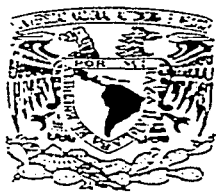


UNIVERSIDAD NACIONAL
AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
"ACATLÁN"

MÉTODOS DE AJUSTE POR NO-RESPUESTA EN ENCUESTAS POR MUESTREO

T E S I N A

PARA OBTENER EL TÍTULO DE:
A C T U A R I O

PRESENTA:

SERGIO CEBALLOS MELO

ASESOR: ACT. LUIS ALEJANDRO TAVERA PÉREZ

NOVIEMBRE 2003

Mamá gracias por todo lo que has hecho por mi
Por ser un ejemplo de integridad, sabiduría y fortaleza
Sobre todo por tu amor
Te amo

Agradecimientos

Este trabajo se pudo realizar en gran parte por el apoyo prestado por el Act. Arturo Blancas Espejo y el Lic. Ramón Saburit Cervantes y su empresa Certiorum C.S., agradezco su apoyo.

También quiero agradecer profundamente los valiosos comentarios de la Act. y Mat. Jessica Marmolejo Hernández que contribuyeron a este trabajo, gracias Jess.

Quiero agradecer a mis amigos con los que cursé la carrera, por los momentos de estudio y apoyo, y sobre todo por su amistad Carlos, Elvia, Alma, Beatriz y Claudia.

A mis amigos con los que he tenido la oportunidad de desempeñarme como actuario y compartir experiencias profesionales y personales Jorge, Maru y Noriega.

Por último agradezco a la Universidad Nacional Autónoma de México y a los contribuyentes del pueblo de México, por haberme proporcionado los medios para desempeñar la carrera de actuario y darme la oportunidad de servir a la sociedad mexicana.

Métodos de ajuste por no-respuesta en encuestas por muestreo.

Resumen

Investigar, analizar, y exponer las técnicas propuestas para ajustar una encuesta por muestreo cuando ocurre no-respuesta, ya sea de unidades de observación completas o de tan sólo en alguna variable de interés. Se exponen el marco teórico de la no-respuesta, el muestreo en dos fases enfocado a ajustar por no-respuesta, métodos de imputación, ponderación y brevemente los métodos paramétricos.

Palabras clave: No-respuesta, muestreo en dos fases, métodos de imputación, ponderación y paramétricos para ajustar por no-respuesta.

Abstract

Research, analyze, and expose some techniques proposed to adjust a sampling survey when non-response occurs, either for unit non-response or item non-response. It is exposed the general theory, use of two-phase sampling for non-response, imputation, weighting methods and parametric models for non-response.

Key words: non-response, two-phase sampling for non-response, imputation, weighting methods and parametric models for non-response.

ÍNDICE

Capítulo I

Introducción

1.1 El error de muestreo y errores de no muestreo en una encuesta por muestreo

1.1.1 Introducción general

El objetivo de una encuesta por muestreo, por lo general, es obtener estimaciones de totales, promedios o proporciones de ciertas características de la población bajo estudio (Cochran 1977).

Si U es el conjunto de una población finita de tamaño N; es decir para una variable *y* de una característica de interés:

$$U = \{y_1, y_2, \ldots, y_N\}$$

El total de una población es:

$$t_U = \sum_{k=1}^{N} y_k$$

El promedio de una población es:

$$\bar{y}_U = \frac{1}{N}\sum_{k=1}^{N} y_k$$

La proporción de la población:

$$p_U = \frac{1}{N}\sum_{k=1}^{N} I_k$$

Donde N es el número de elementos de la población finita, y_k es un valor fijo asociado a una una característica del elemento k o elemento de la población, que

puede ser medido como una variable cualitativa ya sea en escala de intervalo o razón. Por ejemplo, edad, sueldo, etc.

En el caso de la proporción I_i es una variable indicadora donde con valor 1 si el elemento tiene la caracteristica y 0 si no la tiene. Este caso aplica regularmente para variables cualitativas, nominales u ordinales. Por ejemplo, género, escolaridad, etc.

Cuando estos valores del universo son desconocidos, entonces se realiza una encuesta por muestreo, donde mediante un procedimiento se selecciona una muestra aleatoria de la población original de tamaño *n* para hacer una estimación del valor real desconocido.

Asi:

$$\hat{t} = f_1(y_1, y_2, \ldots, y_n)$$

$$\hat{\bar{Y}} = f_2(y_1, y_2, \ldots, y_n)$$

$$\hat{p} = f_3(y_1, y_2, \ldots, y_n)$$

Donde el gorro indica que se trata de una estimación del valor correspondiente (total, media o proporción) en el universo a partir de una muestra, y donde la forma que cada estadístico f_i depende del diseño o estrategia de muestreo utilizado, y de la información auxiliar, si es el caso que haya sido incorporada al diseño. En el presente trabajo el concepto de *diseño de muestreo* será utilizado frecuentemente, visto como la función de probabilidad que describe la selección de las unidades de

observación en muestra. En ocasiones el *diseño de muestreo* será denotado por $P(\cdot)$.

Adicionalmente, es posible construir dichos estimadores con el uso de un vector x_i de k variables auxiliares asociadas a cada unidad de observación. Estos estimadores son llamados de razón o de regresión según sea la metodología usada.

1.1.2 Error de Muestreo

Las estimaciones de totales, promedios y proporciones están sujetas a un error. Donde se le llama *'error'* a la diferencia entre el valor estimado con una muestra en particular contra el valor real de la población, pero desconocido. Este concepto de error es consecuencia directa de la variabilidad entre una muestra y otra, es decir la distribución de muestreo.

Así, el muestreo es una metodología que tiene como una de sus principales características que a las estimaciones se les puede medir lo máximo que pueden diferir del valor real desconocido y la probabilidad que ello ocurra. Matemáticamente:

$$P\left(\hat{y} - \bar{y}_U < \varepsilon\right) = 1 - \alpha$$

Donde:

$\hat{y}$ es el valor promedio estimado a través de la muestra

$\bar{Y}$ es el valor promedio real, pero desconocido de la población

ε es la máxima diferencia entre el valor estimado y el valor verdadero.

$1-\alpha$ es el nivel de confianza.

La misma fórmula aplica para totales y proporciones, realizando las sustituciones correspondientes. El error y el nivel de confianza son las dos medidas para el error de muestreo.

El error de muestreo es la consecuencia directa de contar con una muestra en vez de medir a la población completa. Así, dado un procedimiento de selección de una muestra (producto directo del diseño), es posible calcular el 'error', junto con la probabilidad de que ello ocurra (confianza). Es importante recalcar que la magnitud del error y la confianza son producto de la variabilidad entre una muestra y otra muestra, por lo tanto para su cálculo no toma en consideración la distribución de la población que se toma en la muestra.

1.1.3 Errores de no muestreo

Sin embargo, el error de muestreo no es la única fuente que hace que el valor real y el valor estimado a partir de una muestra difieran. Existen los errores de "no muestreo" que tienen como dificultad que sus distribuciones de probabilidad son desconocidas; así, cuando ocurren estos errores no es posible medirlos y además pueden ser de una magnitud tan considerable que los resultados de una encuesta se tornan inválidos, porque causan que las estimaciones estén sesgadas en una magnitud y dirección desconocidas (Lohr 1999 y Särndal et al. 1992).

Por otra parte, es practica común publicar resultados donde sólo se menciona el error de muestreo; haciendo parecer que los resultados son sumamente confiables. Pero en la realidad no se hace mención alguna de los errores de no muestreo.

Los errores de no muestreo tienen como mayor consecuencia que las estimaciones se encuentran sesgadas, es decir, distorsionadas de forma sistemática. Consecuentemente, para la mayoría de las posibles muestras se obtendrá un valor por arriba o por debajo en una *magnitud desconocida* del valor real, de forma sistemática. Como el sesgo es la consecuencia de los errores de no muestreo, se hará referencia a los diversos tipos de errores de no muestreo en términos de sesgos: Selección sesgada y Medición sesgada.

1.1.3.1 Selección sesgada

Toda selección se realiza a partir de un "marco de muestreo". Un marco de muestreo se define como un listado de la "población objetivo". A su vez, la población objetivo es la población teórica de la que se desea obtener información. Además, la selección sesgada se puede dar por otras causas que serán descritas con mayor detalle a continuación.

Estos son los problemas derivados de un marco de muestreo:

- **El marco de muestreo tiene listados elementos que no pertenecen a la población objetivo.** Por ejemplo, en un directorio de una cámara de comercio, si

éste se actualizara cada seis meses, seguramente con el transcurso del tiempo habrá negocios que ya no estarán en el mercado pero continuarán registrados como si fueran negocios vigentes.

- **Al marco de muestreo le faltan elementos que pertenecen a la población objetivo.** Continuando con el ejemplo del directorio de una cámara de comercio de sus afiliados, igualmente seguramente hay negocios nuevos que no han sido registrados.

Además, también existe selección sesgada cuando una muestra no es seleccionada bajo un procedimiento aleatorio, y por tanto, este tipo de error de no-muestreo desaparece si se utiliza algún método de muestreo. A continuación se muestran los ejemplos más comunes:

- **Muestra por conveniencia,** es aquella donde se seleccionan la muestra en función de la facilidad para obtener la información. Por ejemplo, realizar una encuesta a una persona con aspecto de ser amistosa, ignorando aquellas que no se muestran de buen humor o apuradas.

- **Utilizar un procedimiento de selección que inadvertidamente dependa de alguna característica asociada con las propiedades de interés.** Esto ocurre en muestras obtenidas de voluntarios. Las personas que se ofrecen a participar en un estudio, pueden diferir significativamente de quién no es voluntario.

- **Muestra según el criterio de un "experto".** Por ejemplo, dado el criterio de un experto se selecciona una muestra "representativa". El uso de criterios subjetivos, tiene implícito que el "experto" deliberadamente ignorará a ciertos elementos por no considerarlos representativos de la población. Además, existe la posibilidad que si se tomara la opinión de "otro experto", seguramente él seleccionaría una muestra completamente distinta.

- **Muestra autoseleccionada.** Por ejemplo, el caso de las encuestas que realizan algunas empresas de televisión, donde se invita la gente a llamar, sin importar el número de llamadas, no representa a la opinión pública mexicana, tan sólo representan la opinión de aquellos que decidieron llamar.

También es fuente de error de no-muestreo un error humano en el diseño conceptual de la investigación o una instrucción operativa errónea. Se muestran ejemplos respectivamente a cada caso.

- **Realizar la definición de la población objetivo errónea.** Por ejemplo, si en una encuesta que tenga como propósito conocer la opinión que los consumidores tienen de un producto, y se definiría de forma errónea a la población objetivo (consumidores) como adultos de 18 a 30 años, pero en la realidad los adolescentes de 12 a 18 años son de forman importante consumidores de dicho producto, entonces los resultados de la encuesta pueden estar sesgados.

- **Substituir a un elemento inaccesible por otro que sea accesible.** Por ejemplo, en vez de entrevistar al domicilio designado porque no hay nadie en casa o rechazo la entrevista, entrevistar al vecino de a lado. Es posible, que la característica de interés esté asociada con personas que no tiendan a estar en casa

Por ultimo la No-respuesta, que es el tema que atañe el presente trabajo de investigación, es un error del tipo de selección sesgada. En el capitulo siguiente se efectuará el desarrollo de este tipo de error a detalle.

1.1.3.2 medición sesgada

El otro tipo importante de error de no muestreo ocurre cuando la medición está sesgada, es decir cuando la medición tiene la tendencia de diferir del valor verdadero en alguna dirección. A continuación se listarán los errores de medición más comunes en encuestas a personas:

- **La gente miente.** Por ejemplo, en México cuando es tiempo de elecciones en zonas donde habitan personas sumamente marginadas o las pertenecientes a sindicatos, este tipo de gente puede llegar a mentir sobre su preferencia electoral por miedo a que por represalia le sean suspendidos los beneficios que reciben del gobierno.

- **La gente no entiende la pregunta.** Por ejemplo, una pregunta diciendo "¿Qué proporción del dia ve usted televisión?" Puede llegar a ser interpretada como ¿cuánto tiempo ve usted televisión? ¿Le gusta la televisión? O simplemente no la entiende.

- **La gente olvida.** Por ejemplo, preguntar que programa de televisión vio una persona en la noche anterior, puede llevar a que la gente lo confunda con el programa que vio dos o tres dias anteriores y pero lo reporta al encuestador porque le gustó pero en realidad lo vio en otra ocasión y no lo recuerda.

- **La gente proporciona diferentes respuestas a diferentes entrevistadores.** Por ejemplo, en una encuesta sobre hábitos sexuales, un entrevistador hombre a una mujer seguramente obtendrá diferentes respuestas que si la entrevista la realizara una mujer.

- **La gente tal vez responda según lo que crea puede dar una buena impresión al entrevistador.** Por ejemplo, una pregunta sobre si se está dispuesto a apoyar a un incremento de impuestos para mejorar la asistencia a los grupos más necesitados de la sociedad, tal vez la gente tienda a responder que sí, pero si se hiciera un plebiscito la gente tal vez votaría en sentido contrario al ser secreto el voto. Algunos autores le llaman a esto el "factor de la vergüenza".

- **El entrevistador puede distorsionar la respuesta.** Por ejemplo, leyendo mal las preguntas, anotando respuestas equivocadas, o discutiendo con el entrevistado para proporcione una respuesta en función de lo que el entrevistador cree verdadero o moralmente aceptable.

- **Ciertas palabras pueden significar cosas distintas a personas distintas.** Por ejemplo, la pregunta ¿Es usted dueño de una casa?, La respuesta puede depender de la interpretación del entrevistado del concepto de *dueño* (¿incluye el concepto de dueño cuando todavía se están realizando pagos por la vivienda?) Y de la palabra casa (¿terrenos o departamentos están incluidos?)

1.2 Comentarios introductorios sobre la no-respuesta

Idealmente, en una encuesta por muestreo, diferentes variables son medidas a todas las unidades de observación seleccionadas en la muestra. Sin embargo, datos incompletos o faltantes de dos tipos pueden tener lugar.

En el primer tipo, ninguna de las variables es medida para una unidad dada, este tipo de no-respuesta es llamada no-respuesta de la unidad de observación.

Por otra parte, se tiene el caso donde la mayoría de las preguntas es contestada, pero ciertas preguntas, o no son contestadas o la respuesta obtenida se puede juzgar como que contiene un error grande y por tanto, se borra el dato obtenido durante el procesamiento de la calidad de la información. Por lo general, tales preguntas pueden ser calificadas como del tipo *sensitivas*. Por ejemplo, información sobre los ingresos. Pero también, simplemente es posible que el entrevistado no cuente, sepa o recuerde la información que se le solicita.

Una visión simplista de la consecuencia de la no-respuesta es que se ha obtenido un menor tamaño de muestra que con la que inicialmente se planeó la encuesta. Sin embargo, frecuentemente existen razones para creer, que el grupo de individuos que no responden a una encuesta sistemáticamente difieren de los que si responden. Deming (Cochran 1983) lo expresó en estos términos para las poblaciones humanas "Las personas que no responden a una encuesta son, en cierta forma, y en varios grados diferentes a aquellos que sí responden". Por ende, al realizar estimaciones sólo de la información disponible de los datos derivados

de aquellos que respondieron la encuesta e ignorando la no-respuesta, se puede incurrir en estimaciones sesgadas cuya magnitud es desconocida.

Dos psicólogos, Rosenthal y Rosnow en 1975 (Cochran 1983) resumieron los resultados de diversas investigaciones para identificar el perfil de los individuos que tienden a no responder una encuesta, dentro de la población estadounidense. Así ellos encontraron que las personas que tienden a responder a encuestas (comparadas con los que no responden encuestas) se trata de personas más inteligentes, con mayor educación, más interesados en las preguntas, y se sienten menos apenados o amenazados por las preguntas.

Por otra parte, en resultados de estudios efectuados en el Reino Unido (Lohr 1999) se perfiló a las personas que no responden a las encuestas como: Residentes de Londres, hogares sin automóvil, personas solteras, parejas sin niños, gente mayor, gente divorciada o viuda, originaria de nuevos países de la Commonwealth, personas de menor escolaridad, y subempleados.

Es muy importante que en México se realizara un estudio de esta naturaleza y que fuera auspiciado ya sea por la iniciativa privada o por el sector público. Evidentemente sería un estudio conjunto entre sociólogos, psicólogos y estadísticos, cuya utilidad principal sería tener una idea de la dirección que los sesgos ocasionados por no-respuesta pueden tomar.

1.3 Muestreo por cuota como un tratamiento erróneo para la no-respuesta

El muestreo por cuota, fue una respuesta inicial para atacar la no-respuesta. Esta técnica que fue propuesta por Cherigton, Roper, Gallup y Crossley (Cochran 1983). Formalmente, este método evita la no-respuesta al no requerir que una persona en específico sea entrevistada. En su lugar, se pide al encuestador que busque personas que reúnan ciertas características (en la práctica generalmente se buscan cuotas por edad y sexo).

Sin embargo, el método pierde toda validez científica porque no existen las probabilidades de selección de las unidades que respondieron la encuesta y porque seguramente hay un número importante de unidades de la población objetivo que tuvo probabilidad cero de ser seleccionadas. Por lo tanto, el muestreo por cuota no es considerado en la presente tesina, aunque su uso es común en la práctica en diversas entidades que ejercen el muestreo como su actividad principal, desafortunadamente.

Capítulo II

Marco teórico de la no-respuesta

2.1 Definición de no-respuesta, definición de los 2 tipos y formas de medirla

2.1.1 Introducción

Para formalizar la exposición para la no-respuesta partiremos de la siguiente notación:

En toda encuesta se levantan q variables de estudio a cada unidad de observación k de la muestra; es decir, $y_1,\ldots,y_j,\ldots,y_q$, es decir, q preguntas dentro de un cuestionario a aplicar.

Sea y_{jk} el valor de la variable j para el k-ésimo elemento o unidad de observación de una muestra aleatoria. Además, sea n_s el tamaño de muestra especificado como el necesario para alcanzar un nivel de confianza y error arbitrarios para una muestra s seleccionada bajo un procedimiento aleatorio conocido de selección (diseño de muestreo $P(\cdot)$).

- **Definición de respuesta completa**

Se define como respuesta completa cuando dada la información recabada de una muestra s, (se denota con minúscula indicando que la muestra ya fue seleccionada y se hizo el levantamiento de las q variables a las n_s unidades de

observación), si para todo elemento $k \in s$ se observa un vector *completo* y *consistente* de la forma:

$$y_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk})$$

Donde 'completo' implica que no haya variables sin respuesta. Además, la otra restricción de 'consistencia' de los datos, implica que todas las respuestas de variables cuantitativas o cualitativas se encuentran dentro de un rango válido y consistente con los demás datos. En caso contrario, dicho dato en cuestión debe ser borrado para en su lugar editar un valor en blanco (no-respuesta) en el proceso de control de calidad de los datos.

2.1.2 Los dos tipos de no-respuesta

2.1.2.1 No-respuesta de una unidad de observación:

Si para el k-ésimo elemento de una muestra en *todas* las *q* variables del vector $y_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk})$ no tienen información se define como *no-respuesta de la unidad k de observación*. Es decir, todas las variables no fueron contestadas para ciertas unidades de observación específicas.

Frecuentemente, esto ocurre en los siguientes casos:

- El entrevistado rehusó participar en la encuesta
- No fue localizado

- Fue localizado, pero no fue posible tener acceso al entrevistado
- Otra razón que impidió la obtención de la información

2.1.2.2 No-respuesta Puntual (a preguntas específicas):

Al elemento *k* se le define que tiene *no-respuesta puntual* si para al menos una, pero no todas las *q* variables del vector $y_k = (y_{1k}, \ldots, y_{jk}, \ldots, y_{qk})$ está en blanco, o es inconsistente y se edita a blanco en el proceso de control de calidad.

Por ejemplo, supongase una encuesta donde el k–esimo elemento se negó a declarar su ingreso mensual, lugar de trabajo, u otra pregunta sensitiva o no sensitiva, pero si contestó las demás preguntas.

2.1.3 Definición de conjuntos de respuestas.

Se definen ciertos conjuntos con motivo de establecer medidas descriptivas simples de la magnitud de no-respuesta. El uso de algunos de estos conjuntos también será útil para describir en términos de probabilidad, los tres mecanismos (modelos) de no-respuestas, que se han propuesto y que posteriormente serán descritos.

Dada una muestra s, sea r_j el conjunto de respuestas de las unidades de observación correspondientes a la variable j_j. Se puede escribir:

$$r_j = \{k : k \in S \text{ y } y_{jk} \text{ aceptable se ha registrado}\}$$

En toda encuesta, existen q conjuntos por lo general no idénticos de respuestas, $r_1, r_2, \ldots, r_q$.

A partir de estos conjuntos es posible construir otros para establecer otras medidas de no-respuesta que pueden ser útiles.

Sea r_u el conjunto de unidades de observación que hayan respondido a por lo menos una pregunta.

Por otra parte, el conjunto r_c está compuesto por las unidades de observación que hayan respondido a todas las preguntas.

2.1.4 Una propuesta de medición descriptiva de la no-respuesta

Es necesario realizar el conteo de elementos de los diversos conjuntos mencionados, mismos que serán utilizado para construir diversos indicadores de no-respuesta para una encuesta en específico. Esta cuantificación es la cardinalidad de cada conjunto, Särndal et al. (1992).

La medición de los diversos conjuntos ya definidos en el subcapítulo previo: s, r_u, r_i y r_c, se denota por n_s, n_{r_u}, n_{r_i}, y n_{r_c} respectivamente.

2.1.4.1 Medición de la no-respuesta de las unidades de observación que por lo menos contestaron a una pregunta

Medida de respuesta de las unidades de observación que por lo menos contestaron a una pregunta, sin ponderar:

$$p_{r_u} = \frac{n_{r_u}}{n_s}$$

por ende la no-respuesta de este tipo es el complemento:

$$1 - p_{r_u}$$

En este caso, p_{r_u} mide que tanto éxito se ha tenido en la muestra en obtener al menos una respuesta de los elementos en la muestra seleccionada.

Otra alternativa de medición, es utilizando la ponderación. La medida ponderada es:

$$\tilde{p}_{r_u} = \frac{\sum_{r_u} 1/\pi_k}{\sum_{s} 1/\pi_k}$$

π_k = Probabilidad de inclusión de primer orden del elemento k

donde la medida correspondiente de no-respuesta es:

$$1 - \tilde{p}_{r_u}$$

En donde $\tilde{p}_{r_u}$ se puede interpretar como la estimación de la probabilidad de respuesta de al menos una pregunta o variable, pero en el universo.

2.1.4.2 Medidas de no-respuesta de unidades de observación que tuvieron respuestas completas

$$p_{r_c} = \frac{n_{r_c}}{n_s}$$

Análogamente, la no-respuesta es el complemento:

$1-p_i$

Igualmente, se pueden construir medidas ponderadas usando las probabilidades de inclusión de primer orden.

2.1.4.3 Medidas de no-respuesta para la j-ésima pregunta

Se pueden aplicar las mismas fórmulas para cada una de q preguntas en cuestión, de forma ponderada y sin ponderar.

2.1.5 Otras propuestas de medición de no-respuesta

2.1.5.1 Propuesta de Groves de medición de no-respuesta

Groves (1989) hace la propuesta de medir tres tasas de no-respuesta para encuestas, especificamente para el caso donde todos los miembros del marco de muestreo están enumerados explícitamente para ser medidos (es decir, el caso donde los elementos de observación son directamente seleccionados). Se clasifica a cada entrevista o unidad de observación bajo una y sólo una de las siguientes categorias:

I=	Entrevista completa
P=	Entrevista parcial
NC=	Unidad no contactada pero parte de la muestra especificada
R=	Unidad que rehusó constestar la encuesta
NE=	Unidad que no debería formar parte del marco de muestreo por no ser parte de la población objetivo (Error en el marco de muestreo)
NI=	Unidad no entrevistada por otra causa

Siendo cada una de estas categorías variables indicadoras, para efectos del cálculo de las siguientes tasas.

Así, se define la *tasa de cooperación* como:

$$\frac{\sum I}{\sum I + \sum P + \sum R}$$

que se puede interpretar como la tasa de entrevistas completas con respecto a todos los casos donde se estableció un contacto capaz de ser entrevistado de forma completa. Esta tasa se puede utilizar como una herramienta para medir el grado de efectividad del personal de campo en obtener la información dado que la unidad tuvo la posibilidad de ser entrevistada.

Por otra parte, si se adicionan los casos no contactados en el denominador y las entrevistas parciales en el numerador:

$$\frac{\sum I + \sum P}{\sum I + \sum P + \sum R + \sum NC}$$

Este cociente se puede utilizar para medir la proporción de aquellos casos en muestra que pudieron proveer información, y de hecho la proporcionaron.

Por último, una medición más universalmente aceptada es:

$$\frac{\sum I}{\sum I + \sum P + \sum R + \sum NC + \sum NI}$$

Que es una medida de la proporción donde se contemplan todos los casos contra las encuestas completas.

Igualmente, estas tasas deben construirse de forma ponderada y sin ponderar.

2.1.5.2 Propuesta de medición de no-respuesta de Hidiroglou et al.

Por otra parte, Hidiroglou et al. (1993) proponen las siguientes tasas de no-respuesta. Éstas están definidas de acuerdo a la experiencia de diversas encuestas llevadas a cabo por la agencia de gobierno de estadística de Canadá.

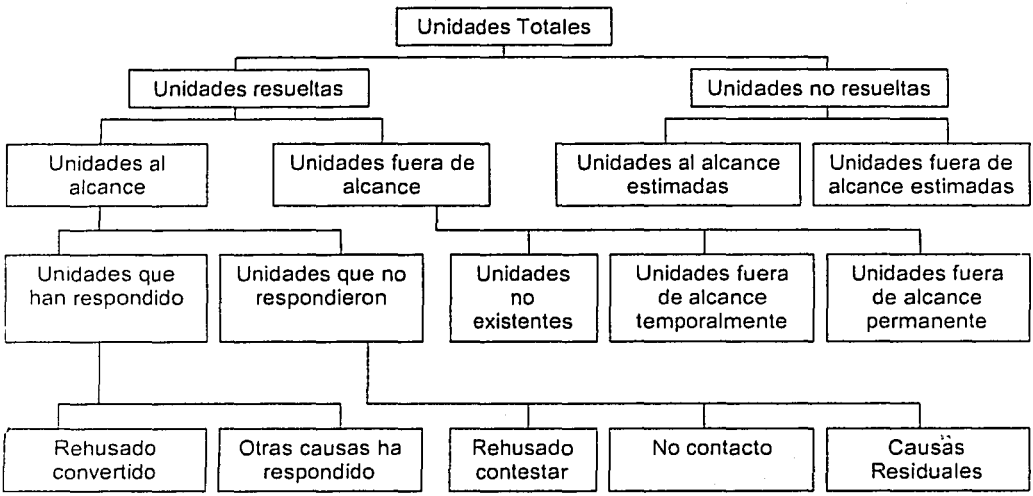
- Se define al *'Total de unidades'* como el número de unidades de observación que se creen como parte de la población objetivo y que forman parte de un marco de muestreo *antes* de que la encuesta se lleve a cabo

- Se define a *'Unidades resueltas'* como aquellas cuya clasificación de pertenecer o no pertenecer a la población objetivo es conocida al momento de levantar la encuesta

- *'Unidades al alcance'* son unidades que sí forman parte de la población objetivo

- *'Unidades fuera de alcance'* es el complemento de las *'unidades al alcance'*. Es decir, no pertenecen a la población objetivo

Así a cada respuesta obtenida se clasifica dentro de una de estas categorías según se muestra en el siguiente cuadro:



Los autores proponen calcular las siguientes tasas (de forma ponderada y sin ponderar):

- *Tasa de fuera de alcance.*- es el cociente del número de unidades de observación fuera de alcance entre el número de unidades resueltas. Esta tasa se interpreta como una medida de la calidad del marco de muestreo utilizado al indicar la proporción de unidades que efectivamente pertenecen a la población objetivo dentro del marco.

- *Tasa de no contacto.*- el cociente del número de unidades de observación al alcance pero que no se estableció contacto y unidades no resueltas entre el número de unidades dentro de alcance y unidades no resueltas. Esta tasa se interpreta como la proporción de la muestra que no fue contactada por el personal de campo.

- *Tasa de rechazo.*- el cociente del número de unidades al alcance que no respondieron entre el número de unidades al alcance. Esta tasa se interpreta como la proporción de unidades que pudieron contestar por estar al alcance pero que a fin de cuentas rechazaron contestar la encuesta por cualquier causa.

- *Tasa de no-respuesta.*- el cociente del número de unidades que no respondieron y que no fueron resueltas entre el número de unidades al alcance y no resueltas. Esta tasa se interpreta como una medida global de no-respuesta independientemente de las causas.

2.1.6 Comentarios adicionales

La elección de cual de las medidas mencionadas se debe calcular para obtener una visión general de la no-respuesta se debe hacer de acuerdo con las necesidades e intereses específicas de cada encuesta y la organización que la realiza.

Las medidas ponderadas y sin ponderar pueden variar significativamente, y se debe hacer el cálculo de ambas para tener una medición más clara de su dimensión en una encuesta.

Se debe destacar que estos indicadores son medidas descriptivas de la magnitud de la no-respuesta ocurrida en un caso en particular, que *no indican de forma alguna* una magnitud del sesgo de las estimaciones.

Por último, se debe reportar explícitamente cual fórmula es la que se está utilizando al momento de reportar la tasa de no-respuesta.

2.2 Mecanismos de respuesta

2.2.1 Antecedentes

Se entiende por mecanismo de respuesta, a un modelo de probabilidad de respuesta condicional de que un individuo responda dado que fue seleccionado en muestra. Se han propuesto tres modelos o mecanismos que describen la probabilidad de respuesta, así en función de cuál es el modelo que mejor describa la probabilidad de respuesta para una muestra en particular se debe decidir cuál es el modelo más apropiado a aplicar.

Sea *r* un subconjunto de una muestra *s* que está conformado por los elementos que respondieron una encuesta. La dificultad inherente es que no es posible determinar con certeza cómo se generan estos conjuntos *r*. En la práctica, se hace el supuesto de que existe una ley de probabilidad que describe la generación del conjunto *r*, es decir, una $P(r \mid s)$ que la probabilidad condicional el que conjunto *r* sea generado dada una muestra especifica seleccionada *s*. No obstante, la distribución de probabilidad del conjunto de respuesta usualmente es desconocida. Así, para obtener conclusiones sólo tomando información de los individuos que han respondido la encuesta, se está obligado a hacerlo a partir de supuestos sobre la distribución de la probabilidad de respuesta. La mayor parte de las veces, dichos supuestos no se pueden verificar. Así se cae en el dilema de que la validez de las inferencias depende de supuestos no verificables. No obstante, en la práctica se tiene la necesidad de construir dichos modelos, buscando que el

mecanismo utilizado se apegue a la realidad lo más posible. Por ende, la selección de modelo adecuado se hace según el criterio y experiencia del investigador sobre el caso en particular en cuestión. Incluso, se puede apoyar la decisión con otros especialistas.

La construcción de los mecanismos de no-respuesta descritos a continuación, se toma de una propuesta hecha por Rosenbaum y Rubin (1983). Ellos proponen utilizar una probabilidad condicional, misma que se utiliza para realizar *un ajuste* a las unidades de observación de la muestra que respondieron para efectuar la estimación de un parámetro (medias, totales o proporciones) de tal forma que se pueda eliminar o reducir el sesgo ocasionado por la no-respuesta. Esta probabilidad condicional es denotada como ϕ_i, y está asociada a cada unidad de observación k de la muestra y es llamada 'calificación de la propensión', que formalmente es la probabilidad condicional de asignar un tratamiento particular a la unidad de observación dado un vector de covariantes. En el caso de la no-respuesta hay dos tratamientos posibles: el primero, es ser una unidad de observación a la que se sí respondió la encuesta, y el segundo es el caso de las unidades donde no hubo respuesta. Se hace uso de un vector de covariantes x_i, que son mediciones observadas previamente a cada unidad de observación cuya información forma parte del marco de muestreo. De hecho, algunas de dichas mediciones pueden ser utilizadas como información auxiliar en el diseño $P(\cdot)$.

Así, se define una variable aleatoria indicadora:

$$R_k = \begin{cases} 1 & \text{Si la unidad k sí responde} \\ 0 & \text{Si la unidad k no responde} \end{cases}$$

Una vez que se ha seleccionado una muestra y hecho el levantamiento, los resultados de esta variable indicadora de respuesta son conocidas para los *n* elementos de la muestra seleccionada.

Así, la probabilidad de que un individuo *k* responda es $\phi_k = P(R_k = 1)$.

Y para la estimación de la calificación de propensión se hace uso de dicho vector de covariantes x_k. La probabilidad de respuesta ϕ_k es utilizada para ajustar la no-respuesta. Más adelante se dan los detalles.

2.2.2 Tres modelos de no-respuesta

Las características de los tres modelos que describen a los mecanismos de no-respuesta según fueron propuestos por Little y Rubin (1987) se describen a continuación:

2.2.2.1 Respuestas faltantes completamente de forma aleatoria (*Missing Completely at Ramdom MCAR por sus siglas en inglés*)

Se identifica este mecanismo de no-respuesta cuando ϕ_k no depende de x_k ni de y_k, ni del diseño de la muestra $P(\cdot)$.

Matemáticamente:

$$P(R_k = 1 \mid \mathbf{X}_k = \mathbf{x}_k) = P(R_k = 1) = \phi_k$$

y $P(R_k = 1 \mid Y_k = y_k) = P(R_k = 1) = \phi_k$

y $P(R_k = 1 \mid S_k = s_k) = P(R_k = 1) = \phi_k$

Donde S es una variable aleatoria de todas las muestras posibles. Y R es la variable aleatoria de las respuestas descrita en 2.2.1. Aunque en el muestreo clásico los valores y_k se consideran fijos y no como producto de una variable aleatoria en sí para efectos de modelar la no-respuesta resulta conveniente considerarlos como resultado de una variable aleatoria.

La independencia de ϕ_k indica que no hay algún patrón común identificable entre las unidades de observación que sí contestaron comparado contra las que no. Esta ausencia de patrón implica que con el sólo uso de las unidades que respondieron la encuesta basta para realizar estimaciones insesgadas, por lo tanto la forma de los estimadores es igual al caso respuesta completa, es decir no se aplica ningún factor de corrección por no-respuesta. Se considera como único impacto de la ausencia de unidades de observación un aumento de la varianza de la estimación, pero con la ventaja de ser una estimación insesgada. De hecho, implícitamente se toma este mecanismo cuando se ignora la no-respuesta. Es decir, al no hacer nada ante una situación de no-respuesta en una encuesta, implícitamente se hace uso de un mecanismo MCAR. En un subcapítulo siguiente se dan más detalles de este caso.

Un ejemplo de la utilización de un modelo MCAR para el caso de encuestas a personas, es cuando se supone que la probabilidad de responder de un individuo k es independiente de su edad, sexo, nivel socio económico, edad, filiación política, etc. (covariantes). Tampoco tiene relación con el diseño de la muestra (por ejemplo, independientes de los estratos, o los conglomerados), ni con el tipo

de preguntas contenidas dentro de la encuesta (es decir, preguntas sensitivas y no sensitivas son respondidas por igual). Dicha dispersión aleatoria y carente de patrones es lo que permite ignorar la no-respuesta y entonces considerar que se realizarán estimaciones no sesgadas utilizando sólo información del conjunto *r* (las unidades de observación que respondieron) sin modificarlos por ningún factor de ajuste.

2.2.2.2 Respuestas faltantes de forma aleatoria dadas variables covariantes, o respuesta ignorable (*Missing at Ramdom Given Covariates MAR por sus siglas en inglés*)

Este mecanismo de no-respuesta ocurre cuando ϕ_i depende de $\mathbf{x}_i$, pero no de y_i.

Matemáticamente:

$$P(R_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \phi_i$$

$$\text{pero } P(R_i = 1 \mid Y_{ij} = y_{ij}) = P(R_i = 1) = \phi_i$$

La dependencia de ϕ_i del vector de covariantes implica que hay un patrón común entre las unidades de observación que no contestaron, pero como es independiente de y_{ij}, se puede modelar para corregir el sesgo ocasionado por la no-respuesta.

Es estos casos se pueden utilizar modelos que ayuden a disminuir el sesgo, (más no a eliminarlo), porque la probabilidad de respuesta ϕ_i es desconocida y sólo se

puede construir una estimación de la misma $\hat{\phi}_i$. Así, entre mejor sea la estimación, mayor será la reducción del sesgo por no-respuesta.

Un ejemplo para este caso, sería cuando la probabilidad de responder depende del sexo (variable covariante), pero no hay dependencia estadística con el diseño de la muestra, ni las respuestas. Así, este tipo de no-respuesta se puede ignorar *después* de que por medio de un modelo se efectúe una corrección. Este modelo se utilizará en el capítulo 5.

2.2.2.3 No-respuesta no ignorable

Este tipo de no-respuesta tiene como mecanismo que la probabilidad de responder ϕ_i depende del valor que pueda tomar la variable de respuesta *pero* que no es posible explicarla completamente por medio del vector de covariantes x_i.

Matemáticamente:

$$P(R_i = 1 \mid Y_i = y_i) = \phi_i .$$

$$\text{pero } P(R_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \phi_i .$$

Cuando se tiene sospecha que puede ocurrir no-respuesta no ignorable, es conveniente implementar diversas estrategias para reducir la no-respuesta en campo y en diseño conceptual al máximo a fin de reducir el sesgo, puesto que los modelos difícilmente podrán corregir el sesgo de forma significativa.

Un ejemplo de este caso, sería en una encuesta electoral cuando la gente que favorece a un candidato tiende a no responder, pero la edad, el sexo, u otra variable disponible no se puede utilizar para explicar la no-respuesta. Es decir, no hay un patrón identificable que permita su corrección.

2.3 Estrategias para prevenir la no-respuesta

Partiendo que se ha seleccionado una muestra aleatoria en un escritorio viene la pregunta ¿qué medidas se pueden tomar para reducir al máximo la no-respuesta? A continuación se enuncian diversas propuestas hechas por diversos autores.

2.3.1 Diseño de experimentos y Control de calidad

Existen dos metodologías ampliamente aceptadas y que son utilizadas en diversos campos de la industria y actividad humana: Diseño de experimentos y Control de Calidad.

Por medio del diseño de experimentos se conocen técnicas que permiten identificar las causas por las que una encuesta en particular se ha obtenido no-respuesta. Además, es posible realizar diseños factoriales para evaluar la interacción entre diversos factores experimentales para poder mejorar el nivel de respuesta.

Respecto al Control de Calidad. Deming (1986) propone el uso de métodos estadísticos para el control de la calidad de los productos o servicios finales de

FALTA

PAGINA

39

Los principales factores que afectan a los entrevistadores son:

- su perfil
- motivación
- capacitación
- carga de trabajo
- método de recolección

Asimismo, las variaciones por tipo de encuesta son:

- demográfica
- económica
- o socio-económica.

Por último, algunos de las variables que tienen impacto sobre el entrevistado para que responda la encuesta son:

- disponibilidad
- la cantidad de preguntas que se la hagan
- su motivación para responder
- cuente con la información que se le pregunta

Se deben tomar todas las medidas pertinentes en cada una de las variables propuestas por Platek como por ejemplo: capacitando adecuadamente a los encuestadores, o en la presentación del encuestador se expongan razones que motiven al entrevistado a responder la encuesta al decirle que su opinión es importante, etc.

Una vez que se haya seleccionado una muestra aleatoria en el escritorio, sólo se tiene una pequeña parte del trabajo, puesto que se debe seleccionar, instruir, capacitar detalladamente, proporcionar estrategias de presentación y convencimiento a los trabajadores de campo. Haciéndoles ver que para que su trabajo dé resultados se deben seguir al pie de la letra todas las especificaciones de selección y medición. Igualmente se debe tener cuidado de no poner una carga excesiva en el entrevistado al hacerle una encuesta muy larga y difícil de contestar. Se debe partir siempre del hecho que quien responde una encuesta en esencia está haciendo un favor.

2.4 Efectos de ignorar la no-respuesta

La no-respuesta casi siempre tiene un efecto importante en una encuesta porque las personas que se rehusan contestar o no fueron accesibles suelen ser diferentes para ciertas variables de las personas que están dispuestas a participar o que si fueron accesibles. Es decir, para las variables donde exista una correlación entre la o las variables de interés y la accesibilidad de las unidades de observación (donde accesibilidad se refiere a elementos que fueron contactados pero no se pudo realizar la medición o a elementos que no se pudieron contactar).

Esto se ve más claro si se toman los tres mecanismos de no-respuesta descritos. Sólo en el caso de se tenga no-respuesta de acuerdo a un mecanismo **MCAR**, ignorar la no-respuesta acarrea sesgo en las estimaciones.

Desafortunadamente es la práctica común ignorar la no-respuesta, bajo el supuesto que las unidades no observadas se asemejan a las unidades si medidas (MCAR) o por simple negligencia.

Capítulo III:

Muestreo en dos fases enfocado a ajustar por no-respuesta

3.1 Teoría de muestreo en dos fases

3.1.1 Introducción

El muestreo en dos fases fue propuesto en 1938 por Neyman, y se usa para dos propósitos diferentes.

El primer propósito, en el caso cuando no se cuenta con un marco de muestreo o se cuenta con uno sin variables auxilares y la actualización de dichas variables es muy costoso, se toma una muestra para medir n variables auxilares que en principio su medición es poco costosa pero correlacionada con las variables reales de interés y_i pero su obtención es más costosa (esto es llamado la primera fase), y con el uso de esta información auxiliar se toma una submuestra obtenida directamente de la muestra de la primera fase (segunda fase) y entonces se mide la variable y_i más costosa, pero con el uso de la información auxilar se procede a realizar una estimación de razón o regresión para mejorar la precisión significativamente. Así, esta metodologia ayuda aumentar la precisión de la estimación o visto de otra forma ayuda a ahorrar tiempo y costos.

El segundo uso que se le puede dar al muestreo en dos fases es evitar estimaciones sesgadas previendo un problema de no-respuesta. La idea es seleccionar una muestra s con un diseño arbitrario $P(\cdot)$ en la primera fase de

tamaño *n*. Se realiza el levantamiento y el resultado segrega la muestra en dos estratos. El estrato de las personas que respondieron y el estrato de aquellos que no respondieron. Entonces, de aquellos que no respondieron se selecciona una submuestra y se hace todo el esfuerzo para obtener a *todos* los elementos de dicha submuestra. La gran ventaja de esta metodología es que es posible construir estimaciones insesgadas pese a que se tenga no-respuesta.

3.1.2 Breve resumen del desarrollo histórico del muestreo en dos fases para no-respuesta

Como se ha mencionado Neyman (1938) fue quién propuso por primera vez el muestreo en dos fases. Por otra parte, Hansen y Hurwitz (1946) fueron quienes propusieron por primera vez el uso del muestreo en dos fases para tratar el problema de la no-respuesta. No obstante, es interesante mencionar que la propuesta de Hansen y Hurwitz está enfocada más bien como una estrategia para reducir el costo de una encuesta, al mandar como primera etapa una encuesta por correo, y de aquellos que no respondieron, seleccionar una submuestra y entonces realizar un levantamiento con encuestadores, que por su puesto es a un costo mucho más alto, pero la estrategia conjunta de realizar una encuesta por correo y con visitas a domicilios, tiene la ventaja de que ahorra costos.

Recientemente, Cochran (1977) explica la no-respuesta con un modelo determinista de una población partida en dos subpoblaciones. Cochran dice "En el estudio de la no-respuesta, es conveniente pensar que la población se divide en dos estratos. El primero consiste en todas las unidades se obtendrían medidas si

las unidades cayeran en muestra. El segundo estrato de las unidades de las que no se obtendrían mediciones". Bajo este modelo, las unidades en el estrato de respuesta, responderían con probabilidad uno, mientras que las otras unidades responderían con probabilidad cero. Este modelo implica que la información sólo puede ser obtenida del estrato de los respondientes y necesariamente de ahí extrapolada a toda la población. Sin embargo, Cochran admite la limitación de este modelo determinista y escribe (1977): "Esta división en dos estratos diferentes es, por su puesto, una sobre simplificación. El azar juega una parte en determinar si una unidad es localizada y medida dado un número de intentos. En una especificación más completa del problema, nosotros asignaríamos a cada unidad una probabilidad que represente la oportunidad de que la unidad fuera medida por algún método de campo, si ésta cayera en muestra". Esta dirección señalada por Cochran, es retomada más recientemente por diversos estudiosos y se considera que el comportamiento de la repuesta con un comportamiento estocástico y no determinista. Así, se toma la perspectiva de suponer modelos para tratar el problema de la no-respuesta.

3.1.3 Breve resumen de la Teoría general del muestreo en dos fases para no-respuesta

3.1.3.1 Elementos principales

A continuación se describirá los elementos principales del muestreo en dos fases según el tratamiento general que hacen del mismo Särndal y Swensson (1987) para la no-respuesta.

Sea una población finita $U = \{y_1, y_2, ..., y_N\}$ de tamaño *N*. Sea *y* la variable de estudio, y sea y_k el valor de la variable *y* para la k-ésima unidad. Se busca estimar el total de la población a partir de una muestra *r*, que fue obtenida por medio de dos fases de selección. Si $A \subseteq U$ es un conjunto de unidades que se observaron, Särndal y Swensson (1987) denotan por simplicidad $\sum_A y_k$ para $\sum_{k \in A} y_k$. Se permite un diseño general para cada una de las dos fases; es decir, las probabilidades de inclusión en cada fase son arbitrarias. Su notación para los diseño de muestra son los siguientes:

a) La muestra de la primera fase ($s \subset U$) de tamaño n_s, no necesariamente fijo, se extrae de un diseño denotado $P_a(\cdot)$, tal que $P_a(\cdot)$ es la probabilidad de selección de *s*. Las probabilidades de inclusión de primer y segundo orden se definen para la primera fase por:

$$\pi_{ak} = \sum_{k \in s} P_a(s), \qquad \pi_{akl} = \sum_{k,l \in s} P_a(s)$$

Donde $\pi_{akk} = \pi_{ak}$. Sea la covarianza de las variable indicadoras de inclusión en muestra dos elementos *k* y *l* denotada como $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$. Se hace el supuesto que $\pi_{ak} > 0$ para toda *k*, y para la estimación de la varianza $\pi_{akl} > 0$ para toda $k = l \in s$.

b) Dada la muestra *s*, la muestra de la segunda fase r ($r \subset s$) de tamaño m_s, no necesariamente fija, es seleccionada de acuerdo con el diseño de muestra $P(\cdot \mid s)$.

tal que $P(r \mid s)$ es la probabilidad condicional de escoger r. Las probabilidades de inclusión de primer y segundo orden dada s se definen:

$$\pi_{k,s} = \sum_{k \in r} P(r \mid s), \qquad \pi_{kl,s} = \sum_{k,l \in r} P(r \mid s)$$

Donde $\pi_{kk,s} = \pi_{k,s}$. Sea igualmente la covarianza $\Delta_{kl,s} = \pi_{kl,s} - \pi_{k,s}\pi_{l,s}$. También se supone que para cualquier muestra s, $\pi_{k,s} > 0$ para toda $k \in s$, y que en la estimación de la varianza $\pi_{kl,s} > 0$ para toda $k \neq l \in s$.

Una vez que se han encontrado las probabilidades de inclusión para la primera y segunda fase y demás elementos mencionados. Se procede a describir la estimación por medio una nueva probabilidad de inclusión que contempla a las probabilidades de las dos fases.

3.1.3.2 Estimación por medio de π^*- sumas expandidas

Se define para toda $k, l \in s$, y para cualquier s

$$\pi_k^* = \pi_{ak}\pi_{k,s} \qquad \pi_{kl}^* = \pi_{akl}\pi_{kl,s}$$

donde $\pi_{kk}^* = \pi_k^*$. Sea $\Delta_{kl}^* = \pi_{kl}^* - \pi_k^*\pi_l^*$. Se definen también los valores y expandidos y los valores Δ por

$$y_k^* = y_k / \pi_{ak}, \quad f_k^* = y_k / \pi_{k,s} = y_k / \pi_k^*, \quad \Delta_{akl} = \Delta_{akl} / \pi_{akl}$$

Note que gorro hacia abajo indica la expansión que corresponde a la primera fase y doble gorro hacia abajo indica la doble expansión por la primera y segunda fases.

Notación: Si $A \subseteq U$ es un conjunto de unidades experimentales, $\sum\sum_A c_{kl}$ significa $\sum_{k \in A} \sum_{l \in A} c_{kl}$

El estimador básico del muestreo en fases, que es llamado estimador $\pi^* ES$ que es el ya definido π^* - sumas expandidas se describe en el resultado siguiente:

En un muestreo en dos fases, un estimador basado en un diseño insesgado del total de la población está dado por el estimado $\pi^* ES$.

$$\bar{t}_{\pi^*} = \sum_s \hat{y}_k = \sum_s y_k / \pi_k^*$$

La varianza poblacional dada por el diseño es:

$$V(\bar{t}_{\pi^*}) = \sum \sum_U \Delta_{kl} y_k y_l + E_a \left\{ \sum \sum_s \Delta_{kl} \hat{y}_k \hat{y}_l \right\}$$

Donde la $E_a(\cdot)$ denota la esperanza con respecto al diseño de muestreo de la fase uno.

Un estimador de la varianza basado en el diseño está dado por:

$$\hat{V}(\bar{t}_{\pi^*}) = \sum \sum_s \Delta_{kl}^* \hat{y}_k \hat{y}_l / \pi_{kl}^*$$

Este estimador $\bar{t}$ permite que se pueda construir un intervalo de confianza 100(1-α) con una distribución normal de forma habitual con la fórmula $\bar{t} \pm z_{1-\alpha/2} \{\hat{V}(\bar{t})\}^{\frac{1}{2}}$.

Estás ecuaciones son generales y aplicables a cualquier diseño de muestreo tanto para la primera fase como para la segunda

El estimador $\pi^{*}ES$, puede ser descrito como del tipo de estimadores de ponderación. Sin embargo, no es el único tipo de estimadores posibles para el caso de muestreo en dos fases. Si se cuenta con información auxiliar, es conveniente usar estimadores de regresión. Un tratamiento general de estimadores de regresión está dado por Särndal y Swensson (1987 paginas 284 - 289). Aquí sólo enunciaremos las características principales y panoramas bajo los cuales es posible utilizar este tipo de estimadores.

Los autores proponen diversas formas de los estimadores de regresión para los siguientes tres casos:

Caso 1: El valor de $x_k = (x_{1k}, \ldots, x_{qk})$ es un vector de información auxiliar disponible sólo para cada unidad de observación de la muestra *s* de la primera fase

Caso 2: El vector x_k está disponible para todas las *k* unidades de toda la población *U*.

Caso 3: Una combinación de los casos 1 y 2. El vector x_k está disponible para todos los miembros de la muestra, por otra parte otra información auxiliar $z_k = (z_{1k}, \ldots, z_{qk})$ (tal vez más débil) es conocida para todos los elementos del universo.

Ante estos tres panoramas es posible construir estimadores de regresión de los totales y la estimación de la varianza. Es importante mencionar que dichos estimadores también hacen uso de los ponderadores $\pi^{*}ES$ y de un modelo de regresión.

Särndal y Swensson (1987) realizaron una simulación de Monter Carlo para comparar los estimadores $\pi^{*}ES$ (ponderación) y los de regresión. Se concluye, que

- Una ventaja de usar un estimador de regresión es que su varianza es menor de forma considerable.
- Los modelos de regresión permiten reducir el sesgo de forma importante, ante la presencia de la no-respuesta, por el uso de variables auxiliares aunque no sea posible identificar correctamente el mecanismo de no-respuesta.

3.2 Muestreo en dos fases para no-respuesta

Esta estrategia de muestreo aplicado para corregir la no-respuesta, fue propuesta por primera vez por Hansen y Hurwitz (1946). La idea es tomar una submuestra de los que no respondieron y *hacer todo el esfuerzo* para obtener respuestas de *todos* los elementos de dicha submuestra. Existen diversos esquemas. Se enunciará la técnica en términos generales. Es decir, un diseño arbitrario se utiliza para extraer la muestra inicial, e igualmente, la segunda etapa de submuestreo se realiza por medio de un diseño arbitrario.

Se limita la discusión al caso de estudio de una sola variable denotada por *y*. Con el objetivo de estimar el total. A continuación se describen los pasos:

1. En la primera etapa, una muestra representada por s_a de tamaño n_a es seleccionada de acuerdo a un diseño arbitrario $P_a(\cdot)$ con probabilidades de inclusión π_{ak} de primer orden y π_{akl} de segundo orden ambas positivas. Además, sea $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$, la covarianza de las variables aleatorias indicadoras de inclusión para los elementos *k* y *l*.

2. A pesar de los esfuerzos para obtener todas las respuestas y_k para todos los elementos de s_a, ocurre no-respuesta (es decir, de algunos elementos no se obtuvo información). Si se toma el supuesto que la respuesta tiene un comportamiento estocástico, entonces existe una función de distribución de respuesta (DR) que gobierna la partición de la muestra en dos categorías de la muestra original s_a en un subconjunto de elementos que respondieron, s_{a1}, de tamaño n_{a1}, y por otra parte, un subconjunto de no-respuesta s_{a2}, de tamaño n_{a2}, donde $n_a = n_{a1} + n_{a2}$. La afirmación de comportamiento estocástico implica que si diferentes muestras s_a fueran obtenidas de forma repetida, el número de elementos en cada categoría variaría entre una encuesta y la otra.

3. Una submuestra grande adecuada s_2 es seleccionada de s_{a2}, por un diseño $P(\cdot \mid s_{a2})$ con probabilidades de inclusión positivas denotadas $\pi_{k|s_{a2}}$ de primer

orden y $\pi_{kl|s_{a2}}$ de segundo orden. Sea la covarianza $\Delta_{kl|s_{a2}} = \pi_{kl|s_{a2}} - \pi_{k|s_{a2}}\pi_{l|s_{a2}}$. Y se toman todas las medidas necesarias y suficientes para tener la respuesta de *todos* los elementos de esta muestra.

El requerimiento de tener respuesta completa de la submuestra s_2, puede ser costoso, pero es la condición necesaria para obtener una estimación del total insesgada.

El conjunto de elementos para los cuales *y* es observada es $s = s_{a1} \cup s_{a2}$. Así, el total $t = \sum_U y_k$ se estima por medio de:

$$\hat{t} = \sum_s \tilde{y}_k = \sum_s y_k / \pi_k^*$$

Donde como $s = s_{a1} \cup s_2$ entonces

$$\pi_k^* = \begin{cases} \pi_{ak} & \text{Si } k \in s_{a1} \\ \pi_{ak}\pi_{k|s_{a2}} & \text{Si } k \in s_{a2} \end{cases}$$

Por lo tanto, este estimador también se puede escribir como:

$$\hat{t} = \sum_{s_{a1}} y_{ak} + \sum_{s_2} \tilde{y}_k$$

Donde $y_{ak} = y_k / \pi_{ak}$

Es decir, se lleva a cabo la expansión de los elementos de cada fase por separado, con su ponderador correspondiente.

Una propiedad importante de $\hat{t}$ es que es un estimador insesgado de *t*, para cualquier distribución del mecanismo de respuesta (DR). Es decir, no es necesario hacer supuesto o modelo alguno sobre cómo se distribuye la respuesta.

La varianza del estimador está dada por:

$$V(\hat{t}) = \sum\sum_{U} \Delta_{akl} \hat{y}_{ak} \hat{y}_{al} + E_{p_a} E_{DR} (\sum\sum_{s_2} \Delta_{kls_{a2}} \hat{y}_k \hat{y}_l \mid s_a)$$

Donde el E_{p_a} hace referencia al valor esperado con el diseño, por otra parte, E_{DR} es el valor esperado respecto a la distribución de respuesta. Así, para la expresión de la varianza el segundo componente no es explícito. Porque no se puede calcular la varianza poblacional si la distribución de la respuesta no es conocida.

Por otra parte, un estimador insesgado de la varianza:

$$\hat{V}(\hat{t}) = \sum\sum_{s} \frac{\Delta_{akl}}{\pi^{*}_{kl}} \hat{y}_{ak} \hat{y}_{al} + \sum\sum_{s_2} \frac{\Delta_{kls_{a2}}}{\pi_{kls_{a2}}} \hat{y}_k \hat{y}_l$$

Donde $s = s_{a1} \cup s_2$

$$\pi^{*}_{kl} = \begin{cases} \pi_{ak}\pi_{kls_{a2}} & \text{Si} k, l \in s_{a2} \\ \pi_{ak}\pi_{ks_{a2}} & \text{Si} k \in s_{a2}, l \in s_{a1} \\ \pi_{ak}\pi_{ls_{a2}} & \text{Si} k \in s_{a1}, l \in s_{a2} \\ \pi_{akl} & \text{Si} k, l \in s_{a1} \end{cases}$$

Observe que este estimador tiene como ventaja que permite su cálculo sin el conocimiento de la distribución de la respuesta.

Igualmente que en caso general se pueden construir estimadores de regresión si se cuenta con información auxiliar para mejorar la precisión de la estimación. Pero como el estimador es insesgado, el uso de un estimador de regresión no sería para corregir el sesgo.

3.3 Alcances y limitaciones del muestreo en dos fases para ajustar por no-respuesta

Los principales alcances del uso de muestreo en dos fases, son que como el diseño de la primera fase es arbitrario, no es necesario recurrir a una segunda fase si se obtienen niveles de no-respuesta muy pequeños. Y así, poder ahorrar recursos.

Para la segunda fase es conveniente tener designados recursos suficientes, porque para obtener estimadores insesgados es necesario obtener a *todos* los elementos de la segunda fase. Para realizar la estimación, es posible realizarla por medio de estimadores simples, de razón o de regresión.

Una limitación importante del muestreo en dos fases es que se parte del supuesto que con el destino de recursos suficiente, será posible obtener respuestas de una submuestra de las unidades de observación que en la primera fase no respondieron. Sin embargo, el obtener información de una unidad de observación, no sólo depende sólo de recursos. Por ejemplo, en el caso de que las unidades de observación fueran personas limitaciones físicas como enfermedad grave, discapacidad para hablar o escuchar, vejez avanzada son fuente de no-respuesta irremediable. Además, otra fuente son personas que se encuentren de viaje, o que existen impedimentos físicos para tener acceso (como ser de un nivel socioeconómico muy alto o muy bajo usualmente impone barreras físicas).

En el caso de que las unidades de observación sean unidades económicas como fabricas o comercios, es posible que haya una negativa a contestar, por razones de seguridad, defensa de intereses, confidencialidad, etc.

El segundo inconveniente del muestreo en dos fases es la posible pérdida de oportunidad en la entrega de resultados porque frecuentemente las encuestas están bajo una fecha límite de entrega, porque tanto realizar dos levantamientos, así como los múltiples esfuerzos inherentes a localizar todas las unidades de la segunda implica un costo en tiempo, que se puede reflejar en un retraso en la entrega.

Capítulo IV

Métodos de imputación

4.1 Marco teórico de la imputación

4.1.1 Introducción

La no-respuesta puntual o a preguntas específicas se pueden atribuir a tres causas en la mayoría de los casos:

- El entrevistador no hace la pregunta (se la salta por error) o no anota la respuesta; es decir, la información está faltante por causas atribuibles al propio entrevistador
- La unidad de observación se niega a responder la pregunta; comúnmente esto está asociado a preguntas que son llamadas sensitivas, por ejemplo, temas relacionados con los ingresos son de lo más común.
- La unidad de observación no puede proveer la información; la mayor parte de las veces se atribuye a olvido o confidencialidad de la información.

Por otra parte, los datos o respuestas puntuales de cada unidad de observación obtenidas deben ser validadas como parte de un proceso de control de calidad. El control se enfoca a dos rubros:

- La información debe ser consistente con otras preguntas; por ejemplo, si una mujer declara tener un hijo debe tener una edad mayor o igual a 12 años.

- Los datos captados pueden estar acotados por un límite superior e inferior para variables cuantitativas, y por otra parte, con una categoría válida para variables cualitativas.

Cuando un dato está dentro de uno de los dos casos arriba mostrados y es identificado en un proceso de control de calidad, la práctica común es borrar el dato que se considera erróneo para considerarlo no-respuesta puntual para la q-ésima pregunta y k-ésima unidad de observación.

La siguiente exposición toman como fuente principalmente a Lohr (1999) y a Särndal et al. (1992).

4.1.2 Marco teórico de la imputación

Supongamos que se tiene una encuesta con q variables de estudio $y_1, \ldots, y_j, \ldots, y_q$. Si tomamos la definición ya vista del conjunto de respuesta r_j como el conjunto de las unidades de observación que respondieron a la j-ésima pregunta; comúnmente se tiene la situación de que los q conjuntos generalmente son de tamaño diferente, es decir, el número de preguntas respondidas difieren entre sí. Por lo tanto, una aproximación posible ante este caso es construir estimaciones separadas de cada variable para cada caso. Sin embargo, esto pocas veces se lleva a cabo porque se multiplica el trabajo y resulta poco práctico.

La imputación se define como asignar valores $\tilde{y}_{jk}$ "artificiales" de acuerdo a una metodología a aquellos valores faltantes para aquellas unidades de observación *k* donde no todas las variables $\tilde{y}_{jk}$ para *j=1, ..., q* son faltantes.

Así, la imputación tiene como principal propósito construir una matriz rectangular completa de datos "limpios". Es decir, se busca que cada conjunto r_j se tenga el mismo número de elementos. Así cada unidad de observación y_k sólo puede estar en uno de dos casos:

- y_k tiene información en todas las *q* variables que le fueron medidas
- y_k no tiene información y por lo tanto forma parte de la no-respuesta por unidad de observación y no forma parte de los conjuntos de respuesta r_j.

Así, se tiene la situación que la no-respuesta puntual es corregida por alguno de los métodos descritos en los subcapítulos siguientes. Además, la no-respuesta de las unidades de observación es tratada con alguno de los métodos descritos en el capítulo 5.

Se considera *"Estadísticamente incorrecto"* tratar los datos que fueron imputados de forma igual a aquellos que son reales para efectos de estimación y otro tipo de análisis. Esto en mucho depende de la calidad de los datos imputados. Así, el error derivado de aplicar métodos estándar de estimación pierde importancia si el valor $\tilde{y}_{jk}$ es cercano al valor real faltante desconocido, pero como los datos

imputados pueden ser un substituto pobre en calidad, cuando están basados en información proveniente de una muestra porque la información de la muestra está sujeta en si misma a un valor esperado y varianza (variabilidad). Por lo tanto, el sesgo y un incremento en la magnitud de la varianza usualmente son consecuencia de la imputación de datos faltantes.

Siempre se deben marcar con una bandera en las bases de datos aquellos datos que fueron imputados a fin de identificarlos y poder distinguirlos de los reales. Incluso, en algunos países está prohibido por ley la utilización de valores imputados, aunque estos sean marcados.

En los últimos años, se han realizado estudios tanto teóricos como aplicados donde se ha demostrado que las técnicas de imputación tienen debilidades considerables.

4.2 Imputación deductiva

Método se refiere aquellos casos, aunque poco común en la práctica, donde un valor imputado $y_{i} = y_{i}$ por medio de una relación lógica. La deducción puede estar basada en otras respuestas obtenidas en el cuestionario para la misma unidad de observación.

Este tipo de imputación puede ser utilizada con éxito en estudios longitudinales (a través del tiempo), asociando los valores que la variables van tomado y en caso de tener un valor faltante pero lógico poder hacer la imputación.

4.3 Imputación por valor medio de celda

Las unidades de observación que respondieron a la encuesta, son divididas en c clases o estratos. Entones, para la variable y_j el valor medio de cada celda es imputado o sustituido para cada no-respuesta puntual dentro de la celda. Evidentemente, para la aplicación de este método es necesario que no haya no-respuesta en la variable que se usa para la creación de las clases o estratos.

Este método de imputación parte del supuesto de que la no-respuesta tiene un mecanismo MAR dentro de cada celda o estrato.

Para la construcción de los estratos, se debe tener conocimiento que se está utilizando una variable tal que tenga una relación con la variable de interés que permita hacer la imputación.

Una desventaja importante del uso de esta metodología, es que el uso de valores medios, disminuye la variabilidad, y al estimar la varianza de las estimaciones de interés, se puede llegar a hacer una subestimación importante de la varianza si un número grande de datos fueron imputados bajo este método.

4.4 Imputación *Hot-deck*

La imputación *Hot-deck* ayuda a crear una variabilidad más auténtica al contrario de lo que pasa con la imputación por valor medio de la celda. Así, en la imputación

hot-deck, las respuestas faltantes son reemplazadas por valores reales de la misma encuesta en cuestión.

El término *hot-deck* (paquete de tarjetas calientes) proviene de los días cuando las computadoras usaban tarjetas perforadas como fuente de datos. Así, el contenedor de las tarjetas las "calentaba" después de su generación, y entonces se hacía la utilización de los mismos datos.

En función del método de selección existen variantes de este método que se enuncian a continuación.

4.4.1 Imputación aleatoria *Hot-deck*

El valor faltante de la variable *j* es reemplazado por un valor y_{ja} tomado de uno que haya respondido a la encuesta, donde el dato a imputar es seleccionado aleatoriamente del conjunto r_j de respuestas.

Este método provee un conjunto de datos para la pregunta *j* con una variabilidad más cercana a la realidad, a pesar de ello, no se pueden utilizar las técnicas estándar para calcular la varianza de las estimaciones.

4.4.4 Imputación secuencial *Hot-deck*

Bajo este procedimiento, dado un valor faltante, se realiza la imputación tomado como donador al dato previo leído en la computadora en la base de datos. Esto se

utiliza partiendo del supuesto que las unidades adyacentes dentro de un mismo grupo tienden a ser más similares a los datos que se obtendrían si se obtuvieran de forma aleatoria. Este supuesto es muy difícil de validar, pero si el investigador tiene elementos que sustenten su hipótesis, entonces se debe utilizar.

4.4.5 Imputación del vecino más cercano *Hot-deck*

Se define una medida de distancia entre las unidades de observación (con el uso de variables auxiliares donde todas las unidades tengan respuestas), y se imputa tomado las variables de aquella unidad de observación que se encuentre más cercana de acuerdo a la función de distancia.

Por ejemplo, si se definiera a edad e ingreso como la medida de distancia entre dos unidades, dado una no-respuesta puntual, se busca otra unidad que tenga la edad e ingreso más cercano y se imputa la respuesta en cuestión.

4.5 Imputación *cold-deck*

Imputación *cold-deck* (tarjetas frías) consiste en utilizar la información de encuestas efectuadas en el pasado, u otra información histórica. Muy poca teoría existe respecto a este método y como la imputación *hot-deck* las estimaciones resultantes se encuentran sesgadas.

Se pueden utilizar las mismas técnicas descritas para el caso de imputación *hot-deck* para seleccionar el dato a ser imputado.

4.6 Imputación por regresión

La imputación por regresión utiliza la relación entre las variables, mientras que los métodos de imputación *hot-deck* y *cold-deck* no lo hacen. Se parte del supuesto que para un dato faltante existen variables predictoras ya sean variables auxilares parte del marco de muestreo disponibles previas al levantamiento, o a variables obtenidas en el levantamiento. Por lo tanto, la ecuación de regresión resultante más del término de error son usados para producir las imputaciones.

El uso de estos modelos parten del supuesto que el mecanismo de no-respuesta es MAR.

4.7 Imputación múltiple

En el caso de la imputación múltiple, cada valor faltante se imputa $m \geq 2$ veces, donde cada imputación se hace cada vez mediante un mismo modelo estocástico. Se crean m bases de datos rectangulares y se analiza cada uno como si no se hubiera realizado imputación alguna. El enfoque de la técnica es que la variabilidad de los diversos resultados proveen al analista una medida de la variabilidad adicional debida a la imputación.

Por otra parte, el uso de imputación múltiple usando diversos modelos de no-respuesta, dan una idea de la sensibilidad de los resultados debidos al uso de los modelos particulares de no-respuesta utilizados.

Rubin (1987 y 1996) propone la metodología para implementar la imputación múltiple.

4.8 Alcances y limitaciones de los métodos de imputación

La principal ventaja del uso de la imputación de datos es que se produce una base de datos que facilita el proceso de estimación y análisis de una encuesta. Por lo tanto, su uso es bastante común en la práctica. No obstante, literatura sobre las propiedades teóricas de los diversos métodos descritos en este capítulo del presente trabajo es escasa. Ford (1983) expone sobre los métodos Hot-deck::

"Dado el hecho de que cada uno de los estudios empíricos está limitado a la investigación de una muestra en particular, una generalización amplia de los resultados es difícil de realizar. Estos estudios soportan, sin embargo, las conclusiones teóricas de que los errores estándar de las estimaciones con información que se imputó con algún método del tipo hot-deck, están subestimados dado que los cálculos suponen que la información originalmente estaba completa. También se indica, que posiblemente no haya mucha mejora en el error cuadrático medio de un estimador *hot-deck*, en comparación con un estimador donde simplemente se omite la información faltante, a menos que haya información auxiliar que se encuentre altamente correlacionada con los datos de la encuesta. Los métodos no muestran de forma consistente que alguno sea mejor que otro".

Otra desventaja del método es que a menos que se tenga un cuidadoso control se puede llegar a perder la distinción entre los valores reales y los imputados. Así, en caso de que haya una duda con los datos, no es posible determinar si se trata de un dato real u artificial creado por un método de imputación.

Rao (1996) y Fay (1996) han propuesto métodos para estimar la varianza de un estimador después de realizar la imputación. En la práctica común se calcula la varianza usando las fórmulas correspondientes al diseño en cuestión como si no hubiera ocurrido imputación. Los autores mencionados proponen el uso del método Jackknife para la estimación de la varianza dada la presencia de datos imputados.

Capítulo V

Métodos de ajuste por ponderación

5.1 Marco teórico de la corrección por ponderación

5.1.1 Introducción

El muestreo en dos fases es una metodología con la que se corrige una situación de no-respuesta y tiene como principal ventaja que tiene como resultado la obtención de estimaciones insesgadas. Sin embargo, si no se recurre a un muestreo en dos fases, y se realiza una encuesta bajo un diseño arbitrario, es casi seguro que haya cierto nivel de no-respuesta.

Los ajustes por ponderación para corregir la no-respuesta hacen uso del modelo de *Respuestas faltantes de forma aleatoria dadas variables covariantes, o respuesta ignorable (Missing at Ramdom Given Covariates MAR).*

Se expondrán los dos principales variaciones de esta metodología: Ajuste por clase, y posestratificación. Si se hace una construcción adecuada de estos modelos y los datos se comportan bajo los supuestos de los mismos, se tiene la desventaja de que las estimaciones resultantes son sesgadas pero en una medida mucho menor al sesgo al que se incurriría si no se hiciera 'nada' (donde no hacer nada implícitamente se asume el modelo MCAR).

La siguiente discusión está limitada al caso donde se tiene no-respuesta de la unidad.

5.1.2 Marco teórico de los métodos por ponderación

Para los métodos por ponderación se toma el modelo MAR. Es decir, se hace le supuesto que la probabilidad de responder está asociada a ciertas características de la población conocidas, que fueron medidas en la muestra. Así, el método consiste en formar *H* estratos en función de alguna variable auxiliar como sexo, edad, nivel socioeconómico etc. para el caso de poblaciones humanas, y se repondera dentro de cada estrato utilizando una estimación de la probabilidad de respuesta ϕ_k. La elección de la variable auxiliar para la formación de estratos se hace bajo el supuesto de que para las variables observadas en cada estrato se asemejan entre sí lo suficiente para *representar* a los que no respondieron.

Sea π_k la probabilidad de inclusión de primer orden de la k-ésima unidad. Se define como el ponderador de una unidad de observación a $w_k = \frac{1}{\pi_k}$ (el recíproco de la probabilidad de selección). Una forma de interpretar el significado de un ponderador, es que éste indica al número de unidades de la población objetivo que representa la k-ésima unidad en muestra. Así usando este método los ponderadores son modificados para ajustar la estimación de un parámetro por no-respuesta.

Sea R_k una variable aleatoria indicadora de si una unidad *k* de observación respondió. Así

$$P(R_k = 1) = \phi_k$$

Es la probabilidad de que la *k* unidad responda a la encuesta.

Por otra parte, sea Z_i una variable aleatoria indicadora de si la k-ésima unidad está en muestra. Por lo tanto, se sabe:

$$P(Z_k = 1) = \pi_k$$

El primer supuesto del modelo se hace al suponer a las variables aleatorias R_k y Z_k son independientes entre sí. Por lo tanto, la probabilidad de que la k-ésima unidad sea seleccionada y efectivamente responda es $\pi_k \phi_k$, siendo el valor del ponderador w modificado por la no-respuesta $w_k = \frac{1}{\pi_k \phi_k}$. Sin embargo, la probabilidad de responder ϕ_k es desconocida porque se ignora la distribución de R_k. Así, tomando el supuesto que se tiene en una encuesta en particular respuesta MAR se puede realizar la *estimación* de dicha probabilidad $\hat{\phi}_k$ y así obtener el ponderador que ajusta por no-respuesta $\hat{w} = \frac{1}{\pi_k \hat{\phi}_k}$.

Los ponderadores están asociados a una partición de la muestra (clases), en función de las diversas categorías que puede tomar una variable covariante como lo son sexo, edad (construyendo agrupaciones), etc. Así, la estimación de la

probabilidad de responder $\hat{\phi}_k$ se supone igual para todos los miembros de cada clase. Misma que será denotada por $\hat{\phi}_c$ para cada *k* de la clase *c*.

5.2 Ajuste ponderado por clase

Una vez formadas las clases, se realiza una subpartción dentro de cada clase en dos categorías: unidades de observación que respondieron y unidades de observación que no respondieron. Sea r_c el número de unidades que respondieron en la clase c, y sea n_c el número de unidades en la muestra seleccionada dentro de la misma clase.

Bajo este esquema se estima la probabilidad de respuesta ϕ_c dentro de cada clase usando la siguiente fórmula:

$$\hat{\phi}_c = \frac{\sum_{k=1}^{r_c} w_k}{\sum_{k=1}^{n_c} w_k}$$

Es decir, se estima la probabilidad de responder como el cociente de los ponderadores de las unidades que respondieron entre la totalidad de unidades de la muestra seleccionada dentro de la clase.

Una vez que se ha estimado la probabilidad de responder y cada elemento de la muestra tiene estimada su probabilidad en función de la clase o estrato al que fue asignado, se procede a modificar los ponderadores de cada unidad que sí respondió, multiplicándolo por el inverso de la probabilidad de responder. Es decir,

$$\hat{w}_k = \frac{1}{\pi_k \hat{\phi}_c}$$

Por lo tanto, el estimador del total bajo este modelo es:

$$\hat{t} = \sum_{k \in s} \hat{w}_k y_k$$

La estimación del inverso de la probabilidad de responder ϕ_k debe cumplir con las siguientes restricciones.

1. Si $\frac{1}{\hat{\phi}_c} > 2$ indica que en la celda en particular hay más individuos que no respondieron a la encuesta que los que respondieron, teniendo como principal consecuencia un incremento en la varianza.

2. Dentro de cada celda, debe haber al menos 30 observaciones (respuestas), para poder utilizar el teorema del límite central.

En caso de que alguna de las dos condiciones o ambas no se cumplan, se deben unir celdas adyacentes para obtener un factor de corrección menor a 2 y que haya más de 30 observaciones.

Para poder desarrollar las fórmulas de la varianza y su estimación es necesario involucrarse con un poco de teoría. Este modelo se puede ver como un muestreo en dos fases donde los elementos de la segunda fase forman parte de un muestreo Bernulli estratificado. Así, se omite el desarrollo de la teoria estando para referencia en Särndal et al. (1992 pagina 581).

5.3 Posestratificación

Esta metodología es muy semejante a la metodología de ajuste ponderado por clase. La diferencia radica en que los universos de las clases son conocidos por una fuente de datos secundarios como por ejemplo el último censo de población y vivienda. Es decir, dado un universo conocido y la suma de ponderadores de las unidades de observación que respondieron, se puede calcular un factor de corrección al ponderador del diseño original por no-respuesta.

De forma general, se realiza una posestratifiación de la muestra en *H* clases, tal que $s = s_1 \cup s_2 \cup ... \cup s_H$. son conjuntos mutuamente excluyentes, donde se conocen por fuentes externas (otro ejemplo sería un muestreo previo que se haya hecho para estimar dominios) $N_1, ... N_h$ los universos de cada clase. Por ejemplo, el número de hombres y mujeres de un dominio de estudio. Así, se va a estimar un factor de ajuste f_i para cada elemento de la muestra que haya respondido, donde:

$$f_i = \frac{N_i}{\text{Suma de todo los pondadores de las unidades que respodieron del estrato}}$$

Por lo tanto,

$$\tilde{w}_i = f_i w_i = f_i \left(\frac{1}{\pi_i}\right)$$

Por lo tanto para la estimación del total se vuelve a utilizar la fórmula:

$$\hat{t} = \sum_{k \in s} \tilde{w}_k y_k$$

El uso de la posestratificación tiene los siguientes tres supuestos:

1. Dentro de cada posestrato, cada unidad seleccionada como parte de la muestra tiene la misma probabilidad de responder. Es decir, la probabilidad de responder ϕ_i^* es la misma para todos los elementos que pertenezcan a una clase.
2. El hecho de que una unida responda o no lo haga, es independiente del comportamiento de las demás unidades de observación dentro de la misma clase.
3. Las unidades que no respondieron dentro de un estrato son semejantes a aquellos que si respondieron.

De hecho, estos supuestos son los mismos al método de ajuste ponderado por clase ya descrito. E igualmente, el factor de ajuste no debe ser mayor a 2 porque si este fuera el caso es indicador de que se tiene menos de la mitad de unidades de observación que respondieron que el de las seleccionadas en muestra. Además, se deben tener al menos 30 unidades que respondieron por estrato para que el teorema de límite central pueda ser utilizado.

5.4 Ajuste de rastrillo

El ajuste por rastrillo es una extensión del ajuste por posestratificación. Se aplica cuando se conocen los universos de dos variables que se cree están relacionadas con la respuesta y se crea una tabla cruzada del siguiente tipo:

W'_{11}	...	...	...	N'_1
...	...	...	...	N'_2
...	...	...	...	...
...	...	...	W'_{lh}	N'_h
N''_1	N''_2	...	N''_l	

Donde cada W'_{ij} es la suma de los ponderadores de las unidades de observación que corresponden a cada celda que respondieron la encuesta en el i-ésimo renglón y la j-ésima columna. La correspondencia está definida por el cruce de las características de las poblaciones. Por ejemplo, alguna categoría de género por algún rango de edad.

La técnica consiste en aplicar un algoritmo iterativo de factores de ajuste f_{ij} a las W'_{ij} de forma tal que de obtenga una W''_{ij} tal que la suma de los renglones y las columnas cuadren con los marginales. Una vez obtenido, se crea un factor de corrección $C_{ij} = \frac{W''_{ij}}{W'_{ij}}$, que debe ser multiplicado a los ponderadores de las observaciones dentro de cada celda, es decir, $\tilde{w}_k = C_{ij} w_k$, si w_k está en la celda *ij*.

Algortimo de ajuste del rastrillo

El siguiente algortimo está descritpo por Holt y Elliot (1991) y por Oh y Scheuren (1983) para ser aplicados al ajuste por no-respuesta.

Sea U un conjunto de elementos de un universo que es conocido por alguna fuente confiable (censo de población y vivienda, o censo económico, etc.) que puede ser partido en h conjuntos mutuamente excluyentes A_i tales que $U = A_1 \cup A_2 \cup ... \cup A_h$. Un ejemplo, sería una población humana, que se puede partir por genero. Por lo tanto, hombres y mujeres es la partición de dicha población. Por otra parte, supóngase que se tiene otra partición del mismo universo dada por *l* conjuntos B_i, tal que $U = B_1 \cup B_2 \cup ... \cup B_l$. Por ejemplo, otra partición de una población humana sería la edad, o grupos de edad. Por último, se denota la cardinalidad de los conjuntos A_i por $N_i^{\cdot}$, y por otra parte B_i por $N_i^{\cdot}$. También, sea la cardinalidad de U denotada por N y sea $W^{\cdot\cdot} = \sum_i \sum_j W_{ij}$, es decir, el universo total estimado por los ponderadores.

Se puede crear una tabla cruzada, donde en la entrada o cruce de cada celda, se asignan la suma de los ponderadores $W_i^{\cdot}$ de las unidades de observación que respondieron la encuesta de cada entrada. Si *c*, es un par ordenado que hace referencia a identificación de la celda en cuestión, es decir, *c=(i, j)*. La suma de los ponderadores de cada celda está denotada por $W_i^{\cdot} = W_{i,j}^{\cdot} = \sum_{k \in c} w_k$ sólo para las unidades *k-ésimas*, que respondieron la encuesta.

W_{11}	...	...	...	N_1^*
...	...	...	...	N_2^*
...	...	...	...	...
...	...	...	$W_{I,h}$	N_h^*
N_1^{**}	N_2^{**}	...	N_I^{**}	N

Partiendo de que cada $W_{i,j}$ se puede interpretar como la estimación del número de unidades en el universo que representan los elementos en muestra. Así, la suma de las $W_{i,j}$ sobre un renglón debería ser igual a su correspondiente N_i^*. Igualmente, la suma de las $W_{i,j}$ pero sobre una columna fija debería corresponder a N_j^{**}. Pero, en el caso de no-respuesta, esto no se cumple. El algoritmo tiene como fin realizar un número finito de iteraciones para encontrar un factor f_{ij} que corrige cada $W_{i,j}$ para la suma corresponda a sus marginales.

5.4.1 Algoritmo del rastrillo

Paso cero

$$f_{i,j}^{(0)} = \frac{N}{W}$$

Multiplicando a cada $W_{i,j}$ por $f_{i,j}^{(0)}$ para crear $W_{i,j}^{(0)} = W_{i,j} \times f_{i,j}^{(0)}$

Para cada renglón sumar

$$\sum_j W_{i,j}^{(0)} = N_{i,\cdot}^{*(0)}$$

Para cada columna sumar

$$\sum_i W_{i,j}^{(0)} = N_{\cdot,j}^{**(0)}$$

Paso t, para t ≥ 1

Crear un factor de corrección para cada renglón i:

$$f_{i.}^{(t)} = \frac{N_{i.}^{(t-1)}}{\sum_{j} W_{i,j}^{(t-1)}}$$

Aplicar el factor de ajuste de cada renglón i

$$W_{i,j}^{(t-1)} \times f_{i.}^{(t)} = W_{i,j}^{(t)}$$

Se procede a sumar los nuevos pesos de cada entrada para obtener el nuevo marginal de los renglones.

$$\sum_{j} W_{i,j}^{(t)} = N_{(i.)}^{(t)}$$

Paso t+1

Ahora se crea un factor de ajuste para cada columna j

$$f_{.j}^{(t+1)} = \frac{N_{.j}^{(t)}}{\sum_{i} W_{i,j}^{(t)}}$$

$$W_{i,j}^{(t)} \times f_{.j}^{(t+1)} = W_{i,j}^{(t+1)}$$

Se procede a sumar, igualmente, los pesos por renglón para obtener los nuevos marginales por columna.

$$\sum_{i} W_{i,j}^{(t+1)} = N_{(.j)}^{(t+1)}$$

La iteración de los pasos t, y t+1 convergen, y se debe detener cuando se alcanza una precisión de centésimas.

Una vez que se han obtenido $W_{i,j}^{(t+h)}$ tal que converjan se construyen factores de ajuste para cada celda $fa_{i,j} = \frac{W_{i,j}^{t+h}}{W_{i,j}^{t}}$. Dicho factor debe multiplicar a cada ponderador w_k para elemento de la muestra según la celda que le corresponda.

Un ciclo completo del algoritmo es una iteración completa de los pasos 0, t y t+1,

5.4.2 Ejemplo

Suponga que se tiene una población humana ficticia de 100 individuos en el universo. Esta se puede partir por sexo (hombre y mujer), asimismo por dos categorías de nivel socio-económico (bajo, y medio-alto) bajo una fuente fidedigna. A continuación se muestra una tabla cruzada.

Incial	**Nivel socio-económico**		
Género	Bajo	Medio-alto	
Hombre	17.6	24	60
Mujer	25.6	16	40
	65	35	100

En el interior de la tabla se observa la suma de los ponderadores de una muestra donde hubo no-respuesta. Esta suma de las entradas es de 83.2, cuando debería ser de 100.

t=0	**Nivel socio-económico**		
Género	Bajo	Medio-alto	
Hombre	21.15	28.85	50
Mujer	30.77	19.23	50
	51.92	48.08	100

t=1	**Nivel socio-económico**		
Género	Bajo	Medio-alto	
Hombre	25.38	34.62	60
Mujer	24.62	15.38	40
	50	50	100

t=2	**Nivel socio-económico**		
Género	Bajo	Medio-alto	
Hombre	33.00	24.23	57.23
Mujer	32.00	10.77	42.77
	65	35	100

Después de 6 iteraciones el algoritmo convergió según se muestra en la siguiente tabla:

t=6	Nivel socio-económico		
Género	Bajo	Medio-alto	
Hombre	34.91	25.09	60.00
Mujer	30.09	9.91	40.00
	65.00	35.00	100

Luego entonces, los factores de ajuste a aplicar son:

fa	Nivel socio-económico	
Género	Bajo	Medio-alto
Hombre	1.983	1.045
Mujer	1.175	0.619

Este modelo opera bajo los siguientes supuestos:

1. Dentro de cada subpoblación, o entrada de la tabla cruzada, las respuestas son generadas por un muestreo Bernulli, e independiente con probabilidad de respuesta $\phi_{ij} > 0$.

2. Los mecanismos de respuesta son independientes entre cada subpoblación.

3. Las probabilidades de respuesta ϕ_{ij} tienen una estructura tales que éstas sólo están determinadas por la fila y columna a la que la unidad cae de forma aleatoria.

Por último, la única condición que se debe cumplir para que el algoritmo converja es que todas las celdas deben tener un valor mayor a cero.

5.5 Método Politz-Simmons

5.5.1 Introducción

En 1949, Politz y Simmons propusieron una metodología que elimina la necesidad de insistir cuando no se encuentra a las unidades experimentales. El diseño de la metodología está diseñado para poblaciones humanas donde la unidad en muestra se busca en su vivienda. Se diseñó esta metodología en un marco de proyectos de investigación de mercado.

5.5.2 Descripción general del método

Muchos individuos no están en casa al momento que un encuestador los busca en su casa. Estos casos con catalogados como "No localizados en casa". Se ha encontrado que la proporción de entrevistas en este caso puede oscilar de 30 a 60% según la experiencia los autores en los años 40 en EUA, así se propone el siguiente método.

La forma más simple de atacar este problema sería, realizar otro intento de visita, una y otra vez, en las viviendas hasta que se obtenga una respuesta. Esta estrategia, tiene como consecuencia un incremento en los costos de la operación, así como retraso del proyecto a un punto tal que realizarlo puede ser sumamente costoso, y por ende insostenible.

Además, esta metodología se enfoca a encuestas donde el interés primario es localizar a un individuo específico de la vivienda. Como por ejemplo, el jefe del hogar (porque hay encuestas que sólo un individuo con esta jerarquía dentro de la

vivienda cuenta con el perfil necesario para responder la encuesta). Por ejemplo, cuando se tiene interés el consumo de ciertos bienes duraderos, etc. En lo subsecuente se le denominará "individuo objetivo".

Si se insiste en varias ocasiones para localizar al individuo objetivo, que en la primer visita no fue localizado, se hace bajo el supuesto de que esta persona estará disponible en alguna otra hora del día en la vivienda, *cuando el personal de campo está en posibilidad de realizar la visita*. Así, individuos objetivo que sólo están en sus viviendas de 10:00 p.m. a 8:00 a.m. prácticamente están fuera de toda encuesta. Pero dejando fuera estos casos extremos, se hace el supuesto que aquellas personas que no se localizaron en la primer visita, pero se pudieron encontrar en la segunda visita, o la tercera, o la cuarta, o la quinta o como última la sexta (una semana de 6 días), son personas que permanece fuera de la vivienda más frecuentemente en diversos niveles. Así, la frecuencia promedio de estar fuera del hogar es mayor en aquellos que respondieron en el segundo intento que en aquellos que fueron localizados en el primer intento. Y así sucesivamente con los demás casos. Partiendo del supuesto que existe una correlación entre las variables de interés y la propensión de estar presente en casa, la técnica tiene como fin reconstruir de un estatus de estar presente en la vivienda el individuo objetivo, con una sola visita e indagando la presencia del individuo objetivo en los 6 días hábiles previos a la entrevista.

Así, por ejemplo, es posible construir seis grupos. Dentro de la semana que se realizó la encuesta

1. Están en la vivienda 1/6 del tiempo
2. Están en la vivienda 2/6=1/3 del tiempo
3. Están en la vivienda 3/6=1/2 del tiempo
4. Están en la vivienda 4/6=2/3 del tiempo
5. Están en la vivienda 5/6 del tiempo
6. Están en la vivienda 6/6=1 del tiempo

Así, a cada individuo de la muestra alcanzada se asigna a sólo a uno de estos grupos. Así, se puede hacer una corrección para la subrepresentación de cada grupo. Es decir, se hace el modelo que sólo un sexto de los individuos del primer grupo fueron entrevistados, un tercio de los individuos del segundo grupo, y así sucesivamente, es decir, se ha estimado la probabilidad de respuesta $\hat{\phi}_i$ para todos los elementos de la muestra que respondieron. Por consecuencia, se deben asignar los siguientes ponderadores:

- 6 para el primer grupo,
- 3 para el segundo grupo
- 2 para el tercer grupo
- 3/2 para el cuarto grupo
- 6/5 para el quinto grupo
- 1 para el sexto grupo.

El objetivo de este método de ponderación es disminuir el sesgo de las estimaciones, aunque no se elimina totalmente. Aunque tiene como principal característica que compensa por la baja representación de personas que no están

frecuentemente en la vivienda. A continuación se enuncia el método de marea más formal.

5.5.3 Resumen del método de Politz-Simmons

1. A cada persona en muestra se visita solo una vez bajo un diseño de muestreo de probabilidad $P(\cdot)$ arbitrario.
2. De cada persona entrevistada, se obtiene información sobre si ha estado presente o no en 6 ocasiones específicas, determinadas de forma aleatoria, incluyendo la de la misma entrevista, lo cual permite estimar la proporción de tiempo que el individuo objetivo se encuentra en la vivienda durante el horario que se está llevado a cabo la entrevista.
3. Los cuestionarios son divididos en seis grupos de acuerdo a la proporción de tiempo que las personas están en el hogar. Es decir, 1/6, 2/6, ...,6/6 del tiempo para los grupos del 1 al 6 respectivamente.
4. El estimador, para cualquier variable de estudio, se produce al ponderar los resultados de cada grupo por el recíproco del porcentaje de tiempo estimado que el individuo objetivo permanece en su hogar. Así, los inversos de las probabilidades de respuesta ϕ_h para h=1 al 6, respectivamente son: 6/1, 6/2, ..., 6/6.

5.5.4 Supuestos del modelo Politz-Simmons

La población de la cual se obtiene la muestra, está restringida a aquellos individuos que están en la vivienda por lo menos en algún momento durante las horas de entrevista establecidas en el diseño. Es decir, la población de la que se

seleccionó la muestra está constituida por aquellas personas que eventualmente podrían ser encontradas durante las horas regulares que se realizan encuestas.

La decisión de definir las horas de entrevista para la encuesta, es sumamente importante para dar sentido a los resultados. Es decir, entre más corto sea el lapso de tiempo para realizar las entrevistas, más grande es el número de personas a las que arbitrariamente se les ha excluido de la población que se ha seleccionado en la muestra por no haberles proporcionado alguna oportunidad de ser seleccionados.

Otro supuesto fuerte de la metodología, es que los entrevistados estarán dispuestos a revelar y / o recordar el lapso de tiempo de su estancia en sus viviendas. Donde lo primero es difícil por cuestiones de seguridad en México actualmente, y lo segundo si no es reportado con exactitud destruye el modelo, pero no es posible evaluar la calidad de dicha información.

5.6 Ventajas y desventajas de los métodos de ponderación

La principal ventaja de los métodos de ponderación es que son metodologías, que si se toman los supuestos correctos, pueden eliminar el sesgo de forma considerable de las estimaciones.

Una característica importante de esta metodología es que la estimación de la probabilidad de responder ϕ se hace para cada unidad de observación, pero dentro de cada unidad se responden k preguntas de interés o variables. Así, si existen probabilidades de responder diferentes por variable para una misma

unidad de observación, entonces el utilizar una sola probabilidad de responder para toda la unidad trae como consecuencia que algunas variables se puede provocar un incremento en el sesgo.

Capítulo VI

Métodos Paramétricos

6.1 Idea general de los métodos parámetricos

Para los casos previos donde se supone que la no-respuesta es ignorable (ya sea bajo un mecanismo MCAR o MAR dados covariantes) en vez de utilizar un método de reponderación es posible construir un modelo paramétrico de los llamados de superpoblaciones.

Un modelo de superpoblación parte del supuesto que la población que se está observando forma parte de otra población más grande, y ésta a su vez, pertence a una más grande y así sucesivamente, hasta que la población sea tan grande como se desee. Es decir, la población bajo estudio está anidada dentro de una serie creciente de poblaciones finitas.

Así, a partir de estos modelos se predicen los valores y_i que forman parte de la muestra. Bajo este enfoque:

- Se desarrolla un modelo para todos los datos.
- Se integran al modelo aquellos componentes que tomen en cuenta el mecanismo propuesto de no-respuesta.

Una vez construido el modelo se realiza la estimación de los parámetros para los modelos por el método de Máxima Verosimilitud, se calcula la varianza de los

parámetros y sus intervalos de confianza, que tienen la ventaja de que toman en cuenta la no-respuesta.

6.2 Alcances y limitaciones de los métodos paramétricos

Las ventajas de los métodos paramétricos respecto a los otros métodos son:

- La utilización de un modelo es flexible y se puede utilizar para incluir cualquier conocimiento que se tenga sobre el mecanismo de no-respuesta.
- Es necesario establecer supuestos sobre la no-respuesta de forma explícita en el modelo, y por lo tanto, algunos de estos supuestos se pueden evaluar.
- Las estimaciones de la varianza que resultan de ajustar el modelo toman en consideración la no-respuesta siempre y cuando sea un modelo adecuado.

Por otra parte, el uso de modelos paramétricos tiene las siguientes limitantes:

- Es necesario tener un dominio pleno de estadística matemática
- Computadora potente.
- Conocimiento de métodos numéricos para optimización

Las ecuaciones de verosimilitud raramente tienen soluciones cerradas. Así, el cálculo de las estimaciones de los parámetros requiere métodos numéricos. De hecho entre más complejo sean ya sea el diseño de muestreo que se haya

utilizado o el mecanismo de respuesta, las funciones de verosimilitud son más difíciles de construir.

Para aquellos interesados en profundizar en el tema Little y Rubin (1987) exponen métodos de Máxima Verosimilitud en general para corregir por no-repuesta.

Conclusiones

Se han mostrado y analizado las diversas propuestas que se han hecho para tratar la no-respuesta desde la década de 1940 hasta nuestros días. La importancia de esta información radica en que como en casi toda encuesta por muestreo, sino que en todas, siempre hay unidades de observación de las que no se puede obtener información o, por otra parte, hubo no-respuesta puntual, siempre es necesario y posible hacer algo para ajustar las estimaciones y disminuir sesgos.

Se debe siempre tener siempre en cuenta que todos los métodos para ajustar una encuesta por no-respuesta, que fueron expuestos en el presente trabajo, se basan en un modelo, por ende, los métodos funcionan en la medida que dicho modelo se asemeje al mecanismo real de no-respuesta que ocurra en cada caso en particular. Así, la mejor herramienta con la que se cuenta para utilizar estos métodos es un profundo conocimiento del fenómeno bajo estudio y la experiencia del responsable de la encuesta.

Por su puesto, ninguno de los métodos mostrados puede sustituir un dato real, y por lo tanto, los esfuerzos de dirección, planeación, organización y control de una encuesta, cuya responsabilidad recae en el cuerpo directivo de la organización, deben contemplar que la no-respuesta es un problema muy serio. Por lo tanto, todo esfuerzo y recurso destinado para reducirla al máximo siempre traerá como beneficio inmediato que la información derivada de la encuesta sea más confiable

y así los clientes o usuarios de la información finales podrán tomar decisiones a favor de los intereses que les compete y por los que contrataron una investigación por medio de una encuesta por muestreo. Además, se generan resultados y acciones que generan un circulo virtuoso en la sociedad y los negocios.

Bibliografía

1. Cochran, W. G. 1977. *Sampling techniques.* New York: Wiley.
2. Cochran, W. G. 1983. Historical perspective. In *Incomplete data in sample surveys.* Vol. 2. Edited by W. G. Madow, I. Olkin., D. B. Rubin, 11-25. New York: Academic Press.
3. Deming, W. E. 1986. *Out of the Crisis.* Cambridge: MIT Press.
4. Fay, R. E. 1996. Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* 91: 490-498.
5. Ford, B. M. 1983. An overview of hot-deck procedures. In *incomplete data in sample surveys.* Vol. 2. Edited by W. G. Madow, I. Olkin., D. B. Rubin, 185-207. New York: Academic Press.
6. Groves, R. M. 1989. *Survey errors and survey cost.* New york: Wiley.
7. Hansen, M. H., and W. N. Hurwitz. 1946. The problem of non-response in sample surveys. *Journal of the American Statistical Association* 41: 517-529.
8. Hidiroglou, M.A., J.D. Drew, and G. B. Gray. 1993. A framework for measuring and reducing nonresponse in surveys. Survey Methology 19: 81-94.
9. Holt, D., and D. Elliot. 1991. Methods of weighting for unit non-response. *Statistician* 40: 333-342.
10. Little, R. J. A., and D. B. Rubin. 1987. *Statistical Analysis with missing data.* New York: Wiley.
11. Lohr, S.L., 1999. *Sampling: design and analysis.* Pacific Grove: International Thompson Publishing Company.
12. Madow, W. G., I. Olkin., D. B. Rubin, eds. 1983. *Incomplete data in sample surveys.* New York: Academic Press.
13. Oh, H. L., and F. J. Scheuren. 1983. Weighting adjustment for unit nonresponse. In *incomplete data in sample surveys.* Vol. 2. Edited by W. G. Madow, I. Olkin., D. B. Rubin, 143-184. New York: Academic Press.
14. Politz, A., and W. Simmons. 1949. An attempt to get the "not at homes" into he sample without callbacks. *Journal of the American Statistical Association* 44: 9-31.
15. Potthoff, R. F., K. G. Manton, and M. A. Woodbury. 1993. Correcting for nonavailability bias in surveys by weighting based on the number of callbacks. *Journal of the American Statistical Association* 88: 1197-1207.
16. Rao, J. N. K. 1996. On variance estimation with imputed survey data. *Journal of the American Statistical Association* 91: 499-506.
17. Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.
18. Sande, I. G. 1983. Hot-deck imputation procedures. In *incomplete data in sample surveys.* Vol. 2. Edited by W. G. Madow, I. Olkin., D. B. Rubin, 339-349. New York: Academic Press.
19. Sarndal, C. E., B. Swensson, and J. Wretman. 1992. *Model assited survey sampling.* New York: Springer-Verlag.