Estimación de la prevalencia de una pregunta sensible multicategórica en poblaciones finitas

Camilo Andrés Mocetón Ramírez
Estadístico

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Mayo de 2017

Estimación de la prevalencia de una pregunta sensible multicategórica en poblaciones finitas

Camilo Andrés Mocetón Ramírez
Estadístico

Trabajo de tesis para optar al título de
Magister en ciencias - Estadística

Director
Leonardo Trujillo Oyola, Ph.D.
Doctor en Estadística

Línea de investigación
Muestreo

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Mayo de 2017

Título en español

Estimación de la prevalencia de una pregunta sensible multicategórica en poblaciones finitas.

Title in English

Estimation of the prevalence of a multicategorical sensitive question in finite populations.

Resumen: A veces, debido a la vergüenza, miedo de tener alguna consecuencia personal como recibir multas, castigo o simplemente porque las personas no quieren revelar su intimidad, los encuestados en un estudio pueden negarse a participar. Por otro lado, algunas personas que responden a la encuesta podrían dar respuestas falsas a algún tipo específico de preguntas, inclusive en estudios hechos por oficinas nacionales de estadística. Para los investigadores, en particular para los estadísticos, el primer problema se conoce como error de no respuesta y el segundo como sesgo en la respuesta. El acceso a información relacionada con una característica sensible en la población induce estos dos problemas particulares: no respuesta y sesgo en las respuestas proporcionadas. Las dos fuentes de error suelen ser un problema cuando la característica de interés a estimar corresponde a preguntas sensibles relacionadas con fenómenos como opinión sobre el aborto, violencia doméstica, eutanasia, fraude y plagio, ingresos, racismo, preferencias sexuales, evasión de impuestos, consumo de drogas, entre muchos otros. Técnicas de Respuesta Aleatorizada (TRAs) y Técnicas de Conteo de Items (TCIs) son útiles para obtener una respuesta confiable, pero también manteniendo la confidencialidad y el anonimato de los encuestados. En particular, las TRAs son diseñadas, principalmente, para estimar la prevalencia de una pregunta sensible en la población con dos respuestas posibles: sí o no. Este trabajo propone un método alternativo para estimar la prevalencia de una pregunta sensible con tres o mas categorías bajo cualquier diseño muestral complejo. Las propiedades de los estimadores propuestos son estudiadas tanto teóricamente como a través de simulaciones Monte Carlo. Una aplicación real a trabajadores administrativos de la Universidad Nacional de Colombia en Bogotá se muestra con el fin de estimar la prevalencia del acoso sexual entre ellos.

Abstract: Sometimes due to embarrassment; fear of having any personal consequences as receiving fines, punishment or simply because people does not want to reveal their intimacy, the respondents in a survey can refuse to participate. On the other hand, some people answering the survey could give false answers for some specific type of questions because they do not want to reveal the truth even in surveys from national statistical offices. For researchers and in particular for statisticians, the first problem is known as a nonresponse error and the second one is known as a bias in the response. Accessing information regarding a sensitive characteristic in the population induces these two particular problems: nonresponse and non-truthful answers. The two sources of error frequently appear to be a problem when the characteristic of interest being estimated corresponds to sensitive questions related to phenomena such as abortion, domestic violence, euthanasia, fraud and plagiarism, income, racism, sexual preferences, tax evasion, use of illegal drugs, among many others. Randomized Response Techniques (RRTs) and Item Count Techniques (ICTs) are useful in order to get a trustful answer but also keeping the confidentiality of the respondents. In particular, RRTs are mostly designed in order to estimate the prevalence of a sensitive question in the population

with two possible answers: yes or no. This thesis proposes an alternative method in order to estimate the prevalence of a sensitive question with three categories or more under any complex survey design. The properties of the proposed estimators are studied both theoretically and through Monte Carlo simulations. An actual application to the staff in a public university in Bogota is shown in order to estimate the prevalence of sexual harassment among them.

Palabras clave: preguntas sensibles; respuesta aleatorizada; modelo aditivo; anonimato; diseños muestrales complejos

Keywords: sensitive questions; randomized response; additive model; anonymity; complex survey designs

Nota de aceptación

Trabajo de tesis

"Mención Meritoria o Laureada"

Jurado

Jurado

Director
Leonardo Trujillo Oyola

Bogotá, D.C., Mayo 26 de 2017

Dedicado a

A mis padres.

Agradecimientos

Agradezco las oraciones y ayuda de mi familia y de mi pareja, que contribuyeron a que este trabajo no solo se desarrollara sino que también llegara a buen término, especialmente le doy las gracias a mis padres y mi abuela que siempre estuvieron presentes apoyándome y brindándome cada cosa que necesité. Sin ustedes esto no hubiera sido posible. También doy un agradecimiento particular al profesor Leonardo Trujillo por su entrega y tutoría.

Índice general

Índice de tablas

Índice de figuras

Introducción

A lo largo de la historia siempre ha existido un problema con la obtención de estimaciones sobre la proporción de una población que tenga ciertas características sensibles como opinión sobre el aborto, evasión de impuestos, inclinaciones sexuales, consumo de drogas, fraude, entre otras. Esto se debe a que en ocasiones los encuestados no responden con la verdad a la pregunta sensible o se niegan a contestarla; generando así sesgos en la información y llevando a conclusiones erróneas sobre la población.

Hay múltiples razones por las que los encuestados responden con mentiras o por las que se niegan a contestar, éstas van desde miedo, timidez o modestia, hasta multas o sanciones. Por estas situaciones, durante la historia, han surgido técnicas para reducir o evitar estas conductas llevando a respuestas más sinceras sin que genere problemas de sesgo y protegiendo el anonimato del encuestado. Entre estas técnicas, se destacan los trabajos de respuesta aleatorizada de Warner (1965) y sus múltiples extensiones de autores como Greenberg y Horvitz (1976) y las técnicas usuales de conteo de ítems Droitcour et al.(1991).

Greenberg (1969) propone un enfoque diferente a lo mencionado por Warner (1965) en el que una pregunta contiene la característica sensible mientras que la otra hace referencia a una pregunta inocua que no tiene nada que ver con el atributo a estimar, logrando así, no afectar la sensibilidad del entrevistado. Luego Moors (1971) demuestra que la técnica de Greenberg (1969) es más eficiente que la de Warner (1965). Una tercera técnica es presentada por Horvitz, Greenberg y Abernathy (1976) en la que hay una mayor protección del anonimato del encuestado al no incluir una pregunta complementaria sino la selección aleatoria a tres proposiciones: (1) la pregunta sensitiva, (2) contestar si y (3) contestar no. Devore (1977) genera una técnica análoga a la de Greenberg (1969) con la diferencia de que la pertenencia al grupo inocuo se establece con probabilidad uno. En Mangat y Singh (1990) el mecanismo aleatorio proporciona n respuestas independientes con dos componentes aleatorias.

Luego de las técnicas de respuesta aleatorizada surgen las técnicas de conteo de items. Droitcour (1991) propone una técnica que consisten en dividir la muestra en dos partes (grupo control y grupo tratamiento), a cada grupo se le da una lista de items y se le pide al encuestado que responda cuántos le molestan, la diferencia entre cada grupo es que en el segundo se encuentra el atributo sensible, luego de esta instrucción se procede a estimar la diferencia de las medias en cuanto al número de items que se reportan como molestos en cada grupo. Hussain (2012) sugiere proporcionar un listado de preguntas a todos los encuestados que consisten de un item sensible y uno no sensible para que cada individuo cuente 1 si posee alguna de las dos características ó 0 sí no, para que al final reporte la sumatoria de todo el listado de preguntas.

En la parte de estimación de preguntas sensibles multicategóricas Warner (1965) propuso dos, en el primero se aplica el método original dos veces y en el segundo se tienen en cuenta todas las posibles combinaciones entre la respuesta verdadera y la respuesta falsa de cada individuo. Un tercer método es presentado por Abul, Ela, Greenberg y Horvitz (1967) en el que se tienen dos submuestras a las que se les pide que contesten "si", según sea el caso, a la pertenencia del individuo en uno de g grupos. Es de notar que dichas técnicas solo tienen en cuenta un diseño muestral aleatorio simple con reempalzamiento.

En múltiples situaciones es posible codificar observaciones con el fin de asegurar privacidad o la no divulgación de valores reales en las respuestas de los encuestados. Ésta codificación puede hacerse mediante la adición o la multiplicación de una variable aleatoria cuya distribución sea conocida. Warner (1971) sugirió que éste enfoque podía ser usado en modelos de técnica de respuesta aleatorizada para el caso de 2 categorías (presencia/ausencia). Sin embargo, adicionalmente este enfoque tiene la restricción de asumir una selección de la muestra mediante muestreo aleatorio simple con reemplazamiento, con el fin de obtener estimadores sobre la proporción de la población con la característica sensible y sus respectivas estimaciones de la varianza. Siguiendo el trabajo de Kim y Flueck (1978) en esta tesis se proponen estimadores y sus correspondientes varianzas para la estimación de la prevalencia de una pregunta sensible multicategórica. El método se ilustra con los resultados de una encuesta hecha a trabajadores administrativos de la Universidad Nacional de Colombia donde se le preguntó a cada uno sí alguna vez había sufrido de acoso sexual en su trabajo teniendo como opción de respuesta tres características: 1) Nunca he sufrido de acoso sexual en mi trabajo 2) Han intentado acosarme sexualmente pero no lo han logrado y 3) He sufrido de acoso sexual en mi trabajo.

CAPÍTULO 1

Marco Teórico

1.1. Técnicas de respuesta aleatorizada

Debido a altas tasas de no respuesta y sesgos en la información recolectada en estudios que incluyen preguntas sensibles, surgen las Técnicas de Respuesta Aleatorizada que protegen la identidad de los encuestados (anonimato) y asi lograr que los mismos estén más dispuestos a contestar lo que se les pregunta. Estas TRA (sigla que se usará para abreviar Técnicas de Respuesta Aleatorizada) tienen su origen en Warner (1965) quién sugirió una técnica en la que se podía estimar el total de personas en una población que tuviera cierta característica de interés cuidando la confidencialidad y anonimato de los encuestados, factores fundamentales a la hora de abordar este tipo de técnicas. A partir de Warner han pasado por la historia múltiples teorías que agregan un valor a esta primera técnica. Es el caso de Greenberg, Abul-ela, Simmons y Horbitz (1969) quienes propusieron una TRA que difería a la original en que mientras una pregunta corresponde al aspecto sensible, la segunda pregunta se reemplaza por una inocua no relacionada con la pregunta sensible, que va a producir una respuesta afirmativa con probabilidad conocida. O Devore (1977) que genra una técnica análoga a la de Greenberg (1969) con la diferencia de que la pertenencia al grupo inocuo se establece con probabilidad uno.

Luego de las TRA surgen las Técnicas de Conteo de Items o TCIs que consisten en contar, de un número de ítems que generan molestia (Droitcour et al. (1991)) o de un listado de preguntas que consisten de un ítem sensible y uno no sensible pedirle a cada individuo que cuente 1 si posee alguna de las dos características ó 0 sí no, para que al final reporte la sumatoria de todo el listado.

También se han propuesto algunas técnicas, explicadas en este capítulo, que hacen referencia a estimar la proporción de individuos con atributos sensibles en preguntas multicategóricas, donde el método de fondo aplicado es el de Warner (1965). Todo estos cambios o mejoras se han dado y probado a lo largo del tiempo, luego, en los siguientes apartados se mencionan los más significativos y por tanto más importantes.

1.1.1. Técnica de Warner

Warner (1965) desarrolló esta técnica para disminuir el sesgo en estudios con una caraterística sensible e incrementar la tasa de respuesta. La técnica consiste de dos preguntas complementarias A (¿Usted pertenece al grupo sensible?) y A^c (¿Usted no pertenece al grupo sensible?). Se asume un muestreo aleatorio simple con reemplazamiento, se le asigna aleatoreamente una de las dos preguntas (A, A^c) a cada encuestado y se le solicita que conteste "si", sí su estado actual coincide con la pregunta seleccionada y "no" en otro caso.

Un estimador insesgado de la proporción de la población de individuos con el ítem sensible es:

$$\hat{\pi}_W = \frac{\hat{\theta} - (1-p)}{2p-1} \tag{1.1}$$

con $p \neq 1/2$.

Donde p es la probabilidad de seleccionar la pregunta A, y $\hat{\theta} = \frac{n'}{n}$ con n' el número de "si" obtenidos en una muestra de tamaño n. La varianza de este estimador es:

$$Var(\hat{\pi}_W) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{1.2}$$

con $p \neq 1/2$.

Por lo tanto, un estimador de esta varianza puede denotarse:

$$\hat{Var}(\hat{\pi}_W) = \frac{\hat{\pi}_W(1-\hat{\pi}_W)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{1.3}$$

con $p \neq 1/2$.

Supuestos

- Todos los eventos son independientes.
- El número de individuos encuestados debe ser grande.
- Los entrevistados deben comprender perfectamente el procedimiento y seguirlo correctamente.

Ventajas y desventajas de la técnica de Warner:

- **Ventajas:**
  - Aumenta la probabilidad de la veracidad en la respuesta de cada entrevistado.
  - Aumenta la confidencialidad del encuestado.
  - Disminuye la tasa de no respuesta.
- **Desventajas:**
  - Aumento en el grado de dificultad de la pregunta.
  - Dificultad al comprender el método de aleatorización.
  - Requiere de muestras de tamaños grandes.

Ver demostración del estimador de Warner, su varianza y la propiedad de insesgamiento del estimador en apéndice de demostraciones.

1.1.2. Técnica de Warner extendida a poblaciones finitas

Särndal, Swensson y Wretman (1992, página 570) extendieron el método de Warner para cualquier diseño muestral probabilístico.

Sea y_k una variable indicadora desconocida para cada individuo en una población de tamaño N que se define:

$$y_k = \begin{cases} 1, & \text{si el k-ésimo individuo pertenece al grupo A;} \\ 0, & \text{si el k-ésimo individuo pertenece al grupo } A^c. \end{cases}$$

Además también se define una variable indicadora para cada individuo de la muestra:

$$x_k = \begin{cases} 1, & \text{si el k-ésimo individuo dentro de la muestra contesta “si”;} \\ 0, & \text{si el k-ésimo individuo dentro de la muestra contesta “no”.} \end{cases}$$

Esta variable es conocida y se observa sin revelar si contestó a A ó A^c, de manera que se protege la confidencialidad del encuestado.

La probabilidad con la que el mecanismo aleatorio selecciona una de las dos preguntas A ó A^c es conocida y su valor es P.

La relación entre x_k y y_k es de la forma:

$$\begin{aligned} Si\ x_k = 1, y_k &= 1 \Rightarrow A \\ Si\ x_k = 1, y_k &= 0 \Rightarrow A^c \\ Si\ x_k = 0, y_k &= 1 \Rightarrow A^c \\ Si\ x_k = 0, y_k &= 0 \Rightarrow A \end{aligned}$$

Lo que quiere decir que en los casos en los que el individuo conteste afirmativamente a la pregunta de pertenecer al grupo A o negativamente a la pregunta de pertenecer al grupo A^c, el individuo expresará la tenencia de la característica sensible.

Ahora para obtener el estimador se tiene:

$$\begin{aligned} P(x_k = 1) &= y_k P + (1 - y_k)(1 - P) \\ &= y_k P + 1 - P - y_k + y_k P \\ &= 2y_k P + 1 - P - y_k \\ &= 1 - P + (2P - 1)y_k \end{aligned}$$

Se iguala a cero y se despeja y_k,

$$\hat{y}_k = \frac{x_k + P - 1}{2P - 1} \tag{1.4}$$

Asumiendo $P \neq 1/2$.

Luego $\hat{y}_k$ es un estimador insesgado de y_k bajo respuesta aleatorizada, es decir,

$$\begin{aligned} E_{RA}[\hat{y}_k] &= E_{RA}\left[\frac{x_k + P - 1}{2P - 1}\right] \\ &= \frac{E_{RA}[x_k] + P - 1}{2P - 1} \\ &= \frac{1 - P + (2P - 1)y_k + P - 1}{2P - 1} \\ &= y_k \end{aligned}$$

$$E_{RA}[\hat{y}_k] = y_k \tag{1.5}$$

con varianza,

$$\begin{aligned} V_{RA}[\hat{y}_k] &= V_{RA}\left[\frac{x_k + P - 1}{2P - 1}\right] \\ &= \frac{V_{RA}[x_k]}{(2P - 1)^2} \\ &= \frac{P(1 - P)}{(2P - 1)^2} \\ &= V_0 \end{aligned}$$

$$V_{RA}[\hat{y}_k] = V_0 \tag{1.6}$$

Sí la presencia / ausencia de la característica sensible fuera conocida para cada individuo en la muestra sin error, el estimador de Horvitz-Thompson sería aplicable en este caso como,

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$$

para cualquier diseño $p(s)$ con probabilidades de inclusión π_k y π_{kl}.

Ver demostración de la propiedad de insesgamiento del estimador de Horvitz-Thompson y la de su varianza en apéndice de demostraciones.

Volviendo al estimador de interés, ya que no se cuenta con y_k sino con una estimación del mismo para cada individuo se tiene que el estimador bajo respuesta aleatorizada se puede escribir como,

$$\hat{t}_{RA} = \sum_s \frac{\hat{y}_k}{\pi_k}$$

con $\hat{y}_k$ igual a 1.4.

Por 1.5 y hallando el valor esperado del estimador se llega a que es insesgado para la población total $t = \sum_U y_k$

$$\begin{aligned} E[\hat{t}_{RA}] &= E_p\left[E_{RA}\left(\sum_s \frac{\hat{y}_k}{\pi_k}|s\right)\right] \\ &= E_p\left[\sum_s \frac{E_{RA}(\hat{y}_k)}{\pi_k}|s\right] \\ &= E_p\left[\sum_s \frac{y_k}{\pi_k}\right] \\ &= \sum_U y_k \\ &= t \end{aligned}$$

donde el subindice p significa "respecto al diseño muestral $p(s)$".

□

Teniendo en cuenta 1.6 la varianza es igual a:

$$\begin{aligned} V[\hat{t}_{RA}] &= V_p[E_{RA}(\hat{t}_{RA}|s)] + E_p[V_{RA}(\hat{t}_{RA}|s)] \\ &= V_p\left[\sum_s \frac{y_k}{\pi_k}\right] + E_p\left[V_{RA}\left(\sum_s \frac{\hat{y}_k}{\pi_k}|s\right)\right] \\ &= \sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l + E_p\left[\sum_s \frac{V_{RA}(\hat{y}_k)}{\pi_k}|s\right] \\ &= \sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l + E_p\left[\sum_s \frac{P(1-P)}{(2P-1)^2\pi_k}|s\right] \\ &= \sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l + \left[\sum_U \frac{1}{\pi_k}\right] V_0 \end{aligned}$$

Donde, $V_0 = V_{RA}[\hat{y}_k]$

□

Nótese que este estimador $\hat{t}_{RA}$ es aplicable bajo cualquier diseño muestral complejo, extendiendo asi, el supuesto de Warner (1965) quien solo lo había considerado para un diseño muestral aleatorio simple con reemplazamiento.
Por otro lado, el término extra en la varianza del estimador puede verse como el precio que se paga por aleatorizar la respuesta y es de notar que será mayor conforme P se acerque a $1/2$ por lo tanto debe elegirse a P tal que se aleje de este número.

1.1.3. Técnica de Greenberg et al.

Esta técnica fue propuesta por Greenberg, Abul-ela, Simmons y Horvitz (1969). Para esta técnica se amplió y perfeccionó el modo de aleatorizar las respuestas de los encuestados. Al igual que la TRA de Warner, el mecanismo aleatorio selecciona una de dos preguntas con la diferencia que Greenberg et al. (1969) considera una segunda pregunta que no tiene nada que ver con la característica de interés, es decir, sí la pregunta sensible es ¿Usted se ha practicado o se practicaría un aborto?, una segunda pregunta inocua podría ser ¿Usted habla inglés? lo que va a producir una respuesta afirmativa con probabilidad conocida.

La pregunta sensitiva vuelve a tener probabilidad p de ser seleccionada, luego la inocua tiene el complemento $1-p$. Sí se asume que la proporción de encuestados con la característica no relacionada π_y es conocida entonces la probabilidad de que el encuestado conteste "si" es

$$P(X_i = 1) = \pi p + \pi_y(1-p) \tag{1.7}$$

Se tiene que n_1 es el número de respuestas "si" de una muestra de tamaño n, luego,

$$\frac{n_1}{n} = \hat{\pi} p + \pi_y(1-p) \tag{1.8}$$

Y al despejar π se llega a que un estimador insesgado de la proporción de la población con la característica sensible está dada por

$$\hat{\pi} = \frac{\frac{n_1}{n} - (1-p)\pi_y}{p} \tag{1.9}$$

con varianza,

$$Var(\hat{\pi}) = \frac{(\pi p + \pi_y(1-p))(1 - \pi p - \pi_y(1-p))}{np^2} \tag{1.10}$$

Ahora, en el caso de que la proporción de la población con la característica no relacionada π_y sea desconocida, Greenberg, Abul-ela, Simmons y Horvitz sugirieron tomar dos muestras independientes bajo muestreo aleatorio simple con reemplazamiento de tamaños n_1 y n_2 con $n_1 + n_2 = n$.

Para la $i - ésima$ muestra se tiene que p_i y $1 - p_i$ denotan las probabilidades de seleccionar las sentencias de poseer la característica sensible A o la no relacionada Y en la decisión aleatoria R_i usada por los encuestados en la $i - ésima$ muestra con $i = 1, 2$.

La probabilidad de que el encuestado conteste "si" está dado por

$$P(X_i = 1) = \pi p_i + \pi_y(1 - p_i) \quad con \quad i = 1, 2 \tag{1.11}$$

Se sabe que,

$$\frac{n_1}{n} = \hat{\pi} p_1 + \pi_y(1 - p_1) \tag{1.12}$$

y

$$\frac{n_2}{n} = \hat{\pi} p_2 + \pi_y(1 - p_2) \tag{1.13}$$

Luego un estimador insesgado de la proporción de individuos con el atributo sensible es

$$\hat{\pi} = \frac{(1 - p_2)\frac{n_1}{n} - (1 - p_1)\frac{n_2}{n}}{p_1 - p_2} \tag{1.14}$$

con varianza,

$$Var(\hat{\pi}) = \frac{1}{(p_1 - p_2)^2}\left[\frac{(1 - p_2)^2)\theta_1(1 - \theta_1)}{n_1} + \frac{(1 - p_1)^2)\theta_2(1 - \theta_2)}{n_2}\right] \tag{1.15}$$

donde $\theta_i = \pi p_i + \pi_y(1 - p_i)$, $i = 1, 2$

1.1.4. Técnica de Horvitz, Greenberg y Abernathy

Horvitz, Greenberg y Abernathy (1976) propusieron en esta técnica no utilizar una pregunta complementaria, sino la opción de asignar mediante un mecanismo aleatorio una de tres alternativas:

- Pregunta sensible
- Contestar "si"
- Contestar "no"

Esto con probabilidades p_1, p_2 y p_3 donde $p_1 + p_2 + p_3 = 1$ teniendo en cuenta las siguientes restricciones:

$$1 > p = p_1 > \frac{1}{2}$$
$$0 < 1 - p_1 - p_2 \leq p_2 < 1 - p_1$$
$$\Leftrightarrow$$
$$1 - p_1 - 2p_2 \leq 0$$
$$p_2 < 1 - p_1$$

Ahora se define X_k como la variable aleatoria del número de individuos en cada grupo

$$\text{Sea } x_k = \begin{cases} y_k, & \text{con probabilidad } p_1; \\ 1, & \text{con probabilidad } p_2; \\ 0, & \text{con probabilidad } 1 - p_1 - p_2. \end{cases}$$

Luego

$$E(x_k) = y_k p_1 + 1 p_2 + 0$$

Despejando se obtiene que el estimador de la proporción de individuos con el atributo sensible es:

$$\hat{y_k} = \frac{x_k - p_2}{p_1} \tag{1.16}$$

donde su varianza está dada por:

$$Var(\hat{y_k}) = \frac{y_k(1 - p_1 - 2p_2)}{p_1} + \frac{p_2(1 - p_2)}{p_1^2} \tag{1.17}$$

1.1.5. Técnica de Devore

Devore (1977) propuso esta técnica, análoga a la de Greenberg (1969), con la diferencia que la probabilidad de que se pertenezca al grupo inocuo W es 1. Se define X_k como la variable aleatoria del número de individuos en cada grupo:

$$\text{Sea } x_k = \begin{cases} y_k, & \text{con probabilidad } p; \\ 1, & \text{con probabilidad } 1 - p. \end{cases}$$

Luego

$$E(x_k) = y_k p + 1 - p \tag{1.18}$$

Despejando se obtiene que el estimador de la proporción de individuos con el atributo sensible es

$$\hat{y_k} = \frac{x_k - (1 - p)}{p} \tag{1.19}$$

donde su varianza está dada por

$$Var(\hat{y_k}) = \frac{(1 - y_k)p(1 - p)}{p^2}$$

1.1.6. Técnica de Mangat y Singh

En este método propuesto por Mangat y Singh (1990) el mecanismo aleatorio proporciona n respuestas independientes con dos componentes aleatorias.

A cada encuestado se le proporcionan dos decisiones, La decisión aleatoria R_1 que consiste de dos estamentos:

1. Pertenencia al grupo sensible
2. Ir a la decisión aleatoria R_2

con probabilidades T y $1-T$ respectivamente. Por otro lado la decisión aleatoria R_2 contiene dos opciones de respuesta:

1. Pertenencia al grupo sensible
2. No pertenencia al grupo sensible

lo que se reduce al método de Warner(1965) con probabilidades p y $1-p$.

En los casos: R_1 opción 1, R_2 opción 1 y R_2 opción 2 se le pide al encuestado que conteste "si" o "no" según su condición.

Ahora, como antes, se define x_k como la variable aleatoria del número de individuos en cada grupo:

$$\text{Sea } x_k = \begin{cases} y_k, & \text{con probabilidad T;} \\ py_k + (1-p)(1-y_k), & \text{con probabilidad 1-T.} \end{cases}$$

Luego la probabilidad de "si" x_k está dada por:

$$E(x_k) = y_k T + (1-T)[py_k + (1-p)(1-y_k)]$$

Finalmente el estimador de la proporción de la población que tienen el atributo sensible es:

$$\hat{y_k} = \frac{x_k - (1-T)(1-p)}{T + (1-T)(2p-1)} \tag{1.20}$$

Con varianza dada por la expresión:

$$Var(\hat{y_k}) = \frac{(1-T)(1-p)[T + p(1-T)]}{[T + (1-T)(2p-1)]^2} \tag{1.21}$$

1.1.7. Técnica de conteo de items

1.1.7.1. Técnica de conteo de ítems usual de Droitcour et al.

Ésta técnica ideada por Droitcour et al. (1991) consiste en tomar dos submuestras independientes de tamaño n_1 y n_2 de una población de tamaño n. Se le da al $i-ésimo$ encuestado de la primera submuestra una lista de g ítems y se le solicita que cuente cuántos de ellos le molestan o incomodan (a este valor se le llama X_i) y análogamente se hace con el $j-ésimo$ encuestado de la segunda submuestra (a este valor se le llama Y_j) con la diferencia de que a esta lista se le agrega la característica o ítem sensible, es decir que tendrán una lista de $(g+1)$ ítems.
Ahora un estimador insesgado de la proporción de la población que respondió de forma afirmativa al ítem sensible es:

$$\hat{y}_k = \bar{Y} - \bar{X} \tag{1.22}$$

Donde $\bar{Y}$ y $\bar{X}$ corresponden a las medias de ambas submuestras. Luego su varianza es

$$Var(\hat{y}_k) = \frac{\pi(1-\pi)}{n_2} + \frac{n\sum_{j=1}^{g}\theta_j\left(1-\sum_{j=1}^{g}\theta_j\right)}{n_1 n_2} + \frac{n\sum_{j,k=1,j\neq k}^{g}\theta_j\theta_k}{n_1 n_2} \tag{1.23}$$

Donde θ_j es la proporción conocida del ítem j en la población.

1.1.7.2. Técnica de conteo de ítems de Hussain, Shah, Shabbir

Hussain, Shah, Shabbir (2012) plantearon que a cada encuestado se le entrega un cuestionario que consiste de g preguntas con $g \geq 2$. Cada pregunta consiste de un ítem sensible y un ítem no sensible y se solicita al encuestado que cuente 1 si posee alguna de las dos características o 0 en otro caso y al final dar el conteo total de las preguntas a las que otorgó un 1.

Sea

$$Z_i = \sum_{j=1}^{g} a_j \tag{1.24}$$

Donde Z_i el conteo total del $i-ésimo$ encuestado y a_i que toma valores 1 y 0 con probabilidades $(1-\pi+\theta_j-\pi\theta_j)$ y $(\pi+\theta_j-\pi\theta_j)$ respectivamente.

Luego el estimador insesgado para la proporción de la poblacion de personas que presentan el ítem sensible es:

$$\hat{\pi}_p = \frac{\bar{Z} - \sum_{j=1}^{g}\theta_j}{g - \sum_{j=1}^{g}\theta_j} \tag{1.25}$$

Y su varianza

$$Var(\hat{\pi}_p) = \frac{\pi(1-\pi)}{n} + \frac{1-\pi}{n\left(g - \sum_{j=1}^{g} \theta_j\right)^2} \left[\left(\sum_{j=1}^{g} \theta_j \right) \left(1 - \sum_{j=1}^{g} \theta_j \right) + \sum_{\substack{j,k=1 \\ j \neq k}}^{g} \theta_j \theta_k \right] \quad (1.26)$$

En algunos casos se puede tener que la proporción de los ítems no sensibles sean iguales, en tal caso se tiene que $\theta_j = \frac{1}{g}$ con lo que resulta la expresión de la varianza de la siguiente forma:

$$Var(\hat{\pi}_p) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)}{ng(g-1)} \quad (1.27)$$

1.2. Técnicas de respuesta aleatorizada para variables multicategóricas

1.2.1. Técnica de Warner doble

Esta técnica consiste en aplicar el método original de Warner (1965) dos veces al mismo grupo de encuestados. La primera aplicación se emplea para estimar π_1 y la segunda para estimar π_2. De esta forma los estimadores para π_1 y π_2 seguirán la expresión,

$$\hat{\pi}_i = \frac{\hat{\theta} - (1 - p_i)}{2p_i - 1} \quad (1.28)$$

con i=1,2 y $p_i \neq 1/2$.

En 1.28 p_i es la probabilidad de seleccionar la pregunta A, y $\hat{\theta} = \frac{n'}{n}$ con n' el número de "si" obtenidos en una muestra de tamaño n. Luego, como la suma de todos los estimadores debe ser igual a 1 se tiene que

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 \quad (1.29)$$

La varianza de $\hat{\pi}_1$ y $\hat{\pi}_2$ está dada por:

$$Var(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n} + \frac{p_i(1-p_i)}{n(2p_i-1)^2} \quad (1.30)$$

con i=1,2 y $p_i \neq 1/2$.

Un estimador de esta varianza está dado por,

$$\hat{Var}(\hat{\pi}_i) = \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n} + \frac{p_i(1-p_i)}{n(2p_i-1)^2} \quad (1.31)$$

De la expresión 1.30 se deriva la varianza de $\hat{\pi}_3$,

$$\begin{aligned} Var(\hat{\pi}_3) &= Var(1-\hat{\pi}_1-\hat{\pi}_2) \\ &= 0+Var(-\hat{\pi}_1-\hat{\pi}_2) \\ &= Var(\hat{\pi}_1)+Var(\hat{\pi}_2)+2cov(\hat{\pi}_1,\hat{\pi}_2) \end{aligned}$$

Luego se reemplaza el resultado de 1.30 para llegar a que,

$$Var(\hat{\pi}_3)=\sum_{i=1}^{2}\left(\frac{\pi_i(1-\pi_i)}{n}+\frac{p_i(1-p_i)}{n(2p_i-1)^2}\right)+2\left(-\frac{\pi_1\pi_2}{n}\right) \tag{1.32}$$

□

Con una varianza estimada dada por,

$$\hat{Var}(\hat{\pi}_3)=\sum_{i=1}^{2}\left(\frac{\hat{\hat{\pi}}_i(1-\hat{\pi}_i)}{n}+\frac{p_i(1-p_i)}{n(2p_i-1)^2}\right)+2\left(-\frac{\hat{\pi}_1\hat{\pi}_2}{n}\right) \tag{1.33}$$

1.2.2. Modelo de contaminación extendido

Este modelo corresponde a una extensión del modelo de Warner (1965). La tabla 1.1 muestra las 6 posibles combinaciones entre grupo reportado y grupo real que puede tener cada individuo donde 1 representa "si" y 0 "no".

TABLA 1.1. Transformaciones de respuesta para el modelo de contaminación extendido

Grupo Real	Grupo Reportado					
	1	2	3	4	5	6
	123	123	123	123	123	123
1	100	100	001	010	001	010
2	010	001	010	100	100	001
3	001	010	100	001	010	100

Cada una de las 6 matrices hace referencia a la respuesta de un encuestado sobre su grupo verdadero y su grupo falso, lo que se conoce como respuesta contaminada. De manera que si un encuestado aleatoriamente selecciona la matriz 2 y sí él pertenece al grupo 2 entonces debe reportar "3". De la tabla 1.1 se puede deducir que $p_i = P(Se\ le\ pida\ al\ encuestado\ en\ el\ grupo\ i\ que\ reporte\ R) = 1/3$ donde $R = 1, 2, 3$ es la respuesta reportada por el encuestado.

Con base en el supuesto anterior se reemplaza $p_i = 1/3$ en el estimador de Warner 1.1 con lo que se genera el estimador del modelo de contaminación extendido,

$$\begin{aligned}\hat{\pi}_i &= \frac{\hat{\theta}_i - (1-(1/3))}{2(1/3)-1} \\ &= \frac{\hat{\theta}_i - (2/3)}{(2/3)-1} \\ &= \frac{(3\hat{\theta}_i - 2)/3}{-1/3}\end{aligned}$$

$$\hat{\pi}_i = -3\hat{\theta}_i + 2 \tag{1.34}$$

Con $i = 1, 2$ donde $\hat{\theta}_i = \frac{n'_i}{n}$ con n'_i el número de "si" obtenidos en una muestra de tamaño n para el grupo i. Luego, como la suma de todos los estimadores debe ser igual a 1 se tiene que,

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 \tag{1.35}$$

Ahora las varianzas se obtienen reemplazando p_i en 1.2, obteniendo,

$$\begin{aligned}Var(\hat{\pi}_i) &= \frac{\pi_i(1-\pi_i)}{n} + \frac{(1/3)(1-(1/3))}{n(2(1/3)-1)^2} \\ &= \frac{\pi_i(1-\pi_i)}{n} + \frac{2/9}{n(2/3-1)^2} \\ &= \frac{\pi_i(1-\pi_i)}{n} + \frac{2/9}{n(1/9)}\end{aligned}$$

Luego,

$$Var(\hat{\pi}_i) = \frac{\pi_i(1-\pi_i)}{n} + \frac{2}{n} \tag{1.36}$$

Con varianza estimada igual a,

$$\hat{Var}(\hat{\pi}_i) = \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n} + \frac{2}{n} \tag{1.37}$$

De la expresión 1.35 se deriva la varianza de $\hat{\pi}_3$,

$$\begin{aligned}Var(\hat{\pi}_3) &= Var(1 - \hat{\pi}_1 - \hat{\pi}_2) \\ &= 0 + Var(-\hat{\pi}_1 - \hat{\pi}_2) \\ &= Var(\hat{\pi}_1) + Var(\hat{\pi}_2) + 2cov(\hat{\pi}_1, \hat{\pi}_2)\end{aligned}$$

Ahora,

$$Var(\hat{\pi}_3) = \frac{\pi_i(1-\pi_i)}{n} + \frac{2}{n} + 2\left(-\frac{1+\pi_1\pi_2}{n}\right) \tag{1.38}$$

con varianza estimada dada por,

$$\hat{Var}(\hat{\pi}_3) = \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n} + \frac{2}{n} + 2\left(-\frac{1+\hat{\pi}_1\hat{\pi}_2}{n}\right) \tag{1.39}$$

1.2.3. Modelo de multiproporciones

Abul, Ela, Greenberg y Horvitz (1967) propusieron este modelo que consiste de 2 submuestras y la siguiente estructura de preguntas:

Q1 Soy un miembro del grupo 1

Q2 Soy un miembro del grupo 2

Q3 Soy un miembro del grupo 3

En la submuestra i con $i = 1, 2$, $Q1$ es seleccionado con probabilidad p_{i1}, $Q2$ con probabilidad p_{i2} y $Q3$ con probabilidad p_{i3} donde,

$$\sum_{j=1}^{3} p_{ij} = 1 \tag{1.40}$$

La probabilidad λ con la que el encuestado contesta "si" a la afirmación seleccionada está dada por:

$$\lambda_i = (p_{i1} - p_{i3})\pi_1 + (p_{i2} - p_{i3})\pi_2 + p_{i3} \qquad con \qquad i = 1, 2 \tag{1.41}$$

Luego, los estimadores están dados por las siguientes expresiones,

$$\hat{\pi}_1 = \frac{(\hat{\lambda}_1 - p_{13})(p_{22} - p_{23}) - (\hat{\lambda}_2 - p_{23})(p_{12} - p_{13})}{(p_{11} - p_{13})(p_{22} - p_{23}) - (p_{21} - p_{23})(p_{12} - p_{13})} \tag{1.42}$$

$$\hat{\pi}_2 = \frac{(\hat{\lambda}_1 - p_{13})(p_{21} - p_{23}) - (\hat{\lambda}_2 - p_{23})(p_{11} - p_{13})}{(p_{11} - p_{13})(p_{22} - p_{23}) - (p_{21} - p_{23})(p_{12} - p_{13})} \tag{1.43}$$

donde $\hat{\lambda}_i$ es la proporción de individuos que contestan "si" en la muestra i con $i = 1, 2$. Luego la estimación para π_3 es,

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 \tag{1.44}$$

Con varianzas,

$$Var(\hat{\pi}_1) = \frac{1}{P^2}\left((p_{22}-p_{23})^2\frac{\lambda_1(1-\lambda_1)}{n_1} + (p_{12}-p_{13})^2\frac{\lambda_2(1-\lambda_2)}{n_2}\right) \tag{1.45}$$

$$Var(\hat{\pi}_2) = \frac{1}{P^2}\left((p_{21}-p_{23})^2\frac{\lambda_1(1-\lambda_1)}{n_1} + (p_{11}-p_{13})^2\frac{\lambda_2(1-\lambda_2)}{n_2}\right) \tag{1.46}$$

$$Var(\hat{\pi}_3) = \frac{1}{P^2}\left((p_{22}-p_{21})^2\frac{\lambda_1(1-\lambda_1)}{n_1} + (p_{12}-p_{11})^2\frac{\lambda_2(1-\lambda_2)}{n_2}\right) \tag{1.47}$$

donde $P = (p_{11}-p_{13})(p_{22}-p_{23}) - (p_{21}-p_{23})(p_{12}-p_{13})$ y n_i es el tamaño de la submuestra i con $i = 1, 2$. Ahora la covarianza entre $\hat{\pi}_1$ y $\hat{\pi}_2$ está dada por,

$$Cov(\hat{\pi}_1, \hat{\pi}_2) = \frac{1}{P^2}\left((p_{22}-p_{23})(p_{23}-p_{21})\frac{\lambda_1(1-\lambda_1)}{n_1} + (p_{12}-p_{13})(p_{13}-p_{11})\frac{\lambda_2(1-\lambda_2)}{n_2}\right) \tag{1.48}$$

CAPÍTULO 2

Modelo aditivo de respuesta aleatorizada en poblaciones finitas

En muchas ocasiones es posible codificar observaciones con el fin de asegurar privacidad o la no divulgación de valores reales de las respuestas de los encuestados. Ésta codificación puede hacerse mediante la adición o la mutliplicación de una variable aleatoria cuya distribución es conocida. El modelo aditivo de Kim y Flueck (1978) supone que en un cuestionamiento directo un encuestado puede elegir ser miembro de uno y solo uno de G grupos. Sin embargo, el encuestado no lo hace cuando se trata de un ítem sensible; por esto se le pide al encuestado que codifique su respuesta con el fin de aumentar la medida de confidencialidad. A continuación se presenta este modelo para el caso de 3 grupos y luego para un número finito G de grupos. En este capítulo se muestra el objetivo principal del trabajo donde se extiende la teoría de Kim y Flueck (1978) para cualquier diseño muestral complejo. El método original solo fue propuesto bajo un diseño muestral aleatorio simple con reemplazamiento.

2.1. Método propuesto para el caso $G = 3$

Siguiendo a Kim y Flueck (1978), quienes propusieron el método con un Muestreo Aleatorio Simple con Reemplazamiento, el presente trabajo busca llevarlo a cualquier diseño complejo $P(s)$. Sea g_k el grupo verdadero del $k-ésimo$ encuestado donde $g_k = 1, 2, 3$ con $k \in U$; y a_k el valor de aumento seleccionado aleatoriamente con $a_k = 1, 2, 3$. Este valor de aumento se puede fijar mediante un mecanismo aleatorio como un dado o una baraja de cartas que tenga 3 posibles valores; de tal modo que si es con una baraja de cartas el encuestado entienda que deberá sumar a su respuesta verdadera este número seleccionado y así codificar su respuesta. Finalmente la respuesta codificada del encuestado cuyo grupo verdadero es g_k tiene la expresión:

$$z_k = g_k + a_k \tag{2.1}$$

De modo que a la respuesta del encuestado se le sumará un número cuya probabilidad es conocida y es la misma para todos los individuos.

Para garantizar mayor confidencialidad, z_k es transformado por el mismo encuestado al valor reportado R_k, donde

$$R_k = \begin{cases} z_k & \text{si } z_k \leq 3, \\ z_k - 3 & \text{si } z_k > 3. \end{cases} \tag{2.2}$$

La tabla 2.1 presenta los valores posibles reportados y su origen $(g_k + a_k)$.

TABLA 2.1. Transformaciones de la respuesta para el modelo aditivo

Número reportado R_k	Origen $(g_k + a_k)$		
1	1+3	2+2	3+1
2	1+1	2+3	3+2
3	1+2	2+1	3+3

Sea $\phi_{g_k} = P(k \in g)$ el parámetro a estimar y $P_a = P(k \textit{ seleccione } a)$ que se asume predeterminado, donde k representa a cualquier encuestado; por otro lado $\lambda_{rk} = P(R_k = r)$ lo que representa la probabilidad de que el encuestado reporte un valor r_k de la tabla anterior. Ahora se considera que aunque P_1, P_2 y P_3 se mantienen fijos para todos los individuos, se tiene solo una muestra probabilística de encuestados y por lo tanto es necesario definir para un individuo en particular su probabilidad λ_{rk} asi:

$$\begin{aligned} \lambda_{1k} &= P_3\phi_{1k} + P_2\phi_{2k} + P_1\phi_{3k} \\ \lambda_{2k} &= P_1\phi_{1k} + P_3\phi_{2k} + P_2\phi_{3k} \\ \lambda_{3k} &= P_2\phi_{1k} + P_1\phi_{2k} + P_3\phi_{3k} \end{aligned}$$

Estas probabilidades λ_{rk} son desconocidas por lo que dependen de valores ϕ_{rk} que también son desconocidos y se quieren estimar.

Demostración:

Se sabe que,

$$\lambda_{1k} + \lambda_{2k} + \lambda_{3k} = 1$$

Luego

$$\lambda_{3k} = 1 - \lambda_{1k} - \lambda_{2k}$$

Ahora se reemplaza a λ_{2k} y a λ_{3k} y se desarrolla el sistema hasta obtener λ_{1k},

$$\begin{aligned} (P_2\phi_{1k} + P_1\phi_{2k} + P_3\phi_{3k}) &= 1 - \lambda_{1k} - (P_1\phi_{1k} + P_3\phi_{2k} + P_2\phi_{3k}) \\ \phi_{1k} &= 1 - (P_2\phi_{1k} + P_1\phi_{2k} + P_3\phi_{3k}) - (P_1\phi_{1k} + P_3\phi_{2k} + P_2\phi_{3k}) \\ &= 1 - P_2\phi_{1k} - P_1\phi_{2k} - P_3\phi_{3k} - P_1\phi_{1k} - P_3\phi_{2k} - P_2\phi_{3k} \\ &= 1 - P_1(\phi_{1k} + \phi_{2k}) - P_2(\phi_{1k} + \phi_{3k}) - P_3(\phi_{2k} + \phi_{3k}) \end{aligned}$$

Ya que,

$$\phi_{1k} + \phi_{2k} + \phi_{3k} = 1$$

y

$$\phi_{3k} = 1 - \phi_{1k} - \phi_{2k}$$

entonces,

$$\begin{aligned}\phi_{1k} &= 1 - P_1(\phi_{1k} + \phi_{2k}) - P_2(\phi_{1k} + 1 - \phi_{1k} - \phi_{2k}) - P_3(\phi_{2k} + 1 - \phi_{1k} - \phi_{2k}) \\ &= 1 - P_1(\phi_{1k} + \phi_{2k}) - P_2(1 - \phi_{2k}) - P_3(1 - \phi_{1k}) \\ &= 1 - P_1\phi_{1k} - P_1\phi_{2k} - P_2 + P_2\phi_{2k} - P_3 + P_3\phi_{1k} \\ &= 1 + \phi_{1k}(P_3 - P_1) + \phi_{2k}(P_2 - P_1) - P_2 - P_3\end{aligned}$$

Finalmente como,

$$P_1 + P_2 + P_3 = 1$$

y

$$P_1 = 1 - P_2 - P_3$$

entonces,

$$\lambda_{1k} = P_1 + (P_3 - P_1)\phi_{1k} + (P_2 - P_1)\phi_{2k}$$

Análogamente se llega a:

$$\lambda_{2k} = P_2 + (P_1 - P_2)\phi_{1k} + (P_3 - P_2)\phi_{2k}$$

□

Luego en forma matricial para cada individuo se tiene,

$$\begin{pmatrix} \lambda_{1k} - P_1 \\ \lambda_{2k} - P_2 \end{pmatrix} = \begin{pmatrix} P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_3 - P_2 \end{pmatrix} \begin{pmatrix} \phi_{1k} \\ \phi_{2k} \end{pmatrix}$$

equivalente a

$$\Lambda_k = P\Phi \tag{2.3}$$

Es de notar que $|P| = 0$ cuando $P_1 = P_2 = P_3 = 1/3$, por lo tanto se asume que $|P| \neq 0$, luego para estimar Φ que corresponde a la matriz de probabilidades de que un individuo en la población pertenezca a cada una de las categorías de la característica sensible, puesto que solo se observa $\hat{\Lambda}$ y los elementos de la matriz P son fijos se tiene:

$$\hat{\Lambda}_k = \begin{pmatrix} \hat{y}_{1k} - P_1 \\ \hat{y}_{2k} - P_2 \end{pmatrix} \tag{2.4}$$

donde

$$\hat{y}_{rk} = \begin{cases} 1 & \text{si } R_k = r \\ 0 & e.o.c. \end{cases}$$

por lo tanto $\hat{y}_{rk} \sim Be(\lambda_{rk})$ con

$$\begin{aligned}E(\hat{y}_{rk}) &= \lambda_{rk} \\ V(\hat{y}_{rk}) &= \lambda_{rk}(1 - \lambda_{rk})\end{aligned}$$

Ahora para llegar al estimador deseado se tiene la siguiente ecuación matricial,

$$\hat{\Phi}_k = P^{-1}\hat{\Lambda}_k \tag{2.5}$$

equivalente a:

$$\begin{pmatrix} \hat{\phi}_{1k} \\ \hat{\phi}_{2k} \end{pmatrix} = \begin{pmatrix} P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_3 - P_2 \end{pmatrix}^{-1} \begin{pmatrix} \hat{y}_{1k} - P_1 \\ \hat{y}_{2k} - P_2 \end{pmatrix}$$

Se tiene que

$$\begin{aligned} \begin{pmatrix} P_3 - P_1 & P_2 - P_1 \\ P_1 - P_2 & P_3 - P_2 \end{pmatrix}^{-1} &= \frac{1}{|P|} \begin{pmatrix} P_3 - P_2 & -(P_2 - P_1) \\ -(P_1 - P_2) & P_3 - P_1 \end{pmatrix}^T \\ &= \frac{1}{|P|} \begin{pmatrix} P_3 - P_2 & P_1 - P_2 \\ P_2 - P_1 & P_3 - P_1 \end{pmatrix}^T \end{aligned}$$

Luego,

$$\begin{pmatrix} \hat{\phi}_{1k} \\ \hat{\phi}_{2k} \end{pmatrix} = \frac{1}{|P|} \begin{pmatrix} P_3 - P_2 & P_1 - P_2 \\ P_2 - P_1 & P_3 - P_1 \end{pmatrix} \begin{pmatrix} \hat{y}_{1k} - P_1 \\ \hat{y}_{2k} - P_2 \end{pmatrix}$$

Ahora, desarrollando el producto matricial se obtiene lo siguiente:

$$\hat{\phi}_{1k} = \frac{1}{|P|}[(P_3 - P_2)(\hat{y}_{1k} - P_1) + (P_1 - P_2)(\hat{y}_{2k} - P_2)] \tag{2.6}$$

$$\hat{\phi}_{2k} = \frac{1}{|P|}[(P_2 - P_1)(\hat{y}_{1k} - P_1) + (P_3 - P_1)(\hat{y}_{2k} - P_2)] \tag{2.7}$$

$$\hat{\phi}_{3k} = 1 - \hat{\phi}_{1k} - \hat{\phi}_{2k} \tag{2.8}$$

Sin embargo, puesto que no necesariamente se tiene la información de todos los individuos en el universo, si no los individuos en una muestra bajo un diseño $p(s)$ con probabilidades de inclusión positivas π_k y π_{kl} el $\phi - estimador$ que señala la proporción de personas en la población que pertenecen al grupo 1 es:

$$\hat{\bar{\phi}}_1 = \frac{\sum_s \hat{\phi}_{1k}/\pi_k}{N} \tag{2.9}$$

Como en 2.9, se tiene análogamente para el grupo 2:

$$\hat{\bar{\phi}}_2 = \frac{\sum_s \hat{\phi}_{2k}/\pi_k}{N} \tag{2.10}$$

y para el grupo 3:

$$\hat{\bar{\phi}}_3 = 1 - \hat{\bar{\phi}}_1 - \hat{\bar{\phi}}_2 \tag{2.11}$$

Ahora, debe mostrarse la propiedad de insesgamiento de dichos estimadores.

Demostración:

Asumiendo N conocido se tiene que,

$$\hat{\Phi}_k = P^{-1}\hat{\Lambda}_k \tag{2.12}$$

Luego la esperanza bajo el mecanismo aleatorio de $\hat{\Phi}_k$ está dada por,

$$\begin{aligned} E_{MA}[\hat{\Phi}_k] &= E_{MA}[P^{-1}\hat{\Lambda_k}] \\ &= P^{-1}E_{MA}[\hat{\Lambda}_k] \\ &= P^{-1}\Lambda_k \end{aligned}$$

Esto pues, $E(\hat{y}_{rk}) = \lambda_{rk}$ y por lo tanto,

$$E_{MA}[\hat{\Lambda}_k] = \begin{pmatrix} \lambda_{1k} - P_1 \\ \lambda_{2k} - P_2 \end{pmatrix} = \Lambda_k$$

Luego,

$$E_{MA}[\hat{\Phi}_k] = \Phi_k$$

Del anterior resultado se obtiene:

$$E_{MA}(\hat{\phi}_{1k}) = \phi_{1k}, \quad E_{MA}(\hat{\phi}_{2k}) = \phi_{2k} \quad y \quad E_{MA}(\hat{\phi}_{3k}) = \phi_{3k} \tag{2.13}$$

Ahora,

$$\begin{aligned} E(\hat{\bar{\phi}}_1) &= E_p\left[E_{MA}\left(\frac{\sum_s \hat{\phi}_{1k}/\pi_k}{N} \mid S\right)\right] \\ &= E_p\left[\frac{\sum_s E_{MA}(\hat{\phi}_{1k})/\pi_k}{N}\right] \\ &= E_p\left[\frac{\sum_s \phi_{1k}/\pi_k}{N}\right] \\ &= \frac{t_{\phi_1}}{N} \\ &= \bar{\phi}_1 \end{aligned}$$

lo que hace referencia a la proporción de personas en la población que pertenecen al grupo 1 y demuestra la propiedad de insesgamiento del estimador.

□

Análogamente se demuestra para $\hat{\bar{\phi}}_2$ y $\hat{\bar{\phi}}_3$.

En cuanto a la varianza del estimador sobre el mecanismo aleatorio:

$$\begin{aligned} V_{MA}(\hat{\Phi}_k) &= V_{MA}(P^{-1}\hat{\Lambda}_k) \\ &= P^{-1}V_{MA}(\hat{\Lambda}_k)(P^{-1})' \end{aligned}$$

donde,

$$V_{MA}(\hat{\Lambda}_k) = \begin{pmatrix} Var(\hat{y}_{1k}) & Cov(\hat{y}_{1k}, \hat{y}_{2k}) \\ Cov(\hat{y}_{1k}, \hat{y}_{2k}) & Var(\hat{y}_{2k}) \end{pmatrix} \tag{2.14}$$

es la matriz de varianzas y covarianzas de $\hat{\Lambda}_k$ bajo el mecanismo aleatorio.

Luego,

$$
\begin{aligned}
V_{MA}(\hat{\Lambda}_k) &= \frac{1}{|P|}\begin{pmatrix} P_3-P_2 & P_1-P_2 \\ P_2-P_1 & P_3-P_1 \end{pmatrix} * \begin{pmatrix} Var(\hat{y}_{1k}) & Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ Cov(\hat{y}_{1k},\hat{y}_{2k}) & Var(\hat{y}_{2k}) \end{pmatrix} \\
&* \frac{1}{|P|}\begin{pmatrix} P_3-P_2 & P_2-P_1 \\ P_1-P_2 & P_3-P_1 \end{pmatrix} \\
&= \frac{1}{|P|^2}\begin{pmatrix} (P_3-P_2)Var(\hat{y}_{1k}) & (P_1-P_2)Var(\hat{y}_{2k}) \\ +(P_1-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_3-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ (P_2-P_1)Var(\hat{y}_{1k}) & (P_3-P_1)Var(\hat{y}_{2k}) \\ +(P_3-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \end{pmatrix} \\
&* \begin{pmatrix} P_3-P_2 & P_2-P_1 \\ P_1-P_2 & P_3-P_1 \end{pmatrix} \\
&= \frac{1}{|P|^2}\begin{pmatrix} (P_3-P_2)(P_3-P_2)Var(\hat{y}_{1k}) & (P_3-P_2)(P_2-P_1)Var(\hat{y}_{1k}) \\ +(P_3-P_2)(P_1-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_3-P_1)(P_3-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ +(P_1-P_2)(P_1-P_2)Var(\hat{y}_{2k}) & +(P_1-P_2)(P_3-P_1)Var(\hat{y}_{2k}) \\ +(P_3-P_2)(P_1-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_1-P_2)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ (P_3-P_2)(P_2-P_1)Var(\hat{y}_{1k}) & (P_2-P_1)(P_2-P_1)Var(\hat{y}_{1k}) \\ +(P_3-P_1)(P_3-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_1-P_2)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ +(P_1-P_2)(P_3-P_1)Var(\hat{y}_{2k}) & +(P_3-P_1)(P_3-P_1)Var(\hat{y}_{2k}) \\ +(P_1-P_2)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_3-P_1)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \end{pmatrix} \\
&= \frac{1}{|P|^2}\begin{pmatrix} (P_3-P_2)^2Var(\hat{y}_{1k}) & (P_3-P_2)(P_2-P_1)Var(\hat{y}_{1k}) \\ +(P_1-P_2)^2Var(\hat{y}_{2k}) & +(P_3-P_1)(P_3-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ +2(P_3-P_2)(P_1-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_1-P_2)(P_3-P_1)Var(\hat{y}_{2k}) \\ & +(P_1-P_2)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ (P_3-P_2)(P_2-P_1)Var(\hat{y}_{1k}) & (P_2-P_1)^2Var(\hat{y}_{1k}) \\ +(P_3-P_1)(P_3-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) & +(P_3-P_1)^2Var(\hat{y}_{2k}) \\ +(P_1-P_2)(P_3-P_1)Var(\hat{y}_{2k}) & +(P_3-P_1)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) \\ +(P_1-P_2)(P_2-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) & \end{pmatrix}
\end{aligned}
$$

Con lo que se obtiene:

$$Var(\hat{\phi}_{1k}) = \frac{(P_3-P_2)^2Var(\hat{y}_{1k}) + 2(P_3-P_2)(P_1-P_2)Cov(\hat{y}_{1k},\hat{y}_{2k}) + (P_1-P_2)^2Var(\hat{y}_{2k})}{|P|^2} \tag{2.15}$$

$$Var(\hat{\phi}_{2k}) = \frac{(P_2-P_1)^2Var(\hat{y}_{1k}) + 2(P_2-P_1)(P_3-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) + (P_3-P_1)^2Var(\hat{y}_{2k})}{|P|^2} \tag{2.16}$$

$$
\begin{aligned}
Cov(\hat{\phi}_{1k},\hat{\phi}_{2k}) &= [(P_3-P_2)(P_2-P_1)Var(\hat{y}_{1k}) + (P_2-P_1)^2Cov(\hat{y}_{1k},\hat{y}_{2k}) \\
&+ (P_3-P_2)(P_3-P_1)Cov(\hat{y}_{1k},\hat{y}_{2k}) + (P_1-P_2)(P_3-P_1)Var(\hat{y}_{2k})]/|P|^2
\end{aligned} \tag{2.17}
$$

Siguiendo el resultado de 2.11 junto con las varianzas y covarianza que se hallaron en el paso anterior se puede llegar a la varianza de $\hat{\phi}_{3k}$.

$$\begin{aligned}
\hat{\hat{\phi}}_{3k} &= 1 - \hat{\hat{\phi}}_{1k} - \hat{\hat{\phi}}_{2k} \\
Var(\hat{\hat{\phi}}_{3k}) &= Var(1 - \hat{\hat{\phi}}_{1k} + \hat{\hat{\phi}}_{2k}) \\
&= 0 + Var(-\hat{\hat{\phi}}_{1k} - \hat{\hat{\phi}}_{2k}) \\
&= Var(\hat{\hat{\phi}}_{1k}) + Var(\hat{\hat{\phi}}_{2k}) + 2Cov(\hat{\phi}_{1k}, \hat{\phi}_{2k}) \\
&= \frac{1}{|P|^2}[(P_3 - P_2)^2 Var(\hat{y}_{1k}) - 2(P_3 - P_2)(P_1 - P_2)Cov(\hat{y}_{1k}, \hat{y}_{2k}) \\
&+ (P_1 - P_2)^2 Var(\hat{y}_{2k}) + (P_2 - P_1)^2 Var(\hat{y}_{1k}) \\
&+ 2(P_2 - P_1)(P_3 - P_1)Cov(\hat{y}_{1k}, \hat{y}_{2k}) + (P_3 - P_1)^2 Var(\hat{y}_{2k}) \\
&+ 2[(P_3 - P_2)(P_2 - P_1)Var(\hat{y}_{1k}) + (P_2 - P_1)^2 Cov(\hat{y}_{1k}, \hat{y}_{2k}) \\
&+ (P_3 - P_2)(P_3 - P_1)Cov(\hat{y}_{1k}, \hat{y}_{2k}) + (P_1 - P_2)(P_3 - P_1)Var(\hat{y}_{2k})]]
\end{aligned}$$

Factorizando se obtienen los dos términos de $Var(\hat{y}_{1k})$ y $Var(\hat{y}_{2k})$, luego para la $Cov(\hat{y}_{1k}, \hat{y}_{2k})$ se desarrolla el producto y se factoriza:

$$\begin{aligned}
Var(\hat{\hat{\phi}}_{3k}) &= \frac{1}{|P|^2}[Var(\hat{y}_{1k})(P_3 - P_2 + P_2 - P_1)^2 + Var(\hat{y}_{2k})(P_1 - P_2 + P_3 - P_1)^2 \\
&+ 2Cov(\hat{y}_{1k}, \hat{y}_{2k})(P_3P_1 - P_2P_3 - P_2P_1 + P_2^2 + P_2P_3 - P_2P_1 - P_1P_3 + P_1^2 \\
&+ P_2^2 + 2P_2P_1 - P_1^2 + P_3^2 - P_3P_1 - P_2P_3 + P_2P_1)] \\
&= \frac{1}{|P|^2}[Var(\hat{y}_{1k})(P_3 - P_1)^2 + Var(\hat{y}_{2k})(P_3 - P_2)^2 \\
&+ 2Cov(\hat{y}_{1k}, \hat{y}_{2k})(-P_3P_1 - P_2P_3 + P_2P_1 + P_3^2) \\
&= \frac{1}{|P|^2}[Var(\hat{y}_{1k})(P_3 - P_1)^2 + Var(\hat{y}_{2k})(P_3 - P_2)^2 \\
&+ 2Cov(\hat{y}_{1k}, \hat{y}_{2k})(P_3(P_3 - P_1) - P_2(P_3 - P_1))
\end{aligned}$$

Obteniendo,

$$Var(\hat{\phi}_{3k}) = \frac{(P_3 - P_1)^2 Var(\hat{y}_{1k}) + 2(P_3 - P_1)(P_3 - P_2)Cov(\hat{y}_{1k}, \hat{y}_{2k}) + (P_3 - P_2)^2 Var(\hat{y}_{2k})}{|P|^2} \tag{2.18}$$

□

Ahora, la varianza de cada $\phi - estimador$ estará dada por:

$$\begin{aligned} V(\hat{\bar{\phi}}_1) &= V_p\left[E_{MA}\left(\frac{\sum_s \hat{\phi}_{1k}/\pi_k}{N} \mid S\right)\right] + E_p\left[V_{MA}\left(\frac{\sum_s \hat{\phi}_{1k}/\pi_k}{N} \mid S\right)\right] \\ &= V_p\left[\frac{\sum_s \phi_{1k}/\pi_k}{N}\right] + E_p\left[\sum_s \frac{1}{N^2\pi_k^2} Var(\hat{\phi}_{1k}) \mid S\right] \end{aligned}$$

Esto es, ya que la respuesta de un índividuo es independiente de la respuesta de otro $(Cov(\hat{\phi}_{1k}, \hat{\phi}_{1l}) = 0)$.

Para facilidad en la notación sea,

$$Var(\hat{\phi}_{1k}) = \frac{Var(\hat{\phi}_{1k})^*}{|P|^2}$$

$$Var(\hat{\phi}_{2k}) = \frac{Var(\hat{\phi}_{2k})^*}{|P|^2}$$

$$Var(\hat{\phi}_{3k}) = \frac{Var(\hat{\phi}_{3k})^*}{|P|^2}$$

Ahora,

$$\begin{aligned} V(\hat{\bar{\phi}}_1) &= V_p\left[\frac{\sum_s \phi_{1k}/\pi_k}{N}\right] + E_p\left[\sum_s \frac{1}{N^2\pi_k^2}\left(\frac{Var(\hat{\phi}_{1k})^*}{|P|^2}\right) \mid S\right] \\ &= \frac{1}{N^2}\sum_U V_p\left(I_k\frac{\phi_{1k}}{\pi_k}\right) + \frac{1}{N^2|P|^2}\sum\sum_U I_k\frac{Var(\hat{\phi}_{1k})^*}{\pi_k^2} \\ &= \frac{1}{N^2}\sum_U \frac{\phi_{1k}^2}{\pi_k^2}\pi_k(1-\pi_k) + \frac{1}{N^2|P|^2}\sum_U \frac{Var(\hat{\phi}_{1k})^*}{\pi_k} \\ &= \frac{1}{N^2}\sum\sum_U \frac{\phi_{1k}^2}{\pi_k^2}\pi_k(1-\pi_k) + \sum\sum_{k\neq l}\frac{\phi_{1k}\phi_{1l}}{\pi_k\pi_l}(\pi_{kl}-\pi_k\pi_l) + \frac{1}{N^2|P|^2}\sum_U \frac{Var(\hat{\phi}_{1k})^*}{\pi_k} \end{aligned}$$

Luego,

$$V(\hat{\bar{\phi}}_1) = \frac{1}{N^2}\sum\sum_U \Delta_{kl}\breve{\phi}_{1k}\breve{\phi}_{1l} + \frac{1}{N^2|P|^2}\sum_U \frac{Var(\hat{\phi}_{1k})^*}{\pi_k} \tag{2.19}$$

Análogamente para $V(\hat{\bar{\phi}}_2)$ y $V(\hat{\bar{\phi}}_3)$.

Finalmente, un estimador de la varianza para el $\phi - estimador$ $\hat{\bar{\phi}}_1$ estará dado por,

$$\hat{V}(\hat{\bar{\phi}}_1) = \frac{1}{N^2}\sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_s \frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k\pi_k}$$

o bien,

$$\hat{V}(\hat{\bar{\phi}}_1) = \frac{1}{N^2}\sum\sum_s \breve{\Delta}_{kl}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_s \frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k^2} \tag{2.20}$$

Donde,

$$\hat{Var}(\hat{\phi}_{1k})^* = (P_3-P_2)^2\hat{Var}(\hat{y}_{1k}) + 2(P_3-P_2)(P_1-P_2)\hat{Cov}(\hat{y}_{1k},\hat{y}_{2k}) + (P_1-P_2)^2\hat{Var}(\hat{y}_{2k})$$

con,

$$\begin{aligned}
\hat{Var}(\hat{y}_{1k}) &= \sum_s \frac{(\hat{y}_{1k}-E(\hat{y}_{1k}))^2}{n-1} \\
\hat{Var}(\hat{y}_{2k}) &= \sum_s \frac{(\hat{y}_{2k}-E(\hat{y}_{2k}))^2}{n-1} \\
\hat{Cov}(\hat{y}_{1k},\hat{y}_{2k}) &= \sum_s \frac{(\hat{y}_{1k}-E(\hat{y}_{1k}))(\hat{y}_{2k}-E(\hat{y}_{2k}))}{n-1}
\end{aligned}$$

Análogamente también se tienen las estimaciones para las varianzas de $\hat{\bar{\phi}}_2$ y $\hat{\bar{\phi}}_3$,

$$\hat{V}(\hat{\bar{\phi}}_2) = \frac{1}{N^2}\sum\sum_s \breve{\Delta}_{kl}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_s \frac{\hat{Var}(\hat{\phi}_{2k})^*}{\pi_k^2} \tag{2.21}$$

$$\hat{V}(\hat{\bar{\phi}}_3) = \frac{1}{N^2}\sum\sum_s \breve{\Delta}_{kl}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_s \frac{\hat{Var}(\hat{\phi}_{3k})^*}{\pi_k^2} \tag{2.22}$$

Ahora, debe mostrarse la propiedad de insesgamiento de los anteriores estimadores de las varianzas.

Demostración:

$$\begin{aligned}
E(\hat{V}(\hat{\breve{\phi}}_1)) &= E_{MA}\left[E_p\left(\frac{1}{N^2}\sum\sum_s \breve{\Delta}_{kl}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_s \frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k^2}\right)\right] \\
&= E_{MA}\left[E_p\left(\frac{1}{N^2}\sum\sum_U I_k I_l \frac{\Delta_{kl}}{\pi_{kl}}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_U I_k \frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k^2}\right)\right] \\
&= E_{MA}\left[\frac{1}{N^2}\sum\sum_U E_p(I_k I_l)\frac{\Delta_{kl}}{\pi_{kl}}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_U E_p(I_k)\frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k^2}\right] \\
&= E_{MA}\left[\frac{1}{N^2}\sum\sum_U \Delta_{kl}\breve{\hat{\phi}}_{1k}\breve{\hat{\phi}}_{1l} + \frac{1}{N^2|P|^2}\sum_U \frac{\hat{Var}(\hat{\phi}_{1k})^*}{\pi_k}\right]
\end{aligned}$$

De 2.13 se sabe que $E(\hat{\phi}_{1k}) = \phi_{1k}$, luego,

$$E(\hat{V}(\hat{\breve{\phi}}_1)) = \frac{1}{N^2}\sum\sum_U \Delta_{kl}\breve{\phi}_{1k}\breve{\phi}_{1l} + \frac{1}{N^2|P|^2}\sum_U \frac{E_{MA}(\hat{Var}(\hat{\phi}_{1k})^*)}{\pi_k}$$

Por inferencia estadística se sabe que $E(s^2) = \sigma^2$ por lo tanto $E_{MA}(\hat{Var}(\hat{\phi}_{1k})^*) = Var(\hat{\phi}_{1k})^*$ llegando así a la insesgadez de la varianza:

$$E(\hat{V}(\hat{\breve{\phi}}_1)) = \frac{1}{N^2}\sum\sum_U \Delta_{kl}\breve{\phi}_{1k}\breve{\phi}_{1l} + \frac{1}{N^2|P|^2}\sum_U \frac{Var(\hat{\phi}_{1k})^*}{\pi_k}$$

Análogamente para $V(\hat{\breve{\phi}}_2)$ y $V(\hat{\breve{\phi}}_3)$.

□

2.1.1. Estimador y varianza bajo un diseño EST-MAS

Un diseño muestral estratificado consiste en dividir una población de tamaño N en subpoblaciones o estratos que sean homogéneos dentro de cada uno teniendo en cuenta alguna variable auxiliar, previamente definida según el estudio, y heterogéneos entre los mismos. Cada uno de estos estratos tendrá tamaños poblacionales $N_1, N_2, ..., N_n$ de tal forma que,

$$N = N_1 + N_2 + ... + N_n$$

Luego, en cada estrato se tomarán muestras de tamaños $n_1, n_2, ..., n_n$ por medio de muestreo aleatorio simple cumpliendo

$$n = n_1 + n_2 + ... + n_n$$

Esta selección por estrato puede hacerse de diferentes formas:

- Seleccionando el mismo número de individuos en cada estrato (Afijación igual)
- Proporcional al tamaño poblacional de cada estrato (Afijación porporcional)
- Proporcional a una variable auxiliar.
- Teniendo en cuenta el costo, el margen de error y el nivel de confianza (Afijación óptima)

De la teoría se sabe que las probabilidades de inclusión de un diseño EST-MAS están dadas por,

$$\begin{aligned} \pi_k &= \frac{n_h}{N_h} \\ \pi_{kl} &= \frac{n_h(n_h-1)}{N_h(N_h-1)} \end{aligned}$$

También se sabe que un estimador insesgado de la media poblacional y su varianza están dados por las expresiones 2.23 y 2.24 respectivamente.

$$\hat{t}_\pi = \sum_{h=1}^{H} W_h \bar{Y}_{sh} \tag{2.23}$$

$$Var(\hat{t}_\pi) = \sum_{h=1}^{H} W_h^2 S^2_{Y_{sh}} \left(\frac{1-f_h}{n_h}\right) \tag{2.24}$$

con,

$$\begin{aligned} n_h &= \textit{Tamaño de muestra en el estrato } h \\ N_h &= \textit{Tamaño de la población del estrato } h \\ f_h &= \frac{n_h}{N_H} \quad \textit{fracción de muestreo} \\ W_h &= \frac{N_h}{N} \quad \textit{ponderación del estrato } h \\ \bar{Y}_{sh} &= \frac{Y_h}{N_h} \\ S_{Y_{sh}} &= \frac{1}{n_h-1}\sum\nolimits_{sh}(y_k-\bar{Y}_{sh})^2 \quad \textit{varianza de la muestra en el estrato } h \end{aligned}$$

Ahora, para llegar al $\phi - \textit{estimador}$ que señala la proporción de individuos en la población que pertenecen al grupo 1 se siguen las expresiones 2.9 a 2.11, reemplazando las probabilidades de inclusión anteriormente descritas, obteniendo

$$\begin{aligned}\hat{\bar{\phi}}_1 &= \frac{\sum_{h=1}^{H}\sum_{sh}\frac{\hat{\phi}_{1k}}{n_h/N_h}}{N} \\ &= \frac{\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{sh}\hat{\phi}_{1k}}{N} \\ &= \frac{\sum_{h=1}^{H}N_h\hat{\bar{\phi}}_{1k}}{N}\end{aligned}$$

Luego,

$$\hat{\bar{\phi}}_1 = \sum_{h=1}^{H} W_h \hat{\bar{\phi}}_{1k} \tag{2.25}$$

Análogamente se tienen los estimadores para el grupo 2 y 3, dados por,

$$\hat{\bar{\phi}}_2 = \sum_{h=1}^{H} W_h \hat{\bar{\phi}}_{2k} \tag{2.26}$$

$$\hat{\bar{\phi}}_3 = \sum_{h=1}^{H} W_h \hat{\bar{\phi}}_{3k} \tag{2.27}$$

Finalmente, un estimador insesgado de la varianza del $\phi - estimador$, utilizando las expresiones 2.20 a 2.22 y reemplazando tanto las probabilidades de inclusión como el estimador insesgado de la varianza para el diseño EST-MAS se obtiene,

$$\begin{aligned}\hat{V}(\hat{\bar{\phi}}_1) &= \sum_{h=1}^{H} W_h^2 S_{Y_{sh}}^2 \left(\frac{1-f_h}{n_h}\right) + \frac{1}{N^2|P|^2}\sum_{sh}\frac{\hat{Var}(\hat{\phi}_{1k})^*}{\left(\frac{n_h}{N_h}\right)^2} \\ &= \sum_{h=1}^{H} W_h^2 S_{Y_{sh}}^2 \left(\frac{1-f_h}{n_h}\right) + \sum_{sh}\frac{N_h^2\hat{Var}(\hat{\phi}_{1k})^*}{N^2 n_h^2|P|^2}\end{aligned}$$

$$\hat{V}(\hat{\bar{\phi}}_1) = \sum_{h=1}^{H} W_h^2 \left(S_{Y_{sh}}^2 \left(\frac{1-f_h}{n_h}\right) + \sum_{sh}\frac{\hat{Var}(\hat{\phi}_{1k})^*}{n_h^2|P|^2}\right) \tag{2.28}$$

Análogamente se tienen los estimadores de las varianzas para $\hat{\bar{\phi}}_2$ y $\hat{\bar{\phi}}_3$,

$$\hat{V}(\hat{\bar{\phi}}_2) = \sum_{h=1}^{H} W_h^2 \left(S_{Y_{sh}}^2 \left(\frac{1-f_h}{n_h}\right) + \sum_{sh}\frac{\hat{Var}(\hat{\phi}_{2k})^*}{n_h^2|P|^2}\right) \tag{2.29}$$

$$\hat{V}(\hat{\bar{\phi}}_3) = \sum_{h=1}^{H} W_h^2 \left(S^2_{Y_{sh}} \left(\frac{1-f_h}{n_h} \right) + \sum_{sh} \frac{\hat{Var}(\hat{\phi}_{3k})^*}{n_h^2 |P|^2} \right) \tag{2.30}$$

2.2. Método propuesto para el caso $G = g$

Luego de ver el caso de 3 categorías se expone la generalización para un estudio en el que se quiera estimar una característica sensible con preguntas que involucren un número finito arbitrario g de items. Para dicha generalización se sabe que g_k es el grupo verdadero del $k - ésimo$ encuestado donde $g_k = 1, ..., g$ con $k \in U$; y a_k es el valor de aumento seleccionado aleatoriamente con $a_k = 1, ..., g$. Luego, la respuesta codificada del encuestado cuyo grupo verdadero es g_k con un valor de aumento a_k dado, tiene la siguiente expresión:

$$z_k = g_k + a_k \tag{2.31}$$

Ahora se transforma z_k a un valor reportado R_k, donde

$$R_k = \begin{cases} z_k & \text{si } z_k \leq g, \\ z_k - g & \text{si } z_k > g. \end{cases} \tag{2.32}$$

La tabla 2.2 presenta los valores posibles reportados y su origen $(g_k + a_k)$.

Tabla 2.2. Transformaciones de la respuesta para el modelo aditivo generalizado a un número finito de categorías

Número reportado R_k	Origen $(g_k + a_k)$				
1	g+1	(g-1)+2	...	2+(g-1)	1+g
2	g+2	(g-1)+3	...	2+g	1+1
⋮	⋮	⋮	⋮	⋮	⋮
(g-1)	g+(g-1)	(g-1)+g	...	2+(g-3)	1+(g-2)
g	g+g	(g-1)+1	...	2+(g-2)	1+(g-1)

Como se ve para $G = 3$ en este caso también es necesario definir para un individuo en particular su probabilidad λ_{rk},

$$\begin{array}{ccccccccccc}
\lambda_{1k} & = & \phi_{gk}P_1 & + & \phi_{(g-1)k}P_2 & + & \dots & + & \phi_{2k}P_{(g-1)} & + & \phi_{1k}P_g \\
\lambda_{2k} & = & \phi_{gk}P_2 & + & \phi_{(g-1)k}P_3 & + & \dots & + & \phi_{2k}P_g & + & \phi_{1k}P_1 \\
\vdots & = & \vdots & + & \vdots & + & \vdots & + & \vdots & + & \vdots \\
\lambda_{(g-1)k} & = & \phi_{gk}P_{(g-1)} & + & \phi_{(g-1)k}P_g & + & \dots & + & \phi_{2k}P_{(g-3)} & + & \phi_{1k}P_{(g-2)} \\
\lambda_{gk} & = & \phi_{gk}P_g & + & \phi_{(g-1)k}P_1 & + & \dots & + & \phi_{2k}P_{(g-2)} & + & \phi_{1k}P_{(g-1)}
\end{array}$$

Luego para estimar ϕ_{gk} es necesario despejarlo del anterior sistema por lo que se tiene,

$$\begin{array}{ccccccccc}
\lambda_{1k} & = & P_1 & + & \phi_{(g-1)k}(P_2 - P_1) & + & \dots & + & \phi_{1k}(P_g - P_1) \\
\lambda_{2k} & = & P_2 & + & \phi_{(g-1)k}(P_3 - P_2) & + & \dots & + & \phi_{1k}(P_1 - P_2) \\
\vdots & = & \vdots & + & \vdots & + & \vdots & + & \vdots \\
\lambda_{(g-1)k} & = & P_{(g-1)} & + & \phi_{(g-1)k}(P_g - P_{g-1}) & + & \dots & + & \phi_{1k}(P_{(g-2)} - P_{(g-1)})
\end{array}$$

pasando a restar P y colocando el sistema en forma matricial se tiene,

$$\begin{pmatrix} \lambda_{1k} - P_1 \\ \lambda_{2k} - P_2 \\ \vdots \\ \lambda_{(g-1)k} - P_{(g-1)} \end{pmatrix} = \begin{pmatrix} P_g - P_1 & P_{(g-1)} - P_1 & \dots & P_2 - P_1 \\ P_1 - P_2 & P_g - P_2 & \dots & P_3 - P_2 \\ \vdots & \vdots & \vdots & \vdots \\ P_{(g-2)} - P_{(g-1)} & P_{(g-3)} - P_{(g-1)} & \dots & P_g - P_{(g-1)} \end{pmatrix} \begin{pmatrix} \phi_{1k} \\ \phi_{2k} \\ \vdots \\ \phi_{(g-1)k} \end{pmatrix}$$

equivalente a

$$\Lambda_k = P\Phi$$

Debe tenerse en cuenta que cuando $P_1 = P_2 = \ldots = P_g = 1/g$, $|P| = 0$, luego se asume que $|P| \neq 0$. Ahora para estimar Φ que corresponde a la matriz de probabilidades de que un individuo en la población pertenezca a cierto grupo y como solo se observa $\hat{\Lambda}$ y los elementos de la matriz P son fijos se tiene:

$$\hat{\Lambda}_k = \begin{pmatrix} \hat{y}_{1k} - P_1 \\ \hat{y}_{2k} - P_2 \\ \vdots \\ \hat{y}_{(g-1)k} - P_(g-1) \end{pmatrix} \tag{2.33}$$

donde, al igual que en $G = 3$,

$$\hat{y}_{rk} = \begin{cases} 1 & \text{si } R_k = r \\ 0 & e.o.c. \end{cases}$$

por lo tanto $\hat{y}_{rk} \sim Be(\lambda_{rk})$ con

$$\begin{aligned} E(\hat{y}_{rk}) &= \lambda_{rk} \\ V(\hat{y}_{rk}) &= \lambda_{rk}(1 - \lambda_{rk}) \end{aligned}$$

Luego para llegar al estimador deseado debe resolverse el siguiente sistema matricial,

$$\hat{\Phi}_k = P^{-1}\hat{\Lambda}_k \tag{2.34}$$

que corresponde a,

$$\begin{pmatrix} \hat{\phi}_{1k} \\ \hat{\phi}_{2k} \\ \vdots \\ \hat{\phi}_{(g-1)k} \end{pmatrix} = \begin{pmatrix} P_g - P_1 & P_{(g-1)} - P_1 & \dots & P_2 - P_1 \\ P_1 - P_2 & P_g - P_2 & \dots & P_3 - P_2 \\ \vdots & \vdots & \vdots & \vdots \\ P_{(g-2)} - P_{(g-1)} & P_{(g-3)} - P_{(g-1)} & \dots & P_g - P_{(g-1)} \end{pmatrix}^{-1} \begin{pmatrix} \lambda_{1k} - P_1 \\ \lambda_{2k} - P_2 \\ \vdots \\ \lambda_{(g-1)k} - P_{(g-1)} \end{pmatrix}$$

donde la estimación para $\hat{\phi}_{gk}$ estará dada por,

$$\hat{\phi}_{gk} = 1 - \hat{\phi}_{1k} - \hat{\phi}_{2k} - ... - \hat{\phi}_{(g-1)k} \tag{2.35}$$

Al igual que en $G = 3$ se sabe que se tiene una muestra bajo un diseño $p(s)$ con probabilidades de inclusión positivas π_k y π_{kl}, luego el $\phi - estimador$ que señala la proporción de personas en la población que pertenecen a cada grupo estará dado por,

$$\hat{\bar{\phi}}_1 = \frac{\sum_s \hat{\phi}_{1k}/\pi_k}{N} \tag{2.36}$$

$$\hat{\bar{\phi}}_2 = \frac{\sum_s \hat{\phi}_{2k}/\pi_k}{N} \tag{2.37}$$

Análogamente para $(g-1)$

$$\hat{\bar{\phi}}_{(g-1)} = \frac{\sum_s \hat{\phi}_{(g-1)k}/\pi_k}{N} \tag{2.38}$$

y para el grupo g se tiene

$$\hat{\bar{\phi}}_g = 1 - \hat{\bar{\phi}}_1 - \hat{\bar{\phi}}_2 - ... - \hat{\bar{\phi}}_{g-1} \tag{2.39}$$

En cuanto a la varianza sobre el mecanismo aleatorio se tiene que

$$\begin{aligned} V_{MA}(\hat{\Phi}_k) &= V_{MA}(P^{-1}\hat{\Lambda}_k) \\ &= P^{-1}V_{MA}(\hat{\Lambda}_k)(P^{-1})' \end{aligned}$$

donde,

$$V_{MA}(\hat{\Lambda}_k) = \begin{pmatrix} Var(\hat{y}_{1k}) & & Cov(\hat{y}_{ik}, \hat{y}_{jk}) \\ & \ddots & \\ Cov(\hat{y}_{ik}, \hat{y}_{jk}) & & Var(\hat{y}_{(g-1)k}) \end{pmatrix} \tag{2.40}$$

con $i \neq j = 1, ..., g-1$, es la matriz de varianzas y covarianzas de $\hat{\Lambda}_k$ bajo el mecanismo aleatorio.

CAPÍTULO 3

Simulación

"La simulación es el proceso de diseñar un modelo de un sistema real y llevar a término experiencias con él, con la finalidad de comprender el comportamiento del sistema o evaluar nuevas estrategias dentro de los límites impuestos por un cierto criterio o un conjunto de ellos para el funcionamiento del sistema." (R.E. Shanon, 1988).

El sistema real del presente estudio es la estimación de la proporción de individuos en la población que tiene cierta característica sensible. Luego, el objetivo es diseñar un modelo que permita evaluar el comportamiento de los estimadores que se hallaron previamente. Esta simulación se hizo por medio del software estadístico R utilizando los paquetes *combinat* y *gtools* que permitieron llegar a varias conclusiones sobre los estimadores.

Como primer paso y como ejemplo se define un tamaño de población de 6 individuos y luego se enumeran todas las posibles muestras, en este caso de tamaño 3, seleccionadas por medio de un muestreo aleatorio simple sin reemplazo, resultando asi un total de 20 muestras ($N = 6$, $n = 3$, $C(6, 3) = 20$) (tabla 3.1). A estas 20 muestras se le adicionaron $3^3 = 27$ posibles codificaciones de las tres respuestas en la muestra y este procedimiento se replicó un numero h de iteraciones donde el máximo h considerado fue de 1.000 iteraciones para un total de 20*27*1.000=540.000 réplicas en la simulación.

En este mismo paso también se definen las probabilidades de adicionar 1, 2 ó 3 a la respuesta real de cada individuo, recordando que estas probabilidades son elegidas por el investigador, que no deben ser iguales y que en total su suma debe ser igual a 1. Para este ejemplo las probabilidades seleccionadas son:

$$p(1) = 0.5 \qquad p(2) = 0.25 \qquad p(3) = 0.25 \tag{3.1}$$

Donde $p(1)$ es la probabilidad de sumar 1 a la respuesta real de cada individuo, $p(2)$ la probabilidad de sumar 2 y $p(3)$ la de sumar 3.

TABLA 3.1. Posibles muestras con MAS sin reemplazo de tamaño 3 para una población de 6 individuos

Muestra	Ind 1	Ind 2	Ind 3
M1	1	2	3
M2	1	2	4
M3	1	2	5
M4	1	2	6
M5	1	3	4
M6	1	3	5
M7	1	3	6
M8	1	4	5
M9	1	4	6
M10	1	5	6
M11	2	3	4
M12	2	3	5
M13	2	3	6
M14	2	4	5
M15	2	4	6
M16	2	5	6
M17	3	4	5
M18	3	4	6
M19	3	5	6
M20	4	5	6

Para poder comparar el estimador y verificar su insesgamiento se definen previamente las frecuencias de cada uno de los 3 grupos dentro de la población para generar la proporción de individuos que poseen cierta característica. Como ejemplo se puede ver la tabla 3.2.

TABLA 3.2. Respuesta verdadera dada por cada individuo

Individuo	Respuesta verdadera
1	1
2	2
3	3
4	3
5	3
6	3

En la tabla 3.2 se observa lo que contesta realmente cada individuo sobre su característica sensible, luego, sus frecuencias se conforman de la siguiente manera: $grupo\ 1 = 1$, $grupo\ 2 = 1$ y $grupo\ 3 = 4$ lo que se traduce en que los parámetros que se quieren estimar son los siguientes,

$$\phi_1 = \frac{1}{6} = 0.1666$$
$$\phi_2 = \frac{1}{6} = 0.1666$$
$$\phi_3 = \frac{4}{6} = 0.6666$$

El siguiente paso consiste en generar todos los posibles números aleatorios entre 1 y 3 que se le suman a cada respuesta verdadera, lo que genera $3^3 = 27$ posibles combinaciones para los 3 individuos de cada una de las 20 muestras en la tabla 3.1 para así llegar a 540 combinaciones posibles (20 x 27) para las cuales se genera la tabla 3.3. Para confirmar que el estimador es insesgado este procedimiento es replicado un número h de veces.

Tabla 3.3. Posibles combinaciones de respuestas para las 20 muestras

Muestra	Respuesta real	Números aleatorios	No. de combinación
{1,2,3}	{1,2,3}	{1,1,1}	1
		{1,1,2}	2
		⋮	⋮
		{3,3,2}	26
		{3,3,3}	27
⋮	⋮	⋮	⋮
{4,5,6}	{3,3,3}	{1,1,1}	514
		{1,1,2}	515
		⋮	⋮
		{3,3,2}	539
		{3,3,3}	540

Tal como se vio en el capítulo 2, el siguiente paso es transformar la respuesta de cada encuestado para proteger su confidencialidad conforme a la expresión 2.2. Antes de calcular los estimadores y sus varianzas es necesario codificar la respuesta transformada a través de variables indicadoras como se muestra en 3.2.

$$\{\underbrace{\overset{ind\ 1}{y_{11}}\ \overset{ind\ 2}{y_{12}}\ \overset{ind\ 3}{y_{13}}}_{respuesta\ 1} - \underbrace{\overset{ind\ 1}{y_{21}}\ \overset{ind\ 2}{y_{22}}\ \overset{ind\ 3}{y_{23}}}_{respuesta\ 2} - \underbrace{\overset{ind\ 1}{y_{31}}\ \overset{ind\ 2}{y_{32}}\ \overset{ind\ 3}{y_{33}}}_{respuesta\ 3}\} \tag{3.2}$$

Donde,

$$y_{gk} = \begin{cases} 1 & \text{si } el\ individuo\ k\ contesta\ codificadamente\ g, \\ 0 & \text{si } e.o.c. \end{cases} \tag{3.3}$$

En dicha codificación las tres primeras componentes conforman las variables indicadoras que los tres individuos en la muestra hayan codificado su respuesta como 1. Es decir que si el primer individuo tiene como respuesta transformada 1, entonces la primer componente de este primer grupo será 1; análogamente para el segundo grupo, se colocará 1 cuando

la respuesta transformada del individuo k sea 2; y por último sucederá lo mismo con las componentes restantes del tercer grupo cuando la respuesta sea 3. (ver ejemplo 3.4)

$$\{\overbrace{\underbrace{1}^{ind\ 1}\ \underbrace{1}^{ind\ 2}\ \underbrace{0}^{ind\ 3}}^{respuesta\ 1} - \underbrace{000}_{respuesta\ 2} - \underbrace{001}_{respuesta\ 3}\} \tag{3.4}$$

Esta codificación corresponde a la muestra en la que el primer individuo tuvo como respuesta transformada 1, el segundo individuo 1 y el tercero 3.

El Diagrama 3.5 clarifica la forma en la que las respuestas de cada encuestado se transforman y codifican con fines de protección de la confidencialidad y el manejo de cada respuesta en la simulación.

$$\begin{array}{cc} Muestra & \{5,1,2\} \\ \Downarrow & \Downarrow \\ Resp.\ verdadera & \{3,1,2\} \\ \Downarrow & \Downarrow \\ No.\ aleatorio & \{1,3,1\} \\ \Downarrow & \Downarrow \\ Suma\ No.\ aleatorio & \{4,4,3\} \\ \Downarrow & \Downarrow \\ Transformacion & \{1,1,3\} \\ \Downarrow & \Downarrow \\ Codificacion & \{110-000-001\} \end{array} \tag{3.5}$$

Luego de conseguir la transformación y codificación de las respuestas para todos los encuestados se procede a calcular los estimadores de cada atributo sensible por individuo mediante los resultados 2.9 - 2.11 y de esta forma obtener un vector de 9 componentes, por individuo, como se muestra en 3.6,

$$\{\hat{\phi}_{11}\hat{\phi}_{12}\hat{\phi}_{13} - \hat{\phi}_{21}\hat{\phi}_{22}\hat{\phi}_{23} - \hat{\phi}_{31}\hat{\phi}_{32}\hat{\phi}_{33}\} \tag{3.6}$$

Después de obtener los estimadores por individuo se hace un promedio por cada grupo de individuos en cada una de las respuestas, lo que genera un vector de tres componentes con las estimaciones de la proporción de individuos que poseen cierto atributo sensible para cada muestra (ver 3.7).

$$\{\underbrace{\hat{\phi}_{11}\hat{\phi}_{12}\hat{\phi}_{13}}_{\hat{\bar{\phi}}_1} - \underbrace{\hat{\phi}_{21}\hat{\phi}_{22}\hat{\phi}_{23}}_{\hat{\bar{\phi}}_2} - \underbrace{\hat{\phi}_{31}\hat{\phi}_{32}\hat{\phi}_{33}}_{\hat{\bar{\phi}}_3}\} \tag{3.7}$$

Finalmente, al obtener el vector de proporciones estimadas por muestra (3.7), se conforma una matriz con todas las muestras en la que cada fila tiene un vector, luego dicha matriz se multiplica por la probabilidad definida por el investigador (3.1) de que cada individuo dentro de cada muestra seleccione un número aleatorio entre 1 y 3, dividido entre el total de muestras sin reemplazo de tamaño $n = 3$ para una población de $N = 6$ individuos (20). Con este cálculo se obtienen los resultados de la tabla 3.4, que comparados con las proporciones definidas al inicio de la simulación demuestran la

insesgadez del estimador.

TABLA 3.4. Estimadores de proporción de atributos sensibles con N=6 y n=3

$\hat{\hat{\phi}}_1$	$\hat{\hat{\phi}}_2$	$\hat{\hat{\phi}}_3$
0.1666	0.1666	0.6666

Habiendo obtenido los estimadores se procede a calcular la varianza de cada uno de ellos por medio de las expresiones 2.20 a 2.22. Este procedimiento se repite por muestra y al final se calcula el promedio de la varianza de todas las muestras para verificar su insesgadez. Para este caso se obtienen las varianzas contenidas en la tabla 3.5.

TABLA 3.5. Varianzas de estimadores de proporción de atributos sensibles con N=6 y n=3

$Var(\hat{\hat{\phi}}_1)$	$Var(\hat{\hat{\phi}}_2)$	$Var(\hat{\hat{\phi}}_3)$
0.925	0.925	1.092

Las siguientes tablas muestran algunos escenarios combinando proporciones reales, probabilidades de sumar un número aleatorio entre 1 y 3, diferentes configuraciones de tamaño de población y muestra, indicando sus respectivas estimaciones de la proporción de individuos con cierta característica sensible y sus varianzas. Dichas tablas fueron generadas con un número h de iteraciones igual a 1000.

TABLA 3.6. Estimadores de proporción de atributos sensibles y sus varianzas con N=6, n=3, p_1=0.5, p_2=0.25 y p_3=0.25

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	1/6=0.166	2/6=0.333	4/6=0.666
ϕ_2	1/6=0.166	3/6=0.5	1/6=0.166
ϕ_3	4/6=0.666	1/6=0.166	1/6=0.166
$\hat{\hat{\phi}}_1$	0.166	0.333	0.666
$\hat{\hat{\phi}}_2$	0.166	0.5	0.166
$\hat{\hat{\phi}}_3$	0.666	0.166	0.166
$Var(\hat{\hat{\phi}}_1)$	0.925	1	1.092
$Var(\hat{\hat{\phi}}_2)$	0.925	1.055	0.925
$Var(\hat{\hat{\phi}}_3)$	1.092	0.925	0.925

La estructura de la tabla 3.6, que es la misma para las demás tablas, consiste de una columna con el nombre "Dato" que hace referencia a las proporciones reales de las características sensibles o parámetros a estimar (ϕ_1, ϕ_2 y ϕ_3), a sus estimadores ($\hat{\hat{\phi}}_1$, $\hat{\hat{\phi}}_2$ y $\hat{\hat{\phi}}_3$) y a sus varianzas ($Var(\hat{\hat{\phi}}_1)$, $Var(\hat{\hat{\phi}}_2)$ y $Var(\hat{\hat{\phi}}_3)$). También cuenta con tres columnas adicionales que exhiben diferentes escenarios en cuanto a configuraciones distintas de proporciones reales de las características sensibles dentro de la población.

Como ejemplo se toma el escenario 1 de la tabla 3.6 en el que están definidas las proporciones reales de las características sensibles de la siguiente manera: $\phi_1 = 1/6 = 0.166$ para el grupo 1, $\phi_2 = 1/6 = 0.166$ para el grupo 2 y $\phi_3 = 1/6 = 0.666$ para el grupo 3; el objetivo de la simulación es obtener el mismo número o acercarse a él; para este caso,

luego de la simulación, se llega a la estimación de cada proporción, obteniendo: $\hat{\hat{\phi}}_1 = 0.166$ para el grupo 1, $\hat{\hat{\phi}}_2 = 0.166$ para el grupo 2 y $\hat{\hat{\phi}}_3 = 0.666$ para el grupo 3, lo que sugiere el insesgamiento de cada uno de los tres estimadores.

Finalmente, en las últimas filas se colocan las varianzas de cada estimador, $Var(\hat{\hat{\phi}}_1) = 0.925$ para el estimador del grupo 1, $Var(\hat{\hat{\phi}}_2) = 0.925$ para el estimador del grupo 2 y $Var(\hat{\hat{\phi}}_3) = 1.092$ para el del grupo 3.

Análogamente se puede hacer la lectura de las demás tablas, donde la diferencia entre las tablas 3.6 a 3.10 es el tamaño de la población y el de la muestra, luego, de las tablas 1 a 5 se presentan probabilidades de sumar un número aleatorio entre 1 y 3 diferentes a las de los escenarios de las tablas 3.6 a 3.10.

TABLA 3.7. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=3, p_1=0.5, p_2=0.25 y p_3=0.25

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.512	0.1	0.391
$\hat{\hat{\phi}}_2$	0.208	0.216	0.433
$\hat{\hat{\phi}}_3$	0.279	0.683	0.175
$Var(\hat{\hat{\phi}}_1)$	1.296	1.102	1.275
$Var(\hat{\hat{\phi}}_2)$	1.162	1.172	1.278
$Var(\hat{\hat{\phi}}_3)$	1.206	1.350	1.153

TABLA 3.8. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=4, p_1=0.5, p_2=0.25 y p_3=0.25

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.505	0.1	0.396
$\hat{\hat{\phi}}_2$	0.207	0.201	0.489
$\hat{\hat{\phi}}_3$	0.287	0.698	0.114
$Var(\hat{\hat{\phi}}_1)$	0.802	0.680	0.779
$Var(\hat{\hat{\phi}}_2)$	0.719	0.719	0.794
$Var(\hat{\hat{\phi}}_3)$	0.746	0.837	0.686

TABLA 3.9. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=3, p_1=0.5, p_2=0.25 y p_3=0.25

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.762	0.352	0.304
$\hat{\hat{\phi}}_2$	0.091	0.412	0.546
$\hat{\hat{\phi}}_3$	0.146	0.234	0.148
$Var(\hat{\hat{\phi}}_1)$	1.552	1.416	1.388
$Var(\hat{\hat{\phi}}_2)$	1.270	1.447	1.499
$Var(\hat{\hat{\phi}}_3)$	1.253	1.347	1.294

TABLA 3.10. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=4, p_1=0.5, p_2=0.25 y p_3=0.25

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.753	0.347	0.3
$\hat{\hat{\phi}}_2$	0.099	0.405	0.551
$\hat{\hat{\phi}}_3$	0.146	0.246	0.148
$Var(\hat{\hat{\phi}}_1)$	1.036	0.941	0.923
$Var(\hat{\hat{\phi}}_2)$	0.837	0.960	1
$Var(\hat{\hat{\phi}}_3)$	0.860	0.903	0.861

Se puede ver que en todas las tablas, la aproximación al parámetro de la población es bastante cercana, comprobando asi, por medio de simulaciones, que el estimador propuesto es insesgado.

Por otro lado, si se comparan pares de tablas como la 3.7 y la 2 es de notar que la varianza es mas pequeña cambiando la configuración de la probabilidad de sumar el número aleatorio, en este caso para la segunda tabla. En esta comparación entre parejas de tablas del presente capítulo y las contenidas en el apéndice B se puede concluir que las segundas siempre generaron una menor varianza. Por este último punto, queda como trabajo posterior, encontrar una combinación de probabilidades que mejoren la varianza de los estimadores.

A continuación en las figuras 3.1, 3.2 y 3.3 se muestra el comportamiento de la estimación de la primera, segunda y tercera categoría, respectivamente, o dicho de otro modo, la estimación de la proporción de individuos que pertenecen al grupo 1, 2 y 3 junto con su varianza. Esta comparación, que se obtiene de las tablas anteriormente vistas, se hace por medio de diferentes tamaños de muestra y población obteniendo fracciones de muestreo distintas que van desde el 15 % hasta el 50 % permitiendo detectar algunas tendencias en las estimaciones.

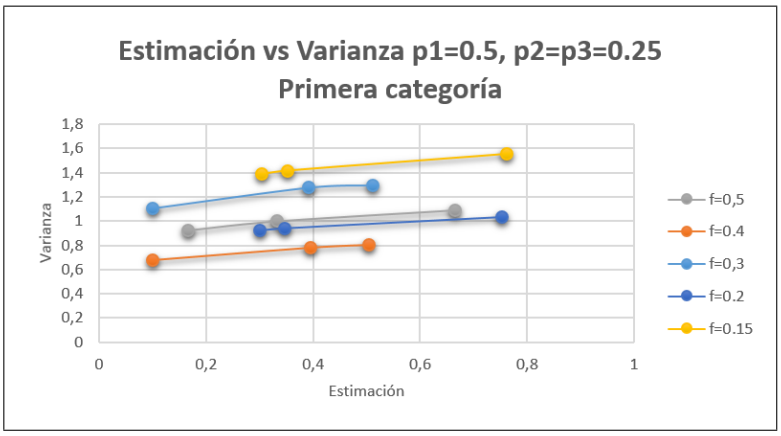

FIGURA 3.1. Comparación entre estimador y su varianza para la categoría 1 bajo diferentes fracciones de muestreo, mostrados en las tablas 3.6 a 3.10

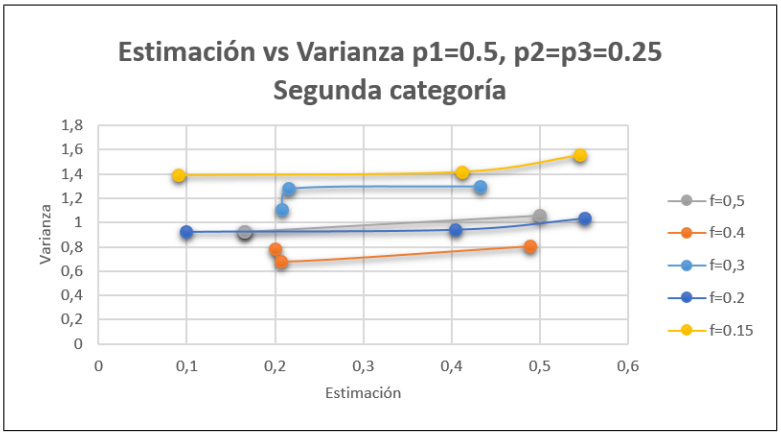

FIGURA 3.2. Comparación entre estimador y su varianza para la categoría 2 bajo diferentes fracciones de muestreo, mostrados en las tablas 3.6 a 3.10

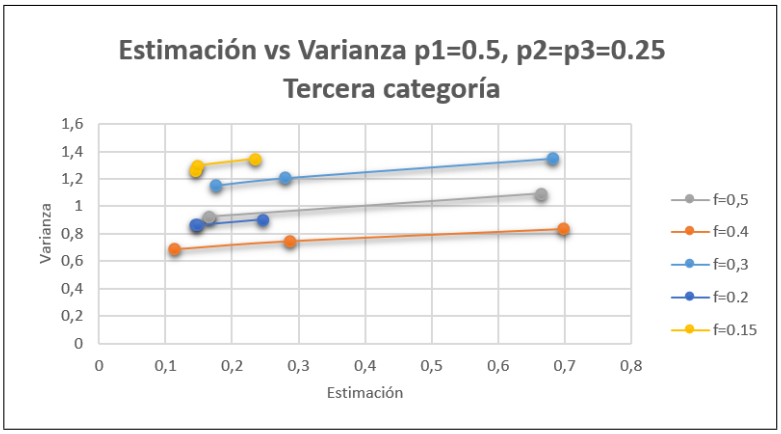

FIGURA 3.3. Comparación entre estimador y su varianza para la categoría 3 bajo diferentes fracciones de muestreo, mostrados en las tablas 3.6 a 3.10

La línea amarilla de las figuras 3.1 a 3.3 señala una fracción de muestreo de 0.15 obtenida de la tabla 3.9 donde se tienen 20 individuos en la población con una muestra de tamaño 3, la línea azul oscura se refiere a la tabla 3.10 con una población de 20 y muestra de 4, en color azul claro se muestra el comportamiento de la estimación de la tabla 3.7 con población de 10 y muestra de 3 individuos, la línea naranja presenta las estimaciones obtenidas y sus varianzas para una fracción de muestreo de 0.4 incluídas en la tabla 3.8 y finalmente se exponen los resultados hallados con la mayor fracción de muestreo (0.5) de la tabla 3.6.

Es de notar, de la figura 3.1 que a medida que la estimación de la proporción del atributo sensible se acerca mas 1 su varianza aumenta en todos los casos de manera gradual, también que a una mayor fracción de muestreo se presenta una varianza mucho menor pero que de igual forma se sigue manteniendo la tendencia de aumento entre estimación y varianza. Este comportamiento de aumento gradual se ve en las demás gráficas de estimación vs varianza para la segunda categoría (gráfica 3.2) y la tercera (gráfica 3.3); sin embargo para la categoría dos, este aumento se intensifica en la última estimación donde, para cada fracción de muestreo, la gráfica muestra un salto.

Para las tablas del apéndice B (1 a 15) se muestran los mismos escenarios de tamaño de población, muestra y porporción de atributos sensibles de las tablas mencionadas en el presente capítulo, cambiando únicamente, la probabilidad de sumar el número aleatorio, dichas tablas se presentan en 3 grupos:

- Grupo 1: Tablas de 1 a 5 con probabilidades de p_1=0.1, p_2=0.6 y p_3=0.3.
- Grupo 2: Tablas de 6 a 10 con probabilidades de p_1=0.6, p_2=0.3 y p_3=0.1.
- Grupo 3: Tablas de 11 a 15 con probabilidades de p_1=0.3, p_2=0.1 y p_3=0.6.

Lo que se concluye de estos grupos de tablas y las contenidas en este capítulo es que las primeras presentan una varianza mas pequeña y que la combinación de cada grupo de probabilidades (0.6, 0.1 y 0.3) entre cada grupo de tablas, no influye de manera significativa en las estimaciones ni en sus varianzas.

En una segunda parte del presente capítulo se muestra la relación existente entre el determinante de la matriz P (Ecuación 2.3) contra la varianza del estimador en cada categoría. Dicha relación es importante ya que la matriz P se constituye por las probabilidades de los valores de aumento determinados por el investigador. Para este ejercicio se utilizó el escenario en el que se tiene una población de tamaño $N = 20$ y una muestra de $n = 4$.

En las figuras 3.4, 3.5 y 3.6 se puede ver que a medida que el determinante de la matriz P se acerca a uno, la varianza del estimador de cada categoría es cada vez menor, y que por el contrario, cuando el determinante es mas cercano a cero la varianza es mayor. Esta condición muestra una relación inversamente proporcional entre el determinante de la matriz P y las varianzas de los estimadores, lo cual en la práctica es bastante útil si se quiere tener una varianza cada vez menor.

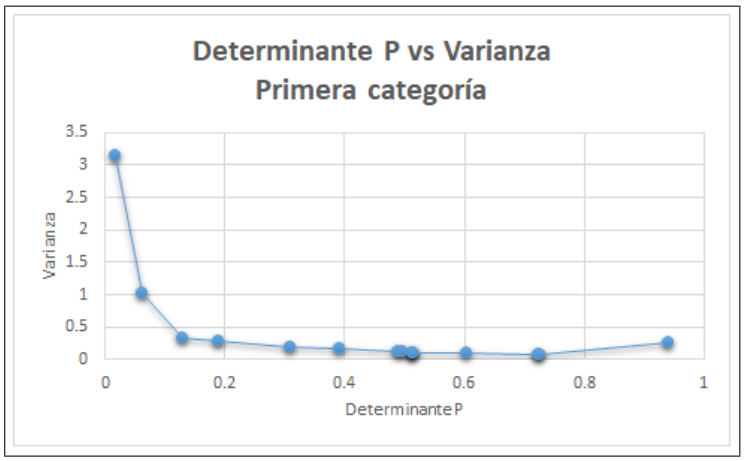

FIGURA 3.4. Comparación entre el determinante de la matriz P y la varianza de la categoría 1 teniendo en cuenta una población N = 20 y una muestra n = 4

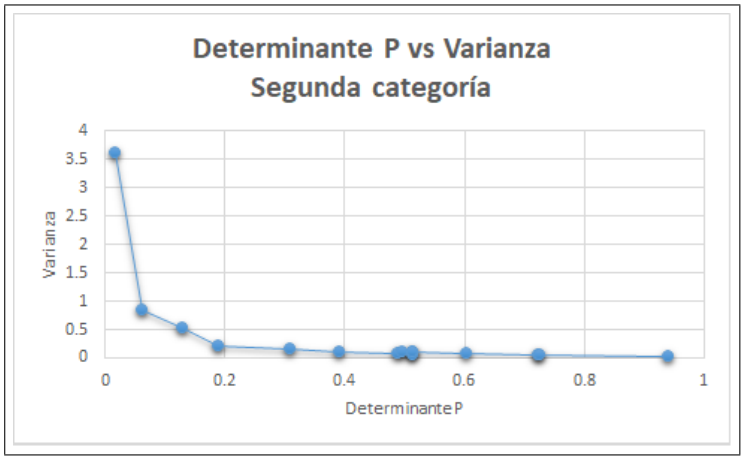

FIGURA 3.5. Comparación entre el determinante de la matriz P y la varianza de la categoría 2 teniendo en cuenta una población N = 20 y una muestra n = 4

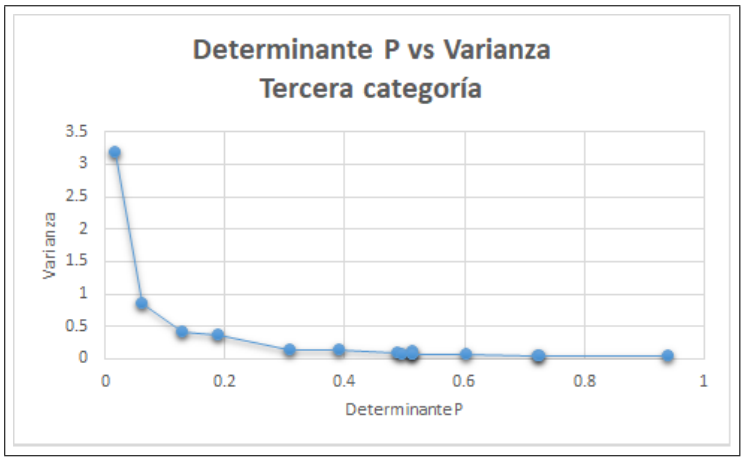

FIGURA 3.6. Comparación entre el determinante de la matriz P y la varianza de la categoría 3 teniendo en cuenta una población N = 20 y una muestra n = 4

CAPÍTULO 4

Aplicación

En esta parte del trabajo se expone la forma en la que la metodología propuesta es llevada a la práctica.

En la actualidad se ha percibido al acoso sexual como un asunto sensible dentro de la población, esto debido a que las víctimas siempre se han sentido atemorizadas a la hora de hablarlo y mas aún de denunciarlo.

Según la directiva 2002/73/CE del parlamento europeo y del consejo de 23 de septiembre de 2002 el acoso sexual se define como:

"la situación en que se produce cualquier comportamiento verbal, no verbal o físico no deseado de índole sexual con el propósito o el efecto de atentar contra la dignidad de una persona, en particular cuando se crea un entorno intimidatorio, hostil, degradante, humillante u ofensivo."

Algunas de las explicaciones a dicha conducta tienen que ver con que las personas que sufren de este delito sienten miedo a perder su empleo, pena al ser señalados o temor a venganza por parte de los victimarios. Esta aplicación pretende dar una idea del porcentaje de trabajadores que a hoy padecen esta violación a sus derechos, ya sea porque intentaron acosarlos y no lo lograron o que finalmente fueron acosados sexualmente, esto minimizando la tasa de no respuesta y protegiendo el anonimato de cada trabajador.

4.1. Objetivo

El objetivo principal del estudio es estimar la proporción de individuos que han sufrido acoso sexual en la población de empleados pertenecientes a la nómina de trabajadores administrativos de la Universidad Nacional de Colombia sede Bogotá en el periodo 2016-II. Con el fin de hacer este estudio útil para los propósitos de esta tesis, se asumirá que esta variable es categórica con tres posibles opciones de respuesta como será descrito más adelante.

4.2. Marco conceptual

4.2.1. Estudios previos

En estudios previos aparece uno realizado en Colombia para el Sensor Yanbal de la Mujer Colombiana (2012) por una firma consultora, en el que se afirma que de la muestra de 600 mujeres y 600 hombres tomada de las principales ciudades del país (Bogotá, Barranquilla, Cali y Medellín) mas del 80 % de los encuestados manifestó que este fenómeno de acoso sexual en el trabajo está presente en su día a día.

En otro estudio realizado por una firma consultora diferente a la anterior y contratada por el Mnisterio de Trabajo de Colombia (2014) que tomó cerca de 1800 empleados de los sectores público y privado, afirmó que el 13 % de los encuestados experimentaron alguna conducta de acoso sexual en el lugar de trabajo. En cuanto al porcentaje de personas que piden algún tipo de ayuda frente a este tipo de situaciones es de notar que es bajo, ya que solamente el 16 % acudió a algún tipo de ayuda o asesoría; de igual forma, es bajo el porcentaje de personas que efectivamente denuncian el acoso sexual en el lugar de trabajo, pues solo el 10 % de los encuestados lo hizo.

En el ámbito internacional, el tema no es muy alejado al caso colombiano pues según un estudio llevado a cabo en España por el Ministerio de trabajo y asuntos sociales de España (2006), teniendo en cuenta a 2.007 mujeres trabajadoras afirmó que el 15 % de ellas han sufrido alguna situación de acoso sexual en el último año pero que más del 40 % no hacen nada o simplemente evitan a la persona.

También, un estudio realizado por la Organización Internacional del Trabajo OIT (2013) afirma que a pesar de la mayor visibilidad que tiene la violencia contra las mujeres en los lugares de trabajo, sigue siendo un problema oculto, pues denunciar estas situaciones sigue siendo traumático para las víctimas y son muy pocas las que lo hacen. Se mencionan algunos datos al rededor del mundo en los que por ejemplo mas del 40 % de las mujeres en los países de la Unión Europea experimentan acoso sexual en su lugar de trabajo, en Asia y el Pacífico indican que del 30 % al 40 % de las trabajadoras reportan alguna forma de acoso sexual y que una cuarta parte de las mujeres y el 16 % de los hombres han sufrido acoso sexual en el lugar de trabajo en Australia.

4.2.2. Población objetivo

La población objetivo considerada se compone por todos los empleados pertenecientes a la nómina de trabajadores administrativos con vinculación pública de la Universidad Nacional de Colombia, Sede Bogotá en el periodo 2016-II, exceptuando docentes con cargos administrativos, contratistas o sus empleados.

4.2.3. Marco muestral

Como marco muestral se emplea el reporte de nómina de los trabajadores administrativos de planta de la Universidad Nacional de Colombia, Sede Bogotá, suministrado por el Departamento de Nutrición y Dietética de la Universidad cuyas variables contenidas son: Nombre del empleado, dependencia, tipo de vinculación con la universidad, edad y correo electrónico. Se cuenta con 1460 registros dentro de los cuales, como se dijo anteriormente,

deben excluirse docentes con cargos administrativos, contratistas o sus empleados, para asi contar con un marco muestral depurado de 1089 individuos.

4.3. Metolodogía

4.3.1. Prueba piloto y operativo en campo

Del marco muestral se tomó una muestra de 36 individuos para realizar una prueba piloto en la que se pudiera decidir cuál mecanismo aleatorio se usaría y que tanta acogida tenía la pregunta y el proceso de aleatorización.

Se utiliza una baraja de cartas como mecanismo aleatorio para decidir el número de incremento que se debe hacer a la respuesta verdadera de cada encuestado. Esta baraja contiene 16 cartas numeradas del 1 al 3 con las siguientes frecuencias:

- 8 cartas para el número 1
- 4 cartas para el número 2
- 4 cartas para el número 3

Lo que significa que la probabilidad de que el encuestado le sume 1 a su respuesta verdadera es 0.5 y que la probabilidad de que le sume 2 ó 3 es de 0.25 para cada caso.

En general la acogida de la pregunta sensible es buena y no se presentan temas como rechazo o molestia a la hora de levantar la información. Para este levantamiento el procedimiento que se lleva a cabo con cada encuestado consiste de 2 partes:

En la primera parte, al inicio de la entrevista, se le explica al encuestado que la pregunta que se le hará, será tratada de un modo diferente en el que nadie podrá saber cuál fue su respuesta, y que el objetivo del estudio es inferir resultados sobre la población y no sobre resultados particulares, de modo tal que se cuida su anonimato y la privacidad de su respuesta.

En la segunda parte, el encuestador debe explicar y seguir los 3 pasos que se enumeran a continuación:

Paso 1 "A continuación voy a leerle una pregunta donde usted debe elegir la opción que mas se adecúe a su respuesta. (Por favor piénsela y no me la diga)

- ¿Ha sufrido usted alguna vez de acoso sexual en su trabajo?
  1 Nunca he sufrido de acoso sexual en mi trabajo
  2 Intentaron acosarme sexualmente pero no lo hicieron.
  3 He sufrido de acoso sexual en mi trabajo.

Paso 2 Ahora, de la siguiente baraja de cartas donde cada carta tienen un número del 1 al 3, va a elegir una, y el valor que le salga lo sumará a su respuesta obtenida anteriormente.

Paso 3 Sí el resultado de la suma del paso anterior es 2 ó 3 repórteme su suma, y sí por el contrario la suma obtenida es 4, 5 ó 6 réstele 3 a este número y reporteme el resultado."

Al final de la prueba piloto se llega a que el 22 % de los encuestados ha sufrido de acoso sexual, valor que se utilizará como proporción para el cálculo del tamaño de la muestra.

4.3.2. Definición de estratos y selección de la muestra

Luego de la prueba piloto se definen 2 estratos de interés, el primero conformado por individuos de sexo masculino y el segundo por individuos de sexo femenino. Con una proporción hallada en la muestra piloto de 0.22, asumiendo un error de 10 %, un nivel de confianza de 90 % y aplicando la ecuación 4.1, que hace referencia al tamaño de muestra de un diseño ESTMAS (Muestreo Aleatorio Simple Estratificado), contenida en Krishnaiah y Rao (1988, página 115),

$$n = \frac{z_{1/2}^2 Np(1-p)}{\varepsilon^2(N-1) + z_{1/2}^2 p(1-p)} \tag{4.1}$$

se llega a un tamaño de 44 individuos. Los cuales, aplicando una afijación óptima (Afijación de Neyman) sin tener en cuenta el costo (Expresión 4.2)

$$n_h = \frac{nN_hP_h(1-P_h)}{\sum_{i=1}^{H} P_i(1-P_i)N_i} \tag{4.2}$$

donde $h = 1, ..., H$ con $H=$ número de estratos, N_h el tamaño de la población del estrato h y P_h la proporción de individuos que han sufrido de acoso sexual en el estrato h, donde $h = 1$ hace referencia al estrato de hombres y $h = 2$ al de mujeres, luego, se obtiene que el tamaño de muestra para cada uno de los dos estratos es de 22 individuos.

4.4. Estimaciones y resultados

Teniendo en cuenta los resultados hallados en el capítulo 2, donde los estimadores de atributos sensibles se presentan con las expresiones 2.25, 2.26 y 2.27 junto con los estimadores de sus varianzas con las expresiones 2.28, 2.29 y 2.30, se llega a la tabla resumen 4.1.

Tabla 4.1. Estimadores y sus varianzas en aplicación sobre acoso sexual

Estrato	Estimadores			Varianzas		
	$\hat{\bar{\phi}}_1$	$\hat{\bar{\phi}}_2$	$\hat{\bar{\phi}}_3$	$\hat{Var}(\hat{\bar{\phi}}_1)$	$\hat{Var}(\hat{\bar{\phi}}_2)$	$\hat{Var}(\hat{\bar{\phi}}_3)$
Total Población	45 %	36 %	18 %	0,087	0,090	0,079
Hombres	27 %	64 %	9 %	0,039	0,046	0,036
Mujeres	64 %	9 %	27 %	0,048	0,044	0,043

En la tabla 4.1 $\hat{\bar{\phi}}_1$, $\hat{\bar{\phi}}_2$ y $\hat{\bar{\phi}}_3$ hacen referencia a la proporción de la población que pertenece a cada grupo (1, 2 ó 3) mientras que $\hat{Var}(\hat{\bar{\phi}}_1)$, $\hat{Var}(\hat{\bar{\phi}}_2)$ y $\hat{Var}(\hat{\bar{\phi}}_3)$ corresponden a la varianza de cada estimador.

Después de aplicar la metodología y obtener los estimadores previamente estudiados se pueden concluir los siguientes puntos:

- Como es de esperarse por estudios previos y por conocimiento de la sociedad actual, las mujeres son las que mas sufren de acoso sexual en su trabajo respecto a los hombres, pues el 27 % de ellas lo han sufrido mientras que para los hombres este porcentaje es de solo el 9 %.
- Poco menos de la tercera parte de los hombres no han sufrido de acoso sexual en el trabajo (27 %) y solo el 9 % señala que fueron víctimas de acoso sexual.
- Para el total poblacional se observa que la varianza es mas alta para el grupo 2 perteneciente a las personas que sintieron que serían víctimas de acoso sexual pero que finalmente no lo sufrieron.
- En general, la estimación de la característica que mide la proporción de la población que ha sufrido de acoso sexual en su trabajo está acorde a lo que señalan los estudios previos en distintas geografías.
- Finalmente, el método utilizado para seleccionar el número aleatortio, que para este caso fue una baraja de cartas, es sencillo y rápido de aplicar, además de que es fácil de entender por parte de los encuestados, sin generar un sesgo adicional a la hora de obtener el número aleatorio, sin embargo debe tenerse en cuenta, que como en todas las TRAs, el método de aleatorización es poco práctico en la realidad.

CAPÍTULO 5

Conclusiones y trabajo futuro

A lo largo de este trabajo se expusieron varias ideas y enfoques distintos en cuanto a la estimación de parámetros sensibles dentro de una población, desde TRAs hasta TCIs, que según las necesidades y recursos del investigador, pueden ser utilizadas para tal fin.

Existe un campo, para el cual las TRA son limitadas en dar una respuesta, es el caso de preguntas sensibles multicategóricas con diseños muestrales complejos; bajo esta condición el estimador propuesto suple esta necesidad, generando un camino para desarrollar y perfeccionar las estimaciones de atributos sensibles en una población.

El estimador propuesto en el segundo capítulo funciona por medio de la adición de una variable aleatoria cuya distribución se conoce, asegurando la privacidad y la no divulgación de valores reales en las respuestas de los encuestados, para dicho estimador y su varianza, en el mismo capítulo, se comprueba analíticamente la propiedad de insesgamiento para luego ejemplificar que por medio de un diseño muestral estratificado aleatorio simple (ESTMAS) es posible llegar al mismo estimador con su varianza y demostrar nuevamente que es insesgado.

En el capítulo tres se constata por medio de simulaciones Monte Carlo y estableciendo diferentes parámetros bajo distintos escenarios, en cuanto a probabilidades de sumar un numero aleatorio entre uno y tres, diferentes configuraciones de tamaño de población y muestra, con un número de iteraciones considerable, se pudo llegar a comprobar que el estimador propuesto seguía siendo insesgado.

Más adelante, en el capítulo cuatro se expone una aplicación real por medio de una encuesta hecha a trabajadores administrativos de la Universidad Nacional de Colombia donde se le preguntó a cada uno sí alguna vez había sufrido de acoso sexual en su trabajo teniendo como opción de respuesta tres características, en esta aplicación se pudo constatar que las estimaciones de parámetros por medio de la metodología propuesta son muy cercanas a las obtenidas en estudios previos, por lo que mediante esta información auxiliar también se puede afirmar que el estimador funciona y que su aplicabilidad es óptima.

Como conclusión final se puede mencionar que el estimador propuesto se ha validado por varios caminos: analíticamente, bajo simulaciones y aplicando el método a un caso real, obteniendo así, resultados satisfactorios en cuanto a la propiedad de insesgamiento y cercanía de la estimación a cifras puntuales propuestas en la parte de simulación y expresadas por otros estudios en el caso de la aplicación.

En la sección dos del capítulo uno donde se mencionan TRAs para variables multicategóricas, se hace necesaria la comparación de dichas técnicas contra la propuesta aunque las demás TRAs solo estén habilitadas para un muestreo aleatorio simple. También, como se ha visto a través del tiempo, los estimadores en las TRAs pueden ser negativos, por lo que se pone sobre la mesa este tipo de situaciones para que sean trabajadas en un futuro.

En cuanto al método propuesto, cabe señalar que será importante abordar el tema de variables multicategóricas cuantitativas, para lo que se ha pensado en la idea de trabajarlo mediante rangos calculados sobre la variable de interés para luego aplicar el mismo método propuesto. En este método también será de mucho interés analizar el comportamiento de las probabilidades de elección del número aleatorio a sumar y poder determinar un criterio de cómo deben ser establecidas dichas probabilidades que permitan una mejora en la varianza de los estimadores.

APÉNDICE

Demostraciones

Demostración Técnica de Warner:

Para llegar al estimador de Warner, se parte del hecho de que cada individuo en una población pertenece a un solo grupo de dos opciones, A y A^c y lo que se quiere es calcular la proporción de individuos que pertencen al grupo A (individuos con la caracterÃstica sensible). Para obtener dicha estimación se escoge un mecanismo aleatorio que permita al encuestado contestar una de dos preguntas con las opciones "si" o "no" sin que el encuestador sepa a cuál de las dos preguntas está dando su respuesta. La probabilidad con la que el mecanismo aleatorio asigna la pregunta sensible A al encuestado se llamará p y será conocida; por el contrario la probabilidad $1 - p$ será la probabilidad de que el mecanismo aleatorio asigne la pregunta A^c.

Sea:

- π = Probabilidad verdadera de pertenecer al grupo A en la población (parámetro a ser estimado).
- p = Probabilidad de contestar a la pregunta sensible. ($p \neq 1/2$)
- $X_i = \begin{cases} 1, & \text{si el i-ésimo encuestado dentro de la muestra contesta "si";} \\ 0, & \text{si el i-ésimo encuestado dentro de la muestra contesta "no".} \end{cases}$

La probabilidad de que el encuestado conteste "si" a la pregunta es,

$$P(X_i = 1) = \pi p + (1 - \pi)(1 - p) \tag{1}$$

equivalente a decir,

$$\begin{aligned} P(X_i = 1) &= p(Tener\ la\ carateristica\ sensible) * p(Contestar\ la\ pregunta\ sensible) \\ &+ p(No\ tener\ la\ carateristica\ sensible) * p(Contestar\ la\ pregunta\ complemento) \end{aligned}$$

y de que conteste "no",

$$P(X_i = 0) = (1 - \pi)p + \pi(1 - p) \tag{2}$$

o dicho de otra forma,

$$\begin{aligned} P(X_i = 0) &= p(No\ tener\ la\ carateristica\ sensible) * p(Contestar\ pregunta\ sensible) \\ &+ p(Tener\ la\ carateristica\ sensible) * p(Contestar\ pregunta\ complemento) \end{aligned}$$

Para estimar el total de individuos que tienen el atributo sensible se tiene que de una muestra de tamaño n, n_1 individuos contestan "si" mientras que $(n - n_1)$ contestan "no" (sin saber sí contestaron a la pregunta A ó A^c) por lo que la función de máxima verosimilitud es,

$$L = [\pi p + (1 - \pi)(1 - p)]^{n_1}[(1 - \pi)p + \pi(1 - p)]^{n-n_1} \tag{3}$$

Para hallar el estimador π se debe maximizar la anterior función de verosimilitud, por lo que el primer paso es aplicar la función logaritmo,

$$\ell = logL = log([\pi p + (1 - \pi)(1 - p)]^{n_1}[(1 - \pi)p + \pi(1 - p)]^{n-n_1}) \tag{4}$$

$$\ell = n_1 log[\pi p + (1 - \pi)(1 - p)] + (n - n_1)log[(1 - \pi)p + \pi(1 - p)] \tag{5}$$

ahora se deriva e iguala a cero para despejar π

$$\begin{aligned} \frac{\partial \ell}{\partial \pi} &= \frac{\partial}{\partial \pi}\ n_1 log[\pi p + (1 - \pi)(1 - p)] + (n - n_1)log[(1 - \pi)p + \pi(1 - p)] \\ &= \frac{\partial}{\partial \pi}\ n_1 log[2\pi p + 1 - p - \pi] + (n - n_1)log[p - 2\pi p + \pi] \\ &= n_1 \frac{2p - 1}{\pi p + (1 - \pi)(1 - p)} + (n - n_1)\frac{1 - 2p}{(1 - \pi)p + \pi(1 - p)} \end{aligned}$$

Igualando a cero se tiene,

$$\begin{aligned}
n_1\frac{2p-1}{\pi p+(1-\pi)(1-p)} &= (n-n_1)\frac{2p-1}{(1-\pi)p+\pi(1-p)} \\
\frac{n-n_1}{n_1} &= \frac{(1-\pi)p+\pi(1-p)}{\pi p+(1-\pi)(1-p)} \\
\frac{n}{n_1}-1 &= \frac{(1-\pi)p+\pi(1-p)}{\pi p+(1-\pi)(1-p)} \\
\frac{n}{n_1}-1 &= \frac{(1-\pi)p+\pi(1-p)}{\pi p+(1-\pi)(1-p)}+1 \\
\frac{n}{n_1}-1 &= \frac{(1-\pi)p+\pi(1-p)+\pi p+(1-\pi)(1-p)}{\pi p+(1-\pi)(1-p)} \\
\frac{n}{n_1}-1 &= \frac{p-\pi p+\pi-\pi p+\pi p+1-p-\pi+\pi p}{\pi p+(1-\pi)(1-p)} \\
\frac{n}{n_1}-1 &= \frac{1}{\pi p+(1-\pi)(1-p)}
\end{aligned}$$

Lo que es igual a:

$$\frac{n_1}{n} = \pi p+(1-\pi)(1-p) \tag{6}$$

la probabilidad de que el encuestado conteste de manera afirmativa a la pregunta.

Ahora se despeja π,

$$\begin{aligned}
\frac{n_1}{n} &= \pi p+1-p-\pi+\pi p \\
\pi(2p-1)+1-p &= \frac{n_1}{n}
\end{aligned}$$

Y finalmente el estimador de la proporción de individuos con la característica sensible es,

$$\hat{\pi} = \frac{n_1}{n(2p-1)}+\frac{p-1}{(2p-1)} \tag{7}$$

O escrito de otra forma

$$\hat{\pi} = \frac{\frac{n_1}{n}+p-1}{2p-1} \tag{8}$$

con $p \neq 1/2$

□

Para hallar el valor esperado del estimador se tiene en cuenta que:

$$n_1 = \sum_n X_i \tag{9}$$

Ahora,

$$\begin{aligned} E[\hat{\pi}] &= E\left[\frac{1}{2p-1}\left((1/n)\sum_n X_i + p - 1\right)\right] \\ &= \frac{1}{2p-1}\left((1/n)\sum_n E[X_i] + p - 1\right) \end{aligned}$$

Debido a que X_i es una variable bernoulli,

$$\sum_n X_i \sim Bin(n, \pi p + (1-\pi)(1-p)) \tag{10}$$

luego

$$\begin{aligned} E[\hat{\pi}] &= \frac{1}{2p-1}[\pi p + (1-\pi)(1-p) + p - 1] \\ &= \frac{1}{2p-1}[\pi p + 1 - p - \pi + \pi p + p - 1] \\ &= \frac{1}{2p-1}[\pi p - \pi + \pi p] \\ &= \frac{1}{2p-1}\pi[2p-1] \\ &= \pi \end{aligned}$$

Lo que demuestra que el estimador es insesgado.

□

Ahora, la varianza será igual a

$$\begin{aligned} Var[\hat{\pi}] &= Var\left[\frac{n_1}{n(2p-1)} + \frac{p-1}{(2p-1)}\right] \\ &= \frac{1}{(2p-1)^2} Var[\frac{n_1}{n} + p - 1] \end{aligned}$$

teniendo en cuenta la igualdad 9,

$$\begin{aligned} Var[\hat{\pi}] &= \frac{1}{n^2(2p-1)^2} Var\left[\frac{\sum_n X_i}{n}\right] \\ &= \frac{1}{n^2(2p-1)^2} Var\left[\frac{\sum_n X_i}{n}\right] \end{aligned}$$

Por 10, se dice que $\sum_n X_i$ tiene como varianza $n[\pi p+(1-\pi)(1-p)][(1-\pi)p+\pi(1-p)]$, luego

$$\begin{aligned}
Var[\hat{\pi}] &= \frac{n[\pi p+(1-\pi)(1-p)][(1-\pi)p+\pi(1-p)]}{n^2(2p-1)^2} \\
&= \frac{4\pi p^2-4\pi^2p^2+4\pi^2p+p-4\pi p+\pi-p^2-\pi^2}{(2p-1)^2n} \\
&= \frac{4\pi p^2-4\pi^2p^2+4\pi^2p-4\pi p+\pi-\pi^2}{(2p-1)^2n}+\frac{p(1-p)}{(2p-1)^2n} \\
&= \frac{1}{n}\left(\frac{4\pi p^2-4\pi p+\pi}{(2p-1)^2}-\frac{4\pi^2p^2-4\pi^2p+\pi^2}{(2p-1)^2}\right)+\frac{p(1-p)}{(2p-1)^2n} \\
&= \frac{1}{n}\left(\frac{\pi(4p^2-4p+1)}{4p^2-4p+1}-\frac{\pi^2(4p^2-4p+1)}{4p^2-4p+1}\right)+\frac{p(1-p)}{(2p-1)^2n} \\
&= \frac{1}{n}(\pi-\pi^2)+\frac{p(1-p)}{(2p-1)^2n} \\
&= \frac{\pi-\pi^2}{n}+\frac{p(1-p)}{(2p-1)^2n}
\end{aligned}$$

Finalmente,

$$Var[\hat{\pi}] = \frac{\pi(1-\pi)}{n}+\frac{p(1-p)}{(2p-1)^2n} \tag{11}$$

Donde el último componente de la ecuación hace referencia a la varianza del mecanismo aleatorio.

□

Estimador de Horvitz-Thompson

Este estimador está dado por:

$$\hat{t}_\pi = \sum_s \frac{y_k}{\pi_k}$$

para cualquier diseño $p(s)$ con probabilidades de inclusión π_k y π_{kl}. Este estimador es insesgado.

Demostración:

El valor esperado del estimador está dado por:

$$\begin{aligned} E[\hat{t}_\pi] &= E\left[\sum_s \frac{y_k}{\pi_k}\right] \\ &= E\left[\sum_U I_k \frac{y_k}{\pi_k}\right] \\ &= \sum_U E(I_k)\frac{y_k}{\pi_k} \end{aligned}$$

como I_k es una variable aleatoria Bernoulli entonces $E(I_k) = \pi_k$, luego,

$$E(\hat{t}_\pi) = \sum_U y_k = t$$

con lo que se demuestra que el estimador de Horvitz-Thompson es insesgado.

□

Ahora, su varianza está dada por

$$\begin{aligned} V[\hat{t}_\pi] &= V\left[\sum_s \frac{y_k}{\pi_k}\right] \\ &= V\left[\sum_U I_k \frac{y_k}{\pi_k}\right] \\ &= \sum_U V\left[I_k \frac{y_k}{\pi_k}\right] + \sum\sum_{k\neq l U} C\left(I_k \frac{y_k}{\pi_k}, I_l \frac{y_l}{\pi_l}\right) \\ &= \sum_U \frac{y_k^2}{\pi_k^2} V(I_k) + \sum\sum_{k\neq l U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l} C(I_k, I_l) \\ &= \sum_U \frac{y_k^2}{\pi_k^2}\pi_k(1-\pi_k) + \sum\sum_{k\neq l U} \frac{y_k}{\pi_k}\frac{y_l}{\pi_l}(\pi_{kl} - \pi_k\pi_l) \\ &= \sum\sum_U \Delta_{kl}\breve{y}_k\breve{y}_l \end{aligned}$$

Luego,

$$V[\hat{t}_\pi] = \sum \sum_U \Delta_{kl} \breve{y}_k \breve{y}_l \tag{12}$$

Donde $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$, $\breve{y}_k = \dfrac{y_k}{\pi_k}$ y $\breve{y}_l = \dfrac{y_l}{\pi_l}$.

□

Ahora,un estimador insesgado para $V[\hat{t}_\pi]$ es:

$$\hat{V}[\hat{t}_\pi] = \sum \sum_s \breve{\Delta}_{kl} \breve{y}_k \breve{y}_l \tag{13}$$

con $\breve{\Delta}_{kl} = \dfrac{\Delta_{kl}}{\pi_{kl}}$

Demostración:

$$\begin{aligned}
E(\hat{V}[\hat{t}_\pi]) &= E\left(\sum \sum_s \breve{\Delta}_{kl} \breve{y}_k \breve{y}_l\right) \\
&= E\left(\sum \sum_U I_k I_l \frac{\Delta_{kl}}{\pi_{kl}} \breve{y}_k \breve{y}_l\right) \\
&= \sum \sum_U E(I_k I_l) \frac{\Delta_{kl}}{\pi_{kl}} \breve{y}_k \breve{y}_l \\
&= \sum \sum_U \Delta_{kl} \breve{y}_k \breve{y}_l \\
&= V[\hat{t}_\pi]
\end{aligned}$$

□

APÉNDICE

Código en R

```
#rm(list=ls())
#### Instalación de paquetes ####
install.packages("combinat")
install.packages("gtools")
require(combinat)
require(gtools)
#### Parámetros iniciales ####
N=6
n=4
#### Probabilidades de sumar 1,2 ó 3 a la respuesta original
(número aleatorio) ####
p1=0.5
p2=0.25
p3=1-p1-p2
#### Número de iteraciones, se multiplica asi para que las matrices A de
números aleatorios y C de valores reales se puedan sumar ####
h=1000*(3^n)
### Frecuencias reales de la población (es con lo que finalmente se compara) ####
prop_1=2
prop_2=3
prop_3=1

############## Llamado de función ##############
piestimador(N,n,p1,p2,p3,h,prop_1,prop_2,prop_3)
################################################

#### FUNCIÓN ####

piestimador<-function(N,n,p1,p2,p3,h,prop_1,prop_2,prop_3)
{
#### Vector de probabilidades del número aleatorio ####
prob=matrix(c(p1,p2,p3),1,3,byrow=TRUE)
```

```
### Matriz "P" definida en el capítulo 2 ####
p=matrix(c(p3-p1,p2-p1,p1-p2,p3-p2),nr=2,nc=2,byrow=TRUE)

#### GENERACIÓN DE MATRIZ DE VALORES REALES C ####

#### Matriz con el total de posibles combinaciones de individuos
multiplicado por el número de iteraciones (N C n)*h ####
C_0=matrix(combn(N,n),nrow=(choose(N,n)*h), ncol=n, byrow=TRUE)

#### Vector de proporciones verdaderas en la población ####
Vector=c(rep(1,prop_1),rep(2,prop_2),rep(3,prop_3))

#### Matriz de valores reales ####
#### Cada bucle asigna una respuesta verdadera a cada individuo ####
#### según el vector calculado en el paso anterior ####
C=matrix(0,ncol=n,nrow=(choose(N,n)*h))
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==1){C[i,j] = Vector[1]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==2){C[i,j] = Vector[2]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==3){C[i,j] = Vector[3]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==4){C[i,j] = Vector[4]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==5){C[i,j] = Vector[5]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==6){C[i,j] = Vector[6]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==7){C[i,j] = Vector[7]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==8){C[i,j] = Vector[8]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==9){C[i,j] = Vector[9]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==10){C[i,j] = Vector[10]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==11){C[i,j] = Vector[11]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==12){C[i,j] = Vector[12]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==13){C[i,j] = Vector[13]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==14){C[i,j] = Vector[14]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==15){C[i,j] = Vector[15]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==16){C[i,j] = Vector[16]}}}
for(i in 1:dim(C_0)[1])
```

```
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==17){C[i,j] = Vector[17]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==18){C[i,j] = Vector[18]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==19){C[i,j] = Vector[19]}}}
for(i in 1:dim(C_0)[1])
{for(j in 1:dim(C_0)[2]){if(C_0[i,j]==20){C[i,j] = Vector[20]}}}

#### GENERACIÓN DE MATRIZ DE VALORES ALEATORIOS A ####

A=matrix((t(permutations(n=3, r=n, repeats.allowed=TRUE))),
            nrow=(choose(N,n)*h), ncol=n,byrow=TRUE)

#### MATRIZ DE PROBABILIDAD TOTAL ####

tabA=matrix(0,nrow=(choose(N,n)*h),ncol=dim(prob)[2])

#### Acá se cuenta el numero de valores que tiene la respuesta en ####
#### cada vector ####
for(i in 1:dim(tabA)[1])
  {tabA[i,]= t(as.matrix(table(factor(A[i,], levels=c(1:dim(prob)[2])))))}

#### Vector de probabilidades repetido NCn*h ####
Pr=as.matrix(rep(prob, (choose(N,n)*h)))

#### Acomoda el vector con filas de 3 columnas (# de categorías) ####
Pr1=t(matrix(Pr, nrow= dim(prob)[2],ncol= (choose(N,n)*h)))

#### p en pr1 elevado al # de veces que se contó en tabA ####
Pad = matrix(0,nrow=(choose(N,n)*h),ncol= dim(prob)[2])
for(i in 1:dim(Pad)[1]){
for(j in 1:dim(Pad)[2]){Pad[i,j]= Pr1[i,j]^tabA[i,j]}}

#### Productoria de las 3 componentes en Pad ####
Pa=as.matrix(apply(Pad,1,prod),(choose(N,n)*h),1)

#### Probabilidad de MAS * Probabilidad de cada muestra según el ####
#### número aleatorio seleccionado####
Ps=(1/choose(N,n))*Pa

#### MATRIZ CON RESPUESTAS TRANSFORMADAS ####

R=A+C
for(i in 1:dim(R)[1]){
  for(j in 1:dim(R)[2]){
    if (R[i,j]>3){R[i,j]=R[i,j]-3} # codifica cada respuesta
  }
}
```

```
if (n == 3) {

#### MATRIZ "y" DE 1'S Y 0'S (RESPUESTAS CODIFICADAS) ####

Y=matrix(0,ncol=9, nrow=dim(R)[1])
for(i in 1:dim(R)[1]){if(R[i,1]==1){Y[i,1] = 1}}
for(i in 1:dim(R)[1]){if(R[i,2]==1){Y[i,2] = 1}}
for(i in 1:dim(R)[1]){if(R[i,3]==1){Y[i,3] = 1}}
for(i in 1:dim(R)[1]){if(R[i,1]==2){Y[i,4] = 1}}
for(i in 1:dim(R)[1]){if(R[i,2]==2){Y[i,5] = 1}}
for(i in 1:dim(R)[1]){if(R[i,3]==2){Y[i,6] = 1}}
for(i in 1:dim(R)[1]){if(R[i,1]==3){Y[i,7] = 1}}
for(i in 1:dim(R)[1]){if(R[i,2]==3){Y[i,8] = 1}}
for(i in 1:dim(R)[1]){if(R[i,3]==3){Y[i,9] = 1}}

#### Vectores con estimaciones por individuo de cada atributo sensible ####

phihat_11=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_11[i,1]=(1/det(p))*(((p3-p2)*(Y[i,1]-p1))+((p1-p2)*(Y[i,4]-p2)))}

phihat_12=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_12[i,1]=(1/det(p))*(((p3-p2)*(Y[i,2]-p1))+((p1-p2)*(Y[i,5]-p2)))}

phihat_13=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_13[i,1]=(1/det(p))*(((p3-p2)*(Y[i,3]-p1))+((p1-p2)*(Y[i,6]-p2)))}

phihat_21=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_21[i,1]=(1/det(p))*(((p2-p1)*(Y[i,1]-p1))+((p3-p1)*(Y[i,4]-p2)))}

phihat_22=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_22[i,1]=(1/det(p))*(((p2-p1)*(Y[i,2]-p1))+((p3-p1)*(Y[i,5]-p2)))}

phihat_23=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihat_23[i,1]=(1/det(p))*(((p2-p1)*(Y[i,3]-p1))+((p3-p1)*(Y[i,6]-p2)))}

phihat_31=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){phihat_31[i,1]=1-phihat_11[i,1]-phihat_21[i,1]}

phihat_32=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){phihat_32[i,1]=1-phihat_12[i,1]-phihat_22[i,1]}

phihat_33=matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){phihat_33[i,1]=1-phihat_13[i,1]-phihat_23[i,1]}
```

```
#### Matriz con phi's estimados ####

phihat=cbind(phihat_11,phihat_12,phihat_13,
             phihat_21,phihat_22,phihat_23,
             phihat_31,phihat_32,phihat_33)

#### Matriz de promedios ####

phihatprom=matrix(0,ncol=3, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihatprom[i,1]=mean(cbind(phihat[i,1],phihat[i,2],phihat[i,3]))}
for(i in 1:dim(Y)[1])
{phihatprom[i,2]=mean(cbind(phihat[i,4],phihat[i,5],phihat[i,6]))}
for(i in 1:dim(Y)[1])
{phihatprom[i,3]=mean(cbind(phihat[i,7],phihat[i,8],phihat[i,9]))}

#### Varianzas estimadas de los phis estimados ####

phihatvar=matrix(0,ncol=3, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihatvar[i,1]=var(c(phihat[i,1],phihat[i,2],phihat[i,3]))}
for(i in 1:dim(Y)[1])
{phihatvar[i,2]=var(c(phihat[i,4],phihat[i,5],phihat[i,6]))}
for(i in 1:dim(Y)[1])
{phihatvar[i,3]=var(c(phihat[i,7],phihat[i,8],phihat[i,9]))}

#### Varianzas * estimadas de los phis estimados ####

var_est_phi_est_1<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_1[i,1]=
    (1/det(p)^2)*((((p3-p2)^2)*(var(Y[i,1:3])))
    +(2*(p3-p2)*(p1-p2)*(cov(Y[i,1:3],Y[i,4:6])))
    +(((p1-p2)^2)*(var(Y[i,4:6]))))}

var_est_phi_est_2<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_2[i,1]=
    (1/det(p)^2)*((((p2-p1)^2)*(var(Y[i,1:3])))
    +(2*(p2-p1)*(p3-p1)*(cov(Y[i,1:3],Y[i,4:6])))
    +(((p3-p1)^2)*(var(Y[i,4:6]))))}

var_est_phi_est_3<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_3[i,1]=
    (1/det(p)^2)*((((p3-p1)^2)*(var(Y[i,1:3])))
    +(2*(p3-p1)*(p3-p2)*(cov(Y[i,1:3],Y[i,4:6])))
    +(((p3-p2)^2)*(var(Y[i,4:6]))))}

}
```

```
if (n == 4) {

#### Matriz "y" de 1's y 0's (respuestas codificadas) ####

  Y=matrix(0,ncol=12, nrow=dim(R)[1])
  for(i in 1:dim(R)[1]){if(R[i,1]==1){Y[i,1] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,2]==1){Y[i,2] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,3]==1){Y[i,3] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,4]==1){Y[i,4] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,1]==2){Y[i,5] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,2]==2){Y[i,6] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,3]==2){Y[i,7] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,4]==2){Y[i,8] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,1]==3){Y[i,9] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,2]==3){Y[i,10] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,3]==3){Y[i,11] = 1}}
  for(i in 1:dim(R)[1]){if(R[i,4]==3){Y[i,12] = 1}}

#### Vectores con estimaciones por individuo de cada atributo sensible ####

  phihat_11=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_11[i,1]=(1/det(p))*(((p3-p2)*(Y[i,1]-p1))+((p1-p2)*(Y[i,5]-p2)))}

  phihat_12=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_12[i,1]=(1/det(p))*(((p3-p2)*(Y[i,2]-p1))+((p1-p2)*(Y[i,6]-p2)))}

  phihat_13=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_13[i,1]=(1/det(p))*(((p3-p2)*(Y[i,3]-p1))+((p1-p2)*(Y[i,7]-p2)))}

  phihat_14=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_14[i,1]=(1/det(p))*(((p3-p2)*(Y[i,4]-p1))+((p1-p2)*(Y[i,8]-p2)))}


  phihat_21=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_21[i,1]=(1/det(p))*(((p2-p1)*(Y[i,1]-p1))+((p3-p1)*(Y[i,5]-p2)))}

  phihat_22=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_22[i,1]=(1/det(p))*(((p2-p1)*(Y[i,2]-p1))+((p3-p1)*(Y[i,6]-p2)))}

  phihat_23=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_23[i,1]=(1/det(p))*(((p2-p1)*(Y[i,3]-p1))+((p3-p1)*(Y[i,7]-p2)))}
```

```
  phihat_24=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihat_24[i,1]=(1/det(p))*(((p2-p1)*(Y[i,4]-p1))+((p3-p1)*(Y[i,8]-p2)))}


  phihat_31=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1]){phihat_31[i,1]=1-phihat_11[i,1]-phihat_21[i,1]}

  phihat_32=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1]){phihat_32[i,1]=1-phihat_12[i,1]-phihat_22[i,1]}

  phihat_33=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1]){phihat_33[i,1]=1-phihat_13[i,1]-phihat_23[i,1]}

  phihat_34=matrix(0,ncol=1, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1]){phihat_34[i,1]=1-phihat_14[i,1]-phihat_24[i,1]}


  #### Matriz con phi's estimados ####

  phihat=cbind(phihat_11,phihat_12,phihat_13,phihat_14,
               phihat_21,phihat_22,phihat_23,phihat_24,
               phihat_31,phihat_32,phihat_33,phihat_34)

  #### Matriz de promedios ####

  phihatprom=matrix(0,ncol=3, nrow=dim(Y)[1])
  for(i in 1:dim(Y)[1])
  {phihatprom[i,1]=mean(cbind(phihat[i,1],phihat[i,2],phihat[i,3],phihat[i,4]))}
  for(i in 1:dim(Y)[1])
  {phihatprom[i,2]=mean(cbind(phihat[i,5],phihat[i,6],phihat[i,7],phihat[i,8]))}
  for(i in 1:dim(Y)[1])
  {phihatprom[i,3]=mean(cbind(phihat[i,9],phihat[i,10],phihat[i,11],phihat[i,12]))}

#### Varianzas estimadas de los phis estimados ####

phihatvar=matrix(0,ncol=3, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1])
{phihatvar[i,1]=var(c(phihat[i,1],phihat[i,2],phihat[i,3],phihat[i,4]))}
for(i in 1:dim(Y)[1])
{phihatvar[i,2]=var(c(phihat[i,5],phihat[i,6],phihat[i,7],phihat[i,8]))}
for(i in 1:dim(Y)[1])
{phihatvar[i,3]=var(c(phihat[i,9],phihat[i,10],phihat[i,11],phihat[i,12]))}

#### Varianzas * estimadas de los phis estimados ####

var_est_phi_est_1<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_1[i,1]=
    (1/det(p)^2)*((((p3-p2)^2)*(var(Y[i,1:4]))))
```

```
   +(2*(p3-p2)*(p1-p2)*(cov(Y[i,1:4],Y[i,5:8])))
   +(((p1-p2)^2)*(var(Y[i,5:8]))))}

var_est_phi_est_2<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_2[i,1]=
    (1/det(p)^2)*((((p2-p1)^2)*(var(Y[i,1:4])))
   +(2*(p2-p1)*(p3-p1)*(cov(Y[i,1:4],Y[i,5:8])))
   +(((p3-p1)^2)*(var(Y[i,5:8]))))}

var_est_phi_est_3<-matrix(0,ncol=1, nrow=dim(Y)[1])
for(i in 1:dim(Y)[1]){var_est_phi_est_3[i,1]=
    (1/det(p)^2)*((((p3-p1)^2)*(var(Y[i,1:4])))
   +(2*(p3-p1)*(p3-p2)*(cov(Y[i,1:4],Y[i,5:8])))
   +(((p3-p2)^2)*(var(Y[i,5:8]))))}
}

phiestimador<-(t(phihatprom)%*%Ps)%*%(1/(h/(3^n)))

#### se divide por el numero de veces que se genera la matriz A, ####
#### es decir h/27 para el caso n=3 ####

#### Varianza estimada de los phis estimados ####

var_est_phiestimador_1<-(1/n)*((1-(n/N))*phihatvar[,1])
                        +(1/n^2)*var_est_phi_est_1
var_est_phiestimador_2<-(1/n)*((1-(n/N))*phihatvar[,2])
                        +(1/n^2)*var_est_phi_est_2
var_est_phiestimador_3<-(1/n)*((1-(n/N))*phihatvar[,3])
                        +(1/n^2)*var_est_phi_est_3
var_est_phiestimador<-cbind(var_est_phiestimador_1,var_est_phiestimador_2,
                            var_est_phiestimador_3)

varianza_est_phiestimador<-(t(var_est_phiestimador)%*%Ps)%*%(1/(h/(3^n)))

#### se divide por el numero de veces que se genera la matriz A, ####
#### es decir h/27 para el caso n=3 ####

return (matrix(c(phiestimador,varianza_est_phiestimador),nrow=3,ncol=2))
}
```

APÉNDICE

Tablas de simulación

Grupo 1

TABLA 1. Estimadores de proporción de atributos sensibles y sus varianzas con N=6, n=3, p_1=0.1, p_2=0.6 y p_3=0.3

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	1/6=0.166	2/6=0.333	4/6=0.666
ϕ_2	1/6=0.166	3/6=0.5	1/6=0.166
ϕ_3	4/6=0.666	1/6=0.166	1/6=0.166
$\hat{\hat{\phi}}_1$	0.166	0.333	0.666
$\hat{\hat{\phi}}_2$	0.166	0.5	0.166
$\hat{\hat{\phi}}_3$	0.666	0.166	0.166
$Var(\hat{\hat{\phi}}_1)$	0.370	0.298	0.330
$Var(\hat{\hat{\phi}}_2)$	0.255	0.362	0.370
$Var(\hat{\hat{\phi}}_3)$	0.330	0.332	0.255

TABLA 2. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=3, p_1=0.1, p_2=0.6 y p_3=0.3

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.511	0.101	0.402
$\hat{\hat{\phi}}_2$	0.182	0.191	0.537
$\hat{\hat{\phi}}_3$	0.305	0.707	0.059
$Var(\hat{\hat{\phi}}_1)$	0.436	0.440	0.336
$Var(\hat{\hat{\phi}}_2)$	0.417	0.302	0.474
$Var(\hat{\hat{\phi}}_3)$	0.364	0.409	0.373

TABLA 3. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=4, p_1=0.1, p_2=0.6 y p_3=0.3

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.505	0.1	0.402
$\hat{\hat{\phi}}_2$	0.189	0.197	0.509
$\hat{\hat{\phi}}_3$	0.304	0.702	0.087
$Var(\hat{\hat{\phi}}_1)$	0.267	0.270	0.219
$Var(\hat{\hat{\phi}}_2)$	0.257	0.187	0.288
$Var(\hat{\hat{\phi}}_3)$	0.224	0.253	0.233

TABLA 4. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=3, p_1=0.1, p_2=0.6 y p_3=0.3

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.748	0.354	0.299
$\hat{\hat{\phi}}_2$	0.099	0.392	0.550
$\hat{\hat{\phi}}_3$	0.152	0.252	0.150
$Var(\hat{\hat{\phi}}_1)$	0.449	0.443	0.397
$Var(\hat{\hat{\phi}}_2)$	0.515	0.486	0.494
$Var(\hat{\hat{\phi}}_3)$	0.326	0.461	0.473

TABLA 5. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=4, p_1=0.1, p_2=0.6 y p_3=0.3

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.750	0.352	0.3
$\hat{\hat{\phi}}_2$	0.097	0.398	0.549
$\hat{\hat{\phi}}_3$	0.151	0.248	0.149
$Var(\hat{\hat{\phi}}_1)$	0.298	0.294	0.263
$Var(\hat{\hat{\phi}}_2)$	0.343	0.325	0.329
$Var(\hat{\hat{\phi}}_3)$	0.216	0.306	0.314

Grupo 2

TABLA 6. Estimadores de proporción de atributos sensibles y sus varianzas con N=6, n=3, p_1=0.6, p_2=0.3 y p_3=0.1

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	1/6=0.166	2/6=0.333	4/6=0.666
ϕ_2	1/6=0.166	3/6=0.5	1/6=0.166
ϕ_3	4/6=0.666	1/6=0.166	1/6=0.166
$\hat{\hat{\phi}}_1$	0.166	0.333	0.666
$\hat{\hat{\phi}}_2$	0.166	0.5	0.166
$\hat{\hat{\phi}}_3$	0.666	0.166	0.166
$Var(\hat{\hat{\phi}}_1)$	0.370	0.298	0.330
$Var(\hat{\hat{\phi}}_2)$	0.255	0.362	0.370
$Var(\hat{\hat{\phi}}_3)$	0.330	0.332	0.255

TABLA 7. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=3, p_1=0.6, p_2=0.3 y p_3=0.1

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.508	0.105	0.374
$\hat{\hat{\phi}}_2$	0.202	0.209	0.469
$\hat{\hat{\phi}}_3$	0.289	0.685	0.156
$Var(\hat{\hat{\phi}}_1)$	0.428	0.434	0.382
$Var(\hat{\hat{\phi}}_2)$	0.418	0.307	0.450
$Var(\hat{\hat{\phi}}_3)$	0.363	0.412	0.398

TABLA 8. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=4, p_1=0.6, p_2=0.3 y p_3=0.1

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.505	0.1	0.395
$\hat{\hat{\phi}}_2$	0.201	0.202	0.493
$\hat{\hat{\phi}}_3$	0.292	0.697	0.110
$Var(\hat{\hat{\phi}}_1)$	0.265	0.269	0.225
$Var(\hat{\hat{\phi}}_2)$	0.260	0.188	0.280
$Var(\hat{\hat{\phi}}_3)$	0.223	0.255	0.237

TABLA 9. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=3, p_1=0.6, p_2=0.3 y p_3=0.1

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.755	0.355	0.302
$\hat{\hat{\phi}}_2$	0.092	0.403	0.547
$\hat{\hat{\phi}}_3$	0.151	0.240	0.150
$Var(\hat{\hat{\phi}}_1)$	0.446	0.440	0.396
$Var(\hat{\hat{\phi}}_2)$	0.513	0.490	0.493
$Var(\hat{\hat{\phi}}_3)$	0.324	0.458	0.472

TABLA 10. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=4, p_1=0.6, p_2=0.3 y p_3=0.1

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.752	0.350	0.3
$\hat{\hat{\phi}}_2$	0.098	0.402	0.550
$\hat{\hat{\phi}}_3$	0.149	0.246	0.149
$Var(\hat{\hat{\phi}}_1)$	0.297	0.294	0.263
$Var(\hat{\hat{\phi}}_2)$	0.343	0.324	0.321
$Var(\hat{\hat{\phi}}_3)$	0.216	0.307	0.314

Grupo 3

TABLA 11. Estimadores de proporción de atributos sensibles y sus varianzas con N=6, n=3, p_1=0.3, p_2=0.1 y p_3=0.6

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	1/6=0.166	2/6=0.333	4/6=0.666
ϕ_2	1/6=0.166	3/6=0.5	1/6=0.166
ϕ_3	4/6=0.666	1/6=0.166	1/6=0.166
$\hat{\hat{\phi}}_1$	0.166	0.333	0.666
$\hat{\hat{\phi}}_2$	0.166	0.5	0.166
$\hat{\hat{\phi}}_3$	0.666	0.166	0.166
$Var(\hat{\hat{\phi}}_1)$	0.370	0.298	0.330
$Var(\hat{\hat{\phi}}_2)$	0.255	0.362	0.370
$Var(\hat{\hat{\phi}}_3)$	0.330	0.332	0.255

TABLA 12. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=3, p_1=0.3, p_2=0.1 y p_3=0.6

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.480	0.093	0.422
$\hat{\hat{\phi}}_2$	0.215	0.199	0.492
$\hat{\hat{\phi}}_3$	0.304	0.707	0.084
$Var(\hat{\hat{\phi}}_1)$	0.430	0.439	0.358
$Var(\hat{\hat{\phi}}_2)$	0.422	0.304	0.456
$Var(\hat{\hat{\phi}}_3)$	0.366	0.412	0.371

TABLA 13. Estimadores de proporción de atributos sensibles y sus varianzas con N=10, n=4, p_1=0.3, p_2=0.1 y p_3=0.6

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	5/10=0.5	1/10=0.1	4/10=0.4
ϕ_2	2/10=0.2	2/10=0.2	5/10=0.5
ϕ_3	3/10=0.3	7/10=0.7	1/10=0.1
$\hat{\hat{\phi}}_1$	0.489	0.1	0.401
$\hat{\hat{\phi}}_2$	0.208	0.199	0.496
$\hat{\hat{\phi}}_3$	0.302	0.7	0.101
$Var(\hat{\hat{\phi}}_1)$	0.267	0.270	0.219
$Var(\hat{\hat{\phi}}_2)$	0.259	0.188	0.283
$Var(\hat{\hat{\phi}}_3)$	0.226	0.253	0.235

TABLA 14. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=3, p_1=0.3, p_2=0.1 y p_3=0.6

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\hat{\phi}}_1$	0.746	0.339	0.298
$\hat{\hat{\phi}}_2$	0.108	0.403	0.552
$\hat{\hat{\phi}}_3$	0.145	0.256	0.148
$Var(\hat{\hat{\phi}}_1)$	0.445	0.445	0.395
$Var(\hat{\hat{\phi}}_2)$	0.519	0.486	0.495
$Var(\hat{\hat{\phi}}_3)$	0.325	0.466	0.472

TABLA 15. Estimadores de proporción de atributos sensibles y sus varianzas con N=20, n=4, p_1=0.3, p_2=0.1 y p_3=0.6

Dato	Escenario 1	Escenario 2	Escenario 3
ϕ_1	15/20=0.75	7/20=0.35	6/20=0.3
ϕ_2	2/20=0.1	8/20=0.4	11/20=0.55
ϕ_3	3/20=0.15	5/20=0.25	3/20=0.15
$\hat{\tilde{\phi}}_1$	0.747	0.346	0.298
$\hat{\tilde{\phi}}_2$	0.103	0.398	0.550
$\hat{\tilde{\phi}}_3$	0.149	0.254	0.151
$Var(\hat{\tilde{\phi}}_1)$	0.297	0.295	0.263
$Var(\hat{\tilde{\phi}}_2)$	0.344	0.323	0.329
$Var(\hat{\tilde{\phi}}_3)$	0.216	0.308	0.314

Bibliografía

- Abul, Ela, A.A., Greenberg, B.G., y Horvitz, D.G., A Multiproportions Randomized Response Model, Journal of the American Statistical Association, 62 (1967), pp. 990-1008.

- "Acoso sexual en el trabajo y masculinidad. Exploración con hombres de la población general: Centroamérica y República Dominicana", estudio realizado por la Organización Internacional del Trabajo, referenciado en: "Presentación del estudio: Acoso sexual en el trabajo y masculinidad: Centroamérica y República Dominicana". Organización Internacional del Trabajo, 27 de Febrero de 2013, página web: `http://www.ilo.org/sanjose/quienes-somos/direcci%C3%B3n/presentaciones/WCMS_205852/lang--es/index.htm`

- Al Sobhi, M.M; Hussain, Z. y Al-Zahrani, B. (2014), General randomized response techniques using Polya's urn process as a randomization device, PLoS ONE 9(12): e115612. doi:10.1371/journal.pone.0115612.

- Cobo, Beatriz (2013), Respuesta aleatoria y técnicas de preguntas indirectas (Tesis de Maestría). Universidad de Granada, Granada, España.

- "El acoso sexual a las mujeres en el ámbito laboral, estudio realizado por Inmark Estudios y Estrategias S. A., mostrado en: "El acoso sexual a las mujeres en el ámbito laboral", Instituto de la mujer de España, 26 de Abril de 2006, página web: `https://www.navarra.es/NR/rdonlyres/D91FE499-4898-4EDD-AA09-213A8AF122EA/153594/MTASEstudioAcosoSexual.pdf`

- "Estudio de Acoso Sexual en el ámbito Laboral", estudio realizado por Consultores en Información - Infométrika S.A.S, referenciado en: "El silencio del acoso sexual en el mundo laboral", Ministerio de trabajo de Colombia, 25 de Noviembre de 2014, página web: `http://www.mintrabajo.gov.co/noviembre/4035-el-silencio-del-acoso-sexual-en-el-mundo-laboral.html.`

- Greenberg, B. G.; Kuebler, R. R., Jr.; Abernathy, J. R. y Hovertiz, D. G. (1971), Application of the randomized response techniques in obtaining quantitative data, Journal of the American Statistical Association, 66, pp. 243-250.

- Greenberg, B.G; Abul-ela, A.A.; Simmons, W.R. y Horvitz, D.C. (1969). The unrelated question RR model: theoretical framework. En Journal of the American Statistical Association, vol. 64, pp. 520-539.

- Gregory R. Warnes, Ben Bolker y Thomas Lumley (2015). gtools: Various R Programming Tools. R package version 3.5.0. http://CRAN.R-project.org/package=gtools

- Gupta, S.; Gupta, B. y Singh, S. (2002), Estimation of sensitivity level of personal interview survey questions, Journal of Statistical Planning and Inference, 100, pp. 239-247.

- Horvitz, D.C.; Greenberg, B.G. y Abernathy, J.R. (1976). Randomized response. A data gathering device for sensitive questions. En International Statistical Review, vol. 44, pp. 181-196.

- Horvitz, D.G.; Shah, B.V. y Simmons, W.R. (1967). The unrelated question randomized response model. Social Statistics Section Proceedings of the American Statistical Association, pp. 65-72.

- Hussain, Z.; Shah, E. y Shabbir, J. (2012), An alternative item count technique in sensitive surveys, Revista Colombiana de Estadistica, 35, pp. 39-54.

- Hussain, Z. (2011), Randomized Response Models in Survey Sampling, Randomized Response Models, VDM Verlag Dr Muller.

- Imai, K. (2011), Multivariate regression analysis for the item count technique, Journal of the American Statistical Association, 106(494), pp. 407-416.

- Kim, J.I. y Flueck, J. A. (1978), An additive randomized response model, Proceedings of the Survey Research Section, American Statistical Association, pp. 351 355.

- Krishnaiah, P. R. y Rao, C. R. (1988). Sampling. Amsterdam: North-Holland.

- Mangat, N.S. y Singh, R. (1990). An alternative randomized response procedures. En Biometrika, vol. 77, pp. 439-442.

- Moors, J.J. (1971). Optimization on the unrelated question in RR model. En Journal of the American Statistical Association, vol. 66, pp. 627-629.

- Parlamento Europeo y Consejo (de 23 de septiembre de 2002). «DIRECTIVA 2002/73/CE relativa a la aplicación del principio de igualdad de trato entre hombres y mujeres en lo que se refiere al acceso al empleo, a la formación y a la promoción profesionales, y a las condiciones de trabajo», p. 3.

- Ryu, J.-B.; Kim, J.-M.; Heo, T.-Y. y Park, C. G. (2005), On stratified randomized response sampling, Model Assisted Statistics and Applications, 1(1), pp. 31-36.

- Särndal C.E.; Swensson B. y Wretman J. (1992). Model Assisted Survey Sampling, Springer-Verlag.

- Scott Chasalow (2012). combinat: combinatorics utilities. R package version 0.0-8. http://CRAN.R-project.org/package=combinat

- "Sensor Yanbal de la mujer colombiana 2012", estudio realizado por Ipsos Napoleon Franco, referenciado en: "¿A qué se debe la inconformidad laboral de la mujer colombiana?", Revista Semana, 14 de Abril de 2012, página web: http://www.semana.com/vida-moderna/articulo/a-que-debe-inconformidad-laboral-mujer-colombiana/256366-3.

- Shannon R.E., 1988, Simulación de Sistemas. Diseño, desarrollo e implementación, Trillas.

- Soberanis,Cruz, V., Ramirez-Valverde, G., Pérez-Elizalde, S. y González-Cossio, F. (2008) Randomized response sampling in finite populations, a unifying approach. En Agrociencia, vol. 42, pp. 537-549.

- Stem, D.E. y Steinhorst, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model. En Journal of the American Statistical Association, vol. 79, num. 387, pp. 555-564.

- Trujillo, L. y González L. M. (2012). Preguntas sensibles en encuestas y metodologías alternativas para asegurar la veracidad y fidelidad de las respuestas. En Revista Ib De La Información Básica Estadística, ISSN: 2256-1552, vol.2 pp.55 - 69.

- Trujillo, L. y Soberanis, V.(2012), Estimación de características sensibles en encuestas por muestreo. XXII Simposio de estadistica, Bucaramanga, Colombia.

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