Revista Ciencia Aplicada
Conocimiento Multidisciplinario ISSN: 3122-3664

Artículo Científico

Consecuencias y estrategias ante la no respuesta: propuesta de una regla de submuestreo

Octaviano Juárez Romero[1], Pablo O. Juárez Moreno[2]

[1]Facultad de Matemáticas, Universidad Autónoma de Guerrero, Acapulco 39650, Guerrero, México.
[2]Área de Estadística, Facultad de Sociología, Universidad Autónoma de Guerrero, Av. Paseo de la cañada, Guerrero, México.

Resumen: La presencia de la no respuesta en los diseños de muestreo probabilístico constituye una problemática que puede generar estimaciones sesgadas respecto de los valores poblacionales. Para mitigar este problema, la teoría estadística recurre al submuestreo en el estrato de las unidades no respondientes con el fin de incorporar la información faltante de manera insesgada. En este trabajo se analizan las consecuencias de la no respuesta en la inferencia muestral y las estrategias para disminuirla, y se propone una nueva regla de submuestreo más eficiente diseñada para complementar la regla clásica de Bouza. Con base en la teoría de submuestreo, se evaluó el comportamiento de la nueva propuesta frente a otras tres reglas clásicas. Los resultados indican que la nueva regla es más eficiente bajo escenarios de baja a moderada no respuesta respecto de las otras reglas. Se concluye que la regla propuesta es más eficiente que las otras reglas cuando hay moderada no respuesta, es complementaria a la regla de Bouza y elimina la dependencia de parámetros fijados por el muestreador.

Palabras clave: fracción de muestreo, inferencia en no respondientes, simulación numérica y gráfica, muestreo en dos fases.

PRELIMINAR

1. Introducción

Al aplicar un diseño muestral en una investigación, el muestreador espera que ocurran dos eventos para obtener buenas estimaciones: que todos los individuos $u_{ij,i=1,2,\ldots,k_s,j=1,2,\ldots,q}$ en su muestra estén disponibles para responder la encuesta y que respondan a todas las preguntas de la encuesta. Con esto, aspira a que se obtenga una matriz de dimensión $k_s \bullet q$, donde $k_s = 1, 2, \ldots, n$,

$$\begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1q} \\ u_{21} & u_{22} & \cdots & u_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k_s1} & u_{k_s2} & \cdots & u_{k_sq} \end{bmatrix}$$

Esto es, que todos los individuos u_{ij}, de la muestra s de tamaño n hayan respondido a las q preguntas. En un caso individual, el vector fila de respuestas q-dimensional del k-ésimo elemento en la muestra es dado por,

Volumen 2, no. 2

Recibido: agosto 10, 2026
Aceptado: octubre 21, 2026
autor de correspondencia:
18006@uagro.mx

https://doi.org/10.66482/km88hy79

$$y_k = \left(y_{k1}, \ldots, y_{kq}\right), k = 1, .., n.$$

A pesar de una buena planeación de la encuesta, el vector de respuesta suele presentar faltantes, ya que en la práctica es difícil que todos los encuestados contesten todas las preguntas por lo que el encuestador tendrá inconvenientes para realizar estimaciones cercanas a la realidad. Entre los problemas más comunes del muestreo por encuestas aparece el sesgo y la no respuesta [1-3]. Esta última se presenta cuando no se encuentra disponible el elemento a encuestar, cuando un cuestionario enviado por algún medio (correo postal, correo electrónico, formulario digital, aplicación o teléfono móvil) no es contestado o por cualquier otro motivo. Al haber entradas faltantes en la matriz de dimensión $k_s xq$, surgen dificultades teóricas y prácticas para realizar inferencia.

Si un encuestado no contesta una pregunta, entonces hay no respuesta en alguna entrada de alguna de las q columnas, pero si el individuo u_{k_s} no es localizado o se rehúsa a contestar toda la encuesta, genera en la matriz de datos una línea que queda completamente en blanco, es decir, no hay respuestas en las q preguntas. En ambos casos existen datos faltantes, motivo por el cual el muestreador debe decidir si realizar inferencia o no, a pesar de la no respuesta. Es evidente que, si hay un porcentaje alto de no respuesta, las estimaciones no proporcionarán información confiable o estarán sesgadas con respecto a la característica poblacional a investigar.

En este documento se abordan las consecuencias teóricas que existen al haber no respuesta al aplicar una encuesta bajo la teoría de muestreo, que son el sesgo y la ausencia de estimadores insesgados. Se describen estrategias a posteriori de la no respuesta y las implicaciones de su uso. Asimismo, se realiza una revisión de las tres reglas de submuestreo existentes y se define en cada una de ellas la fracción de muestreo en los no respondientes, su varianza y la esperanza de la varianza (error), a partir de esta revisión se propone una nueva regla de submuestreo. Por último, se realiza una simulación numérica y gráfica para evaluar y comparar la eficiencia de la nueva regla con las tres existentes.

2. Consecuencia de la no respuesta

Un conjunto de no respondientes puede ser grande o pequeño dependiendo del cuidado que se haya tenido en las distintas etapas de la planeación y del diseño muestral utilizado. A pesar de que los muestreadores utilicen distintas estrategias para reducir la no respuesta, al final, por lo general existirá. En este trabajo solo se considera este inconveniente como inherente al diseño de muestreo. Teniendo en cuenta lo anterior, si se utilizan estimadores insesgados para realizar la inferencia y solo se considera a los respondientes como buenos representantes de la muestra original s, se tendrá un sesgo en la estimación al no considerar la información de los no respondientes. De esto último se puede deducir que, en la presencia de la no respuesta, se tiene como consecuencia que no haya estimadores insesgados.

PRELIMINAR

2.1 Sesgo de la no respuesta

El principal problema de la no respuesta es el sesgo que genera [1]. Se considera una población finita $U = \{u_1, \dots, u_N\}$ de tamaño N, la cual se puede clasificar en dos estratos, aquellos que respondieron N_R y aquellos que no respondieron N_{NR}, con sus respectivas medias $\bar{y}_{RU}$ y $\bar{y}_{NRU}$. El promedio poblacional $\bar{y}_U$ está dado por:

$$\bar{y}_U = \frac{N_R}{N}\bar{y}_{RU} + \frac{N_{NR}}{N}\bar{y}_{NRU}$$

por lo que,

$$Sesgo = E[\bar{y}_R] - \bar{y}_U \approx \frac{N_{NR}}{N}(\bar{y}_{RU} - \bar{y}_{NRU}) \quad (1)$$

donde $\bar{y}_R$ representa un estimador aproximadamente insesgado de la media en el estrato N_R, utilizando solamente los datos de los respondientes. El sesgo expresado en (1) se reduce cuando $\bar{y}_{NRU}$ se aproxima a $\bar{y}_{RU}$, aunque no se conoce $\bar{y}_{NRU}$. Otra posibilidad es que el cociente $\frac{N_{NR}}{N}$ sea pequeño, esto es, que exista una cantidad pequeña de no respuesta.

2.2 Ausencia de estimadores insesgados ante la no respuesta en la práctica

Como se ha mencionado, una consecuencia de la presencia de no respuesta es la ausencia de estimadores insesgados, tal como lo muestran Särndal y colaboradores [4]. Para ello, denotan por θ_{qis} la probabilidad de que el *i-ésimo* elemento responda a la pregunta q dada la muestra s seleccionada. En primera instancia, se supone que la probabilidad θ_{qis} es conocida y positiva para todos los elementos i de la muestra s.

Con la notación en [4], para el total poblacional $t = \sum_U y_i$, un π-estimador natural $\hat{t}_\pi = \sum_S {y_i}/{\pi_i}$ $donde\ \pi_i = P_r(i \epsilon S)$ es la probabilidad de inclusión del $i - ésimo$ elemento. Sea $\pi_i^* = \pi_{ai}\pi_{i|S_a}$, el π^*- estimador es $\hat{t}_* = \sum_S {y_i}/{\pi_i^*} = \sum_S \breve{\breve{y}}_i$.

En el marco del muestreo de dos fases, para el total del q-ésimo ítem, $t_q = \sum_U y_{qi}$, se define un candidato a π^*-estimador insesgado dado en la expresión (2),

$$\hat{t}_{q\pi^*} = \sum_{r_q}\left(\frac{y_{qi}}{\pi_i}\right)\theta_{qis} \quad (2)$$

donde $r_q = \{i: i \epsilon s\ y\ y_{iq}$ es registrada$\}$.

Dado que $E(\hat{t}_{q\pi*}|s) = \sum_s {y_{qi}}/{\pi_i} = \hat{t}_{q\pi}$, donde $\hat{t}_{q\pi}$ es el estimador para la variable q dado π. Así se muestra que $\hat{t}_{q\pi*}$ es insesgado,

$$E(\hat{t}_{q\pi*}) = E_d E(\hat{t}_{q\pi*}|s) = E_d(\hat{t}_{q\pi}) = t_q \quad (3)$$

PRELIMINAR

donde $E_d(\cdot)$ denota la esperanza dado el diseño muestral *d* para obtener la muestra s.

El resultado dado en (3) parte del supuesto que las probabilidades θ_{qis} son conocidas. El recíproco, $1/\theta_{qis}$, en el estimador dado en (2) puede ser considerado como un ponderador para eliminar el sesgo de la no respuesta. Sin embargo, estas ponderaciones no son conocidas en la práctica y, en presencia de no respuesta, las técnicas usuales no funcionan para obtener estimadores insesgados.

Ante la presencia de no respuesta no se pueden encontrar estimadores insesgados, pero sí estimadores aproximadamente insesgados (con sesgo reducido). Para ello, es necesario disponer con estimaciones de las probabilidades de respuesta $\hat{\theta}_{qis}$, utilizando los datos disponibles de la encuesta, datos auxiliares y haciendo supuestos sobre la distribución de la respuesta.

2.3 Qué hacer con la no respuesta

Para disminuir el sesgo y evitar errores de muestreo derivados de la no respuesta, existen distintos métodos o estrategias de estimación. Se debe mencionar que estas consideraciones son a *posteriori* una vez que se presenta la no respuesta. Un método para reducir la no respuesta o el sesgo a *priori* es el uso de técnicas de respuestas aleatorizadas [5-7]. Este trabajo solo aborda la no respuesta una vez observada. En las siguientes líneas se enuncian las diferentes estrategias para tratar la no respuesta: ignorar la no respuesta, submuestreo en los no respondientes, métodos de imputación de datos y modelos de superpoblación e información auxiliar.

2.3.1 Ignorar la no respuesta

Una mala práctica al realizar muestreo por encuesta es ignorar la no respuesta. Esta práctica suele deberse al desconocimiento de sus implicaciones o por la reducción de costos en el proceso de realizar una encuesta. Difícilmente ignorar la no respuesta producirá estimaciones verdaderas, debido a que la distribución de los no respondientes puede ser distinta a la de los respondientes. Es decir, un estimador de la media poblacional que considera solo a los respondientes puede diferir de la media poblacional verdadera a consecuencia de esa subrepresentación ocasionada por la falta de información de los no respondientes. Por tanto, el procedimiento puede producir estimaciones que no reflejen adecuadamente la distribución poblacional.

Otra forma de ignorar la no respuesta es cuando se contempla el supuesto de que los datos faltantes, en la presencia de la no respuesta, están asociados a la aleatoriedad al considerar independientes las probabilidades de respuesta a los datos aún no observados. Este supuesto difícilmente se cumple [8], ya que las probabilidades de respuesta están correlacionadas con los resultados objetivos por lo que los datos faltantes no están faltando debido a la aleatoriedad.

PRELIMINAR

2.3.2 Submuestreo en los no respondientes

Ignorar la no respuesta no es una estrategia que genere resultados confiables o cercanos al valor verdadero. Hansen-Hurwitz [9] propusieron el primer método para tratar este problema, llamado submuestreo, con la finalidad de reducir el error de estimación. Para la descripción del método de submuestreo, se considera que se ha utilizado un muestreo aleatorio simple sin reemplazo para seleccionar la muestra *s* de la población finita $U = \{u_1, u_2, \dots, u_N\}$. Ante el problema de la no respuesta, la población puede ser particionada en dos estratos $U = U_1 + U_2$, tales que:

$$U_1 = \{u \epsilon U \mid u \text{ responde en la primera visita}\}$$

$$U_2 = \{u \epsilon U \mid u \text{ no responde en la primera visita}\}$$

donde, $|U_1| = N_1$, $|U_2| = N_2$ y $N = N_1 + N_2$.

La característica de interés *Y*, de la cual se quiere estimar la media poblacional $\bar{Y}$ mediante el promedio de los elementos en la muestra de tamaño *n* es,

$$\bar{y} = \frac{1}{n}\sum_{i \epsilon s} y_i \tag{4}$$

Cuando se planifica la encuesta se considera que $N_2 = 0$, pero esto usualmente no ocurre. Por ello, se modifica (4). Con este propósito, se particiona la muestra *s* en la forma correspondiente a la partición de la población *U*,

$$s_i = \{u \epsilon s | u \epsilon U_i\}, para\ i = 1,2$$

y se extrae una submuestra s_2' de s_2.

Bajo el supuesto de que todos los elementos de la submuestra s_2' den su respuesta, un estimador insesgado para $\bar{Y}$ es dado en (5),

$$\bar{y}' = w_1 \bar{y}_1 + w_2 \bar{y}_2' \tag{5}$$

donde $w_i = {}^{n_i}/_{n}$ con $|s_i| = n_i$ para $i = 1,2$. Las medias muestrales son,

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i \epsilon s_1} y_i \ , \quad \bar{y}_2' = \frac{1}{n_2'}\sum_{i \epsilon s_2'} y_i .$$

La varianza del estimador en (5) al fijar *s*, está dada por

$$V = V(\bar{y}') = \frac{N-n}{nN} S^2 + w_2^2 \frac{n_2 - n_2'}{n_2' n_2} S_2^2 \tag{6}$$

donde $S_2^2 = \frac{1}{N_2 - 1}\sum_{i \epsilon U_2}(y_i - \bar{Y}_2)^2$, con $\bar{Y}_2$ la media poblacional del estrato U_2.

PRELIMINAR

En esta situación se enfrenta al problema de determinar la fracción de submuestreo adecuada en el conjunto de los no respondientes. Para determinar el tamaño de la submuestra s_2' de s_2, esto es, determinar n_2', existen tres reglas clásicas que se describirán más adelante; además, se propondrá la extensión de una de ellas para mejorar su eficiencia.

2.3.3 Métodos de imputación de datos

Para resolver el problema de la no respuesta en preguntas de una encuesta, se han propuesto distintos métodos de imputación de datos con la finalidad de tener completa la matriz de datos. Se supone que la encuesta comprende q variables de interés $y_1, \dots, y_j \dots, y_q$, que corresponden a q preguntas. Sea s la muestra probabilística y r_j, $j = 1, \dots, q$ el conjunto de respuestas para la variable y_j. Entonces, la unión $r_u = r_1 \cup r_2 \cup \dots \cup r_q$ corresponde al conjunto de individuos que respondieron al menos una pregunta del cuestionario. Por otro lado, la intersección $r_c = r_1 \cap r_2 \cap \dots \cap r_q$ corresponde a los individuos que dieron respuesta a las q preguntas, esto es, que se tienen las respuestas del i-ésimo individuo $y_i = (y_{i1}, y_{i2}, \dots, y_{iq})$. En consecuencia, $r_u - r_c$ representan el conjunto de preguntas sin respuesta y $s - r_u$ representan a las unidades de no respuesta, se supone que ambos son no vacíos. Se denomina *enfoque de conjunto de respuesta* cuando el conjunto r_j es utilizado para la estimación de la j-ésima variable. En este enfoque, la estimación de los totales $t_j = \sum_U y_{ji}$, $j = 1,2, \dots, q$ se puede realizar pregunta por pregunta mediante el enfoque de ajuste ponderado [4].

La imputación de datos consiste en generar un valor $\tilde{y}_{ji}$ para un no respondiente y_{ji}, tal que $i \epsilon r_u - r_c \; y \; j = 1,2, \dots q$. Para este propósito existen diversos métodos que pueden emplear el conjunto de respuestas o información auxiliar. Entre los métodos de imputación más utilizados se encuentran la imputación deductiva, por media global, por media de celda, *hot deck*, *cold deck*, por regresión y la imputación múltiple, entre otros [10-11].

2.3.4 Modelo de superpoblación e información auxiliar

Los modelos de superpoblación permiten abordar el problema de la no respuesta en unidades completas, en contraste con la imputación de datos aplicada para la no respuesta parcial (algunas de las preguntas). Chaudhury y Stenger [12] tratan el problema de la no respuesta en unidades, mediante el uso de modelos superpoblacionales e información auxiliar.

Suponen que se cuenta con k variables auxiliares $x_1, x_2, \dots, x_k$, altamente correlacionadas con la variable de estudio Y. Los valores de las variables auxiliares constituyen la matriz X de dimensión $N \times k$ esto es,

$$X = [X_{ij}], i = 1,2, \dots, N, j = 1,2, \dots, k$$

PRELIMINAR

donde el *i-ésimo* renglón se denota por $x_i = (x_{i1}, x_{i2}, \dots, x_{ik}), i = 1,2, \dots, N$. Se selecciona una muestra s de tamaño n con probabilidad $p(s)$ y probabilidades de inclusión $\pi_i > 0$. Por tanto, se obtiene en cada muestra una submatriz X_s de dimensión $n \times k$ de X. Se supone que $n_1 < n$ elementos de la muestra s respondieron, por lo que se construye la submatriz X_r de $n_1 \times k$ de X_s.

El modelo establecido en [12] es $E_m(Y) = X\beta$, $V_m(Y) = \sigma^2 V$, donde $Y = (Y_1, Y_2, \dots, Y_n)'$, $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$, $\sigma^2 > 0$ y $V \epsilon M(N, N)$, es una matriz diagonal con σ^2 desconocida y V conocida.

Bajo el supuesto de respuesta completa, proponen el estimador,

$$t_s = \sum_{i=1}^{N} \hat{\mu}_i \tag{7}$$

donde $\hat{\mu}_i = x_i \hat{\beta}_s$, con $\hat{\beta}_s = (X_s' \pi_s^{-1} V_s^{-1} X_s)^{-1} (X_s' \pi_s^{-1} V_s^{-1} Y_s)$ y, $\pi_s = \text{diag}(\pi_i)$ para $i \epsilon s$; $V_s = \text{diag}(V_{ii})$ para $i \epsilon s$; $Y_s = \left(Y_{i_1}, Y_{i_2}, \dots, Y_{i_n}\right)'$ para $i \epsilon s$, se supone que existen todas las inversas de las matrices.

El estimador dado en (7), se puede expresar de la forma,

$$t_s = U_s' Y_s = \sum_{i \epsilon s} U_{si} Y_i \tag{8}$$

donde

$$U_s' = 1_N' X \left(X_s' \pi_s^{-1} V_s' X_s\right)^{-1} X_s' \pi_s^{-1} V_s^{-1}. \tag{9}$$

La expresión (9) es un vector n-dimensional con componentes U_{si}.

En la presencia de la no respuesta, esto es, que se tenga solamente una submuestra s_1 de tamaño $n_1 < n$, se emplea el estimador,

$$\tilde{t}_s = \sum_{i \epsilon s_1} U_{si} Y_i + \sum_{i \epsilon s - s_1} U_{si} \bar{Y}_i \tag{10}$$

Para calcular (10), es necesario construir las submatrices $X_{s_1}', \pi_{s_1}^{-1}, V_{s_1}^{-1}$ correspondientes a las matrices $X_s', \pi_s^{-1}, V_s^{-1}$ y el subvector Y_{s_1} del vector Y_s, eliminando las entradas de los elementos en $s - s_1$, para el cálculo de

$$\hat{\beta}_{s_1} = \left(X_{s_1}' \pi_{s_1}^{-1} V_{s_1}' X_{s_1}\right)^{-1} \left(X_{s_1}' \pi_{s_1}^{-1} V_{s_1}' Y_{s_1}^{-1}\right), \quad \hat{Y}_i = x_i \hat{\beta}_{s_1}$$

satisfaciendo la igualdad,

PRELIMINAR

$$\tilde{t}_s = \sum_{i \epsilon s_1} U_{s_1 i} Y_i = t_{s_1}$$

siendo $U_{s_1 i}$ el i-ésimo elemento del vector U'_{s_1} de dimensión $1 \times n_1$, dado por:

$$U'_{s_1} = 1'_N X \left(X'_{s_1} \pi_{s_1}^{-1} V'_{s_1} X_{s_1} \right)^{-1} X'_{s_1} \pi_{s_1}^{-1} V_{s_1}^{-1}.$$

Con ello, se obtiene un estimador $\tilde{t}_s$ de insesgadez asintótica mediante su propuesta de utilizar un modelo superpoblacional con información auxiliar fuertemente correlacionada con la variable de interés.

Los métodos o estrategias mencionadas anteriormente cumplen con su objetivo que es solventar la no respuesta. Su utilidad en la práctica dependerá de las necesidades del investigador o del contexto donde se sitúe la investigación, por ejemplo, cuando se tienen pocos datos faltantes o pocas entradas de no respuesta es mejor realizar una imputación de datos; cuando se pueden identificar los no respondientes, es adecuado realizar un submuestreo; si la no respuesta afecta unidades completas, es mejor utilizar modelos de superpoblación. Sin embargo, ignorar la no respuesta ante su presencia no es una estrategia para considerar, ya que la distribución de las unidades no respondientes puede diferir de la correspondiente a las unidades respondientes y esto ocasionaría estimaciones sesgadas a la realidad.

3. Reglas de submuestreo existentes

Se ha indicado que hay distintas estrategias para enfrentar la no respuesta una vez que se presenta. Sin embargo, no es conveniente utilizar métodos puramente matemáticos como la imputación de datos ya que ocasiona problemas metodológicos y estadísticos al ponderar matemáticamente datos o entradas faltantes, esta afirmación se fundamenta en razones similares a las expuestas para desaconsejar la estrategia de ignorar la no respuesta.

Dado que la estrategia de submuestreo es la única que considera el remuestreo en los no respondientes, esto conlleva ventajas sobre las otras estrategias. Si se logra respuesta completa en el submuestreo, se obtienen estimaciones que consideran toda la información contenida en la muestra, lo que permite obtener estimadores insesgados de los parámetros de la población de estudio. Por consiguiente, se abordará el submuestreo ante la presencia de no respuesta con las tres reglas existentes: Hansen-Hurwitz [9], Srinath [13] y Bouza [14]. Se describen las tres reglas para determinar el tamaño n'_2 de la submuestra s'_2.

3.1 Regla de Hansen-Hurwitz

La regla que proponen Hansen y Hurwitz [9] para determinar el tamaño de la submuestra s'_2 de s_2, está dada por la relación,

PRELIMINAR

$$n_2^{'} = \frac{n_2}{K} \tag{11}$$

con una constante $K > 1$, siguiendo la definición de Hansen-Hurwitz [9] y siendo fijado por el muestreador. Para analizar el efecto de esta regla, se sustituye la relación en (11) en la expresión de la varianza dada en (6),

$$V_{H-H}(\bar{y}^{'}) = \frac{N-n}{nN}S^2 + \frac{n_2(K-1)}{n^2}S_2^2. \tag{12}$$

El valor esperado de la varianza en (12) es,

$$E(V_{H-H}) = \frac{N-n}{nN}S^2 + W_2\frac{K-1}{n}S_2^2$$

3.2 Regla de Srinath

La regla de Srinath [13], para determinar el valor n_2' de la submuestra s_2' de s_2, en función de un parámetro *H*, está dada por:

$$n_2^{'} = \frac{n_2^2}{(Hn + n_2)}. \tag{13}$$

Los valores del parámetro *H* se encuentran en el intervalo $(0, \infty)$. De igual forma que en la regla anterior, se sustituye (13) en la varianza del estimador dada en (6) para obtener:

$$V_S(\bar{y}^{'}) = \frac{N-n}{nN}S^2 + \frac{H}{n}S_2^2.$$

Al sustituir el valor de n_2' se pierde la aleatoriedad de n_2 dependiendo solo del parámetro *H*, por lo que la esperanza de la varianza se mantiene sin cambios.

3.3 Regla de Bouza

Con base en la muestra *s*, la regla de Bouza [14] calcula el tamaño a tomar en los no respondientes con:

$$n_2' = n_2 w_2 \tag{14}$$

donde w_2 es la definida y utilizada en la expresión (6), siendo la ponderación de la muestra observada s_2.

En las dos reglas anteriores existe una dependencia del valor n_2' con respecto a los parámetros *K* y *H*, respectivamente. En Bouza [14] n_2' depende de los resultados de la encuesta.

Ahora se sustituye el valor de n_2' dado en (14) en la expresión de la varianza (6) dando como resultado:

PRELIMINAR

$$V_B(\bar{y}') = \frac{N-n}{nN}S^2 + \frac{n_1}{n^2}S_2^2$$

por lo que la varianza esperada bajo [14] es:

$$E(V_B) = \frac{N-n}{nN}S^2 + \frac{W_1}{n}S_2^2$$

3.4 Regla de submuestreo complementaria a la de Bouza

Como lo ha señalado Bouza [14], su regla minimiza el error de muestreo y que no dependa de la fracción de muestreo fijada por el muestreador, sino del tamaño en el estrato de los no respondientes. Sin embargo, su regla presenta una menor eficiencia cuando la no respuesta es de moderada a baja. Es decir, cuando $n_2 \leq n_1\, 0.5$ la precisión del estimador y la eficiencia de la varianza disminuyen cuando en el estrato de los no respondientes se submuestrea a pocos elementos.

Por lo anterior, se presenta una nueva regla complementaria a la de Bouza [14] con el propósito de mejorarla eficiencia y reducir el error de muestreo cuando hay moderada o poca no respuesta.

La regla propuesta modifica la fracción de submuestreo de Bouza [14] mediante la expresión:

$$n_2' = n_2 - \frac{n_2^2}{n}, \qquad n_2' < n_2\,. \tag{15}$$

Sustituyendo (15) en (6) se obtiene su varianza

$$V_J(\bar{y}') = \left(\frac{(N-n)}{nN}\right)S^2 + \frac{w_2^2}{n_1}S_2^2$$

Ahora se deduce la esperanza de $V_J(\bar{y}')$

$$E[V_J] = E\left[\left(\frac{(N-n)}{nN}\right)S^2 + \frac{w_2^2}{n_1}S_2^2\right] = \left(\frac{(N-n)}{nN}\right)S^2 + \frac{S_2^2}{n^2\, n_1}E[n_2^2]$$

$$= \left(\frac{(N-n)}{nN}\right)S^2 + \left(\frac{W_2W_1}{n\, n_1}\frac{N-n}{N-1} + \frac{W_2^2}{n_1}\right)S_2^2$$

y si se toma una N suficientemente grande, entonces $N-n \approx N-1$, por lo que la expresión del error sería,

$$\left(\frac{S^2}{n}\right) + \left(\frac{W_2W_1}{n\, n_1} + \frac{W_2^2}{n_1}\right)S_2^2$$

PRELIMINAR

Se aplicará la misma equivalencia a las demás reglas.

Dado el uso de muestreo simple aleatorio sin reemplazo n_2 es una variable con distribución hipergeométrica con esperanza y varianza

$$E[n_2] = n\,W_2$$

$$Var[n_2] = n\,W_2W_1\,\frac{N-n}{N-1}$$

Por tanto,

$$E\left[\frac{n_2^2}{n^2 n_1}\right] = \frac{1}{n^2 n_1}\left[Var(n_2) + \left(E(n_2)\right)^2\right] = \frac{W_2W_1}{n\,n_1}\frac{N-n}{N-1} + \frac{W_2^2}{n_1}.$$

4. Ganancia en precisión de las reglas con la propuesta

Para evaluar la precisión del tamaño de muestra en los no respondientes, y siguiendo el trabajo de Bouza [14], se comparan las varianzas y las esperanzas de las varianzas de las tres reglas con la propuesta.

Ganancia en precisión de Bouza [14] con la nueva regla considerando sus varianzas

$$G\left[\left(V_B(\bar{y}')\right),\left(V_J(\bar{y}')\right)\right] = V_B(\bar{y}') - V_J(\bar{y}') = \frac{N-n}{nN}S^2 + \frac{n_1}{n^2}S_2^2 - \left[\left(\frac{(N-n)}{nN}\right)S^2 + \frac{w_2^2}{n_1}S_2^2\right]$$

entonces,

$$\frac{n_1}{n^2} > \frac{w_2^2}{n_1} \Leftrightarrow \; n_1 > n_2$$

Esta desigualdad siempre será válida cuando en [14] se tome $n_1 > n_2$. Tomando esta fracción de muestreo de los no respondientes, la varianza es más pequeña usando la regla propuesta que la de Bouza [14].

Ahora comparando los errores, se tiene,

$$G\left[E(V_B), E\left[V_J\right]\right] = \frac{N-n}{nN}S^2 + \frac{W_1}{n}S_2^2 - \left[\left(\frac{(N-n)}{nN}\right)S^2 + \left(\frac{W_2W_1\frac{N-n}{N-1}}{n\,n_1} + \frac{W_2^2}{n_1}\right)S_2^2\right]$$

$$\frac{W_1}{n} > \frac{W_2W_1}{n\,n_1} + \frac{W_2^2}{n_1} \Leftrightarrow \; n_1 > n_2$$

De la misma forma sucede cuando se compara la varianza en la regla de Hansen-Hurwitz [9] con la propuesta, esto es,

$$G\left[V_{H-H}(\bar{y}'), V_J(\bar{y}')\right] = \frac{N-n}{nN}S^2 + \frac{n_2(K-1)}{n^2}S_2^2 - \left[\left(\frac{(N-n)}{nN}\right)S^2 + \frac{w_2^2}{n_1}S_2^2\right]$$

PRELIMINAR

$$\frac{n_2(K-1)}{n^2} > \frac{w_2^2}{n_1} \Leftrightarrow k > \frac{n_2}{n_1} + 1$$

y sus errores

$$G\left[E(V_{H-H}), E\left(V_J(\bar{y}')\right)\right] = \frac{N-n}{nN}S^2 + W_2\frac{K-1}{n}S_2^2 - \left[\left(\frac{(N-n)}{nN}\right)S^2 + \left(\frac{W_2W_1\frac{N-n}{N-1}}{n\,n_1} + \frac{W_2^2}{n_1}\right)S_2^2\right]$$

Cuando se considera la expresión

$$w_2\frac{k-1}{n} > \frac{w_2w_1\frac{N-n}{N-1}}{nn_1} + \frac{w_2^2}{n_1} \Leftrightarrow k > \frac{N-n}{N-1} + \frac{n_2}{n_1} + 1$$

como está definido para $k > 1$, de igual manera para esta desigualdad se cumple cuando $n_1 > n_2$.

Comparando la regla de Srinath [13] con la nueva,

$$G\left[V_S(\bar{y}'), V_J(\bar{y}')\right] = V_S(\bar{y}') - V_J(\bar{y}')$$

$$\Rightarrow \frac{H}{n} > \frac{w_2^2}{n_1} \Leftrightarrow H > \frac{n_2}{n_1}.$$

Como el término de la derecha de la desigualdad es un número menor que uno, esta desigualdad se mantendrá siempre y cuando $H > 0.5$, por lo que la propuesta es más eficiente comparada con la regla de Srinath [13] y con las mismas consideraciones de las reglas anteriores.

Y sus errores,

$$G\left[E\left(V_S(\bar{y}')\right), E\left(V_J(\bar{y}')\right)\right] = E\left(V_S(\bar{y}')\right) - E\left(V_J(\bar{y}')\right)$$

$$\Rightarrow \frac{H}{n} > \frac{W_2W_1\frac{N-n}{N-1}}{n\,n_1} + \frac{W_2^2}{n_1} \Leftrightarrow H > \frac{n_2}{n_1}$$

5. Simulación y resultados

Con el propósito de evaluar la eficiencia de las varianzas bajo cada regla, se realizó una simulación numérica con 1000 iteraciones y se promediaron las varianzas utilizando, $Ef = \left(\frac{V_l(\bar{y}')}{V_p(\bar{y}')}\right), l \neq p$. En esta simulación se consideró únicamente el coeficiente correspondiente a los no respondientes en cada varianza, ya que los demás términos se mantienen constantes. En la Tabla 1 se muestran las varianzas bajo cada regla y considerando una muestra de $n = 1000$ y valores mínimos en los parámetros *K* y H que

PRELIMINAR

hacen más eficientes esas reglas. De manera aleatoria se definió el tamaño de n_2 que bajo la teoría de submuestreo es la parte desconocida. Al fijar $n = 1000$ se pretende visualizar el comportamiento en el estrato de los no respondientes. Si el tamaño de la muestra se incrementa, el orden de eficiencia de las varianzas bajo cada regla no cambia, solo en el resultado numérico el exponente se incrementaría de manera negativa.

Tabla 1. Eficiencia de las reglas de submuestreo tomando como base n=1000 para distintos rangos de n_2.

		Varianzas				
Regla	$n = 1000$	$10 \leq n_2 \leq 100$	$10 \leq n_2 \leq 250$	$10 \leq n_2 \leq 400$	$10 \leq n_2 \leq 500$	$10 \leq n_2 \leq 990$
Hansen-Hurwitz	$k = 1.5$	0.000027	0.000065	0.000104	**0.000129**	**0.000253**
Srinath	*h=0.5*	0.0005	0.00050	0.00050	0.00050	0.00050
Hansen-Hurwitz	$k = 2.5$	0.000082	0.000196	0.000313	0.000389	0.000759
Srinath	*h=1.5*	0.0015	0.00150	0.00150	0.00150	0.00150
Hansen-Hurwitz	$k = 5$	0.00022	0.000523	0.000836	0.001038	0.002024
Srinath	*h=3*	0.003	0.00300	0.00300	0.00300	0.00300
Hansen-Hurwitz	$k = 10$	0.000495	0.001177	0.001881	0.002335	0.004556
Srinath	*h=6*	0.006	0.00600	0.00600	0.00600	0.00600
Bouza		0.000944	0.000869	0.000790	0.000740	0.000493
Propuesta		**0.000004**	**0.000027**	**0.000081**	0.000142	0.003416

Lo observado en la Tabla 1 muestra que la regla propuesta es más eficiente cuando la tasa de no respuesta es baja o moderada. En estas condiciones, a medida que aumentan los parámetros K y H de las reglas de Hansen-Hurwitz [9] y Srinath [13], respectivamente, estas son menos eficientes que la propuesta cuando hay de moderada a baja no respuesta. Cuando la tasa de no respuesta es elevada, la regla propuesta es menos eficiente que la de Bouza [14]. Este resultado indica que la regla propuesta es complementaria a la de Bouza.

Para evaluar la eficiencia de las reglas junto con la propuesta, se realiza una simulación gráfica para visualizar el comportamiento de ellas y cómo la nueva regla complementa de manera eficiente a la de Bouza [14]. Se graficaron las varianzas de las reglas incrementando los no respondientes.

PRELIMINAR

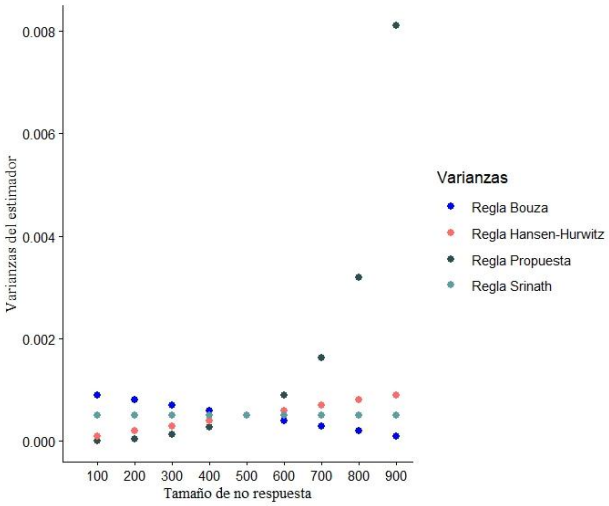


Figura 1. Varianzas bajo las reglas.

La Figura 1 concuerda con lo visto en la simulación numérica. Es decir, la nueva regla es más eficiente cuando se tiene de moderada a baja no respuesta, a su vez, la regla de Bouza [14] es más eficiente, y se complementa con la propuesta, cuando hay mucha no respuesta.

6. Discusión y conclusión

En este trabajo se realizó un estudio de las consecuencias de la no respuesta y qué hacer ante su presencia. También se analizaron las reglas clásicas de submuestreo, cuyo estudio fue iniciado por Hansen y Hurwitz [9] y ampliada por Srinath [13] y Bouza [14], teniendo como resultado una nueva regla de submuestreo más eficiente.

Las reglas de Hansen-Hurwitz [9] y Srinath [13] tienen parámetros los cuales el investigador deberá fijar. Para realizar esta acción debe considerar factores subjetivos, económicos o logísticos, entre otros, y dependiendo de la circunstancia y el valor fijado para tales parámetros, puede que beneficie o perjudique las estimaciones. Por ello, la regla de Bouza [14] y la propuesta no consideran esos factores subjetivos, sino que el tamaño a remuestrear va en función de los respondientes con lo que se busca obtener mejores estimaciones, como se observa en la simulación numérica y gráfica.

La nueva regla es más eficiente que las otras reglas comparadas cuando el tamaño de los no respondientes es de moderado a bajo; esto representa una ventaja importante sobre las otras reglas y sobre la imputación de datos. Ante esta situación, es preferible utilizar el submuestreo con la regla propuesta que recurrir a la imputación de datos, bajo los supuestos estudiados, debido a que hay métodos de imputación que sesgan la información al ponderarla.

PRELIMINAR

Las reglas clásicas [9,13,14] tienen un comportamiento estable de eficiencia cuando hay poca, moderada o mucha no respuesta. Por otro lado, la propuesta no presenta esa estabilidad y no se recomienda utilizarla cuando la no respuesta es alta. Sin embargo, ante la presencia de elevada no respuesta en s, el investigador tendrá que replantearse por razones económicas si es conveniente usar submuestreo para enfrentarla. Debido a que el costo de la encuesta podría aproximarse al doble de lo planteado al inicio de la investigación ya que prácticamente se volvería a encuestar a la mayoría de los elementos de la muestra.

La mayor ventaja que presenta la nueva regla sobre todas las demás, se debe a que es más eficiente cuando se tiene moderada o poca no respuesta, a saber, que el número de no respondientes sea menor al 50 % del tamaño muestral como se observó en los resultados de la simulación y se mostró en la ganancia de la precisión entre las reglas. Esto es de gran importancia ya que en la práctica la proporción de no respuesta suele ser relativamente baja, son casos puntuales que representan un porcentaje mínimo del tamaño de la muestra. En el caso de que $n_2 > n_1$, el investigador debe considerar que existe un error metodológico (mala capacitación de los encuestadores, mala elección del diseño muestral, instrumento de recolección de datos inadecuado, no se cuenta con un marco muestral bueno, etc.) en la planeación de la encuesta que resultaría en reevaluar alguna o todas sus fases. Se recuerda al lector que en este trabajo la no respuesta se considera inherente al diseño de muestreo y no debido a la mala planeación en alguna de las fases de la encuesta, ya que esto aborda distintos temas y no solo el diseño muestral.

Como trabajos subsecuentes para reducir la no respuesta y el sesgo en la respuesta se puede estudiar conjuntamente las reglas de submuestreo y las técnicas de respuesta aleatorizada. Para esto, se debe submuestrear en los no respondientes y a ellos se les pedirá enmascarar su respuesta verdadera, generalmente sensible, con una técnica de respuesta aleatorizada.

En conclusión, el submuestreo ante la no respuesta convierte la problemática del sesgo en una labor controlada dentro del diseño probabilístico, asegurando que se siga reflejando fielmente la diversidad y las características de toda la población objetivo. La regla propuesta demuestra ser capaz de eliminar la dependencia de parámetros fijados por el muestreador, ofreciendo una formulación más objetiva. Los resultados numéricos y gráficos indican que la nueva regla presenta menor varianza y error de muestreo que las reglas clásicas cuando la no respuesta es baja o moderada. El estudio no busca reemplazar del todo las contribuciones clásicas, sino integrar un esquema complementario. Mientras que la regla de Bouza [14] sigue siendo indispensable en presencia de niveles altos de no respuesta, la regla formulada en este trabajo es más eficiente y económica cuando la no respuesta es poca o moderada.

PRELIMINAR

Referencias

1. S. L. Lohr, Sampling: Design and Analysis. Boca Raton, FL, EE. UU.: Chapman and Hall/CRC, 2021. https://doi.org/10.1201/9780429298899
2. S. C. Gupta y V. K. Kapoor, Fundamentals of Mathematical Statistics. Nueva Delhi, India: Sultan Chand & Sons, 2020.
3. W. G. Cochran, Técnicas de Muestreo. Nueva York, NY, EE. UU.: John Wiley & Sons, Inc., 1971.
4. C. E. Särndal, B. Swensson y J. Wretman, Model Assisted Survey Sampling. Nueva York, NY, EE. UU.: Springer-Verlag, 1992. https://doi.org/10.1007/978-1-4612-4378-6
5. M. Azeem, "Introducing a weighted measure of privacy and efficiency for comparison of quantitative randomized response models," Pakistan Journal of Statistics, vol. 39, no. 3, pp. 377–385, 2023.
6. A. Ioannidis, A. Litke y N. K. Papadakis, "A randomized response framework to achieve differential privacy in medical data," Electronics, vol. 14, no. 326, 2025. https://doi.org/10.3390/electronics14020326
7. C. N. Bouza-Herrera, P. O. Juárez-Moreno, A. Santiago-Moreno y J. M. Sautto-Vallejo, "A quantitative compulsory randomized response technique. Simulation with sensitive data on violence in Mexico," Investigación Operacional, vol. 46, no. 4, pp. 555–570, 2025.
8. D. Marella y D. Pfeffermann, "Accounting for Non-ignorable Sampling and Non-response in Statistical Matching," International Statistical Review, vol. 91, no. 2, pp. 269-293, 2022. https://doi.org/10.1111/insr.12524
9. M. H. Hansen y W. N. Hurwitz, "The problem of non-responses in survey sampling," Journal of the American Statistical Association, vol. 41, no. 236, pp. 517–529, 1946. https://doi.org/10.1080/01621459.1946.10501894
10. R. J. A. Little y D. B. Rubin, Statistical Analysis with Missing Data, 3a ed. Hoboken, NJ, EE. UU.: Wiley, 2019. https://doi.org/10.1002/9781119482260
11. C. N. Bouza y C. Viada, "Imputation of individual values of a variable using product predictors," Investigación Operacional, vol. 42, no. 3, pp. 321–325, 2021.
12. A. Chaudhury y H. Stenger, Survey Sampling: Theory and Methods. Boca Raton, FL, EE. UU.: CRC Press, 2005. https://doi.org/10.1201/9781420028638
13. K. P. Srinath, "Multiphase sampling in nonresponse problems," Journal of the American Statistical Association, vol. 66, no. 335, pp. 583–586, 1971. https://doi.org/10.1080/01621459.1971.10482310
14. C. N. Bouza, "Sobre el problema de la fracción de submuestreo para el caso de las no respuestas," Trabajos de Estadística y de Investigación Operativa, vol. 32, no. 2, pp. 30–36, 1981. https://doi.org/10.1007/BF02888736

PRELIMINAR