MUESTREO DE RESPUESTA ALEATORIZADAS CON PROBABILIDADES DESIGUALES: EL ESTIMADOR DE RAO-HARTLEY-COCHRAN

Víctor H. Soberanis Cruz[1]-, Jaime D. Cuevas Domínguez
Universidad de Quintana Roo, Colonia del Bosque, Chetumal, Quintana Roo, México, CP 77000.

ABSTRACT
The technique of Randomized Response has been introduced to reduce the risk of evasive answer or no-response in survey samplings of sensitive issues to estimate the total of individuals with that sensitive characteristic. We also know that highly efficient strategies of estimation require strong auxiliary information. Hence, in this article we obtain and compare two estimators: The Rao-Hartley-Cochran (RHC) estimator with Warner´s model and the RHC estimator with Greenberg´s model, both in a finite population setting.
We will show via simulation that $\hat{t}_{A1,RHC,\pi}$, the RHC π-estimator in Warner´s model is less efficient than $\hat{t}_{A2,RHC,\pi}$, the RHC π -estimator in Greenberg's model

KEY WORDS: π-Estimators, Randomized Response, Sensitive Question, Warner's Technique, RHC scheme.

MSC: 62D05

RESUMEN
La técnica de Respuestas Aleatorizadas (RR) ha sido introducida para reducir el riesgo de evasión o no respuesta de preguntas sensitivas en encuestas con muestreo para estimar el total de individuos con alguna característica sensitiva. Reconociendo que las estrategias de alta eficiencia en la estimación para estos casos requieren de variables auxiliares fuertes, en este artículo se propone al correspondiente estimador de Rao-Hartley-Cochran (RHC) tanto para el Modelo de Warner como para el Modelo de Greenberg.
Se muestra, mediante simulación, que el estimador que obtenemos de RHC en el Modelo de Greenberg resulta más eficiente que el del Modelo de Warner.

1. INTRODUCCIÓN

En estudios de asuntos sensitivos usando muestreo de encuestas, por ejemplo, uso de drogas, evasión de impuestos, preferencias sexuales, honestidad en exámenes finales en estudiantes universitarios, opinión respecto a autoridades, etc., muchos entrevistados rehúsan a participar en la encuesta o proporcionan respuestas falsas o respuestas condicionadas [10], ocasionando que la precisión y confiabilidad de los estimadores se alteren de un manera importante. La técnica de Respuesta Aleatorizada (RR) introducida por Warner [11] es una posible solución para la protección del anonimato del entrevistado.
Warner [11] desarrolló una técnica llamada Respuestas Aleatorizadas la cual garantiza el anonimato del entrevistado y consiste en un mecanismo aleatoria que selecciona una de dos preguntas complementarias, siendo la pregunta1 "¿Tienes la característica A?" y siendo la pregunta2 "¿Tienes la característica $\overline{A}$?, donde A es la característica sensitiva en estudio y $\overline{A}$ es la ausencia de tal característica sensitiva. El entrevistado contestará SI o NO y el entrevistador nunca sabrá que pregunta contestó el entrevistado, y es en este sentido que se protege el anonimato del entrevistado. Greenberg et.al [2] proponen como alternativa al modelo de Warner[11] de preguntas complementarias un modelo de respuestas aletorizadas pero con dos preguntas no relacionadas, donde la pregunta no relacionada es inocua. La comparación de estos dos modelos se ha hecho en el marco de Poblaciones Infinitas así como la determinación de la(s) probabilidad(es) óptima(s) en la

[1] vsobera@uqroo.mx, jaicueva@uqroo.mx)

selección de las preguntas en el mecanismo aleatorio . Gupta y Shabir [3] (2004) discuten la estimación simultanea del nivel de sensibilidad de la pregunta sensitiva y la de la respuesta promedio de la característica sensitiva pero en el marco de poblaciones infinitas. Por otra parte, Naddeo y Pisani [6] (2005) aunque discuten el problema de la detectabilidad imperfecta para estimar totales usando Muestreo Adaptativo por Conglomerados y una aproximación basada en puro diseño, tiene el inconveniente de los conteos replicados en dos etapas del Muestro Adaptativo, ya que las preguntas son sensitivas. Tenemos también la propuesta de Jea-Bok Ryu, *et.al* [4] (2005) que utiliza muestreo estratificado más el modelo de Mangat y Singh, y aunque mejora los modelos de muestreo estratificado con asignación proporcional y con asignación optima, tiene el inconveniente de ser un muestreo con reemplazo. Gjestvang y Singh [1] (2006) consideran un modelo de respuestas aleatorizadas que, en efecto, resulta más eficiente que los modelos de Warner, Mangat y Singh, y el de Mangath entre otros; pero la introducción de los parámetros de su modelo no resulta de fácil implementación y por otro lado lo desarrollan en el marco de poblaciones infinitas. El trabajo de Saha [] (2006) considera un esquema Respuestas Aleatorizadas en dos etapas para poblaciones finitas y además propone condiciones bajo las cuales su modelo produce estimaciones más precisas de los parámetros con respecto a los modelos de Warner, Mangat y Singh, y Mangat; sin embargo no considera la introducción de variables auxiliares para más aún incrementar la eficiencia de sus estimaciones.
Uno de los objetivos de este trabajo es la construcción del Estimador de Rao-Hartley-Cochran [7] tanto en el Método de Preguntas Complementarias de Warner [11] como en el Método que Preguntas No-relacionadas de Greenberg [2]. El otro objetivo de nuestro trabajo será la comparación de estos estimadores en el esquema del Muestreo Aleatorio Simple. El desarrollo de los estimadores a comparar lo haremos con base en la Teoría de los estimadores-π [9].

2. PREGUNTAS COMPLEMENTARIAS (MODELO DE WARNER)

2.1. El procedimiento de muestreo

Consideremos una población consistente de N elementos y para simplificar, el k-ésimo elemento de la población será representado por su etiqueta k. De manera que denotaremos a la población finita como

$$U = \{1,2,\dots,N\}$$

En este trabajo el tamaño de la población N se supondrá conocido.
Sea *y* una variable que mide alguna característica sensitiva, y sea y_k el valor de *y* para el k-ésimo elemento de la población. Así y_k es desconocida pero no aleatoria [12]. Además $y_k = 1$ si el k-ésimo individuo de la población tiene la característica sensitiva **A**, y $y_k = 0$ si el k-ésimo individuo no tiene la característica sensitiva **A**. Lo que se desea es estimar $t_A = \sum_U y_k$, total de los individuos en la población con la característica sensitiva **A.**
Sea p_t la probabilidad de extraer a la t-ésima unidad en la primera extracción de la población U. Por ejemplo, si estamos muestreando con probabilidades proporcionales al tamaño de x_t, $p_t = \frac{x_t}{\sum_{k=1}^{N} x_k}$. El procedimiento de muestreo consiste de las siguientes tres etapas [3]:

(a) Particionamos la población aleatoriamente en n grupos de tamaños $N_1, N_2, \dots, N_n \ \ donde \ \ N_1 + N_2 + \cdots + N_n = N$, $con\ n\ el\ tamaño\ de\ la\ muestra\ y\ no\ necesariamente\ n/N.$

(b) Extraer una muestra de tamaño uno con probabilidad p_t de cada une de los grupos independientemente.

(c) La metodología de Warner [11] es como sigue: Se implementa un mecanismo aleatorio (**RC**) tal que con probabilidad *P* el entrevistado selecciona la proposición Q_A: Tengo la característica **A** y con probabilidad *1-P* selecciona la proposición complementaria $Q_{\bar{A}}$: No tengo la característica **A.** El entrevistado responde "Si"(=1) o "No"(=0) sin que el entrevistador sepa que pregunta se esta contestando y de este modo se garantiza el anonimato del entrevistado.

2.2. El estimador propuesto y su varianza

Para cualquier elemento k, se define

$$Z_k = \begin{cases} y_k & con\ prob.\ \ P; \\ 1 - y_k & con\ prob.\ 1 - P \end{cases}$$

Así que

$$E_{RC}(Z_k) = \mathrm{y_k P} + (1 - \mathrm{y_k})$$

Asumiendo que P ≠ 1/2, se define [12]

$$\hat{Z}_k = \frac{Z_k + P - 1}{2P - 1}$$

Para cada k, $\hat{Z}_k$ es insesgado para y_k con respecto al mecanismo aleatorio RC; esto es, $E_{RC}(\hat{Z}_k) = \mathrm{y_k}$ y la varianza es:

$$V_{RC}(\hat{Z}_k) = \frac{\mathrm{P(1-P)}}{(2\mathrm{P}-1)^2} = \mathrm{V_0}$$

El mecanismo de selección de la muestra propuesto por Rao-Hartley-Cochran [7] intenta la selección de una muestra de tamaño *n* con probabilidades proporcionales a una variable *x*, sin reemplazo. El uso de esta variable auxiliar (x_t) aumenta la eficiencia del muestreo. La población U se particiona aleatoriamente en *n* subgrupos disjuntos por pares $U_i,\ i = 1,2,\dots,n$. Así, U_1 es una muestra aleatoria simple de *U*, U_2 es una muestra aleatoria simple de $U - U_1$, y así sucesivamente. Sea N_i el tamaño del grupo U_i. La muestra se selecciona como sigue. Un solo elemento se selecciona de cada grupo con probabilidad proporcional al tamaño de *x*: Al elemento *k* en el grupo U_i se le da la probabilidad $\frac{x_k}{N_i \bar{x}_{U_i}}, donde\ \bar{x}_{U_i} = \frac{\sum_{k\in U_i} x_k}{N_i}$. Sea k_i el elemento seleccionado del grupo U_i . El estimador de Rao-Hartley-Cochran para $t_A = t_y = \sum_U y_k$ viene dado por

$$\hat{t}_{A1,RHC} = \sum_{i=1}^{n} \frac{\hat{Z}_{k_i}}{p_{k_i}} \pi_i$$

Donde $p_{k_i} = \frac{x_{k_i}}{t_x}, t_x = \sum_{i=1}^{n}\sum_{t=1}^{N_i} x_{it},\ \pi_i = \frac{\sum_{t=1}^{N_i} x_{it}}{t_x}, i = 1,2,\dots,n.$ La insesgades de $\hat{t}_{A1,RHC}$ se obtiene de la insesgades de $\hat{t}_{Y,RHC} = \sum_{i=1}^{n} \frac{y_{k_i}}{p_{k_i}} \pi_i$. De hecho

$$E_{RC}(\hat{t}_{A1,RHC}) = \mathrm{E_{RC}}\left(\sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\hat{\mathrm{Z}}_\mathrm{i}}{\mathrm{p_i}/\pi_\mathrm{i}}\right) = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\mathrm{E_{RC}}(\hat{\mathrm{Z}}_\mathrm{i})}{\mathrm{p_i}/\pi_\mathrm{i}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \frac{\mathrm{y_i}}{\mathrm{p_i}/\pi_\mathrm{i}}$$

De modo que

$$\begin{aligned} E(\hat{t}_{A1,RHC}) &= E_1\{E_2[E_{RC}(\hat{t}_{A1,RHC})]\} \\ &= E_1E_2\left(\sum_{i=1}^{n} \frac{y_i}{p_i/\pi_i}\right) = Y = t_A. \end{aligned}$$

Para el cálculo de la varianza definimos

$$I_{G_{i;t}} = \begin{Bmatrix} 1 & si\ G_i \ni t; \\ 0 & de\ otra\ forma \end{Bmatrix}$$

y

$$I_{G_{i;t,t'}} = \begin{Bmatrix} 1 & si\ G_i \ni t, t'; \\ 0 & de\ otra\ forma \end{Bmatrix}$$

De manera que

$$E_1\left(I_{G_{i;t}}\right) = Pr\left\{I_{G_{i;t}} = 1\right\} = \frac{N_i}{N}$$

y

$$E_1\left(I_{G_{i;t,t'}}\right) = Pr\left\{I_{G_{i;t,t'}} = 1\right\} = Pr\{G_i \ni t \& t'\} = \frac{N_i(N_i - 1)}{N(N-1)}$$

Se tiene

$$V(\hat{t}_{A1,RHC}) = E_1\{E_2[V(\hat{t}_{A1,RHC})]\} + E_1\{V_2[E_{RC}(\hat{t}_{A1,RHC})]\} + V_1\{E_2[E_{RC}(\hat{t}_{A1,RHC})]\}$$
$$= E_1\{V_2[E_{RC}(\hat{t}_{A1,RHC})]\} + V_1\{E_2[E_{RC}(\hat{t}_{A1,RHC})]\} + E_1\{E_2[V_{RC}(\hat{t}_{A1,RHC})]\}$$
$$= V\left(\sum_{i=1}^{n} \frac{y_i}{p_i/\pi_i}\right) + E_1\{E_2[V_{RC}(\hat{t}_{A1,RHC})]\}$$

Ahora

$$V_{RC}(\hat{t}_{A1,RHC}) = \sum_{i=1}^{n} \frac{V_{RC}(\hat{Z}_i)}{\left(p_i/\pi_i\right)^2} = \sum_{i=1}^{n} \frac{V_0}{\left(p_i/\pi_i\right)^2}$$

De modo que

$$E_2[V_{RC}(\hat{t}_{A1,RHC})] = V_0 \sum_{i=1}^{n} E_2\left[\frac{1}{\left(p_i/\pi_i\right)^2}\right] = V_0 \sum_{i=1}^{n}\left\{\sum_{t=1}^{N_i} \frac{\pi_i^2}{p_t^2}\frac{p_t}{\pi_i}\right\} = V_0 \sum_{i=1}^{n}\left\{\sum_{t=1}^{N_i} \frac{\pi_i}{p_t}\right\} = V_0 \sum_{i=1}^{n} \pi_i \sum_{t=1}^{N_i} \frac{1}{p_t}$$
$$= V_0 \sum_{i=1}^{n}\left(\sum_{t=1}^{N} p_t I_{G_i;t}\right)\left(\sum_{t'=1}^{N} \frac{1}{p_{t'}} I_{G_i;t'}\right) = V_0 \sum_{i=1}^{n}\left(\sum_{t=1}^{N} I_{G_i;t} + \sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}} I_{G_i;t,t'}\right)$$
$$= V_0 \sum_{i=1}^{n}\left(N_i + \sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}} I_{G_i;t,t'}\right) = V_0\left[N + \sum_{i=1}^{n}\left(\sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}} I_{G_i;t,t'}\right)\right]$$

Por tanto

$$E_1\{E_2[V_{RC}(\hat{t}_{A1,RHC})]\} = V_0\left[N + \sum_{i=1}^{n} E_1\left(\sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}} I_{G_i;t,t'}\right)\right] = V_0\left[N + \sum_{i=1}^{n}\left(\frac{N_i(N_i-1)}{N(N-1)}\sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}}\right)\right]$$
$$= V_0\left[N + \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)}\sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}}\right]$$

Finalmente nos queda

$$V(\hat{t}_{A1,RHC}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)}\left(\sum_{t=1}^{N} \frac{y_t^2}{p_t} - Y^2\right) + V_0\left[N + \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)}\sum_{t=1}^{N}\sum_{\substack{t'=1 \\ t'\neq t}}^{N} \frac{p_t}{p_{t'}}\right].$$

2.3. Estimador de la varianza

Un estimador insesgado para la varianza de $\hat{t}_{A1,RHC}$ viene dado por

$$v(\hat{t}_{A1,RHC}) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2}\left[\left(\sum_{i=1}^{n} \frac{\hat{Z}_i}{p_i^2/\pi_i} - \hat{t}_{A1,RHC}^2\right) + 2V_0 \sum_{i=1}^{n} \frac{1}{\left(p_i/\pi_i\right)^2}\right]$$

En efecto

$$V_{RC}(\hat{t}_{A1,RHC}) = E_{RC}(\hat{t}_{A1,RHC}^2) - \left(\sum_{i=1}^{n} \frac{y_i}{p_i/\pi_i}\right)^2$$

$$\Rightarrow \qquad E_{RC}(\hat{t}_{A1,RHC}^2) = \hat{Y}_{RHC}^2 + \quad V_0 \sum_{i=1}^{n} \frac{1}{\left(p_i/\pi_i\right)^2}$$

De modo que

$$E_{RC}\left(v(\hat{t}_{A1,RHC})\right) = \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2}.\left[\left(\sum_{i=1}^{n} \frac{y_i}{p_i^2/\pi_i} - \hat{Y}_{RHC}^2\right) - V_0 \sum_{i=1}^{n} \frac{1}{\left(p_i/\pi_i\right)^2} + 2V_0 \sum_{i=1}^{n} \frac{1}{\left(p_i/\pi_i\right)^2}\right]$$
$$= \frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2}\left[\left(\sum_{i=1}^{n} \frac{y_i}{p_i^2/\pi_i} - \hat{Y}_{RHC}^2\right) + V_0 \sum_{i=1}^{n} \frac{1}{\left(p_i/\pi_i\right)^2}\right]$$

Así

$$E\left(v\left(\hat{t}_{A1,RHC}\right)\right) = E_1E_2\left\{\frac{\sum_{i=1}^{n} N_i^2 - N}{N^2 - \sum_{i=1}^{n} N_i^2}\left[\left(\sum_{i=1}^{n}\frac{y_i}{p_i^2/\pi_i} - \hat{Y}_{RHC}^2\right) + V_0\sum_{i=1}^{n}\frac{1}{\left(p_i/\pi_i\right)^2}\right]\right\}$$

$$= V\left(\hat{Y}_{RHC}\right) + V_0E_1E_2\left(\sum_{i=1}^{n}\frac{1}{\left(p_i/\pi_i\right)^2}\right) = V\left(\hat{Y}_{RHC}\right) + V_0\left[N + \left(\frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)}\right)\sum_{t=1}^{N}\sum_{t\neq t'}^{N}\frac{p_t}{p_{t'}}\right] \quad = V\left(\hat{t}_{A1,RHC}\right).$$

3. TÉCNICA DE PREGUNTAS NO RELACIONADAS CON $t_W = \sum_U w_k$ CONOCIDO.

3.1 El procedimiento de muestreo, el estimador y su varianza

El procedimiento de muestreo es como en 2.1 excepto que en lugar de (c) tenemos:

(d) Queremos estimar $t_A = \sum_U y_k$ considerando que $t_W = \sum_U w_k$, total de la característica **W no relacionada** con la característica sensitiva ***A,*** es conocida. El mecanismo aleatorio RC [1] será tal que Q_A es elegido con probabilidad P y Q_W con probabilidad 1-P. Es importante señalar que la característica **W** es no sensitiva, es innocua.

Definimos

$$Z_k = \begin{Bmatrix} y_k & con\ probabilidad\ P \\ w_k & con\ probabilidad\ 1-P \end{Bmatrix}$$

Por lo tanto

$$E_{RC}(Z_k) = y_kP + w_k(1-P) \quad \equiv \theta_k$$

También

$$V_{RC}(Z_k) = \theta_k(1-\theta_k) = P(1-P)d_k$$

donde $d_k = \begin{Bmatrix} 0 & si\ \ y_k = w_k \\ 1 & de\ otro\ modo \end{Bmatrix}$.

Si $t_\theta = \sum_{k\in U}\theta_k$, entonces $t_\theta = Pt_A + (1-P)t_W$ y $t_A = \frac{t_\theta - (1-P)t_W}{P}$.

El estimador-RHC para t_θ es:

$$\hat{t}_{\theta,RHC} = \sum_{i=1}^{n}\frac{Z_{k_i}}{p_{k_i}}\pi_i$$

El estimador $\hat{t}_{\theta,RHC}$ es insesgado:

$$E\left(\hat{t}_{\theta,RHC}\right) = E_1E_2E_{RC}\left(\sum_{i=1}^{n}\frac{Z_{k_i}}{p_{k_i}}\pi_i\right)$$

$$= E_1E_2\left(\sum_{i=1}^{n}\frac{\theta_{k_i}}{p_{k_i}}\pi_i\right) \quad = t_\theta.$$

También

$$V_{RC}\left(\hat{t}_{\theta,RHC}\right) = V_{RC}\left(\sum_{i=1}^{n}\frac{Z_i}{p_i/\pi_i}\right) = \sum_{i=1}^{n}\frac{V_{RC}(Z_i)}{\left(p_i/\pi_i\right)^2} \quad = P(1-P)\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}$$

De modo que

$$V\left(\hat{t}_{\theta,RHC}\right) = E_1V_2E_{RC}\left(\hat{t}_{\theta,RHC}\right) + V_1E_2E_{RC}\left(\hat{t}_{\theta,RHC}\right) + E_1E_2V_{RC}\left(\hat{t}_{\theta,RHC}\right)$$

$$= E_1V_2\left(\sum_{i=1}^{n}\frac{\theta_i}{p_i/\pi_i}\right) + V_1E_2\left(\sum_{i=1}^{n}\frac{\theta_i}{p_i/\pi_i}\right) + E_1E_2\left[P(1-P)\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}\right]$$

$$= \mathrm{V}\left(\sum_{i=1}^{n}\frac{\theta_i}{p_i/\pi_i}\right) + E_1E_2\left[P(1-P)\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}\right] = \frac{\sum_{i=1}^{n} N_i^2 - N}{N(N-1)}\left(\sum_{t=1}^{N}\frac{\theta_t^2}{p_t} - t_\theta^2\right) + E_1E_2\left[P(1-P)\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}\right].$$

Pero

$$E_1E_2\left[\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}\right]=E_1\left\{\sum_{i=1}^{n}E_2\left[\frac{d_i}{\left(p_i/\pi_i\right)^2}\right]\right\}=E_1\left(\sum_{i=1}^{n}\sum_{t=1}^{N_i}\frac{\pi_i^2}{p_t^2}d_i\frac{p_t}{\pi_t}\right)=E_1\left(\sum_{i=1}^{n}\sum_{t=1}^{N_i}\frac{d_t}{p_t/\pi_i}\right)$$

$$=E_1\left(\sum_{i=1}^{n}\pi_i\sum_{i=1}^{N_i}\frac{d_t}{p_t}\right)=E_1\left\{\sum_{i=1}^{n}\left(\sum_{t=1}^{N}p_t\,I_{G_i;t}\right)\left(\sum_{t=1}^{N}\frac{d_{i'}}{p_{t'}}I_{G_i;t}\right)\right\}$$

$$=E_1\left\{\sum_{i=1}^{n}\left(\sum_{t=1}^{N}d_iI_{G_i;t}+\sum_{t=1}^{N}\sum_{\substack{t'=1\\t'\neq t}}^{N}\frac{p_t}{p_{t'}}d_{t'}I_{G_i;t,t'}\right)\right\}$$

$$=\sum_{i=1}^{n}\sum_{t=1}^{N}d_t\frac{N_i}{N}+\sum_{i=1}^{n}\frac{N_i(N_i-1)}{N(N-1)}\left(\sum_{t=1}^{N}\sum_{\substack{t'=1\\t'\neq t}}^{N}\frac{p_t}{p_{t'}}d_{t'}\right)=\sum_{t=1}^{N}d_t+\frac{\sum_{i=1}^{n}N_i^2-N}{N(N-1)}\left(\sum_{t=1}^{N}\sum_{\substack{t'=1\\t'\neq t}}^{N}\frac{p_t}{p_{t'}}d_{t'}\right)$$

Así

$$V\left(\hat{t}_{\theta,RHC}\right)=\frac{\sum_{i=1}^{n}N_i^2-N}{N(N-1)}\left(\sum_{t=1}^{N}\frac{\theta_t^2}{p_t}-t_\theta^2\right)+P(1-P)\left[\sum_{t=1}^{N}d_t+\frac{\sum_{i=1}^{n}N_i^2-N}{N(N-1)}\left(\sum_{t=1}^{N}\sum_{\substack{t'=1\\t'\neq t}}^{N}\frac{p_t}{p_{t'}}d_{t'}\right)\right]$$

$$=V_1\left(\hat{t}_{\theta,RHC}\right)+V_2\left(\hat{t}_{\theta,RHC}\right).$$

Un estimador para $V_2\left(\hat{t}_{\theta,RHC}\right)$ esta dado por $v_2\left(\hat{t}_{\theta,RHC}\right)=P(1-P)\sum_{i=1}^{n}\frac{d_i}{\left(p_i/\pi_i\right)^2}$ de manera que para obtener un estimador para $V\left(\hat{t}_{\theta,RHC}\right)$ solamente requerimos un estimador para $V_1\left(\hat{t}_{\theta,RHC}\right)$.

Finalmente el estimador-RHC para t_A es:

$$\hat{t}_{A2,RHC}=\frac{\hat{t}_{\theta,RHC}-(1-P)t_W}{P}=\frac{1}{P}\hat{t}_{\theta,RHC}-\frac{1-P}{P}t_W\,.$$

4. EJEMPLO DE APLICACIÓN DE LOS RESULTADOS DE LA SECCIÓN 3

```
#RR PREGUNTAS COMPLEMENTARIAS con Probabilidades Desiguales(Rao-Hartley-
   Cochran)
#Simulación1.
N<-400
Iset<-1:N
n<-40
r<-N/n
A<-300
Y1<-c(rep(0,N-A),rep(1,A))
Y<-sample(Y1,N,replace=F)
Z<-rep(2,N)
P<-0.7
pRC<-runif(N)
for (k in 1:N){if (pRC[k]<P) (Z[k]<-Y[k]) else (Z[k]<-1-Y[k])}
#
x<-sample(18:50,N,replace=T)
Zhat<-(Z-(1-P))/(2*P-1)
M<-250
df<-array(rep(0,N*5*M),dim=c(N,5,M))
for (l in 1:M)(df[,,l]<-matrix(c(Iset,x,Y,Z,Zhat),nrow=N))
#
U<-array(rep(0,N*M),dim=c(r,n,M))
for ( l in 1:M)(U[,,l]<-matrix(sample(Iset,N,replace=F),nrow=n))
X<-array(rep(0,N*M),dim=c(r,n,M))
p<-array(rep(0,N*M),dim=c(r,n,M))
s<-matrix(rep(0,n*M),nrow=n)
for (l in 1:M){for (k in 1:n)(X[,k,l]<-x[U[,k,l]])&(p[,k,l]<-
   X[,k,l]/sum(X[,,l]))&(s[k,l]<-sample(U[,k,l],1,prob=p[,k,l]))}
#
df.s<-array(rep(0,n*5*M),dim=c(n,5,M))
for ( l in 1:M)(df.s[,,l]<-df[s[,l],,l])
#
```

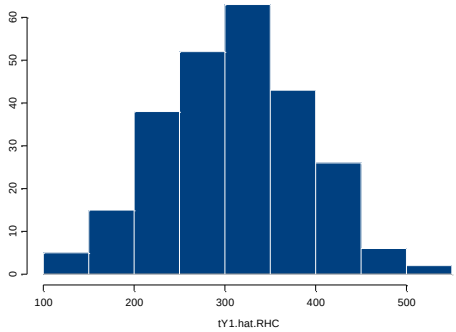
```
x.s<-matrix(rep(0,n*M),nrow=n)
Y.s<-matrix(rep(0,n*M),nrow=n)
Z.s<-matrix(rep(0,n*M),nrow=n)
Zhat.s<-matrix(rep(0,n*M),nrow=n)
for(l in 1:M){(x.s[,l]<-df.s[,2,l])&(Y.s[,l]<-df.s[,3,l])&(Z.s[,l]<-
   df.s[,4,l])&(Zhat.s[,l]<-df.s[,5,l])}
#
t.x<-matrix(rep(0,n*M),nrow=n)
for( l in 1:M){for (k in 1:n)(t.x[k,l]<-sum(X[,k,l]))}
tY1.hat.RHC<-rep(0,M)
for( l in 1:M)(tY1.hat.RHC[l]<-sum((Zhat.s[,l]/x.s[,l])*t.x[,l]))
tY1.hat.RHC<-floor(tY1.hat.RHC)
```

RESULTADOS DE LA SIMULACIÓN1

```
***  Summary Statistics for data in:  tY1.hat.RHC ***

X
    Min: 119.00000
1st Qu.: 255.25000
   Mean: 311.73600
 Median: 308.00000
3rd Qu.: 363.75000
                                    Max: 520.00000

Std Dev.:  77.35509
```



#RR PREGUNTAS NO RELACIONADAS con Probabilidades Desiguales(Rao-Hartley-Cochran)

```
#Simulación2.
N<-400
Iset<-1:N
n<-40
r<-N/n
A<-300
Y1<-c(rep(0,N-A),rep(1,A))
Y<-sample(Y1,N,replace=F)
B<-50
W1<-c(rep(0,N-B),rep(1,B))
W<-sample(W1,N,replace=F)
tW<-sum(W)
Z<-rep(2,N)
P<-0.7
pRC<-runif(N)
for (k in 1:N){if (pRC[k]<P) (Z[k]<-Y[k]) else (Z[k]<-W[k])}
```

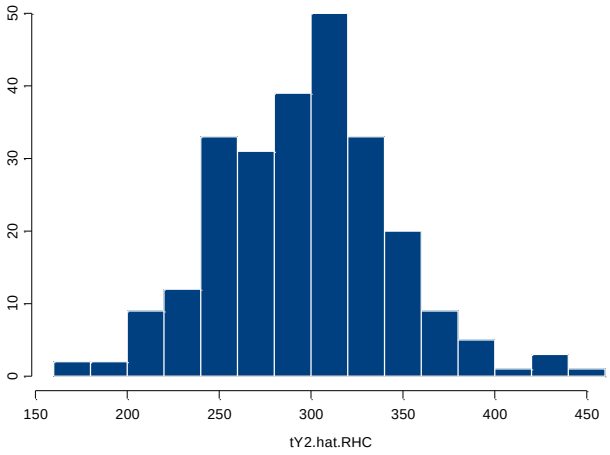
```
#
x<-sample(18:50,N,replace=T)
theta<-P*Y+(1-P)*W
M<-250
df<-array(rep(0,N*6*M),dim=c(N,6,M))
for (l in 1:M)(df[,,l]<-matrix(c(Iset,x,W,Y,Z,theta),nrow=N))
#
U<-array(rep(0,N*M),dim=c(r,n,M))
for ( l in 1:M)(U[,,l]<-matrix(sample(Iset,N,replace=F),nrow=n))
X<-array(rep(0,N*M),dim=c(r,n,M))
p<-array(rep(0,N*M),dim=c(r,n,M))
s<-matrix(rep(0,n*M),nrow=n)
for (l in 1:M){for (k in 1:n)(X[,k,l]<-x[U[,k,l]])&(p[,k,l]<-
   X[,k,l]/sum(X[,,l]))&(s[k,l]<-sample(U[,k,l],1,prob=p[,k,l]))}
#
df.s<-array(rep(0,n*6*M),dim=c(n,6,M))
for ( l in 1:M)(df.s[,,l]<-df[s[,l],,l])
#
x.s<-matrix(rep(0,n*M),nrow=n)
for(l in 1:M){(x.s[,l]<-df.s[,2,l])&(W.s[,l]<-df.s[,3,l])&(Y.s[,l]<-
   df.s[,4,l])&(Z.s[,l]<-df.s[,5,l])&(theta.s[,l]<-df.s[,6,l])}
W.s<-matrix(rep(0,n*M),nrow=n)
Y.s<-matrix(rep(0,n*M),nrow=n)
Z.s<-matrix(rep(0,n*M),nrow=n)
theta.s<-matrix(rep(0,n*M),nrow=n)
for(l in 1:M){(x.s[,l]<-df.s[,2,l])&(W.s[,l]<-df.s[,3,l])&(Y.s[,l]<-
   df.s[,4,l])&(Z.s[,l]<-df.s[,5,l])&(theta.s[,l]<-df.s[,6,l])}
#
t.x<-matrix(rep(0,n*M),nrow=n)
for( l in 1:M){for (k in 1:n)(t.x[k,l]<-sum(X[,k,l]))}
tY2.hat.RHC<-rep(0,M)
for( l in 1:M)(tY2.hat.RHC[l]<-(1/P)*sum((Z.s[,l]/x.s[,l])*t.x[,l])-((1-
   P)/P)*tW)
tY2.hat.RHC<-floor(tY2.hat.RHC)
tY2.hat.RHC
```

RESULTADOS DE LA SIMULACIÓN2

```
***  Summary Statistics for data in:  tY2.hat.RHC ***

X
    Min: 176.00000
1st Qu.: 262.75000
   Mean: 296.85600
 Median: 299.00000
3rd Qu.: 324.75000
    Max: 452.00000

Std Dev.:  46.58683
```



RECOMENDACIONES: El estudio de simulación muestra que el estimador $\hat{t}_{A2,RHC,\pi}$ de Greenberg tiene menor varianza que el estimador $\hat{t}_{A1,RHC,\pi}$ de Warner.

RECEIVED NOVEMBER 2009
REVISED JULY 2010

REFERENCIAS

[1] GJESTVANG, C.R, and SINGH, S. (2006):A new randomized response model. **J. R. Statist. Soc.** B 68, Part 3, 523–530.

[2] GREENBERG B.C., ABUL-ELA, ABDEL-LATIF, A., SIMMONS, W.R., and HORVITZ, D.C. (1969) The unrelated questions randomized response model: Theoretical framework. **Journal of the American Statistical Association,** 64, 520-539.

[3] GUPTA S., SHABBIR J. (2004): Sensitivity estimation for personal interview survey questions. **Statistica**, anno lxiv, n. 4..

[4] JEA-BOK RYU, JONG-MIN KIM, TAE-YOUNG HEO and CHUN GUN PARK. On stratified randomized response sampling. **Model Assisted Statistics and Applications,** 1. 31-36

[5] MOORS, J.J.A (1971): Optimization of the unrelated question RR model. **J. Amer. Statist. Assoc.** 70, 80-83.

[6] NADDEO, S., and PISANI, C. (2005): Two-stage adaptive cluster sampling. **Statistical Methods & Applications,** 14: 3-10.

[7] RAO,J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962): On a simple procedure of unequal probability sampling without replacement. **Journal of the Royal Statistical Society** B 24**,** 482-491**.**

[8] SAHA, A. (2006): A generalized Two-Stage randomized response procedure in complex sample surveys. **Aust. N. Z. J. Stat.** 48, 429–443

[9] SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J. (1992): **Model Assisted Survey Sampling**. New York: Springer Verlag.

[10] SUDMAN, S., and BRADBURN, N.M. (1974): **Response Effects in Surveys**, Aldine, Chicago.

[11] WARNER, S. L. (1965), Randomized Response: A Survey Technique for Eliminating Evasive Answer

Bias, **Journal of the American Statistical Association**, 60, 63-69.

[12] WRETMAN, J.K., SARNDAL, C.E., CASSEL,C.M.,(1977): **Foundations of Inference in Survey Sampling.** Wiley, New York