UN MODELO DE MUESTREO PARA TRATAMIENTO DE PREGUNTAS EVASIVAS CON ALTA SENSIBILIDAD

Víctor H. Soberanis Cruz, Jaime D. Cuevas Domínguez

vsobera@uqroo.mx
División de Ciencias e Ingeniería, Universidad de Quintana Roo
Boulevard Bahía s/n, esquina Ignacio Comonfort, Col. Del Bosque
Chetumal Quintana Roo, C.P. 77019

RESUMEN

Este trabajo presenta e ilustra la aplicación de una técnica de muestreo con respuestas aleatorizadas, que permite obtener información sobre asuntos sensibles o delicados en poblaciones finitas como una alternativa al modelo inicial propuesto por Warner. La técnica presentada resulta mejor que la del modelo inicial en el sentido de que es más precisa (menor variación).

Palabras Claves: Muestreo probabilístico, preguntas sensitivas, respuesta aleatorizada.

INTRODUCCIÓN

Los métodos estadísticos más populares (clásicos), tienen una gran utilidad en diversos campos entre ellos en la medición de las variables físicas utilizadas en Ingeniería, donde se utilizan para encontrar la variabilidad sistemática del instrumento de medición y la variabilidad aleatoria propia del instrumento, apoyados por patrones de referencia nacional e internacional de estos instrumentos. Con estos métodos se puede determinar el nivel de incertidumbre de la estimación de la precisión y exactitud.

Sin embargo, si se trata de variables que se miden con una encuesta, cuya veracidad de la respuesta depende de muchos factores personales del encuestado, se requieren de nuevas consideraciones y métodos para determinar la incertidumbre.

Existen varios cuerpos de conocimiento cuando se trata del estudio de la medición de las respuestas en las encuestas, uno de ellos dirigido principalmente al tipo de respuesta numérica como la edad, los ingresos y al tipo de respuesta ordinal como la percepción de un servicio, el nivel de confort en un edificio etc. En estos casos se pueden utilizar métodos desarrollados a partir del análisis factorial exploratorio, el análisis factorial confirmatorio y la modelación de las ecuaciones estructurales (Cuevas *et al* 2007), considerando en la modelación, un error de medida de las respuestas, que permite construir indicadores de validez y confiabilidad del instrumento, y en consecuencia de los resultados de las encuestas.

Un segundo nicho de conocimiento en el estudio de la medición de las respuestas en las encuestas, desde el punto de vista estadístico, es el abordaje del problema de evasión de respuesta o la falsedad en preguntas sensitivas o difíciles de contestar para el encuestado, como son: Uso de drogas, preferencias sexuales, honestidad en los exámenes, honestidad en el trabajo, honestidad en comisiones de promociones o estímulos para los profesores etc.

Esta dificultad en la respuesta estriba en la posibilidad de la violación del anonimato en el proceso de la encuesta, que pudiera herir la susceptibilidad del encuestado.

Una forma de proteger el anonimato del entrevistado, consiste en garantizar que el entrevistador no conozca que pregunta se está respondiendo.

Warner (1965) desarrolló un modelo, conocido actualmente como modelo W, cuya característica principal es la técnica llamada Respuestas Aleatorizadas (RR), la cual pretende garantizar el anonimato del entrevistado mediante un mecanismo aleatorio que selecciona una de dos preguntas

complementarias, por ejemplo: la pregunta 1.- ¿Copiaste en el examen?, siendo la pregunta 2.- ¿No copiaste en el examen? El entrevistado contestará V (verdadero) o Falso (F), de acuerdo a la pregunta seleccionada, y el entrevistador registra la respuesta pero nunca sabrá que pregunta respondió el encuestado, protegiéndose el anonimato del entrevistado.

El método genérico consiste en:

Extraer una muestra aleatoria de tamaño n de acuerdo a un diseño de muestreo seleccionado (aleatorio simple, estratificado, conglomerados, probabilidades proporcionales a los tamaños, etc.)

Cada elemento de la muestra o entrevistado, seleccionará una de las dos preguntas complementarias, y responderá V (verdadero) o F (falso)

Mediante el modelo estadístico W se estima el total de personas con la característica sensible.

Basado en los trabajos iniciales de Warner, se han desarrollados otros modelos de muestreo para encuestas de respuestas aleatorizadas. Greeberg *et al* (1969) proponen como alternativa al modelo de W de preguntas complementarias, un modelo con un mecanismo de selección aleatorio de la pregunta a realizarse al entrevistado entre dos preguntas no relacionadas entre sí, conocido como modelo U (unrelated), donde una pregunta es sobre el asunto sensitivo y la segunda pregunta no relacionada es inocua, es decir no es sensible al entrevistado. De forma tal que un entrevistado puede responder una de las dos preguntas como por ejemplo: ¿Copiaste en el examen?, ¿Te inscribiste en deportes?

Un tercer modelo es propuesto por Morton (Horvitz, 1976), que en este trabajo se menciona como modelo Mu (Morton unrelated), permite mayor protección del anonimato del entrevistado: La selección aleatoria se realizará entre tres propuestas a) Una propuesta sensitiva; b) la palabra Si, sin relación sensitiva y c) la palabra No, sin relación sensitiva. Teniendo cada una la probabilidad de ser escogida de P_1, P_2, P_3 donde $P_1 + P_2 + P_3 = 1$.

Con la intención de clarificar el proceso de selección del modelo Mu, suponga que las tres proposiciones que pueden ser seleccionadas, son a) Copió en el examen, b) Si, c) No. Suponga también que un entrevistado escoge aleatoriamente la propuesta o pregunta b), dando como resultado un Si, que no tiene relación aparente con el asunto de copiar en el examen, solo los entrevistados que por suerte les corresponda responder la propuesta a), procederán a responder sobre el asunto sensitivo: Debido a que el entrevistador no conoce sobre que propuestas contestaron, el entrevistado puede sentir garantizado el anonimato y contestar con veracidad.

Otros modelos de respuestas aleatorizadas han sido estudiados, entre otros un modelo C (Soberanis 2008) para relacionar la pregunta inocua del modelo U y la variable sensitiva.

En este trabajo se utiliza el modelo Mu para el caso de poblaciones finitas, y mediante un caso calculan los indicadores estadísticos apropiados.

MATERIALES Y MÉTODOS

a) Definición del problema

La población a la que está dirigida el estudio, es una población finita de tamaño N= 700 estudiantes. La variable de interés representada por la letra $\boldsymbol{y}$, mide una característica sensitiva, en este caso mediante la respuesta a la pregunta ¿Has consumido alguna droga ilegal? Así $\boldsymbol{y_k}$ representa la respuesta de la pregunta sensitiva del entrevistado cuyo índice en la población $\{1, 2, ..., N\}$ es k y puede tomar solo dos valores, 0 para ausencia o negación de la característica, y 1 para la existencia o aceptación de dicha característica. Nótese que $\boldsymbol{y_k}$ es una variable desconocida, pero no es aleatoria.

El objetivo del método es estimar la totalidad de los individuos con la característica sensitiva, es decir para este trabajo estimar cuantos estudiantes de la población finita han consumido drogas ilegales. Este total se representa como t_A.

b) Procedimiento de muestreo.

Se utiliza un esquema de muestreo aleatorio simple sin reemplazo, para determinar el tamaño de la muestra y seleccionar a los individuos a entrevistar. Se define el mecanismo aleatorio (RC), de forma tal que cada elemento de la muestra selecciona aleatoriamente una de las tres proposiciones: (1) has consumido drogas ilegales, (2) una instrucción que dice Si y (3) una instrucción que dice No. Con probabilidades de ser escogidas P_1, P_2, P_3 respectivamente, donde $P_1 + P_2 + P_3 = 1$.

c) Cálculo del estimador y la varianza del estimador.

El estimador propuesto de t_A , $\hat{t}_A$, resulta insesgado, por tanto un buen indicador de referencia es la varianza del estimador. Para mostrar que en efecto $\hat{t}_A$, el estimador que proponemos para t_A, es insesgado procedemos como sigue:

Sea Z_k la variable aleatoria que toma los valores $y_k, 1, 0$ con probabilidades P_1, P_2, P_3 respectivamente.

Así, el valor esperado con este mecanismo aleatorio para Z_k asumiendo que se conoce y_k, se puede calcular como sigue, y definirse como θ_k:

$$E_{RC}(Z_k|y_k) = y_k P_1 + P_2 \equiv \theta_k \quad (1)$$

Si definimos $t_\theta = \sum_{k \in U} \theta_k$ (2)

Entonces $t_\theta = P_1 t_A + NP_2$ (3)

$$t_A = \frac{t_\theta - NP_2}{P_1} \quad (4)$$

El estimador para t_θ es:

$$\hat{t}_\theta = \sum_s \frac{Z_k}{\pi_k} \quad (5)$$

Donde π_k es la probabilidad de que el *k*-ésimo elemento de la población caiga en la muestra. Se puede verificar que el estimador $\hat{t}_\theta$ es insesgado (Särndal et al 1992). Finalmente el estimador insesgado para t_A es

$$\hat{t}_A = \frac{1}{P_1}\hat{t}_\theta - \frac{NP_2}{P_1} \quad (6)$$

d) Se procede a realizar los cálculos con la ayuda de Excel y comentar los resultados.

DESARROLLO Y RESULTADOS

El diseño de muestreo elegido es un muestreo aleatorio simple sin reemplazo, para un tamaño de población N=700 y tamaño de muestra n=140, definiendo a f = n/N.

Mediante un sistema de sorteo simple con un marco que contenga los 700 individuos se eligen los 140 individuos a ser entrevistados.

Posteriormente se procede al siguiente mecanismo aleatorio: en una caja se colocan cien tarjetas indistinguibles, 70 de ellas con la proposición sensitiva ¿Has consumido drogas ilegales?, 9 con la instrucción Si y 21 con la instrucción No. El individuo selecciona una tarjeta al azar y responde Si o No, de acuerdo a la tarjeta, se registra su respuesta asignando un 1 por cada Si respondido, y un cero por cada No. El individuo regresa la tarjeta al mazo y revuelve.

Este mecanismo aleatorio define una variable aleatoria que denominamos Z_k , donde el índice k pertenece a la muestra, que puede tomar valores de y_k con una probabilidad de P_1 , de 1 con una probabilidad P_2, de 0 con una probabilidad $1\text{-}P_1\text{-}P_2$. Note que Z_k es dicotómica o binaria y solo toma valores de 1 y 0.

Los valores z_k , $k \in s$ que son las realizaciones de la variable aleatoria Z_k , para este ejercicio son:

1 1 1 0 1 1 1 0 1 1 1 0 0 0 1 0 1 1 1 1 1 1 0 0 1
1 0 0 0 0 1 1 1 1 0 1 1 1 1 1 1 0 0 1 1 0 1 1 1 0 1
1 1 1 0 1 0 1 0 0 1 01 1 1 0 0 1 1 1 1 1 1 1 1 1 1
1 1 1 0 0 0 0 1 0 0 1 0 1 1 0 1 1 1 1 1 1 0 0 0 0 1
1 1 0 0 0 1 1 1 0 1 0 0 0 1 0 1 1 0 0 1 0 1 1 1 1 1
1 1 1 0 1 10 1 1 0 1 1 1.

Para estimar el total, se considera el tipo de muestreo aleatorio simple para poblaciones finitas ($\pi_k = \frac{n}{N} = f$) y el mecanismo aleatorio Mu (Soberanis, 2008) resultando.

$$\hat{t}_A = (1/P_1)(1/f)\sum_{S \ni k} Z_k - NP_2/P_1 \quad (7)$$

Donde P_1, P_2 se calcula con la cantidad de tarjetas, resultando 0.70 y 0.09 respectivamente: Se consultó en la literatura y no se encontró alguna referencia que indique la magnitud de las probabilidades para mejorar el proceso, siendo un campo interesante aún en desarrollo.

La expresión $\sum_{s \ni k} z_k$ representa la suma de los 1 en la muestra, que en este caso es de 90 y $f = n/N = 0.2$ ya que n=140 y N=700.

Al realizar el cálculo del número estimado de estudiantes que han consumido droga ilegal de esa población, de 700 estudiantes, resulta ser de 553.

Es posible calcular la varianza del estimador para los casos como este que se elige una selección de muestra aleatoria simple para poblaciones finitas y mecanismo aleatorio Mu (Soberanis 2008).

$$\hat{v}(\hat{t}) = \frac{1}{P_1^2}\left\{(Z_S)^t Q_S (Z_S) + \left(\frac{(1-P_1-P_2)}{f^2}\right)\sum_{S \ni k}(Z_k - P_2) + \frac{(N^2 P_2)(1-P_2)}{n}\right\} \quad (8)$$

Donde Q_S es una matriz n x n con elementos 1-*f* en la diagonal y elementos 1-((N-1)/(n-1))*f*, fuera de la diagonal, Z_S es el vector de las observaciones de Z y $(Z_S)^t$ es su vector traspuesto.

Los cálculos arrojan una varianza:

$$\hat{V}(\hat{t}) = 1111.733$$

y en consecuencia una desviación estándar de 33.34.

CONCLUSIONES

El objetivo de este artículo es presentar los resultados de aplicar un modelo usando un muestreo de encuestas para asuntos sensitivos, mediante una técnica que garantice la protección del anonimato del entrevistado. Generándose las conclusiones, que en los siguientes párrafos se describen.

El procedimiento permite visualizar que la técnica garantiza el anonimato del entrevistado, esperándose que no evada la pregunta y conteste con veracidad.

El modelo permite obtener un buen estimador del total de individuos con la característica sensitiva.

Los modelos RR dependen del tipo de muestreo para seleccionar la muestra y de la técnica del mecanismo aleatorio, en este caso se utilizó el modelo Mu ya que produce un estimador insesgado cuya varianza resulta menor que la del estimador propuesto por Warner (Soberanis, 2008).

Los autores de este trabajo consideran que este tema es pertinente y tiene un gran potencial para los estudios que se requieren para medir de forma más adecuada los niveles de opacidad, delincuencia y adicciones, que agobia a las ciudades de México.

REFERENCIAS

Cuevas J.,Soberanis V., Acosta R., (2007) Modelación de ecuaciones estructurales para la evaluación de la calidad percibida en el proceso de enseñanza-aprendizaje. *Revista Cubana de Educación Superior*, XXVII, 13-20.

Greenberg B.G., et al. (1969) The unrelated question RR model: Theoretical Framework. *Journal of the American Statistical Association,* 64, 520-539.

Horvitz D.C. (1976). RR: A data gathering device for sensitive question. *International Statistical Review*, 44, 181-196.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

Soberanis V., *et al* (2008). Muestreo de Respuestas Aleatorizadas en Poblaciones

Finitas: Un Enfoque Unificador. *Agrociencia*, 42-5, 537-549.

Warner, S (1965). Randomized Response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60, 63-69.