UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

VARIANZA DEL ESTIMADOR DEL TOTAL BAJO MUESTREO SISTEMÁTICO DE ALGUNAS POBLACIONES SIMÉTRICAS

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
MONICA TINAJERO BRAVO

DIRECTOR DE LA TESIS:
DRA. GUILLERMINA ESLAVA GÓMEZ,
FACULTAD DE CIENCIAS

MIEMBROS DEL COMITÉ TUTOR
DRA. ELIANE R. RODRIGUES, INSTITUTO DE MATEMÁTICAS

DR. IGNACIO MÉNDEZ RAMÍREZ, INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y SISTEMAS

MÉXICO, D. F. ABRIL DE 2014.

AGRADECIMIENTOS

La realización de mis estudios de doctorado, que culminan con la presentación de esta tesis, ha sido posible por el apoyo de diferentes instancias, cada una de las cuales ha jugado un papel fundamental.

En primer lugar agradezco a la Universidad Nacional Autónoma de México, en la cual me he formado, que a través de su infraestructura, personal y profesores, me dió la oportunidad de continuar en la aventura del conocimiento y la investigación. También agradezco al Consejo Nacional de Ciencia y Tecnología por otorgarme una beca para llevar a cabo estos estudios.

Una persona muy importante en el desarrollo de esta tesis es la Dra. Guillermina Eslava Gómez, quien con su dirección, orientación y apoyo hizo posible llevar a buen fin esta tarea, le agradezco mucho el tiempo y el esfuerzo dedicados a lo largo de estos años.

También doy las gracias a los sinodales Dr. Luis Manuel Cruz Orive, Dr. Martín Humberto Félix Medina, Dra. Eliane Regina Rodrigues y Dr. Emilio López Escobar, quienes con sus valiosos comentarios y sugerencias enriquecieron este trabajo. En especial, al Dr. Luis Manuel Cruz Orive, quien además de ser una guía a distancia, también me orientó y apoyó durante las visitas de trabajo académico que realicé a la Universidad de Cantabria. Asimismo, agradezco el apoyo del Dr. Ignacio Méndez Ramírez, miembro del Comité Tutor.

No puedo dejar de mencionar a mis compañeros y amigos del doctorado Jessica, Eunice y Ricardo, con quienes he compartido no sólo conocimientos sino bellos momentos.

De igual manera reitero mi gratitud a mis papás, quienes han sido incondicionales en las decisiones que he tomado, así como a mis hermanos quienes siempre me han alentado a continuar.

Por último, a mi esposo e hijas, quienes me han acompañado en este camino, agradezco su paciencia durante las jornadas largas de trabajo. Paco, muchas gracias por todo el apoyo y la comprensión que me has brindado. Zairis, Vale y Cassy, gracias por ser el motor que me impulsa a superarme y emprender nuevas tareas.

Varianza del estimador del total bajo muestreo sistemático de algunas poblaciones simétricas

Mónica Tinajero Bravo
Posgrado en Ciencias Matemáticas, UNAM
Tutor: Dra. Guillermina Eslava Gómez

Abril de 2014

Resumen

El muestreo sistemático puede generar estimaciones de la media o del total poblacional más precisas que el muestreo aleatorio simple, sin embargo su eficiencia dependerá de las propiedades de la población y el orden de la misma. En este trabajo se obtuvieron dos resultados de investigación. El primero de ellos se refiere a tres condiciones suficientes para que la varianza del estimador del total bajo muestreo sistemático de una población fija sea nula. En el segundo resultado se derivó una expresión para la varianza del muestreo sistemático y del aleatorio simple, considerando una superpoblación que tiene una distribución uniforme.

En el contexto de poblaciones fijas y finitas, las condiciones suficientes, aunque no necesarias, para que la varianza sea cero son: i) el tamaño de muestra y el de la población deben de ser pares, ii) la población esté ordenada de manera equilibrada, como aquí se denomina y iii) los valores ordenados de la población sean simétricos con respecto a su media. En el caso de una población continua o función, se encontraron condiciones análogas para que la integral de la función en un intervalo cerrado se estime sin error. El conjunto de muestras posibles bajo muestreo sistemático, cuando la población se ordena como en ii, es el mismo que se obtendría bajo lo que se denomina muestreo sistemático equilibrado (del inglés *balanced systematic sampling*) cuando la población se ordena de manera creciente.

El segundo resultado se obtuvo siguiendo el enfoque de superpoblaciones, suponiendo una distribución uniforme. Se derivó la expresión de la varianza total del predictor del total bajo a) muestro sistemático considerando una población ordenada de manera equilibrada, b) muestreo sistemático considerando un orden creciente y c) muestreo aleatorio simple. Demostrándose que la varianza del primer diseño es menor que la del segundo diseño y ésta a su vez menor que la del muestreo aleatorio simple. Mediante simulaciones, se mostró que lo anterior también se cumple para las

distribuciones de Laplace y normal, así como para diferentes valores que se consideraron del parámetro de forma de la distribución normal generalizada, cuya función de densidad es simétrica con respecto a la media.

En la práctica, las condiciones i y ii se podrían controlar, la tercera está dada y es difícil que se cumpla; por ello se comparó empíricamente la eficiencia del muestreo sistemático, con respecto a la del aleatorio simple usando cinco conjuntos de datos. En cuatro de las cinco poblaciones analizadas, el muestreo sistemático cuando la población se ordenó de manera equilibrada fue más preciso, que cuando la población se ordenó de forma creciente; en la quinta población esto también sucedió para tamaños de muestra pequeños, mientras que para la mayoría de los tamaños de muestra grandes, el muestreo sistemático bajo el orden creciente fue más preciso que el orden equilibrado. Ambos diseños fueron más eficientes que el aleatorio simple.

Bajo muestreo sistemático no existen estimadores insesgados de la varianza, usando los cinco conjuntos de datos, se compararon tres estimadores que se sugieren en la literatura, no se observó un mejor estimador en términos del sesgo.

Palabras clave: Estadísticas de orden, fraccionador suave, muestreo sistemático equilibrado, orden de la población, población simétrica, superpoblación, varianza.

Índice general

Capítulo 1

Introducción

El muestreo sistemático es un método de selección que consiste en muestrear puntos igualmente espaciados a lo largo de una secuencia de elementos, o bien de un dominio continuo, iniciando en un punto determinado de forma aleatoria. En este trabajo sólo se tratará el caso del muestreo sistemático unidimensional y no se abordará el muestreo sistemático en más dimensiones.

Este diseño se usa tanto en poblaciones finitas, como por ejemplo las poblaciones humanas, así como también en poblaciones continuas, como por ejemplo un tumor o la materia gris de un cerebro humano, en cuyo caso puede ser de interés la estimación del volumen o el área.

Una población finita de N elementos se denota por el conjunto $U = \{u_1, \ldots, u_N\}$, donde u_i corresponde al $i-$ésimo elemento, por simplicidad u_i generalmente se representa sólo por su etiqueta i, por lo que la población finita queda como $U = \{1, \ldots, N\}$; el valor de la variable de intéres y del $i-$ésimo elemento se representa por y_i, $i = 1, \ldots, N$. En este tipo de poblaciones puede ser de interés estimar, con base en una muestra, la suma o total poblacional

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i.$$

La forma básica de seleccionar una muestra sistemática de una población finita consiste en elegir, con probabilidad $1/T$, un elemento r entre los primeros T de la población, y a partir de éste, tomar a los elementos situados cada T unidades. Al número r se le denomina arranque aleatorio y a $T \in \mathbb{N}$ se le conoce como salto o

intervalo muestral, con $N = nT + c$ y c un entero que satisface $0 \leq c < T$. Bajo este esquema de selección, la muestra queda determinada por las unidades etiquetadas como

$$s_r = \{r, r+T, \ldots, r+(n_s-1)T\},$$

donde n_s es una variable aleatoria que toma los valores $n+1$ si $r \leq c$ ó n si $c < r \leq T$ y corresponde al tamaño de la muestra.

Cuando la población es continua, la característica de interés toma un valor en cada punto del espacio muestral y por ende se puede tratar como una función f. En este caso, un parámetro de interés puede ser la integral de la función,

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

donde $f : \mathbb{R} \mapsto \mathbb{R}^+$, es una función acotada no aleatoria, de cuadrado integrable en un intervalo $A = [a, b]$, esto es $\int_A |f(x)|^2 \, dx < \infty$, y fuera de A es cero.

Si t se estima con base en una muestra sistemática s_r, $f(x)$ se observará en aquellos puntos tales que $x \in s_r$, con

$$s_r = \{r + (j-1)T, j \in \mathbb{Z}\} \cap A,$$

donde r es una realización de una variable uniforme en $[0, T]$ y el tamaño de muestra n_s es una variable aleatoria.

En la práctica, el muestreo sistemático es atractivo principalmente por dos razones: i) por su simplicidad operativa, y ii) porque, para cierta clase de poblaciones, genera estimadores más eficientes que el muestreo aleatorio simple. Su eficiencia depende de las propiedades de la población y del orden de la misma. Algunas veces el muestreo sistemático es más preciso, ya que refleja cualquier estratificación implícita que exista en el orden del marco de muestreo, sin embargo, es necesario tener información sobre la estructura de la población para usarlo de manera eficiente. La primer investigación al respecto fue hecha por Madow y Madow [34] quienes obtuvieron expresiones para la varianza del muestreo sistemático que dependen de las autocorrelaciones entre los valores de la población.

Como lo han señalado Cochran [6], Yates [60] e Iachan [25], 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, para estimar la varianza del estimador del total de manera insesgada. No obstante, algunas de las aproximaciones propuestas en la literatura, para estimar la varianza a partir de una muestra de una población finita, se han obtenido bajo el enfoque de superpoblaciones, suponiendo que los valores de la población $y_1, y_2, \ldots, y_N$ provienen o son realizaciones de una superpoblación de variables aleatorias $Y_1, Y_2, \ldots, Y_N$.

El enfoque de superpoblaciones también se usa para hacer comparaciones teóricas de la varianza del muestreo sistemático respecto a otros esquemas de selección. Uno de los trabajos en este sentido, y que constituyó la base para desarrollos posteriores, corresponde a Cochran [6], en el que se supone que los elementos de la población están correlacionados. Generalizaciones al modelo que considera Cochran se encuentran en artículos como Quenouille [44], Gautschi [15], Hájek [23] e Iachan [26]. Otro tipo de poblaciones muy estudiadas en la literatura son aquellas que presentan tendencia, ver por ejemplo Cochran [7], Gautschi [15], Singh, Jindal y Garg [51], Bellhouse y Rao [2], así como Bartolucci y Montanari [1].

En el caso de poblaciones continuas, uno de los primeros trabajos se debe a Yates [60], quien usa el muestreo sistemático para a) estimar el área bajo la función de densidad gaussiana o normal, y b) estimar la proporción de una línea que posee cierto atributo, la línea está divida en secciones, algunas de éstas poseen el atributo y el resto no. Con base en el planteamiento de Yates, Moran [38] obtiene una expresión exacta para la varianza de estimador del área cuando se conoce la función a integrar. Posteriormente, Matheron [35], [36] con sus métodos transitivos, proporciona una alternativa para derivar aproximaciones de la varianza del muestreo sistemático, aun cuando la función no se conoce precisamente. Reexaminando la teoría de Matheron, Souchet [52], Kiêu [32], Kiêu et al. [33] y García-Fiñana y Cruz-Orive [14] derivan expresiones para la varianza del estimador que dependen de la suavidad de la función, suponiendo que ésta es suave a trozos.

Con la finalidad de reducir el error de muestreo, autores como Sethi [48, 49], Murthy [39], Sampath et al. [46], han propuesto el uso de una variante del muestreo sistemático: el muestreo sistemático equilibrado (BSS del inglés *balanced systematic sampling*). Sethi [48, 49] demostró que cuando se eligen dos unidades mediante BSS de una población ordenada de manera creciente, la varianza del estimador del total

es la mínima, en comparación con cualquier otra manera de elegir dos unidades. También para reducir varianza, Gundersen [20] describe un algoritmo de selección llamado el fraccionador suave (del inglés *smooth fractionator*), el cual consiste en fraccionar un objeto en unidades, ordenarlas de manera equilibrada y aplicar un muestreo sistemático. El conjunto de muestras que se obtiene bajo este algoritmo es el mismo que el que se obtiene bajo BSS cuando las unidades están ordenadas de manera creciente.

La historia de esta tesis es como sigue. Inicialmente se quería estimar la varianza del estimador del total en el muestreo sistemático, por lo que además de explorar los estimadores ya existentes en el caso de poblaciones finitas, usando algunos conjuntos de datos, se analizó una aproximación que se usa en la Estereología, en el contexto de poblaciones continuas. Esto nos llevó al orden simétrico de la población que se propone en el fraccionador suave, y por consiguiente, al muestreo sistemático equilibrado. Primero fue de interés encontrar la varianza del estimador del área de la inversa de la función de distribución normal, debido a la importancia de esta distribución; haciendo algunos supuestos, se encontró que la varianza era cero, lo cual nos guió para encontrar condiciones para que la varianza del muestreo sistemático, a partir de una población fija, sea nula. Siguiendo con esta manera de ordenar a la población, mediante el enfoque de superpoblaciones, se derivó la varianza suponiendo una distribución uniforme. Se intentó lo mismo para las distribuciones normal y de Laplace, pero como las expresiones involucradas en la varianza son muy complicadas, el comportamiento de la varianza se analizó mediante aproximaciones, así como a través de un estudio de simulación.

Los resultados de investigación obtenidos en este trabajo están contenidos en los capítulos 4 y 5. La tesis en su conjunto está organizada de la siguiente manera. En el capítulo 2 se describen con mayor precisión los principales resultados de los trabajos citados en esta introducción. En el capítulo 3 se describe el muestreo sistemático a partir de una población finita, se presenta el estimador de Horvitz-Thompson del total de la población, así como su varianza; también se presenta la descomposición de la variación total en la población. En el caso de una población continua, se describe el algoritmo de selección y se da el estimador del área bajo el gráfico de una función continua, acotada y fija, así como la varianza del estimador; por último, se describe el muestreo sistemático equilibrado. En el capítulo 4 se proporcionan tres condiciones

suficientes bajo las cuales la estimación del total en el caso de una población finita, o bien del área, en el caso continuo, tienen varianza cero. En el capítulo 5 se obtiene explícitamente, o por simulación, la varianza del predictor del total bajo muestreo sistemático desde el enfoque de superpoboblaciones, suponiendo tres modelos que pertenecen a la familia de la distribución normal generalizada: uniforme, Laplace y normal o gaussiana, así como dos órdenes: el equilibrado y el creciente, que se usa frecuentemente en la práctica del muestreo porque se supone que existe una relación entre la variable de interés y y la variable auxiliar x, que se usa para ordenar la población. En el capítulo 6 se presentan tres de las aproximaciones propuestas en la literatura para estimar la varianza del sistemático. Estos tres estimadores se comparan empíricamente en el capítulo 7 usando cinco conjuntos de datos, también se evalúa empíricamente la eficiencia del muestreo sistemático en relación con la del aleatorio simple. En el capítulo 8 se ofrecen algunas conclusiones, así como futuras líneas de investigación. Finalmente, en el anexo se presenta la demostración más detallada de la varianza bajo muestreo equilibrado suponiendo la distribución uniforme.

Se sometió un artículo (Tinajero et al. [57]) relacionado con el trabajo del capítulo 4, el cual ya fue aceptado, en el que en la Proposición 1 se da una condición suficiente para que la varianza del estimador bajo muestreo sistemático sea cero en el contexto de poblaciones continuas. En el caso de poblaciones discretas, suponiendo que el tamaño de muestra y de poblacion son pares y el intervalo muestral es entero, en la Proposición 3 se establece una condición análoga; en el Corolario 3 se demuestra que la condición que se proporciona es equivalente a las que se encontraron en la Proposición 4.1 de esta tesis.

Es muy importante aclarar que todas las proposiciones que se presentan corresponden a contribuciones de carácter original y por ello se ofrecen las demostraciones respectivas.

Capítulo 2

Antecedentes

El enfoque de superpoblaciones se ha usado en el muestreo sistemático para i) comparar las varianzas del estimador usado en este diseño con las de estimadores usados en otros esquemas, y ii) aproximar la varianza del sistemático, al no existir un estimador insesgado de la misma. Uno de los primeros trabajos que considera una superpoblación se debe a Cochran [6], donde se supone un modelo en el que los elementos de la población están correlacionados serialmente y se compara la varianza del muestreo sistemático con otros esquemas de selección. Posteriormente, Quenouille [44] y Hájek [23] generalizan el modelo planteado por Cochran. Gautschi [15] e Iachan [26] también parten del trabajo de Cochran y comparan el muestreo sistemático de varios arranques aleatorios con el de uno solo.

Poblaciones que presentan una tendencia también han sido estudiadas. Por ejemplo Cochran [7] considera tres modelos para la tendencia y proporciona en cada caso un estimador de la varianza. Gautschi [15] supone un modelo lineal y obtiene la varianza bajo muestreo sistemático, estratificado y aleatorio simple. Singh, Jindal y Garg [51] comparan la varianza del muestreo sistemático, el sistemático modificado y el aleatorio simple considerando una tendencia lineal. Bellhouse y Rao [2] también suponen que la población presenta una tendencia lineal, comparan el muestreo sistemático con el sistemático modificado, el sistemático equilibrado, el sistemático centrado y un ajuste en el estimador de la media. Bartolucci y Montanari [1] suponen un modelo lineal y descomponen la varianza del muestreo sistemático en una parte que dependende del componente sistemático del modelo y otra que depende del componente estocástico del mismo. Suponiendo una superpoblación que presenta

una tendencia parabólica, Sampath et al [46] comparan tres estrategias de muestreo: sistemático, sistemático equilibrado y sistemático modificado, cambiando el peso que tienen cuatro unidades en la muestra en el estimador del total.

En el caso de poblaciones continuas, un trabajo pionero se debe a Yates [60], quien trata el muestreo sistemático aplicado a la función de densidad gaussiana y una línea dividida en secciones. Moran [38] encuentra una expresión para la varianza del estimador del área cuando se conoce la función a integrar. Matheron [35], [36] deriva aproximaciones de la varianza del muestreo sistemático, aunque se desconozca la función. Souchet [52], Kiêu [32], Kiêu et al. [33] y García-Fiñana y Cruz-Orive [14] encuentran expresiones para la varianza del estimador bajo condiciones más generales.

Gundersen [20], describe un procedimiento de selección que denomina el fraccionador suave. Las muestras que se obtienen son las mismas que bajo lo que Sethi [48] llama muestreo sistemático equilibrado. Sethi [48], [49] demostró que este es el método óptimo de selección si $n = 2$.

A continuación se presentan los resultados de los trabajos citados anteriormente.

Cochran [6] estudió poblaciones cuyos elementos están correlacionados serialmente, donde la correlación entre Y_i y Y_{i+u}, ρ_u, es una función positiva monótona decreciente que depende sólo de u. Explícitamente, supone que los elementos de la población finita y_i, $i = 1, 2, \ldots, N = nT$ con $T \in \mathbb{N}$ son generados de una superpoblación $Y_1, \ldots, Y_N$ bajo el modelo m según el cual

a) $E_m(Y_i) = \mu$

b) $V_m(Y_i) = \sigma^2$

c) $E_m(Y_i - \mu)(Y_{i+u} - \mu) = \sigma^2 \rho_u$,

donde $0 \leq \rho_v \leq \rho_u$ con $v > u$, $E_m(\cdot)$ y $V_m(\cdot)$ denotan el valor esperado y la varianza bajo el modelo, respectivamente. Encontró que el valor esperado de la varianza de la media bajo muestreo sistemático está dado por

$$\begin{aligned}\sigma^2_{SY} &= E_m\left[V_{SY}\left(\bar{y}\right)\right] \\ &= \frac{\sigma^2}{n}\left(1 - \frac{1}{T}\right)\left[1 - \frac{2}{N(T-1)}\sum_{u=1}^{N-1}(N-u)\rho_u + \frac{2T}{n(T-1)}\sum_{u=1}^{n-1}(n-u)\rho_{Tu}\right],\end{aligned}$$

donde $\bar{y} = \sum_{i \in s_r} y_i/n$ es el estimador de la media poblacional, s_r denota la muestra seleccionada, $n = n_s$ es el tamaño de muestra, el cual es fijo en este caso, y $V_{SY}(\bar{y})$ es la varianza de $\bar{y}$ bajo muestreo sistemático.

Análogamente, derivó expresiones para el valor esperado de la varianza cuando se considera un muestreo estratificado, donde se elige una unidad en cada uno de los n estratos en los que se divide la población, σ^2_{ST}, y para el valor esperado de la varianza bajo un muestreo aleatorio simple con tamaño de muestra n, σ^2_{SI}, las cuales también son funciones lineales de las correlaciones:

$$\sigma^2_{ST} = E_m\left[V_{ST}(\bar{y})\right] = \frac{\sigma^2}{n}\left(1 - \frac{1}{T}\right)\left[1 - \frac{T}{T(T-1)}\sum_{u=1}^{T-1}(T-u)\rho_u\right],$$

$$\sigma^2_{SI} = E_m\left[V_{SI}(\bar{y})\right] = \frac{\sigma^2}{n}\left(1 - \frac{1}{T}\right)\left[1 - \frac{2}{N(N-1)}\sum_{u=1}^{N-1}(N-u)\rho_u\right].$$

Mostró que no se puede establecer un resultado general acerca de la eficiencia relativa del muestreo sistemático respecto a los otros dos diseños, a menos que se hagan supuestos adicionales sobre la forma específica de la población. Demostró que si además se supone que el correlograma es convexo (cóncavo hacia arriba), esto es,

$$\rho_{i-1} + \rho_{i+1} - 2\rho_i \geq 0 \quad \text{con} \quad i = 2, \ldots, nT - 2,$$

entonces $\sigma^2_{SY} \leq \sigma^2_{ST} \leq \sigma^2_{SI}$.

Obtuvo una expresión para σ^2_{SY} cuando la correlación ρ_u es de tipo lineal y exponencial. Finalmente señala que en el caso exponencial, $\rho_u = e^{-\lambda u}$, si n y T son grandes una estimación de $\sigma^2_{SY} = n^{-1}\sigma^2\left[1 - 2/(T\lambda) + 2/\exp(T\lambda - 1)\right]$ se obtiene usando la varianza muestral en lugar de σ^2, estimando ρ_T, $u = T$, mediante la correlación entre los elementos en la muestra y despejando para obtener una estimación de λ.

Quenouille [44] señala que las expresiones para las varianzas que encontró Cochran [6] se pueden obtener bajo las condiciones más generales. Supone un modelo en dos etapas en el que cada y_i es una muestra de una superpoblación que satisface:

a) $E_m(Y_i) = \mu_i$ y $V_m(Y_i) = \sigma_i^2$, $i = 1, \ldots, N$

b) $E_m(\mu_i) = \mu$ y $V_m(\mu_i) = \sigma^2$, $i = 1, \ldots, N$

c) $E_m(\mu_i - \mu)(\mu_j - \mu) = \sigma^2 \rho_{ij}$, $i, j = 1, \ldots, N$ con $i \neq j$

d) $\rho_u = \frac{1}{N-u} \sum_{i=1}^{N-u} \rho_{i,i+u}$.

En este caso las varianzas que obtuvo Cochran [6], σ^2_{SY}, σ^2_{ST} y σ^2_{SI}, aumentan en la cantidad $(nN)^{-1}(1 - 1/T)\sum_{i=1}^{N} \sigma_i^2$. También extiende los resultados para el muestreo en dos dimensiones, obtiene expresiones para la precisión del muestreo sistemático, del muestreo estratificado y del muestreo aleatorio simple, o bien combinaciones de dos de estos diseños, por ejemplo sistemático en una dirección y estratificado en la otra.

Hájek [23] consideró un modelo más general que el de Cochran [6], que incorpora una variable auxiliar x_i, $i = 1, \ldots, N$, cuya especificación es la siguiente

a) $E_m(Y_i) = \mu x_i$

b) $E_m(Y_i - \mu x_i)^2 = \sigma^2 x_i^2$

c) $E_m(Y_i - \mu x_i)(Y_{i+u} - \mu x_{i+u}) = \sigma^2 x_i x_{i+u} \rho_u$,

donde ρ_u es una función convexa que depende de u. Mostró que bajo muestreo sistemático con probabilidad proporcional a la medida de tamaño x_i y tamaño de muestra fijo, el estimador de Horvitz-Thompson $\hat{t}_\pi = \sum_{i \in s} y_i/\pi_i$ tiene varianza mínima, donde $\pi_i = nx_i/(x_1 + x_2 + \cdots + x_N)$, $i = 1, \ldots, N$, corresponde a la probabilidad de inclusión del $i-$ésimo elemento en la muestra.

Gautschi [15] compara el muestreo sistemático cuando se eligen k arranques aleatorios, con el muestreo sistemático con un arranque aleatorio, ambos diseños con el mismo tamaño de muestra, $n = kl$ con $k, l \in \mathbb{N}$. Si $\sigma^2_{SY}(k)$ denota el valor esperado de la varianza de la media bajo un muestreo sistemático con k arranques aleatorios, demuestra que $\sigma^2_{SY} \leq \sigma^2_{SY}(k)$ tanto en el caso de una población con tendencia lineal, como en el de una población cuya correlación entre elementos es una función decreciente y convexa, como la que supone Cochran [6].

Iachan [26] considera que la población proviene de un proceso estacionario de segundo orden, específicamente, supone que los valores de la población son observaciones en $i = 1, \ldots, N$ generados de un proceso $Y(i)$ tal que

a) $E_m[Y(i)] = 0$

b) $E_m[Y(i)Y(i+u)] = \sigma^2\rho_u$, $i, u \in \mathbb{R}$.

Demuestra que si se cumplen

i) $\sum_{u=1}^{\infty} |\rho_u| < \infty$

ii) $\rho_u \geq \rho_v$ con $u < v$,

entonces, $\sigma^2_{SY}(k) \leq \sigma^2_{SI}$ para todo $k \geq 1$ cuando $N \to \infty$. La condición sobre convexidad de ρ_u, que se supone en Cochran [6], no es necesaria para que el muestreo sistemático con k arranques aleatorios sea más preciso que el muestreo aleatorio simple. Si además de las dos condiciones anteriores, se cumple que

iii) ρ_u es convexa,

iv) ρ_u es continuamente diferenciable en $(0, \infty)$, con derivada ρ'_u, y $\sum_{u=1}^{\infty} |u\rho'_u| < \infty$,

entonces $\sigma^2_{SY}(k_1) \leq \sigma^2_{SY}(k_2)$ si $k_1 \leq k_2$.

Otra clase de poblaciones son aquéllas que presentan cierta tendencia. En Cochran [7] §8.11, se presentan tres modelos para superpoblaciones que suponen una tendencia más un efecto aleatorio:

$$Y_i = \mu_i + \epsilon_i, \ i = 1, \ldots, N,$$

donde μ_i es una función de i, $E(\epsilon_i) = 0$, $E(\epsilon_i^2) = \sigma_i^2$ y $E(\epsilon_i\epsilon_j) = 0$ $(i \neq j)$.

El primer modelo corresponde a la población en orden aleatorio, en cuyo caso $\mu_i = \mu$ es una constante; el segundo se refiere a efectos de estratificación, en el cual μ_i es constante dentro de cada estrato de T unidades; el tercero supone una tendencia lineal, $\mu_i = \mu + \beta i$. Para cada uno se proporciona el estimador de la varianza de la media de una población finita.

Gautschi[15] supone que $Y_1, \ldots, Y_N$ son variables no correlacionadas cuyos valores esperados cambian linealmente con i, específicamente para $i = 1, \ldots, N$ se supone que:

a) $E_m(Y_i) = \mu + \beta i$

b) $V_m(Y_i) = \sigma^2$

c) $C_m(Y_i, Y_j) = 0$ con $i \neq j$.

Considerando este modelo obtiene que

$$\sigma^2_{SY} = \frac{\beta^2(T^2-1)}{12} + \frac{T-1}{nT}\sigma^2,$$
$$\sigma^2_{ST} = \frac{\beta^2(T^2-1)}{12n} + \frac{T-1}{nT}\sigma^2,$$
$$\sigma^2_{SI} = \frac{\beta^2(T-1)(nT+1)}{12} + \frac{T-1}{nT}\sigma^2.$$

Entonces, $\sigma^2_{ST} \leq \sigma^2_{SY} \leq \sigma^2_{SI}$. La igualdad se cumple sólo si $n = 1$.

En el trabajo publicado por Singh et al. [51], se propone el muestreo sistemático modificado (MSS del inglés *modified systematic sampling*), el cual consiste en seleccionar unidades que son equidistantes del inicio y del fin de la población. Las unidades en muestra serán las etiquetadas por $r + iT$, $N + 1 - r - iT$ con $i = 0, 1, \ldots, n/2 - 1$ si n es par; si n es impar, la muestra estará conformada por las unidades $r+iT$, $N+1-r-iT$ y $r+(n-1)T/2$, $i = 0, 1, \ldots, (n-1)/2-1$. Cuando la población presenta una tendencia lineal, como la que supone Cochran [7], encontraron que el muestreo sistemático modificado es el de menor varianza, le sigue el muestreo sistemático y el de mayor varianza corresponde al muestreo aleatorio simple, esto es, $\sigma^2_{MSS} \leq \sigma^2_{SY} \leq \sigma^2_{SI}$.

Bellhouse y Rao [2] comparan el valor esperado del error cuadrático medio del estimador de la media bajo cinco métodos, uno de ellos corresponde al muestreo sistemático y los otros cuatro a modificaciones de este. Los métodos modificados consisten en lo siguiente.

a) Ajustar el estimador de la media asignándole un peso diferente a las observaciones de los extremos, $\bar{y}' = \bar{y} + [2(n-1)T]^{-1}(2r - T - 1)(y_r - y_{r+(n-1)T})$.

b) Usar el muestreo sistemático modificado (MSS) propuesto por Singh et al. [51].

c) Usar el muestreo sistemático equilibrado (BSS del inglés *balanced systematic sampling*), el cual consiste en seleccionar los elementos $r+2Ti$ y $2T(i+1)-r+1$ con $i = 0, 1, \ldots, n/2-1$ si n es par, o bien los elementos $r+2Ti$, $2T(i+1)-r+1$ y $r+(n-1)T$ con $i = 0, 1, \ldots, (n-1)/2-1$ si n es impar. Este método coincide con el MSS cuando $n = 2$.

d) Usar el muestreo sistemático centrado (CSS del inglés *centred systematic sampling*) , que consiste en usar el muestreo sistemático con $r = (T+1)/2$ si T es impar o elegir con probabilidad $1/2$ entre $r = T/2$ ó $r = T/2 + 1$ si T es par.

Bajo el supuesto que la superpoblación presenta una tendencia lineal, como el de Gautshi[15], los autores encuentran que los métodos modificados dan lugar a un error cuadrático medio menor con respecto al muestreo sistemático. También consideran otros modelos:

a) Tendencia lineal y variación periódica, $E_m(Y_i) = \mu + \beta i + x_i$ donde x_i es una función periódica de i, $i = 1, \ldots, N$.

b) Tendencia parabólica, $E_m(Y_i) = \mu + \beta_1 i + \beta_2 i^2$.

   En ambos casos $V_m(Y_i) = \sigma^2$ y $C_m(Y_i, Y_j) = 0$ con $i \neq j$.

c) Autocorrelación serial, como el estudiado por Cochran [6].

En el primer caso, el muestreo sistemático fue el menos eficiente y el más eficiente el que usa el estimador $\bar{y}'$. En el caso de la tendencia parabólica el más eficiente fue el CSS si n es impar, seguido por el muestreo sistemático estimando la media mediante $\bar{y}'$. Cuando los elementos están correlacionados serialmente, el método de menor varianza fue el CSS, le siguen el muestreo sistemático, el MSS y por último el BSS. La varianza de $\bar{y}'$ resultó similar a la del muestreo sistemático sin modificación alguna.

Bartolucci y Montanari [1] suponen que la población finita es una realización de una superpoblación que sigue el modelo lineal

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i, \ i = 1, \ldots, N,$$

donde el error ϵ tiene las siguientes propiedades

$$E_m(\epsilon_i) = 0 \text{ y } E_m(\epsilon_i \epsilon_j) = \sigma_{ij}^2.$$

Encuentran que bajo este modelo,

$$E_m\left[V_{SY}\left(\bar{y}\right)\right] = E_m\left[V_{SY}^t\left(\bar{y}\right)\right] + E_m\left[V_{SY}^s\left(\bar{y}\right)\right],$$

esto es, el valor esperado de la varianza del estimador de la media puede descomponerse en el valor esperado de $V_{SY}^{t}(\cdot)$, la varianza debida al componente sistemático del modelo, $\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$, y la debida al componente estocástico ϵ_i, $V_{SY}^{s}(\cdot)$.

Cuando $E_m(\epsilon_i \epsilon_j) = 0$, $E_m(\epsilon_i^2) = \sigma^2$, $p = 1$ y $x_{1i} = i$ se tiene el modelo con tendencia lineal mencionado anteriormente, $Y_i = \alpha + \beta i + \epsilon_i$, en este caso

$$E_m\left[V_{SY}^{t}(\bar{y})\right] = \frac{\beta^2(T^2-1)}{12} \text{ y}$$

$$E_m\left[V_{SY}^{s}(\bar{y})\right] = \frac{T-1}{nT}\sigma^2.$$

También proponen dos estimadores insesgados bajo el modelo, para estimar la parte de la varianza debida al componente sistemático del modelo, uno de ellos se basa en promedios móviles y el otro en polinomios locales.

Sampath et al. [46] consideran el problema de la estimación del total en presencia de una tendencia parabólica, bajo tres diseños: muestreo sistemático, BSS y MSS y el siguiente modelo de superpoblación

$$Y_i = \mu + \beta_1 i + \beta_2 i^2 + \epsilon_i, \ i = 1, \ldots, N,$$

donde $E_m(\epsilon_i) = 0$, $V_m(\epsilon_i) = \sigma^2 i^g$ con g una constante predeterminada y $C_m(\epsilon_i, \epsilon_j) = 0$ con $i \neq j$, $(i = 1, \ldots, N)$.

Derivan el estimador del total, modificando los pesos de cuatro de las unidades seleccionadas, en el sistemático y MSS estas unidades corresponden a la más cercana del inicio, la más cercana del fin y las dos más cercanas al centro; en el BSS las unidades en muestra del primer bloque de $2T$ elementos y las dos seleccionadas en el último bloque. Obtienen el error cuadrático medio de los estimadores con los pesos modificados. Señalan que debido a la complejidad de las expresiones encontradas, la comparación de las tres estrategias se hace por medio de un estudio numérico, eligiendo diferentes valores de n, N, T y g. Concluyen que el menor error cuadrático medio corresponde a: i) muestreo sistemático si $g = 0$, ii) BSS si $g = 2, 3$ y iii) BSS si $g = 1$ siempre y cuando n sea muy grande en comparación con T, en otro caso, el sistemático es mejor.

En el caso de poblaciones continuas, uno de los trabajos pioneros corresponde al de Yates [60], quien analiza dos ejemplos:

a) Una línea dividida en secciones, algunas de las cuales poseen cierto atributo. Se desea estimar la proporción de la línea que posee el atributo. Si se considera sólo una sección de longitud c que posee el atributo, se puede representar mediante la función

$$f(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq c \\ 0 & \text{en otro caso.} \end{cases}$$

b) La función de densidad normal o gaussiana, $f(x) = (2\pi\sigma^2)^{-1/2} \exp(-x^2/(2\sigma^2))$. Se quiere estimar el area bajo la curva.

En el primer caso, obtiene que la varianza de la proporción bajo muestreo sistemático, depende de la relación entre la longitud de la sección, c, y el intervalo muestral T. Si c es pequeña comparada con T, entonces la eficiencia del muestreo estratificado con una unidad por estrato de longitud T es muy similar a la del muestreo sistemático; cuando $c > T$, el muestreo sistemático es más eficiente que el muestreo estratificado; el autor comenta que el caso c cercano a T no se había analizado en detalle, pero se esperaba en general lo mismo. Cuando la función corresponde a la densidad gaussiana, de manera numérica, encuentra que el error al estimar la integral es muy pequeño, menor al 1.5 %.

En su trabajo, Yates también trata el problema de la estimación de la varianza para poblaciones finitas, propone dividir la muestra en segmentos, que denomina muestras sistemáticas parciales, para obtener comparaciones independientes y con ellas una estimación de la varianza. A partir de la comparación entre diferentes estimadores aplicados a diferentes poblaciones, concluye que no es posible establecer una estimación del error confiable de manera general.

En el tema de integración numérica, Moran [38] con base en el problema planteado por Yates [60], obtiene la varianza al estimar la integral $t = \int_{\mathbb{R}} f(x)dx$ mediante la suma de los valores que toma la función cuando se muestrea sistemáticamente sobre el espacio muestral:

$$\hat{t} = T \sum_{j=-\infty}^{\infty} f(r + jT), \tag{2.1}$$

donde r sigue una distrubución uniforme en el intervalo $[0, T]$ y $f(x)$ es una función de variación acotada y Lebesgue integrable en $-\infty < x < \infty$. Moran [38] encontró que la varianza de (2.1) está dada por

$$V_{SY}\left(\hat{t}\right) = T \sum_{j=-\infty}^{\infty} \int_{\mathbb{R}} f(r) f(r + jT) dr - t^2. \tag{2.2}$$

Ya que la expresión (2.2) sólo se puede evaluar directamente en casos particulares, encuentra una expresión equivalente en términos de la transformada de Fourier

$$V_{SY}\left(\hat{t}\right) = \sum_{j=-\infty}^{\infty} \left| F\left(\frac{j}{T}\right)\right|^2 - t^2 = 2 \sum_{j=1}^{\infty} \left| F\left(\frac{j}{T}\right)\right|^2, \tag{2.3}$$

donde $F(w) = \int_{\mathbb{R}} \exp(-2\pi i w x) f(x) dx$ corresponde a la transformada de Fourier de la función f.

Deriva expresiones para $V_{SY}\left(\hat{t}\right)$ considerando tres funciones específicas:

a) La función de densidad normal $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$

b) La función rectangular $f(x) = 1/\ell$ si $-\ell/2 \leq x \leq \ell/2$

c) La función de densidad de Cauchy $f(x) = [\pi(1 + x^2)]^{-1}$.

Posteriormente, motivada por las aplicaciones en el área de Geoestadística, surgió la teoría transitiva de G. Matheron [35], [36]. Los métodos basados en esta teoría son no paramétricos y asintóticos, constituyen una alternativa para estimar la varianza del muestreo sistemático. Un área en la que estos métodos han sido una herramienta muy importante es en la Estereología, ciencia que trata de la teoría y metodología del muestreo geométrico (Gual-Arnau y Cruz-Orive [19]).

Como lo señala Cruz-Orive [8], la teoría de Matheron tiene ventajas prácticas, ya que se puede usar para integrales de funciones arbitrarias y las aproximaciones de la varianza se pueden calcular a partir de los datos; por el contrario, para evaluar la varianza obtenida por Moran (ecuación (2.3)) se debe de conocer explícitamente la función.

Matheron [35] muestra que $|F(w)|^2$ corresponde a la transformada de Fourier $G(w)$ del covariograma $g(h) = \int_{\mathbb{R}} f(x) f(x + h) dx$ asociado a la función f. Obtiene

(p. 72) la siguiente expresión alternativa para la varianza

$$V_{SY}\left(\hat{t}\right) = 2\sum_{j=1}^{\infty} G\left(\frac{j}{T}\right). \tag{2.4}$$

El covariograma $g(h)$ es una función positiva definida en todo su soporte, esto es, para cualquier conjunto de puntos $\lambda_1, \ldots, \lambda_K \in \mathbb{C}$ y de números reales $h_1, \ldots, h_K$ se cumple que

$$\sum_i \sum_j \lambda_i \lambda_j g(h_i - h_j) \geq 0.$$

Como consecuencia de esta propiedad también se tiene que $g(0) \geq 0$, $|g(h)| \leq g(0)$ y $g(-h) = g(h)$. Matheron usa aproximaciones adecuadas de g cerca del origen y las aplica a la expresión (2.4) para aproximar la varianza $V_{SY}\left(\hat{t}\right)$.

Una expresión equivalente a (2.2) está dada por (ver Matheron [35], p. 73)

$$V_{SY}\left(\hat{t}\right) = T\sum_{j=-\infty}^{\infty} g(jT) - \int_{\mathbb{R}} g(h)dh, \tag{2.5}$$

la cual puede verse como el error al aproximar la integral del covariograma por medio de una suma. Intuitivamente, esta fórmula muestra que la varianza del estimador sólo depende de T y del covariograma g, la varianza es más pequeña conforme T decrece o bien conforme f, y por lo tanto g, es menos irregular. Matheron [36] (p. 27), menciona que los modelos teóricos de covariogramas tienen irregularidades cerca de cero, de $b-a$ y de $a-b$. Señala que es posible demostrar que si T es pequeño, la varianza (2.5) puede expresarse como la suma de:

a) El término de extensión, que depende del comportamiento de $g(h)$ cerca del origen.

b) El término *Zitterbewegung* que depende de $g(h)$ cerca de $b-a$, el cual fluctúa alrededor de cero.

Propone que para calcular la varianza se use sólo el término de extensión, argumenta que aunque el *Zitterbewegung* puede no ser despreciable, fluctúa alrededor de cero, y por ello se puede ignorar, además no se puede calcular con una sola muestra. Para estimar la varianza, propone que se use un modelo téorico para g, por ejemplo

un modelo polinomial, el cual debe de ser una función positiva definida y debe de aproximarse al covariograma experimental o de los datos

$$g^*(kT) = T \sum_{j=-\infty}^{\infty} f(r+jT)f(r+jT+kT).$$

En el área de la Estereología, Souchet [52] y Kiêu [32] reexaminaron los resultados de Matheron y derivan una versión refinada de la fórmula estándar de Euler-MacLaurin, que proporciona una expansión del error al aproximar la integral de una función por datos discretos. Para ello suponen que la función es integrable, con soporte acotado en $\mathbb{R}$ y suave a trozos. A la función f integrable y con soporte acotado la refieren como la función de medida.

Bajo estos supuestos Kiêu [32], p. 43, y Kiêu et al. [33], p. 269, obtienen la varianza del estimador de t. Señalan que cuando T es suficientemente pequeño, el término de extensión, $V_E(\hat{t})$, es una aproximación de $V_{SY}(\hat{t})$ ya que representa la tendencia de la varianza,

$$V_E(\hat{t}) = (-1)^q T^{2q+2,T} P_{2q+2}(0) \sum_{v \in Df^{(q)}} [Sf^{(q)}(v)]^2, \tag{2.6}$$

donde

q es el orden de la primera derivada de f no continua,

$P_l(0) = B_l(0)/l!$ con $B_l(0)$ el $l-$ésimo número de Bernoulli,

$Sf^{(q)}(x) := \lim_{y \to x^+} f^{(q)}(y) - \lim_{y \to x^-} f^{(q)}(y)$, $x \in \mathbb{R}$, es la amplitud del salto de la $q-$ésima derivada de f, $f^{(q)}$,

$Df^{(q)} = \left\{x : Sf^{(q)}(x) \neq 0\right\}$ es el soporte de $Sf^{(q)}$.

García-Fiñana y Cruz-Orive [14] 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 q ya no necesariamente es un entero no negativo. La aproximación se obtiene a través de un refinamiento de la fórmula de Euler-MacLaurin usando herramientas de cálculo fraccional. La expresión analítica depende del parámetro de suavización $q \in [0,1]$, que generalmente se desconoce y

habrá que estimarlo o bien suponerlo. La varianza de extensión con $q \in [0,1]$ es muy similar a (2.6) donde se suma sobre las singularidades v_i de $f^{(q)}$:

$$V_E(\hat{t}) = \frac{T^{2q+2} P_{2q+2,T}(0)}{\cos(\pi q)} \sum_{v_i} [S f^{(q)}(v_i)]^2.$$

También en el campo de la Estereología, Gundersen [20] propone un algoritmo de selección llamado el fraccionador suave (del inglés *smooth fractionator*), el cual consiste en aplicar un muestreo sistemático a unidades físicamente separadas, cada una de las cuales tiene asociada una medición (variable auxiliar) relacionada con la variable de interés. Esta población se ordena simétricamente, con un pico en el centro y saltos mínimos entre unidades, de acuerdo a la información auxiliar. Su nombre proviene de la idea de fraccionar un objeto en N unidades o piezas y después ordenarlas de manera tal que la diferencia en la variable auxiliar entre una unidad y la siguiente sea lo más pequeña posible. La eficiencia del esquema la ilustra mediante seis ejemplos, uno con datos simulados y cinco con datos reales. Señala que para tamaños de muestra iguales o mayores a diez unidades, la varianza del estimador bajo el fraccionador suave es muy pequeña. Menciona que los aspectos importantes para la eficiencia del diseño, en orden de importancia, son a) la garantía de simetría y de que haya una sola moda, y b) los saltos mínimos; el sesgo potencial y el ruido de la variable auxiliar son menos relevantes, aunque entre mayor sea la relación entre las variables auxiliar y la de interés, mayor será la eficiencia del fraccionador suave.

Las muestras sistemáticas que se obtienen a partir de la población ordenada como en el fraccionador suave, son las mismas que se obtienen bajo el muestreo sistemático equilibrado (BSS) aplicado una población ordenada de manera creciente. Gundersen refiere el trabajo de Murthy [39], quien además señala que este esquema de selección fue usado en la Encuesta Nacional de la India desde 1955 para propósitos de estudios especiales y refiere el trabajo de Sethi [49].

Sethi [48, 49] demostró que si una población con un número par de elementos, N, se ordena de manera creciente de acuerdo a la característica de interés, entonces la manera óptima de elegir un par de unidades, es hacerlo mediante BSS con $n = 2$. La elección es óptima porque mínimiza cualquier función de pérdida que sea no decreciente y cóncava hacia abajo, en particular minimiza la varianza. En el Capítulo 3 se precisará el resultado de Sethi.

Capítulo 3

Muestreo sistemático de una población fija

Una estrategia de muestreo es la combinación del método de selección y de estimación. Existen tres clases de estrategias: basada en el diseño, basada en el modelo y asistida por el modelo; cuya diferencia radica en la fuente de aleatoriedad que utilizan para dar una estructura estocástica.

En la que se basa en el diseño, los valores $y_1, y_2, \ldots, y_N$ de la variable de interés, o bien la función $f(x)$, se consideran fijos pero desconocidos; la fuente de aleatoriedad se halla en la selección de la muestra, por lo que las probabilidades de selección introducidas con el diseño se usan para determinar valores esperados, varianzas y sesgos de los estimadores. En el muestreo sistemático por ejemplo, el estimador de la media es una variable aleatoria porque generalmente varía de muestra en muestra.

La segunda clase de estrategia se basa en el modelo, en la cual los valores $y_1, \ldots, y_N$ asociados con la población de N elementos, se ven como una posible realización de variables aleatorias $Y_1, \ldots, Y_N$. Las propiedades de los estimadores dependen de la función de densidad de probabilidad conjunta $f_{Y_1,\ldots,Y_N}(y_1, \ldots, y_N; \underline{\mu})$ de las N variables aleatorias, la cual tiene el vector de parámetros desconocidos $\underline{\mu}$. El modelo de la superpoblación o simplemente el modelo, denotado por m, se referirá a las condiciones asociadas con la clase de distribuciones a la que pertenece $f_{Y_1,\ldots,Y_N}(y_1, \ldots, y_N; \underline{\mu})$. Por ejemplo, un modelo m podría enunciar que $Y_1, \ldots, Y_N$ son variables aleatorias idénticamente distribuidas con media μ, varianza σ^2 y covarianza entre Y_i y Y_j dada por $\sigma^2 \exp(-\lambda(j-i))$, $i < j$.

Existen diferentes razones que justifican alguno de estos enfoques. En favor del basado en el diseño, que se aplica ampliamente en la práctica para la estimación de parámetros de poblaciones finitas, se argumenta que en muchos casos se conoce muy poco acerca de la población como para establecer algún modelo y se conocen las propiedades estadísticas de los estimadores como su valor esperado y varianza, las cuales no dependen de un modelo. Adicionalmente, la aleatorización en la selección protege contra un sesgo de selección y las muestras aleatorias son vistas como objetivas (Särndal et al. [54], §1.10).

En la aproximación basada en el modelo se argumenta (Hannan [41]) que aunque la información auxiliar se usa para la estimación, por ejemplo en el estimador de razón o para el diseño de muestreo, como en la estratificación, su uso se justifica formalmente hasta que se supone algún modelo que relacione la variable auxiliar x con la de interés y. El uso de información proveniente de encuestas por muestreo para propósitos analíticos no se puede hacer sólo en el marco basado en el diseño.

La controversia que inicialmente existió entre ambos enfoques, en la que se veían como extremos, se ha suavizado considerablemente en las últimas dos décadas, hasta que actualmente la teoría y práctica del muestreo también se basa en una combinación de ambos. Así, una tercera estrategia, en la que la teoría del muestreo se apoya en modelos, es el muestreo asistido por el modelo (del inglés *model assisted survey sampling*); un ejemplo es suponer un modelo para la no respuesta o errores de medición, reconociendo el diseño muestral. Este enfoque es el adoptado en el libro de texto de Särndal et al. [54]. Algunas notas sobre el desarrollo histórico de estos enfoques se pueden consultar en Särndal [53], en donde además se proporcionan varias referencias.

A continuación se presenta el estimador del total y su precisión bajo el muestreo sistemático desde el enfoque basado en el diseño, primero se considera una población finita y después una población continua; puesto que el número de elementos en el primer caso es finito, y en el segundo infinito, es necesario hacer algunas distinciones entre ambas poblaciones en aspectos como la selección y el error de estimación. En el Capítulo 5 se trata la varianza del muestreo sistemático suponiendo superpoblaciones.

3.1. Estimador del total de una población finita y su varianza

La mayor parte de la teoría clásica del muestreo, motivada por censos y encuestas a poblaciones humanas, trata de la selección y estimación a partir de una muestra proveniente de una población finita y fija.

Sea y_i el valor de la variable de interés para la $i-$ésima unidad de la población, u_i, $i = 1, \ldots, N$. Cuando se elige el número aleatorio r con igual probabilidad del conjunto $\{1, \ldots, T\}$, la muestra sistemática consiste en el subconjunto de unidades

$$s_r = \{u_i : i = r + (j-1)T \leq N; j = 1, 2, \ldots, n_s\},$$

donde $T \in \mathbb{N}$ es el salto o intervalo muestral y corresponde a la parte entera de N/n, con $N = nT + c$ y c un entero que satisface $0 \leq c < T$; n_s es el tamaño de la muestra o cardinalidad de s_r y es una variable aleatoria tal que

$$n_s = \begin{cases} n+1 & \text{si } r \leq c \\ n & \text{en otro caso.} \end{cases}$$

El valor esperado y varianza del tamaño de muestra son $E(n_s) = N/T$ y $V(n_s) = (N/T - [N/T])(1 - N/T + [N/T])$, respectivamente; aquí $[N/T]$ denota la parte entera de N/T y sólo se usará esta notación cuando así se indique.

Gráficamente este método de selección puede representarse como sigue

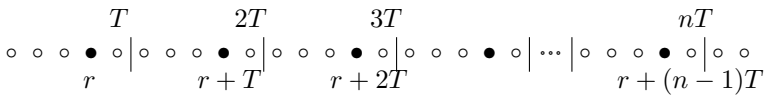


Figura 3.1: Selección de una muestra sistemática de una población finita. Los círculos negros indican las unidades seleccionadas, $n_s = n$.

Cuando no haya ambigüedad s_r se denotará también por s y por simplicidad el $i-$ésimo elemento sólo se representará por la etiqueta i, por lo que la muestra está dada por

$$s_r = \{r, r + T, \ldots, r + (n_s - 1)T\}.$$

El conjunto de T muestras posibles $s_1, s_2, \ldots, s_T$ es una partición de la población, esto es $U = \bigcup_{r=1}^{T} s_r$. Entonces el total poblacional $t = \sum_{i=1}^{N} y_i$ puede reescribirse como:

$$t = \sum_{r=1}^{T} t_{s_r}$$

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

El muestreo sistemático es equivalente a seleccionar, con igual probabilidad, uno de T conglomerados con n ó $n+1$ unidades cada uno. Esto se esquematiza en el Cuadro 3.1, cuando $c = 0$, en el que cada conglomerado está representado por una columna y cada una de las T columnas tiene la misma probabilidad de ser escogida como la muestra. Otra manera de visualizar el muestreo sistemático es dividir a la población en n grupos, en este caso los renglones de la tabla y luego seleccionar el $r-$ésimo elemento de cada renglón o grupo.

Cuadro 3.1: Posibles muestras bajo muestreo sistemático

	Muestra				
	s_1	$\cdots$	s_r	$\cdots$	s_T
Valores	y_1	$\cdots$	y_r	$\cdots$	y_T
de y	y_{T+1}	$\cdots$	y_{T+r}	$\cdots$	y_{2T}
	$\vdots$		$\vdots$		$\vdots$
	$y_{(n-1)T+1}$	$\cdots$	$y_{(n-1)T+r}$	$\cdots$	y_N
Total muestral	t_{s_1}	$\cdots$	t_{s_r}	$\cdots$	t_{s_T}

En los capítulos 4 y 5 se supondrá que $T = N/n \in \mathbb{N}$, o equivalentemente $c = 0$, porque el tamaño de muestra queda fijo, simplificando el problema. Otras ventajas de que $T \in \mathbb{N}$ son que la media muestral es un estimador insesgado de la media poblacional y la probabilidad de que la unidad i pertenezca a la muestra es la misma para todo $i = 1, \ldots, N$.

Si N/n no es un entero, la media muestral es un estimador sesgado de $\overline{y}_U$, sin embargo el sesgo es pequeño y puede despreciarse si el tamaño de muestra es grande, Cochran [7] señala que probablemente este sesgo sea despreciable si $n > 50$, y que no se espera que sea grande si n es pequeña.

Además, si $N/n \notin \mathbb{N}$, la condición $T \in \mathbb{N}$ podría llevar a un tamaño de muestra muy diferente al deseado en algunas situaciones. Por ejemplo si $N = 160$ y se desea seleccionar 60 unidades, entonces si $T = 2$ se elegirían 80 unidades, mientras que si $T = 3$ se elegirían 53 ó 54. Si N es grande comparado con n, la diferencia es menor.

En la práctica, cuando se requiere un tamaño de muestra n y $N/n \notin \mathbb{N}$, se puede optar por alguna de las siguientes alternativas (ver Särndal et al. [54], p. 77):

- Se eliminan c unidades al azar de la población, de tal manera que $N = Tn$. Estas c unidades se pueden considerar en un estrato, o bien dejarse fuera de la selección.
- Se usa un muestreo sistemático circular. En este caso se elige un número aleatorio r entre 1 y N como punto de inicio, las siguientes unidades en la muestra serán $r + (j-1)T$ si $r + (j-1)T \leq N$ ó $r + (j-1)T - N$ si $r + (j-1)T > N$, con $j = 1, \ldots, n$ y T el entero más cercano a N/n.
- Se usa un intervalo fraccional. Sean $T = N/n$ y $r \sim U(0, T)$, la muestra consistirá en aquellos elementos k que cumplen $k - 1 < r + (j-1)T \leq k$ con $j = 1, \ldots, n$.

Al muestreo sistemático Murthy [39] también lo llama muestreo sistemático lineal para distinguirlo de algunas modificaciones al mismo, como el muestreo sistemático circular, en este trabajo por simplicidad se le referirá como muestreo sistemático.

En correspondencia con el algoritmo de selección, se encuentra el proceso de estimación del parámetro de interés. Un parámetro μ es una característica de la población y generalmente es desconocido. Este puede verse como una función de los valores $y_1, y_2, \ldots, y_N$, esto es,

$$\mu = f(y_1, \ldots, y_N).$$

Su valor exacto puede obtenerse si se mide la población completa, si no hay errores de medición, ni tampoco no respuesta. Algunos parámetros de interés son el total de una variable de estudio, la media, la mediana y el coeficiente de correlación entre dos variables.

Un estimador de μ es una función de los valores de la muestra,

$$\hat{\mu} = f(y_{i_1}, \ldots, y_{i_n}) \text{ con } i_j \in s.$$

Por ejemplo, un estimador de la media poblacional, $\overline{y}_U = \sum_{i=1}^{N} y_i/N$, es la media muestral

$$\overline{y} = \frac{1}{n}\sum_{i\in s} y_i,$$

y un estimador de la varianza poblacional $s_U^2 = \sum_{i=1}^{N} \left(y_i - \overline{y}_U\right)^2/(N-1)$ es la varianza muestral

$$s^2 = \frac{1}{n-1}\sum_{i\in s} \left(y_i - \overline{y}\right)^2.$$

Un estimador es lineal si se trata de una combinación lineal de los valores y_i $i \in s$,

$$\hat{\mu} = \sum_{i\in s} \alpha_i y_i, \quad \alpha_i \in \mathbb{R} \tag{3.1}$$

con α_i $(i \in s)$ constantes. El estimador $\overline{y}$ es lineal; en cambio, el estimador s^2 no es lineal. Otra propiedad importante es la insesgadez. Un estimador es insesgado si

$$E(\hat{\mu}) = \mu.$$

Horvitz y Thompson [24] derivaron un estimador del total poblacional, $t = \sum_{i\in U} y_i$, en el caso de una muestra aleatoria sin reemplazo de un universo finito. Encontraron que en la clase de los estimadores lineales de la forma dada en (3.1), donde α_i es usado como un peso para el elemento i siempre que este haya sido seleccionado en la muestra, el único, y por lo tanto el mejor estimador linealmente insesgado es

$$\hat{t}_\pi = \sum_{i\in s} \frac{y_i}{\pi_i}, \tag{3.2}$$

donde π_i se denomina probabilidad de inclusión de primer orden, corresponde a la probabilidad de que la unidad i pertenezca a la muestra y se obtiene a partir de lo que en la literatura se denomina diseño muestral, ver por ejemplo Särndal et al. [54].

Un diseño muestral, denotado por $p(s)$, es una función tal que $p(s_r)$ es la probabilidad de elegir la muestra s_r bajo el esquema de selección usado. Esta función va del conjunto de todas las muestras de U, $\mathcal{L} = \{s_1, s_2, \ldots, s_{2^N}\}$, a un conjunto de probabilidades $\{p(s_1), p(s_2), \ldots, p(s_{2^N})\}$, esto es,

$$p(s) : \{s_1, s_2, \ldots, s_{2^N}\} \to \{p(s_1), p(s_2), \ldots, p(s_{2^N})\}.$$

Como $p(s)$ es una distribución de probabilidades en $\mathcal{L}$, se tiene que

$$p(s) \geq 0 \ \forall \ s \in \mathcal{L} \ \text{ y } \ \sum_{s \in \mathcal{L}} p(s) = 1.$$

El subconjunto de $\mathcal{L}$ compuesto de aquellas muestras s para las cuales $p(s) > 0$ constituye el conjunto de muestras posibles.

La probabilidad π_i se determina como sigue

$$\pi_i = \Pr(i \in s) = \sum_{s \ni i} p(s),$$

aquí $s \ni i$ denota que se suma sobre aquellas muestras s que contienen al elemento i. Análogamente, la probabilidad de que los elementos i y j sean incluidos en la muestra, llamada probabilidad de inclusión de segundo orden, está dada por

$$\pi_{ij} = \Pr(i \& j \in s) = \sum_{s \ni i \& j} p(s).$$

Al estimador (3.2) se le ha llamado de Horvitz-Thompson, en Särndal et al. [54] también se le refiere como estimador-π. Si $\pi_i > 0$ para toda $i \in U$, entonces la varianza del estimador-π está dada por

$$V\left(\hat{t}_\pi\right) = \sum_U \sum_U \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) y_i y_j = \sum_U \sum_U \frac{\pi_{ij}}{\pi_i \pi_j} y_i y_j - \left(\sum_U y_i\right)^2. \tag{3.3}$$

Horvitz y Thompson [24] también proporcionan un estimador insesgado de $V\left(\hat{t}_\pi\right)$ si $\pi_{ij} > 0$ para toda $i, j \in U$,

$$\widehat{V}\left(\hat{t}_\pi\right) = \sum_s \sum_s \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j}\right) y_i y_j. \tag{3.4}$$

Yates y Grundy [62] encontraron una expresión alternativa para la varianza dada en (3.3) cuando el tamaño de muestra es fijo:

$$V\left(\hat{t}_\pi\right) = -\frac{1}{2} \sum_U \sum_U \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

Un estimador insesgado de esta varianza si $\pi_{ij} > 0$ para toda $i, j \in U$ es

$$\widehat{V}\left(\hat{t}_\pi\right) = -\frac{1}{2} \sum_s \sum_s \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

El estimador Horvitz-Thompson también puede usarse en el muestreo con reemplazo para estimar el total poblacional de manera insesgada, en este caso la probabilidad de inclusión π_i es la probabilidad de que el $i-$ésimo elemento sea seleccionado al menos una vez. En el caso de selección con reemplazo, existe otro estimador insesgado de t debido a Hansen y Hurwitz (ver Särndal et al. [54], resultado 2.9.1):

$$\hat{t}_{pwr} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}, \tag{3.5}$$

donde m es el número de extracciones, k_i corresponde al elemento seleccionado en la $i-$ésima extracción y p_k denota la probabilidad de seleccionar el elemento k en la $i-$ésima extracción, como el muestreo es con reemplazo $p_{k_i} = p_k \ \forall i = 1, \ldots, m$. No se puede generalizar cuál de los dos estimadores tiene menor varianza, ésta dependerá de los valores $y_1, \ldots, y_N$.

En el caso del muestreo sistemático, las probabilidades de inclusión de primer y segundo orden son:

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

En consecuencia, el estimador del total (3.2) y su varianza (3.3) se reducen a las expresiones siguientes (Särndal et al. [54], resultado 3.4.1)

$$\hat{t}_\pi = \hat{t}_{s_r} = T\sum_{i\in s_r} y_i = T t_{s_r}, \tag{3.6}$$

$$V_{SY}\left(\hat{t}_\pi\right) = T\sum_{r=1}^{T}\left(t_{s_r} - \frac{t}{T}\right)^2 = \frac{1}{T}\sum_{r=1}^{T}\left(T t_{s_r} - t\right)^2. \tag{3.7}$$

Sin embargo, ya que no se cumple la condición $\pi_{ij} > 0$ para toda $i, j \in U$, la ecuación (3.4) no puede usarse para estimar la varianza, por lo que se han obtenido aproximaciones haciendo algunos supuestos sobre la estructura de la población, usando algún método de remuestreo o bien modificando el esquema de selección. En el capítulo 6 se darán tres de estas aproximaciones.

3.2. Orden de la población y eficiencia

La eficiencia del muestreo sistemático también depende de la manera en que la población está ordenada. De acuerdo con la expresión (3.7), la varianza $V_{SY}\left(\hat{t}_\pi\right)$ tiende a cero si los totales muestrales t_{s_r} son aproximadamente iguales, es decir, si el orden de los N elementos de la población es tal que las T muestras sistemáticas tienen totales similares.

Esto también puede observarse al descomponer la variación total, o suma total de cuadrados (SST, SS del inglés *sum of squares* y T de *total*), en la variación dentro de la muestra sistemática (SSW, W del inglés *within*) y la variación entre muestras sistemáticas (SSB, B del inglés *between*), esto es,

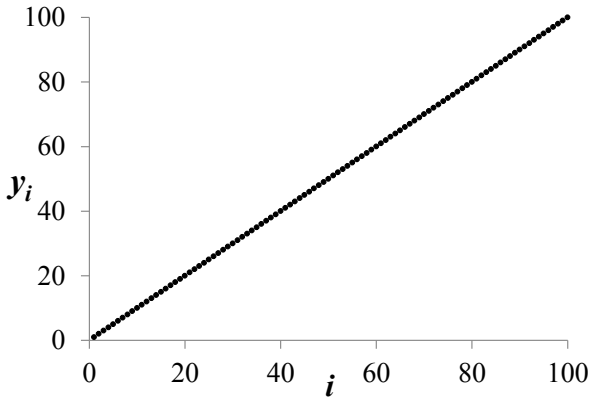
$$\begin{aligned} SST = (N-1)s_U^2 = \sum_{i=1}^{N}\left(y_i - \overline{y}_U\right)^2 &= \sum_{r=1}^{T}\sum_{i\in s_r}\left(y_i - \overline{y}_{s_r}\right)^2 + n\sum_{r=1}^{T}\left(\overline{y}_{s_r} - \overline{y}_U\right)^2 \\ &= \sum_{r=1}^{T}\sum_{i\in s_r}\left(y_i - \overline{y}_{s_r}\right)^2 + \frac{1}{N}\frac{\sum_{r=1}^{T}\left(Tt_{s_r} - t\right)^2}{T} \\ &= SSW + SSB, \end{aligned}$$

donde $\overline{y}_{s_r} = \sum_{i\in s_r} y_i/n = t_{s_r}/n$ es la media de la muestra r (ver Särndal et al. [54], p. 78). Nótese que $V_{SY}\left(\hat{t}_\pi\right) = N \times SSB$.

Como la varianza poblacional s_U^2 es fija, a medida que las muestras son más heterogéneas al interior, entonces SSW aumenta y SSB disminuye, por lo que el muestreo sistemático es más eficiente. La heterogeneidad dentro de las muestras sistemáticas puede lograrse si los n grupos en los que se dividió la población son homogéneos al interior. Murthy [39], §5.5, señala que un buen arreglo debería asegurar que unidades similares con respecto a y_i estén juntas, lo cual sugiere que un posible orden es el creciente, en contraparte, una crítica al orden creciente es que las primeras muestras subestiman al total poblacional y las últimas muestras lo sobreestiman.

La variable de interés y_i $(i = 1, \ldots, N)$ generalmente se desconoce, en este caso la población se puede ordenar de acuerdo con alguna variable auxiliar x_i $(i = 1, \ldots, N)$, es deseable que la información auxiliar tenga la mayor relación posible con la de interés para que este diseño sea eficiente.

Con la finalidad de mostrar cómo la varianza del muestreo sistemático depende



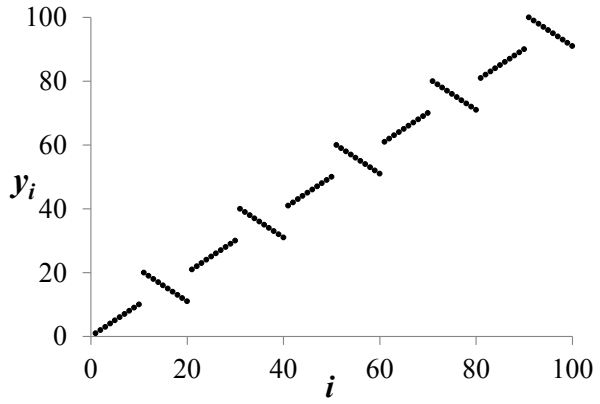
	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	1	2	3	4	5	6	7	8	9	10
	11	12	13	14	15	16	17	18	19	20
	21	22	23	24	25	26	27	28	29	30
	31	32	33	34	35	36	37	38	39	40
	41	42	43	44	45	46	47	48	49	50
	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70
	71	72	73	74	75	76	77	78	79	80
	81	82	83	84	85	86	87	88	89	90
	91	92	93	94	95	96	97	98	99	100
t_{s_r}	460	470	480	490	500	510	520	530	540	550

Figura 3.2: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada de manera creciente. $V_{SY}\left(\hat{t}_\pi\right) = 82{,}500$, $SSW = 82{,}500$ y $SSB = 825$.

del orden de la población, se presenta el siguiente ejemplo que se basa en el Ejemplo 3.4.2 de Särndal et al. [54] y en Muzquiz [40] §3.1.

Ejemplo 3.2.1. Supónga una población con valores $1, 2, \ldots, 100$ y se desea estimar el total, $t = \sum_{i=1}^{100} y_i = \sum_{i=1}^{100} i = 5050$. Para ello se eligen muestras sistemáticas de tamaño $n = 10$ y se comparan seis órdenes de la población, los tres primeros fueron tomados de Särndal et al. [54] y los restantes de Muzquiz [40]:

a) Creciente
b) Alternando de manera creciente y decreciente por bloques
c) Creciente dentro de bloques



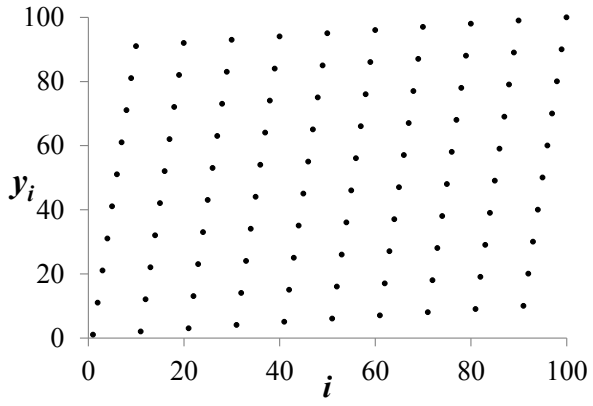
	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	1	2	3	4	5	6	7	8	9	10
	20	19	18	17	16	15	14	13	12	11
	21	22	23	24	25	26	27	28	29	30
	40	39	38	37	36	35	34	33	32	31
	41	42	43	44	45	46	47	48	49	50
	60	59	58	57	56	55	54	53	52	51
	61	62	63	64	65	66	67	68	69	70
	80	79	78	77	76	75	74	73	72	71
	81	82	83	84	85	86	87	88	89	90
	100	99	98	97	96	95	94	93	92	91
t_{s_r}	505	505	505	505	505	505	505	505	505	505

Figura 3.3: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada alternando de manera creciente y decreciente por bloques. $V_{SY}\left(\hat{t}_\pi\right) = 0$, $SSW =$ 83,325 y $SSB = 0$.

d) Triangular
e) Decreciente y creciente
f) Triangular dentro de bloques.

En las Figuras 3.2 a 3.7 se muestra la población ordenada por cada uno de los seis criterios y los valores de y_i para las $T = 10$ muestras posibles según el orden.

La varianza del estimador del total, según el orden de la población, se presenta en el Cuadro 3.2. Como puede observarse el mejor orden no es único, hay cuatro casos donde la varianza es cero: b) alternar de manera creciente y decreciente por

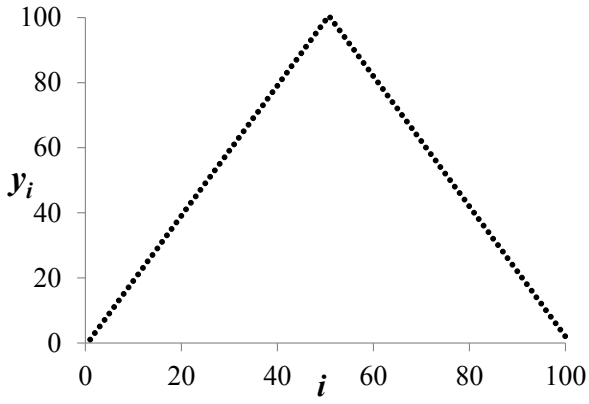


	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	1	11	21	31	41	51	61	71	81	91
	2	12	22	32	42	52	62	72	82	92
	3	13	23	33	43	53	63	73	83	93
	4	14	24	34	44	54	64	74	84	94
	5	15	25	35	45	55	65	75	85	95
	6	16	26	36	46	56	66	76	86	96
	7	17	27	37	47	57	67	77	87	97
	8	18	28	38	48	58	68	78	88	98
	9	19	29	39	49	59	69	79	89	99
	10	20	30	40	50	60	70	80	90	100
t_{s_r}	55	155	255	355	455	555	655	755	855	955

Figura 3.4: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada de manera creciente dentro de bloques. $V_{SY}\left(\hat{t}_\pi\right) = 8{,}250{,}000$, $SSW = 825$ y $SSB = 82{,}500$.

bloques, d) triangular, e) decreciente y creciente y f) triangular dentro de bloques.

Los dos arreglos de la población que generaron una varianza del estimador positiva fueron el orden creciente y el orden creciente dentro de bloques. Cuando la población se ordena de manera creciente las primeras muestras, $s_1, \ldots, s_5$, subestiman el total y las muestras $s_6, \ldots, s_{10}$, sobreestiman el total poblacional. La varianza del estimador del total bajo el orden creciente fue de 82,500, dando lugar a un coeficiente de variación del total de 5.7 %, la varianza del estimador bajo el orden creciente dentro de bloques fue de 8,250,000 y el coeficiente de variación del

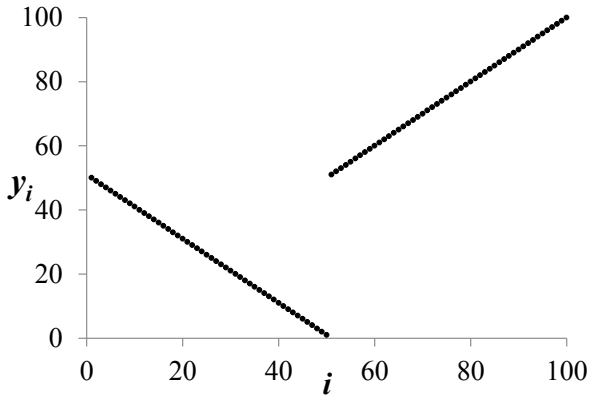


	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	1	3	5	7	9	11	13	15	17	19
	21	23	25	27	29	31	33	35	37	39
	41	43	45	47	49	51	53	55	57	59
	61	63	65	67	69	71	73	75	77	79
	81	83	85	87	89	91	93	95	97	99
	100	98	96	94	92	90	88	86	84	82
	80	78	76	74	72	70	68	66	64	62
	60	58	56	54	52	50	48	46	44	42
	40	38	36	34	32	30	28	26	24	22
	20	18	16	14	12	10	8	6	4	2
t_{s_r}	505	505	505	505	505	505	505	505	505	505

Figura 3.5: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada de manera triangular. $V_{SY}\left(\hat{t}_\pi\right) = 0$, $SSW = 83{,}325$ y $SSB = 0$.

total de 56.9 %. Si en lugar de efectuar un muestreo sistemático se selecciona una muestra aleatoria simple sin reemplazo (SI por *sampling irrestricted*) del mismo tamaño, la varianza correspondiente sería 757,500, es decir, 9.18 veces la varianza del muestreo sistemático aplicado al orden creciente. Como ya se mencionó, el muestreo sistemático no siempre es más eficiente que el muestreo aleatorio simple, en este caso, cuando la población se ordenó de manera creciente dentro de bloques la varianza fue casi 11 veces la del aleatorio simple.

Muzquiz [40] además de encontrar que los órdenes dados en los incisos b, d, e, y f tienen varianza cero, calculó la varianza bajo muestreo sistemático para otras seis

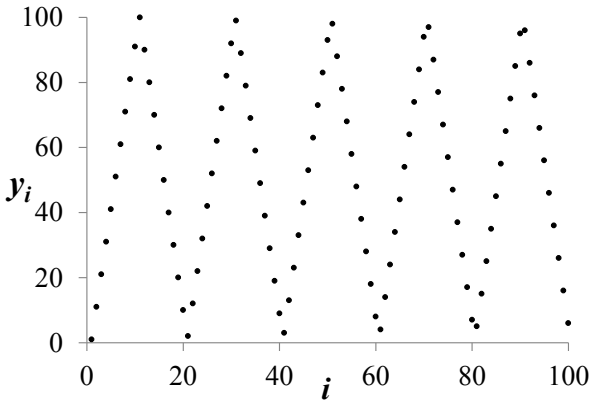


	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	50	49	48	47	46	45	44	43	42	41
	40	39	38	37	36	35	34	33	32	31
	30	29	28	27	26	25	24	23	22	21
	20	19	18	17	16	15	14	13	12	11
	10	9	8	7	6	5	4	3	2	1
	51	52	53	54	55	56	57	58	59	60
	61	62	63	64	65	66	67	68	69	70
	71	72	73	74	75	76	77	78	79	80
	81	82	83	84	85	86	87	88	89	90
	91	92	93	94	95	96	97	98	99	100
t_{s_r}	505	505	505	505	505	505	505	505	505	505

Figura 3.6: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada decreciente y creciente. $V_{SY}\left(\hat{t}_\pi\right) = 0$, $SSW = 83{,}325$ y $SSB = 0$.

maneras de acomodar la población de este ejemplo, incluyendo el orden creciente. Encontró que todas tuvieron varianza positiva pero menor a la del SI. Entre estas seis formas, cuando la población ordenada quedó como $\{100, 98, \ldots, 2, 99, 97, \ldots, 1\}$ generó una varianza mayor a la del orden creciente; cuando la población ordenada quedó como $\{50, 49, \ldots, 1, 100, 99, \ldots, 51\}$ la varianza del estimador es igual a la del orden creciente. Para este ejemplo, el orden que induce a una varianza cero no es único, más aun, no se encontró cuál es el número de formas de ordenar a las unidades que llevan a este resultado.

En la siguiente sección se describirá una variante del muestreo sistemático que coinci-



	Muestra									
	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
y_i	1	11	21	31	41	51	61	71	81	91
	100	90	80	70	60	50	40	30	20	10
	2	12	22	32	42	52	62	72	82	92
	99	89	79	69	59	49	39	29	19	9
	3	13	23	33	43	53	63	73	83	93
	98	88	78	68	58	48	38	28	18	8
	4	14	24	34	44	54	64	74	84	94
	97	87	77	67	57	47	37	27	17	7
	5	15	25	35	45	55	65	75	85	95
	96	86	76	66	56	46	36	26	16	6
t_{s_r}	505	505	505	505	505	505	505	505	505	505

Figura 3.7: Muestras posibles de la población con valores $1, 2, \ldots, 100$ ordenada de manera triangular dentro de bloques. $V_{SY}\left(\hat{t}_\pi\right) = 0$, $SSW = 83,325$ y $SSB = 0$.

de con el muestreo sistemático cuando la población se ordena de manera equilibrada, como aquí se llama, y que corresponde al orden triangular en el ejemplo anterior.

3.3. Muestreo sistemático equilibrado

El muestreo sistemático equilibrado (BSS) probablemente fue denominado así por Sethi al describir el método óptimo para seleccionar un par de unidades, con diferentes probabilidades, ordenadas de manera creciente o decreciente. Sethi [48] §9.4 señala "this method of selection shall be called Balanced Systematic Sampling (of size two)".

Cuadro 3.2: Varianza del estimador del total $t = \sum_{i=1}^{100} i$, según tipo de orden

Orden	$V_{SY}\left(\hat{t}_\pi\right)$	$V_{SI}\left(\hat{t}_\pi\right)$	SSW	SSB	SST
a) Creciente	82,500	757,500	82,500	825	83,325
b) Alternando por bloques	0	757,500	83,325	0	83,325
c) Creciente dentro de bloques	8,250,000	757,500	825	82,500	83,325
d) Triangular	0	757,500	83,325	0	83,325
e) Decreciente y creciente	0	757,500	83,325	0	83,325
f) Triangular dentro de bloques	0	757,500	83,325	0	83,325

Es Murthy [39], p. 165, quien describe esta técnica de muestreo para $n > 2$, la cual consiste en dividir la población en $n/2$ grupos de $2T$ unidades cada uno y seleccionar de cada grupo un par de unidades equidistantes del inicio y del fin del grupo de manera sistemática. Si r se elige con igual probabilidad del conjunto $\{1, \ldots, T\}$, en el primer grupo de $2T$ unidades, las unidades en la muestra serán las etiquetadas por r y $2T - r + 1$; en el segundo grupo las unidades en muestra corresponderán a $r + 2T$ y $4T - r + 1$, y así sucesivamente. Entonces, la muestra s'_r constará de las unidades

$r + 2Ti$ y $2T(i + 1) - r + 1$ con $i = 0, 1, \ldots, n/2 - 1$ si n es par, o bien

$r + 2Ti$, $2T(i + 1) - r + 1$ y $r + (n - 1)T$ con $i = 0, 1, \ldots, (n - 1)/2 - 1$ si n es impar.

Por otra parte, Gundersen [20] en lo que él denomina el fraccionador suave (del inglés *smooth fractionator*), propone aplicar un muestreo sistemático a unidades acomodadas simétricamente con el objetivo de tener un diseño eficiente. La población se ordena simétricamente o de manera equilibrada, como aquí se le referirá, de la siguiente manera:

$y_{(1)}, y_{(3)}, \ldots, y_{(N-3)}, y_{(N-1)}, y_{(N)}, y_{(N-2)}, \ldots, y_{(4)}, y_{(2)}$ si N es par,

$y_{(1)}, y_{(3)}, \ldots, y_{(N-2)}, y_{(N)}, y_{(N-1)}, y_{(N-3)}, \ldots, y_{(4)}, y_{(2)}$ si N es impar

donde $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$. La población del Ejemplo 3.2.1 está acomodada de esta manera en el orden llamado triangular.

Cuando la población se ordena de manera equilibrada, se obtiene el mismo conjunto de muestras sistemáticas, $\{s_1, s_2, \ldots, s_T\}$, que cuando se aplica muestreo sistemático equilibrado a una población ordenada de manera creciente, $\{s'_1, s'_2, \ldots, s'_T\}$, es decir, son dos algoritmos de selección distintos que producen el mismo conjunto de muestras. Adicionalmente, $p(s_r) = p(s'_r) = 1/T$ $(r = 1, \ldots, T)$, entonces ambos algoritmos corresponden al mismo diseño. Esto se muestra en el ejemplo siguiente.

Ejemplo 3.3.1. Suponga que se tiene una población con $N = 20$ elementos y se seleccionará una muestra de tamaño $n = 4$ mediante BSS, donde la población está ordenada de manera creciente, esto es

$$y_{(1)}, y_{(2)}, \ldots, y_{(19)}, y_{(20)}.$$

Entonces, las posibles muestras y los respectivos valores que toma y para cada una de estas son los siguientes.

Muestra	Valores de y
$s'_1 = \{1, 10, 11, 20\}$	$y_{(1)}, y_{(10)}, y_{(11)}, y_{(20)}$
$s'_2 = \{2, 9, 12, 19\}$	$y_{(2)}, y_{(9)}, y_{(12)}, y_{(19)}$
$s'_3 = \{3, 8, 13, 18\}$	$y_{(3)}, y_{(8)}, y_{(13)}, y_{(18)}$
$s'_4 = \{4, 7, 14, 17\}$	$y_{(4)}, y_{(7)}, y_{(14)}, y_{(17)}$
$s'_5 = \{5, 6, 15, 16\}$	$y_{(5)}, y_{(6)}, y_{(15)}, y_{(16)}$.

Si la población está ordenada de manera equilibrada, esto es

$$y_{(1)}, y_{(3)}, \ldots, y_{(17)}, y_{(19)}, y_{(20)}, y_{(18)}, \ldots, y_{(4)}, y_{(2)},$$

y se selecciona una muestra sistemática, las posibles muestras y los valores de y son

Muestra	Valores de y
$s_1 = \{1, 11, 20, 10\}$	$y_{(1)}, y_{(11)}, y_{(20)}, y_{(10)}$
$s_2 = \{3, 13, 18, 8\}$	$y_{(3)}, y_{(13)}, y_{(18)}, y_{(8)}$
$s_3 = \{5, 15, 16, 6\}$	$y_{(5)}, y_{(15)}, y_{(16)}, y_{(6)}$
$s_4 = \{7, 17, 14, 4\}$	$y_{(7)}, y_{(17)}, y_{(14)}, y_{(4)}$
$s_5 = \{9, 19, 12, 2\}$	$y_{(9)}, y_{(19)}, y_{(12)}, y_{(2)}$.

Como puede verse en este ejemplo $\{s_1, s_2, s_3, s_4, s_5\} = \{s'_1, s'_2, s'_3, s'_4, s'_5\}$, con $p(s_r) = p(s'_r) = 1/5$ $(r = 1, \ldots, 5)$.

Murthy [39], p. 165, señala que el BSS reduce la varianza del estimador de la media cuando la población presenta una tendencia lineal. Demuestra que en una población hipotética cuyos valores de interés siguen una progresión aritmética, $y_i = a + bi$, $i = 1, \ldots, N$, si N y n son pares, entonces la varianza del estimador de la media es igual a cero si la muestra se obtiene usando BSS.

Sethi [48, 49] demostró que si una población de N elementos, con N par, se ordena de manera creciente de acuerdo a la característica de interés, es decir, $y_1 \leq y_2 \leq \cdots \leq y_N$, y si L es una función de pérdida que para todo $|u| \leq |u'|$ satisface:

$$L(|A + u'|) + L(|A - u'|) \geq L(|A + u|) + L(|A - u|), \text{ con } A, u, u' \in \mathbb{R}, \quad (3.8)$$

entonces la manera óptima de elegir una muestra de tamaño 2, o un par óptimo como él lo llama, es elegir a las unidades equidistantes del inicio y del fin del arreglo, etiquetadas por r y $N + 1 - r$, donde r se elige con probabilidad $2/N$ del conjunto $\{1, \ldots, N/2\}$, este algoritmo de selección corresponde al caso del BSS con $n = 2$. Esta elección es óptima en el sentido que mínimiza cualquier función de pérdida como L. Sethi señala que en particular, la varianza del estimador del total o de la media la cumple.

Para demostrar lo anterior, Sethi propone dos conjuntos de muestras posibles ℓ y ℓ'. Entre las $N/2$ muestras posibles, ℓ contiene las muestras $s_1 = (1, i)$ y $s_2 = (j, N)$, mientras que ℓ' contiene las muestras $s_1' = (1, N)$ y $s_2' = (i, j)$, el resto de las muestras son las mismas en ambos conjuntos.

Sea e_{ij} el error absoluto de la estimación del total basada en el par (i, j), esto es,

$$e_{ij} = \left| \frac{N}{2}(y_i + y_j) - t \right|.$$

Si se elige s_1 el error absoluto está dado por

$$\begin{aligned} e_{1i} &= \left| \frac{N}{2}(y_1 + y_i) - t \right| \\ &= \left| \frac{N}{4}(y_1 + y_i + y_j + y_N) - t - \frac{N}{4}\left[(y_j + y_N) - (y_1 + y_i)\right] \right| \\ &= |A - e_1| \end{aligned}$$

donde $A = \frac{N}{4}(y_1 + y_i + y_j + y_N) - t$ y $e_1 = \frac{N}{4}\left[(y_j + y_N) - (y_1 + y_i)\right]$.

Análogamente, el error absoluto para cada uno de los pares en s_2, s_1', s_2' se puede escribir como sigue,

$$e_{jN} = |A + e_1|$$

$$e_{1N} = |A - e_2|$$

$$e_{ij} = |A + e_2|$$

donde $e_2 = \frac{N}{4}\left[(y_1 + y_N) - (y_i + y_j)\right]$.

Como puede verse $|e_2| < |e_1|$. Entonces, si la función de pérdida L cumple la condición (3.8), se tiene que

$$L(|A + e_1|) + L(|A - e_1|) \geq L(|A + e_2|) + L(|A - e_2|),$$

o equivalentemente,

$$L(|e_{jN}|) + L(|e_{1i}|) \geq L(|e_{ij}|) + L(|e_{1N}|).$$

Por lo tanto, la función de pérdida para el conjunto ℓ' es menor que para el conjunto ℓ. Siguiendo este procedimiento repetidamente se llega a que el conjunto con todos los pares de unidades equidistantes, $\ell_0 = \{(r, N + 1 - r), \quad r = 1, \ldots, N/2\}$, es el conjunto óptimo para cualquier función de pérdida que satisfaga (3.8). Señala que las funciones no decrecientes y cóncavas hacia abajo satisfacen la condición (3.8). Este es el caso de $L(e) = ke^2$ y por lo tanto de la varianza del estimador del total o de la media. Entonces, ℓ_0 lleva a una varianza mínima.

Sethi [49], p. 317, menciona que para $n = 2$, la varianza del estimador será igual a cero si la población es simétrica, textualmente "... if the population were symmetrical the risk would reduce to zero."

Como se vió en §3.1, el muestreo sistemático es equivalente a seleccionar uno de T conglomerados con n unidades cada uno. Bajo esta nomenclatura, el problema del par óptimo analizado por Sethi, corresponde a formar T conglomerados de tamaño 2 y elegir uno de ellos aleatoriamente. También analiza la mejor manera de formar conglomerados con $n > 2$ unidades.

A través de un ejemplo en el que compara dos poblaciones ordenadas de manera creciente, con $N = 6$ elementos cada una y $n = 3$, muestra que la manera óptima

de seleccionar las dos muestras posibles es diferente entre ambas poblaciones. Los valores que toma la variable de interés y para cada una de las poblaciones hipotéticas se presentan en el Cuadro 3.3.

Cuadro 3.3: Poblaciones hipotéticas, ejemplo Sethi [49] ,§6

Población	y_1	y_2	y_3	y_4	y_5	y_6
I	1	2	4	8	16	32
II	1	2	3	4	5	6

En la población I, la manera óptima de formar los dos conglomerados es $s_1 = \{u_1, u_2, u_6\}$ y $s_2 = \{u_3, u_4, u_5\}$. Esta no es la mejor manera de elegir las dos muestras en la población II. Concluye que para $n > 2$ se deberán de hacer otras consideraciones además del orden de la población para obtener una manera de formar conglomerados óptima.

Describe un método para obtener lo que define como un arreglo de conglomerados cercano al óptimo como sigue. Supóngase que hay $u_1, u_2, \ldots, u_{Tn}$ unidades en la población y ésta se divide en T conglomerados de n unidades cada una, sean dos de los conglomerados que tienen los siguientes valores de la característica y.

Conglomerado i, $C_i : y_{i1} \leq y_{i2} \leq \cdots \leq y_{in}$

Conglomerado i', $C_i : y_{i'1} \leq y_{i'2} \leq \cdots \leq y_{i'n}$.

Si sucede que

$$y_{i1} < y_{i'j} \quad j = 1, 2, \ldots, n, \tag{3.9}$$

entonces intercambiando u_{ij} con $u_{i'j}$ en los dos conglomerados llevará a una reducción de la varianza. Un arreglo de conglomerados que no satisface la condición (3.9) será denominado arreglo de conglomerados cercano al óptimo (del inglés *nearly optimum clustering*).

Señala que una manera simple de obtener un arreglo de conglomerados cercano al óptimo es la siguiente:

i. Ordenar las unidades en orden creciente de acuerdo a la característica de interés.

ii. Dividirlas en n estratos homogéneos de T unidades cada uno. Las primeras T unidades forman el primer estrato, las segundas T el segundo y así sucesivamente.

iii. Tomar los estratos en pares. El arreglo sugerido tiene conglomerados que incluyen un par óptimo de unidades de cada uno de los pares de estratos.

Hay un gran número de maneras de formar un arreglo de conglomerados cercano al óptimo, dependiendo de los pares de estratos que se formen. Uno de ellos es

$C_i : u_{2kT+i},\ u_{2(k+1)T+1-i}$ donde $k = 0, 1, \ldots$ toma los valores para los cuales exista una unidad en la población.

Es importante señalar que el BSS corresponde a este último arreglo de conglomerados.

Murthy [39], p. 167, señala que una mejor manera de seleccionar una muestra de $n = 2^m$ elementos, $m \in \mathbb{Z}^+$, a partir de una población con un número de elementos N, múltiplo de 2^m, puede ser formando las $N/2$ posibles muestras de tamaño $n = 2$ bajo BSS, luego usar esos $N/2$ pares para formar $N/4$ muestras de $n = 4$ unidades, ordenando los $N/2$ pares de manera creciente de acuerdo al total de la característica de interés y tomando esos pares de una manera equilibrada. Siguiendo el procedimiento anterior sucesivamente, se pueden formar todas las posibles muestras de tamaño $2, 4, 8, \ldots$, y finalmente tomar una de las muestras con igual probabilidad. A continuación se presenta un ejemplo en el que se sigue el método anterior y se compara con el BSS.

Ejemplo 3.3.2. Sea una población hipotética de $N = 20$ elementos y se desea seleccionar una muestra de $n = 4 = 2^2$ elementos. Los valores y_i $(i = 1, \ldots, 20)$ que toma la variable de interés, ordenados de manera creciente son los siguientes:

$$\{1,\ 2,\ 5,\ 6,\ 6,\ 7,\ 8,\ 8,\ 9,\ 10,\ 10,\ 12,\ 13,\ 15,\ 15,\ 18,\ 19,\ 20,\ 22,\ 25\}$$

Aplicando el método señalado por Murthy, los 10 pares óptimos y su respectivo total son como sigue.

r	Par óptimo	Total	r	Par óptimo	Total
1	$(1, 25)$	26	6	$(7, 15)$	22
2	$(2, 22)$	24	7	$(8, 15)$	23
3	$(5, 20)$	25	8	$(8, 13)$	21
4	$(6, 19)$	25	9	$(9, 12)$	21
5	$(6, 18)$	24	10	$(10, 10)$	20

Los pares ordenados de manera creciente son

Orden	Par óptimo	Total	Orden	Par óptimo	Total
1	$(10, 10)$	20	6	$(6, 18)$	24
2	$(8, 13)$	21	7	$(2, 22)$	24
3	$(9, 12)$	21	8	$(5, 20)$	25
4	$(7, 15)$	22	9	$(6, 19)$	25
5	$(8, 15)$	23	10	$(1, 25)$	26

Aplicando nuevamente el BSS a estos 10 pares, finalmente se tienen las siguientes muestras posibles

Muestra	Valores y_i, $i \in s_r$	Total muestral, t_{s_r}
s_1	$\{10, 10, 1, 25\}$	46
s_2	$\{8, 13, 6, 19\}$	46
s_3	$\{9, 12, 5, 20\}$	46
s_4	$\{7, 15, 2, 22\}$	46
s_5	$\{8, 15, 6, 18\}$	47

Bajo BSS los valores de y en cada una de las muestras posibles y su respectivo total muestral son

Muestra	Valores y_i, $i \in s_r$	Total muestral, t_{s_r}
s_1	$\{1, 10, 10, 25\}$	46
s_2	$\{2, 9, 12, 22\}$	45
s_3	$\{5, 8, 13, 20\}$	46
s_4	$\{6, 8, 15, 19\}$	48
s_5	$\{6, 7, 15, 18\}$	46

La varianza del estimador del total bajo el primer método es $V_{MUR}\left(\hat{t}_\pi\right) = 4$ y bajo BSS es $V_{BSS}\left(\hat{t}_\pi\right) = 24$. Como se esperaba, el método sugerido por Murthy es más

eficiente que el BSS, sin embargo, una desventaja del método es que el tamaño de muestra debe de satisfacer $n = 2^m$. Hay $T = N/n$ muestras posibles, al igual que en el BSS.

No debe de confundirse el muestreo sistemático equilibrado con otro método de selección denominado muestreo equilibrado (del inglés *balanced sampling*), el cual no es objeto de estudio de este trabajo. Uno de los primeros trabajos en los que se menciona este método corresponde a Yates [61], §3.13, quien propone un método para seleccionar muestras equilibradas. De acuerdo con la definición dada en Tillé [56], el muestreo equilibrado es un método aleatorio de selección de unidades de una población, que genera una muestra tal que los estimadores Horvitz-Thompson de los totales de las p variables auxiliares, son iguales a los verdaderos totales de la población, esto es, un diseño muestral es equilibrado en las variables auxiliares $X_1, \ldots, X_p$ si y sólo si satisface las ecuaciones

$$\sum_{i \in s} \frac{x_{ij}}{\pi_i} = \sum_{i \in U} x_{ij}$$

para toda $s \in \{s_1, s_2, \ldots, s_M\}$ tal que $p(s) > 0$ y para toda $j = 1, \ldots, p$.

3.4. Estimador de la integral de una función acotada y su varianza

El muestreo sistemático también se aplica en áreas como integración numérica, Geoestadística, análisis de imágenes y Estereología, en donde la población de interés puede estar definida en un dominio continuo, por ejemplo un tumor, un corte de tejido, un campo agrícola o un bosque. En esta situación la variable de estudio se puede expresar como una función f y un parámetro de interés puede ser la integral de la función $t = \int_{\mathbb{R}} f(x)dx$, que en muchas situaciones corresponde al área. En esta sección se verá el estimador de t y su varianza cuando la función cumple ciertas condiciones.

Sea $f : \mathbb{R} \mapsto \mathbb{R}$ una función acotada y fija, de cuadrado integrable en un intervalo $A = [a, b]$, esto es $\int_A |f(x)|^2 \, dx < \infty$, y fuera de A es cero. Cuando se realiza un muestreo sistemático con un intervalo muestral $T > 0$, $f(x)$ se observa en la muestra $s = s_r = \{r + (j-1)T, j \in \mathbb{Z}\} \cap A$ con $r \sim U(0, T)$, el tamaño de muestra n_s es una

variable aleatoria. Gráficamente se tiene que

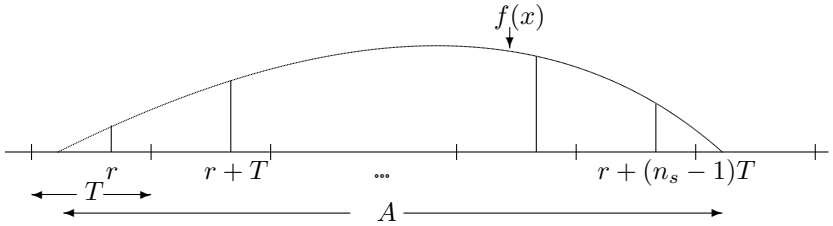


Figura 3.8: Selección sistemática de una población continua. Las unidades seleccionadas corresponden a $r, r+T, \ldots, r+(n_s-1)T$.

Un estimador de t, análogo al estimador (3.6), es:

$$\hat{t} = T \sum_{x \in s_r} f(x). \tag{3.10}$$

El estimador anterior es insesgado, como se demuestra a continuación, tomando como base la dada por Cruz-Orive [8].

$$\begin{aligned} E\left[\hat{t}\right] &= E\left[T \sum_{x \in s_r} f(x)\right] \\ &= E\left[T \sum_{j \in \mathbb{Z}} f(r + [j-1]T)\right], \end{aligned}$$

puesto que $f(x)$ se observa en un número finito de puntos:

$$\begin{aligned} E\left[\hat{t}\right] &= T \sum_{j \in \mathbb{Z}} E\left[f(r + [j-1]T)\right] \\ &= T \sum_{j \in \mathbb{Z}} \int_0^T f(r + [j-1]T) \frac{1}{T} dr \\ &= \sum_{j \in \mathbb{Z}} \int_{(j-1)T}^{jT} f(v) dv \\ &= \int_{\mathbb{R}} f(x) dx = t. \quad \square \end{aligned}$$

Una representación muy conocida de la varianza de (3.10) está dada por la expresión (ver Matheron [35], p. 73)

$$V_{SY}\left(\hat{t}\right) = T \sum_{k=-\infty}^{\infty} g(kT) - \int_{\mathbb{R}} g(h) dh$$

donde $g(h) = \int_{\mathbb{R}} f(x)f(x+h)dx$ es el covariograma de la función f. Esta varianza puede verse como el error al aproximar la integral del covariograma por la suma $T\sum g(kT)$.

Kiêu [32] y Kiêu et al. [33] suponen que la función f es suave a trozos de orden $(q, 1)$, donde q el orden de la primera derivada de f no continua. Si $f^{(k)}$ denota la $k-$ésima derivada de f, la amplitud de su salto está dado por

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

siempre que el límite exista. Su soporte es el conjunto $Df^{(k)} = \left\{x : Sf^{(k)}(x) \neq 0\right\}$.

Una función f de soporte acotado es suave a trozos (q, s), $q,\ s \in \mathbb{N} \cup \{0\}$, si

a) $Df^{(k)} = \emptyset$, $k = 0, 1, ..., q-1$ y $Df^{(q)} \neq \emptyset$ y

b) $f^{(k)}$ tiene un número finito de saltos y estos son finitos, $k = q, q+1, ..., q+s$.

Con base en la teoría transitiva de Matheron, estos autores derivan una expresión para $V_{SY}\left(\hat{t}\right)$:

$$\begin{aligned} V_{SY}(\hat{t}) = (-1)^q T^{2q+2} P_{2q+2,T}(0) \sum_{v\in Df^{(q)}} (Sf^{(q)}(v))^2 \\ + (-1)^q T^{2q+2} \sum_{\substack{u,v\in Df^{(q)},\\ v-u\neq 0}} P_{2q+2,T}(v-u) Sf^{(q)}(u) Sf^{(q)}(v) + o(T^{2q+2}), \end{aligned} \tag{3.11}$$

donde

$P_{l,T}(x) = P_l\left(\frac{x}{T} - \left[\frac{x}{T}\right]\right)$ es el polinomio modificado de Bernoulli de orden l, $[x]$ denota la parte entera de x. Este polinomio es una función con período T que se define por $P_l(x) = B_l(x)/l!$ con $B_l(x)$ el $l-$ésimo polinomio de Bernoulli, $l \geq 1$.

La ecuación (3.11) se representa también de la siguiente manera

$$V_{SY}(\hat{t}) = V_E(\hat{t}) + Z(T) + o(T^{2q+2}),$$

donde

i. $V_E(\hat{t}) = (-1)^q T^{2q+2} P_{2q+2,T}(0) \sum_{v \in Df^{(q)}} (Sf^{(q)}(v))^2$ es el llamado término de extensión y sólo depende de las amplitudes de las transiciones de $f^{(q)}$.

ii. $Z(T) = (-1)^q T^{2q+2} \sum_{\substack{u,v \in Df^{(q)}, \\ v-u \neq 0}} P_{2q+2,T}(v-u) Sf^{(q)}(u) Sf^{(q)}(v)$ 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. $o(T^{2q+2})$ es una función tal que $\lim_{T \to 0} o(T^{2q+2})/T^{2q+2} = 0$.

Cuando T es suficiente pequeño, $V_E(\hat{t})$ constituye una buena aproximación de $V_{SY}(\hat{t})$ ya que representa la tendencia de la varianza.

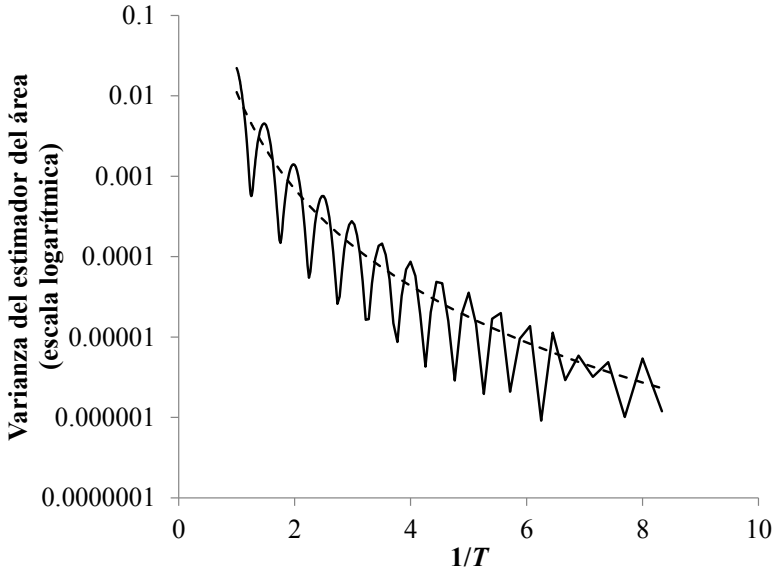


Figura 3.9: Término de extensión, $V_E(\hat{t}) = T^4/90$ y *Zitterbewegung*, $Z(T) = -T^4 B_4 (2/T - [2/T])/3$ del estimador del área bajo la función $f(x) = 2x - x^2$, $x \in [0, 2]$. La línea - - - corresponde a $V_E(\hat{t})$ y la línea —— a $V_E(\hat{t}) + Z(T)$.

Ejemplo 3.4.1. Sea $f(x) = 2x - x^2$ con $x \in [0, 2]$ y $t = \int_0^2 f(x)dx$. En este caso $q = 1$ ya que $f^{(q)}(x)$ no es continua en $x = 0$ y $x = 2$; luego, las transiciones son $Sf^{(1)}(0) = 2$ y $Sf^{(1)}(2) = 2$. Entonces, $V_E(\hat{t}) = T^4/90$ y $Z(T) = -T^4 B_4 (2/T - [2/T])/3$, donde $B_4(x) = x^2(x-1)^2 - 1/30$.

El término de extensión y la suma de este más el *Zitterbewegung* se muestran en la Figura 3.9. En ella se observa que si se ignoran los términos de orden superior

$o(T^4)$, V_E (línea punteada) corresponde a la tendencia de $V_{SY}(\hat{t})$; el comportamiento ondulatorio que se observa en la línea continua, $V_E + Z_(T)$, se debe al *Zitterbewegung*.

Souchet [52] y Kiêu [32] demuestran que si la función f es suave a trozos (q, s), entonces su covariograma g es suave a trozos $(2q+1, s)$ y que existe una relación entre las transiciones del covariograma y las transiciones de la función

$$Sg^{(2q+1)}(c) = -(-1)^q \sum_{\substack{u,v \in Df^{(q)}, \\ v-u=c}} Sf^{(q)}(u) Sf^{(q)}(v).$$

Como el covariograma tiene la propiedad de que $g(h) = g(-h)$, las transiciones del covariograma alrededor del origen se pueden expresar en términos del covariograma, $Sg^{(2q+1)}(0) = 2g^{(2q+1)}(0^+)$. Por lo tanto, en términos del covariograma el término de extensión puede escribirse como

$$V_E(\hat{t}) = -\frac{2B_{2q+2}}{(2q+2)!} T^{2q+2} g^{(2q+1)}(0^+) \tag{3.12}$$

donde

$g^{(2q+1)}(0^+) = \frac{-(-1)^q}{2} \sum_{t \in Df^{(q)}} Sf^{(q)}(t)$.

$B_l = B_l(0)$ es el $l-$ésimo número de Bernoulli. Los primeros números de Bernoulli son: $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_4 = -1/30$ y los números impares son igual a cero, $B_3 = B_5 = B_7 = \cdots = 0$.

Como puede apreciarse, la varianza del estimador depende de q, el orden de la primera derivada de f no continua, así como de la magnitud de sus saltos o transiciones. Si T es pequeña, la tasa de convergencia de la varianza de $\hat{t}$ es de orden T^{2q+2}, esto es, $V_{SY}(\hat{t}) \leq cT^{2q+2}$ con c una constante. Entonces, en el peor de los casos, cuando f no es continua ($q = 0$), la convergencia es de orden T^2, mientras que en el muestreo aleatorio simple es de orden T.

Si la varianza del estimador se aproxima mediante el término de extensión, para obtener un estimador de la varianza es necesario conocer q el grado de suavidad de la función, así como tener una estimación de la $2q+1-$ésima derivada del covariograma cerca del origen. Ya que generalmente $g^{(2q+1)}(0^+)$ se desconoce, se puede suponer un modelo, Kiêu [32] obtiene una aproximación del covariograma usando la fórmula de Taylor, como se verá en el Capítulo 6.

Capítulo 4

Condiciones suficientes para que la varianza del estimador sea cero

La varianza del estimador del total de una población finita $V_{SY}\left(\hat{t}_{\pi}\right)$ dada en la expresión (3.7) depende del orden de la población, así también la varianza del estimador del área bajo una función, $V_{SY}\left(\hat{t}\right)$, proporcionada en la ecuación (3.11) depende de la suavidad de la función. Para saber si el muestreo sistemático es más eficiente que el aleatorio simple es necesario tener información sobre la estructura de la población.

En el desarrollo de este trabajo se encontraron tres condiciones, aunque no necesarias al menos suficientes, bajo las cuales la varianza del muestreo sistemático es cero. La primera se refiere al tamaño de la muestra y al tamaño de la población, o bien al intervalo muestral, la segunda al orden de la población y la tercera a una propiedad de simetría en la variable de interés. Estas condiciones se establecen de manera precisa en la proposición 4.1, en el caso de una población finita, y en la proposición 4.2, en el caso de una población continua.

4.1. Total de una población finita

Como se mencionó anteriormente, Sethi [49] señala que la varianza del estimador del total será igual a cero si la población es simétrica, si tiene un número N par de unidades, se ordena de manera creciente y se seleccionan dos unidades equidistantes del inicio y del fin de este orden, r y $N+1-r$, donde r es una realización de la distribución uniforme discreta, con $r = 1, \ldots, N/2$. Murthy [39] §5.9d muestra que

en una población de N elementos, con $y_i = a + bi$ $(i = 1, \ldots, N)$ y N par, si se elige una muestra de tamaño n par mediante muestreo sistemático equilibrado (BSS), entonces la varianza del estimador de la media es cero.

Sethi no contempla el caso $n > 2$ y la población que considera Murthy es muy específica. Como se verá en el siguiente ejemplo, estos resultados sobre la varianza nula se pueden extender a tamaños de muestra pares y otras poblaciones.

Ejemplo 4.1.1. Un conjunto hipotético de 20 datos. Suponga una población de $N = 20$ unidades cuyos valores y_i, $i = 1, \ldots, 20$, una vez ordenados (ver figura 4.1), son los siguientes :

$$\{1,\ 6,\ 12,\ 26,\ 32,\ 38,\ 43,\ 53,\ 65,\ 71,\ 73,\ 68,\ 62,\ 48,\ 42,\ 36,\ 31,\ 21,\ 9,\ 3\}.$$

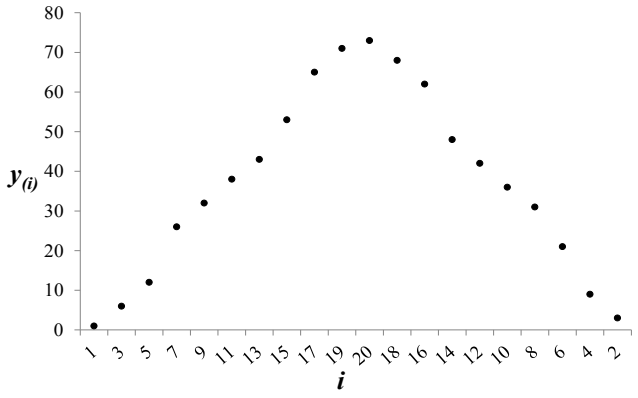


Figura 4.1: Valores y_i, $i = 1, \ldots, 20$, del Ejemplo 4.1.1, ordenados de manera equilibrada. Si $n = 2$ o $n = 4$, entonces $V_{SY}\left(\hat{t}_\pi\right) = 0$.

Hay cinco posibles muestras sistemáticas de tamaño $n = 4$, los valores observados en cada una de ellas y el total muestral se presentan en el Cuadro 4.1. Como el estimador del total $\hat{t}_{s_r} = 740$, $r = 1, \ldots, 5$, entonces $V_{SY}\left(\hat{t}_\pi\right) = 0$. Es fácil verificar que si $n = 2$, el estimador del total $\hat{t}_{s_r} = 740$, $r = 1, \ldots, 10$, entonces la varianza del estimador es cero.

En este ejemplo, el total se estima sin error por varias razones: a) la población hipotética tiene la característica de que es simétrica con respecto a su media, esto es, la distancia entre $\overline{y}_U - y_{(i)}$ es la misma que entre $y_{(N-i+1)} - \overline{y}_U$, $i = 1, \ldots, 10$,

b) la manera como se ordenó la población, c) N y n son números pares y d) se eligió una muestra sistemática. Estas condiciones se establecen de manera formal en la siguiente proposición.

Cuadro 4.1: Valores y_i en las cinco muestras sistemáticas del ejemplo 4.1.1

Muestra	Valores y_i, $i \in s_r$	Total muestral, t_{s_r}	$\hat{t}_\pi$
s_1	$\{1, 38, 73, 36\}$	148	5(148)
s_2	$\{6, 43, 68, 31\}$	148	5(148)
s_3	$\{12, 53, 62, 21\}$	148	5(148)
s_4	$\{26, 65, 48, 9\}$	148	5(148)
s_5	$\{32, 71, 42, 3\}$	148	5(148)

Proposición 4.1. *Sea y_i, $i = 1, \ldots, N$, el valor de la variable de interés para la $i-$ésima unidad en una población con N unidades, n el tamaño de muestra y T el intervalo muestral. Si se satisface que:*

i) *N y n son números pares y $T = N/n \in \mathbb{Z}^+$.*

ii) *La población se ordena de manera equilibrada, esto es,*

$$y_{(1)}, y_{(3)}, \ldots, y_{(N-1)}, y_{(N)}, \ldots, y_{(4)}, y_{(2)}, \quad \text{donde } y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}. \tag{4.1}$$

iii) *Los valores ordenados $y_{(i)}$ son simétricos respecto a la media poblacional $\overline{y}_U$, esto es, cumplen*

$$y_{(N-i+1)} - \overline{y}_U = \overline{y}_U - y_{(i)}, \quad i = 1, \ldots, N/2 \tag{4.2}$$

donde $\overline{y}_U = \sum_{i=1}^{N} y_i/N$.

Entonces, bajo muestreo sistemático, el estimador del total poblacional $\hat{t}_\pi = T \sum_{i \in s_r} y_i$ tiene varianza cero, es decir, $V_{SY}\left(\hat{t}_\pi\right) = 0$.

Demostración. Considerando muestreo sistemático y que la población se ordena de acuerdo con (4.1), los elementos en la muestra son:

$$\begin{aligned} s_r = &\{2r-1, 2r-1+2T, \ldots, 2r-1+2(n/2-2)T, 2r-1+2(n/2-1)T, \\ &2(n/2)T-2r+2, 2(n/2-1)T-2r+2, \ldots, 4T-2r+2, 2T-2r+2\}. \end{aligned}$$

donde r representa el arranque aleatorio y se eligió con probabilidad $1/T$ del conjunto $\{1,\ldots,T\}$.

Suponiendo (4.2) se tiene que

$$\begin{aligned}
\hat{t}_\pi &= T\sum_{i\in s_r} y_{(i)} \\
&= T\sum_{i=1}^{n/2} y_{(2r-1+2[i-1]T)} + T\sum_{i=1}^{n/2} y_{(2[n/2-i+1]T-2r+2)} \\
&= T\sum_{i=1}^{n/2} y_{(2r-1+2[i-1]T)} + T\sum_{i=1}^{n/2} y_{(N-2[i-1]T-2r+2)} \\
&= T\sum_{i=1}^{n/2} y_{(2r-1+2[i-1]T)} + T\sum_{i=1}^{n/2} \left[2\overline{y}_U - y_{(2[i-1]T+2r-1)}\right] \\
&= T2\overline{y}_U\frac{n}{2} = t.
\end{aligned}$$

Por lo tanto, $V_{SY}\left(\hat{t}_\pi\right) = 0$. □

La proposición anterior también se cumple si en lugar de ordenar a la población de acuerdo con (4.1) se ordena de cualquiera de las siguientes maneras

$$y_{(2)}, y_{(4)}, \ldots, y_{(N)}, y_{(N-1)}, \ldots, y_{(3)}, y_{(1)},$$

$$y_{(N)}, y_{(N-2)}, \ldots, y_{(2)}, y_{(1)}, \ldots, y_{(N-3)}, y_{(N-1)}, \quad \textit{donde } y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}.$$

Sin embargo, existen otros órdenes que también producen una varianza cero, como se establece en el Corolario 4.1.

Como se mostró en §3.3, cuando la población está ordenada de manera creciente, el BSS corresponde al mismo diseño que el muestreo sistemático cuando la población está ordenada de manera equilibrada, $y_{(1)}, y_{(3)}, \ldots, y_{(N-1)}, y_{(N)}, \ldots, y_{(4)}, y_{(2)}$, entonces,

$$V_{SY}\left(\hat{t}_\pi\right) = V_{BSS}\left(\hat{t}_\pi\right) = 0,$$

esto es, si en lugar de ordenar la población siguiendo la condición ii), se ordena de manera creciente y se mantienen las otras dos condiciones, el total se estima de manera exacta bajo muestreo sistemático equilibrado.

Corolario 4.1. *Si se mantienen las condiciones* i) *y* iii) *de la Proposición 4.1, la varianza del estimador,* $V_{SY}\left(\hat{t}_\pi\right)$*, será nula para cualquier orden tal que si* $y_{(i)} \in s_r$*,* $y_{(N+1-i)}$ *también pertenecerá a la muestra.*

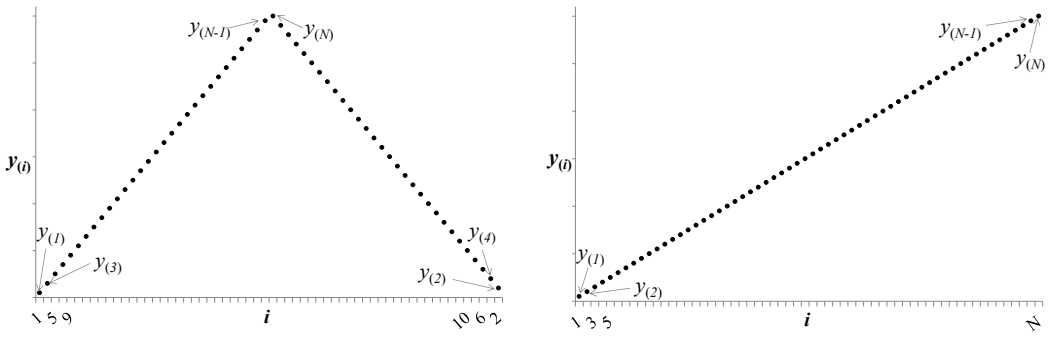
Demostración. De la condición iii) se tiene que $y_{(i)} + y_{(N-i+1)} = 2\overline{y}_U$ y como $y_{(i)} \in s_r$ y $y_{(N+1-i)} \in s_r$, entonces

$$\begin{aligned}\hat{t}_\pi &= T\sum_{i\in s_r} y_{(i)} \\ &= T\frac{n}{2}2\overline{y}_U = t.\end{aligned}$$

Por lo tanto, $V_{SY}\left(\hat{t}_\pi\right) = 0$. □

A continuación se da otro ejemplo que cumple con las condiciones de la Proposición 4.1.

Ejemplo 4.1.2. Los primeros N **números naturales**. Sea una población con valores $1, 2, \ldots, N$, con N par, n el tamaño de muestra y $T = N/n \in \mathbb{N}$ el intervalo muestal, donde la población se ordena de manera equilibrada como en la condición (4.1) (ver Figura 4.2 (a)). Si se selecciona una muestra bajo muestreo sistemático con n par, entonces $V_{SY}\left(\hat{t}_\pi\right) = 0$.



(a) Orden equilibrado (b) Orden creciente

Figura 4.2: Población con valores $1, 2, \ldots, N$, según orden.

Antes de verificar las condiciones de la proposición, suponga que $N = 20$ y $n = 4$. Los valores y_i en cada una de las cinco posibles muestras son $s_1 : \{1, 11, 20, 10\}$, $s_2 : \{3, 13, 18, 8\}$, $s_3 : \{5, 15, 16, 6\}$, $s_4 : \{7, 17, 14, 4\}$ y $s_5 = \{9, 19, 12, 2\}$. Como

$\hat{t}_\pi = t = 210$ para todas las muestras, la varianza del estimador es cero. Se puede probar que esto es cierto en general, con $T = N/n \in \mathbb{N}$, N y n números pares.

Para mostrar que la varianza es cero, por medio de la proposición, sólo falta verificar la condición (4.2). Como $\overline{y}_U = (N+1)/2$, entonces

$$y_{(N-i+1)} - \overline{y}_U = N - i + 1 - \frac{N+1}{2} = \frac{N+1}{2} - i = \overline{y}_U - y_{(i)}, \quad i = 1, 2, \ldots, N.$$

Por lo tanto, $V_{SY}\left(\hat{t}_\pi\right) = 0$.

Como puede verse en el ejemplo 3.2.1 en el que $N = 100$, los órdenes dados en b), e) y f) también producen una varianza igual a cero. En estos casos no se cumple la condición (4.1), que se refiere al orden equilibrado de la población, sin embargo, estos órdenes cumplen la condición dada en el Corolario 4.1.

Por otra parte, si la población se ordena de manera creciente, como se muestra en la Figura 4.2 (b), entonces

$$\begin{aligned}\hat{t}_\pi = \hat{t}_{s_r} &= T \sum_{i \in s_r} y_i \\ &= T \sum_{i=1}^{n} \left[r + T(i-1)\right] = N\left[r + \frac{N}{2}\left(\frac{n-1}{n}\right)\right].\end{aligned}$$

La varianza del estimador depende de r y por lo tanto $V_{SY}\left(\hat{t}_\pi\right) > 0$.

Otro ejemplo hipotético en el que la varianza del estimador es cero y no se cumplen las condiciones (4.1) y (4.2), es el siguiente. Supóngase una población que toma los valores $\{1, 3, 2, 3, 4, 4, 5, 5, 5, 5, 7, 6, 6, 8, 7, 13, 11, 12, 9, 9\}$, si se selecciona una muestra sistemática de $n = 4$ elementos, se puede ver fácilmente que $\hat{t}_\pi = \hat{t}_{s_r} = 125$, $r = 1, \ldots, 5$ y por ende $V_{SY}\left(\hat{t}_\pi\right) = 0$.

4.2. Área bajo una función

De manera análoga a las condiciones que se establecieron en la Proposición 4.1 para que la varianza del estimador del total de una población finita sea nula, en la Proposición 4.2 se darán las condiciones para que el área bajo el gráfico de la función f se estime de manera exacta, antes se considera un ejemplo.

Ejemplo 4.2.1. Área de un triángulo isósceles. Sea $f(x)$ la función (ver Figura 4.3) definida por

$$f(x) = \begin{cases} 2x & \text{si } x \in [0, 1/2] \\ 2 - 2x & \text{si } x \in [1/2, 1]. \end{cases}$$

Si se selecciona una muestra bajo muestreo sistemático con n par y $T = 1/n$, entonces el área del triángulo, $t = \int_0^1 f(x)dx$, se estima de manera exacta, como se verá en seguida.

El estimador de t está dado por

$$\begin{aligned} \hat{t} &= T \sum_{x_i \in s_r} f(x_i) \\ &= T \sum_{i=1}^{n/2} f\left(r + (i-1)T\right) + T \sum_{i=n/2+1}^{n} f\left(r + (i-1)T\right) \\ &= T \sum_{i=1}^{n/2} 2\left[r + (i-1)T\right] + T \sum_{i=n/2+1}^{n} \left\{2 - 2\left[r + (i-1)T\right]\right\}. \end{aligned}$$

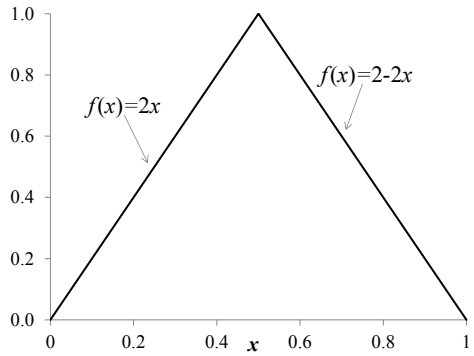


Figura 4.3: Triángulo isósceles. Si $T = 1/n$ y n es par, $\hat{t}$ estima de manera exacta el área del triángulo.

Haciendo un cambio de variable en los límites de la segunda suma

$$\begin{aligned} \hat{t} &= T \sum_{i=1}^{n/2} 2\left[r + (i-1)T\right] + T \sum_{i=1}^{n/2} \left\{2 - 2\left[r + (i-1+n/2)T\right]\right\} \\ &= 2T\frac{n}{2} - 2T^2 \frac{n}{2}\frac{n}{2} = \frac{1}{2} = t. \end{aligned}$$

Por lo tanto, el área bajo $f(x)$ se estima de manera exacta, esto es, $V_{SY}\left(\hat{t}\right) = 0$. Este ejemplo se presenta en Mattfeldt [37], donde se compara la precisión del muestreo sistemático de seis modelos geométricos de $f(x)$, uno de los cuales corresponde al triángulo isósceles.

Las condiciones para que la integral de $f(x)$ se estime sin error se refieren al tamaño de muestra, al intervalo muestral y a dos tipos de simetrías. La primera de estas simetrías corresponde a que la función debe de ser simétrica con respecto a la recta vertical que pasa por el punto medio del intervalo $A = [a, 2b - a]$ (rango de integración); un ejemplo se muestra en la Figura 4.4 (b), en la que la gráfica es simétrica con respecto a la recta $y = b$. Si la función es simétrica con respecto al eje de las ordenadas, se trata de una función par.

La segunda trata de que la función sea simétrica en $[a, b]$ con respecto al punto $(x_0 = (a + b)/2, f(x_0))$; en la Figura 4.4 (c) se presenta una función simétrica con respecto al punto x_0, en donde puede verse que la distancia entre $f(x_0 + \epsilon)$ y $f(x_0)$ es la misma que entre $f(x_0 - \epsilon)$ y $f(x_0)$. Un caso particular corresponde a una función impar la cual es simétrica con respecto al origen, $(x_0 = 0, f(x_0) = 0)$. Estos requerimientos se precisan a continuación.

Proposición 4.2. *Sean $f(x)$ una función que es de cuadrado integrable en un intervalo $A = [a, 2b-a]$ y fuera de A es cero, n el tamaño de muestra y T el intervalo muestral. Si se satisface que:*

i) *n es un número par y $T = 2(b - a)/n$, con $a < b$.*

ii) *La función tiene la forma siguiente*

$$f(x) = \begin{cases} w(x) & \text{si } x \in [a, b] \\ w(2b - x) & \text{si } x \in [b, 2b - a]. \end{cases} \tag{4.3}$$

iii) *Para $x \in [a, b]$, $w(x)$ es simétrica respecto al punto $(x_0 = (a + b)/2, w(x_0))$, esto es,*

$$w\left(x_0 + \epsilon\right) - w\left(x_0\right) = w\left(x_0\right) - w\left(x_0 - \epsilon\right) \tag{4.4}$$

para todo ϵ tal que $0 \leq \epsilon \leq (b - a)/2$.

Entonces, bajo muestreo sistemático, $\hat{t} = T\sum_{x\in s_r} f(x)$ *estima el área* $t = \int_{\mathbb{R}} f(x)dx$ *sin error, esto es,* $V_{SY}\left(\hat{t}\right) = 0$.

Demostración. Como la función es de cuadrado integrable, el área t y la varianza del estimador $\hat{t}$ existen. La muestra está formada por las unidades

$$s_r = \{x_i = a + r + (i-1)T; i = 1, \ldots, n\}$$

donde r es un número que se elige con igual probabilidad en el intervalo $[0, T]$. Entonces

$$\begin{aligned}\hat{t} &= T\sum_{x_i\in s_r} f(x_i) \\ &= T\sum_{i=1}^{n/2} f\left(a + r + (i-1)T\right) + T\sum_{i=n/2+1}^{n} f\left(a + r + (i-1)T\right).\end{aligned} \tag{4.5}$$

El segundo sumando de la última igualdad en (4.5) se puede reescribir como

$$\begin{aligned}\sum_{i=n/2+1}^{n} f\left(a + r + (i-1)T\right) &= \sum_{i=n/2+1}^{n} w\left(2b - a - r - (i-1)T\right) \\ &= \sum_{i=1}^{n/2} w\left(2b - a - r - (i + n/2 - 1)T\right) \\ &= \sum_{i=1}^{n/2} w\left(b - (r + (i-1)T)\right).\end{aligned}$$

Como la condición de simetría dada en (4.4) es equivalente a $w\left(2x_0 - x\right) = 2w\left(x_0\right) - w\left(x\right)$, entonces

$$\begin{aligned}\sum_{i=1}^{n/2} w\left(b - (r + (i-1)T)\right) &= \sum_{i=1}^{n/2} w\left(a + b - (a + r + (i-1)T)\right) \\ &= \sum_{i=1}^{n/2}\left[2w\left(\frac{a+b}{2}\right) - w\left(a + r + (i-1)T\right)\right] \\ &= nw\left(\frac{a+b}{2}\right) - \sum_{i=1}^{n/2} w\left(a + r + (i-1)T\right).\end{aligned} \tag{4.6}$$

Sustituyendo la última igualdad de (4.6) en la ecuación (4.5) se tiene

$$\begin{aligned}\hat{t} &= T\sum_{i=1}^{n/2} w\left(a+r+(i-1)T\right) + Tnw\left(\frac{a+b}{2}\right) - T\sum_{i=1}^{n/2} w\left(a+r+(i-1)T\right) \\ &= Tnw\left(\frac{a+b}{2}\right).\end{aligned}$$

A partir de la condición iii) se obtiene que $w(x_o) = \left[w\left(x_0+\epsilon\right) + w\left(x_0-\epsilon\right)\right]/2$, tomando $\epsilon = (b-a)/2$ se tiene que $w\left([a+b]/2\right) = [w(a)+w(b)]/2$, luego

$$\hat{t} = (b-a)\left[w(a)+w(b)\right] = t.$$

Es decir, el parámetro de interés t se estima de manera exacta para cualquier muestra s_r, $r = 1, \ldots, T$, esto es

$$V_{SY}\left(\hat{t}\right) = 0. \qquad \square$$

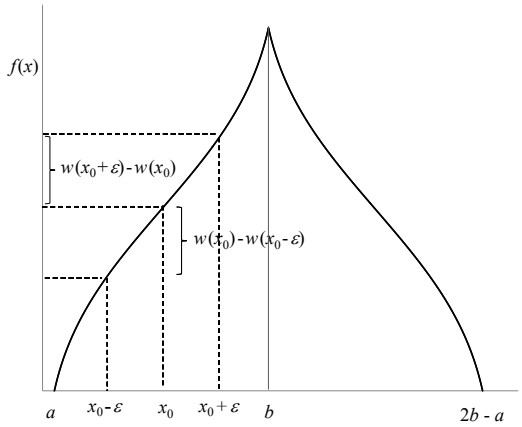
De la definición de $f(x)$ y la condición (4.4), la función $w(2b-x)$ con $x \in [b, 2b-a]$ también será simétrica con respecto al punto $(x_0 = (3b-a)/2, w(x_0))$. Los supuestos dados en (4.1) y (4.3) se refieren a simetría en el orden de la población, lo que aquí se refiere como orden equilibrado, mientras que las condiciones (4.2) y (4.4) suponen simetría con respecto a un punto. En el caso continuo, en la Figura 4.4 (a) se muestra una función que cumple las condiciones (4.3) y (4.4), en el panel (b) se presenta una función que sólo cumple la condición dada en (4.3) y en el panel (c) una función que sólo cumple la dada en (4.4).

A continuación se dan dos ejemplos en los que el área bajo la función se estima de manera exacta.

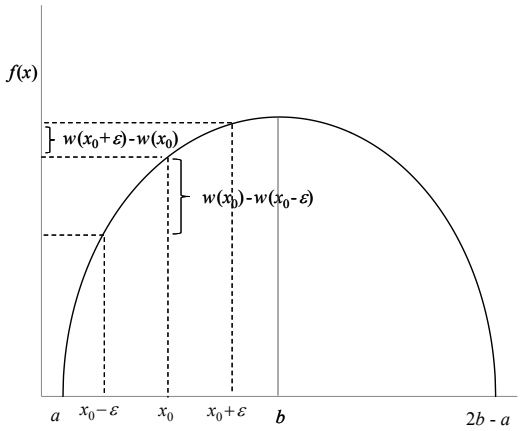
Ejemplo 4.2.2. Función triangular. Sea $f(x)$ la función que se define como

$$f(x) = \begin{cases} m(x-c_1)+c_2 & \text{si } x \in [a,b] \\ m(2b-x-c_1)+c_2 & \text{si } x \in [b, 2b-a], \end{cases}$$

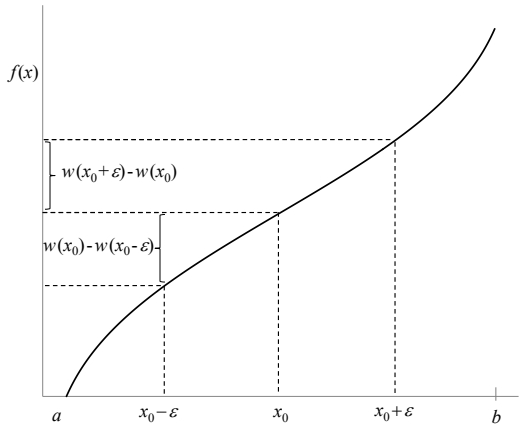
donde m, c_1 y c_2 son constantes (ver Figura 4.5 (a)). Entonces el área bajo $f(x)$ se estima de manera exacta cuando se elige una muestra sistemática con n par y $T = 2(b-a)/n$.



(a) Función que cumple las condiciones (4.3) y (4.4).



(b) Función que cumple la condición (4.3), pero no la (4.4) .



(c) Función que cumple la condición (4.4), pero no la (4.3).

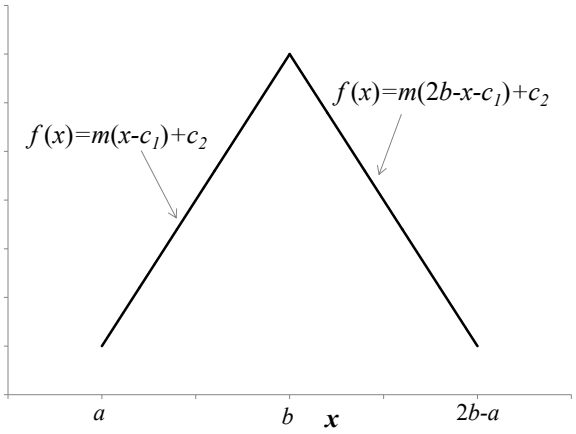
Figura 4.4: Ejemplos de funciones para ilustrar las condiciones de la Proposición 4.2.

La condición dada en (4.3) se cumple por definición de $f(x)$, enseguida se verifica que el supuesto dado en (4.4) también se cumple. Sea $x \in [a, b]$ y $0 \leq \epsilon \leq (b-a)/2$, entonces

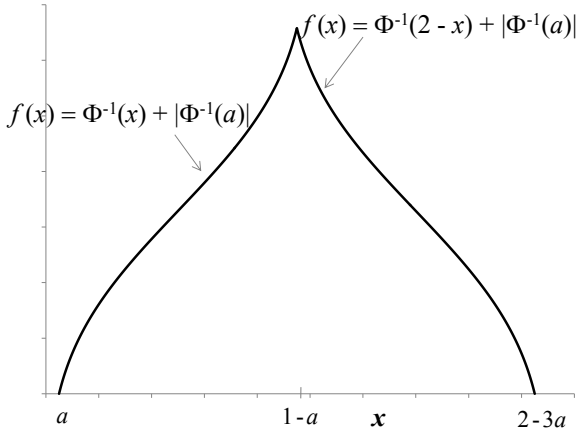
$$w(x_0 + \epsilon) - w(x_0) = m(x_0 + \epsilon - c_1) + c_2 - m(x_0 - c_1) - c_2 = m\epsilon,$$

$$w(x_0) - w(x_0 - \epsilon) = m(x_0 - c_1) + c_2 - m(x_0 - \epsilon - c_1) - c_1 = m\epsilon.$$

Como se verifican las condiciones, entonces $V_{SY}\left(\hat{t}\right) = 0$. El Ejemplo 4.2.1 es un caso particular con $m = 2$, $a = 0$, $b = 1/2$, $c_1 = 0$ y $c_2 = 0$.



(a) Función triangular, donde m, c_1 y c_2 son constantes.



(b) Inversa de la función de distribución de una variable aleatoria continua y simétrica $\Phi^{-1}(x)$. Para fines ilustrativos se graficó la inversa de la distribución normal.

Figura 4.5: Funciones señaladas en los ejemplos 4.2.2 y 4.2.3 en los que $V_{SY}\left(\hat{t}\right) = 0$ si n es par, $T = 2(b-a)/n$ en el panel (a) y $T = 2(1-2a)/n$ en el panel (b).

Ejemplo 4.2.3. Inversa de la función de distribución de una variable

aleatoria. Sea Z una variable aleatoria continua y simétrica con función de densidad dada por $f_Z(z)$ y sea $f(x)$ la función definida mediante

$$f(x) = \begin{cases} \Phi^{-1}(x) + |\Phi^{-1}(a)| & \text{si } 0 < a \leq x \leq 1-a \\ \Phi^{-1}(2-2a-x) + |\Phi^{-1}(a)| & \text{si } 1-a \leq x \leq 2-3a \end{cases}$$

donde $\Phi^{-1}(x)$ denota la inversa de la función de distribución de la variable aleatoria Z para el cuantil x, es decir, $x = \int_{-\infty}^{z} f_Z(s)ds$ con $z = \Phi^{-1}(x)$ (ver Figura 4.5 (b)). Si se selecciona una muestra sistemática con n par y $T = 2(1-2a)/n$, entonces $\hat{t}$ estima de manera exacta el área $t = 2(1-2a)\left|\Phi^{-1}(a)\right|$, esto es, $V_{SY}\left(\hat{t}\right) = 0$.

El supuesto dado en (4.3) se cumple por definición de $f(x)$. Falta verificar que el supuesto de que $w(x)$ es simétrica con respecto al punto $(x_0 = \frac{1}{2}, w(x_0))$ se satisface. Sea $x \in [a, 1-a]$ y $0 \leq \epsilon \leq (1-2a)/2$, entonces

$$\begin{aligned} w\left(x_0+\epsilon\right) - w\left(x_0\right) &= \Phi^{-1}\left(x_0+\epsilon\right) + \left|\Phi^{-1}(a)\right| - \Phi^{-1}\left(x_0\right) - \left|\Phi^{-1}(a)\right| \\ &= \Phi^{-1}\left(\frac{1}{2}+\epsilon\right), \text{ y} \end{aligned}$$

$$\begin{aligned} w\left(x_0\right) - w\left(x_0-\epsilon\right) &= \Phi^{-1}\left(x_0\right) + \left|\Phi^{-1}(a)\right| - \Phi^{-1}\left(x_0-\epsilon\right) - \left|\Phi^{-1}(a)\right| \\ &= -\Phi^{-1}\left(\frac{1}{2}-\epsilon\right) \\ &= \Phi^{-1}\left(\frac{1}{2}+\epsilon\right). \end{aligned}$$

Por lo tanto, $V_{SY}\left(\hat{t}_\pi\right) = 0$.

Es importante mencionar que los resultados dados en las proposiciones 4.1 y 4.2 sólo constituyen una guía teórica porque involucran el conocimiento de la variable de interés y, para toda la población, o bien de la función $f(x)$ en el intervalo de integración, en cuyo caso el parámetro de interés se podría calcular de manera exacta sin necesidad de seleccionar una muestra.

Adicionalmente, si se sabe que la población finita es simétrica respecto a su media, o que la función es simétrica respecto al punto medio del intervalo $[a, b]$, el parámetro de interés puede calcularse de manera exacta, sin necesidad de seleccionar una muestra, si se conocen y_1 y y_N en el caso discreto, $f(a)$ y $f(b)$ o bien $f\left((a+b)/2\right)$ en el caso continuo. No obstante, si estos puntos se desconocen, pero se conocen y_2 y

y_{N-1} o $f(a+\Delta)$ y $f(b-\Delta)$, también se puede calcular de manera exacta el parámetro. En general, lo mismo sucede para cualquier muestra bajo BSS con $n = 2$.

Sin embargo, seleccionar una muestra de un número par de unidades a partir de una población finita con un número de elementos múltiplo del tamaño de muestra, o bien, de una población continua con un intervalo muestral igual a $2(b-a)/n$, podría ayudar a disminuir la varianza del estimador del total bajo muestreo sistemático, aunque esto no necesariamente se cumple en todos los casos. En el siguiente capítulo se demuestra que esta estrategia funciona en el caso de la distribución uniforme que es una variable aleatoria simétrica y en el capítulo 7, usando cinco conjuntos de datos, se muestra que la estrategia también funciona en la mayoría de los casos.

Capítulo 5

Varianza del muestreo sistemático suponiendo una superpoblación

El enfoque basado en una superpoblación constituye una herramienta importante en el desarrollo de la teoría y práctica del muestreo. El concepto de superpoblación tiene diferentes usos (ver por ejemplo Cassel et al. [5]), entre los que se encuentra utilizarlo simplemente como un instrumento matemático, no asociado con algún proceso físico o creencia subjetiva, para hacer derivaciones teóricas explícitas. En este capítulo se obtiene la varianza del muestreo sistemático, o una aproximación de ésta, suponiendo cuatro modelos de superpoblación específicos, y se compara respecto al muestreo aleatorio simple. Los resultados de carácter original se presentan y demuestran en las proposiciones respectivas.

Bajo esta estrategia, la muestra puede verse como un proceso en dos etapas, cada una de las cuales contribuye a la aleatoriedad: por una parte los valores observados $y_1, \ldots, y_N$ son realizaciones de las variables $Y_1, \ldots, Y_N$, las cuales siguen un modelo de superpoblación m, y por la otra, la muestra s es seleccionada de acuerdo con el diseño de muestreo $p(\cdot)$.

El modelo m supone que $Y_1, \ldots, Y_N$ son variables aleatorias independientes e idénticamente distribuidas con función de densidad conjunta $f_{Y_1,\ldots,Y_N}(y_1, \ldots, y_N; \underline{\mu}) = \prod_{i=1}^{N} f_{Y_i}(y_i; \underline{\mu})$, donde $\underline{\mu}$ es el vector de parámetros de $f_{Y_1,\ldots,Y_N}(\cdot)$. Se consideran cuatro distribuciones para Y_i $(i = 1, \ldots, N)$: uniforme, Laplace, normal y normal generalizada.

El diseño $p(\cdot)$ corresponde al muestreo aleatorio simple, o bien, al muestreo sistemático. Como en este último diseño el orden de la población es importante, se supondrán dos órdenes: equilibrado y creciente. Por esta razón es necesario introducir las estadísticas de orden, como se explica a continuación.

Si los valores de la población finita $y_1, \ldots, y_N$ se consideran realizaciones de la función de densidad $f_{Y_1,\ldots,Y_N}(\cdot)$ y si $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$ son las observaciones ordenadas de manera creciente, entonces el valor de $y_{(i)}$ cambiará en cada población finita; a la variable aleatoria $Y_{(i)}$ correspondiente a $y_{(i)}$ se le conoce como la $i-$ésima estadística de orden, $i = 1, \ldots, N$. Formalmente, las estadísticas de orden se definen como sigue.

Definición 5.1. Sean $Y_1, \ldots, Y_N$ variables aleatorias reales. Sean $Y_{(1)} \leq \cdots \leq Y_{(N)}$ que denotan los valores ordenados de $Y_1, \ldots, Y_N$. Entonces $Y_{(1)}, \ldots, Y_{(N)}$ son llamadas las estadísticas de orden de $Y_1, \ldots, Y_N$.

5.1. Predictor de Horvitz-Thompson

Bajo el esquema de superpoblaciones la suma o total, $\theta = \sum_{i=1}^{N} Y_i$, es una variable aleatoria, a diferencia de los capítulos 3 y 4, en los que el total t es una cantidad fija. Cuando se elige una muestra sistemática, el predictor de Horvitz-Thompson de θ es

$$\hat{\theta}_\pi = \hat{\theta}_{s_r} = T \sum_{i \in s_r} Y_i,$$

donde T es el intervalo muestral, N corresponde al tamaño de la población, n es el tamaño de muestra y $s = s_r$ denota la muestra seleccionada. Por las razones señaladas en el capítulo 4, primera parte, se supondrá que $T = N/n \in \mathbb{N}$.

Fuller [12] §2.9, señala que el término predictor se emplea en algunas áreas de la estadística para denotar una función de observaciones que aproxima una cantidad aleatoria desconocida, en este caso θ, y se usa el término estimador cuando aproxima una cantidad fija desconocida, y que algunos autores usan el término estimador en ambos casos. Aquí $\hat{\theta}_\pi$ denota el predictor de la suma de las variables aleatorias y $\hat{t}_\pi$ el estimador de la suma o total de las observaciones; en términos de las variables aleatorias $Y_1, \ldots, Y_N$, $\hat{t}_\pi$ es condicional a $Y_i = y_i$, $i = 1, \ldots, N$.

Suponiendo que la población está ordenada bajo un criterio específico, entonces el predictor de Horvitz-Thompson puede escribirse como

$$\hat{\theta}_\pi = \hat{\theta}_{s_r} = T \sum_{i \in s_r} Y_{(i)}, \tag{5.1}$$

donde $Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}$ denotan las estadísticas de orden de las variables aleatorias $Y_1, Y_2, \ldots, Y_N$.

Como se verá en §5.3, para obtener la varianza de este predictor, se necesitan el valor esperado y la covarianza de las estadísticas de orden bajo el modelo m.

El valor esperado de la $i-$ésima estadística de orden de $Y_1, Y_2, \ldots, Y_N$ variables aleatorias independientes, con función de distribución $F_Y(y)$, es

$$\mu_i = E\left[Y_{(i)}\right] = C_{i,N} \int_0^1 F^{-1}(u) u^{i-1} (1-u)^{N-i} du, \quad i = 1, \ldots, N$$

donde $C_{i,N} = N!/\left[(i-1)!(N-i)!\right]$ y $u = F_Y(y)$ es absolutamente continua. La covarianza de $Y_{(i)}$ y $Y_{(j)}$, $i \leq j \leq N$, se obtiene mediante

$$\begin{aligned} \mu_{ij} &= C\left[Y_{(i)}, Y_{(j)}\right] \\ &= C_{i,j,N} \int_0^1 \int_0^v \left[F^{-1}(u) - \mu_{(i)}\right] \left[F^{-1}(v) - \mu_{(j)}\right] u^{i-1} (v-u)^{j-i-1} (1-v)^{N-j} du dv, \end{aligned}$$

donde $C_{i,j,N} = N!/\left[(i-1)!(j-i-1)!(N-j)!\right]$, ver por ejemplo David y Nagaraja [10], §3.

5.2. Modelos de superpoblación y órdenes considerados

Los modelos que se suponen corresponden a la distribución uniforme, Laplace, normal y normal generalizada. Se trabajó con estas distribuciones porque sus funciones de densidad son simétricas, característica similar a la tercera condición de las proposiciones 4.1 y 4.2, para que la varianza del estimador sea cero en el caso de una población fija.

Inicialmente, era de interés la distribución normal o gaussiana, dada la importancia que tiene en la estadística; sin embargo, como lo señalan David y Nagaraja [10]

(p. 40), los primeros momentos de las estadísticas de orden se pueden obtener de manera explícita sólo para distribuciones muy simples como la uniforme y la exponencial. Por ello, se trabajó con la distribución uniforme, cuyas expresiones son cerradas y sencillas. Posteriormente, se consideró la distribución de Laplace; en este caso, se utilizaron aproximaciones de la media y la covarianza de las estadísticas de orden, porque las expresiones exactas son complicadas y no fue posible simplificar la expresión de la varianza de $\hat{\theta}_\pi$. Después, se retomó la distribución normal usando también aproximaciones. Finalmente, por medio de un estudio de simulación, la varianza del predictor se comparó suponiendo la distribución normal generalizada.

Se suponen dos maneras de ordenar la población: equilibrada y creciente. La primera se eligió porque autores como Sethi [49] (para $n = 2$), Murthy [39], así como Gundersen [20], encontraron que es una manera de reducir varianza. La segunda manera se usa comúnmente en la práctica porque para reducir la varianza del muestreo sistemático, se deben de colocar unidades similares juntas, y ésta es una manera de hacerlo; también se podría usar un orden decreciente y se obtendrían resultados análogos.

a) Orden equilibrado. Las variables aleatorias $Y_1, Y_2, \ldots, Y_N$, con N par, ordenadas de manera equilibrada quedan de la siguiente manera

$$Y_{(1)}, Y_{(3)}, \ldots, Y_{(N-3)}, Y_{(N-1)}, Y_{(N)}, Y_{(N-2)}, \ldots, Y_{(4)}, Y_{(2)}.$$

donde $Y_{(i)}$, $i = 1, \ldots, N$ denota la $i-$ésima estadística de orden. También se podrían ordenar de manera inversa, esto es,

$$Y_{(2)}, Y_{(4)}, \ldots, Y_{(N-2)}, Y_{(N)}, Y_{(N-1)}, Y_{(N-3)}, \ldots, Y_{(3)}, Y_{(1)}.$$

Un ejemplo se muestra en la Figura 5.1 (a) en la que los valores $y_1, y_2, \ldots, y_N$ se generaron a partir de una distribución normal y se ordenaron de manera equilibrada.

Bajo el orden equilibrado y muestreo sistemático, las estadísticas de orden en la muestra cuando el arranque aleatorio es $r = 1, 2, \ldots, T$ son

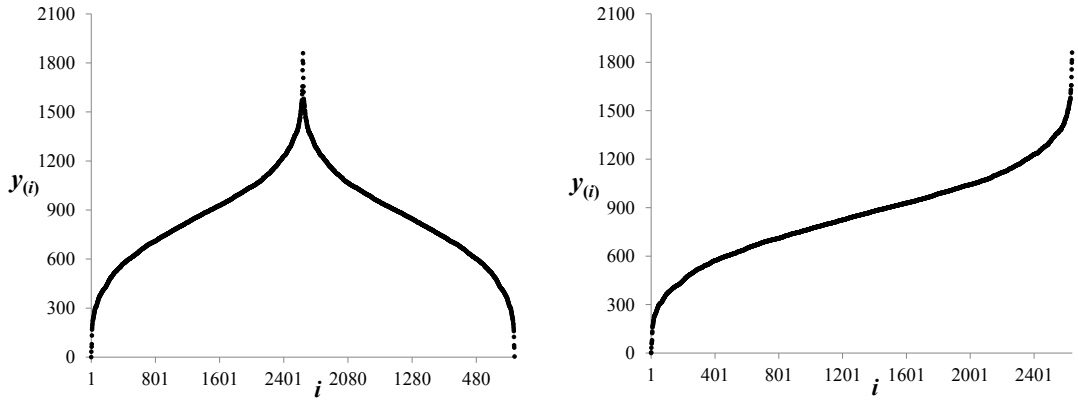
$$Y_{(2r-1)}, Y_{(2r-1+2T)}, \ldots, Y_{(2r-1+2(n/2-2)T)}, Y_{(2r-1+2(n/2-1)T)},$$

$$Y_{(2(n/2)T-2r+2)}, Y_{(2(n/2-1)T-2r+2)}, \ldots, Y_{(4T-2r+2)}, Y_{(2T-2r+2)}.$$

Por lo tanto, la muestra resultante s_r consta de las estadísticas de orden etiquetadas mediante

$$s_r = \{2r-1, 2r-1+2T, \ldots, 2r-1+2(n/2-2)T, 2r-1+2(n/2-1)T, \\ 2(n/2)T-2r+2, 2(n/2-1)T-2r+2, \ldots, 4T-2r+2, 2T-2r+2\}.$$

Si por ejemplo $N = 20$, $n = 4$ y $r = 2$ entonces la muestra queda conformada por las estadísticas de orden $Y_{(3)}, Y_{(13)}, Y_{(18)}, Y_{(8)}$, o equivalentemente por las $Y_{(i)}, i \in s_r$ donde $s_r = \{3, 13, 18, 8\}$. Puesto que el muestreo sistemático equilibrado bajo un orden creciente y el muestreo sistemático cuando la población está ordenada de manera equilibrada, corresponden al mismo diseño, en lo sucesivo al muestreo sistemático bajo el orden equilibrado se le denotará de manera abreviada por BSS.



(a) Orden equilibrado, $y_{(1)}, y_{(3)}, \ldots, y_{(4)}, y_{(2)}$

(b) Orden creciente, $y_{(1)}, y_{(2)}, \ldots, y_{(N)}$

Figura 5.1: Valores $y_1, y_2, \ldots, y_N$ generados de una distribución normal con media $\mu = 850$ y varianza $\sigma^2 = 75{,}625$, $N = 2640$, según orden donde $y_{(1)} \leq \cdots \leq y_{(N)}$.

b) Orden creciente. El conjunto de variables aleatorias $Y_1, Y_2, \ldots, Y_N$ ordenadas en forma creciente queda como

$$Y_{(1)}, Y_{(2)}, \ldots, Y_{(N)}.$$

En la Figura 5.1 (b) se muestran los valores $y_1, y_2, \ldots, y_N$ generados de una distribución normal, ordenados en forma creciente.

Bajo el orden creciente y muestreo sistemático, las variables en la muestra son

$$Y_{(r)}, Y_{(r+T)}, \ldots, Y_{(r+(n-1)T)},$$

es decir, las estadísticas de orden en la muestra están identificadas por

$$s_r = \{r, r+T, r+2T, \ldots, r+(n-1)T\}.$$

Retomando el ejemplo donde $N = 20$, $n = 4$ y $r = 2$ entonces las estadísticas de orden en la muestra son $Y_{(2)}, Y_{(7)}, Y_{(12)}, Y_{(17)}$, o equivalentemente por las $Y_{(i)}, i \in s_r$ donde $s_r = \{2, 7, 12, 17\}$.

En adelante al muestreo sistemático bajo el orden creciente se le denotará de manera abreviada por SYC. No hay que confundir esta notación con la que algunos autores dan al muestreo sistemático circular o bien al sistemático centrado.

5.3. Varianza total bajo el modelo y el diseño

Como hay dos etapas de aleatoriedad, la debida al modelo y la debida al diseño, el operador esperanza se puede aplicar de manera conjunta bajo ambas etapas o bien bajo sólo una de ellas. En lo sucesivo $E_p(\cdot)$ denotará el valor esperado y $V_p(\cdot)$ la varianza bajo el diseño p, mientras que $E_m(\cdot)$ y $V_m(\cdot)$ referirán al valor esperado y la varianza bajo el modelo m.

El valor esperado bajo el diseño y el modelo conjuntamente, denominado valor esperado total, puede expresarse como

$$E_{p,m}(\cdot) = E_p\left[E_m(\cdot|s)\right].$$

De manera análoga, la varianza total o conjunta bajo el modelo y el diseño, denominada por Isaki y Fuller [28] varianza anticipada, está dada por

$$\begin{aligned} V_{p,m}(\cdot) &= E_{p,m}\left[(\cdot) - E_{p,m}(\cdot)\right]^2 \\ &= E_p\left[V_m(\cdot|s)\right] + V_p\left[E_m(\cdot|s)\right]. \end{aligned} \tag{5.2}$$

Si los momentos bajo el modelo para la superpoblación son finitos, como en el caso de las distribuciones uniforme, Laplace y normal, y el muestreo es no informativo,

es decir, las probabilidades de selección son independientes de la variable de interés, entonces se puede intercambiar el orden del operador esperanza, por lo que el valor esperado y la varianza total también pueden expresarse como

$$E_{p,m}(\cdot) = E_{m,p}(\cdot) = E_m\left[E_p(\cdot)\right],$$

$$V_{p,m}\left(\cdot\right) = V_{m,p}\left(\cdot\right) = E_m\left[V_p\left(\cdot\right)\right] + V_m\left[E_p\left(\cdot\right)\right]. \tag{5.3}$$

Se encontró que en el caso del muestreo sistemático la varianza conjunta del predictor $\hat{\theta}_\pi$ se reduce a una expresión que, además de ciertas cantidades fijas, contiene el valor esperado del producto de estadísticas de orden, como se establece en la Proposición 5.3.1; mientras que cuando se efectua un muestreo aleatorio simple, la varianza conjunta depende únicamente de una cantidad fija como se demuestra en la Proposición 5.3.2.

Proposición 5.3.1. *Sean $Y_1, Y_2, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2. Sean $Y_{(1)}, \ldots, Y_{(N)}$ las estadísticas de orden de las N variables aleatorias, n el tamaño de muestra y $T = N/n$. Considérese además que la población está ordenada por algún criterio. Entonces, la varianza total del predictor de $\theta = \sum_{i=1}^{N} Y_i$ bajo muestreo sistemático corresponde a*

$$V_{SY,m}\left(\hat{\theta}_\pi\right) = T\left[N\sigma^2 + (N - Nn)\mu^2 + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}Y_{(j)}\right]\right]. \tag{5.4}$$

Demostración. Usando la expresión (5.3) cuando el muestreo es sistemático, $p = \text{SY}$, la varianza total del predictor es

$$V_{SY,m}\left(\hat{\theta}_\pi\right) = E_m\left[V_{SY}\left(\hat{\theta}_\pi\right)\right] + V_m\left[E_{SY}\left(\hat{\theta}_\pi\right)\right]. \tag{5.5}$$

A continuación se obtienen cada uno de los componentes de la varianza total. El valor esperado bajo el diseño del predictor $\hat{\theta}_\pi$ está dado por

$$\begin{aligned} E_{SY}\left(\hat{\theta}_\pi\right) &= \frac{\sum_{r=1}^{T}\hat{\theta}_{s_r}}{T} = \sum_{r=1}^{T}\sum_{i\in s_r} Y_{(i)} \\ &= \sum_{i=1}^{N} Y_{(i)} = \sum_{i=1}^{N} Y_i = \theta. \end{aligned}$$

Tomando la varianza con respecto al modelo

$$\begin{aligned} V_m\left[E_{SY}\left(\hat{\theta}_\pi\right)\right] &= V_m\left(\sum_{i=1}^{N} Y_i\right) \\ &= \sum_{i=1}^{N} V\left(Y_i\right) = N\sigma^2. \end{aligned} \tag{5.6}$$

Por otra parte, calculando el primer sumando de la varianza total,

$$\begin{aligned} V_{SY}\left(\hat{\theta}_\pi\right) &= \frac{\sum_{r=1}^{T}\left(\hat{\theta}_{s_r}-\theta\right)^2}{T} = \frac{\sum_{r=1}^{T}\hat{\theta}_{s_r}^2}{T} - \theta^2 \\ &= T\sum_{r=1}^{T}\left(\sum_{i\in s_r} Y_{(i)}\right)^2 - \sum_{i=1}^{N} Y_i^2 - 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Y_iY_j \\ &= T\sum_{r=1}^{T}\left[\sum_{i\in s_r} Y_{(i)}^2 + \sum\sum_{\substack{i,j\in s_r \\ i\neq j}} Y_{(i)}Y_{(j)}\right] - \sum_{i=1}^{N} Y_i^2 - 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Y_iY_j. \end{aligned}$$

Encontrando el valor esperado con respecto al modelo

$$\begin{aligned} E_m\left[V_{SY}\left(\hat{\theta}_\pi\right)\right] =& T\sum_{r=1}^{T}\left[\sum_{i\in s_r} E_m\left[Y_{(i)}^2\right] + \sum\sum_{\substack{i,j\in s_r \\ i\neq j}} E_m\left[Y_{(i)}Y_{(j)}\right]\right] \\ & - \sum_{i=1}^{N} E_m\left(Y_i^2\right) - 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E_m\left(Y_iY_j\right) \\ =& T\left[\sum_{i=1}^{N} E_m\left[Y_{(i)}^2\right] + \sum_{r=1}^{T}\sum\sum_{\substack{i,j\in s_r \\ i\neq j}} E_m\left[Y_{(i)}Y_{(j)}\right]\right] \\ & - \sum_{i=1}^{N} E_m\left(Y_i^2\right) - 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} E_m\left(Y_iY_j\right) \\ =& T\left[N\sigma^2 + N\mu^2 + \sum_{r=1}^{T}\sum\sum_{\substack{i,j\in s_r \\ i\neq j}} E_m\left[Y_{(i)}Y_{(j)}\right]\right] \\ & - N\sigma^2 - N\mu^2 - 2\mu^2\left[\frac{N(N-1)}{2}\right] \\ =& T\left[N\sigma^2 + (N-Nn)\mu^2 + \sum_{r=1}^{T}\sum\sum_{\substack{i,j\in s_r \\ i\neq j}} E_m\left[Y_{(i)}Y_{(j)}\right]\right] - N\sigma^2. \end{aligned} \tag{5.7}$$

Sustituyendo (5.6) y (5.7) en la varianza total (5.5) finalmente se llega al resultado propuesto. □

En términos de las covarianzas la expresión (5.4) es equivalente a

$$\begin{aligned} V_{SY,m}\left(\hat{\theta}_\pi\right) =& T\left[N\sigma^2 + (N - Nn)\mu^2 + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum C_m\left[Y_{(i)}, Y_{(j)}\right]\right] \\ &+ T\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}\right] E_m\left[Y_{(j)}\right]. \end{aligned} \tag{5.8}$$

Es importante señalar que la expresión (5.4) es válida para un orden dado, cualquiera que este sea, como puede verse en la demostración. Ahora bien, si $Y_1, \ldots, Y_N$ son variables aleatorias simétricas con respecto a su media, μ, n es un número par y el orden es equilibrado, entonces la suma de productos de valores esperados dada en (5.8) puede simplificarse como sigue

$$\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}\right] E_m\left[Y_{(j)}\right] = Tn^2\mu - \sum_{i=1}^{N} E_m^2\left[Y_{(i)}\right]. \tag{5.9}$$

Para verificar la igualdad anterior, considérese que en el caso de una variable aleatoria simétrica, con respecto a su media, se cumple que

$$E_m\left[Y_{(i)}\right] - \mu = -E_m\left[Y_{(N+1-i)}\right] + \mu,$$

luego

$$E_m\left[Y_{(i)}\right] + E_m\left[Y_{(N+1-i)}\right] = 2\mu.$$

Bajo BSS si $i \in s_r$, también $N + 1 - i \in s_r$, entonces

$$n\mu = \sum_{i\in s_r} E_m\left[Y_{(i)}\right],$$

$$n^2\mu^2 = \left\{\sum_{i\in s_r} E_m\left[Y_{(i)}\right]\right\}^2 = \sum_{i\in s_r} E_m^2\left[Y_{(i)}\right] + \sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}\right] E_m\left[Y_{(j)}\right],$$

sumando sobre las muestras posibles se tiene

$$Tn^2\mu^2 = \sum_{r=1}^{T}\sum_{i\in s_r} E_m^2\left[Y_{(i)}\right] + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}\right]E_m\left[Y_{(j)}\right],$$

despejando en la igualdad anterior se llega a la expresión (5.9).

Proposición 5.3.2. *Sean $Y_1, Y_2, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2, $i = 1, \ldots, N$. Sean n el tamaño de muestra y $T = N/n \in \mathbb{N}$ el intervalo muestral. Entonces, bajo muestreo aleatorio simple (SI), la varianza total del predictor $\hat{\theta}_\pi$ es*

$$V_{SI,m}\left(\hat{\theta}_\pi\right) = TN\sigma^2. \tag{5.10}$$

Demostración. La varianza total del predictor $\hat{\theta}_\pi$ está dada por

$$V_{SI,m}\left(\hat{\theta}_\pi\right) = E_{SI}\left[V_m\left(\hat{\theta}_\pi|s\right)\right] + V_{SI}\left[E_m\left(\hat{\theta}_\pi|s\right)\right].$$

Puesto que se selecciona una muestra con muestreo aleatorio simple, entonces $\hat{\theta}_\pi$ puede escribirse como

$$\hat{\theta}_\pi = T\sum_{i\in s_r} Y_i.$$

Luego, el valor esperado bajo el modelo es

$$\begin{aligned} E_m\left(\hat{\theta}_\pi|s\right) &= E_m\left(T\sum_{i\in s} Y_i|s\right) \\ &= T\sum_{i\in s} E_m(Y_i) = Tn\mu = N\mu. \end{aligned}$$

Calculando la varianza bajo el diseño de la expresión anterior

$$V_{SI}\left[E_m\left(\hat{\theta}_\pi|s\right)\right] = V_{SI}\left[N\mu\right] = 0. \tag{5.11}$$

Por otro lado, la varianza bajo el modelo es

$$\begin{aligned} V_m\left(\hat{\theta}_\pi|s\right) &= V_m\left(T\sum_{i\in s} Y_i|s\right) \\ &= T^2\sum_{i\in s} V_m\left(Y_i\right) \\ &= T^2 n\sigma^2 = TN\sigma^2. \end{aligned}$$

Tomando el promedio sobre las muestras posibles

$$\begin{aligned} E_{SI}\left[V_m\left(\hat{\theta}_\pi|s\right)\right] &= E_{SI}\left[TN\sigma^2\right] \\ &= TN\sigma^2. \end{aligned} \tag{5.12}$$

Finalmente, sumando (5.11) y (5.12) se obtiene el resultado dado en (5.10). □

A partir de las ecuaciones (5.4) y (5.10), es fácil ver que

$$V_{SY,m}\left(\hat{\theta}_\pi\right) - V_{SI,m}\left(\hat{\theta}_\pi\right) = T(N-Nn)\mu^2 + T\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}Y_{(j)}\right].$$

Si se cumple la desigualdad

$$(N-Nn)\mu^2 < -\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_m\left[Y_{(i)}Y_{(j)}\right],$$

entonces el muestreo sistemático es más eficiente que el aleatorio simple. A continuación se comparan las varianzas del BSS, SYC y SI suponiendo una distribución uniforme para la superpoblación.

5.4. Modelo de superpoblación uniforme

En este apartado se deriva la varianza del predictor de la suma de N variables aleatorias bajo el modelo y el diseño. Se supone la distribución uniforme y que la población se ordena de forma equilibrada, o de manera creciente. En las siguientes secciones se consideran las distribuciones de Laplace, normal y normal generalizada.

Para distinguir entre los cuatro modelos, en lugar del subíndice m se usará ξ para el modelo que supone la distribución uniforme, δ para la distribución de Laplace, λ para la distribución normal y Λ para la distribución normal generalizada. Para diferenciar los diseños, en lugar del subíndice p, BSS denotará el muestreo sistemático cuando la población se ordena de manera equilibrada, SYC al muestreo sistemático cuando la población se ordena de manera creciente y SI al muestreo aleatorio simple.

Se dice que la variable aleatoria continua Y tiene una distribución uniforme en el intervalo $[u_1, u_2]$, también denotada por $Y \sim U(u_1, u_2)$, si su función de densidad

está dada por:

$$f_Y(y) = \frac{1}{u_1 - u_2}, \quad y \in [u_1, u_2].$$

El valor esperado y la varianza de Y son

$$E(Y) = \frac{u_1 + u_2}{2}, \quad V(Y) = \frac{(u_2 - u_1)^2}{12}.$$

Es fácil ver que $Y \sim U(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$ tiene función de densidad

$$f_Y(y) = \frac{1}{2\sqrt{3}\sigma}, \quad y \in [\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]$$

con $E(Y) = \mu$ y $V(Y) = \sigma^2$.

A la variable $Y \sim U(0, 1)$ se le llama distribución uniforme estándar, en este caso $E(Y) = 1/2$ y $V(Y) = 1/12$.

Una característica de las estadísticas de orden de la distribución uniforme es que sus valores esperados, varianzas y covarianzas poseen expresiones cerradas.

Proposición 5.4.1. *Sea ξ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes con distribución uniforme en el intervalo $[\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera equilibrada (BSS), entonces la varianza total del predictor $\hat{\theta}_\pi = T \sum_{i \in s_r} Y_{(i)}$ es igual a*

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]\sigma^2. \tag{5.13}$$

Demostración. La demostración se presenta de manera resumida, para mayor detalle puede consultarse el Anexo. Sin pérdida de generalidad suponga que $Y_1, Y_2, \ldots, Y_N \sim U(0, 1)$ son variables independientes y $X_i = Y_{(i)}$ denota la $i-$ésima estadística de orden de las N variables, entonces el valor esperado, la varianza y la covarianza bajo el modelo son respectivamente (ver por ejemplo Hastings [22]):

$$E_\xi[X_i] = \frac{i}{N+1}, \quad V_\xi[X_i] = \frac{i(N+1-i)}{(N+1)^2(N+2)} \text{ y}$$

$$C_\xi\left[X_i, X_j\right] = \frac{i(N+1-j)}{(N+1)^2(N+2)} \text{ si } i < j.$$

La varianza total del predictor $\hat{\theta}_\pi$ es

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = E_{BSS}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] + V_{BSS}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right]. \tag{5.14}$$

La varianza condicional del predictor bajo el modelo, si se considera que la muestra está fija, está dada por

$$\begin{aligned}
V_\xi\left(\hat{\theta}_\pi|s\right) &= V_\xi\left(T\sum_{i\in s_r} Y_{(i)}\,|s_r\right) \\
&= T^2\left\{\sum_{i\in s_r} V_\xi\left[Y_{(i)}\,|s_r\right] + \sum_{i\neq j\in s_r} C_\xi\left[Y_{(i)}, Y_{(j)}\,|s_r\right]\right\} \\
&= T^2\left\{\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) + \sum_{i=n/2+1}^{n} V_\xi\left(X_{-2r+2+2(n-i+1)T}\right)\right. \\
&\quad + 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&\quad + 2\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n} C_\xi\left(X_{2r-1+2(i-1)T}, X_{-2r+2+2(n-j+1)T}\right) \\
&\quad \left. +2\sum_{i=n/2+1}^{n-1}\sum_{j=i+1}^{n} C_\xi\left(X_{-2r+2+2(n-i+1)T}, X_{-2r+2+2(n-j+1)T}\right)\right\}.
\end{aligned}$$

Cambiando los límites de las sumas para que el subíndice i corra desde uno,

$$\begin{aligned}
V_\xi\left(\hat{\theta}_\pi|s\right) &= T^2\left\{\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) + \sum_{i=1}^{n/2} V_\xi\left(X_{N-2r+2+2(1-i)T}\right)\right. \\
&\quad + 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&\quad + 2\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{N-2r+2+2(1-j)T}\right) \\
&\quad \left. +2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{N-2r+2+2(1-i)T}, X_{N-2r+2+2(1-j)T}\right)\right\}.
\end{aligned} \tag{5.15}$$

Haciendo $a = 2r-1-2T$, $b = N-2r+2+2T = N+1-a$ y $c = [(N+1)^2(N+2)]^{-1}$ y sustituyendo las expresiones para $E_\xi[X_i]$, $V_\xi[X_i]$ y $C_\xi[X_i, X_j]$ se tiene que

$$\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) = c\left[ab\frac{n}{2} + (2Tb-2Ta)\sum_{i=1}^{n/2} i - 4T^2\sum_{i=1}^{n/2} i^2\right] \tag{5.16}$$

$$\sum_{i=1}^{n/2} V_\xi\left(X_{N-2r+2+2(1-i)T}\right) = c\left[ab\frac{n}{2} + (2Tb-2Ta)\sum_{i=1}^{n/2} i - 4T^2\sum_{i=1}^{n/2} i^2\right] \tag{5.17}$$

$$\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) = \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{a+2jT}\right)$$

luego,

$$\begin{aligned}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{a+2jT}\right) = c\Bigg\{&ab\frac{n^2}{4} - ab\frac{n}{2} - Ta\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}-1\right)\\ &+\left[-ab+Ta+Tbn-T^2n\left(\frac{n}{2}+1\right)\right]\sum_{i=1}^{n/2-1} i\\ &+\left[Ta-2Tb+2T^2\right]\sum_{i=1}^{n/2-1} i^2 + 2T^2\sum_{i=1}^{n/2-1} i^3\Bigg\}\end{aligned} \tag{5.18}$$

$$\begin{aligned}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{N-2r+2+2(1-i)T}, X_{N-2r+2+2(1-j)T}\right) &= \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{b-2iT}, X_{b-2jT}\right)\\ &= \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{a+2jT}\right).\end{aligned}$$

Debido a que $C_\xi(X_i, X_j)$ es válida si $i < j$, para calcular la suma de covarianzas

$$\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{N-2r+2+2(1-j)T}\right) = \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right), \tag{5.19}$$

se consideran cuatro casos y se desarrollan cada uno de ellos: i) T par con $r \leq T/2$, ii) T par y $r \geq T/2+1$, iii) T impar con $r \leq (T+1)/2$ y iv) T impar con $r \geq (T+1)/2+1$.

i) Sea T par y $r \leq T/2$

$$\begin{aligned}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) =& a^2\frac{n^2}{4} + a^2\frac{n}{2} + Ta\frac{n^3}{8} + Ta\frac{3n^2}{4} + Tan \\ &+ \left[-a^2 - Ta + T^2\frac{n^2}{2} + 3T^2n + 4T^2\right]\sum_{i=1}^{n/2} i \\ &+ \left[-Ta - 6T^2 - 2T^2n\right]\sum_{i=1}^{n/2} i^2 + 2T^2\sum_{i=1}^{n/2} i^3 \\ &+ \frac{2}{4}Tbn^2 + Tbn - b^2\frac{n}{2} \\ &+ \left[b^2 - Tb - 2T^2n - Tbn - 4T^2\right]\sum_{i=1}^{n/2} i \\ &+ \left[-Tb + 2nT^2 + 6T^2\right]\sum_{i=1}^{n/2} i^2 - 2T^2\sum_{i=1}^{n/2} i^3.\end{aligned} \tag{5.20}$$

Sustituyendo los resultados obtenidos en (5.16), (5.17), (5.18) y (5.20) en la expresión (5.15) y simplificando se obtiene

$$V_\xi\left(\hat{\theta}_\pi | s_r\right) = \frac{2N^2}{(N+1)(N+2)n}\left[\frac{Nn}{24} - \frac{N}{6n} + \frac{n}{8} - \frac{3}{4} + r\right].$$

ii) Sea T par y $r \geq T/2 + 1$

$$\begin{aligned}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) =& a^2\frac{n^2}{4} + Ta\frac{n^3}{8} + Ta\frac{n^2}{4} \\ &+ \left[-a^2 - Ta + T^2\frac{n^2}{2} + T^2n\right]\sum_{i=1}^{n/2} i \\ &+ \left[-Ta - 2T^2 - 2T^2n\right]\sum_{i=1}^{n/2} i^2 + 2T^2\sum_{i=1}^{n/2} i^3 \\ &+ \left(b^2 - Tb - Tbn\right)\sum_{i=1}^{n/2} i - 2T^2\sum_{i=1}^{n/2} i^3 \\ &+ \left[-Tb + 2nT^2 + 2T^2\right]\sum_{i=1}^{n/2} i^2.\end{aligned} \tag{5.21}$$

Sustituyendo las expresiones halladas en (5.16), (5.17), (5.18) y (5.21) en la ecuación (5.15), y simplificando se obtiene

$$V_\xi\left(\hat{\theta}_\pi|s_r\right) = \frac{2N^2}{(N+1)(N+2)n}\left[\frac{Nn}{24} + \frac{5N}{6n} + \frac{n}{8} + \frac{3}{4} - r\right].$$

Luego, el valor esperado bajo el diseño es

$$\begin{aligned} E_{BSS}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] &= \frac{1}{T}\sum_{r=1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right) \\ &= \frac{1}{T}\left\{\sum_{r=1}^{T/2} V_\xi\left(\hat{\theta}_\pi|s_r\right) + \sum_{r=T/2+1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right)\right\} \\ &= \frac{N^2}{12(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]. \end{aligned} \tag{5.22}$$

Obteniendo el segundo término de la varianza total

$$E_\xi\left(\hat{\theta}_\pi|s\right) = E_\xi\left[T\sum_{i\in s_r} Y_{(i)}\right] = T\sum_{i\in s_r} E_\xi\left[Y_{(i)}\right] = T\frac{n}{2} = \frac{N}{2},$$

como se esperaba.

Como $E_\xi\left(\hat{\theta}_\pi|s\right)$ es una constante, entonces el segundo término de la varianza total (5.14) es cero, esto es

$$V_{BSS}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right] = 0. \tag{5.23}$$

Y reemplazando (5.22) y (5.23) en (5.14) se obtiene la expresión de la varianza total

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{12(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]. \tag{5.24}$$

Cuando T es impar, las expresiones para la suma de covarianzas en (5.19) son las mismas que las halladas cuando T es par, como se muestra en el anexo. Por lo tanto, la varianza total para cualquier $T \in \mathbb{N}$ está dada por la ecuación (5.24).

Cuando $Y_i^* \sim U(0,1)$, entonces $E(Y_i^*) = 1/2$ y $V(Y_i^*) = 1/12$. Si se supone $Y_i \sim U(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$, $E(Y_i) = \mu$ y $V(Y_i) = \sigma^2$. Es fácil ver que en términos de Y_i^*, $Y_i = \mu - \sqrt{3}\sigma + 2\sqrt{3}\sigma Y_i^*$ con $V(Y_i) = 12\sigma^2 V(Y_i^*)$.

Finalmente, la varianza total considerando una media μ y varianza σ^2 es

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N+3+2\frac{N}{n^2}\right]\sigma^2. \qquad \square$$

Si además se considera un tamaño de población lo suficientemente grande para suponer que $N \approx N+3$, entonces la expresión anterior se reduce a

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) \approx N\left[1+\frac{2}{n^2}\right]\sigma^2. \tag{5.25}$$

Proposición 5.4.2. *Sea ξ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes con distribución uniforme en el intervalo $[\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera creciente (SYC), entonces la varianza total del predictor $\hat{\theta}_\pi = T\sum_{i\in s_r} Y_{(i)}$ es igual a*

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)}\left[1+\frac{N}{n^2}\right]\sigma^2. \tag{5.26}$$

Demostración. La varianza total del predictor $\hat{\theta}_\pi$ es

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = E_{SYC}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] + V_{SYC}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right].$$

A continuación se derivan expresiones para cada una de esas varianzas y valores esperados cuando $Y_i \sim U(0,1)$.

$$\begin{aligned} V_\xi\left(\hat{\theta}_\pi|s\right) &= V_\xi\left(T\sum_{i\in s_r} Y_{(i)}\,|s_r\right) \\ &= T^2\left[\sum_{i\in s_r} V_\xi\left[Y_{(i)}\,|s_r\right] + \sum_{i\neq j\in s_r} C_\xi\left[Y_{(i)}, Y_{(j)}\,|s_r\right]\right] \\ &= T^2\left[\sum_{i=1}^{n} V_\xi\left(X_{r+(i-1)T}\right) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_\xi\left(X_{r+(i-1)T}, X_{r+(j-1)T}\right)\right] \end{aligned} \tag{5.27}$$

donde $X_i = Y_{(i)}$. Haciendo $a = r - T$, $b = N + 1 - r + T = N + 1 - a$ y $c = [(N+1)^2(N+2)]^{-1}$, el primer sumando de la ecuación (5.27) se puede reescribir

como

$$\begin{aligned}\sum_{i=1}^{n} V_\xi\left(X_{r+(i-1)T}\right) &= \sum_{i=1}^{n} V_\xi\left(X_{a+iT}\right)\\ &= c\sum_{i=1}^{n}(a+iT)(b-iT)\\ &= c\left(abn - aT\sum_{i=1}^{n} i + bT\sum_{i=1}^{n} i - T^2\sum_{i=1}^{n} i^2\right).\end{aligned} \tag{5.28}$$

Sustituyendo a, b y c en el segundo sumando de (5.27) se tiene

$$\begin{aligned}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_\xi\left(X_{r+(i-1)T}, X_{r+(j-1)T}\right) =& \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_\xi\left(X_{a+iT}, X_{a+jT}\right)\\ =& c\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(a+iT)(b-jT)\\ =& c\left\{abn(n-1) - \frac{aT}{2}n(n+1)(n-1)\right.\\ &+\left[-ab+\frac{aT}{2}+bTn-\frac{T^2}{2}n(n+1)\right]\sum_{i=1}^{n-1} i\\ &\left.+\left[\frac{aT}{2}-bT+\frac{T^2}{2}\right]\sum_{i=1}^{n-1} i^2 + \frac{T^2}{2}\sum_{i=1}^{n-1} i^3\right\}.\end{aligned} \tag{5.29}$$

Sustituyendo las expresiones (5.28) y (5.29) en (5.27) y simplificando se obtiene

$$\begin{aligned}V_\xi\left(\hat{\theta}_\pi|s_r\right) =& T^2 c\left\{-\frac{T^2n^2(n+1)^2}{4} - \frac{2Ta}{3}n(n+1)\left(n-\frac{1}{4}\right)\right.\\ &\left.+\frac{Tb}{3}n\left(n+\frac{1}{2}\right)(n+1) + abn^2\right\}\\ =& T^2 c\left\{\frac{N(n+1)}{6}\left[-\frac{6N(n+1)}{4} - 4a\left(n-\frac{1}{4}\right) + 2b\left(n+\frac{1}{2}\right)\right] + abn^2\right\}.\end{aligned} \tag{5.30}$$

Obteniendo el valor esperado bajo el diseño

$$E_{SYC}(a) = E_{SYC}(r-T) = \frac{T+1}{2} - T = \frac{1-T}{2}$$

$$
\begin{aligned}
E_{SYC}(b) &= E_{SYC}(N+1-a) = N+1-\frac{1-T}{2} = N+1+\frac{T-1}{2} \\
E_{SYC}(ab) &= E_{SYC}\left(-NT-T-T^2+(N+1+2T)r-r^2\right) \\
&= -NT-T-T^2+(N+1+2T)\frac{T+1}{2}-\frac{2T^2+3T+1}{6} \\
&= \frac{3N(1-T)+2(1-T^2)}{6}.
\end{aligned}
$$

Entonces, el valor esperado bajo el modelo de la varianza que se obtuvo en (5.30) es

$$
\begin{aligned}
E_{SYC}\left[V_\xi\left(\hat{\theta}_\pi|s_r\right)\right] &= T^2c\left\{\frac{N(n+1)}{6}\left[-\frac{6N(n+1)}{4}-4E_{SYC}\left(a\right)\left(n-\frac{1}{4}\right)\right.\right. \\
&\qquad \left.\left.+2E_{SYC}\left(b\right)\left(n+\frac{1}{2}\right)\right]+E_{SYC}\left(ab\right)n^2\right\} \\
&= T^2c\left\{\frac{N(n+1)}{6}\left[-\frac{6N(n+1)}{4}-4\left(\frac{1-T}{2}\right)\left(n-\frac{1}{4}\right)\right.\right. \\
&\qquad \left.\left.+2\left(N+1+\frac{T-1}{2}\right)\left(n+\frac{1}{2}\right)\right]+\frac{3N(1-T)+2(1-T^2)}{6}n^2\right\}.
\end{aligned}
$$

Efectuando operaciones se llega a

$$
E_{SYC}\left[V_\xi\left(\hat{\theta}_\pi|s_r\right)\right] = \frac{N^2}{12(N+1)^2}\left[N+2+\frac{N}{n^2}\right]. \tag{5.31}
$$

Por otro lado, el valor esperado del predictor, bajo el modelo, es

$$
\begin{aligned}
E_\xi\left(\hat{\theta}_\pi|s\right) &= E_\xi\left(T\sum_{i\in s_r}Y_{(i)}\,|s_r\right) \\
&= T\sum_{i=1}^{n}E_\xi\left(X_{r+(i-1)T}\right) \\
&= T\sum_{i=1}^{n}\frac{r+(i-1)T}{N+1} \\
&= \frac{N}{2n(N+1)}\left[2rn+Nn-N\right],
\end{aligned}
$$

$$
\begin{aligned}
V_{SYC}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right] &= V_{SYC}\left[\frac{N}{2n(N+1)}\left[2rn+Nn-N\right]\right] \\
&= \frac{N^2}{4n^2(N+1)^2}4n^2V_{SYC}(r) \\
&= \frac{N^2}{(N+1)^2}\left(\frac{T^2-1}{12}\right) \\
&= \frac{N^2}{12(N+1)^2}\left(\frac{N^2}{n^2}-1\right).
\end{aligned} \tag{5.32}
$$

A partir de los resultados (5.31) y (5.32) se obtiene la varianza total

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{12(N+1)}\left[1+\frac{N}{n^2}\right].$$

Finalmente, considerando que $E_\xi(Y_i) = \mu$ y $V_\xi(Y_i) = \sigma^2$ se tiene la expresión dada en (5.26). □

Si el tamaño de la población es lo suficientemente grande para suponer que $N \approx N+2$, la expresión anterior se simplifica quedando de la siguiente manera

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) \approx N\left[1+\frac{N}{n^2}\right]\sigma^2. \tag{5.33}$$

Por otra parte, de acuerdo con lo establecido en la Proposición 5.3.2, bajo el modelo ξ y muestreo aleatorio simple la varianza del predictor es

$$V_{SI,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{n}\sigma^2. \tag{5.34}$$

A partir de las fórmulas (5.25), (5.33) y (5.34), se observa claramente que $V_{BSS,\xi} < V_{SYC,\xi} < V_{SI,\xi}$, es decir, el muestreo sistemático cuando la población se ordena de forma equilibrada, es más eficiente que el muestreo sistemático cuando la población se ordena de manera creciente y éste es más eficiente que el muestreo aleatorio simple.

Si $CV^2_{p,\xi} = V_{p,\xi}\left(\hat{\theta}_\pi\right)/E^2_{p,\xi}\left(\hat{\theta}_\pi\right)$ denota el coeficiente de variación al cuadrado, en términos de la notación de Landau, es fácil ver que $CV^2_{BSS,\xi} = O(1/N)$, $CV^2_{SYC,\xi} = O(1/n^2)$ y $CV^2_{SI,\xi} = O(1/n)$, esto es, $CV^2_{BSS,\xi}$ es de orden $1/N$, $CV^2_{SYC,\xi}$ es de orden $1/n^2$ y $CV^2_{SI,\xi}$ es de orden $1/n$. Por lo que conforme n aumenta, $CV^2_{BSS,\xi}$ y $CV^2_{SYC,\xi}$ serán parecidos.

A partir de las varianzas totales dadas en las expresiones (5.13), (5.26) y (5.34), es sencillo demostrar que si $n = N$, entonces $V_{BSS,\xi} = V_{SYC,\xi} = V_{SI,\xi}$, como se esperaba, y que la desigualdad se cumple para cualesquiera $n < N$ y N pares, tales que $T = N/n \in \mathbb{N}$, como se demuestra en la Proposición 5.4.3 en el Anexo.

A manera de ejemplo, en los Cuadros 5.1 y 5.2 se presenta $V_{p,\xi}\left(\hat{\theta}_\pi\right)$ para poblaciones de $N = 120$ unidades y de $N = 1,200$, ambas con $\mu = 0$ y $\sigma^2 = 1$; no se presentan todos los tamaños de muestra posibles, tales que $T \in \mathbb{N}$, porque no aportan información adicional. En los cuadros puede observarse que:

a) La varianza del muestreo sistemático, orden equilibrado, es mucho menor que la de los otros dos diseños, sobre todo cuando el tamaño de muestra es pequeño. En el caso de una población con $N = 120$ elementos, la eficiencia relativa del BSS con respecto al SI oscila entre 2 y 40; cuando $N = 1200$, la eficiencia varía entre 2 y 400. La eficiencia relativa del SYC con respecto al SI cambia entre 2 y 6 si $N = 120$, en el caso de $N = 1200$, la eficiencia está entre 2 y 17.

b) La varianza del BSS es muy parecida para cualquier tamaño de muestra a partir de 6 u 8 unidades.

Cuadro 5.1: Varianza total del predictor, bajo el diseño y el modelo ξ : $Y_1, \ldots, Y_N$ tienen distribución uniforme con $\mu = 0$ y $\sigma^2 = 1$, $N = 120$

$\frac{n}{N} \times 100$	n	$V_{BSS,\xi}\left(\hat{\theta}_\pi\right)$	$V_{SYC,\xi}\left(\hat{\theta}_\pi\right)$	$V_{SI,\xi}\left(\hat{\theta}_\pi\right)$
1.67	2	178.5	3689.3	7200
3.33	4	134.6	1011.6	3600
5.00	6	126.5	515.7	2400
6.67	8	123.6	342.1	1800
8.33	10	122.3	261.8	1440
16.67	20	120.6	154.7	720
25.00	30	120.2	134.9	480
33.33	40	120.1	127.9	360
50.00	60	120.0	123.0	240

$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]$,
$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)}\left[1 + \frac{N}{n^2}\right]$ y $V_{SI,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{n}$.

En presencia de una tendencia lineal, los T posibles valores estimados de la media siguen un comportamiento monótono. Autores como Bellhouse y Rao [2] y Singh et al. [51] comparan el estimador de la media, bajo alguna variante del muestreo sistemático o bien del estimador, suponiendo que la superpoblación sigue el modelo lineal $Y_i = a + bi + \epsilon_i$ con $E(\epsilon_i) = 0$, $E(\epsilon_i^2) = \sigma_{\epsilon_i}^2$ y $E(\epsilon_i \epsilon_j) = 0 \; \forall i \neq j$, $i, j = 1, \ldots, N$. Bellhouse y Rao obtienen que la varianza del estimador del muestreo sistemático equilibrado es menor que la del sistemático con un orden creciente. Singh et al. obtienen que el muestreo sistemático con un orden creciente es más eficiente que el

muestreo aleatorio simple. El modelo que aquí se plantea, es decir, que la población sigue una distribución uniforme y se ordena de manera creciente, es similar al modelo anterior pero sin el supuesto que los errores son no correlacionados. Esto puede verse si la $i-$ésima estadística de orden se reescribe como:

$$Y_{(i)} = Y_{(i-1)} + \epsilon_i, \quad i = 1, \ldots, N,$$

donde $\epsilon_1, \ldots, \epsilon_N$ son variables aleatorias no negativas. Es fácil demostrar que los errores no son independientes ya que $E_m(\epsilon_i \epsilon_{i-1}) = 12\sigma^2(N+1)^{-1}(N+2)^{-1}$ y $E_m(\epsilon_i) E_m(\epsilon_{i-1}) = 12\sigma^2(N+1)^{-2}$.

Cuadro 5.2: Varianza total del predictor, bajo el diseño y el modelo ξ : $Y_1, \ldots, Y_N$ tienen distribución uniforme con $\mu = 0$ y $\sigma^2 = 1$, $N = 1200$

$\frac{n}{N} \times 100$	n	$V_{BSS,\xi}\left(\hat{\theta}_\pi\right)$	$V_{SYC,\xi}\left(\hat{\theta}_\pi\right)$	$V_{SI,\xi}\left(\hat{\theta}_\pi\right)$
0.17	2	1798.50	360899.3	720000
0.33	4	1349.62	91124.1	360000
0.50	6	1266.50	41165.7	240000
0.67	8	1237.40	23680.3	180000
0.83	10	1223.94	15587.0	144000
1.67	20	1205.98	4796.0	72000
3.33	40	1201.49	2098.3	36000
5.00	60	1200.66	1598.7	24000
6.67	80	1200.37	1423.8	18000
8.33	100	1200.24	1342.9	14400
16.67	200	1200.06	1235.0	7200
25.00	300	1200.02	1215.0	4800
33.33	400	1200.01	1208.0	3600
50.00	600	1200.00	1203.0	2400

$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]$,
$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)}\left[1 + \frac{N}{n^2}\right]$ y $V_{SI,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{n}$.

5.5. Modelo de superpoblación Laplace

En esta sección se obtiene la varianza del predictor del total suponiendo que la población tiene una distribución doble exponencial, la cual fue descubierta por Pierre

Laplace en 1774 y por ello también se le conoce como distribución de Laplace; recibe otros nombres como distribución exponencial de dos colas, exponencial bilateral o primera ley del error de Poisson. Su función de densidad, expresada en función de su media μ y su desviación estándar σ, es

$$f_Y(y) = \frac{1}{\sqrt{2}\sigma} \exp\left(-\frac{\sqrt{2}\,|y-\mu|}{\sigma}\right),\ y \in \mathbb{R}.$$

La forma estándar se obtiene si $\mu = 0$ y $\sigma^2 = 2$, quedando

$$f_Y(y) = \frac{1}{2} \exp\left(-|y|\right).$$

Existen expresiones exactas para los primeros dos momentos y productos de las estadísticas de orden de esta distribución (consultar por ejemplo Govindarajulu [17], p. 250 o bien en Johnson [29], p. 170-171):

$$E\left[Y_{(k)}\right] = \frac{1}{2^N}\left\{\sum_{i=0}^{k-1}\binom{N}{i} S_1(k-i, N-i) - \sum_{i=k}^{N}\binom{N}{i} S_1(i-k+1, i)\right\},\ 1 \le k \le N \tag{5.35}$$

$$E\left[Y_{(k)}^2\right] = \frac{1}{2^N}\left\{\sum_{i=0}^{k-1}\binom{N}{i} S_2(k-i, N-i) + \sum_{i=k}^{N}\binom{N}{i} S_2(i-k+1, i)\right\},\ 1 \le k \le N \tag{5.36}$$

$$\begin{aligned} E\left[Y_{(k)}Y_{(l)}\right] =& \frac{1}{2^N}\left\{\sum_{i=0}^{k-1}\binom{N}{i} S_3(k-i, l-i, N-i)\right. \\ &- \sum_{i=k}^{l-1}\binom{N}{i} S_1(i-k+1, i) S_1(l-i, N-i) \\ &\left.+ \sum_{i=l}^{N}\binom{N}{i} S_3(i-l+1, i-l+1, i)\right\},\quad 1 \le k < l \le N \end{aligned} \tag{5.37}$$

donde

$$S_1(k, N) = \sum_{i=N-k+1}^{N} \frac{1}{i}, \quad S_2(k, N) = \sum_{i=N-k+1}^{N} \frac{1}{i^2} + \left[S_1(k, N)\right]^2 \text{ y}$$

$$S_3(k,l,N) = \sum_{i=N-k+1}^{N} \frac{1}{i^2} + S_1(k,N)S_1(l,N).$$

Usando las expresiones anteriores Govindarajulu [17] tabuló los valores de la media, varianza y covarianzas de las estadísticas de orden de la distribución de Laplace para $N \leq 20$ unidades. Debido a la complejidad de las fórmulas (5.35), (5.36) y (5.37), aunque la expresión para $V_{p,m}\left(\hat{\theta}_\pi\right)$ es cerrada, no nos fue posible encontrar una expresión simplificada o simple.

Para encontrar una expresión más sencilla para la varianza total $V_{p,m}\left(\hat{\theta}_\pi\right)$ se usaron aproximaciones asintóticas de los primeros momentos de las estadísticas de orden. Éstas se obtuvieron siguiendo Hastings, et al. [22], quien señala que si $f_Y(y)$ es la función de densidad de la variable aleatoria Y, una expresión asintótica para la media m_i de la $i-$ésima estadística de orden de una muestra de tamaño N, se obtiene resolviendo la siguiente ecuación para m_i

$$\int_{-\infty}^{m_i} f(y)dy = \frac{i}{N+1}.$$

Análogamente, la varianza asintótica m_{ii} está dada por

$$m_{ii} = \frac{i(N+1-i)}{N(N+1)^2\left[f(m_i)\right]^2},$$

y la covarianza asintótica m_{ij} entre $Y_{(i)}$ y $Y_{(j)}$ es

$$m_{ij} = \frac{i(N+1-j)}{N(N+1)^2 f(m_i)f(m_j)}, \quad \text{con } i \leq j.$$

Obsérvese que como hay dos etapas de aleatoriedad, N corresponde al tamaño de la población o número de elementos que se eligen de la superpoblación, mientras que n denotará el tamaño de muestra o número de elementos que se eligen de la población.

En el caso de la distribución de Laplace estas ecuaciones toman formas cerradas como se establece en la siguiente proposición.

Proposición 5.5.1. *Sean $Y_{(1)}, \ldots, Y_{(N)}$ las estadísticas de orden de $Y_1, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas con función de densidad*

$f_Y(y) = 2^{-1}\exp(-|y|)$ y N par. Entonces una aproximación asintótica para $E\left[Y_{(i)}\right]$, $(i = 1, \ldots, N)$, es

$$m_i = \begin{cases} \ln\left(\frac{2i}{N+1}\right) & \text{si } 1 \leq i \leq N/2 \\ -\ln\left(\frac{2(N+1-i)}{N+1}\right) & \text{si } N/2+1 \leq i \leq N, \end{cases} \tag{5.38}$$

y una expresión asintótica para la covarianza entre $Y_{(i)}$ y $Y_{(j)}$ está dada por

$$m_{ij} = \begin{cases} \frac{N+1-j}{Nj} & \text{si } i \leq j \leq N/2 \\ \frac{i}{N(N+1-i)} & \text{si } N/2+1 \leq i \leq j \\ \frac{1}{N} & \text{si } i \leq N/2 < j. \end{cases} \tag{5.39}$$

Demostración. Para encontrar m_i se debe de resolver

$$\frac{1}{2}\int_{-\infty}^{m_i} \exp(-|y|)\, dy = \frac{i}{N+1}.$$

Si $i \leq N/2$ se tiene que

$$\frac{1}{2}\exp(m_i) = \frac{i}{N+1},$$

despejando

$$m_i = \ln\left(\frac{2i}{N+1}\right).$$

Si $i \geq N/2 + 1$ entonces

$$1 - \frac{1}{2}\exp(-m_i) = \frac{i}{N+1}$$

luego,

$$m_i = -\ln\left(\frac{2(N+1-i)}{N+1}\right).$$

Una aproximación para la covarianza se encuentra evaluando

$$m_{ij} = \frac{i(N+1-j)}{N(N+1)^2 f(m_i) f(m_j)} \text{ si } i \leq j.$$

Si $i \leq j \leq N/2$

$$\begin{aligned} m_{ij} &= \frac{i(N+1-j)}{N(N+1)^2 \frac{1}{2}\exp\left[\ln\left(\frac{2i}{N+1}\right)\right] \frac{1}{2}\exp\left[\ln\left(\frac{2j}{N+1}\right)\right]} \\ &= \frac{N+1-j}{Nj}, \end{aligned}$$

si $N/2+1 \leq i \leq j$

$$\begin{aligned} m_{ij} &= \frac{i(N+1-j)}{N(N+1)^2 \frac{1}{2} \exp\left[\ln\left(\frac{2(N+1-i)}{N+1}\right)\right] \frac{1}{2} \exp\left[\ln\left(\frac{2(N+1-j}{N+1}\right)\right]} \\ &= \frac{i}{N(N+1-i)}, \end{aligned}$$

si $i \leq N/2 < j$

$$\begin{aligned} m_{ij} &= \frac{i(N+1-j)}{N(N+1)^2 \frac{1}{2} \exp\left[\ln\left(\frac{2(N+1-j)}{N+1}\right)\right] \frac{1}{2} \exp\left[\ln\left(\frac{2i}{N+1}\right)\right]} \\ &= \frac{1}{N}. \end{aligned}$$

Por lo tanto, m_i y m_{ij} están dadas por (5.38) y (5.39), respectivamente. □

La aproximación de $E\left[Y_{(i)}\right]$ es mejor para las estadísticas de orden cercanas al centro, es decir, para i cercana a $N/2$, que para las de los extremos, por ello se encontró una fórmula recurrente para m_i en términos de $m_{N/2}$.

Proposición 5.5.2. *Sean $Y_{(1)}, \ldots, Y_{(N)}$ las estadísticas de orden de $Y_1, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas con función de densidad $f_Y(y) = 2^{-1} \exp(-|y|)$ y N par. Entonces, otra aproximación asintótica para $E\left[Y_{(i)}\right]$, $m_i^{'}$ $(i = 1, \ldots, N)$, es*

$$m_i^{'} = m_{N/2} - \sum_{k=1}^{N/2-i} \frac{1}{N/2-k} \text{ si } 1 \leq i \leq N/2,$$

análogamente, para las estadísticas de orden mayores a $N/2$, se tiene que

$$m_{N/2+i}^{'} = -m_{N/2} + \sum_{k=1}^{i-1} \frac{1}{N/2-k}, \text{ si } 1 \leq i \leq N/2,$$

donde $m_{N/2} = \ln[N/(N+1)]$.

Demostración. Sea $i > N/2$, entonces

$$\begin{aligned} m_{i+1} &= -\ln\left(\frac{2(N-i)}{N+1}\right) \\ &= -\ln\left(\frac{2}{N+1}\right) - \ln(N-i). \end{aligned}$$

De la expresión $\sum_{k=1}^{N} N^{-1} = \ln(N)+\gamma+(2N)^{-1}+O(N^{-2})$, donde γ es la constante de Euler-Mascheroni, $\gamma = \lim_{k\to\infty}\left(\sum_{i=1}^{k} i^{-1} - \ln k\right)$ (ver por ejemplo Sedgewick [47], p.28), es fácil ver que $\ln(N+1-i) \approx \ln(N-i) + [2(N+1-i)]^{-1} + [2(N-i)]^{-1}$. Sustituyendo esta aproximación en la última igualdad se tiene que

$$\begin{aligned} m_{i+1} &\approx -\ln\left(\frac{2}{N+1}\right) - \left[\ln(N+1-i) - \frac{1}{2(N-i)} - \frac{1}{2(N+1-i)}\right] \\ &\approx -\ln\left(\frac{2}{N+1}\right) - \left[\ln(N+1-i) - \frac{1}{(N-i)}\right] \\ &= m_i + \frac{1}{(N-i)}. \end{aligned}$$

Empleando esta relación de manera recurrente se tiene

$$\begin{aligned} m_{N/2+2} &\approx m_{N/2+1} + \frac{1}{N/2-1} \\ m_{N/2+3} &\approx m_{N/2+1} + \frac{1}{N/2-1} + \frac{1}{N/2-2} \\ &\vdots \\ m_{N/2+i} &\approx m_{N/2+1} + \frac{1}{N/2-1} + \cdots + \frac{1}{N/2-(i-1)}. \end{aligned}$$

Por lo tanto

$$m'_{N/2+i} = m_{N/2+1} + \sum_{k=1}^{i-1} \frac{1}{N/2-k}, \text{ con } i = 1, \ldots, N/2.$$

Puesto que la función de densidad de Laplace es simétrica, entonces $m_{N/2+1} = -m_{N/2}$. Luego

$$m'_{N/2+i} = -m_{N/2} + \sum_{k=1}^{i-1} \frac{1}{N/2-k}, \text{ con } i = 1, \ldots, N/2. \tag{5.40}$$

Por el mismo argumento de simetría $m'_i = -m'_{N+1-i}$, luego si $1 \leq i \leq N/2$ y usando la ecuación (5.40)

$$\begin{aligned} m'_i &= -m'_{N/2+N/2+1-i} \\ &= m_{N/2} - \sum_{k=1}^{N/2-i} \frac{1}{N/2-k}, \text{ con } i = 1, \ldots, N/2. \end{aligned}$$

□

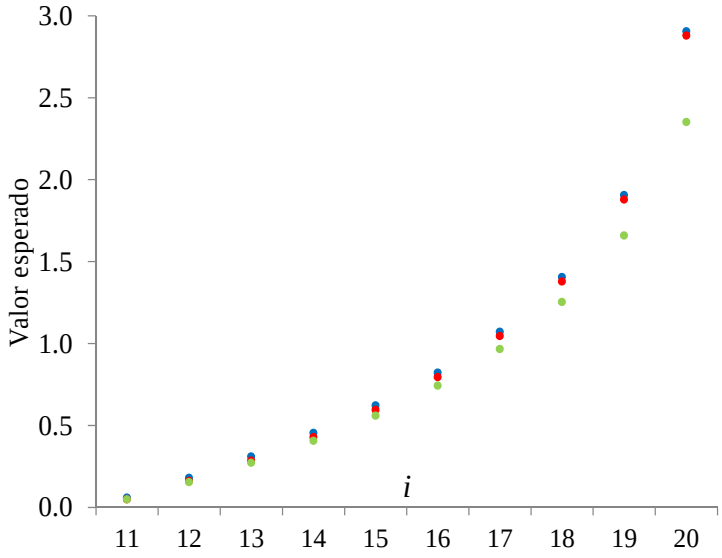


Figura 5.2: Valor esperado • $E\left[Y_{(i)}\right]$ y aproximaciones • $m_i = -\ln\left(\frac{2(N+1-i)}{N+1}\right)$ con $i = 11,\ldots,20$ y • $m'_{N/2+i} = -m_{N/2} + \sum_{k=1}^{i-1} \frac{1}{N/2-k}$ con $i = 1,\ldots,10$ donde $Y_{(1)},\ldots,Y_{(N)}$, $N = 20$, son las estadísticas de orden de $Y_1,\ldots,Y_N$ con distribución de Laplace, $\mu = 0$ y $\sigma^2 = 2$.

En la Figura 5.2 se comparan las dos aproximaciones y el valor esperado, este último tomado de Govindarajulu [17], para una población de $N = 20$ unidades. Como puede observarse, la segunda aproximación m'_i es mejor que m_i, sobre todo cuando i se acerca a N; aunque no se graficó, lo mismo sucede para las estadísticas de orden con $i \leq N/2$, conforme i se acerca a 1, la aproximación m'_i es más cercana al valor esperado exacto.

Usando las dos aproximaciones anteriores, se obtuvo la varianza total, bajo el diseño y el modelo Laplace, como se establece en las siguientes dos proposiciones.

Proposición 5.5.3. *Sea δ el modelo bajo el cual $Y_1,\ldots,Y_N$ son variables aleatorias independientes con función de densidad $f_Y(y) = (\sqrt{2}\sigma)^{-1}\exp\left(-\sqrt{2}\,|y-\mu|\,/\sigma\right)$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera equilibrada (BSS), entonces una*

aproximación de la varianza total del predictor $\hat{\theta}_\pi = T\sum_{i\in s_r} Y_{(i)}$ *está dada por*

$$\begin{aligned} V_{BSS,\delta}\left(\hat{\theta}_\pi\right) \approx & (N+1)\sigma^2\left[\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right) - \frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right) + \frac{1}{N}\left(1-\gamma-\frac{1}{n}\right)\right] \\ & + \frac{N}{n}\sigma^2\left[\ln\left(\frac{N}{2}-1\right) - \frac{Nm_{N/2}^2}{2} + (N-2)m_{N/2} + \frac{1}{N-2} + \gamma + \frac{5}{2}\right] \\ = & V_{BSS,\delta}^{aprox}\left(\hat{\theta}_\pi\right) \end{aligned} \tag{5.41}$$

donde $\gamma = \lim_{k\to\infty}\left(\sum_{i=1}^{k} i^{-1} - \ln k\right) \approx 0.5772156649$ *es la constante de Euler-Mascheroni y* $m_{N/2} = \ln\left[N/(N+1)\right]$.

Demostración. A partir de la ecuación de la varianza total (5.8) se tiene que

$$\begin{aligned} V_{BSS,\delta}\left(\hat{\theta}_\pi\right) = & T\left[N\sigma^2 + (N-Nn)\mu^2 + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum C_\delta\left[Y_{(i)}, Y_{(j)}\right]\right] \\ & + T\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_\delta\left[Y_{(i)}\right]E_\delta\left[Y_{(j)}\right]. \end{aligned} \tag{5.42}$$

En seguida se obtienen la suma de covarianzas y productos de medias usando las aproximaciones asintóticas bajo δ. Por simplicidad en la notación se usará $X_i = Y_{(i)}$ con $i = 1, 2, \ldots, N$ y sin pérdida de generalidad se supone que $E_\delta(Y_i) = \mu = 0$ y $V_\delta(Y_i) = \sigma^2 = 2$.

$$\begin{aligned} \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i<j}}\sum C_\delta\left(X_i, X_j\right) &\approx \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i<j}}\sum m_{ij} \\ &= \sum_{r=1}^{T}\left[\sum_{\substack{i,j\in s_r \\ i<j\leq N/2}}\sum m_{ij} + \sum_{\substack{i,j\in s_r \\ N/2+1\leq i<j}}\sum m_{ij} + \sum_{\substack{i,j\in s_r \\ i\leq N/2<j}}\sum m_{ij}\right] \\ &= 2\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i<j\leq N/2}}\sum m_{ij} + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\leq N/2<j}}\sum m_{ij} \\ &= 2\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i<j\leq N/2}}\sum \frac{N+1-j}{Nj} + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\leq N/2<j}}\sum \frac{1}{N} \\ &= \frac{2(N+1)}{N}\sum_{i=1}^{n/2-1}\sum_{j=1}^{T}\frac{i}{iT+j} - \frac{Tn}{2N}\left(\frac{n}{2}-1\right) + \frac{Tn^2}{4N}. \end{aligned}$$

Usando la aproximación $\sum_{i=1}^{k} i^{-1} \approx \ln k + (2k)^{-1} + \gamma$ donde γ es la constante de Euler-Mascheroni, $= \lim_{k\to\infty} \left(\sum_{i=1}^{k} i^{-1} - \ln k\right) \approx 0.5772156649$, y simplificando se llega a

$$\begin{aligned}\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum C_\delta\left(X_i, X_j\right) \approx & \frac{N+1}{N}\left[n\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right) - 2\ln\left(\left(\frac{n}{2}\right)!\right)\right.\\ & \left. + \frac{n}{N}\left(1-\gamma\right) - \frac{1}{N}\right] + \frac{1}{2}.\end{aligned} \tag{5.43}$$

Como se vio en la expresión (5.9), en el caso del muestreo sistemático, orden equilibrado, y tomando en cuenta que la distribución es simétrica con $\mu = 0$, se tiene que

$$\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i\neq j}}\sum E_\delta\left(X_i\right) E_\delta\left(X_j\right) = -\sum_{i=1}^{N}\left[E_\delta(X_i)\right]^2 = -2\sum_{i=N/2+1}^{N}\left[E_\delta(X_i)\right]^2.$$

Luego,

$$\begin{aligned}-\sum_{i=N/2+1}^{N}\left[E_\delta(X_i)\right]^2 &\approx -\sum_{i=1}^{N/2}\left[m'_{N/2+i}\right]^2\\ &= -\sum_{i=1}^{N/2}\left[-m_{N/2} + \sum_{k=1}^{i-1}\frac{1}{N/2-k}\right]^2\\ &= -\frac{Nm_{N/2}^2}{2} + 2\left(\frac{N}{2}-1\right)m_{N/2} - \sum_{i=1}^{N/2-1}\frac{1}{i} - 2\sum_{i=1}^{N/2-2}\frac{i}{i+1}\\ &= -\frac{Nm_{N/2}^2}{2} + 2\left(\frac{N}{2}-1\right)m_{N/2} - \sum_{i=1}^{N/2-1}\frac{1}{i} - 2\sum_{i=1}^{N/2-2}\left(1-\frac{1}{i+1}\right)\\ &\approx -\frac{Nm_{N/2}^2}{2} + (N-2)m_{N/2} - N + \ln\left(\frac{N}{2}-1\right) + \frac{1}{N-2} + 2 + \gamma.\end{aligned} \tag{5.44}$$

Sustituyendo (5.43) y (5.44) en (5.42) se tiene que

$$\begin{aligned} V_{BSS,\delta}\left(\hat{\theta}_\pi\right) &\approx 2T\left\{N+\left(\frac{N+1}{N}\right)\left[n\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right)-2\ln\left(\left(\frac{n}{2}\right)!\right)\right]\right\} \\ &\quad + 2T\left\{\left(\frac{N+1}{N}\right)\left[+\frac{n}{N}(1-\gamma)-\frac{1}{N}\right]+\frac{1}{2}-+\frac{N}{2}m_{N/2}^2\right\} \\ &\quad + 2T\left\{(N-2)m_{N/2}-N+\ln\left(\frac{N}{2}-1\right)+\frac{1}{N-2}+2+\gamma\right\} \\ &= 2(N+1)\left[\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right)-\frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right)+\frac{1}{N}\left(1-\gamma-\frac{1}{n}\right)\right] \\ &\quad + \frac{N}{n}\left[2\ln\left(\frac{N}{2}-1\right)-Nm_{N/2}^2+2(N-2)m_{N/2}+\frac{2}{N-2}+2\gamma+5\right]. \end{aligned} \tag{5.45}$$

Si en lugar de $E_\delta(Y_i)=0$ y $V_\delta(Y_i)=2$ se tiene que $E_\delta(Y_i)=\mu$ y $V_\delta(Y_i)=\sigma^2$, entonces a partir de (5.45) se llega finalmente al resultado

$$\begin{aligned} V_{BSS,\delta}\left(\hat{\theta}_\pi\right) &\approx (N+1)\sigma^2\left[\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right)-\frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right)+\frac{1}{N}\left(1-\gamma-\frac{1}{n}\right)\right] \\ &\quad + \frac{N}{n}\sigma^2\left[\ln\left(\frac{N}{2}-1\right)-\frac{Nm_{N/2}^2}{2}+(N-2)m_{N/2}+\frac{1}{N-2}+\gamma+\frac{5}{2}\right] \\ &= V_{BSS,\delta}^{aprox}\left(\hat{\theta}_\pi\right). \end{aligned}$$

□

Si el tamaño de la población es tal que $N \approx N+1$, entonces la expresión anterior se reduce a la siguiente

$$V_{BSS,\delta}\left(\hat{\theta}_\pi\right) \approx N\sigma^2\ln\left(\frac{n}{2}\right)+\frac{N}{n}\sigma^2\left[\ln\left(\frac{N}{2}\right)-2\ln\left(\left(\frac{n}{2}\right)!\right)+Nm_{N/2}+\gamma+2.5\right]. \tag{5.46}$$

Proposición 5.5.4. *Sea δ el modelo bajo el cual $Y_1,\ldots,Y_N$ son variables aleatorias independientes con función de densidad $f_Y(y)=(\sqrt{2}\sigma)^{-1}\exp\left(-\sqrt{2}\,|y-\mu|/\sigma\right)$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T=N/n\in\mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera creciente (SYC), entonces una*

aproximación de la varianza total del predictor $\hat{\theta}_\pi = T\sum_{i\in s_r} Y_{(i)}$ *está dada por*

$$\begin{aligned} V_{SYC,\delta}\left(\hat{\theta}_\pi\right) \approx{}& (N+1)\sigma^2\left[\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right) - \frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right) + \frac{1}{N}\left(1-\gamma-\frac{1}{n}\right)\right] \\ &+ \frac{N}{n}\sigma^2\left[N - c_0\ln\left(\frac{N}{2}-1\right) - 2c_0\ln\left(\left(\frac{N}{2}-1\right)!\right) - \frac{Nc_0^2}{2} - \frac{c_0}{N-2}\right. \\ &- \gamma c_0 + \frac{1}{2} - \sum_{i=1}^{N/2-1} c^2(i) + \sum_{r=1}^{T}\left(\sum_{i=1}^{n/2} c(r+iT-T-1)\right)^2 \\ &\left. - \sum_{r=1}^{T}\sum_{i=1}^{n/2} c(r+iT-T-1)\sum_{i=1}^{n/2} c(N/2-r-iT+T)\right] \\ ={}& V_{SYC,\delta}^{aprox}\left(\hat{\theta}_\pi\right) \end{aligned} \tag{5.47}$$

donde $c_0 = m_{N/2} - \ln(N/2-1) - (N-2)^{-1}$, $c(0) = 0$, $c(i) = \ln i + (2i)^{-1}$ si $1 \le i \le N/2-1$ y $\gamma = \lim_{k\to\infty}\left(\sum_{i=1}^{k} i^{-1} - \ln k\right) \approx 0.5772156649$ es la constante de Euler-Mascheroni.

Demostración. A partir de la ecuación(5.8) se tiene que

$$\begin{aligned} V_{SYC,\delta}\left(\hat{\theta}_\pi\right) ={}& T\left[N\sigma^2 + N(1-n)\mu^2 + 2\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum C_\delta\left[Y_{(i)}, Y_{(j)}\right]\right] \\ &+ 2T\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum E_\delta\left[Y_{(i)}\right]E_\delta\left[Y_{(j)}\right]. \end{aligned} \tag{5.48}$$

A continuación se encuentran la suma de covarianzas y productos de medias usando las aproximaciones asintóticas bajo δ. Por simplicidad en la notación se usará $X_i = Y_{(i)}$ con $i = 1, 2, \ldots, N$ y sin pérdida de generalidad se supone que $E_\delta(Y_i) = \mu = 0$ y $V_\delta(Y_i) = \sigma^2 = 2$.

$$\begin{aligned}
\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum C_\delta\left(X_i,X_j\right) \approx & \sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_{r+(i-1)T,r+(j-1)T} \\
= & \sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=i+1}^{n/2} m_{r+(i-1)T,r+(j-1)T} \\
& + \sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n} m_{r+(i-1)T,r+(j-1)T} \\
& + \sum_{r=1}^{T}\sum_{i=n/2+1}^{n-1}\sum_{j=i+1}^{n} m_{r+(i-1)T,r+(j-1)T} \\
= & 2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=i+1}^{n/2} m_{r+(i-1)T,r+(j-1)T} \\
& + \sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n} m_{r+(i-1)T,r+(j-1)T} \\
= & 2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=i+1}^{n/2} \frac{N+1-[r+(j-1)T]}{N[r+(j-1)T]} + \sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n}\frac{1}{N} \\
= & 2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=i+1}^{n/2}\left[\frac{N+1}{N}\frac{1}{r+(j-1)T}-\frac{1}{N}\right] + \sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n}\frac{1}{N} \\
= & \frac{2(N+1)}{N}\sum_{r=1}^{T}\sum_{i=1}^{n/2-1}\frac{i}{iT+r} - \frac{2}{N}\sum_{r=1}^{T}\sum_{i=1}^{n/2}\left(\frac{n}{2}-i\right)+\frac{n}{4}.
\end{aligned}$$

Usando la aproximación $\sum_{i=1}^{k} i^{-1} \approx \ln k + (2k)^{-1} + \gamma$ y simplificando se llega a

$$\begin{aligned}
\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum C_\delta\left(X_i,X_j\right) \approx & \frac{N+1}{N}\left[n\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right) - 2\ln\left(\left(\frac{n}{2}\right)!\right) + \frac{n}{N}\left(1-\gamma\right) - \frac{1}{N}\right] \\
& + \frac{1}{2}.
\end{aligned} \tag{5.49}$$

Por otra parte,

$$\begin{aligned}
\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum E_\delta\left(X_i\right)E_\delta\left(X_j\right) \approx & \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum m_i^{'}m_j^{'} \\
= & \sum_{r=1}^{T}\sum_{i=1}^{n/2-1} m_{r+(i-1)T}^{'}\sum_{j=i+1}^{n/2} m_{r+(j-1)T}^{'} \\
& + \sum_{r=1}^{T}\sum_{i=1}^{n/2} m_{r+(i-1)T}^{'}\sum_{j=n/2+1}^{n} m_{r+(j-1)T}^{'} \\
& + \sum_{r=1}^{T}\sum_{i=n/2+1}^{n-1} m_{r+(i-1)T}^{'}\sum_{j=i+1}^{n} m_{r+(j-1)T}^{'} \\
= & 2\sum_{r=1}^{T}\sum_{i=1}^{n/2-1} m_{r+(i-1)T}^{'}\sum_{j=i+1}^{n/2} m_{r+(j-1)T}^{'} \\
& + \sum_{r=1}^{T}\sum_{i=1}^{n/2} m_{r+(i-1)T}^{'}\sum_{j=n/2+1}^{n} m_{r+(j-1)T}^{'} \\
= & \sum_{r=1}^{T}\left[\sum_{i=1}^{n/2} m_{r+(i-1)T}^{'}\right]^2 - \sum_{i=1}^{N/2}(m_i^{'})^2 \\
& - \sum_{r=1}^{T}\sum_{i=1}^{n/2} m_{r+(i-1)T}^{'}\sum_{i=1}^{n/2} m_{\frac{N}{2}+1-r-(i-1)T}^{'}.
\end{aligned} \tag{5.50}$$

Usando la proposición 5.5.2 se tiene que para $i = 1, \ldots, N/2$

$$m_i^{'} = m_{N/2} - \sum_{k=1}^{N/2-i}\frac{1}{N/2-k} \approx c_0 + c(i-1).$$

donde $c_0 = -m_{N/2} - \ln\left(N/2-1\right) - (N-2)^{-1}$, $c(i) = \ln i + (2i)^{-1}$ si $1 \leq i \leq N/2-1$ y $c(0) = 0$.

Sustituyendo esta aproximación en la última igualdad de (5.50),

$$\begin{aligned}
\sum_{r=1}^{T}\sum_{\substack{i,j\in s_r\\ i<j}}\sum m_i^{'}m_j^{'} \approx & \sum_{r=1}^{T}\left[\sum_{i=1}^{n/2}\left[c_0 + c(r+iT-T-1)\right]\right]^2 \\
& - \sum_{r=1}^{T}\sum_{i=1}^{n/2}\left[c_0 + c(r+iT-T-1)\right]\sum_{i=1}^{n/2}\left[c_0 + c\left(\frac{N}{2} - r - iT + T\right)\right] \\
& - \sum_{i=1}^{N/2}\left[c_0 + c(i-1)\right]^2.
\end{aligned}$$

Efectuando operaciones

$$\sum_{r=1}^{T}\sum_{\substack{i,j\in sr\\ i<j}}\sum m_i^{'}m_j^{'} \approx -\frac{N}{2}c_0^2 - 2c_0\left\{\ln\left(\left(\frac{N}{2}-1\right)!\right)+\frac{1}{2}\left[\ln\left(\frac{N}{2}-1\right)+\gamma+\frac{1}{N-2}\right]\right\}$$
$$-\sum_{i=1}^{N/2-1}c^2(i)+\sum_{r=1}^{T}\left(\sum_{i=1}^{n/2}c(r+iT-T-1)\right)^2$$
$$-\sum_{r=1}^{T}\sum_{i=1}^{n/2}c(r+iT-T-1)\sum_{i=1}^{n/2}c(N/2-r-iT+T). \tag{5.51}$$

Sustituyendo (5.49) y (5.51) en (5.48) se obtiene

$$V_{SYC,\delta}\left(\hat{\theta}_\pi\right) \approx 2(N+1)\left[\left(1-\frac{1}{N}\right)\ln\left(\frac{n}{2}\right)-\frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right)+\frac{1}{N}\left(1-\gamma-\frac{1}{n}\right)\right]$$
$$+\frac{2N}{n}\left[N-c_0\ln\left(\frac{N}{2}-1\right)-2c_0\ln\left(\left(\frac{N}{2}-1\right)!\right)-\frac{Nc_0^2}{2}\right.$$
$$-\frac{c_0}{N-2}-\gamma c_0+\frac{1}{2}-\sum_{i=1}^{N/2-1}c^2(i)+\sum_{r=1}^{T}\left(\sum_{i=1}^{n/2}c(r+iT-T-1)\right)^2$$
$$\left.-\sum_{r=1}^{T}\sum_{i=1}^{n/2}c(r+iT-T-1)\sum_{i=1}^{n/2}c(N/2-r-iT+T)\right].$$

Si en lugar de $E_\delta(Y_i)=0$ y $V_\delta(Y_i)=2$ se tiene que $E_\delta(Y_i)=\mu$ y $V_\delta(Y_i)=\sigma^2$, finalmente se llega al resultado proporcionado en (5.47). □

Si además se considera N grande para poder suponer que $N-1\approx N$, entonces

$$V_{SYC,\delta}\left(\hat{\theta}_\pi\right) \approx N\sigma^2\left[\ln\left(\frac{n}{2}\right)-\frac{2}{n}\ln\left(\left(\frac{n}{2}\right)!\right)\right]+\sigma^2\left(1-\gamma-\frac{1}{n}\right)$$
$$+\frac{N}{n}\sigma^2\left[N-c_0\ln\left(\frac{N}{2}-1\right)-2c_0\ln\left(\left(\frac{N}{2}-1\right)!\right)-\frac{Nc_0^2}{2}\right.$$
$$-\frac{c_0}{N-2}-\gamma c_0+\frac{1}{2}-\sum_{i=1}^{N/2-1}c^2(i)+\sum_{r=1}^{T}\left(\sum_{i=1}^{n/2}c(r+iT-T-1)\right)^2$$
$$\left.-\sum_{r=1}^{T}\sum_{i=1}^{n/2}c(r+iT-T-1)\sum_{i=1}^{n/2}c(N/2-r-iT+T)\right].$$

A diferencia del modelo uniforme en el que se encontró que el coeficiente de variación del predictor del total es de orden $O(1/N)$, $O(1/n^2)$ y $O(1/n)$ bajo BSS, SYC y SI, respectivamente, tanto en el caso del modelo normal como en el de la distribución de Laplace, no fue posible determinar el orden. Con la finalidad de ilustrar la eficiencia de los tres diseños, así como de explorar el grado de aproximación de la varianza numérica, ésta se comparó con la obtenida mediante un estudio de simulación para poblaciones de $N = 120$ y $N = 1200$ elementos.

Como se tiene en la expresión (5.2)

$$V_{p,m}\left(\hat{\theta}_\pi\right) = E_{p,m}\left[\hat{\theta}_\pi - E_{p,m}\left(\hat{\theta}_\pi\right)\right]^2.$$

Si $Y_1, \ldots, Y_N$ tienen media μ, entonces

$$E_{p,m}\left[\hat{\theta}_\pi\right] = E_m\left[E_p\left(\hat{\theta}_\pi\right)\right] = TE_m\left[E_p\left(\sum_{i\in s_r} Y_{(i)}\right)\right].$$

En el caso del muestreo sistemático

$$E_{p,m}\left[\hat{\theta}_\pi\right] = E_m\left(\sum_{i=1}^{N} Y_{(i)}\right) = E_m\left(\sum_{i=1}^{N} Y_i\right) = N\mu.$$

Por lo tanto,

$$V_{p,m}\left(\hat{\theta}_\pi\right) = E_{p,m}\left[\hat{\theta}_\pi - N\mu\right]^2.$$

Para el estudio de simulación se generaron $K = 10,000$ poblaciones de tamaño N, 120 y 1200, cada una bajo el modelo δ con $\mu = 0$ y $\sigma^2 = 1$, la varianza simulada se calculó usando $V_{p,\delta}^{sim}\left(\hat{\theta}_\pi\right)$ como sigue.

$$V_{p,\delta}\left(\hat{\theta}_\pi\right) = E_{p,\delta}\left[\hat{\theta}_\pi\right]^2 \approx \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{r=1}^{T}\left(Tt_{s_r}^k\right)^2 = V_{p,\delta}^{sim}\left(\hat{\theta}_\pi\right)$$

donde p se refiere al diseño: BSS ó SYC, la muestra s_r es la obtenida según el diseño p y $t_{s_r}^k = \sum_{i\in s_r} y_i^k$ corresponde al total de la $r-$ésima muestra en la $k-$ésima población generada.

En los Cuadros 5.3 y 5.4 se presentan los resultados numéricos. Se observa lo siguiente.

a) Existen diferencias entre las varianzas simuladas y las aproximadas por fórmula. En el caso del BSS, la diferencia entre ambas varianzas es menor al 9 % de la varianza simulada. En el caso del SYC, la diferencia entre la varianza simulada y la aproximación es menor al 4 % de la varianza simulada. Esta diferencia se debe al efecto de la aproximación y al de la simulación, ambos efectos están confundidos y se desconoce si éstos son aditivos.

b) La varianza bajo muestreo sistemático es menor cuando la población se ordena de manera equilibrada en comparación con un orden creciente.

c) La varianza del muestreo sistemático bajo cualquiera de los dos órdenes es menor que la del muestreo aleatorio simple. La eficiencia relativa del BSS con respecto al SI varía entre 2 y 21 si $N = 120$ y entre 2 y 150 si $N = 1200$. La eficiencia relativa del SYC con respecto al SI está entre 1.2 y 3 cuando $N = 120$ y entre 1.2 y 6 cuando $N = 1200$.

d) La varianza del muestreo sistemático bajo el orden equilibrado se estabiliza para $n \geq 20$ (en el caso de $N = 120$) y $n \geq 40$ (en el caso de $N = 1200$).

Cuadro 5.3: Varianza total del predictor, bajo el diseño y el modelo $\delta : Y_1, \ldots, Y_N$ tienen distribución de Laplace con $\mu = 0$ y $\sigma^2 = 1$, $N = 120$

n	$V^{sim}_{BSS,\delta}(\hat{\theta}_\pi)$	$V^{aprox}_{BSS,\delta}(\hat{\theta}_\pi)$	$V^{sim}_{SYC,\delta}(\hat{\theta}_\pi)$	$V^{aprox}_{SYC,\delta}(\hat{\theta}_\pi)$	$V_{SI,\delta}(\hat{\theta}_\pi)$	Diferencia[a/] BSS	Diferencia[a/] SYC
2	341.6	371.2	5,910.3	5,954.1	7,200	8.7	0.7
4	212.4	227.0	2,203.5	2,226.0	3,600	6.9	1.0
6	174.2	183.6	1,230.9	1,245.8	2,400	5.4	1.2
8	156.6	163.3	820.9	832.8	1,800	4.3	1.4
10	146.7	151.8	606.0	616.5	1,440	3.5	1.7
20	129.5	131.0	267.7	275.4	720	1.2	2.9
30	124.9	125.0	189.8	196.5	480	0.1	3.5
40	122.9	122.3	159.4	165.5	360	0.5	3.8
60	121.2	120.1	135.7	140.9	240	0.9	3.8

[a/] Corresponde a la diferencia relativa $\frac{|V^{aprox}_{p,\delta} - V^{sim}_{p,\delta}|}{V^{sim}_{p,\delta}} \times 100$, p : BSS, SYC.

Cuadro 5.4: Varianza total del predictor, bajo el diseño y el modelo $\delta : Y_1, \ldots, Y_N$ tienen distribución de Laplace con $\mu = 0$ y $\sigma^2 = 1$, $N = 1200$

n	$V^{sim}_{BSS,\delta}(\hat{\theta}_\pi)$	$V^{aprox}_{BSS,\delta}(\hat{\theta}_\pi)$	$V^{sim}_{SYC,\delta}(\hat{\theta}_\pi)$	$V^{aprox}_{SYC,\delta}(\hat{\theta}_\pi)$	$V_{SI,\delta}(\hat{\theta}_\pi)$	Diferencia[a/] BSS	SYC
2	4,806.2	5,085.4	593,113.5	592,479.0	720,000	5.8	0.1
4	2,822.0	2,958.5	217,413.3	217,159.4	360,000	4.8	0.1
6	2,204.6	2,296.5	118,013.0	117,866.1	240,000	4.2	0.1
8	1,914.3	1,981.0	75,864.6	75,795.2	180,000	3.5	0.1
10	1,745.8	1,798.8	53,691.5	53,646.6	144,000	3.0	0.1
20	1,435.9	1,458.0	18,314.8	18,294.3	72,000	1.5	0.1
40	1,298.8	1,307.3	6,576.6	6,574.6	36,000	0.7	0.0
60	1,259.1	1,262.5	3,883.3	3,889.4	24,000	0.3	0.2
80	1,241.3	1,241.8	2,828.5	2,834.1	18,000	0.0	0.2
100	1,231.1	1,230.1	2,299.5	2,305.7	14,400	0.1	0.3
200	1,213.5	1,209.0	1,514.5	1,519.9	7,200	0.4	0.4
300	1,208.7	1,202.9	1,346.5	1,351.0	4,800	0.5	0.3
400	1,206.7	1,200.1	1,282.8	1,287.0	3,600	0.5	0.3
600	1,205.0	1,200.0	1,234.6	1,237.9	2,400	0.4	0.3

[a/] Corresponde a la diferencia relativa $\frac{|V^{aprox}_{p,\delta} - V^{sim}_{p,\delta}|}{V^{sim}_{p,\delta}} \times 100$, p : BSS, SYC.

5.6. Modelo de superpoblación normal

En este apartado se obtendrá una aproximación para la varianza conjunta cuando la superpoblación tiene una distribución normal. La función de densidad de la distribución normal con media μ y varianza σ^2 está dada por

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{\sigma^2}\right), \ y \in \mathbb{R},$$

su forma estándar corresponde a $\mu = 0$ y $\sigma^2 = 1$, en cuyo caso la función de densidad queda como sigue

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right).$$

El cálculo de la varianza total $V_{p,m}\left(\hat{\theta}_\pi\right)$, depende de las varianzas y covarianzas de las estadísticas de orden bajo el modelo. Las estadísticas de orden de la distribución

normal han sido estudiadas ampliamente en la literatura, sin embargo, no existe una expresión cerrada para sus momentos.

Autores como Jones [30] y Godwin [16] han derivado expresiones o cantidades numéricas exactas para los primeros dos momentos cuando el tamaño de la población es menor o igual a 10. Para tamaños de población mayores se han encontrado aproximaciones difíciles de manipular de manera algebraica, por ejemplo, Blom [4] propone aproximar el valor esperado de la $i-$ésima estadística de orden de N variables aleatorias mediante

$$E\left[Y_{(i)}\right] \approx \Phi^{-1}\left(\frac{i-\alpha}{N-2\alpha+1}\right),$$

donde $\Phi(z) = (2\pi)^{-1/2}\int_{-\infty}^{z}\exp\left(-z^2/2\right)dz$ es la función de distribución de la normal estándar y α es un parámetro que hace más o menos precisa la aproximación. Blom tabuló el valor de α requerido para que la aproximación genere el valor exacto de $E\left[Y_{(i)}\right]$ para diferentes valores de i y N, hace la conjetura que $\alpha \in (0.33, 0.5)$ y sugiere usar $\alpha = 3/8 = 0.375$.

Otra parte importante del trabajo que se ha realizado para $N > 10$ es numérico: se han tabulado valores esperados con cierta precisión, por ejemplo, Teichroew [55] generó tablas con el valor esperado y productos de las estadísticas de orden para $N \leq 20$ con 10 decimales, Harter [21] tabuló el valor esperado para $N \leq 400$ con 5 decimales. Royston [45] proporciona un algoritmo para calcular el valor esperado; Davis y Stephens [11], así como Shea y Scallon [50] publicaron algoritmos para aproximar la matriz de varianza-covarianza de las estadísticas de orden.

De manera alternativa, al igual que en el modelo de Laplace, en este trabajo se obtuvieron aproximaciones de la media, varianza y covarianza de las estadísticas de orden siguiendo la aproximación asintótica señalada en Hastings et al. [22]. Para aproximar la distribución normal estándar se tomó como base la dada por Polya [43], que es una expresión simple y muy precisa para estimar la función de distribución acumulada de esta variable aleatoria.

Proposición 5.6.1. *Sean $Y_{(1)}, \ldots, Y_{(N)}$ las estadísticas de orden de $Y_1, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas con función de densidad*

$f_Y(y) = (2\pi)^{-1/2} \exp(-y^2/2)$. *Una aproximación asintótica para* $E\left[Y_{(i)}\right]$, m_i, *es*

$$m_i = \begin{cases} -\left[-\frac{\pi}{2}\ln(4p_iq_i)\right]^{\frac{1}{2}} & \text{si } 0 < p_i < \frac{1}{2} \\ \left[-\frac{\pi}{2}\ln(4p_iq_i)\right]^{\frac{1}{2}} & \text{si } \frac{1}{2} \leq p_i < 1, \end{cases}$$

donde $p_i = i/(N+1)$ *y* $q_i = 1 - p_i$. *Una expresión asintótica para la covarianza entre* $Y_{(i)}$ *y* $Y_{(j)}$ $(i \leq j)$, m_{ij}, *está dada por*

$$m_{ij} = \frac{2\pi(N+1)^{\pi}}{2^{\pi}N(N+1)^2} \frac{i^{1-\frac{\pi}{4}}}{j^{\frac{\pi}{4}}} \frac{(N+1-j)^{1-\frac{\pi}{4}}}{(N+1-i)^{\frac{\pi}{4}}}, \tag{5.52}$$

y la aproximación asintótica para la varianza de la $i-$*ésima estadística de orden,* m_{ii}, *es*

$$m_{ii} = \frac{2\pi(N+1)^{\pi}}{2^{\pi}N(N+1)^2} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}}.$$

Demostración. La aproximación m_i debe de satisfacer

$$\Phi(m_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{m_i} \exp\left(-\frac{y^2}{2}\right) dy = \frac{i}{N+1} = p_i.$$

No existe una expresión cerrada para esta integral, en la literatura se pueden encontrar varias aproximaciones, algunas son más precisas para valores de m_i cercanos a 0, otras para valores en las colas, algunas son complicadas y otras más simples. En este caso se usará la expresión derivada por Polya [43], que es simple y muy precisa:

$$p = \Phi(z) \approx \frac{1}{2}\left\{1 + \left[1 - \exp\left(-\frac{2z^2}{\pi}\right)\right]^{\frac{1}{2}}\right\} \text{ si } z \geq 0,$$

el autor señala que el error en la estimación de $\Phi(z)$ es menor a 0.71 % de la integral aproximada. Despejando z se llega a

$$z = \Phi^{-1}(p) \approx \left[-\frac{\pi}{2}\ln(4pq)\right]^{\frac{1}{2}} \text{ si } \frac{1}{2} \leq p < 1.$$

A partir de esta última ecuación es fácil ver que

$$m_i = \begin{cases} -\left[-\frac{\pi}{2}\ln(4p_iq_i)\right]^{\frac{1}{2}} & \text{si } 0 < p_i \leq \frac{1}{2} \\ \left[-\frac{\pi}{2}\ln(4p_iq_i)\right]^{\frac{1}{2}} & \text{si } \frac{1}{2} \leq p_i < 1, \end{cases}$$

evaluando la función de densidad de probabilidad de la normal estándar en m_i se tiene que

$$f(m_i) = \frac{1}{\sqrt{2\pi}} (4p_i q_i)^{\frac{\pi}{4}} \text{ si } 0 < p_i < 1.$$

La aproximación para la covarianza está dada por

$$m_{ij} = \frac{i(N+1-j)}{N(N+1)^2 f(m_i) f(m_j)} \text{ si } i \leq j.$$

Sustituyendo $f(m_i)$

$$\begin{aligned} m_{ij} &= \frac{i(N+1-j)}{N(N+1)^2 \frac{(4p_i q_i)^{\frac{\pi}{4}}}{\sqrt{2\pi}} \frac{(4p_j q_j)^{\frac{\pi}{4}}}{\sqrt{2\pi}}} \\ &= \frac{2\pi}{2^\pi N(N+1)^2} i(N+1-j) \left[\frac{i}{N+1} \frac{N+1-i}{N+1} \frac{j}{N+1} \frac{N+1-j}{N+1} \right]^{-\frac{\pi}{4}} \\ &= \frac{2\pi(N+1)^\pi}{2^\pi N(N+1)^2} \frac{i^{1-\frac{\pi}{4}}}{j^{\frac{\pi}{4}}} \frac{(N+1-j)^{1-\frac{\pi}{4}}}{(N+1-i)^{\frac{\pi}{4}}} \text{ si } i \leq j. \end{aligned}$$

Haciendo $j = i$ en la aproximación para m_{ij} se tiene

$$m_{ii} = \frac{2\pi(N+1)^\pi}{2^\pi N(N+1)^2} \frac{i^{1-\frac{\pi}{4}}}{i^{\frac{\pi}{4}}} \frac{(N+1-i)^{1-\frac{\pi}{4}}}{(N+1-i)^{\frac{\pi}{4}}} = \frac{2\pi(N+1)^\pi}{2^\pi N(N+1)^2} i^{1-\frac{\pi}{2}} (N+1-i)^{1-\frac{\pi}{2}} \quad \square$$

Usando estas aproximaciones se derivó una expresión para la varianza total de $\hat{\theta}_\pi$, como se establece en las dos proposiciones siguientes.

Proposición 5.6.2. *Sea λ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes con función de densidad $f_Y(y) = (\sqrt{2\pi}\sigma)^{-1} \exp\left[-(y-\mu)^2/\sigma^2\right]$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera equilibrada (BSS), entonces una*

aproximación de la varianza conjunta del predictor $\hat{\theta}_\pi = T\sum_{i\in s_r} Y_{(i)}$ *está dada por*

$$\begin{aligned}
V_{BSS,\lambda}\left(\hat{\theta}_\pi\right) \approx T\sigma^2 \frac{2\pi(N+1)^\pi}{2^\pi N(N+1)^2} \Bigg\{ & \sum_{i=1}^{N} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}} \\
& + 4\sum_{r=1}^{T}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}} \frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=T'/2+1}^{T}\sum_{i=1}^{n/2} \Bigg[\frac{(-2r+2iT+2)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r-2iT+N-1)^{\frac{\pi}{4}}} \\
& - \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r+2iT+2)^{\frac{\pi}{4}}} \frac{(2r-2iT+N-1)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \Bigg] \Bigg\} \\
= & V^{aprox}_{BSS,\lambda}\left(\hat{\theta}_\pi\right)
\end{aligned} \tag{5.53}$$

donde $a = -2T-1$, $b = N+2+2T = N+1-a$ y $T' = \begin{cases} T & \text{si } T \text{ es par} \\ T+1 & \text{si } T \text{ es impar.} \end{cases}$

Demostración. La varianza conjunta de $\hat{\theta}$ es igual a

$$V_{BSS,\lambda}\left(\hat{\theta}_\pi\right) = E_{BSS}\left[V_\lambda\left(\hat{\theta}_\pi|s\right)\right] + V_{BSS}\left[E_\lambda\left(\hat{\theta}_\pi|s\right)\right]. \tag{5.54}$$

En seguida se aproximan cada una de estos términos. Sin pérdida de generalidad, supóngase que $E_\lambda[Y_i] = 0$, $V_\lambda[Y_i] = 1$ y $X_i = Y_{(i)}$ denota la $i-$ésima estadística de orden, entonces

$$\begin{aligned}
V_\lambda\left(\hat{\theta}_\pi|s\right) &= V_\lambda\left(T\sum_{i\in s_r} Y_i\,|s_r\right) \\
&= T^2\left\{\sum_{i\in s_r} V_\lambda\left[Y_{(i)}\,|s_r\right] + \sum_{i\neq j\in s_r} C_\lambda\left[Y_{(i)}, Y_{(j)}\,|s_r\right]\right\}.
\end{aligned} \tag{5.55}$$

La suma de varianzas es aproximadamente

$$\sum_{i\in s_r} V_\lambda\left(Y_{(i)}\,|s_r\right) \approx k\sum_{i\in s_r} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}}$$

donde $k = 2\pi(N+1)^{\pi}/\left[2^{\pi}N(N+1)^2\right]$. Desarrollando la suma de covarianzas dada en (5.55)

$$\begin{aligned}
\sum_{i\neq j\in s_r} C_\lambda\left[Y_{(i)}, Y_{(j)}\,|s_r\right] =& 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\lambda\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&+ 2\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n} C_\lambda\left(X_{2r-1+2(i-1)T}, X_{-2r+2+2(n-j+1)T}\right) \\
&+ 2\sum_{i=n/2+1}^{n-1}\sum_{j=i+1}^{n} C_\lambda\left(X_{-2r+2+2(n-i+1)T}, X_{-2r+2+2(n-j+1)T}\right) \\
=& 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\lambda\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&+ 2\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\lambda\left(X_{2r-1+2(i-1)T}, X_{N-2r+2+2(1-j)T}\right) \\
&+ 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\lambda\left(X_{N-2r+2+2(1-i)T}, X_{N-2r+2+2(1-j)T}\right).
\end{aligned}$$

Haciendo $a = -2T-1$, $b = N+2+2T = N+1-a$ en la primera suma de covarianzas y usando la aproximación dada en (5.52) se tiene que

$$\begin{aligned}
&\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\lambda\left(X_{2r+2iT+a}, X_{2r+2jT+a}\right) \\
&\qquad \approx k\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\frac{(N+1-2r-2jT-a)^{1-\frac{\pi}{4}}}{(N+1-2r-2iT-a)^{\frac{\pi}{4}}} \qquad (5.56) \\
&\qquad = k\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}}.
\end{aligned}$$

Análogamente, sustituyendo a y b en la tercera suma de covarianzas y usando la aproximación para m_{ij},

$$\begin{aligned}
&\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\lambda\left(X_{-2r-2iT+b}, X_{-2r-2jT+b}\right) \\
&\qquad \approx k\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}}\frac{(N+1+2r+2iT-b)^{1-\frac{\pi}{4}}}{(N+1+2r+2jT-b)^{\frac{\pi}{4}}}
\end{aligned}$$

$$= k \sum_{i=1}^{n/2-1} \sum_{j=i+1}^{n/2} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}}. \tag{5.57}$$

Sustituyendo a, b y usando la aproximación para m_{ij} en la segunda suma de covarianzas, cuando T es par y $r \leq T/2$, se tiene que

$$\begin{aligned}
&\sum_{i=1}^{n/2} \sum_{j=1}^{n/2} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\quad = \sum_{i=1}^{n/2} \sum_{j=1}^{n/2-i+1} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\qquad + \sum_{i=1}^{n/2} \sum_{j=n/2-i+2}^{n/2} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\quad \approx k \sum_{i=1}^{n/2} \sum_{j=1}^{n/2-i+1} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}} \frac{(N+1+2r+2jT-b)^{1-\frac{\pi}{4}}}{(N+1-2r-2iT-a)^{\frac{\pi}{4}}} \\
&\qquad + k \sum_{i=1}^{n/2} \sum_{j=n/2-i+2}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(N+1-2r-2iT-a)^{1-\frac{\pi}{4}}}{(N+1+2r+2jT-b)^{\frac{\pi}{4}}},
\end{aligned}$$

luego,

$$\begin{aligned}
&\sum_{i=1}^{n/2} \sum_{j=1}^{n/2} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\quad \approx k \sum_{i=1}^{n/2} \sum_{j=1}^{n/2-i+1} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}} \frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
&\qquad + k \sum_{i=1}^{n/2} \sum_{j=n/2-i+2}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}.
\end{aligned} \tag{5.58}$$

Si T es par y $r \geq T/2+1$

$$\begin{aligned}
&\sum_{i=1}^{n/2} \sum_{j=1}^{n/2} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\quad = \sum_{i=1}^{n/2} \sum_{j=1}^{n/2-i} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right) \\
&\qquad + \sum_{i=1}^{n/2} \sum_{j=n/2-i+1}^{n/2} C_\lambda \left(X_{2r+2iT+a}, X_{-2r-2jT+b}\right)
\end{aligned}$$

$$
\begin{aligned}
\approx & k\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(N+1+2r+2jT-b)^{1-\frac{\pi}{4}}}{(N+1-2r-2iT-a)^{\frac{\pi}{4}}} \\
& + k\sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(N+1-2r-2iT-a)^{1-\frac{\pi}{4}}}{(N+1+2r+2jT-b)^{\frac{\pi}{4}}} \\
= & k\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& + k\sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}.
\end{aligned}
\tag{5.59}
$$

Tomando el valor esperado bajo el diseño

$$
E_{BSS}\left[V_\lambda\left(\hat{\theta}_\pi|s\right)\right] = \frac{1}{T}\sum_{r=1}^{T}\left[T^2\sum_{i\in s_r}V_\lambda\left[Y_{(i)}\,|s_r\right] + T^2\sum_{i\neq j\in s_r}C_\lambda\left[Y_{(i)},Y_{(j)}\,|s_r\right]\right].
$$

Sustituyendo las aproximaciones obtenidas en (5.56), (5.57), (5.58) y (5.59)

$$
\begin{aligned}
E_{BSS}\left[V_\lambda\left(\hat{\theta}_\pi|s\right)\right] \approx & Tk\left[\sum_{i=1}^{N} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}}\right. \\
& + 4\sum_{r=1}^{T}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=1}^{T/2}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=1}^{T/2}\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \\
& + 2\sum_{r=T/2+1}^{T}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
& \left. + 2\sum_{r=T/2+1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\right].
\end{aligned}
$$

Por otra parte, para calcular el segundo término de la expresión (5.54), primero se tiene que derivar el valor esperado bajo el modelo,

$$
E_\lambda\left(\hat{\theta}_\pi|s\right) = E_\lambda\left(T\sum_{i\in s_r}Y_{(i)}\right) = T\sum_{i\in s_r}E_\lambda\left[Y_{(i)}\right] = 0
$$

Como la varianza de una constante es cero, entonces $V_{BSS}\left[E_\lambda\left(\hat{\theta}_\pi|s\right)\right]=0$. Por lo tanto,

$$
\begin{aligned}
V_{BSS,\lambda}\left(\hat{\theta}_\pi\right) =& E_{BSS}\left[V_\lambda\left(\hat{\theta}_\pi|s\right)\right] \\
\approx& T\frac{2\pi(N+1)^\pi}{2^\pi N(N+1)^2}\left\{\sum_{i=1}^{N} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}}\right. \\
&+4\sum_{r=1}^{T}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
&+2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
&+2\sum_{r=1}^{T}\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \\
&+2\sum_{r=T/2+1}^{T}\sum_{i=1}^{n/2}\left[\frac{(-2r+2iT+2)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}}\frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r-2iT+N-1)^{\frac{\pi}{4}}}\right. \\
&\left.\left.-\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r+2iT+2)^{\frac{\pi}{4}}}\frac{(2r-2iT+N-1)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}}\right]\right\}.
\end{aligned}
\tag{5.60}
$$

Cuando T es impar, las expresiones (5.56) y (5.57) se mantienen. La aproximación dada en (5.58) se cumple si $r \leq (T+1)/2$ y la aproximación (5.59) es válida si $r \geq (T+1)/2+1$, luego

$$
\begin{aligned}
E_{BSS}\left[V_\lambda\left(\hat{\theta}_\pi|s\right)\right] =& \frac{1}{T}\sum_{r=1}^{T}\left[T^2\sum_{i\in s_r}V_\lambda\left[Y_{(i)}|s_r\right]+T^2\sum_{i\neq j\in s_r}C_\lambda\left[Y_{(i)},Y_{(j)}|s_r\right]\right] \\
& Tk\left[\sum_{i=1}^{N} i^{1-\frac{\pi}{2}}(N+1-i)^{1-\frac{\pi}{2}}\right. \\
&+4\sum_{r=1}^{T}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}}\frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
&+2\sum_{r=1}^{(T+1)/2}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}\frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}}\frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}}
\end{aligned}
$$

$$
\begin{aligned}
\approx &+ 2 \sum_{r=1}^{(T+1)/2} \sum_{i=1}^{n/2} \sum_{j=n/2-i+2}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \\
&+ 2 \sum_{r=(T+1)/2+1}^{T} \sum_{i=1}^{n/2} \sum_{j=1}^{n/2-i} \frac{(2r+2iT+a)^{1-\frac{\pi}{4}}}{(-2r-2jT+b)^{\frac{\pi}{4}}} \frac{(2r+2jT+a)^{1-\frac{\pi}{4}}}{(-2r-2iT+b)^{\frac{\pi}{4}}} \\
&+ 2 \sum_{r=(T+1)/2+1}^{T} \sum_{i=1}^{n/2} \sum_{j=n/2-i+1}^{n/2} \frac{(-2r-2jT+b)^{1-\frac{\pi}{4}}}{(2r+2iT+a)^{\frac{\pi}{4}}} \frac{(-2r-2iT+b)^{1-\frac{\pi}{4}}}{(2r+2jT+a)^{\frac{\pi}{4}}} \Bigg].
\end{aligned}
$$

Considerando que $V_{BSS}\left[E_\lambda\left(\hat{\theta}_\pi|s\right)\right] = 0$, para T impar se tiene que $V_{BSS,\lambda}\left(\hat{\theta}_\pi\right)$ se aproxima por una expresión similar a la dada en (5.60) (T par), pero con $r = (T+1)/2+1, \ldots, T$ en el último sumando.

Finalmente, si $E_\lambda[Y_i] = \mu$ y $V_\lambda[Y_i] = \sigma^2$ en lugar de media cero y varianza unitaria, se tiene el resultado de la proposición. □

Proposición 5.6.3. *Sea λ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes con función de densidad $f_Y(y) = (\sqrt{2\pi}\sigma)^{-1} \exp\left(-(y-\mu)^2/\sigma^2\right)$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera creciente (SYC), entonces una aproximación de la varianza conjunta del predictor $\hat{\theta}_\pi = T\sum_{i\in s_r} Y_{(i)}$ está dada por*

$$
\begin{aligned}
V_{SYC,\lambda}\left(\hat{\theta}_\pi\right) \approx & T\sigma^2 \Bigg\{ N + 2k \sum_{r=1}^{T} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{[r+iT-T]^{1-\frac{\pi}{4}}}{[r+jT-T]^{\frac{\pi}{4}}} \frac{[-r-jT+d]^{1-\frac{\pi}{4}}}{[-r-iT+d]^{\frac{\pi}{4}}} \\
& + 2\sum_{r=1}^{T} \left(\sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(-r-iT+d)\right] \right\}^{\frac{1}{2}} \right)^2 \\
& - 2\sum_{r=1}^{T} \left(\sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(-r-iT+d)\right] \right\}^{\frac{1}{2}} \right. \\
& \left. \sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(-r-iT+e)\,(r+iT+f)\right] \right\}^{\frac{1}{2}} \right) \\
& + \pi \left[\frac{N}{2} \ln 4 - N\ln(N+1) + \ln(N!) \right] \Bigg\} \\
= & V_{SYC,\lambda}^{aprox}\left(\hat{\theta}_\pi\right)
\end{aligned}
\tag{5.61}
$$

donde $c = -\pi/2 \ln[4/(N+1)^2]$, $d = N+1+T$, $e = N/2+1+T$ y $f = N/2-T$.

Demostración. Como se señaló en la ecuación (5.8), la varianza total está dada por

$$\begin{aligned} V_{SYC,\lambda}\left(\hat{\theta}_\pi\right) =& T\left[N + (N-Nn)\mu^2 + \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum C_\lambda\left[Y_{(i)}, Y_{(j)}\right]\right] \\ &+ \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_\lambda\left[Y_{(i)}\right] E_\lambda\left[Y_{(j)}\right]. \end{aligned} \tag{5.62}$$

Usando la notación $X_i = Y_{(i)}$ y suponiendo que $E_\lambda[Y_i] = 0$ y $V_\lambda[Y_i] = 1$, la suma de covarianzas se aproxima mediante

$$\begin{aligned} \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum C_\lambda\left(X_i, X_j\right) =& 2\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} C_\lambda\left(X_{r+(i-1)T}, X_{r+(j-1)T}\right) \\ \approx& 2\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_{r+(i-1)T,r+(j-1)T}. \end{aligned}$$

Sustituyendo el valor de m_{ij}, la suma anterior queda como

$$\begin{aligned} \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum C_\lambda\left(X_i, X_j\right) \approx& 2k\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{[r+(i-1)T]^{1-\frac{\pi}{4}}}{[r+(j-1)T]^{\frac{\pi}{4}}}\frac{[N+1-r-(j-1)T]^{1-\frac{\pi}{4}}}{[N+1-r-(i-1)T]^{\frac{\pi}{4}}} \\ =& 2k\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{[r+iT-T]^{1-\frac{\pi}{4}}}{[r+jT-T]^{\frac{\pi}{4}}}\frac{[-r-jT+d]^{1-\frac{\pi}{4}}}{[-r-iT+d]^{\frac{\pi}{4}}}, \end{aligned} \tag{5.63}$$

donde $d = N+1+T$. La suma de productos de valores esperados es

$$\begin{aligned} \sum_{r=1}^{T}\sum_{\substack{i,j\in s_r \\ i\neq j}}\sum E_\lambda\left(X_i\right) E_\lambda\left(X_j\right) =& 2\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E_\lambda\left(X_{r+(i-1)T}\right) E_\lambda\left(X_{r+(j-1)T}\right) \\ \approx& 2\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_{r+(i-1)T}m_{r+(j-1)T} \\ =& 2\sum_{r=1}^{T}\left[\sum_{i=1}^{n/2} m_{r+(i-1)T}\right]^2 - 2\sum_{i=1}^{N/2} m_i^2 \\ & - 2\sum_{r=1}^{T}\sum_{i=1}^{n/2} m_{r+(i-1)T}\sum_{i=1}^{n/2} m_{\frac{N}{2}+1-r-(i-1)T}. \end{aligned}$$

Se demostró que si $0 < p_i \leq 1/2$ la aproximación asintótica del valor esperado bajo el modelo es $m_i = -\left[-\pi/2 \ln(4p_iq_i)\right]^{1/2} = -\left\{c - \pi/2 \ln[i(N+1-i)]\right\}^{1/2}$ donde $c = -\pi/2 \ln\left[4/(N+1)^2\right]$, entonces

$$\begin{aligned}
&\sum_{r=1}^{T} \sum_{\substack{i,j \in s_r \\ i \neq j}} \sum E_\lambda(X_i) E_\lambda(X_j) \\
&\qquad \approx 2 \sum_{r=1}^{T} \left(\sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(N+1-r-iT+T)\right] \right\}^{\frac{1}{2}} \right)^2 \\
&\qquad\quad - 2 \sum_{i=1}^{N/2} \left\{ c - \frac{\pi}{2} \ln\left[i(N+1-i)\right] \right\} \\
&\qquad\quad - 2 \sum_{r=1}^{T} \sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(N+1-r-iT+T)\right] \right\}^{\frac{1}{2}} \\
&\qquad\quad \sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[\left(\frac{N}{2} + 1 - r - iT + T \right) \left(\frac{N}{2} + r + iT - T \right) \right] \right\}^{\frac{1}{2}}
\end{aligned}$$

Haciendo $d = N+1+T$, $e = N/2+1+T$, $f = N/2-T$ y simplificando el segundo sumando

$$\begin{aligned}
&\sum_{r=1}^{T} \sum_{\substack{i,j \in s_r \\ i \neq j}} \sum E_\lambda(X_i) E_\lambda(X_j) \\
&\qquad \approx 2 \sum_{r=1}^{T} \left(\sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(-r-iT+d)\right] \right\}^{\frac{1}{2}} \right)^2 \\
&\qquad + \pi \left[\frac{N}{2} \ln 4 - N \ln(N+1) + \ln(N!) \right] \\
&\qquad - 2 \sum_{r=1}^{T} \sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(r+iT-T)(-r-iT+d)\right] \right\}^{\frac{1}{2}} \\
&\qquad \sum_{i=1}^{n/2} \left\{ c - \frac{\pi}{2} \ln\left[(-r-iT+e)(r+iT+f)\right] \right\}^{\frac{1}{2}}.
\end{aligned} \tag{5.64}$$

Sustituyendo (5.63) y (5.64) en la varianza total (5.62) se llega a

$$
\begin{aligned}
V_{SYC,\lambda}\left(\hat{\theta}_\pi\right) \approx & T\left\{N + 2k\sum_{r=1}^{T}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{[r+iT-T]^{1-\frac{\pi}{4}}}{[r+jT-T]^{\frac{\pi}{4}}}\frac{[-r-jT+d]^{1-\frac{\pi}{4}}}{[-r-iT+d]^{\frac{\pi}{4}}}\right. \\
& + 2\sum_{r=1}^{T}\left(\sum_{i=1}^{n/2}\left\{c-\frac{\pi}{2}\ln\left[(r+iT-T)(-r-iT+d)\right]\right\}^{\frac{1}{2}}\right)^2 \\
& - 2\sum_{r=1}^{T}\left(\sum_{i=1}^{n/2}\left\{c-\frac{\pi}{2}\ln\left[(r+iT-T)(-r-iT+d)\right]\right\}^{\frac{1}{2}}\right. \\
& \left.\left.\sum_{i=1}^{n/2}\left\{c-\frac{\pi}{2}\ln\left[(-r-iT+e)\,(r+iT+f)\right]\right\}^{\frac{1}{2}}\right) \\
& \left. + \pi\left[\frac{N}{2}\ln 4 - N\ln(N+1) + \ln(N!)\right]\right\}.
\end{aligned}
$$

Por último, considerando que $E_\lambda[Y_i] = \mu$ y $V_\lambda[Y_i] = \sigma^2$ en lugar de media cero y varianza unitaria, se obtiene la aproximación dada en (5.61). □

Las expresiones de las varianzas (5.53) y (5.61) se evaluaron numéricamente para poblaciones de tamaño $N = 120$ y 1200, estos cálculos se compararon con la varianza simulada $V_{p,\lambda}^{sim}\left(\hat{\theta}_\pi\right)$. Análogamente al caso de la distribución de Laplace, la varianza total, es

$$V_{p,\lambda}\left(\hat{\theta}_\pi\right) = E_{p,\lambda}\left[\hat{\theta}_\pi - N\mu\right]^2.$$

Entonces, la varianza por simulación se obtuvo generando $K = 10,000$ poblaciones de tamaño N bajo la distribución normal con $\mu = 0$ y $\sigma^2 = 1$ y calculando:

$$V_{p,\lambda}^{sim}\left(\hat{\theta}_\pi\right) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{r=1}^{T}\left(Tt_{s_r}^k\right)^2,$$

donde p : BSS, SYC. La muestra s_r es la obtenida según el diseño p y $t_{s_r}^k = \sum_{i\in s_r} y_i^k$ corresponde al total de la $r-$ésima muestra en la $k-$ésima población generada siguiendo el modelo λ.

Cuadro 5.5: Varianza total, bajo el diseño y el modelo $\lambda : Y_1, \ldots, Y_N$ tienen distribución normal con $\mu = 0$ y $\sigma^2 = 1$, $N = 120$

n	$V^{sim}_{BSS,\lambda}(\hat{\theta}_\pi)$	$V^{aprox}_{BSS,\lambda}(\hat{\theta}_\pi)$	$V^{sim}_{SYC,\lambda}(\hat{\theta}_\pi)$	$V^{aprox}_{SYC,\lambda}(\hat{\theta}_\pi)$	$V_{SI,\lambda}(\hat{\theta}_\pi)$	Diferencia[a/] BSS	Diferencia[a/] SYC
2	208.4	188.1	4,904.0	4,968.3	7,200	9.8	1.3
4	153.7	142.3	1,630.2	1,708.9	3,600	7.4	4.8
6	139.0	130.4	870.6	942.4	2,400	6.2	8.3
8	132.5	125.6	572.6	636.0	1,800	5.2	11.1
10	129.0	122.8	424.2	480.1	1,440	4.7	13.2
12	127.0	121.3	338.7	388.8	1,200	4.5	14.8
20	123.5	118.7	205.2	239.8	720	3.8	16.9
24	122.8	118.5	180.3	210.1	600	3.5	16.6
30	122.1	117.8	158.9	183.8	480	3.5	15.6
40	121.5	117.7	141.8	160.8	360	3.2	13.4
60	121.1	117.1	128.9	141.5	240	3.3	9.8

[a/] Corresponde a la diferencia relativa $\frac{|V^{aprox}_{p,\lambda} - V^{sim}_{p,\lambda}|}{V^{sim}_{p,\lambda}} \times 100$, p : BSS, SYC.

Los resultados se presentan en los Cuadros 5.5 y 5.6, en éstos se aprecia lo siguiente:

a) Cuando la población se ordena de manera equilibrada, la varianza aproximada es menor que la varianza obtenida cuando se simula. La diferencia entre la varianza simulada y la aproximada oscila entre el 3 y 15 % de la simulada.

b) Cuando la población se ordena de manera creciente, la varianza aproximada es mayor que la obtenida mediante simulación. La diferencia entre ambas varianzas oscila entre el 1 y 31 % de la simulada. Como se mencionó en la distribución de Laplace, la diferencia entre las varianzas se debe a dos efectos confundidos: al efecto de la aproximación de las medias y covarianzas de las estadísticas de orden y al efecto de la simulación.

c) La varianza del muestreo aleatorio simple es mayor que la del BSS y la del SYC. La varianza del SI es al menos dos veces la del muestreo sistemático cuando la población se ordena de manera equilibrada y al menos 1.5 veces cuando la población se ordena de manera creciente. Si $N = 120$, la varianza del aleatorio simple es a lo más 35 veces la del BSS y 4 veces la del SYC. Si

$N = 1200$, la varianza del aleatorio es a lo más 306 veces la del BSS y 10 veces la del SYC.

d) En el caso del BSS, la varianza se estabiliza para $n \geq 20$.

Cuadro 5.6: Varianza total, bajo el diseño y el modelo $\lambda : Y_1, \ldots, Y_N$ tienen distribución normal con $\mu = 0$ y $\sigma^2 = 1$, $N = 1200$

n	$V^{sim}_{BSS,\lambda}(\hat{\theta}_\pi)$	$V^{aprox}_{BSS,\lambda}(\hat{\theta}_\pi)$	$V^{sim}_{SYC,\lambda}(\hat{\theta}_\pi)$	$V^{aprox}_{SYC,\lambda}(\hat{\theta}_\pi)$	$V_{SI,\lambda}(\hat{\theta}_\pi)$	Diferencia[a/] BSS	Diferencia[a/] SYC
2	2,348.8	1,989.0	486,471.4	489,194.4	720,000	15.3	0.6
4	1,675.8	1,472.6	156,155.6	160,242.2	360,000	12.1	2.6
6	1,483.4	1,334.1	79,133.9	82,955.4	240,000	10.1	4.8
8	1,398.1	1,274.5	48,703.2	52,131.8	180,000	8.8	7.0
10	1,351.6	1,242.7	33,434.4	36,505.0	144,000	8.1	9.2
20	1,271.3	1,190.2	10,692.7	12,656.6	72,000	6.4	18.4
40	1,240.2	1,171.6	3,934.7	5,061.4	36,000	5.5	28.6
60	1,231.9	1,167.0	2,512.7	3,296.0	24,000	5.3	31.2
80	1,228.2	1,165.2	1,979.7	2,575.5	18,000	5.1	30.1
100	1,226.2	1,164.2	1,721.0	2,198.8	14,400	5.1	27.8
200	1,222.9	1,162.6	1,354.2	1,579.3	7,200	4.9	16.6
300	1,222.0	1,162.3	1,280.4	1,415.4	4,800	4.9	10.5
400	1,221.7	1,162.2	1,253.4	1,342.0	3,600	4.9	7.1
600	1,221.4	1,162.0	1,233.5	1,274.6	2,400	4.9	3.3

[a/] Corresponde a la diferencia relativa $\frac{|V^{aprox}_{p,\lambda} - V^{sim}_{p,\lambda}|}{V^{sim}_{p,\lambda}} \times 100$, p : BSS, SYC.

Para los tres modelos de superpoblación que se consideraron: uniforme, Laplace y normal, se mostró que el BSS y el SYC son más eficientes que el SI; también se observó que al ordenar la población de manera equilibrada, se reduce la varianza del predictor del total en comparación con un orden creciente. Esto se cumple si el tamaño de muestra es un número par y si $T = N/n$ es un entero; si N/n no es un entero entonces la varianza total tendría otro componente debido a la varianza del tamaño de muestra que sería una variable aleatoria; bajo BSS si n es un número impar, no se eliminaría el segundo término de la varianza total, $V_{BSS}\left[E_m(\hat{\theta}_\pi \,|s\,)\right]$, como sucede cuando n es par.

Además del orden de la población, otra condición importante para la mayor eficiencia del BSS es que las variables aleatorias sean simétricas, como las tres anteriores. Aunque no se demostró que $V_{BSS,m}\left(\hat{\theta}_\pi\right) < V_{SYC,m}\left(\hat{\theta}_\pi\right)$ para cualquier variable aleatoria simétrica, en la siguiente sección se mostrará por medio de simulaciones que esta condición se cumple para la distribución normal generalizada.

5.7. Modelo de superpoblación normal generalizada

En esta sección se supone que la variable de interés tiene una distribución más general: la distribución normal generalizada. La uniforme, Laplace y normal, vistas en las secciones anteriores, son casos particulares de esta distribución.

Una variable aleatoria Y sigue una distribución normal generalizada si su función de densidad está dada por

$$f_Y(y) = \frac{1}{2\Gamma(1+1/\omega)A(\omega,\sigma)} \exp\left\{-\left|\frac{y-\mu}{A(\omega,\sigma)}\right|^{\omega}\right\},\ y \in \mathbb{R},\ \omega > 0,\ \sigma > 0 \text{ y } \mu < \infty \tag{5.65}$$

donde $A(\omega,\sigma) = \left[\sigma^2\Gamma(1/\omega)/\Gamma(3/\omega)\right]^2$ y $\Gamma(\cdot)$ denota la función gamma. En este modelo el parámetro μ denota la media y σ^2 denota la varianza de la variable aleatoria, el parámetro de forma ω es una medida relacionada con la curtosis de la función de densidad. La distribución de Laplace o doble exponencial es un caso particular cuando $\omega = 1$, $\omega = 2$ corresponde a la distribución normal y el caso límite $\omega \to \infty$ converge a la distribución uniforme en el intervalo $[\mu - \sqrt{3}\sigma,\ \mu + \sqrt{3}\sigma]$.

En la Figura 5.3 se presenta esta función de densidad para cinco valores del parámetro ω, en la que se observa que conforme ω aumenta la curva se va haciendo más plana.

Con excepción de casos particulares como la distribución de Laplace, $\omega = 1$, y la uniforme, $\omega \to \infty$, no se tienen expresiones cerradas para la varianza y la covarianza de las estadísticas de orden de la distribución normal generalizada, para cualquier valor de ω. Por esta razón la varianza total $V_{p,\Lambda}\left(\hat{\theta}_\pi\right)$ se aproximó por simulación.

Sea Λ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes

con función de densidad dada por (5.65), con parámetros $\mu = 0$, $\sigma = 1$ y $\omega = 1, 2, 3, 10$. Para cada valor de ω se generaron $K = 10,000$ poblaciones bajo el modelo Λ con $N = 1,200$ unidades cada una. La varianza total se aproximó mediante

$$V_{p,\Lambda}\left(\hat{\theta}_\pi\right) = E_{p,\Lambda}\left[\hat{\theta}_\pi - E_{p,\Lambda}\left(\hat{\theta}_\pi\right)\right]^2,$$

como $E_{p,\Lambda}\left(\hat{\theta}_\pi\right) = N\mu = 0$, entonces

$$V_{p,\Lambda}\left(\hat{\theta}_\pi\right) \approx \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{r=1}^{T}\left(Tt_{s_r}^k\right)^2 = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T}\sum_{r=1}^{T}\left(Tt_{s_r}^k\right)^2 = V_{p,\Lambda}^{sim}\left(\hat{\theta}_\pi\right)$$

donde $t_{s_r}^k = \sum_{i\in s_r} y_i^k$ corresponde al total de la $r-$ésima muestra obtenida según el diseño, en la $k-$ésima población generada bajo el modelo Λ. En §5.4 se obtuvo la varianza para la distribución uniforme, es decir, cuando $\omega \to \infty$. El diseño p corresponde al muestreo sistemático con un orden equilibrado (BSS) y al muestreo sistemático con un orden creciente (SYC).

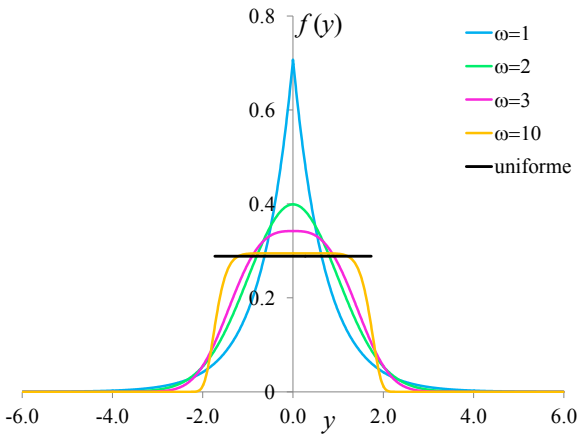
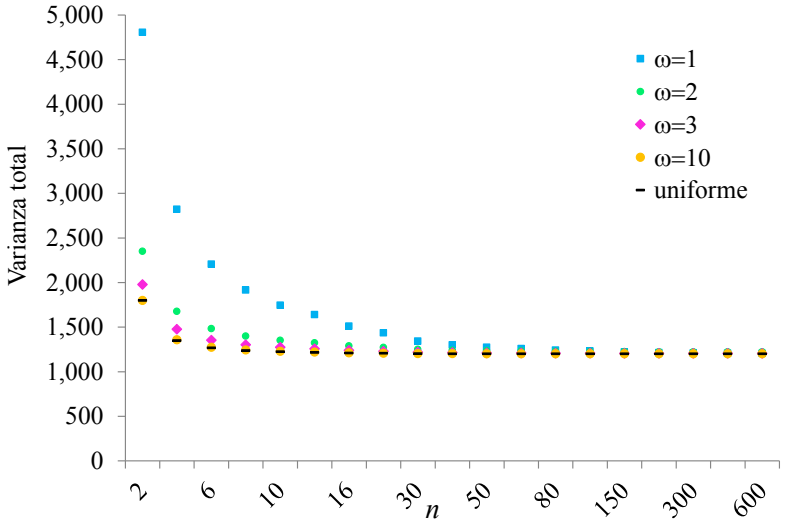
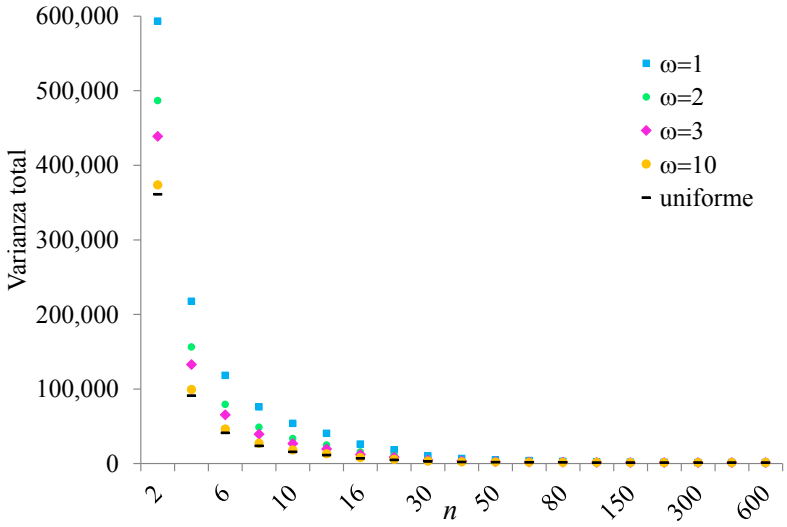


Figura 5.3: Función de densidad de la variable aleatoria normal generalizada con parámetros $\mu = 0$, $\sigma = 1$ y $\omega = 1$ (Laplace), $\omega = 2$ (normal), $\omega = 3$, $\omega = 10$ y $\omega \to \infty$ (uniforme).

Los resultados de la simulación se muestran en la Figura 5.4 (a) y (b). En ambas gráficas se aprecia lo siguiente.



(a) Muestreo sistemático, orden equilibrado (BSS)



(b) Muestreo sistemático, orden creciente (SYC)

Figura 5.4: Varianza $V_{p,\Lambda}^{sim}\left(\hat{\theta}_\pi\right)$ según valor de ω y varianza $V_{p,\xi}\left(\hat{\theta}_\pi\right)$ para $\omega \to \infty$ donde $p : BSS, SYC$; $\omega = 1$ corresponde a la distribución de Laplace, $\omega = 2$ a la normal y $\omega \to \infty$ a la uniforme; $N = 1200$.

a) Para n fijo, conforme el valor de ω aumenta, la varianza del predictor disminuye; la menor varianza corresponde a la distribución uniforme y la mayor a la distribución de Laplace.

b) Para un valor de ω fijo, conforme el tamaño de muestra aumenta, la varianza disminuye, el menor cambio es a partir de $n = 20$ unidades.

c) Al comparar la varianza entre los dos diseños, se encontró que el muestreo sistemático y un orden equilibrado es más eficiente que el muestreo sistemático y un orden creciente, esto se cumple para cualquier tamaño de muestra y valor del parámetro de forma ω. La eficiencia del BSS es menor conforme el tamaño de muestra aumenta (ver Cuadro 5.7); por ejemplo, si $N = 1200$ y $\omega = 3$, la eficiencia relativa del BSS respecto al SYC varía entre 221.9 si $n = 2$ y 1.01 si $n = 600$.

Cuadro 5.7: Eficiencia relativa del muestreo sistemático equilibrado, respecto al sistemático y orden creciente, $N = 1200$

$\frac{n}{N} \times 100$	Tamaño de muestra n	Eficiencia relativa $^{/a}$ Laplace $\omega = 1$	normal $\omega = 2$	$\omega = 3$	$\omega = 10$	uniforme $\omega \to \infty$
0.2	2	123.41	207.12	221.87	207.63	200.67
0.3	4	77.04	93.18	89.92	73.20	67.52
0.5	6	53.53	53.35	48.24	36.42	32.50
0.7	8	39.63	34.83	30.33	21.87	19.14
0.8	10	30.75	24.74	21.02	14.76	12.74
1.0	12	24.66	18.61	15.56	10.75	9.20
1.3	16	17.13	11.87	9.74	6.63	5.64
1.7	20	12.75	8.41	6.85	4.67	3.98
2.5	30	7.40	4.64	3.79	2.67	2.33
3.3	40	5.06	3.17	2.63	1.96	1.75
4.2	50	3.82	2.45	2.08	1.62	1.48
5.0	60	3.08	2.04	1.77	1.43	1.33
6.7	80	2.28	1.61	1.44	1.25	1.19
8.3	100	1.87	1.40	1.29	1.16	1.12
12.5	150	1.42	1.19	1.13	1.07	1.05
16.7	200	1.25	1.11	1.08	1.04	1.03
25.0	300	1.11	1.05	1.03	1.02	1.01
33.3	400	1.06	1.03	1.02	1.01	1.01
50.0	600	1.02	1.01	1.01	1.00	1.00

Modelo $\Lambda : Y_1, \ldots, Y_N$ son variables aleatorias independientes con función de distribución normal generalizada y parámetros $\mu = 0$, $\sigma^2 = 1$ y $\omega = 1, 2, 3, 10$ y $\omega \to \infty$.

$^{a/}$ Eficiencia relativa $= V^{sim}_{SYC,\Lambda}\left(\hat{\theta}_\pi\right) / V^{sim}_{BSS,\Lambda}\left(\hat{\theta}_\pi\right)$ si $\omega = 1, 2, 3, 10$

Eficiencia relativa $= V_{SYC,\xi}\left(\hat{\theta}_\pi\right) / V_{BSS,\xi}\left(\hat{\theta}_\pi\right)$ si $\omega \to \infty$.

Cuadro 5.8: Eficiencia relativa del muestreo sistemático, orden equilibrado, respecto al aleatorio simple, $N = 1200$

$\frac{n}{N} \times 100$	Tamaño de muestra n	Eficiencia relativa $^{/a}$ Laplace $\omega = 1$	normal $\omega = 2$	$\omega = 3$	$\omega = 10$	uniforme $\omega \to \infty$
0.2	2	149.8	306.5	364.1	400.2	400.3
0.3	4	127.6	214.8	244.0	265.4	266.7
0.5	6	108.9	161.8	177.3	188.6	189.5
0.7	8	94.0	128.7	138.4	144.8	145.5
0.8	10	82.5	106.5	113.1	117.3	117.7
1.0	12	73.2	90.6	95.4	98.4	98.6
1.3	16	59.6	69.8	72.7	74.3	74.4
1.7	20	50.1	56.6	58.6	59.6	59.7
2.5	30	35.8	38.4	39.4	39.9	39.9
3.3	40	27.7	29.0	29.7	29.9	30.0
4.2	50	22.6	23.3	23.8	24.0	24.0
5.0	60	19.1	19.5	19.9	20.0	20.0
6.7	80	14.5	14.7	14.9	15.0	15.0
8.3	100	11.7	11.7	11.9	12.0	12.0
12.5	150	7.9	7.8	8.0	8.0	8.0
16.7	200	5.9	5.9	6.0	6.0	6.0
25.0	300	4.0	3.9	4.0	4.0	4.0
33.3	400	3.0	2.9	3.0	3.0	3.0
50.0	600	2.0	2.0	2.0	2.0	2.0

Modelo $\Lambda : Y_1, \ldots, Y_N$ son variables aleatorias independientes con función de distribución normal generalizada y parámetros $\mu = 0$, $\sigma^2 = 1$ y $\omega = 1, 2, 3, 10$ y $\omega \to \infty$.

$^{a/}$ Eficiencia relativa $= V_{SI,\Lambda}\left(\hat{\theta}_\pi\right) / V^{sim}_{BSS,\Lambda}\left(\hat{\theta}_\pi\right)$ si $\omega = 1, 2, 3, 10$

Eficiencia relativa $= V_{SI,\xi}\left(\hat{\theta}_\pi\right) / V_{BSS,\xi}\left(\hat{\theta}_\pi\right)$ si $\omega \to \infty$.

En términos de la eficiencia relativa del BSS con respecto al muestreo aleatorio simple, en el Cuadro 5.8 puede verse lo siguiente

a) La eficiencia disminuye conforme el tamaño de muestra aumenta.

b) La ganancia de este tipo de muestreo es notablemente mayor cuando el tamaño de muestra es pequeño, $n < 10$.

c) Si $n \geq 20$, la eficiencia para diferentes valores del parámetro ω es muy parecida,

siendo mayor la correspondiente a la distribución uniforme.

d) La menor eficiencia para los diferentes valores de ω y tamaños de muestra considerados fue de 2 y la mayor de 400.

Cuadro 5.9: Eficiencia relativa del muestreo sistemático, orden creciente, respecto al aleatorio simple, $N = 1200$

	Tamaño de	Eficiencia relativa $^{/a}$				
$\frac{n}{N} \times 100$	muestra	Laplace	normal			uniforme
	n	$\omega = 1$	$\omega = 2$	$\omega = 3$	$\omega = 10$	$\omega \to \infty$
0.2	2	1.2	1.5	1.6	1.9	2.0
0.3	4	1.7	2.3	2.7	3.6	4.0
0.5	6	2.0	3.0	3.7	5.2	5.8
0.7	8	2.4	3.7	4.6	6.6	7.6
0.8	10	2.7	4.3	5.4	7.9	9.2
1.0	12	3.0	4.9	6.1	9.2	10.7
1.3	16	3.5	5.9	7.5	11.2	13.2
1.7	20	3.9	6.7	8.6	12.8	15.0
2.5	30	4.8	8.3	10.4	14.9	17.2
3.3	40	5.5	9.1	11.3	15.3	17.2
4.2	50	5.9	9.5	11.5	14.8	16.2
5.0	60	6.2	9.6	11.2	13.9	15.0
6.7	80	6.4	9.1	10.3	12.0	12.6
8.3	100	6.3	8.4	9.3	10.4	10.7
12.5	150	5.5	6.6	7.0	7.5	7.6
16.7	200	4.8	5.3	5.6	5.8	5.8
25.0	300	3.6	3.7	3.9	3.9	4.0
33.3	400	2.8	2.9	2.9	3.0	3.0
50.0	600	1.9	1.9	2.0	2.0	2.0

Modelo $\Lambda : Y_1, \ldots, Y_N$ son variables aleatorias independientes con función de distribución normal generalizada y parámetros $\mu = 0$, $\sigma^2 = 1$ y $\omega = 1, 2, 3, 10$ y $\omega \to \infty$.

$^{a/}$ Eficiencia relativa $= V_{SI,\Lambda}\left(\hat{\theta}_\pi\right) / V^{sim}_{SYC,\Lambda}\left(\hat{\theta}_\pi\right)$ si $\omega = 1, 2, 3, 10$

Eficiencia relativa $= V_{SI,\xi}\left(\hat{\theta}_\pi\right) / V_{SYC,\xi}\left(\hat{\theta}_\pi\right)$ si $\omega \to \infty$.

La eficiencia relativa del SYC con respecto al muestreo aleatorio simple se presenta en el Cuadro 5.9. Se observa que

a) La eficiencia relativa muestra un patrón en forma de "U" invertida, es decir, creciente hasta llegar a cierta n y decreciente a partir de este punto; el punto de cambio varía según el valor de ω: cuando $\omega = 1$ el patrón cambia en $n = 80$ y disminuye hasta llegar a $n = 30$ o $n = 40$ para la distribución uniforme.

b) Al igual que en el BSS, la mayor eficiencia se obtiene cuando la población tiene una distribución uniforme.

c) La menor eficiencia fue de 1.2 y la mayor de 17.

Capítulo 6

Aproximaciones de la varianza del estimador del total

La desventaja principal del muestreo sistemático es que no existe un estimador insesgado de la varianza del estimador del total basado en una sola muestra y cualquier estimación depende de los supuestos que se hagan sobre la forma de la población muestreada. En la literatura se han desarrollado varias aproximaciones, mismas que se pueden agrupar en a) modificar el esquema de selección al elegir más de un arranque aleatorio, b) utilizar algún método de remuestreo y c) hacer algún supuesto sobre la estructura de la población de interés.

El primer tipo de aproximación consiste en seleccionar k réplicas de muestras sistemáticas de tamaño $n' = n/k$ cada una, lo cual implica seleccionar k arranques aleatorios de manera independiente. A estas réplicas se les conoce como muestras sistemáticas interpenetrantes. Este método genera un estimador insesgado de la varianza, pero tiene la desventaja de que se pierde precisión si el número de replicas es pequeño (por ejemplo 2 ó 3) o si el número de réplicas se aumenta dejando el tamaño total de muestra fijo n (Iachan [25], Gautschi [15]).

Si se tienen k submuestras, el estimador de la varianza corresponde al de un muestreo aleatorio simple de k conglomerados de un total de T, con n' elementos cada uno, esto es,

$$v_{SY}^{k}\left(\hat{t}_{\pi}\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{k}\sum_{\alpha=1}^{k}\frac{\left(\bar{y}_{\alpha}-\bar{y}\right)^2}{k-1},$$

donde $\bar{y}_{\alpha}$ es la media muestral de la $\alpha-$ésima submuestra sistemática y $\bar{y}$ es la media

muestral basada en $n = n'k$ elementos.

La varianza también se ha aproximado usando métodos de remuestreo como jackknife, bootstrap y grupos aleatorios. Estos procedimientos se han vuelto comunes en los últimos años ya que se ha demostrado su utilidad en diseños muestrales como el muestreo estratificado bietápico, así como para estimadores lineales. El uso de estos métodos se ha incrementado en la práctica debido al desarrollo y disponibilidad de programas computacionales.

Uno de los métodos más usados es el de jackknife, cuyas propiedades teóricas han sido estudiadas y ha revelado un buen comportamiento en diversos estudios empíricos. Este procedimiento fue introducido por Quenouille en 1949 como un método para reducir el sesgo de un estimador en el contexto de poblaciones infinitas y Tukey en 1958 sugirió esta técnica para estimar varianza, Durbin en 1959 considera esta técnica en el caso de poblaciones finitas (Wolter [59], §4.1).

La estimación de la varianza se hace considerando diferentes submuestras o réplicas de la muestra completa, calculando la varianza del estimador a partir de la variabilidad entre las estimaciones del parámetro de interés para las diferentes réplicas. En el caso del muestreo sistemático cada submuestra se obtiene al eliminar un grupo de unidades de la muestra total, una práctica común consiste en quitar una unidad a la vez, por ello el número de submuestras posibles es n. Existen varios estimadores de varianza jackknife (ver por ejemplo Särndal [54], §11.5), uno de ellos corresponde al siguiente

$$v_{JAC}\left(\hat{t}_\pi\right) = \frac{n-1}{n}\sum_{i\in s_r}\left(\hat{t}_{\pi(i)} - \hat{t}_\pi\right)^2,$$

donde $\hat{t}_{\pi(i)}$ es el estimador de Horvitz-Thompson del total poblacional excluyendo la $i-$ésima unidad.

Otro tipo de aproximación de la varianza se ha obtenido al hacer algún supuesto sobre la estructura de la población de interés. Cochran [6] por ejemplo, al suponer una superpoblación en la cual la correlación entre dos unidades separadas u unidades es una función exponencial, $\rho_u = e^{-\lambda u}$, obtiene que si n y T son grandes, una estimación de la varianza es $v_{SY}\left(\hat{t}_\pi\right) = s^2\left[1 - 2(T\lambda)^{-1} + 2(e^{T\lambda} - 1)^{-1}\right]/n$ donde se estima ρ_T con la muestra y se despeja para obtener una estimación de λ.

Berger [3] propone un estimador de varianza para un muestreo sistemático basado

en regresión. Estima una regresión en donde la variable de interés es la dependiente y las independientes corresponden a variables correlacionadas con la variable de interés, el estimador considera los residuos de la regresión, el coeficiente de regresión ponderado y la matriz de varianza-covarianza de las variables independientes.

Opsomer et al. [42] proponen un estimador de la varianza bajo un modelo no paramétrico, usando regresión polinomial local como el método de estimación. El modelo para la superpoblación que consideran es

$$Y_i = m(x_i) + v(x_i)^{1/2} e_i,\ i = 1, \ldots, N$$

donde $m(\cdot)$ y $v(\cdot)$ son funciones continuas y acotadas. Los errores e_i $(i = 1, \ldots, N)$ son variables aleatorias independientes con media cero y varianza unitaria. Para estimar las funciones m y v usan una regresión polinomial local.

Dependiendo de los supuestos sobre la población se han obtenido diferentes estimadores de la varianza del muestreo sistemático, que son en general insesgados bajo el modelo de superpoblación que se supone, Iachan [25]. Wolter [58, 59] compara ocho estimadores sesgados que en su experiencia son representativos de los que resultan útiles en la práctica:

a) Tratar la muestra sistemática como una muestra aleatoria simple sin reemplazo.
b) Estimar la varianza con base en diferencias sucesivas entre los valores muestrales.
c) Suponer que la muestra sistemática proviene de un muestreo aleatorio estratificado con dos unidades seleccionadas de cada estrato de $2T$ unidades.
d) Considerar la diferencia de segundo orden entre los datos en la muestra.
e) Suponer que la correlación entre dos elementos separados u unidades es de tipo exponencial.
f) Considerar diferencias de observaciones sucesivas hasta de orden 4.
g) Considerar diferencias hasta de orden 8.
h) Hacer una partición de la muestra en k submuestras de tamaño n/k.

El autor analiza las propiedades empíricas de estos estimadores, así como algunas de sus propiedades teóricas suponiendo cinco modelos de superpoblaciones: aleatorio, tendencia lineal, efectos de estratificación, autocorrelacionado de orden uno y efectos periódicos. Sugiere que si no se puede suponer algún modelo para la población, se

usen los estimadores señalados en b) y c), los cuales parecen ser buenas aproximaciones para diferentes clases de poblaciones.

A continuación se precisan tres de las aproximaciones encontradas en la literatura y que se usarán en la comparación empírica que se hará en el capítulo siguiente. Se eligieron estos tres estimadores por dos razones: i) autores como Wolter [58] los recomiendan y ii) tuvieron menor sesgo en una comparación efectuada en el desarrollo de esta tesis con algunos conjuntos de datos reales. Los primeros dos se derivaron en el muestreo de poblaciones finitas y el tercero en el contexto de poblaciones continuas.

i. Estimador que toma la primera diferencia entre los datos (FDI del inglés *first differences*). Se basa en diferencias sucesivas de los valores muestrales.

$$v_{FDI}\left(\hat{t}_\pi\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=2}^{n}\frac{\left(y_i-y_{i-1}\right)^2}{2(n-1)}, \tag{6.1}$$

donde $n/N = 1/T$ es la fracción de muestreo. Cuando no haya lugar a dudas y para simplificar la notación se usará indistintamente $i \in s_r$ o $i = 1, 2, \ldots, n$ para denotar que la unidad i pertenece a la muestra.

Fuller [13], resultado 5.3.5, señala que este estimador resulta de suponer un modelo local en el que dos observaciones adyacentes tienen la misma media, es decir, $Y_i = \mu_j + e_i$, $i \in [j, j+1]$, con $E(e_i) = 0$, $E(e_i^2) = \sigma_i^2$ y $E(e_i e_j) = 0$ $(i \neq j)$. Corresponde al estimador que se basa en diferencias sucesivas del trabajo de Wolter [58].

ii. Estimador que se basa en la diferencia de segundo orden entre los datos muestrales (SDI del inglés *second differences*)

$$v_{SDI}\left(\hat{t}_\pi\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\sum_{i=3}^{n}\frac{\left(y_i-2y_{i-1}+y_{i-2}\right)^2}{6(n-2)}. \tag{6.2}$$

Cochran [7] señala que este estimador resulta de suponer un modelo lineal para la población, es decir, $Y_i = \mu + bi + e_i$ con $E(e_i) = 0$, $E(e_i^2) = \sigma_i^2$ y $E(e_i e_j) = 0$ $(i \neq j)$. Además, es uno de los estimadores que Wolter [58] recomienda. Fuller [13], resultado 5.3.7, propone un estimador muy similar:

$$v_{SDI'}\left(\hat{t}_\pi\right) = N^2\left(1-\frac{n}{N}\right)\frac{1}{n}\left[\sum_{i=3}^{n}\frac{\left(y_i-2y_{i-1}+y_{i-2}\right)^2}{6n} + \frac{\left(y_1-y_2\right)^2+\left(y_{n-1}-y_n\right)^2}{2n}\right].$$

iii. Estimador basado en el covariograma (COV). Como se vio en §3.4, cuando se quiere estimar el área bajo una función f en el intervalo A, $t = \int_A f(x)dx$, una aproximación de la varianza del estimador, si T es suficientemente pequeño, está dada por el término de extensión señalado en la ecuación (3.12). Haciendo $\beta_{T,2q+1} = T^{2q+1} g^{(2q+1)}(0^+)/(2q+1)!$, $V_E(\hat{t})$ se puede reescribir como

$$V_E(\hat{t}) = -\frac{B_{2q+2}}{(q+1)} T \beta_{T,2q+1} \tag{6.3}$$

donde $q \in \mathbb{Z}^+ \cup \{0\}$ es el orden de la primera derivada de la función f que presenta saltos, B_l es el $l-$ésimo número de Bernoulli y $g^{(2q+1)}$ es la $(2q+1)-$ésima derivada del covariagrama de f.

Para estimar el término de extensión, Kiêu [32] obtiene una aproximación del covariograma cerca del origen, $g^{(2q+1)}(0^+)$. Esta aproximación a) considera que las estimaciones del covariograma están disponibles en un conjunto discreto de valores, b) usa la fórmula de Taylor y c) aplica el método de mínimos cuadrados para estimar los coeficientes involucrados en la expansión de Taylor. El estimador resultante tiene la forma

$$v_E(\hat{t}) = -\frac{B_{2q+2}}{(q+1)} T \sum_k \lambda_k C_k \tag{6.4}$$

donde $C_k = \sum_{i=1}^{n-k} y_i y_{i+k}$.

Posteriormente, García-Fiñana y Cruz-Orive [14] generalizan la ecuación (6.4) para $q \in [0, 1]$

$$v_E\left(\hat{t}\right) = \alpha(q)\left[3C_0 - 4C_1 + C_2\right] T^2$$

donde $\alpha(q) = \Gamma(2q+2)\zeta(2q+2)\cos(q\pi)/\left[(2\pi)^{2q+2}(1-2^{2q-1})\right]$, $\zeta(w) = \sum_{k=1}^{\infty} k^{-w}$ denota la función Zeta de Riemann y q es una constante o parámetro de suavización que se puede estimar mediante

$$\hat{q} = \frac{1}{2\ln k} \ln\left(\frac{3C_0 - 4C_k + C_{2k}}{3C_0 - 4C_1 + C_2}\right) - \frac{1}{2}. \tag{6.5}$$

No existe una guía para el valor apropiado de $k = 2, 3, \ldots$, pero los autores recomiendan $k = 2, 4$.

En el caso de un conjunto de datos discretos, la función f no es continua, por lo que puede suponerse que $q = 0$ y por ende $\alpha(0) = 1/12$. Si además se considera la corrección por finitud, $(1 - n/N)$, el estimador toma la forma siguiente

$$v_{COV}\left(\hat{t}_\pi\right) = N^2\left(1 - \frac{n}{N}\right)\frac{1}{n^2}\frac{[3C_0 - 4C_1 + C_2]}{12}. \tag{6.6}$$

Como lo señala Wolter [59], p. 302, el factor de corrección por finitud no es un componente necesario en los estimadores de la varianza, pues la expresión de la varianza no incluye de manera explícita el factor de corrección por finitud. No obstante, si n/N es pequeña, el factor tendrá un efecto pequeño en la varianza, si por otra parte n/N se acerca a uno, la varianza tenderá a cero. En el caso del enfoque basado en el covariograma, aproximaciones que arrojan varianza cero cuando el tamaño muestral se acerca al de la población se obtienen en Cruz-Orive y Gual Arnau [18] y en Cruz-Orive [9].

En el próximo capítulo, usando cinco conjuntos de datos, se comparan las tres aproximaciones de la varianza señaladas arriba, así como la varianza del estimador, considerando tres diseños: BSS, SYC y SI.

Capítulo 7

Evaluación empírica de la varianza del muestreo sistemático

En este capítulo se hace un análisis empírico usando cinco conjuntos de datos, cuatro reales y uno simulado. Los objetivos del análisis son i) comparar la eficiencia relativa del muestreo sistemático, considerando dos órdenes de la población: equilibrado y creciente, con respecto al aleatorio simple y ii) comparar los tres estimadores de la varianza que se señalan en §6.1 en términos de su sesgo.

Entre los estimadores existentes en la literatura se utilizan $v_{FDI}\left(\hat{t}_\pi\right)$ dado en la expresión (6.1), $v_{SDI}\left(\hat{t}_\pi\right)$ en la expresión (6.2) y $v_{COV}\left(\hat{t}_\pi\right)$ en la expresión (6.6).

Como el grado de simetría de la población juega un papel importante en la eficiencia del sistemático (orden simétrico), para cada conjunto de datos se calculó el siguiente coeficiente de asimetría

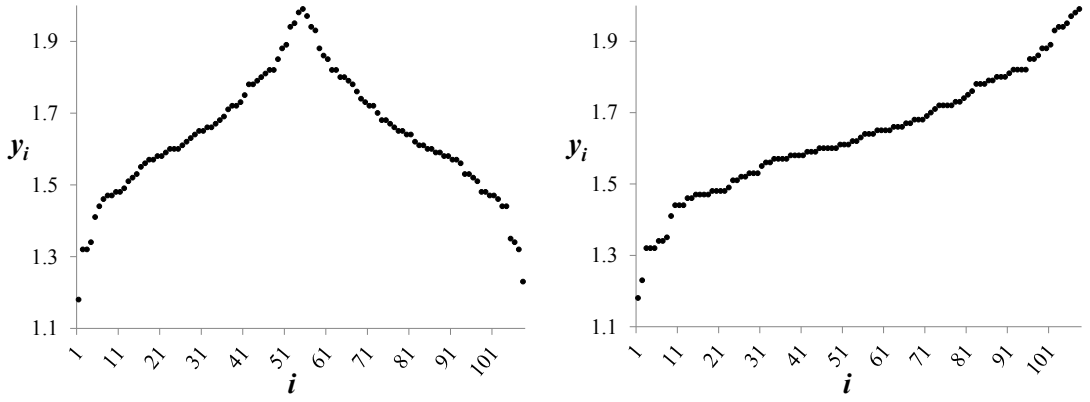
$$\gamma_3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_U)^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2\right]^{3/2}}$$

donde y_i es el valor de la variable de interés para el $i-$ésimo elemento y $\overline{y}_U$ es la media poblacional. En general un coeficiente de asimetría de la variable Y se define mediante (ver Kendall [31], p. 88):

$$\gamma_3^{'} = \frac{\mu_3}{\sigma^3} = \frac{E\left[(Y-\mu)^3\right]}{\left\{E\left[(Y-\mu)^2\right]\right\}^{3/2}}$$

En el caso de una distribución simétrica, como la normal generalizada, $\gamma_3^{'} = 0$.

Los cinco conjuntos de datos son independientes y diferentes en cuanto a la variable de interés, fuente de información y tamaño de población. De los cinco ejemplos dos tienen un número de unidades pequeño, un poco más de 100 elementos, otros dos un número de elementos más grande, alrededor de 2,500 y el quinto tiene $N = 32{,}000$ unidades. Estos se describen con mayor detalle a continuación.



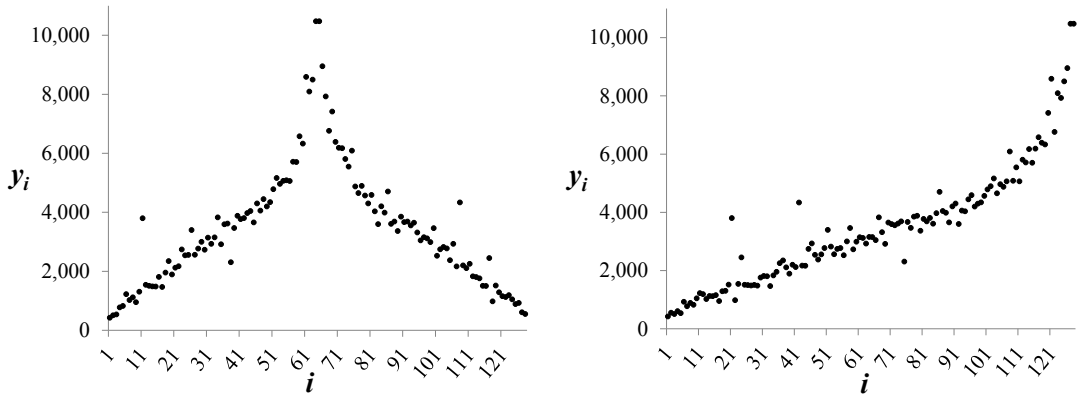
(a) Orden equilibrado (b) Orden creciente

Figura 7.1: Datos simulados de una distribución normal. Población ordenada de acuerdo con y_i, $y_i : i-$ésimo valor generado ($i = 1, \ldots, N$), $N = 108$. Coeficiente de asimetría $\gamma_3 =$ -0.08.

a) Datos simulados de una distribución normal con $N = 108$. En este caso no se cuenta con información auxiliar y la variable de interés y_i corresponde al $i-$ésimo dato generado ($i = 1, 2, \ldots, N$) bajo una distribución normal. En la Figura 7.1 (a) se presenta esta población ordenada de manera equilibrada con respecto a y_i, y en el panel (b) se presentan los datos ordenados de forma creciente. El coeficiente de asimetría de la población, γ_3, fue de -0.08, valor muy cercano a cero que es el valor que toma en una población simétrica.

b) La población correspondiente a $N = 128$ aldeas comprendidas en un *tehsil* (unidad administrativa que comprende aldeas y pueblos) en el estado de Madras, India. La información aparece en el libro de Murthy [39], la cual fue recolectada durante los Censos de 1951 y 1961, en §5.8e y §5.9f esta información se usa para comparar el desempeño de dos estimadores, uno de los cuales es v_{FDI}, así como para comparar la eficiencia de diferentes tipos de muestreo sis-

temático: BSS (equilibrado), CSS (centrado) y LSS (lineal), este último es al que aquí simplemente se ha llamado sistemático. En este ejercicio la variable auxiliar para ordenar las unidades, x_i, fue la población en 1951 en la $i-$ésima aldea y la variable de interés y_i corresponde a la población en 1961 en la aldea i, $i = 1, 2, \ldots, N$. En la Figura 7.2 (a) se muestra la gráfica de este conjunto de datos ordenado de manera equilibrada con respecto a x_i y en (b) ordenado de manera creciente. El coeficiente de asimetría es $\gamma_3 = 1.03$.



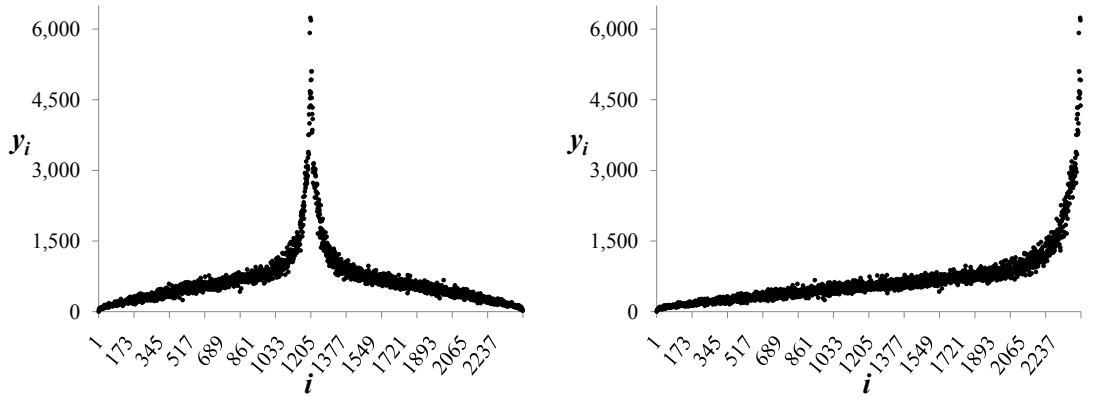
(a) Orden equilibrado

(b) Orden creciente

Figura 7.2: Población en aldeas comprendidas en un *tehsil* en el estado de Madras, India. Información ordenada de acuerdo con x_i, y_i : población en 1960 en la $i-$ésima aldea, x_i : población en 1950 en la $i-$ésima aldea ($i = 1, \ldots, N$), $N = 128$. Coeficiente de asimetría $\gamma_3 = 1.03$.

c) El número de niños entre 5 y 13 años de edad que asisten a la escuela, $N = 2,400$. La información proviene de la muestra censal del XII Censo General de Población y Vivienda 2000 [27] recolectada en México por el Instituto Nacional de Estadística y Geografía (INEGI), misma que se proporciona a nivel de individuo (un poco más de 10 millones de registros) y para fines del ejercicio se agregó a nivel municipal (2,400 municipios). El INEGI proporciona la condición de asistencia o no a la escuela para cada uno de los integrantes de 5 años y más del hogar seleccionado, la variable y_i se construyó contando el número de niños entre 5 y 13 años que asisten a la escuela en el municipio i ($i = 1, 2, \ldots, N$). También con la muestra censal se formó la variable auxiliar x_i que corresponde a la población en el $i-$ésimo municipio, $i = 1, 2, \ldots, N$. Para

fines del ejercicio la información anterior constituye la población, la variable de interés se presenta en la Figura 7.3 (a) ordenada de forma equilibrada de acuerdo con x_i y en (b) ordenada de manera creciente. La población no es simétrica puesto que el coeficiente de asimetría es $\gamma_3 = 3.63$.



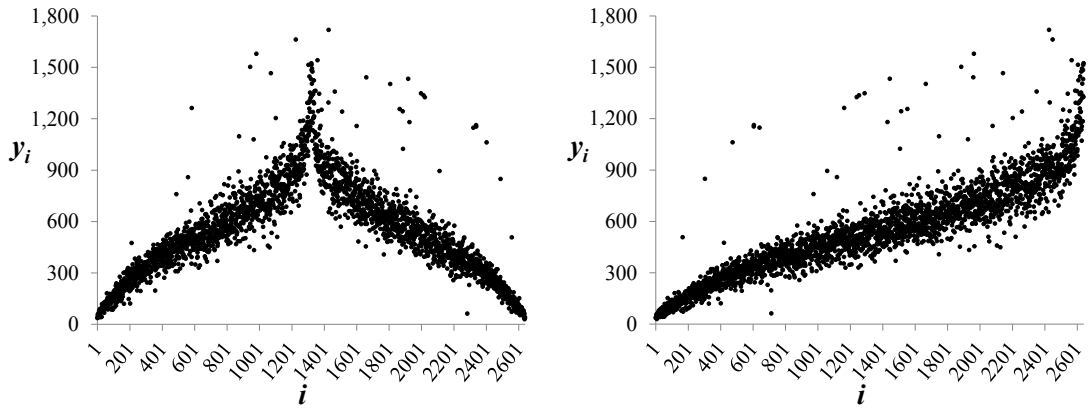
(a) Orden equilibrado

(b) Orden creciente

Figura 7.3: Niños entre 5 y 13 años que asisten a la escuela en el año 2000. Valores ordenados de acuerdo con x_i, y_i : número de niños entre 5 y 13 años que asisten a la escuela en el $i-$ésimo municipio, x_i : población en el municipio i $(i = 1, \ldots, N)$, $N = 2,400$. Coeficiente de asimetría $\gamma_3 = 3.63$.

d) El número de votantes que acudieron a expresar su preferencia electoral en el estado de Michoacán, $N = 2,640$. La información se tomó del Programa de Resultados Preliminares (PREP2000) y corresponde a las elecciones para Presidente de la República Mexicana que se llevaron a cabo en el año 2000. La variable auxiliar x_i que se usó para ordenar la población fue la lista nominal en la sección electoral i y la variable de interés y_i fue el total de votos en la $i-$ésima sección, $i = 1, 2, \ldots, N$. Este total de votos comprende además de los votos obtenidos por cada partido político, los votos por los candidatos no registrados y los votos anulados. La lista nominal en la sección se refiere al número de ciudadanos debidamente registrados en el Padrón Electoral a quienes ya se les ha entregado su credencial para votar y está vigente. En la figura 7.4 (a) se muestra la población ordenada de manera equilibrada por la variable auxiliar x_i $(i = 1, \ldots, N)$, en el panel (b) se presenta la población ordenada de manera creciente. En este ejemplo el coeficiente de asimetría fue

de 0.56.

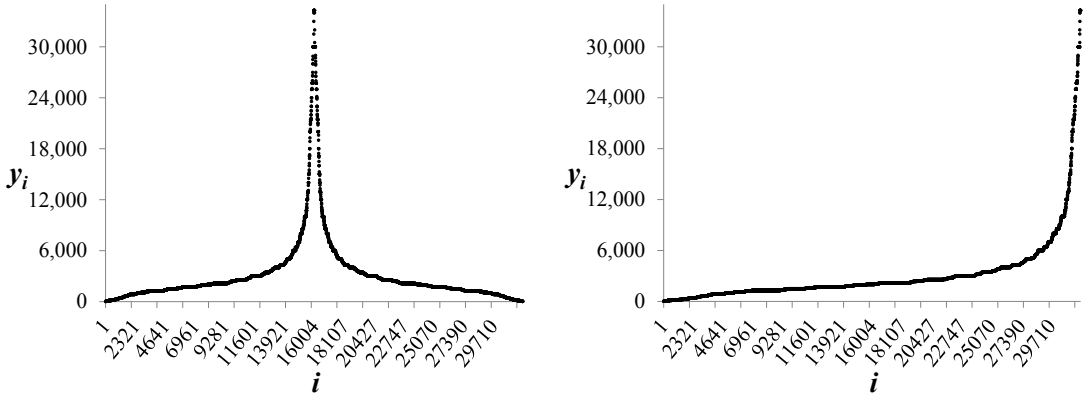


(a) Orden equilibrado

(b) Orden creciente

Figura 7.4: Número de votos en la elección Presidencial del 2000 en el estado de Michoacán. Población ordenada de acuerdo con x_i, y_i : número de votos para Presidente en las elecciones del año 2000 en la $i-$ésima sección, x_i : lista nominal en la $i-$ésima sección $(i = 1, \ldots, N)$, $N = 2,640$. Coeficiente de asimetría $\gamma_3 = 0.56$.

e) El ingreso de personas de 12 años y más en el estado de Aguascalientes, con $N = 32,000$ elementos. La información base también fue recopilada en la muestra censal del XII Censo General de Población y Vivienda 2000 [27]. La variable de interés, y_i, se refiere al ingreso total de la $i-$ésima persona que tiene al menos 12 años de edad, $i = 1, 2, \ldots, N$. Este ingreso considera el ingreso por trabajo, pensión, ayuda de familiares, ayuda de procampo o progresa y otro tipo de ingresos como beca, renta e intereses bancarios. En ese ejercicio no se usó información auxiliar por lo que la población se ordenó según el ingreso del individuo, como se muestra en la Figura 7.5. Entre los ejemplos considerados fue la población más asimétrica con respecto a su media, $\gamma_3 = 4.19$.



(a) Orden equilibrado

(b) Orden creciente

Figura 7.5: Ingreso de personas de 12 años y más en el estado de Aguascalientes. Población ordenada de acuerdo con y_i, y_i : ingreso total de la $i-$ésima persona ($i = 1, \ldots, N$), $N = 32,000$. Coeficiente de asimetría $\gamma_3 = 4.19$.

Con estas cinco poblaciones se calcularon las varianzas bajo BSS, SYC y SI, los resultados se presentan en la sección siguiente.

7.1. Eficiencia relativa

Cuando la población es simétrica respecto a su media y se ordena de manera equilibrada, la varianza del estimador del total bajo muestreo sistemático es cero si N y n son pares y $T = N/n$ es entero, como se demostró en la Proposición 4.1. Adicionalmente, de acuerdo con lo que se mostró en el capítulo 5, para distribuciones simétricas como la uniforme, normal y Laplace, la varianza de muestreo bajo BSS es menor que la del SYC y ésta que la del SI.

En las poblaciones reales la condición de simetría con respecto a la media difícilmente se cumple, como sucede con las cinco poblaciones mencionadas al inicio del capítulo, por ello en esta sección se evalua la eficiencia del muestreo sistemático usando esos datos.

Para ello se calcularon la varianza del muestreo sistemático cuando la población se ordena de manera equilibrada, la del sistemático bajo un orden creciente y la del muestreo aleatorio simple, denotadas por $V_{BSS}\left(\hat{t}_\pi\right)$, $V_{SYC}\left(\hat{t}_\pi\right)$ y $V_{SI}\left(\hat{t}_\pi\right)$, re-

spectivamente, las primeras dos están dadas por la expresión (3.7) y la tercera por $V_{SI}\left(\hat{t}_\pi\right) = (1 - n/N)s_U^2/n$.

La comparación entre los diseños se hizo usando tres cocientes de varianzas como indicador de la eficiencia:

i) La eficiencia relativa del muestreo sistemático equilibrado con respecto al aleatorio simple, $\frac{V_{SI}\left(\hat{t}_\pi\right)}{V_{BSS}\left(\hat{t}_\pi\right)}$.

ii) La eficiencia del sistemático cuando la población se ordena de manera creciente en relación con el aleatorio simple, $\frac{V_{SI}\left(\hat{t}_\pi\right)}{V_{SYC}\left(\hat{t}_\pi\right)}$.

iii) La eficiencia del sistemático equilibrado, respecto al sistemático bajo un orden creciente, $\frac{V_{SYC}\left(\hat{t}_\pi\right)}{V_{BSS}\left(\hat{t}_\pi\right)}$.

Los resultados se muestran en los cuadros 7.1 a 7.5 para la mayoría de los tamaños de muestra tales que $T = N/n \in \mathbb{N}$, no se incluyen algunos tamaños de muestra porque no aportan información adicional. Lo que se observa en éstos se puede resumir como sigue.

- En los cinco ejercicios que se realizaron, el muestreo sistemático bajo un orden equilibrado fue más eficiente que el muestreo aleatorio simple. Esto se cumplió inclusive para las variables que se alejan de la simetría como el ingreso. La mínima eficiencia relativa del BSS respecto al SI fue de 1.8; para los casos en que $n \geq 10$ la mínima eficiencia fue de 3.9.

- En las cinco poblaciones, cuando cada una de estas se ordenó de manera creciente, el muestreo sistemático también fue más eficiente que el aleatorio simple. La menor eficiencia del SYC respecto al SI fue de 1.2 y 2.3 considerando $n \geq 10$.

- En cuatro de los conjuntos de datos analizados, la varianza del estimador para una muestra sistemática fue menor cuando la población se ordenó de manera equilibrada que cuando se ordenó de manera creciente. La excepción fue la población referente al número de votos, donde el BSS resultó más eficiente que el SYC cuando $n < 40$, si $n \geq 40$ el resultado se invierte siendo más eficiente el SYC, con excepción de $n = 264$ y $n = 528$.

Cuadro 7.1: Eficiencia relativa. Valores simulados de una distribución normal, $N = 108$

Tamaño de muestra	Eficiencia relativa $\frac{V_{SI}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$	$\frac{V_{SI}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$	$\frac{V_{SYC}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$
2	101.5	1.4	71.0
4	93.4	2.5	37.2
6	139.4	2.9	48.9
12	156.0	6.5	24.1
18	261.8	7.4	35.6
36	144.9	10.1	14.3
54	212.4	17.3	12.3

Cuadro 7.2: Eficiencia relativa. Población (1961) en aldeas en un *tehsil*, $N = 128$

Tamaño de muestra	Eficiencia relativa $\frac{V_{SI}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$	$\frac{V_{SI}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$	$\frac{V_{SYC}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$
2	5.7	1.3	4.2
4	7.5	2.1	3.5
8	20.3	3.8	5.4
16	69.7	5.9	11.8
32	209.8	17.0	12.3
64	108,633.5	9.6	11,316.8

- En el caso de los datos simulados de la distribución normal, el de la población en aldeas que pertenecen a un *tehsil* y el ingreso en el estado de Aguascalientes, la eficiencia del SYC respecto al SI aumenta conforme aumenta el tamaño de muestra, sólo dos tamaños de muestra no siguieron este patrón: $n = 64$ en el ejemplo de la población y $n = 800$ en el del ingreso. Sólo en el ejemplo de la población en el *tehsil* la eficiencia del BSS respecto al SI mostró el patrón anterior.

- En la población generada a partir de la distribución normal la eficiencia del BSS respecto al SYC resultó mayor a 93 para todos los tamaños de muestra,

situación que se esperaba puesto que los datos se generaron de una distribución simétrica, γ_3 = -0.08. El hecho de que la población se ordenara con respecto a la variable de interés también contribuyó a la eficiencia del sistemático.

- En los datos sobre el ingreso, la eficiencia del BSS (ó SYC) respecto al SI es mayor a 146 para tamaños de muestra grandes, $n \geq 200$ en el caso del BSS y $n \geq 1,600$ en el caso del SYC. Este resultado sorprende un poco debido a que el coeficiente de asimetría de esta población fue el más alto, entre las cinco que se analizaron; pero hay que recordar que la población se ordenó respecto a la variable de interés.

Cuadro 7.3: Eficiencia relativa. Niños entre 5 y 13 años que asisten a la escuela, $N = 2,400$

Tamaño de muestra	Eficiencia relativa $\frac{V_{SI}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$	$\frac{V_{SI}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$	$\frac{V_{SYC}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$
2	1.9	1.2	1.7
4	2.2	1.4	1.6
6	2.8	1.7	1.7
8	3.4	2.0	1.7
10	4.1	2.3	1.8
20	8.2	3.7	2.2
40	17.7	5.4	3.3
60	21.0	7.3	2.9
80	19.9	8.4	2.4
100	27.9	12.8	2.2
160	24.3	14.3	1.7
200	28.0	19.8	1.4
300	57.5	25.8	2.2
400	24.0	19.5	1.2
480	21.5	18.5	1.2
600	40.9	105.5	0.4
800	11.1	9.2	1.2
1200	52.5	136.4	0.4

- Llama la atención que la eficiencia del BSS respecto al SI, en el ejemplo de la

población en el tehsil cuando $n = 64$, es altísima (mayor a 100,000), lo mismo sucede en el ejemplo del ingreso cuando $n = 8000, 16000$; la posible explicación a esto es que el total casi se estima de manera exacta bajo BSS, por lo que $V_{BSS}\left(\hat{t}_\pi\right)$ es muy pequeña, menor a un dígito en ambos casos.

De acuerdo con las observaciones anteriores, el diseño con menor varianza fue el BSS siguiéndole el SYC, salvo algunos tamaños de muestra; en todos los casos el diseño de mayor varianza fue el SI. Ahora surge la interrogante ¿cómo estimar la varianza del BSS o del SYC puesto que no existe un estimador insesgado de ella? En la siguiente sección se comparan la aproximación propuesta y tres tomadas de la literatura, para los cinco conjuntos de datos.

Cuadro 7.4: Eficiencia relativa. Número de votos en el estado de Michoacán, $N = 2,640$

Tamaño de muestra	Eficiencia relativa		
	$\frac{V_{SI}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$	$\frac{V_{SI}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$	$\frac{V_{SYC}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$
2	6.4	1.4	4.4
4	6.3	1.9	3.3
6	6.8	2.6	2.7
10	7.0	3.5	2.0
20	7.5	4.8	1.6
30	7.4	5.5	1.3
40	6.1	6.3	1.0
60	6.5	9.3	0.7
80	6.4	7.1	0.9
110	7.6	9.9	0.8
176	4.7	4.8	1.0
220	5.3	19.1	0.3
264	8.9	5.1	1.8
330	8.7	14.3	0.6
440	3.1	21.2	0.1
528	11.4	2.7	4.2
660	4.4	12.1	0.4
880	2.9	10.1	0.3
1320	2.3	471.7	<0.1

Cuadro 7.5: Eficiencia relativa. Ingreso en el estado de Aguascalientes, $N = 32,000$

Tamaño de muestra	Eficiencia relativa		
	$\frac{V_{SI}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$	$\frac{V_{SI}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$	$\frac{V_{SYC}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$
2	1.8	1.2	1.6
4	2.4	1.4	1.6
8	3.4	2.0	1.8
10	3.9	2.3	1.7
20	6.6	3.2	2.1
40	12.9	4.8	2.7
64	21.0	6.4	3.3
80	33.0	7.9	4.2
100	52.3	9.1	5.7
200	152.8	17.9	8.5
320	216.9	29.6	7.3
400	2,953.3	35.5	83.2
500	1,209.0	44.3	27.3
640	669.0	58.6	11.4
800	3,328.0	54.3	61.3
1000	2,740.5	80.7	34.0
1600	1,921.1	146.1	13.1
2000	3,357.5	199.4	16.8
3200	29,970.3	261.1	114.8
4000	8,634.2	311.6	27.7
6400	28,839.9	553.0	52.2
8000	113,960.1	593.8	191.9
16000	258,481.4	962.3	268.6

7.2. Varianza estimada

En esta sección se compara la varianza exacta del estimador bajo muestreo sistemático, $V_{BSS}\left(\hat{t}_\pi\right)$ ó $V_{SYC}\left(\hat{t}_\pi\right)$ dependiendo del orden de la población, con los tres estimadores de la varianza señalados en el capítulo 6, para cada uno de las cinco universos.

Específicamente para cada población, estimador y tamaño de muestra se cal-

culó el promedio de las varianzas estimadas como sigue,

$$\overline{v}_{\alpha}\left(\hat{t}_{\pi}\right) = \frac{\sum_{r=1}^{T} v_{\alpha}\left(\hat{t}_{s_r}\right)}{T}. \tag{7.1}$$

donde α es igual a

- FDI, diferencia de orden uno entre los datos en la muestra
- SDI, diferencias de orden dos entre los datos en la muestra
- COV, covariograma.

La población se ordena de forma equilibrada o creciente según la variable auxiliar x_i, o bien la variable de interés y_i, $i = 1, \ldots, N$, según se indicó en la primera sección de este capítulo.

En la descripción de los resultados también se hará referencia al sesgo del estimador, esto es,

$\overline{v}_{\alpha}\left(\hat{t}_{\pi}\right) - V_{BSS}\left(\hat{t}_{\pi}\right)$ en el caso del BSS y

$\overline{v}_{\alpha}\left(\hat{t}_{\pi}\right) - V_{SYC}\left(\hat{t}_{\pi}\right)$ en el caso del SYC.

Es importante mencionar que todas las variables de interés se estandarizaron de tal manera que tienen media cero y varianza unitaria, esto es, $\overline{y}_U = \sum_{i=1}^{N} y_i/N = 0$ y $s_U^2 = \sum_{i=1}^{N} \left(y_i - \overline{y}_U\right)^2/(N-1) = 1$. Los resultados se muestran en las Figuras 7.6 a 7.10 y una descripción de ellos se da a continuación.

- En la Figura 7.6 (a) se presentan los resultados para las 108 observaciones generadas de la distribución normal ordenados de manera equilibrada. Se observa que para la mayoría de los tamaños de muestra, en promedio, el estimador $v_{SDI}\left(\hat{t}_{s_r}\right)$ es el más cercano al valor exacto $V_{BSS}\left(\hat{t}_{\pi}\right)$. En promedio, los tres estimadores sobreestiman la varianza exacta.

  En la Figura (b) se observa que cuando la población se ordena de manera creciente, $\overline{v}_{FDI}\left(\hat{t}_{\pi}\right)$ y $\overline{v}_{COV}\left(\hat{t}_{\pi}\right)$ son los más próximos a la varianza exacta. Con excepción de $n = 2$, en promedio, los estimadores subestiman la varianza del SYC.

- En la Figura 7.7 (a) y (b) se grafican los cálculos referentes a la población en 1961 en 128 villas comprendidas en un *tehsil*, según el orden. En promedio, el estimador con menor sesgo corresponde a $v_{COV}\left(\hat{t}_{s_r}\right)$; cuando $n = 64$ el sesgo de los tres estimadores es grande, en este caso $V_{BSS}\left(\hat{t}_{\pi}\right)$ es muy cercana a cero. Como puede verse en el panel (a) los tres estimadores sobreestiman la varianza del BSS.

  En el caso del SYC (ver panel (b)), los estimadores $v_{FDI}\left(\hat{t}_{s_r}\right)$ y $v_{COV}\left(\hat{t}_{s_r}\right)$ son los de menor sesgo. Las tres la subestiman para la mayoría de los tamaños de muestra.

- Como puede apreciarse en la Figura 7.8 (a), que se basa en el número de niños que asiste a la escuela, $\overline{v}_{COV}\left(\hat{t}_{\pi}\right)$ es el más cercano a la varianza exacta si $n < 200$, cuando $n \geq 200$, $\overline{v}_{FDI}\left(\hat{t}_{\pi}\right)$ y $\overline{v}_{SDI}\left(\hat{t}_{\pi}\right)$ son los más próximos. Con excepción de algunos tamaños de muestra, $v_{FDI}\left(\hat{t}_{s_r}\right)$ y $v_{SDI}\left(\hat{t}_{s_r}\right)$ sobreestiman la varianza del BSS.

  En el panel (b) de la misma figura se observa que $v_{FDI}\left(\hat{t}_{s_r}\right)$ es el de menor sesgo. En promedio, las tres aproximaciones subestiman la varianza del SYC para casi todos los tamaños de muestra.

- En la Figura 7.9 (a) se presenta el promedio de las varianzas usando la información sobre el número de votos, en esta se observa que $v_{FDI}\left(\hat{t}_{s_r}\right)$ y $v_{SDI}\left(\hat{t}_{s_r}\right)$ son los de menor sesgo. Estos dos estimadores sobreestiman la varianza del BSS, con excepción de algunos tamaños de muestra.

  Respecto al SYC (ver panel (b)), si $n \leq 176$ los estimadores $v_{FDI}\left(\hat{t}_{s_r}\right)$ y $v_{SDI}\left(\hat{t}_{s_r}\right)$ son los de menor sesgo, mientras que si $n \geq 220$ en promedio $v_{COV}\left(\hat{t}_{s_r}\right)$ es el más cercano. Este último subestima la varianza, con excepción $n = 1320$; $v_{FDI}\left(\hat{t}_{s_r}\right)$ y $v_{SDI}\left(\hat{t}_{s_r}\right)$ la subestiman si $n \leq 40$ y la sobreestiman para casi todo $n \geq 60$.

- En la Figura 7.10 se presentan los cálculos para los datos referentes al ingreso. Cuando la población se ordena equilibradamente (ver panel (a)) $\overline{v}_{COV}\left(\hat{t}_{\pi}\right)$ es el más cercano a la varianza del BSS. Para la mayoría de los tamaños de muestra, las tres aproximaciones son mayores a la varianza exacta.

  El promedio de las estimaciones cuando la población se ordena de manera creciente se muestran en la gráfica (b), como puede observarse, $\overline{v}_{COV}\left(\hat{t}_{\pi}\right)$ es

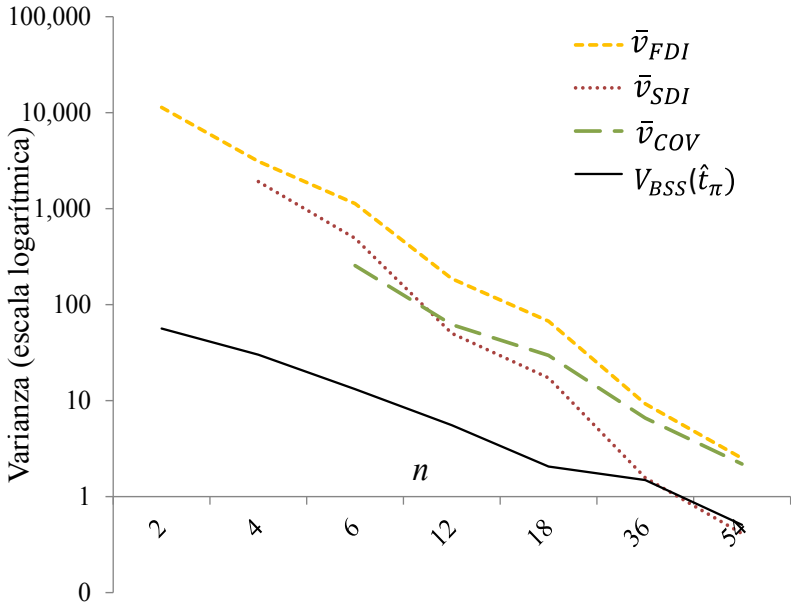
el más cercano a la varianza del SYC. Los tres estimadores subestiman la varianza en la mayoría de los casos.

Como ya lo han señalado diversos autores, no existe un mejor estimador para todas las situaciones, sin embargo, los estimadores $v_{SDI}\left(\hat{t}_\pi\right)$ y $v_{COV}\left(\hat{t}_\pi\right)$ tuvieron menor sesgo cuando la población se ordena de manera equilibrada. Si la población se ordena en forma creciente, $v_{FDI}\left(\hat{t}_\pi\right)$ y $v_{COV}\left(\hat{t}_\pi\right)$ son los que tuvieron un menor sesgo.

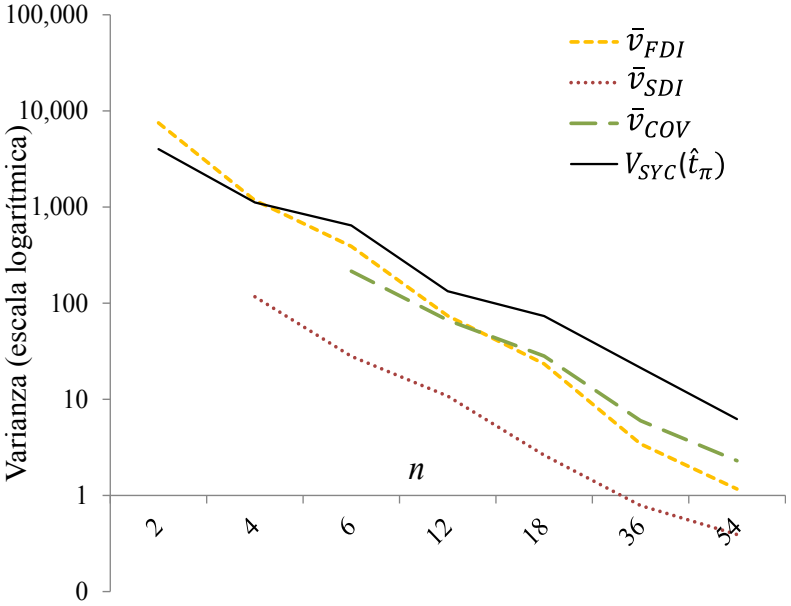
Si se desea una estimación conservadora de la varianza, en el caso del BSS los tres estimadores sobreestiman la varianza en tres poblaciones, en las otras dos la sobreestiman $v_{FDI}\left(\hat{t}_\pi\right)$ y $v_{SDI}\left(\hat{t}_\pi\right)$. En el caso del SYC, las tres aproximaciones subestiman la varianza en la mayoría de las situaciones.

La elección del estimador debe hacerse considerando las propiedades de la población muestreada, es recomendable usar algún estimador como los aquí señalados en lugar del que supone un muestreo aleatorio simple, pues como puede notarse en la eficiencia relativa, la varianza de este último es mayor que la varianza del muestreo sistemático.

Se obtuvieron resultados análogos usando los sesgos relativos $\frac{\overline{v}_\alpha\left(\hat{t}_\pi\right)-V_{BSS}\left(\hat{t}_\pi\right)}{V_{BSS}\left(\hat{t}_\pi\right)}$ y $\frac{\overline{v}_\alpha\left(\hat{t}_\pi\right)-V_{SYC}\left(\hat{t}_\pi\right)}{V_{SYC}\left(\hat{t}_\pi\right)}$, los cuales se presentan en los cuadros 7.6 a 7.10. También puede observarse que en algunos casos el sesgo relativo crece conforme el tamaño de muestra aumenta, esto se puede explicar porque cada estimador supone que la variable de interés tiene cierta estructura, si esta variable se aleja de los supuestos, el sesgo puede aumentar. Para mayor referencia al respecto puede consultarse Wolter [58, 59], en donde se proporciona el sesgo relativo esperado de ocho diferentes estimadores suponiendo cuatro modelos de población.

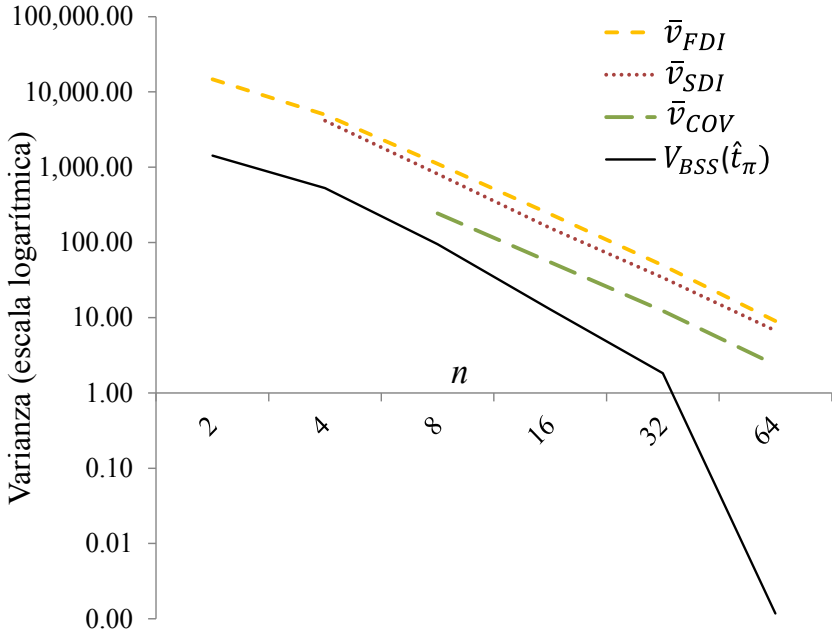


(a) Orden equilibrado

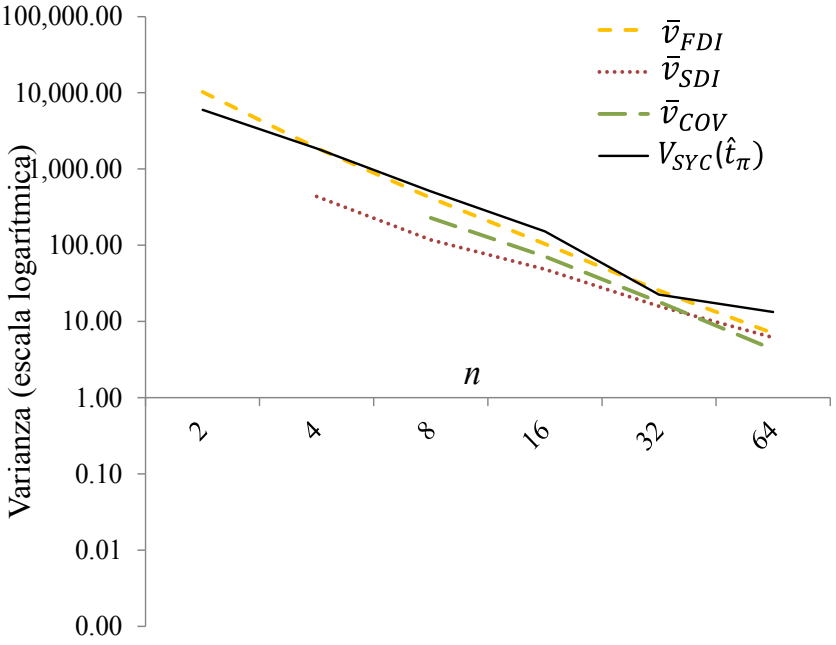


(b) Orden creciente

Figura 7.6: Varianzas del estimador del total bajo muestreo sistemático, calculadas con $N = 108$ observaciones generadas de una distribución normal. Cada línea punteada representa el promedio $\overline{v}_\alpha = \frac{\sum_{r=1}^{T} v_\alpha(\hat{t}_{s_r})}{T}$, según el tamaño de muestra; se presentan líneas en lugar de puntos para una mejor visualización.

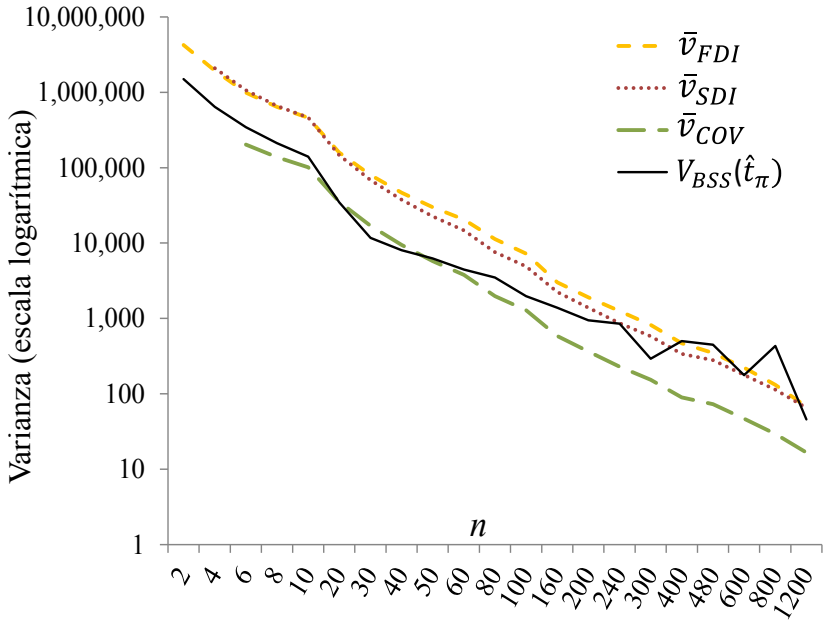


(a) Orden equilibrado

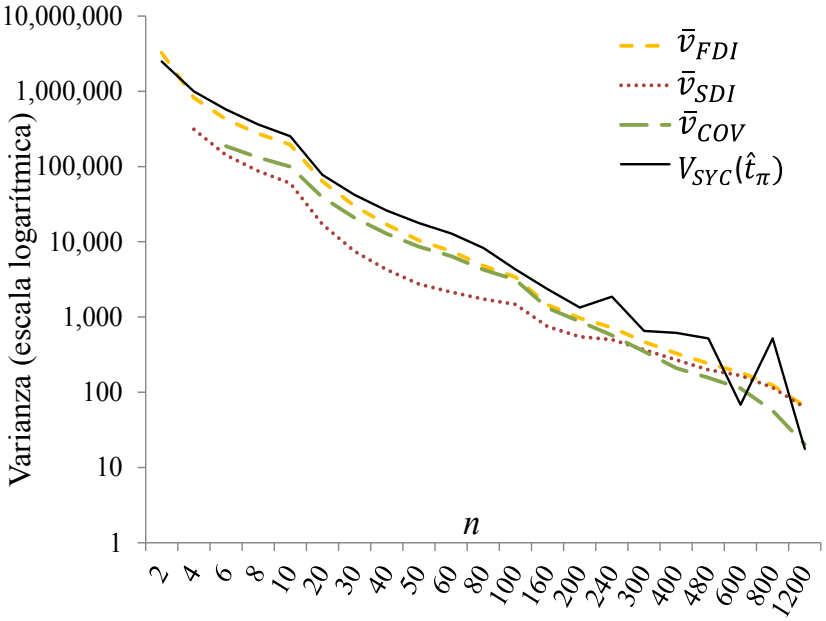


(b) Orden creciente

Figura 7.7: Varianzas del estimador del total bajo muestreo sistemático, calculadas con la población de $N = 128$ aldeas comprendidas en un *tehsil* en el estado de Madras, India. Cada línea punteada representa el promedio $\overline{v}_\alpha = \frac{\sum_{r=1}^{T} v_\alpha(\hat{t}_{sr})}{T}$, según el tamaño de muestra; se presentan líneas en lugar de puntos para una mejor visualización.

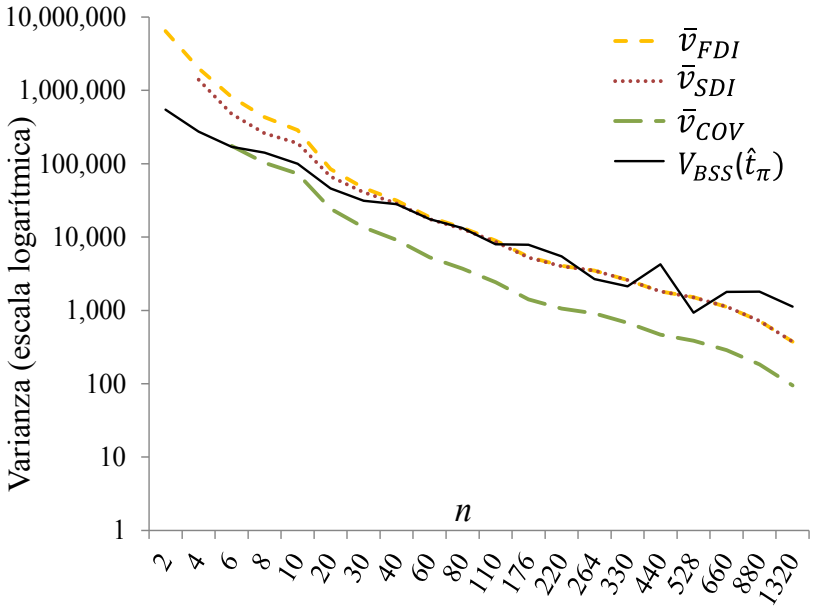


(a) Orden equilibrado

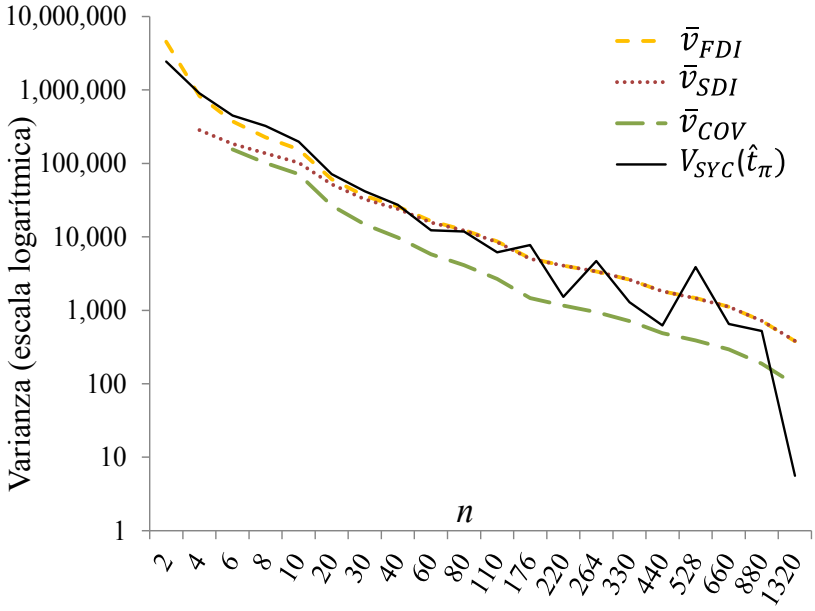


(b) Orden creciente

Figura 7.8: Varianzas del estimador del total bajo muestreo sistemático, calculadas con el número de niños entre 5 y 13 años que asisten a la escuela en $N = 2,400$ municipios. Cada línea punteada representa el promedio $\overline{v}_\alpha = \frac{\sum_{r=1}^{T} v_\alpha(\hat{t}_{s_r})}{T}$, según el tamaño de muestra; se presentan líneas en lugar de puntos para una mejor visualización.

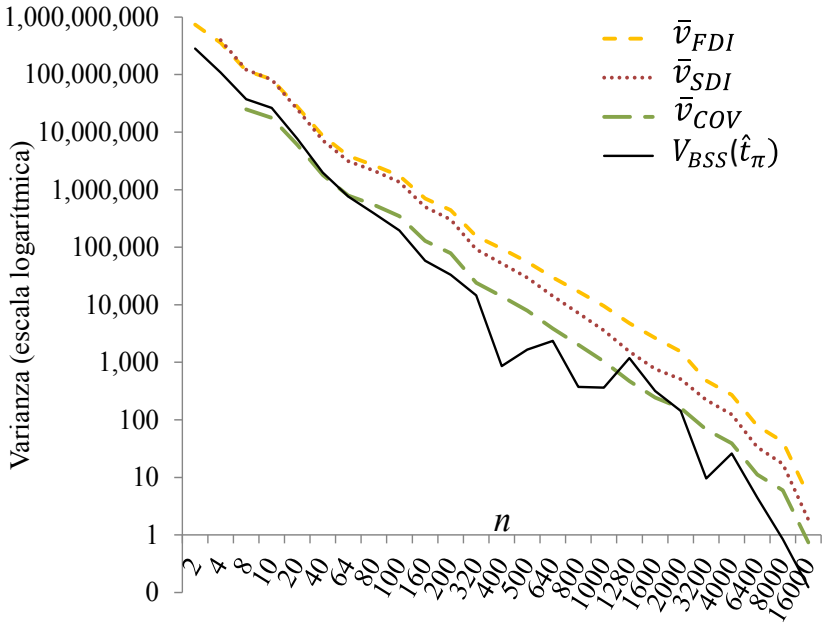


(a) Orden equilibrado

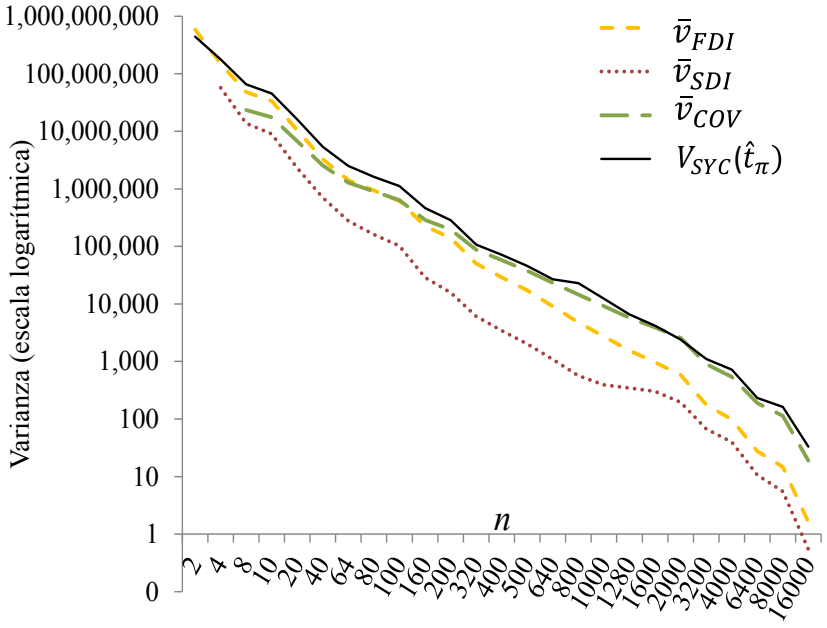


(b) Orden creciente

Figura 7.9: Varianzas del estimador del total bajo muestreo sistemático, calculadas con el número de votantes que acudieron a votar en $N = 2,640$ secciones del estado de Michoacán. Cada línea punteada representa el promedio $\overline{v}_\alpha = \frac{\sum_{r=1}^{T} v_\alpha(\hat{t}_{s_r})}{T}$, según el tamaño de muestra; se presentan líneas en lugar de puntos para una mejor visualización.



(a) Orden equilibrado



(b) Orden creciente

Figura 7.10: Varianzas del estimador del total bajo muestreo sistemático, calculadas con el ingreso de $N = 32,000$ personas de 12 años y más en el estado de Aguascalientes. Cada línea punteada representa el promedio $\overline{v}_\alpha = \frac{\sum_{r=1}^{T} v_\alpha(\hat{t}_{s_r})}{T}$, según el tamaño de muestra; se presentan líneas en lugar de puntos para una mejor visualización.

Cuadro 7.6: Sesgo relativo. Valores simulados de una distribución normal, $N = 108$

Tamaño de	Orden equilibrado [a/]			Orden creciente [b/]		
muestra	FDI	SDI	COV	FDI	SDI	COV
2	201.0			0.9		
4	102.7	62.8		0.0	-0.9	
6	85.0	36.5	18.4	-0.4	-1.0	-0.7
12	32.8	8.1	10.2	-0.4	-0.9	-0.5
18	31.7	7.4	13.4	-0.7	-1.0	-0.6
36	5.2	0.1	3.4	-0.8	-1.0	-0.7
54	3.9	-0.2	3.3	-0.8	-0.9	-0.6

Sesgo relativo: [a/] $\frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{BSS}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$, α : FDI, SDI, COV
[b/] $\frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{SYC}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$.

Cuadro 7.7: Sesgo relativo. Población (1961) en aldeas en un *tehsil*, $N = 128$

Tamaño de	Orden equilibrado [a/]			Orden creciente [b/]		
muestra	FDI	SDI	COV	FDI	SDI	COV
2	9.3			0.7		
4	8.4	6.9		0.0	-0.8	
8	10.7	7.6	1.6	-0.2	-0.8	-0.6
16	17.5	11.0	3.2	-0.3	-0.7	-0.5
32	26.1	17.6	5.7	0.1	-0.3	-0.2
64	7691.3	5719.3	2010.0	-0.5	-0.5	-0.7

Sesgo relativo: [a/] $\frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{BSS}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$, α : FDI, SDI, COV
[b/] $\frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{SYC}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$.

Cuadro 7.8: Sesgo relativo. Niños 5 - 13 años que asisten a la escuela, $N = 2,400$

Tamaño de	Orden equilibrado [a/]			Orden creciente [b/]		
muestra	FDI	SDI	COV	FDI	SDI	COV
2	1.8			0.3		
4	2.0	2.3		-0.2	-0.7	
6	1.9	2.1	-0.4	-0.3	-0.8	-0.7
8	2.0	2.1	-0.4	-0.2	-0.8	-0.6
10	2.3	2.3	-0.3	-0.2	-0.8	-0.6
20	3.6	3.2	0.0	-0.2	-0.8	-0.5
30	5.8	4.9	0.4	-0.3	-0.8	-0.5
40	4.8	3.7	0.2	-0.4	-0.8	-0.5
50	3.8	2.6	-0.1	-0.4	-0.8	-0.5
60	3.6	2.3	-0.1	-0.4	-0.8	-0.5
80	2.2	1.2	-0.4	-0.4	-0.8	-0.5
100	2.7	1.5	-0.3	-0.2	-0.7	-0.3
160	1.2	0.6	-0.6	-0.4	-0.7	-0.4
200	1.0	0.5	-0.6	-0.3	-0.6	-0.3
240	0.5	0.0	-0.7	-0.6	-0.7	-0.7
300	1.8	1.0	-0.5	-0.3	-0.4	-0.5
400	-0.1	-0.3	-0.8	-0.5	-0.6	-0.7
480	-0.2	-0.4	-0.8	-0.5	-0.6	-0.7
600	0.3	0.0	-0.7	1.7	1.4	0.7
800	-0.7	-0.7	-0.9	-0.8	-0.8	-0.9
1200	0.5	0.4	-0.6	2.7	2.5	0.2

Sesgo relativo: [a/] $\frac{\overline{v}_\alpha(\hat{t}_\pi) - V_{BSS}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$, α : FDI, SDI, COV

[b/] $\frac{\overline{v}_\alpha(\hat{t}_\pi) - V_{SYC}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$.

Cuadro 7.9: Sesgo relativo. Número de votos en el estado de Michoacán, $N = 2,640$

Tamaño de	Orden equilibrado [a/]			Orden creciente [b/]		
muestra	FDI	SDI	COV	FDI	SDI	COV
2	10.7			0.9		
4	6.3	4.1		-0.1	-0.7	
6	3.8	1.8	0.0	-0.2	-0.6	-0.7
8	2.0	0.8	-0.3	-0.3	-0.6	-0.7
10	1.9	0.9	-0.3	-0.2	-0.5	-0.6
20	0.8	0.4	-0.5	-0.1	-0.3	-0.6
30	0.5	0.3	-0.6	-0.1	-0.2	-0.6
40	0.1	0.0	-0.7	-0.1	-0.1	-0.6
60	0.1	0.0	-0.7	0.3	0.3	-0.5
80	0.0	0.0	-0.7	0.0	0.0	-0.7
110	0.1	0.1	-0.7	0.4	0.4	-0.6
176	-0.3	-0.3	-0.8	-0.3	-0.4	-0.8
220	-0.3	-0.3	-0.8	1.7	1.7	-0.2
264	0.3	0.3	-0.7	-0.3	-0.3	-0.8
330	0.2	0.2	-0.7	1.0	1.0	-0.4
440	-0.6	-0.6	-0.9	1.9	1.9	-0.2
528	0.6	0.6	-0.6	-0.6	-0.6	-0.9
660	-0.4	-0.4	-0.8	0.7	0.7	-0.5
880	-0.6	-0.6	-0.9	0.4	0.4	-0.6
1320	-0.7	-0.7	-0.9	66.9	67.5	16.5

Sesgo relativo: $^{a/}\ \frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{BSS}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$, α : FDI, SDI, COV

$^{b/}\ \frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{SYC}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$.

Cuadro 7.10: Sesgo relativo. Ingreso en el estado de Aguascalientes, $N = 32,000$

Tamaño de	Orden equilibrado [a/]			Orden creciente [b/]		
muestra	FDI	SDI	COV	FDI	SDI	COV
2	1.6			0.3		
4	2.3	2.7		-0.2	-0.7	
8	2.2	2.3	-0.3	-0.3	-0.8	-0.6
10	2.2	2.1	-0.3	-0.3	-0.8	-0.6
20	2.6	2.3	-0.2	-0.4	-0.9	-0.6
40	3.4	2.6	-0.1	-0.4	-0.9	-0.5
64	4.1	3.1	0.0	-0.4	-0.9	-0.5
80	5.8	4.5	0.4	-0.4	-0.9	-0.4
100	7.9	6.0	0.8	-0.5	-0.9	-0.4
160	11.1	7.6	1.2	-0.5	-0.9	-0.4
200	12.3	8.1	1.4	-0.5	-0.9	-0.3
320	9.6	5.3	0.6	-0.5	-0.9	-0.2
400	108.9	60.4	15.3	-0.6	-1.0	-0.2
500	32.4	16.7	3.7	-0.6	-1.0	-0.2
640	11.7	5.0	0.7	-0.7	-1.0	-0.1
800	44.3	18.2	4.4	-0.8	-1.0	-0.4
1000	25.3	8.9	1.8	-0.8	-1.0	-0.3
1280	3.1	0.3	-0.6	-0.8	-0.9	-0.1
1600	7.3	1.4	-0.2	-0.8	-0.9	-0.1
2000	9.7	2.6	0.1	-0.8	-0.9	0.1
3200	48.9	22.1	6.1	-0.8	-0.9	-0.2
4000	9.5	3.8	0.5	-0.9	-0.9	-0.3
6400	16.7	6.6	1.5	-0.9	-1.0	-0.2
8000	48.9	19.4	6.1	-0.9	-1.0	-0.3
16000	38.4	13.9	5.0	-1.0	-1.0	-0.4

Sesgo relativo: $^{a/}\ \frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{BSS}(\hat{t}_\pi)}{V_{BSS}(\hat{t}_\pi)}$, α : FDI, SDI, COV

$^{b/}\ \frac{\overline{v}_\alpha(\hat{t}_\pi)-V_{SYC}(\hat{t}_\pi)}{V_{SYC}(\hat{t}_\pi)}$.

Capítulo 8

Conclusiones y trabajo futuro

Los principales resultados de investigación de este trabajo se refieren a condiciones bajo las cuales el muestreo sistemático es más eficiente que el muestreo aleatorio simple. Éstos pueden resumirse como sigue:

- En el caso de poblaciones fijas, finitas o continuas, se encontraron tres condiciones suficientes bajo las cuales la varianza del muestreo sistemático es cero. Cuando la población es finita estas condiciones son i) N y n sean pares y $T = N/n \in \mathbb{N}$, ii) la población se ordene de manera equilibrada y iii) los valores ordenados de la variable de interés sean simétricos respecto a la media poblacional. Cuando se trata de una función acotada en $[a, 2b - a]$, si i) n es par y $T = 2(b-a)/n$, ii) la función es simétrica respecto a la recta $x = b$ y iii) es simétrica con respecto al punto $\left(x_0 = \frac{a+b}{2}, f(x_0)\right)$ para $x \in [a, b]$, entonces el área bajo $f(x)$ se estima sin error. Es importante subrayar que las condiciones señaladas son suficientes pero no son necesarias, esto es, puede ser que no se cumplan y que la varianza sea cero.

- En el caso de superpoblaciones, suponiendo que la población sigue una distribución uniforme, se derivó la expresión de la varianza del predictor del total, bajo el modelo y el diseño, considerando tres diseños: muestreo sistemático bajo un orden equilibrado de la población (BSS), muestreo sistemático bajo un orden creciente de la población (SYC) y muestreo aleatorio simple (SI); se demostró que la varianza del muestreo sistemático es menor cuando la población se ordena de manera equilibrada que en forma creciente, y ésta a su vez es menor que la del muestreo aleatorio simple. Mediante simulaciones esto mismo

se mostró para otros dos modelos: la distribución de Laplace y la distribución normal. También se encontraron aproximaciones a la varianza del predictor del total suponiendo estas distribuciones.

- Suponiendo la distribución normal generalizada, cuya función de densidad es simétrica y que tiene como casos particulares las distribuciones uniforme, normal y Laplace, mediante simulaciones se mostró que entre los valores que se consideraron del parámetro de forma, ω, la varianza del modelo de Laplace fue la mayor y la varianza de la uniforme la menor. Al comparar entre los tres diseños, el más eficiente corresponde al muestreo sistemático bajo el orden equilibrado, le sigue el sistemático bajo un orden creciente y por último el muestreo aleatorio simple.

Estos resultados son teóricos puesto que involucran el conocimiento de la variable de interés para toda la población o de la distribución que tiene la misma. No obstante, a partir de ellos, se puede sugerir que el BSS es más eficiente que el SYC y éste que el SI si la población es simétrica.

Entonces, ¿qué sucede si la variable de interés no es simétrica con respecto a su media, o bien, si la población se ordena usando información auxiliar? En este trabajo se compararon las varianzas del BSS, SYC y SI usando cinco conjuntos de datos, que son diferentes en cuanto al número de elementos, naturaleza de la información y disponibilidad de información auxiliar. Aunque las cinco poblaciones tienen diferente grado de simetría, medida por el coeficiente de asimetría de la variable de interés, en cuatro de ellas la varianza del sistemático fue menor cuando la población se ordenó de manera equilibrada que cuando se ordenó de manera creciente. En los cinco ejercicios ambos diseños fueron más eficientes que el aleatorio simple; la menor eficiencia del BSS con respecto al SI fue de 1.8 y la menor del SYC con respecto al SI fue de 1.2.

Análogamente, en el caso de la distribución normal generalizada, para los valores de ω y de n que se consideraron, la menor eficiencia del BSS con respecto al SI fue de 2 y la menor eficiencia del SYC con respecto al SI fue de 1.2.

En la práctica, lo anterior lleva a sugerir que si el tamaño de muestra n y el de población N son pares, se puede tener un intervalo muestral $T = N/n$ entero y se dispone de una variable auxiliar para ordenar la población, entonces se prefiera

muestrear sistemáticamente a la población ordenada de manera equilibrada, aunque no se garantiza que esto funcione en todas las poblaciones.

Un problema práctico del uso del muestreo sistemático es que no existe un estimador insesgado de la varianza. De forma exploratoria, en este trabajo se compararon tres aproximaciones existentes en la literatura, usando cinco conjuntos de datos, no se encontró un mejor estimador en términos del sesgo.

En los cinco ejercicios, así como en las superpoblaciones analizadas, la varianza del muestreo aleatorio simple es mayor que la del sistemático, en consecuencia, usar la estimación de la varianza del aleatorio simple como aproximación de la del sistemático puede llevar a una sobreestimación, por lo que se recomienda elegir una aproximación de las existantes usando el conocimiento que se tiene sobre la estructura de la población. Sin embargo, no se debe de olvidar que existen poblaciones como la del Ejemplo 3.2.1, ordenada de manera creciente dentro de bloques, en las cuales la varianza del muestreo sistemático es mayor a la del aleatorio simple.

Se considera que los hallazgos de este trabajo pudieran servir de partida para investigaciones posteriores, como las siguientes:

1. Demostrar si para cualquier superpoblación cuya función de densidad sea simétrica, se cumple que el BSS es más eficiente que el SYC y este más eficiente que el SI. Esto se demostró para la distribución uniforme y mediante simulaciones se mostró para la normal y Laplace.

2. Analizar el desempeño del muestreo sistemático equilibrado en la estimación de áreas bajo funciones que no satisfacen completamente las condiciones de la proposición 4.2, por ejemplo la función inversa de la distribución de una variable aleatoria no simétrica como la ji-cuadrada.

3. Encontrar condiciones necesarias para que la varianza del muestreo sistemático de una población fija sea cero o menor que la de otros diseños, por ejemplo el aleatorio simple.

4. Caracterizar las formas de ordenar a la población que ayuden a reducir la varianza del muestreo sistemático, de manera equilibrada es una de ellas.

5. En este trabajo no se abordó el muestreo sistemático con probabilidades desiguales, sería de interés generalizar las condiciones para que este diseño tenga

una varianza mínima.

Anexo

Proposición 5.4.1. *Sea ξ el modelo bajo el cual $Y_1, \ldots, Y_N$ son variables aleatorias independientes con distribución uniforme en el intervalo $[\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma]$. Sean n y N números pares que denotan el tamaño de muestra y el de la población, respectivamente, y $T = N/n \in \mathbb{N}$ el intervalo muestral. Si se selecciona una muestra sistemática de la población ordenada de manera equilibrada (BSS), entonces la varianza total del predictor $\hat{\theta}_\pi = T \sum_{i \in s_r} Y_{(i)}$ es igual a*

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]\sigma^2.$$

Demostración. Sin pérdida de generalidad suponga que $Y_1, Y_2, \ldots, Y_N \sim U(0,1)$ son variables aleatorias independientes y $X_i = Y_{(i)}$ denota la i-ésima estadística de orden de las N variables, entonces el valor esperado, la varianza y la covarianza bajo el modelo son respectivamente:

$$E_\xi[X_i] = \frac{i}{N+1}$$

$$V_\xi[X_i] = \frac{i(N+1-i)}{(N+1)^2(N+2)}$$

$$C_\xi[X_i, X_j] = \frac{i(N+1-j)}{(N+1)^2(N+2)} \text{ si } i < j.$$

La varianza total del predictor $\hat{\theta}_\pi$ es

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = E_{BSS}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] + V_{BSS}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right]. \tag{8.1}$$

A continuación se encontrarán expresiones para cada una de esas varianzas y valores esperados.

$$
\begin{aligned}
V_\xi\left(\hat{\theta}_\pi|s\right) &= V_\xi\left(T\sum_{i\in s_r} Y_{(i)}\,|s_r\right) \\
&= T^2\left\{\sum_{i\in s_r} V_\xi\left[Y_{(i)}\,|s_r\right] + \sum_{i\neq j\in s_r} C_\xi\left[Y_{(i)}, Y_{(j)}\,|s_r\right]\right\} \\
&= T^2\left\{\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) + \sum_{i=n/2+1}^{n} V_\xi\left(X_{-2r+2+2(n-i+1)T}\right)\right. \\
&\quad + 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&\quad + 2\sum_{i=1}^{n/2}\sum_{j=n/2+1}^{n} C_\xi\left(X_{2r-1+2(i-1)T}, X_{-2r+2+2(n-j+1)T}\right) \\
&\quad \left. +2\sum_{i=n/2+1}^{n-1}\sum_{j=i+1}^{n} C_\xi\left(X_{-2r+2+2(n-i+1)T}, X_{-2r+2+2(n-j+1)T}\right)\right\} \qquad (8.2) \\
&= T^2\left\{\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) + \sum_{i=1}^{n/2} V_\xi\left(X_{N-2r+2+2(1-i)T}\right)\right. \\
&\quad + 2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) \\
&\quad + 2\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{N-2r+2+2(1-j)T}\right) \\
&\quad \left. +2\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{N-2r+2+2(1-i)T}, X_{N-2r+2+2(1-j)T}\right)\right\}.
\end{aligned}
$$

Haciendo $a = 2r - 1 - 2T$, $b = N - 2r + 2 + 2T = N + 1 - a$ y $c = \frac{1}{(N+1)^2(N+2)}$ y desarrollando el primer término de la última igualdad se tiene que

$$
\begin{aligned}
\sum_{i=1}^{n/2} V_\xi\left(X_{2r-1+2(i-1)T}\right) &= \sum_{i=1}^{n/2} V_\xi\left(X_{a+2iT}\right) \\
&= c\sum_{i=1}^{n/2}(a+2iT)(N+1-a-2iT) \\
&= c\sum_{i=1}^{n/2}(a+2iT)(b-2iT)
\end{aligned}
$$

$$= c\left[ab\frac{n}{2} + (2Tb - 2Ta)\sum_{i=1}^{n/2} i - 4T^2\sum_{i=1}^{n/2} i^2\right]. \qquad (8.3)$$

Del segundo término de la última igualdad en (8.2)

$$\begin{aligned}\sum_{i=1}^{n/2} V_\xi\left(X_{N-2r+2+2(1-i)T}\right) &= \sum_{i=1}^{n/2} V_\xi\left(X_{b-2iT}\right)\\ &= c\sum_{i=1}^{n/2}(b-2iT)(a+2iT)\\ &= c\left[ab\frac{n}{2} + (2Tb - 2Ta)\sum_{i=1}^{n/2} i - 4T^2\sum_{i=1}^{n/2} i^2\right].\end{aligned}$$

Como puede verse $V_\xi\left(X_{2r-1+2(i-1)T}\right) = V_\xi\left(X_{N-2r+2+2(1-i)T}\right)$. Por otro lado, sustituyendo a, b y c en la primer covarianza de (8.2) se tiene que

$$\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{2r-1+2(j-1)T}\right) = \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{a+2jT}\right)$$

efectuando operaciones,

$$\begin{aligned}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{a+2jT}\right) &= \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}\frac{(a+2iT)(N+1-a-2jT)}{(N+1)^2(N+2)}\\ &= c\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}(a+2iT)(b-2jT)\\ &= c\left\{ab\frac{n^2}{4} - ab\frac{n}{2} - Ta\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}-1\right)\right. \qquad (8.4)\\ &\quad + \left[-ab + Ta + Tbn - T^2 n\left(\frac{n}{2}+1\right)\right]\sum_{i=1}^{n/2-1} i\\ &\quad \left. + \left[Ta - 2Tb + 2T^2\right]\sum_{i=1}^{n/2-1} i^2 + 2T^2\sum_{i=1}^{n/2-1} i^3\right\}.\end{aligned}$$

Sustituyendo a, b y c y desarrollando la tercer covarianza de (8.2) se tiene lo siguiente

$$\begin{aligned}\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} C_\xi\left(X_{b-2iT}, X_{b-2jT}\right) &= \sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2} \frac{(b-2jT)(N+1-b+2iT)}{(N+1)^2(N+2)} \\ &= c\sum_{i=1}^{n/2-1}\sum_{j=i+1}^{n/2}(b-2jT)(a+2iT) \\ &= c\Bigg\{ab\frac{n^2}{4} - ab\frac{n}{2} - Ta\frac{n}{2}\left(\frac{n}{2}+1\right)\left(\frac{n}{2}-1\right) \\ &\quad + \left[-ab+Ta+Tbn-T^2n\left(\frac{n}{2}+1\right)\right]\sum_{i=1}^{n/2-1} i \\ &\quad + \left[Ta-2Tb+2T^2\right]\sum_{i=1}^{n/2-1} i^2 + 2T^2\sum_{i=1}^{n/2-1} i^3\Bigg\}.\end{aligned} \tag{8.5}$$

Ya que $C_\xi\left(X_i, X_j\right)$ es válida para $i<j$, para derivar la covarianza

$$\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{2r-1+2(i-1)T}, X_{N-2r+2+2(1-j)T}\right) = \sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right), \tag{8.6}$$

se consideran cuatro casos: i) T par con $r \leq \frac{T}{2}$, ii) T par y $r \geq \frac{T}{2}+1$, iii) T impar con $r \leq \frac{T+1}{2}$ y iv) T impar con $r \geq \frac{T+1}{2}+1$.

i) Sea T par y $r \leq \frac{T}{2}$

$$\begin{aligned}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) &= \sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\ &\quad + \sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\ &= c\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}(a+2iT)(a+2jT) \\ &\quad + c\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2}(b-2jT)(b-2iT)\end{aligned}$$

$$
\begin{aligned}
&= c\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}\left(a^2+2aTj+2aTi+4T^2ij\right)\\
&\quad + c\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2}\left(b^2-2bTi-2bTj+4T^2ij\right)\\
&= a^2\frac{n^2}{4}+a^2\frac{n}{2}+Ta\frac{n^3}{8}+Ta\frac{3n^2}{4}+Tan\\
&\quad + \left[-a^2-Ta+T^2\frac{n^2}{2}+3T^2n+4T^2\right]\sum_{i=1}^{n/2} i\\
&\quad + \left[-Ta-6T^2-2T^2n\right]\sum_{i=1}^{n/2} i^2+2T^2\sum_{i=1}^{n/2} i^3\\
&\quad + \frac{2}{4}Tbn^2+Tbn-b^2\frac{n}{2}\\
&\quad + \left[b^2-Tb-2T^2n-Tbn-4T^2\right]\sum_{i=1}^{n/2} i\\
&\quad + \left[-Tb+2nT^2+6T^2\right]\sum_{i=1}^{n/2} i^2-2T^2\sum_{i=1}^{n/2} i^3.
\end{aligned}
\tag{8.7}
$$

Sustituyendo los resultados encontrados en (8.3), (8.4), (8.5) y (8.7) en la última igualdad de (8.2) y simplificando se obtiene

$$
V_\xi\left(\hat{\theta}_\pi|s_r\right)=2T^2c\,(a+b)\,n\left[(a+b)\,\frac{n}{8}-(b-a)\,\frac{1}{4}-\frac{T}{12}\,(n+2)\,(n-5)\right].
$$

Ya que $a+b=N+1$ y $b-a=N+4T+3-4r$ la ecuación anterior se puede escribir como

$$
\begin{aligned}
V_\xi\left(\hat{\theta}_\pi|s_r\right) &= 2T^2c\,(N+1)\,n\left[(N+1)\,\frac{n}{8}-(N+4T+3-4r)\,\frac{1}{4}-\frac{T}{12}\,(n+2)\,(n-5)\right]\\
&= 2T^2c\,(N+1)\,n\left[\frac{Nn}{24}-\frac{N}{6n}+\frac{n}{8}-\frac{3}{4}+r\right],
\end{aligned}
$$

sustituyendo $c=\frac{1}{(N+1)^2(N+2)}$ finalmente se llega a

$$
V_\xi\left(\hat{\theta}_\pi|s_r\right)=\frac{2N^2}{(N+1)(N+2)n}\left[\frac{Nn}{24}-\frac{N}{6n}+\frac{n}{8}-\frac{3}{4}+r\right]\ \text{si } T \text{ es par y } r\leq\frac{T}{2}.
$$

ii) Sea T par y $r \geq \frac{T}{2}+1$

$$\begin{aligned}
\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) &= \sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\
&\quad + \sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\
&= c\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i}\left(a^2+2aTj+2aTi+4T^2ij\right) \\
&\quad + c\sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2}\left(b^2-2bTi-2bTj+4T^2ij\right) \\
&= a^2\frac{n^2}{4}+Ta\frac{n^3}{8}+Ta\frac{n^2}{4} \\
&\quad + \left[-a^2-Ta+T^2\frac{n^2}{2}+T^2n\right]\sum_{i=1}^{n/2} i \\
&\quad + \left[-Ta-2T^2-2T^2n\right]\sum_{i=1}^{n/2} i^2 + 2T^2\sum_{i=1}^{n/2} i^3 \\
&\quad + \left(b^2-Tb-Tbn\right)\sum_{i=1}^{n/2} i + \left[-Tb+2nT^2+2T^2\right]\sum_{i=1}^{n/2} i^2 \\
&\quad - 2T^2\sum_{i=1}^{n/2} i^3.
\end{aligned} \tag{8.8}$$

Sustituyendo (8.3), (8.4), (8.5) y (8.8) en la última igualdad de (8.2) y simplificando se obtiene

$$V_\xi\left(\hat{\theta}_\pi|s_r\right) = 2T^2c\left(a+b\right)n\left[\left(a+b\right)\frac{n}{8}+\left(b-a\right)\frac{1}{4}-\frac{T}{12}\left(n+2\right)\left(n+1\right)\right].$$

Puesto que $a+b=N+1$ y $b-a=N+4T+3-4r$ la ecuación anterior se puede escribir como

$$V_\xi\left(\hat{\theta}_\pi|s_r\right) = 2T^2c\left(N+1\right)n\left[\left(N+1\right)\frac{n}{8}+\left(N+4T+3-4r\right)\frac{1}{4}-\frac{T}{12}\left(n+2\right)\left(n+1\right)\right].$$

Sustituyendo $c=\frac{1}{(N+1)^2(N+2)}$ y efectuando operaciones

$$V_\xi\left(\hat{\theta}_\pi|s_r\right) = \frac{2N^2}{(N+1)(N+2)n}\left[\frac{Nn}{24}+\frac{5N}{6n}+\frac{n}{8}+\frac{3}{4}-r\right] \text{ si } T \text{ es par y } r \geq \frac{T}{2}+1.$$

Luego, el valor esperado sobre el diseño es

$$\begin{aligned} E_{BSS}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] &= \frac{1}{T}\sum_{r=1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right) \\ &= \frac{1}{T}\left\{\sum_{r=1}^{T/2} V_\xi\left(\hat{\theta}_\pi|s_r\right) + \sum_{r=T/2+1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right)\right\} \\ &= \frac{1}{T}\frac{2N^2}{(N+1)(N+2)n}\left\{\sum_{r=1}^{T/2}\left[\frac{Nn}{24} - \frac{N}{6n} + \frac{n}{8} - \frac{3}{4} + r\right]\right. \\ &\quad \left. + \sum_{r=T/2+1}^{T}\left[\frac{Nn}{24} + \frac{5N}{6n} + \frac{n}{8} + \frac{3}{4} - r\right]\right\} \\ &= \frac{N^2}{12(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]. \end{aligned} \tag{8.9}$$

Encontrando el segundo término de la varianza total

$$\begin{aligned} E_\xi\left(\hat{\theta}_\pi|s\right) &= E_\xi\left[T\sum_{i\in s_r} Y_{(i)}\right] \\ &= T\sum_{i\in s_r} E_\xi\left[Y_{(i)}\right] \\ &= T\frac{n}{2} \\ &= \frac{N}{2}. \end{aligned} \tag{8.10}$$

Ya que $E_\xi\left(\hat{\theta}_\pi|s\right)$ es una constante, entonces

$$V_{BSS}\left[E_\xi\left(\hat{\theta}_\pi|s\right)\right] = 0. \tag{8.11}$$

Por lo tanto, reemplazando (8.9) y (8.11) en (8.1) se tiene que la varianza total es

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{12(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]. \tag{8.12}$$

Cuando T es impar, las expresiones para (8.6) son las mismas que las halladas

cuando T es par, como se verá a continuación. Si $r \leq \frac{T+1}{2}$ se tiene que

$$\begin{aligned}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) &= \sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) + \sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\ &= c\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i+1}\left(a^2 + 2aTj + 2aTi + 4T^2ij\right) \\ &\quad + c\sum_{i=1}^{n/2}\sum_{j=n/2-i+2}^{n/2}\left(b^2 - 2bTi - 2bTj + 4T^2ij\right)\end{aligned}$$

en donde la última igualdad está dada por (8.7). Análogamente, si $r \geq \frac{T+1}{2} + 1$

$$\begin{aligned}\sum_{i=1}^{n/2}\sum_{j=1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) &= \sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) + \sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2} C_\xi\left(X_{a+2iT}, X_{b-2jT}\right) \\ &= c\sum_{i=1}^{n/2}\sum_{j=1}^{n/2-i}\left(a^2 + 2aTj + 2aTi + 4T^2ij\right) \\ &\quad + c\sum_{i=1}^{n/2}\sum_{j=n/2-i+1}^{n/2}\left(b^2 - 2bTi - 2bTj + 4T^2ij\right).\end{aligned}$$

Los términos de esta última igualdad corresponden a los dados en la expresión (8.8). El valor esperado bajo el diseño de la varianza bajo el modelo en el caso en que T es impar es igual al caso de T par como se verá enseguida

$$\begin{aligned}E_{BSS}\left[V_\xi\left(\hat{\theta}_\pi|s\right)\right] &= \frac{1}{T}\sum_{r=1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right) \\ &= \frac{1}{T}\left\{\sum_{r=1}^{\frac{T+1}{2}} V_\xi\left(\hat{\theta}_\pi|s_r\right) + \sum_{r=\frac{T+1}{2}+1}^{T} V_\xi\left(\hat{\theta}_\pi|s_r\right)\right\} \\ &= \frac{1}{T}\frac{2N^2}{(N+1)(N+2)n}\left\{\sum_{r=1}^{\frac{T+1}{2}}\left[\frac{Nn}{24} - \frac{N}{6n} + \frac{n}{8} - \frac{3}{4} + r\right]\right. \\ &\quad \left. + \sum_{r=\frac{T+1}{2}+1}^{T}\left[\frac{Nn}{24} + \frac{5N}{6n} + \frac{n}{8} + \frac{3}{4} - r\right]\right\} \\ &= \frac{N^2}{12(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right].\end{aligned}$$

Por lo tanto, la varianza anticipada para cualquier $T \in \mathbb{Z}^+$ está dada por la ecuación (8.12). Cuando $Y_i^* \sim U(0,1)$, entonces $E(Y_i^*) = 1/2$ y $V(Y_i^*) = 1/12$, si se supone

$Y_i \sim U(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$, entonces $E(Y_i) = \mu$ y $V(Y_i) = \sigma^2$. Es fácil ver que en términos de la distribución uniforme estándar $Y_i = \mu - \sqrt{3}\sigma + 2\sqrt{3}\sigma Y_i^*$ y por ende $V(Y_i) = 12\sigma^2 V(Y_i^*)$. Finalmente, la varianza total considerando una media μ y varianza σ^2 es

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]\sigma^2. \qquad \square$$

Proposición 5.4.3. *Sean N y n números pares, tales que $T = N/n \in \mathbb{N}$, entonces se cumple que*

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) \leq V_{SYC,\xi}\left(\hat{\theta}_\pi\right) \leq V_{SI,\xi}\left(\hat{\theta}_\pi\right).$$

Demostración. Primero se demostrará que si $n = 2, 4, \ldots, N/2$, el BSS es más preciso que el SYC.

$$\begin{aligned}
\frac{N}{n} &> 1 \\
\frac{N^2}{n^2} &> 1 \\
\left(\frac{N}{N+2}\frac{N}{n^2}\right) &> \frac{1}{N+2} \\
\left(1 - \frac{2}{N+2}\right)\frac{N}{n^2} &> \frac{1}{N+2} \\
\frac{N}{n^2} &> \frac{1}{N+2} + \left(\frac{2}{N+2}\right)\frac{N}{n^2} \\
1 + \frac{N}{n^2} &> 1 + \frac{1}{N+2} + \left(\frac{2}{N+2}\right)\frac{N}{n^2} \\
1 + \frac{N}{n^2} &> \frac{1}{N+2}\left(N + 3 + \frac{2N}{n^2}\right) \\
\frac{N^2}{N+1}\left(1 + \frac{N}{n^2}\right) &> \frac{N^2}{(N+1)(N+2)}\left[N + 3 + \frac{2N}{n^2}\right] \\
\frac{N^2}{N+1}\left(1 + \frac{N}{n^2}\right)\sigma^2 &> \frac{N^2}{(N+1)(N+2)}\left[N + 3 + \frac{2N}{n^2}\right]\sigma^2
\end{aligned}$$

Por lo tanto,

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) > V_{BSS,\xi}\left(\hat{\theta}_\pi\right).$$

Ahora se demostrará la segunda parte de la desigualdad

$$\begin{aligned} N + \frac{N^2}{n^2} &< \frac{N^2}{n^2} \\ \Rightarrow N\left(1 + \frac{N}{n^2}\right) &< \frac{N^2}{n^2} \\ \Rightarrow N\left(1 + \frac{N}{n^2}\right)\sigma^2 &< \frac{N^2}{n^2}\sigma^2 \end{aligned}$$

Luego,

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) < V_{SI,\xi}\left(\hat{\theta}_\pi\right).$$

Por otra parte, sustituyendo $n = N$ en las expresiones

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)(N+2)}\left[N + 3 + 2\frac{N}{n^2}\right]\sigma^2,$$

$$V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{(N+1)}\left[1 + \frac{N}{n^2}\right]\sigma^2,$$

$$V_{SI,\xi}\left(\hat{\theta}_\pi\right) = \frac{N^2}{n}\sigma^2,$$

se tiene que

$$V_{BSS,\xi}\left(\hat{\theta}_\pi\right) = V_{SYC,\xi}\left(\hat{\theta}_\pi\right) = V_{SI,\xi}\left(\hat{\theta}_\pi\right) = N\sigma^2,$$

con lo cual se completa la demostración. □

Bibliografía

[1] F. Bartolucci and G. E. Montanari. A new class of unbiased estimators of the variance of the systematic sample mean. *Statistical Planning and Inference*, 136:1512–1525, 2006.

[2] D. R. Bellhouse and J. N. K. Rao. Systematic sampling in the presence of a trend. *Biometrika*, 62(3):694–697, 1975.

[3] Y. G. Berger. A variance estimator for systematic sampling from a deliberately ordered population. *Communications in Statistics. Theory and Methods*, 34:1533–1541, 2005.

[4] G. Blom. *Statistical Estimates and Transformed Beta Variables*. John Wiley and Sons, 1958.

[5] C. M. Cassel, C. E. Särndal, and J. H. Wretman. *Foundations of Inference in Survey Sampling*. Wiley, 1977.

[6] W. G. Cochran. Relative accuracy of systematic and stratified random samples for a certain class of populations. *The Annals of Mathematical Statistics*, 17(2):164–177, 1946.

[7] W. G. Cochran. *Sampling Techniques*. Wiley, 1977.

[8] L. M. Cruz-Orive. On the precision of systematic sampling: a review of matheron's transitive methods. *Journal of Microscopy*, 153:315–333, 1989.

[9] L. M. Cruz-Orive. A general variance predictor for Cavalieri slices. *Journal of Microscopy*, 222:158–165, 2006.

[10] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley, 2003.

[11] C. S. Davis and M. A. Stephens. Algorithm as 128: Approximating the covariance matrix of normal order statistics. *Applied Statistics*, 27(2):206–212, 1978.

[12] W. A. Fuller. *Introduction to Statistical Time Series*. Wiley Series in Probability and Statistics, 1996.

[13] W. A. Fuller. *Sampling Statistics*. Wiley, John and Sons, 2009.

[14] M. García-Fiñana and L. M. Cruz-Orive. Improved variance prediction for systematic sampling on R. *Statistics*, 38(3):243–272, 2004.

[15] W. Gautschi. Some remarks on systematic sampling. *The Annals of Mathematical Statistics*, 28(2):385–394, 1957.

[16] H. J. Godwin. Some low moments of order statistics. *The Annals of Mathematical Statistics*, 20(2):279–285, 1949.

[17] Z. Govindarajulu. Best linear estimates under symmetric censoring of the parameters of double exponential population. *Journal of the American Statistical Association*, 61(313):248–258, 1966.

[18] X. Gual-Arnau and L. M. Cruz-Orive. Variance prediction under systematic sampling with geometric probes. *Adv. Appl. Prob.*, 30:889–903, 1998.

[19] X. Gual-Arnau and L. M. Cruz-Orive. Systematic sampling on the circle and on the sphere. *Adv. Appl. Prob.*, 32:628–647, 2000.

[20] H. J. G. Gundersen. The smooth fractionator. *Journal of Microscopy*, 207:191–210, 2002.

[21] H. L. Harter. Expected values of normal order statistics. *The Annals of Mathematical Statistics*, 48(1/2):151–165, 1961.

[22] C. Hastings, F. Mosteller, J. W. Tukey, and C. P. Winsor. Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, 18(3):413–426, 1947.

[23] J. Hájek. Optimal strategy and other problems in probability sampling. *Casopis pro Pestování Matematiky*, 84(4):387–423, 1959.

[24] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[25] R. Iachan. Systematic sampling: A critical review. *International Statistical Review*, 50(3):293–303, 1982.

[26] R. Iachan. Asymptotic theory of systematic sampling. *The Annals of Statistics*, 11(3):959–969, 1983.

[27] INEGI. Tabulados básicos nacionales y por entidad federativa. Base de datos y tabulados de la muestra censal XII Censo General de Población y Vivienda 2000, México, 2001. Disco compacto.

[28] C. T. Isaki and W. A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982.

[29] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, Vol. I.* John Wiley and Sons, 1994.

[30] H. L. Jones. Exact lower moments of order statistics in small samples from a normal distribution. *The Annals of Mathematical Statistics*, 19(2):270–273, 1948.

[31] M. G. Kendall and A. Stuart. *The advanced theory of Statisticis. Volume 1. Distribution theory.* Griffin, 1977.

[32] K. Kiêu. Three lectures on sistematic geometric sampling. Memoirs, 13. University of Aarhus, 1997.

[33] K. Kiêu, S. Souchet, and J. Istas. Precision of systematic sampling and transitive methods. *Statistical Planning and Inference*, 77:263–279, 1999.

[34] W. G. Madow and L. H. Madow. On the theory of systematic sampling I. *Annals of Mathematics Statistics*, 15(1):1–24, 1944.

[35] G. Matheron. Les variables régionaliseés et leur estimation. Masson, Paris, 1965.

[36] G. Matheron. The theory of regionalized variables and its applications. Les Cahiers du Centre de Morpholophie Mathématique de Fontainebleau, No. 5. École Nationale Supérieure des Mines de Paris, Fontainebleau, 1971.

[37] T. Mattfeldt. The accuracy of one-dimensional systematic sampling. *Journal of Microscopy*, 153:301–313, 1989.

[38] P. A. P. Moran. Numerical integration by systematic sampling. *Proc. Camb. Phil. Soc.*, 46:111–115, 1950.

[39] M. N. Murthy. *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, 1967.

[40] C. Muzquiz Fragoso. Eficiencia relativa del muestreo sistemático bajo diversos ordenamientos de las unidades poblacionales. Tesis de maestría, Universidad Nacional Autónoma de México, México, 2004.

[41] G. Nathan. Superpopulation models in survey sampling. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1575–1577. Springer-Verlag, 2011.

[42] J. D. Opsomer, M. Francisco-Fernández, and X. Li. Model-based nonparametric variance estimation for systematic sampling. *Scandinavian Journal of Statistics*, 39(3):528–542, 2012.

[43] G. Polya. Remarks on computing the probability integral in one and two dimensions. In J. Neyman, editor, *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, pages 63–78. University of Press, 1946.

[44] M. H. Quenouille. Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3):355–375, 1949.

[45] J. P. Royston. Algorithm AS 177: Expected normal order statistics (exact and approximate). *Applied Statistics*, 31(2):161–165, 1982.

[46] S. Sampath, V. Varalakshmi, and R. Geetha. Estimation under systematic sampling schemes for parabolic populations. *Journal of Applied Statistics*, 36(11):1281–1292, 2009.

[47] R. Sedgewick and P. Flajolet. *An introduction to the analysis of algorithms*. Addison-Wesley, 2013.

[48] V. K. Sethi. *Contributions to Stratified Sampling and Some Related Problems*. Tesis doctoral, Instituto de Ciencias Sociales, Universidad de Agra, India, 1963.

[49] V. K. Sethi. On optimum paring of units. *Sankhya: The Indian Journal of Statistics, Series B*, 27(3/4):315–320, 1965.

[50] B. L. Shea and A. J. Scallon. Remark AS R72: A remark on algorithm AS 128. Approximating the covariance matrix of normal order statistics. *Applied Statistics*, 37(1):151–155, 1988.

[51] D. Singh, K. K. Jindal, and J. N. Garg. On modified systematic sampling in the presence of a trend. *Biometrika*, 55(3):541–546, 1968.

[52] S. Souchet. Précision de l'estimateur de Cavalieri. Rapport de stage, D.E.A. de statistiques et modèles aléatoires appliqués à la finance, Université Paris-VII, Laboratoire de Biométrie, INRA-Versailles, 1995.

[53] C. E. Särndal. Design-based and model-based inference in suvey sampling. *Scandinavian Journal of Statistics*, 5(1):27–52, 1978.

[54] C. E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, 1992.

[55] D. Teichroew. Tables of expected values of order statistics and products of order statistics for samples of size twenty and less from the normal distribution. *The Annals of Mathematical Statistics*, 27(2):410–426, 1956.

[56] Y. Tillé. *Sampling Algorithms*. Springer, 2006.

[57] M. Tinajero-Bravo, G. Eslava-Gómez, and L. M. Cruz-Orive. Conditions for exact Cavalieri estimation. *Image Analysis and Stereology*, por aparecer.

[58] K. M. Wolter. An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, 79:781–790, 1984.

[59] K. M. Wolter. *Introduction to Variance Estimation*. Springer-Verlag, 2007.

[60] F. Yates. Systematic sampling. *Philosophical Transactions of the Royal Society of London*, 241(834):345–377, 1948.

[61] F. Yates. *Sampling Methods for Censuses and Surveys.* Hafner Publishing Company, New York, third edition, 1960.

[62] F. Yates and P. M. Grundy. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):253–261, 1953.