Un Nuevo Estimador Insesgado de la Razón para Variables Proporcionales: Análisis de su Poca Efectividad

Antonio Mayorquin Garcia
Técnicas de Muestreo, Especialidad 2024
IIMAS, UNAM

2024-09-20

Resumen

Este informe presenta un nuevo estimador insesgado de la razón para dos variables aleatorias proporcionales. La ventaja teórica es su Aunque el estimador clásico de la razón puede ser sesgado, los experimentos numéricos mostraron que dicho sesgo no es significativo en la práctica, incluso con tamaños de muestra moderados. El estimador propuesto R', aunque teóricamente insesgado, presenta una mayor varianza cuando los valores de las variables X_i son pequeños, lo que lo hace menos preciso en comparación con el estimador clásico R. Conforme los valores de X_i se alejan de cero, el rendimiento del estimador R' mejora, pero no se observaron ventajas claras sobre el estimador clásico en los escenarios evaluados. En general, el estimador clásico R demostró ser una opción más confiable y flexible para la estimación de la razón, mientras que el uso de R' debe ser evaluado cuidadosamente según las características de los datos.

1. Introducción

En muestreo estadístico, los estimadores de razón son comúnmente usados cuando dos variables, $\mathcal{Y}$ y $\mathcal{X}$, están correlacionadas, frecuentemente proporcionales entre sí, como en el caso $\mathcal{Y} \propto \mathcal{X}$. El estimador clásico de la razón R ofrece una estimación de esta proporcionalidad. Sin embargo, es bien sabido que este estimador es sesgado, especialmente en muestras pequeñas. Aunque

este sesgo se reduce con tamaños de muestra grandes debido a la aproximación normal, presenta una limitación significativa en aplicaciones prácticas. Este informe propone un estimador alternativo, R', que es insesgado y más fácil de manejar en términos de cálculo de varianza.

2. Definiciones y Fundamentos Teóricos

Sean $\mathcal{X}$ y $\mathcal{Y}$ variables aleatorias positivas tales que $\mathcal{Y} = \alpha\mathcal{X} + \epsilon$, donde ϵ es un término de error aleatorio con valor esperado 0 y varianza σ^2. Para una población de tamaño N, el estimador de la razón R se define como:

$$R = \frac{\sum_{i=1}^{N} Y_i}{\sum_{i=1}^{N} X_i} = \frac{\bar{Y}}{\bar{X}}$$

La razón de esto es poder llegar a la siguiente ecuación

$$Y = RX, \qquad \bar{Y} = R\bar{X}$$

Sin embargo, se puede demostrar que R es un estimador sesgado de α, particularmente en muestras pequeñas. La relación con α se observa viendo que para cada $i \in [0, n]$:

$$Y_i = \alpha X_i + \epsilon_i \implies \sum_{i=1}^{N} Y_i = \alpha \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} \epsilon_i \implies R = \alpha + \frac{\sum_{i=1}^{N} \epsilon_i}{\sum_{i=1}^{N} X_i}$$

3. Estimador Insesgado Propuesto

Proponemos un estimador alternativo R' definido como:

$$R' = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{X_i}$$

Al analizar su valor esperado y varianza, mostramos que este estimador es insesgado para α. Su relacióncon α se ve por lo siguiente:

$$\frac{Y_i}{X_i} = \alpha + \frac{\epsilon_i}{X_i} \implies R' = \alpha + \frac{1}{N} \sum_{i=1}^{N} \frac{\epsilon_i}{X_i}$$

4. Prueba de Insesgadez

Para demostrar que R' es insesgado, consideremos la siguiente expresión:

Sea Z_i una variable aleatoria de Bernoulli que indica si el elemento i está incluido en la muestra, con $E[Z_i] = \frac{n}{N}$. El estimador $\hat{R}'$, calculado a partir de una muestra de tamaño n, se da como:

$$\hat{R}' = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{X_i} = \frac{1}{n} \sum_{i=1}^{N} Z_i \frac{Y_i}{X_i}$$

El valor esperado de $\hat{R}'$ es:

$$E[\hat{R}'] = \frac{1}{n} \sum_{i=1}^{N} E[Z_i] \frac{Y_i}{X_i} = \frac{1}{n} \sum_{i=1}^{N} \frac{n}{N} \frac{Y_i}{X_i} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i}{X_i} = R'$$

Por lo tanto, $E[\hat{R}'] = R'$, lo que muestra que $\hat{R}'$ es un estimador insesgado de α.

5. Varianza de R'

La varianza del estimador $\hat{R}'$ se puede derivar de manera similar. Utilizando técnicas estándar, la varianza está dada por:

$$\text{Var}(\hat{R}') = \left(1 - \frac{n}{N}\right) \frac{1}{n} S^2(R')$$

Donde $S^2(R')$ es la varianza muestral de los valores $\frac{Y_i}{X_i}$, definida como:

$$S^2(R') = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{Y_i}{X_i} - R'\right)^2$$

6. Comparación de Varianza entre R y R'

Ahora comparamos la varianza del estimador clásico R con la del estimador propuesto R'. La varianza del estimador clásico está dada por:

$$\text{Var}(\hat{R}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{N} \frac{(Y_i - RX_i)^2}{N-1} \frac{1}{\bar{X}^2}$$

Usando la suposición de proporcionalidad $Y_i = \alpha X_i + \epsilon_i$, podemos expresar la diferencia de varianza entre R y R' como:

$$\left|\mathrm{Var}(\hat{R}') - \mathrm{Var}(\hat{R})\right| \approx \left(1 - \frac{n}{N}\right)\frac{1}{n}\sum_{i=1}^{N}\frac{\epsilon_i^2\left|\frac{1}{X_i^2} - \frac{1}{\bar{X}^2}\right|}{N-1}$$

Aplicando la desigualdad de Hölder con $p \to 1$ y $q \to \infty$, y suponiendo que $\sum_{i=1}^{N}\epsilon_i^2 \approx E[\epsilon] = n\sigma^2$, podemos acotar la diferencia en la varianza de la siguiente manera:

$$\left|\mathrm{Var}(\hat{R}') - \mathrm{Var}(\hat{R})\right| \leq \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{N-1}\max_i\left|\frac{1}{X_i^2} - \frac{1}{\bar{X}^2}\right|$$

Esta expresión proporciona una cota superior clara de la diferencia entre las varianzas de los dos estimadores, indicando que R' tiene una varianza cercana a la de R bajo suposiciones razonables.

7. Ejemplo Numérico

Para ilustrar el desempeño del estimador propuesto R' en comparación con el estimador clásico R, se ha llevado a cabo un experimento numérico considerando diferentes valores de desplazamiento en las variables X_i (denotado como x_{offset}).

Se generó una población de tamaño $N = 300$, donde las variables X_i siguen una distribución uniforme en el intervalo $x_{\text{offset}} + [0, 20$, y las variables Y_i se generaron usando el modelo $Y_i = \alpha X_i + \epsilon_i$, con $\alpha = 1{,}5$ y ϵ_i siguiendo una distribución normal con media 0 y varianza $\sigma^2 = 0{,}1$. Para cada valor de x_{offset} (de 0 a 15 en pasos de 1), se tomaron muestras de tamaño $n = 40$ y se repitió el experimento $k = 200$ veces.

Los estimadores R y R' fueron calculados para cada muestra, y sus valores medios, junto con sus desviaciones estándar, fueron graficados para cada valor de x_{offset}. El código empleado está en el Anexo A.

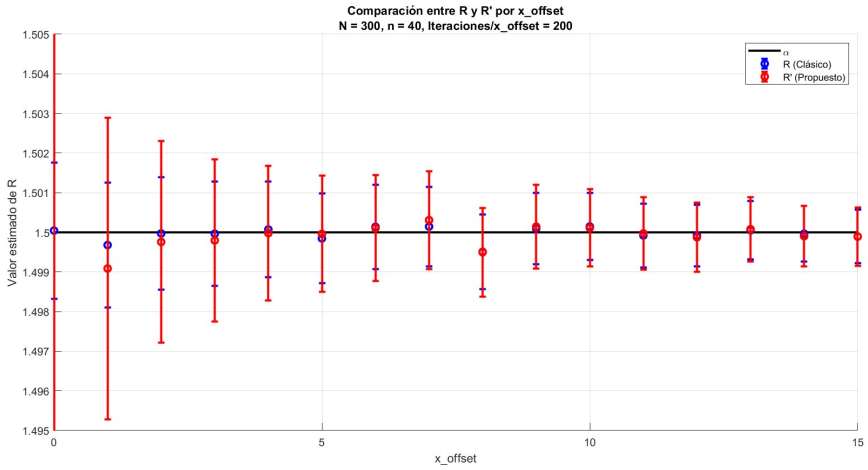

Figura 1: Comparación entre los estimadores R y R' para diferentes valores de x_{offset}, con $N = 300$, $n = 40$ y $k = 200$ iteraciones.

La Figura 1 ilustra la comparación entre los valores medios de R y R', con sus respectivas barras de error que representan la desviación estándar. Es evidente que, conforme x_{offset} aumenta, el rendimiento de R' mejora y se aproxima más al valor de α, reduciendo su varianza en comparación con R.

8. Discusión

El análisis numérico mostró que, en general, el estimador clásico R tiene una mayor precisión para aproximar el valor verdadero de $\alpha = 1{,}5$ en comparación con el estimador propuesto R', especialmente cuando los valores de x_{offset} son pequeños (es decir, cuando los X_i se aproximan más a cero). Este comportamiento se puede explicar por la estructura de la varianza en el estimador R', la cual incluye un término de división por X_i^2. Cuando los valores de X_i son cercanos a cero, este término aumenta significativamente, incrementando la varianza y reduciendo la precisión del estimador R'. En contraste, cuando los valores de x_{offset} son mayores, la varianza de R' disminuye considerablemente y su desempeño se vuelve más comparable al de R.

El sesgo teórico del estimador clásico R no parece ser un factor relevante en este experimento, ya que el estimador R mostró un comportamiento prácticamente insesgado en todos los escenarios, lo que plantea dudas sobre la ventaja de utilizar el estimador propuesto R' en este contexto. Además, R' podría no ofrecer la misma flexibilidad en la derivación de otras estimaciones,

como la estimación poblacional, lo que es una fortaleza del estimador clásico R.

Finalmente, sería útil realizar un análisis teórico más profundo de R' bajo diferentes distribuciones de X y ϵ, así como en diferentes configuraciones de muestreo, para determinar si existen escenarios en los que el estimador propuesto pueda tener ventajas claras sobre R.

9. Conclusión

Este informe presentó un nuevo estimador insesgado R' para la razón de dos variables proporcionales. Aunque el estimador R' resuelve teóricamente el problema de sesgo inherente al estimador clásico R, los resultados numéricos sugieren que el estimador clásico R tiene un mejor rendimiento en términos de precisión en una amplia gama de escenarios. La ventaja teórica de R' de ser insesgado se ve opacada por su alta varianza cuando X_i se aproxima a cero, lo que lo hace menos atractivo en situaciones donde los valores de X_i son bajos.

En conclusión, el estimador clásico R sigue siendo una opción preferible en la mayoría de los casos, y las ventajas del estimador R' no se manifiestan claramente en los experimentos realizados. Futuros trabajos deberían enfocarse en evaluar el desempeño de R' en otros esquemas de muestreo y en distribuciones alternativas, para encontrar posibles aplicaciones en las que el estimador propuesto supere al estimador clásico.

10. Referencias

- Romero Mares, P. I. (2024). Técnicas de Muestreo I. *Departamento de Probabilidad y Estadística, IIMAS, UNAM.* Recuperado de `https://drive.google.com/file/d/1r7E25z-Zrh0wldwyjQxyGvBSKbITtTEe/view`

A. Anexo: Código en MATLAB

En esta sección se presenta el código en MATLAB utilizado para realizar los cálculos y generar los gráficos comparativos entre los estimadores R y R'. El código se ejecutó para una población de tamaño $N = 300$, un tamaño de muestra $n = 40$, y $k = 200$ iteraciones, con diferentes valores de x_offset que varían entre 0 y 15.

Listing 1: Código MATLAB para la comparación de los estimadores R y R'.

```
% Parámetros
N = 300; % Tamaño de la población
n = 40;  % Tamaño de la muestra
alpha = 1.5; % Valor verdadero de la constante de proporcionalidad
sigma2 = 0.1; % Varianza de los errores
mu_epsilon = 0; % Media de los errores
k = 200; % Número de iteraciones

x_offset_List = 0:15; % Lista de valores para x_offset

% Inicialización de matrices para almacenar los resultados para cada x_offset
mean_R_values = zeros(length(x_offset_List), 1);
mean_R_prime_values = zeros(length(x_offset_List), 1);
std_R_values = zeros(length(x_offset_List), 1);
std_R_prime_values = zeros(length(x_offset_List), 1);

% Bucle sobre los valores de x_offset
for idx = 1:length(x_offset_List)
    x_offset = x_offset_List(idx);

    % Inicialización de los vectores para almacenar resultados en cada iteración
    R_values = zeros(k,1);
    R_prime_values = zeros(k,1);
    var_R_values = zeros(k,1);
    var_R_prime_values = zeros(k,1);

    % Bucle sobre las iteraciones
    for i = 1:k
        % Generación de los datos
        X = rand(N, 1) * 20 + x_offset; % X_i, generado de una distribución uniforme entre 0 y 20 + x_offset
        epsilon = normrnd(mu_epsilon, sqrt(sigma2), N, 1); % Errores epsilon_i, distribuidos normalmente
        Y = alpha * X + epsilon; % Y_i generado según el modelo Y_i = alpha * X_i + epsilon_i

        % Muestra aleatoria
        sample_indices = randperm(N, n); % Selección aleatoria de n índices
        X_sample = X(sample_indices); % Muestra de X
        Y_sample = Y(sample_indices); % Muestra de Y

        % Cálculo del estimador clásico R
        R = sum(Y_sample) / sum(X_sample);

        % Cálculo del estimador propuesto R'
```

```
        R_prime = mean(Y_sample ./ X_sample);

        % Cálculo de las varianzas
        var_R =       (1 - n/N)*(1/n)*sum((Y_sample-R*X_sample).^2)/(N-1) / mean(X_sample)^2;
        var_R_prime = (1 - n/N)*(1/n)*sum((Y_sample./X_sample-R_prime).^2)/(N-1);

        % Almacenar los valores de R, R' y sus varianzas
        R_values(i) = R;
        R_prime_values(i) = R_prime;
        var_R_values(i) = var_R;
        var_R_prime_values(i) = var_R_prime;
    end

    % Cálculo de las medias y desviaciones estándar para este x_offset
    mean_R_values(idx) = mean(R_values);
    mean_R_prime_values(idx) = mean(R_prime_values);
    std_R_values(idx) = sqrt(mean(var_R_values)); % Desviación estándar de R
    std_R_prime_values(idx) = sqrt(mean(var_R_prime_values)); % Desviación estándar de R'
end

% Graficar los resultados
figure;
hold on;
plot([x_offset_List(1), x_offset_List(end)], [alpha, alpha], 'k', 'LineWidth',2, 'DisplayName', "\alpha")
% Graficar R y R' con sus desviaciones estándar para cada x_offset
errorbar(x_offset_List, mean_R_values, std_R_values, 'o','Color', 'b', 'LineWidth', 2, 'DisplayName', 'R (Clásico)');
errorbar(x_offset_List, mean_R_prime_values, std_R_prime_values, 'o','Color', 'r', 'LineWidth', 2, 'DisplayName', 'R'' (Propuesto)');

% Configuración del gráfico
xlabel('x\_offset');
ylabel('Valor estimado de R');
ylim([alpha-0.005, alpha+0.005])
title({['Comparación entre R y R'' por x\_offset'], ['N = ', num2str(N), ', n = ', num2str(n), ', Iteraciones/x\_offset = ', num2str(k)]});
legend('show');
grid on;
hold off;
```