Demostraciones para Diseño de Muestreo Sistemático

1. Demostración de una Identidad Algebraica para la Varianza

Objetivo: Demostrar la siguiente igualdad, que es un paso intermedio en la derivación de la varianza del estimador.

$$a\sum_{\xi=1}^{a}\left(t_{m_\xi}-\frac{t_y}{a}\right)^2 = a\sum_{\xi=1}^{a}t_{m_\xi}^2 - t_y^2$$

Para esta demostración, se utiliza la propiedad fundamental de que la suma de los totales de todas las a muestras sistemáticas posibles es igual al total poblacional, t_y. Es decir:

$$\sum_{\xi=1}^{a}t_{m_\xi} = t_y$$

Demostración:

$$\begin{aligned}
& a\sum_{\xi=1}^{a}\left(t_{m_\xi}-\frac{t_y}{a}\right)^2 \\
&= a\sum_{\xi=1}^{a}\left(t_{m_\xi}^2 - 2t_{m_\xi}\frac{t_y}{a} + \left(\frac{t_y}{a}\right)^2\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \sum_{\xi=1}^{a}2t_{m_\xi}\frac{t_y}{a} + \sum_{\xi=1}^{a}\frac{t_y^2}{a^2}\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \frac{2t_y}{a}\sum_{\xi=1}^{a}t_{m_\xi} + \frac{t_y^2}{a^2}\sum_{\xi=1}^{a}1\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \frac{2t_y}{a}\sum_{\xi=1}^{a}t_{m_\xi} + \frac{t_y^2}{a^2}\cdot a\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \frac{2t_y}{a}(t_y) + \frac{t_y^2}{a}\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \frac{2t_y^2}{a} + \frac{t_y^2}{a}\right) \\
&= a\left(\sum_{\xi=1}^{a}t_{m_\xi}^2 - \frac{t_y^2}{a}\right) \\
&= a\sum_{\xi=1}^{a}t_{m_\xi}^2 - a\cdot\frac{t_y^2}{a} \\
&= a\sum_{\xi=1}^{a}t_{m_\xi}^2 - t_y^2
\end{aligned}$$

2. Demostración de la Varianza en Términos del Coeficiente de Correlación Intramuestral

Resultado 7: La varianza de un diseño sistemático se puede escribir alternativamente como:

$$V(\hat{t}_\pi) = \frac{N^2 S^2_{YU}}{n}[(1-f)+(n-1)\delta]$$

con δ el coeficiente de varianza (o correlación) intramuestral.

Definiciones Clave:

- N: Tamaño de la población.
- n: Tamaño de la muestra.
- $k = N/n$: Intervalo de muestreo y número de muestras sistemáticas posibles.
- $f = n/N$: Fracción de muestreo.
- S^2_{YU}: Varianza de la población, $\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y}_U)^2$.
- $V_{MAS}(\hat{t}_\pi)$: Varianza del estimador del total en Muestreo Aleatorio Simple, que es $\frac{N^2 S^2_{YU}}{n}(1-f)$.

Estrategia de Demostración: La demostración parte de la descomposición de la varianza total (ANOVA), que separa la variabilidad total de la población en la variabilidad *entre* las muestras sistemáticas y la variabilidad *dentro* de las mismas.

Demostración: Partimos de la identidad de la descomposición de la Suma de Cuadrados Totales (SCT):

$$SCT = SCE + SCD \tag{1}$$

donde:

- $SCT = \sum_{i=1}^{N}(y_i - \bar{Y}_U)^2 = (N-1)S^2_{YU}$ (Suma de Cuadrados Total).
- $SCE = \sum_{\xi=1}^{k} n(\bar{y}_{s_\xi} - \bar{Y}_U)^2$ (Suma de Cuadrados **Entre** muestras).
- $SCD = \sum_{\xi=1}^{k}\sum_{i\in s_\xi}(y_i - \bar{y}_{s_\xi})^2$ (Suma de Cuadrados **Dentro** de las muestras).

La varianza del estimador del total en muestreo sistemático, $V(\hat{t}_\pi)$, se define a partir de la variabilidad de las medias de todas las k muestras posibles:

$$V(\hat{t}_\pi) = V(N\bar{y}_s) = N^2 V(\bar{y}_s) = N^2 \frac{1}{k}\sum_{\xi=1}^{k}(\bar{y}_{s_\xi} - \bar{Y}_U)^2 = \frac{N^2}{k}\frac{SCE}{n} \tag{2}$$

Sustituyendo $k = N/n$, obtenemos una relación directa entre la varianza y la variabilidad entre muestras:

$$V(\hat{t}_\pi) = \frac{N^2}{(N/n)n} SCE = N \cdot SCE \tag{3}$$

Despejando SCE de la identidad ANOVA: $SCE = SCT - SCD$. Sustituimos en la ecuación de la varianza:

$$V(\hat{t}_\pi) = N \cdot (SCT - SCD) = N\left((N-1)S^2_{YU} - \sum_{\xi=1}^{k}(n-1)S^2_{w\xi}\right) \tag{4}$$

donde $S^2_{w\xi}$ es la varianza dentro de la ξ-ésima muestra.

Esta es una expresión exacta de la varianza. La fórmula del **Resultado 7** es una reorganización de esta expresión, muy útil para la interpretación. Se obtiene al relacionar la varianza del muestreo sistemático (V_{SYS}) con la del Muestreo Aleatorio Simple (V_{MAS}):

$$V_{SYS} = V_{MAS}[1 + (n-1)\rho_w]$$

donde ρ_w (o δ en nuestra notación) es el coeficiente de correlación intramuestral, que mide qué tan parecidas son las unidades dentro de una misma muestra sistemática en comparación con la población general.

La fórmula del ejercicio separa explícitamente el componente de MAS:

$$V(\hat{t}_\pi) = \underbrace{\frac{N^2 S_{YU}^2}{n}(1-f)}_{V_{MAS}(\hat{t}_\pi)} + \underbrace{\frac{N^2 S_{YU}^2}{n}(n-1)\delta}_{\text{Ajuste por diseño sistemático}}$$

Esta forma es una re-expresión donde δ se define para que la igualdad sea cierta. Compara la eficiencia del diseño sistemático con el MAS. Si δ es negativo (las muestras son heterogéneas internamente), el muestreo sistemático es más preciso que el MAS. Si δ es positivo (las muestras son homogéneas), es menos preciso. Si $\delta \approx 0$, su eficiencia es similar.