PESQUIMAT , Revista de la F.C.M. de la
Universidad Nacional Mayor de San Marcos
Vol. XVIII N°1, pp. 29-36, Lima - Perú, Abril 2015

EL MODELO DE RESPUESTA NO ALEATORIZADA: UNA APLICACIÓN DEL MODELO PARALELO

Olga Lidia Solano Dávila [1] **, *Manuel Hilario Reyes Liñan*** [2]

(Recibido: 02/04/2015 - Aceptado: 18/05/2015)

Resumen: En el presente trabajo presentamos los resultados de una investigación realizada en la Facultad de Ciencias Matemáticas de la UNMSM, para investigar la proporción de personas que han consumido drogas, porcentaje de personas que consumen alcohol todos los fines de semana,y porcentaje de personas que copian en los exámenes, utilizando la teoría del modelo de respuesta no aleatorizada, el modelo paralelo (Tian, 2012;Tian & Liu 2014). En el diseño muestral, se utilizó el muestreo aleatorio estratificado con afijación proporcional al tamaño de cada estrato (Scheffer y Mendenhall, 2007) considerando como estratos a las Escuelas Académicas Profesionales (E.A.P.) de la FCM. Para el cálculo del tamaño de muestra se consideró un límite para el error de estimación de 3,03 %, con un nivel de confianza del 95 % y la información proporcionada por la Dirección Académica de la FCM, de los alumnos matriculados el primer semestre del año académico 2015, el tamaño de muestra fue de 573 alumnos, repartidos en forma proporcional a las cuatro E.A.P. de la FCM.

Palabras clave: Modelo de resposta no aleatorizada, modelo paralelo, pregunta delicada, pregunta no relacionada.

RESPONSE MODEL NON RANDOMIZED: AN APPLICATION OF PARALLEL MODEL

Abstract: In this paper we present the results of research conducted in the Faculty of Mathematics (FCM) at the University Nacional Mayor de San Marcos in Peruvian to investigate the proportion of people who have tried drugs, percentage of people who drink alcohol every weekend, and percentage of people who copy in exams using theory the nonrandomized response model, the parallel model. The parallel model (Tian, 2012;Tian & Liu 2014). Stratified random sampling with allocation proportional to the size of each stratum was used in the sample design (Scheffer and Mendenhall, 2007), considering as strata to the Professional Academic School (E.A.P) of the FCM. For the calculation of sample size limit to the estimation error of 3,03 %, with a confidence level of 95 % and the information provided by the Academic Board of the FCM, of the students enrolled in the first half of the year it was considered academic 2014, the sample size was 573 students, distributed in proportion to the four EAP of the FCM.

Keywords: Resposta model nonrandomized, parallel model, delicate question, unrelated question.

[1]UNMSM, Facultad de Ciencias Matemáticas, e-mail:osolanod@unmsm.edu.pe
[2]Hospital Santa Rosa, e-mail: reyes@unmsm.edu.pe

1. Introducción

Existen multitud de estudios realizados para mejorar la calidad y veracidad de las respuestas obtenidas sobre temas sensibles, como consumo de drogas, relaciones sexuales con más de dos personas, consumo de alcohol, copia de exámenes, etc. Frecuentemente los investigadores obtienen resistencias a la participación en estos estudios u obtienen respuestas falsas de parte de los entrevistados. (Warner, 1965) realizó la primera propuesta para obtener respuestas válidas ante preguntas embarazosas basándose en la realización de dos preguntas mutuamente excluyentes (por ejemplo "A: Declaré mis ingresos extraordinarios el año pasado"; "B: No declaré mis ingresos extraordinarios el año pasado"). Posteriormente, se desarrollaron otros métodos basados en (Warner, 1965), como el denominado "método de alternativa forzada" de Tracy (Fox & Tracy, 1986). El modelo de respuesta no aleatorizada es un método especialmente diseñado para asegurar privacidad a los entrevistados en el estudio de temas sensibles, delicados o embarazosos. Se intenta con ello evitar el sesgo de los entrevistados en ciertas conductas hacia la respuesta socialmente más deseable. Se ha utilizado para analizar temas como copiar en los exámenes, insolvencia, fraudes, haber sido arrestado, conducir bajo los efectos del alcohol, tener un hijo fuera del matrimonio, aborto, etc.

Recientemente, (Tian & Liu, 2014) propuso el nuevo modelo de respuesta no aleatorizada (MRNA), llamado el modelo paralelo, para estimar la proporción desconocida $\pi = \mathrm{P}\ (\mathrm{Y} = 1)$, de personas con una característica sensible, mediante la introducción de dos variables aleatorias dicotómicas no sensibles U y W tal que Y, U y W son mutuamente independientes (Tian & Liu, 2014) ha desarrollado un marco general de análisis y diseño para el modelo paralelo de respuesta no aleatorizada.

El MRNA, el cual utiliza una o dos variables no sensibles (ejemplo, fecha de nacimiento del entrevistado o el último dígito del número de teléfono del entrevistado), combinado con una o dos variables sensibles para formar una tabla de contingencia incompleta y obtener indirectamente de los entrevistados respuestas sensibles (Takahasi & Sakasegawa, 1977); (Tian, Yu, Tang, & Geng, 2007b); (Tian, Tang, Liu, Tan, & Tang, 2011); (Yu, Tian, & Tang, 2008); (Tan, Tian, & Tang, 2009); (Tang, Tian, Tang, & Liu, 2009), el diseño no requiere de mecanismos de aleatorización.

El objetivo del presente trabajo es presentar los resultados de uan investigación realizada en la FCM de la UNMSM utilizando el modelo de respuesta no aleatorizada, el modelo paralelo, en el contexto descrito con la finalidad de que cualquier investigador interesado en usar esta metodología conozca los aspectos fundamentales de la misma.

2. Metodología

Modelo de respuesta no aleatorizada - el modelo paralelo

Tian (2012) hizo una propuesta del modelo paralelo, el cual es una versión del modelo de la pregunta no relacionada. Sea $\{Y = 1\}$ denota una clase de personas que tienen una característica (por ejemplo, consumo de drogas) y $\{Y = 0\}$ denota la clase complementaria. El objetivo es estimar la proporción de los elementos que pertenecen a la clase $\{Y = 1\}$, $\pi = P(Y = 1)$. Suponga que U y W son dos variables dicotomicas no sensibles y Y, U y W son independientes entre si con $q = P(U = 1)$ y $p = P(W = 1)$ conocidas.

El diseño de la encuesta para el modelo paralelo

Entrevistadores pueden diseñar un cuestionario en el formato mostrado en el lado izquierdo de la Tabla 1, y pedir a cada entrevistado conectar los dos círculos por una línea recta, si él/ella pertenece a uno de los dos círculos o conectar los dos cuadrados por una línea recta si él/ella pertenece a uno de los dos cuadrados. Note que todas las clases $\{W = 1\}, \{W = 0\}, \{U = 1\}, \{U = 0\}$ son clases no sensibles, así $\{U = 1, W = 0\} \cup \{Y = 1, W = 1\}$ es también una sub

clase no sensible. Por lo tanto, si el entrevistado pertenece a la clase sensíble no es conocido por los entrevistadores. Por lo tanto, si él entrevistado pertenece a la clase sensible no és conocido por los entrevistados. Las probabilidades de las celdas correspondientes son mostradas en el lado derecho de la Tabla 1. Las tres variables binarias U, Y y W son independientes, la probabilidad conjunta es el producto de dos probabilidades marginales correspondientes.

Tabla 1: El modelo paralelo y las correspondientes celdas de probabilidades

Categoría	$W=0$	$W=1$	Categoría	$W=0$	$W=1$	Marginal
$U=0$			$U=0$	$(1-q)(1-p)$		$1-q$
$U=1$			$U=1$	$q(1-p)$		q
$Y=0$			$Y=0$		$(1-\pi)p$	$1-\pi$
$Y=1$			$Y=1$		πp	π
		Marginal		$1-p$	p	1

En relación a la Tabla 1, se define una variable aleatoria de Bernoulli como

$$Y^P = \begin{cases} 1, \text{ si los cuadrados son conectados} \\ 0, \text{ si los círculos son conectados} \end{cases}$$

Donde el superíndice "P" representa a la variable de Bernoulli para el modelo paralelo. Por lo tanto, las probabilidades de $Y^P = 1$ y $Y^P = 0$ están dadas por:

$P[Y^P = 1] = q(1-p) + \pi p$ y $P[Y^P = 0] = (1-q)(1-p) + (1-\pi)p$,

respectivamente.

Si $Y_{obs} = \left\{ y_i^P : i = 1, ..., n \right\}$ denota los datos observados para los n encuestados, entonces la función de probabilidad para π es:

$L_P(\pi \ / \ Y_{obs}) = \prod \left[q(1-p+\pi p]^{y_i^P} \ [(1-q)\,(1-p) + (1-\pi)p] \ ^{1-y_i^P} \right.$

Como consecuencia el estimador de máxima verosimilitud de π es:

$$\widehat{\pi}_P = \frac{\overline{y}^P - q(1-p)}{p} \tag{1}$$

donde $\overline{y}^P = (1/n)\sum y_i^P$.Se verifica que $\widehat{\pi}_P$ es un estimador insesgado de π y la varianza de $\widehat{\pi}_P$ está dado por:

$$Var(\widehat{\pi}_P) = \frac{\delta(1-\delta)}{np^2},$$

donde $\delta \approx q(1-p) + \pi p$.

De acuerdo con el teorema del límite central:

$$\frac{\widehat{\pi}_{_P} - \pi}{\sqrt{Var(\widehat{\pi}_{_P})}} = \frac{n\widehat{\pi}_{_P} - \pi}{\sqrt{n}\delta(1-\delta)/p} \sim N(0,1), \quad \text{cuando} \quad n \to \infty.$$

Intervalo de Confianza de Wald para π

Por el Teorema de Límite Central y $\widehat{\pi}_P$ especificado en (1):

$\frac{\widehat{\pi}_{_P} - \pi}{\sqrt{Var(\widehat{\pi}_{_P})}} \sim N(0,1)$, cuando n$\to \infty$.

El Intervalo de Confianza, de Wald, para π al 100(1-α) % de confianza es:

$\left[\widehat{\pi}_{_P} - Z_{\alpha/2}\sqrt{Var(\widehat{\pi}_{_P})} \ , \ \widehat{\pi}_{_P} + Z_{\alpha/2}\sqrt{Var(\widehat{\pi}_{_P})} \right]$

3. Resultados y Discusión

El presente estudio fue realizado en la Facultad de Ciencias Matemáticas, de la UNMSM. Se aplicó el modelo de respuesta aleatorizada - el modelo paralelo - a los alumnos matriculados el

Semestre académico del 2014-I, para investigar la proporción de personas que ha consumido drogas, porcentaje de personas que consumen alcohol todos los fines de semana, y porcentaje de personas que copian en los exámenes, utilizando la teoría del modelo de respuesta no aleatorizada, el modelo paralelo (Tian, 2012;Tian & Liu 2014). Para el cálculo del tamaño de la muestra se utilizó el muestreo aleatorio estratificado (Scheaffer; Mendenhall; Ott, 2007) con un límite para el error de estimación de 3,03 % y la información proporcionada por la Dirección Académica de la FCM. Se consideró a cada E.A.P. como un estrato: Investigación Operativa, Estadística, Computación Científica y Matemática. Los estudiantes fueron seleccionados aleatoriamente hasta completar el tamaño de la muestra igual a 573. **Las variables elegidas para este estudio fueron las siguientes**:

1. Prevalencia de vida de drogas

Preguntas sensibles: "Durante mi vida he consumido drogas (marihuana, pasta básica de cocaína, éxtasis) por lo menos una vez" (Y=1)

•"Nunca he consumido drogas (marihuana, pasta básica de cocaína, éxtasis)" (Y=0)

Preguntas no sensibles: •Trabajo actualmente (U =1)

•No trabajo actualmente (U =0)

•Utilizo página Web (W=1)

•No utilizo página Web (W=0)

2. Prevalencia de sexo

Preguntas sensibles: •"Durante mi vida he mantenido relaciones sexuales con más de dos personas (simultáneamente o no)" (Y =1)

•"Durante mi vida jamás he mantenido relaciones sexuales con más de dos personas (simultáneamente o no)" (Y=0)

Preguntas no sensibles: Elegí mi carrera por vocación (U=1)
•No elegí mi carrera por vocación (U= 0)
•He interrumpido mis estudios universitarios por lo menos una vez (W=1)
•Nunca he interrumpí mis estudios universitarios (W = 0)

3. Consumo actual de alcohol

Preguntas sensibles: •" Consumo alcohol (cerveza, vino, sangría, etc.) con frecuencia (todos los fines de semana)" (Y =1)
•"No consumo alcohol (cerveza, vino, sangría, etc.) con frecuencia (todos los fines de semana)" (Y = 0)

Preguntas no sensibles:• El servicio de la clínica universitaria es bueno (U = 1)
•El servicio de la clínica universitaria no es bueno (U = 0)
•Estudié en colegio público (W = 1)
•No estudié en colegio público (W = 0)

4. Copia en los exámenes

Preguntas sensibles: •"He copiado en los exámenes por lo menos una vez " (Y =1)
•"Nunca he copiado en los exámenes " (Y = 0)
•El servicio del comedor es bueno (U=1)
•El servicio del comedor no es bueno (U=0)
•Me gusta estudiar solo (W = 1)
•No me gusta estudiar solo (W = 0)

Cada una de las preguntas consideradas anteriormente fueron organizadas en una tabla de contingencia.

Tamaño de muestra

De acuerdo a Registros Académicos de la Dirección Académica los Alumnos Matriculados el Semestre 2014 - I en la Facultad de Ciencias Matemáticas de la UNMSM, está dividida en cuatro Escuelas Académico Profesionales(ver Tabla 2), de acuerdo a esta información se decidió considerar a cada Escuela Académico Profesional como un estrato, en total tenemos cuatro estratos o Escuelas. El esquema de muestreo que se utilizó fue el Muestreo Aleatorio Estratificado, con afijación proporcional de acuerdo a la cantidad de alumnos en cada uno de los estratos, donde cada estrato es una Escuela Académico Profesional de la Facultad de Ciencias Matemáticas de la UNMSM.

Tabla 2: Distribución de los estudiantes de Pre-Grado según Escuela Académico profesional
Semestre 2014-I - FCM-UNMSM

Escuela Académico Profesional	Frecuencia	Porcentaje
Matemática	484	31,35 %
Estadística	275	17,81 %
Investigación Operativa	476	30,83 %
Computación Científica	309	20,01 %
Total	1544	100,00 %

Se utilizó el Muestreo Aleatorio Estratificado (Scheaffer; Mendenhall; Ott, 2007), con un nivel de confianza del 95 % y un limite para el error de estimación de 3,03 %, el tamaño de muestra fue de 573 alumnos matriculados el Semestre 2014-I. El modelo de respuesta no aleatorizada requiere un tamaño de muestra más grande que el método convencional. La entrevista se realizó del 02 al 15 de junio del año 2014. La distribución de la muestra por Escuela Académico Profesional se muestra en la Tabla 3.

Tabla 3: Distribución de la muestra según Escuela Académico Profesional
Semestre 2014-I - FCM-UNMSM

Escuela Académico Profesional	Frecuencia
Matemática	181
Estadística	101
Investigación Operativa	175
Computación Científica	116
Total	573

El estimador del modelo paralelo de respuesta no aleatorizada para la"Prevalencia de vida en

drogas de un alumno de la Facultad de Ciencias Matemáticas", cuyo estimador fue desarrollado anteriormente (ver ecuación 1). El promedio de los dos cuadrados conectados en la muestra es: $\overline{y}_P = \frac{176}{573} = 0,307$.
La probabilidad de que un estudiante utiliza página Web es: $q = P(U = 1) = 0,319$.

La probabilidad de que un estudiante trabaje actualmente es: $p = P(W = 1) = 0,592$,

substituyendo en la ecuación (1), tenemos:

$$\widehat{\pi}_P = \frac{\overline{y}_P - q(1-p)}{p} = \frac{0,307 - 0,319(1-0,592)}{0,592} = 0,299.$$

De este resultado se desprende que, la prevalencia de vida en drogas de los alumnos fue de

0,299.
De forma análoga se calculó la proporción de personas que consumen alcohol todos los finales

de semana, donde $\overline{y}_P = \frac{338}{573} = 0,590$ y tenemos información de $q = P(U = 1) = 0,600$ (la probabilidad de que un alumno estudie en una escuela pú

blica) y $p = P(W = 1) = 0,280$ (probabilidad de que un estudiante manifestó que el servicio de la clinica es bueno), substituyendo en la ecuación (1), tenemos que :

$\widehat{\pi}_P = \frac{0,590 - 0,600(1-0,280)}{0,280} = 0,5643$.
La proporción de personas que consumen alcohol todos los fines de semana es 0,5643.

De forma análoga se calculo la proporción de personas que copian en los exámenes, donde

$\overline{y}_P = \frac{261}{573} = 0,455$ y tenemos información de $q = P(U = 1) = 0,511$(la probabilidad de que un alumno manifiesta que el servicio del comedor es bueno) y $p = P(W = 1) = 0,220$ (probabilidad de que un estudiante estudie solo), substituyendo en la ecuación (1), tenemos que :
$\widehat{\pi}_P = \frac{0,455 - 0,511(1-0,220)}{0,220} = 0,2565$.La proporción de personas que copian en los exámenes es

0,2565.

Tabla 4: Resultados de la investigación: estimación, error estándar e intervalo de confianza

Proporción de personas que	Estimación	Error estándar	I.C. 95 %
consumen drogas	0,299	0,033	$[0,2352; 0,3628]$
consumen alcohol	0,5643	0,073	$[0,4200; 0,7077]$
copian en los exámenes	0,2565	0,095	$[0,0734; 0,4441]$

En la Tabla 4 se observa que utilizando el modelo de respuesta no aleatorizada - modelo paralelo - la proporción de personas que han consumido drogas alguna vez en su vida es 0,299; el 56,43 % de los entrevistados consumen alcohol todos los fines de semana y el 25,65 % de los entrevistados han copiado en los exámenes por lo menos una vez en su vida.

4. Conclusión

El 66,7 % de los alumnos entrevistados pertenecen al género masculino mientras que el 33,3 % al femenino. La edad promedio de los estudiantes fue de 23,21 años.El modelo de respuesta no aleatorizada - el modelo paralelo - permitio inplementar una serie de técnicas de muestreo estadístico, que se mostro eficiente en la selecci´pn y ejecución de la investigación.
La estimación de la "Proporción de personas que consumieron drogas alguna vez en su vidaçon el modelo de respuesta no aleatorizada fue de 0,299, la estimación del "Porcentaje de personas que consumen alcohol todos los fines de semana"fue de 56,43 % y la estimación del "Porcentaje de personas que han copiado en los exámenes alguna vez en su vida"fue de 25,65 %.
El modelo de respuesta aleatorizada es más eficiente cuando es utilizada en investigaciones donde

el problema es altamente sensible y para el cual se requiere que el tamaño de muestra es mayor que el método tradicional.

De los resultados se concluye que el Modelo Paralelo, fue útil y provechosa en una investigación de la realidad de nuestra sociedad, identificar el comportamientos en los estudiantes de la FCM frente a preguntas sensibles, se recomienda seguir aplicando los modelos de respuesta no aleatorizada. en investigaciones de nuestra realidad.Las desventajas del MRNA está en el hecho de que los gastos en la capacitación de los entrevistadores (el entrenamiento en la técnica) y el tiempo que requiere la entrevista para explicar la técnica al entrevistado es más alta con respecto al método tradicional. Se recomienda seguir experimentando estos modelos, en muestras más grandes y en temas en donde la pregunta sea efectivamente altamente sensible o muy comprometedor.Estos modelos también se podrían implementar utilizando la inferencia estadística bayesiana.

Financiamiento

Los autores expresamos nuestro agradecimiento al Consejo Superior de Investigaciones de la Universidad Nacional Mayor de San Marcos por el apoyo financiero para la ejecución del estudio motivo de la presente publicación.

REFERENCIAS BIBLIOGRÁFICAS

[1] Reyes, H. (2014). El modelo de respuesta aleatorizada para estudiar el comportamiento de los estudiantes de la FCM frente a preguntas sensiblesn.Tesis de licenciado en estadística.Facultad de Ciencias Matemáticas. UNMSM, Lima, Perú.

[2] Scheaffer, R., Mendenhall, W. y Ott, L. (2007). *Elementos de muestreo. Grupo* Edit. International Thomson Paraninfo S.A., Madrid, España.

[3] Solano, O. et al. (2010). *Modelo de resposta aleatorizada:aplicação do modelo de Simmons.* Revista Brasileira de Biometria, São Paulo, Vol 28, N^{o}4: 43-51.

[4] Takahasi, K.,Sakasegawa, H. (1977). *A randomized response technique without making use of any randomized device.* Annals of the Institute of Statistical Mathematics. Vol 29:N^{o}1:1-8.

[5] Tan, M.,Tian, G.L., Tang, M.L. (2009). *Sample surveys with sensitive questions: a non-randomized response approch.* The American Statistician. Vol 63, N^{o}1:9-16.

[6] Tang, M.L.,Tian, G.L., Tang, N.S.,Liu, Z.Q. (2009). *A new non-randomized multi-category response model for surveys with a single sensitive question: design and analysis.* Journal of the Korean Statistical Society . Vol 38, N^{o}1:339-349.

[7] Tian, G.L.,Liu, Y. (2014). *Sample size determination for the parallel model in a survey.* Journal of the Korean Statistical Society . Vol 43, 235-249.

[8] Tian, G.L.,Yu,J.W., Tang, M.L., Geng, Z. (2007b). *A new non-randomized model for analyzing sensitive questions with binary outcomes.* Statistics in Medicine. Vol 26, 4238-4252.

[9] Warner, S.L. (1965). *Randomized response: a survey technique for eliminating evasive answer bias.* Journal of the American Statistical Association. Vol 60, 63-69.