v) dv \\
&= \frac{1}{T} \int_{\Phi_a}^{\Phi_a + \frac{n}{2}T} \Phi^{-1}(v) dv \\
&= \frac{1}{T} \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(v) dv
\end{aligned}$$

donde se hizo el cambio de variable $v = \Phi_a + [u + i - 1]T$.

Análogamente, efectuando el cambio de variable $v = 2 - 2(1 - \Phi_b) + \Phi_a + [u + i - 1]T = 2\Phi_b - \Phi_a - [u + i - 1]T$, se tiene que el valor esperado del segundo sumando es

$$\begin{aligned}
\sum_{i=n/2+1}^n E(y_i) &= \sum_{i=n/2+1}^n \int_0^1 \Phi^{-1}(2\Phi_b - \Phi_a - [u + i - 1]T) du \\
&= \sum_{i=n/2+1}^n -\frac{1}{T} \int_{2\Phi_b - \Phi_a - [i-1]T}^{2\Phi_b - \Phi_a - iT} \Phi^{-1}(v) dv \\
&= -\frac{1}{T} \int_{2\Phi_b - \Phi_a - \frac{n}{2}T}^{2\Phi_b - \Phi_a - nT} \Phi^{-1}(v) dv \\
&= -\frac{1}{T} \int_{\Phi_b}^{\Phi_a} \Phi^{-1}(v) dv \\
&= \frac{1}{T} \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(v) dv.
\end{aligned}$$

Por lo tanto,

$$\sum_{i=1}^n E(y_i) = \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(v) dv.$$

Por otra parte, el valor esperado de la variable al cuadrado es:

$$\begin{aligned}
\sum_{i=1}^n E(y_i^2) &= 2 \sum_{i=1}^{n/2} \int_0^1 \{\Phi^{-1}(\Phi_a + [u + i - 1]T)\}^2 du \\
&= \frac{2}{T} \sum_{i=1}^{n/2} \int_{\Phi_a + [i-1]T}^{\Phi_a + iT} \{\Phi^{-1}(v)\}^2 dv \\
&= \frac{2}{T} \int_{\Phi_a}^{\Phi_a - [n/2]T} \{\Phi^{-1}(v)\}^2 dv \\
&= \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \{\Phi^{-1}(v)\}^2 dv.
\end{aligned}$$

Para obtener el valor esperado de $y_i y_j$ se deben de considerar 3 casos, dependiendo de los valores que tomen los subíndices i y j ,

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n E(y_i y_j) = s1 + s2 + s3 \quad (34)$$

donde:

$$s1 = \sum_{i=1}^{n/2-1} \sum_{j=i+1}^{n/2} \int_0^1 \Phi^{-1}(\Phi_a + [u + i - 1]T) \Phi^{-1}(\Phi_a + [u + j - 1]T) du \quad (35)$$

$$s2 = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^n \int_0^1 \Phi^{-1}(\Phi_a + [u + i - 1]T) \Phi^{-1}(2\Phi_b - \Phi_a - [u + j - 1]T) du \quad (36)$$

$$s3 = \sum_{i=n/2+1}^{n-1} \sum_{j=i+1}^n \int_0^1 \Phi^{-1}(2\Phi_b - \Phi_a - [u + i - 1]T) \Phi^{-1}(2\Phi_b - \Phi_a - [u + j - 1]T) du. \quad (37)$$

Haciendo los cambios de variable $v = \Phi_a + [u + i - 1]T$ y $j = i + k$ en (35) y simplificando se llega a que

$$\begin{aligned} s1 &= \sum_{i=1}^{n/2-1} \sum_{k=1}^{n/2-i} \frac{1}{T} \int_{\Phi_a + [i-1]T}^{\Phi_a + iT} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \\ &= \frac{1}{T} \left\{ \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_a + T} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \right. \\ &\quad + \sum_{k=1}^{n/2-2} \int_{\Phi_a + T}^{\Phi_a + 2T} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \\ &\quad \left. + \dots + \sum_{k=1}^1 \int_{\Phi_a + [n/2-2]T}^{\Phi_a + [n/2-1]T} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \right\} \\ &= \frac{1}{T} \left\{ \int_{\Phi_a}^{\Phi_a + [n/2-1]T} \Phi^{-1}(v) \Phi^{-1}(v + T) dv \right. \\ &\quad + \int_{\Phi_a}^{\Phi_a + [n/2-2]T} \Phi^{-1}(v) \Phi^{-1}(v + 2T) dv \\ &\quad \left. + \dots + \int_{\Phi_a}^{\Phi_a + T} \Phi^{-1}(v) \Phi^{-1}(v + [n/2 - 1]T) dv \right\} \\ &= \frac{1}{T} \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_a + [n/2-k]T} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv \\ &= \frac{1}{T} \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b - kT} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv. \end{aligned} \quad (38)$$

Si $v = \Phi_a + [u + i - 1]T$ y $j = i + k$, entonces $2\Phi_b - \Phi_a - [u + j - 1]T = 2\Phi_b - kT - v$.
Sustituyendo en (36) se tiene lo siguiente

$$\begin{aligned}
s_2 &= \sum_{i=1}^{n/2} \sum_{k=n/2+1-i}^{n-i} \frac{1}{T} \int_{\Phi_a+[i-1]T}^{\Phi_a+iT} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - kT) dv \\
&= \frac{1}{T} \left\{ \sum_{k=n/2}^{n-1} \int_{\Phi_a}^{\Phi_a+T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - kT) dv \right. \\
&\quad + \sum_{k=n/2-1}^{n-2} \int_{\Phi_a+T}^{\Phi_a+2T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - kT) dv \\
&\quad + \cdots + \left. \sum_{k=1}^{n/2} \int_{\Phi_a+[n/2-1]T}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - kT) dv \right\} \\
&= \frac{1}{T} \left\{ \int_{\Phi_a}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2]T) dv \right. \\
&\quad + \int_{\Phi_a+T}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 - 1]T) dv \\
&\quad + \cdots + \int_{\Phi_a+[n/2-1]T}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - T) dv \\
&\quad + \int_{\Phi_a}^{\Phi_a+[n/2-1]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 + 1]T) dv \\
&\quad + \int_{\Phi_a}^{\Phi_a+[n/2-2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 + 2]T) dv \\
&\quad + \cdots + \left. \int_{\Phi_a}^{\Phi_a+T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n - 1]T) dv \right\} \\
&= \frac{1}{T} \left\{ \sum_{k=1}^{n/2} \int_{\Phi_a+[k-1]T}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 - k + 1]T) dv \right. \\
&\quad + \left. \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_a+[n/2-k]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 + k]T) dv \right\} \\
&= \frac{1}{T} \left\{ \sum_{k=0}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_a+[n/2]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 - k]T) dv \right. \\
&\quad + \left. \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_a+[n/2-k]T} \Phi^{-1}(v) \Phi^{-1}(2\Phi_b - v - [n/2 + k]T) dv \right\}
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{T} \left\{ \sum_{k=0}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \right. \\
&\quad \left. + \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right\}. \tag{39}
\end{aligned}$$

Si en la ecuación (37) se hacen los cambios de variable $v = 2\Phi_b - \Phi_a - [u + j - 1]T$ y $j = i + k$, se tiene que

$$\begin{aligned}
s3 &= \sum_{i=n/2+1}^{n-1} \sum_{k=1}^{n-i} -\frac{1}{T} \int_{2\Phi_b-\Phi_a-[i-1]T}^{2\Phi_b-\Phi_a-[i]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \\
&= \sum_{i=n/2+1}^{n-1} \sum_{k=1}^{n-i} \frac{1}{T} \int_{2\Phi_b-\Phi_a-[i]T}^{2\Phi_b-\Phi_a-[i-1]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \\
&= \frac{1}{T} \left\{ \sum_{k=1}^{n/2-1} \int_{2\Phi_b-\Phi_a-[n/2+1]T}^{2\Phi_b-\Phi_a-[n/2]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \right. \\
&\quad + \sum_{k=1}^{n/2-2} \int_{2\Phi_b-\Phi_a-[n/2+2]T}^{2\Phi_b-\Phi_a-[n/2+1]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \\
&\quad + \cdots + \left. \sum_{k=1}^1 \int_{2\Phi_b-\Phi_a-[n-1]T}^{2\Phi_b-\Phi_a-[n-2]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \right\} \\
&= \frac{1}{T} \left\{ \int_{2\Phi_b-\Phi_a-[n-1]T}^{2\Phi_b-\Phi_a-[n/2]T} \Phi^{-1}(v) \Phi^{-1}(v - T) dv \right. \\
&\quad + \int_{2\Phi_b-\Phi_a-[n-2]T}^{2\Phi_b-\Phi_a-[n/2]T} \Phi^{-1}(v) \Phi^{-1}(v - 2T) dv \\
&\quad + \cdots + \left. \int_{2\Phi_b-\Phi_a-[n/2+1]T}^{2\Phi_b-\Phi_a-[n/2]T} \Phi^{-1}(v) \Phi^{-1}(v - [n/2 - 1]T) dv \right\} \\
&= \frac{1}{T} \sum_{k=1}^{n/2-1} \int_{2\Phi_b-\Phi_a-[n-k]T}^{2\Phi_b-\Phi_a-[n/2]T} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \\
&= \frac{1}{T} \sum_{k=1}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(v - kT) dv \\
&= \frac{1}{T} \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(v + kT) dv. \tag{40}
\end{aligned}$$

Sustituyendo (38), (39) y (40) en la ecuación (34)

$$\begin{aligned} \sum_{i=1}^{n-1} \sum_{j=i+1}^n E(y_i y_j) &= \frac{1}{T} \left\{ 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \right. \\ &\quad + \sum_{k=0}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \\ &\quad \left. + \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right\}. \end{aligned}$$

Sustituyendo los valores esperados en (33) y efectuando operaciones se obtiene

$$\begin{aligned} V_{WTM}(\hat{t}_\pi^*) &= T^2 \left\{ \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \{\Phi^{-1}(v)\}^2 dv \right. \\ &\quad + \frac{2}{T} \left[2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \right. \\ &\quad + \sum_{k=0}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \\ &\quad \left. + \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right] \\ &\quad \left. + 2n|a| \frac{2}{T} \int_{\Phi_a}^{\Phi_b} \Phi^{-1}(v) dv + n^2 a^2 \right\} - t^2, \end{aligned}$$

simplicando se llega finalmente a

$$\begin{aligned} V_{WTM}(\hat{t}_\pi^*) &= 2T \left\{ \int_{\Phi_a}^{\Phi_b} \{\Phi^{-1}(v)\}^2 dv \right. \\ &\quad + 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\ &\quad + \sum_{k=0}^{n/2-1} \int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv \\ &\quad \left. - \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv \right\} \\ &\quad - \{t - 2|a|(\Phi_b - \Phi_a)\}^2. \end{aligned} \tag{41}$$

Cuando $\Phi_b = 1 - \Phi_a$ y considerando que en el caso de la distribución normal $\Phi^{-1}(v) =$

$-\Phi^{-1}(1-v)$, entonces

$$\begin{aligned}
\int_{\Phi_a+kT}^{\Phi_b} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v + kT) dv &= \int_{\Phi_a+kT}^{1-\Phi_a} \Phi^{-1}(v) \Phi^{-1}(1-v+kT) dv \\
&= - \int_{\Phi_a+kT}^{1-\Phi_a} \Phi^{-1}(1-v) \Phi^{-1}(1-v+kT) dv \\
&= \int_{1-\Phi_a-kT}^{\Phi_a} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&= - \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv,
\end{aligned}$$

$$\begin{aligned}
\int_{\Phi_a}^{\Phi_b-kT} \Phi^{-1}(v) \Phi^{-1}(\Phi_a + \Phi_b - v - kT) dv &= \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(1-v-kT) dv \\
&= - \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(1-v) \Phi^{-1}(1-v-kT) dv \\
&= \int_{1-\Phi_a}^{\Phi_a+kT} \Phi^{-1}(v) \Phi^{-1}(v-kT) dv \\
&= \int_{1-\Phi_a-kT}^{\Phi_a} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&= - \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv.
\end{aligned}$$

Luego, reemplazando las últimas dos integrales en (41) se llega a

$$\begin{aligned}
V_{WTM}(\hat{t}_\pi^*) &= 2T \left\{ \int_{\Phi_a}^{1-\Phi_a} \{\Phi^{-1}(v)\}^2 dv \right. \\
&\quad + 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&\quad - 2 \sum_{k=1}^{n/2-1} \int_{\Phi_a}^{1-\Phi_a-kT} \Phi^{-1}(v) \Phi^{-1}(v+kT) dv \\
&\quad \left. - \int_{\Phi_a}^{1-\Phi_a} \Phi^{-1}(v) \Phi^{-1}(v) dv \right\} \\
&\quad - \{t - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= - \{t - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= - \{2|a|(\Phi_b - \Phi_a) - 2|a|(\Phi_b - \Phi_a)\}^2 \\
&= 0.
\end{aligned}$$

6.3. Estimador de \hat{t}_π^* , WTM población finita

Un estimador insesgado de t^* bajo un muestreo sistemático, está dado por la expresión

$$\begin{aligned}\hat{t}_\pi^* &= T \sum_{i \in s_r} y_i^* \\ &= T \sum_{i \in s_r} (|a| + y_i)\end{aligned}$$

donde

$$T = \frac{N}{n}$$

$$s_r = r, r + T, r + 2T, \dots, r + (n - 1)T.$$

Luego,

$$\begin{aligned}\hat{t}_\pi^* &= T \sum_{i=1}^n (|a| + y_i) \\ &= Tn|a| + T \sum_{i=1}^n y_i \\ &= N|a| + T \sum_{i=1}^{n/2} y_i + T \sum_{i=n/2+1}^n y_i \\ &= N|a| + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=1}^{n/2} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T - 1]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T]\Delta)]^2 \right\} \right] + \\ &\quad + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=n/2+1}^n \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - [r + (i - 1)T]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a + [r + (i - 1)T - 1]\Delta)]^2 \right\} \right] \\ &= N|a| + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=1}^{n/2} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T - 1]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T]\Delta)]^2 \right\} \right] + \\ &\quad + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=1}^{n/2} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a - [r + (n/2 + i - 1)T]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(2 - 3\Phi_a + [r + (n/2 + i - 1)T - 1]\Delta)]^2 \right\} \right].\end{aligned}$$

Ya que $2 - 3\Phi_a - [r + (n/2 + i - 1)T]\Delta = 1 - \Phi_a - [r + (i - 1)T]\Delta$, $2 - 3\Phi_a - [r + (n/2 + i - 1)T - 1]\Delta =$

$1 - \Phi_a - [r + (i - 1)T - 1]\Delta$ y $\Phi^{-1}(1 - v) = -\Phi^{-1}(v)$, entonces

$$\begin{aligned} \hat{t}_\pi^* &= N|a| + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=1}^{n/2} \left[\exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T - 1]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [\Phi^{-1}(\Phi_a + [r + (i - 1)T]\Delta)]^2 \right\} \right] + \\ &\quad + \frac{T}{\sqrt{2\pi}\Delta} \sum_{i=1}^{n/2} \left[\exp \left\{ -\frac{1}{2} [-\Phi^{-1}(\Phi_a - [r + (i - 1)T]\Delta)]^2 \right\} + \right. \\ &\quad \left. - \exp \left\{ -\frac{1}{2} [-\Phi^{-1}(\Phi_a + [r + (i - 1)T - 1]\Delta)]^2 \right\} \right] \\ &= N|a|. \end{aligned}$$

Referencias

- [1] Y. G. Berger. A variance estimator for systematic sampling from a deliberately ordered population. *Communications in Statistics. Theory and Methods*, 34:1533–1541, 2005.
- [2] W. G. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Annals of Mathematics Statistics*, 17:164–177, 1946.
- [3] W. G. Cochran. *Sampling Techniques*. Wiley, 1977.
- [4] A. W. Fuller. *Sampling Statistics*. Wiley, John and Sons, 2009.
- [5] M. García-Fiñana. *Muestreo Sistemático en R*. Tesis doctoral, Universidad de Cantabria, España, 2000.
- [6] M. García-Fiñana and L. M. Cruz-Orive. Improved variance prediction for systematic sampling on R. *Statistics*, 38(3):243–272, 2004.
- [7] W. Gautschi. Some remarks on systematic sampling. *The Annals of Mathematical Statistics*, 28(2):385–394, 1957.
- [8] H. J. G. Gundersen. The smooth fractionator. *Journal of Microscopy*, 207:191–210, 2002.
- [9] R. Iachan. Systematic sampling: A critical review. *International Statistical Review*, 50(3):293–303, 1982.
- [10] K. Kiêu. Three lectures on sistematic geometric sampling. *Memoirs*, 13. University of Aarhus, 1997.

- [11] K. Kiêu and M. Mora. Stereological estimation of mean volume: precision of three simple sampling designs. Reporte técnico, Unité Mathématiques et Informatique Appliquées, Institut National de la Recherche Agronomique, Versailles, 2005.
- [12] K. Kiêu, S. Souchet, and J. Istas. Precision of systematic sampling and transitive methods. *Statistical Planning and Inference*, 77:263–279, 1999.
- [13] W. G. Madow and L. H. Madow. On the theory of systematic sampling I. *Annals of Mathematics Statistics*, 15(1):1–24, 1944.
- [14] P.C. Mahalanobis. Recent experiments in statistical sampling in the indian statistical institute. *Royal Statistical Society*, 109(4):325–370, 1946.
- [15] G. Matheron. Les variables régionalisées et leur estimation. Masson, Paris, 1965.
- [16] M. H. Quenouille. Problems in plane sampling. *Annals of Mathematics Statistics*, 20(3):355–375, 1949.
- [17] C. E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, 1992.
- [18] P. V. Sukhatme, B. V. Sukhatme, and Asok. *Sampling Theory of Surveys with Applications*. Iowa State University Press, 1984.
- [19] K. M. Wolter. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79:781–790, 1984.
- [20] K. M. Wolter. *Introduction to Variance Estimation*. Springer-Verlag, 2007.
- [21] F. Yates. Systematic sampling. *Philosophical Transactions of the Royal Society of London*, 241(834):345–377, 1948.
- [22] F. Yates. *Sampling Methods for Censuses and Surveys*. Charles Griffin and Company Limited, 1981.