UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

DEPARTAMENTO DE ESTADISTICA E INVESTIGACION OPERATIVA

TRATAMIENTO ESTADISTICO DE LA NO RESPUESTA EN POBLACIONES FINITAS

TESIS DOCTORAL

José Martos Peinado

SEVILLA, 1995

TRATAMIENTO ESTADISTICO DE LA NO RESPUESTA EN POBLACIONES FINITAS

Memoria presentada por el Licenciado José Martos Peinado para optar al Grado de Doctor en Matemáticas.

Año 1995.

VºBº

Director de la Tesis

Prof. Dr.D.Jose Mª Caridad y Ocerín

VºBº

Director de la Tesis

Prof. Dr.D.Jose J. Rodriguez Alcaide

VºBº

Tutor de la Tesis

Prof. Dr. D. Rafael Infante Macias

DEPARTAMENTO DE ESTADISTICA E INVESTIGACION OPERATIVA

FACULTAD DE MATEMATICAS

UNIVERSIDAD DE SEVILLA

AGRADECIMIENTOS

Agradezco, ante todo, a los profesores Dr. José María Caridad y Ocerin y Dr. Jose Javier Rodríguez Alcaide, la posibilidad de haber podido trabajar con ellos en estos últimos años, aprendiendo de sus conocimientos y adquiriendo con la investigación dirigida una visión más amplia y práctica de la Estadística y su aplicación a la Economia. Así mismo, como Directores de ésta Tesis, su permanente disposición en la dirección del trabajo; sus observaciones, sugerencias y oportunas correcciones, han sido decisivas para la culminación de ésta labor de investigación.

Al profesor Dr. Rafael Infante Macias, que como Tutor de Tesis, me ha facilitado el camino, tanto en la inscripción de ésta en el Departamento de Estadística que el dirige, como por su colaboración en la finalización del presente trabajo.

Así mismo, agradezco al Instituto de Estadística de Andalucía y en especial al profesor Dr. Andrés Arroyo Pérez, por presentar el objeto de investigación como campo prioritario para dicho Instituto y por la ayuda aportada en la revisión bibliográfica.

Al profesor Dr. César Hervás, responsable, en gran medida de mi dedicación a la Estadística y junto con el Dr. Antón García , del entusiasmo que tengo para la investigación. El contacto personal con ambos, a diario, me han resultado de inestimable valor.

A la Consejería de Educación y Ciencia de la Junta de Andalucía y la Universidad de Córdoba, que por el acuerdo alcanzado entre las mismas, tengo concedida una comisión de servicios desde 1990 lo que me ha permitido la dedicación necesaria en un trabajo de éstas características y un enriquecimiento intelectual fructífero e inestimable.

A la Unidad de Economia, al Departamento de Matemática Aplicada y de Estadística, Econometría e Investigación Operativa y a la Facultad de Ciencias Económicas y Empresariales (ETEA), todas ellas de la Universidad de Córdoba, que me han facilitado el marco intelectual adecuado para el desarrollo de ésta investigación.

Tambien debo agradecer a los profesores Dr. Gustavo Gómez, Dr. Clemente Mata, Dr. Manuel Sanchez y Dr. Valeriano Domenech y demás miembros del Departamento de Producción Animal de la Facultad de Veterinaria (Universidad de Córdoba), por la deferencia que han tenido commigo en estos años de convivencia así como el brindarme su amistad, que me honra.

Al Dr. Joaquín Reyes, que ha sido el que ha hecho posible la conexión con el Centro de Cálculo de la Universidad de Córdoba; sus sugerencias han sido decisivas en la elaboración de los gráficos y tablas del presente trabajo.

Por último, quiero agradecer el apoyo incondicional de mi familia que han tenido que sufrir, soportar y aliviar el desgaste que este tipo de trabajos lleva inherente.

A mi esposa e hijos.

A la memoria de mis padres.

INDICE

INDICE

INTRODUCCION

INTRODUCCION

El muestreo de una población está dirigido a obtener información acerca de una o varias características de las unidades que la componen. Esta información se obtiene a través de una encuesta y en la mayoría de los casos de una parte de la población, a la que llamamos muestra, a partir de la cual proponemos estimadores sobre dichas características de forma que proporcionan la información más eficiente.

El objetivo usual pues de una encuesta por muestreo es el de obtener una estimación, puntual o por intervalo, de algunos de los parámetros poblacionales. Un problema surge cuando al seleccionar una muestra, algunas de las unidades no responden para una característica en particular, o tambien cuando el marco no está totalmente depurado.

Hablaremos de no respuesta cuando no se obtienen observaciones, total o parcialmente, de las unidades seleccionadas y designadas para la muestra. La presencia de ésta, en la encuesta, crea problemas y entre otros, en la estimación de los parámetros poblacionales, con la presencia del sesgo, el cual, surge al no ser la información faltante una muestra aleatoria de los datos que se buscan y cuya magnitud es desconocida.

En el **capítulo I**, nuestro objetivo es doble. En una primera parte estudiamos los errores en las encuestas, las causas y efectos que produce la no respuesta sobre la estimación de un parámetro poblacional, tanto a nivel puntual como por intervalo. Y en una segunda parte, estudiamos los principales procedimientos en las distintas fases del muestreo, para tratar la no respuesta.

En la estimación puntual, es conocido (Kish, 1965 y Cochran, 1977) que al no proveer la muestra, con no respuesta, información acerca del parámetro poblacional de los que no responden, el tamaño del sesgo nos es desconocido. Utilizando el concepto de sesgo relativo (Kish, 1965) así como el estudio que hace Plateck (1986), sobre la posibilidad de estimar adecuadamente el nivel (o tasa) de respuesta esperado en una encuesta, dependiendo de las condiciones en que se realiza ésta, nosotros proponemos expresiones para el sesgo relativo del estimador muestral, tanto de la media como del total poblacionales. Así mismo, comparando las expresiones de ambos sesgos, llegamos a la conclusión, entre otras de índole práctica, de que a igualdad de tasas de no respuesta, es mejor estimar el total que la media.

Para la estimación por intervalo, al desconocer el tamaño del sesgo, hace imposible la asignación de intervalos de confianza útiles para la estimación de los parámetros poblacionales (Azorín y Sanchez-Crespo, 1986). No obstante, el efecto del sesgo en la exactitud de un estimador es estudiado experimentalmente por varios autores. Así, Raj (1968) y Cochran (1977) comparan, con sus respectivos niveles de confianza, el intervalo asociado a un estimador insesgado con el de un estimador sesgado, siendo el error máximo admisible el mismo para ambos, como precisión mínima a exigir de los resultados e igual a un factor que depende del error estándar del estimador muestral que estemos utilizando. Llegan experimentalmente a la conclusión, entre otras, de que el nivel de confianza del segundo es menor que el del primero.

Así mismo, Hansen, Hurwitz y Madow (1953) y posteriormente Kish (1965) comparan, con sus respectivos niveles de confianza, el intervalo asociado a un estimador insesgado con el de un estimador sesgado, siendo el error máximo admisible del primero igual a un factor que depende del error estándar y, el del segundo, de un factor que depende del error total, definido como la raíz cuadrada del error cuadrático medio del estimador muestral que estemos utilizando. Llegan experimentalmente a la conclusión de que el efecto del sesgo es despreciable siempre que el valor absoluto del sesgo del estimador no sobrepase su error estándar.

Nosotros estudiamos de una forma genérica, dada su complementariedad, las relaciones de orden entre los niveles de confianza o las regiones de error, indistintamente, existentes entre los tres tipos de intervalos indicados anteriormente y demostramos que la región de error asociada al segundo intervalo es mayor que la del primero, consecuencia a la que llegan Cochran y Raj experimentalmente. Así mismo que la región de error del tercero es inferior a la del segundo y finalmente que la relación de orden entre el tercero y el primero depende del cociente entre el valor absoluto del sesgo y el error estándar. Ampliamos la solución propuesta por Hansen, Hurwitz y Madow, indicada anteriormente, y demostramos que es posible identificar los intervalos asociados al estimador sesgado e insesgado, con unas características definidas, no siendo necesario que el sesgo del estimador no sobrepase su error estándar.

Por otro lado, para el tratamiento de la no respuesta en una encuesta, dos son los caminos que se siguen, según la etapa de realización en la que nos encontremos, y que en ocasiones son complementarias. Hay un camino inicial y es el de evitar la no respuesta que, será posible, cuando nos encontremos en la fase de planificación, de diseño o de campo; en ésta última fase damos una visión de las técnicas empleadas, como revisión bibliográfica, desde la visitas adicionales (Demming, 1953, Kish y Hess, 1959 y Cochran, 1977) a las técnicas de respuesta aleatorizada (Ladoux, 1962 y Warner, 1965). El segundo camino es el de aceptar que hay no respuesta y buscar soluciones de cómo tratarla; las técnicas de depuración (Villán y Bravo, 1990 y Villar, 1992) estudian las respuestas inconsistentes

y las de imputación (Bailar, Bailey y Corby, 1978 y Rubin, 1986) las técnicas para sustituir unidades con no respuesta; para ésta fase de procesamiento o validación damos una visión de las técnicas empleadas.

Los métodos de muestreo más conocidos y utilizados consideran estimadores que utilizan sólo los valores observados de la característica en estudio a los que denominamos estimadores directos o expandidos. Para el caso de muestras con no respuesta ¿ cuáles son los estimadores muestrales que estiman a los parámetros poblacionales media y total ?; en la bibliografía consultada, sólo Sanchez-Crespo (1976), propone un estudio comparativo de varios estimadores muestrales.

En el **capítulo II**, dentro de los diferentes tipos de muestreo y considerando que hay elementos de la muestra que no responden, nosotros hemos elegido tres de ellos :

Para el primero, **muestreo sin reposición aleatorio simple** (m.a.s.) proponemos un conjunto de estimadores directos de la media y el total, cada uno con una hipótesis que lo caracteriza. Estudiamos la sesgadez o insesgadez y deducimos una expresión razonada de su varianza y el de un estimador de ésta; así mismo, los comparamos en cuanto a precisión y exactitud (Cochran, 1977) o eficiencia (Cramer, 1960) y acuracidad (Azorín y Sanchez-Crespo, 1986), respectivamente.

En la elección de estos estimadores han influido diversos factores, siendo el principal el considerar las diferentes situaciones en las que nos podemos encontrar en muestras con unidades que no responden y su correspondiente tratamiento : Sustitución por cero; sustitución con una variable auxiliar, eliminación, eliminación pero considerando el tamaño de la muestra en su totalidad ,sustitución por la media y duplicación ; así mismo, al considerar el submuestreo como una opción más frente a la no respuesta planteamos el estimador de Hansen y Hurwitz (1946).

Demostraremos bajo ciertas condiciones que, de los estimadores sesgados, el que presenta menor varianza, es más acurado que el más eficiente de los insesgados y, por consiguiente, preferible a éste (Cansado, 1983) . Así mismo, estudiamos la eficiencia relativa (Oh y Scheuren, 1983) y, utilizando las consecuencias obtenidas en el capítulo anterior, estudiamos los intervalos de confianza para los estimadores sesgados, considerando un error máximo admisible que depende del error total. Proponemos bajo qué condiciones el intervalo asociado al estimador con menor varianza es el que mejor se aproxima al intervalo asociado al estimador insesgado.

Con los otros dos métodos de muestreo elegidos, **muestreo estratificado y muestro de conglomerados bietápicos**, elegimos los dos estimadores insesgados, de entre los estudiados para una m.a.s. y planteamos las mismas hipótesis para cada uno de los estratos ó conglomerados y proponemos estimadores directos del conjunto de la muestra para la media y el total. Estudiamos la sesgadez o insesgadez y deducimos una expresión razonada de su varianza y el de un estimador de ésta.

Dado que uno de los estimadores propuestos para cada uno de los estratos o conglomerados es el de Hansen y Hurwitz, lo que estamos proponiendo es una generalización de éste estimador para estos dos últimos tipos de muestreo. Además, para el muestreo estratificado y con éste estimador, proponemos diversas consideraciones según que la estratificación sea proporcionada o desproporcionada, estudiando en éste último caso dos cuestiones (para la última y a nivel de m.a.s., Raj (1968) propone una función de coste análoga con éste estimador):

a) La afijación óptima, como medio de asignar fracciones de muestreo a los estratos con el objetivo de obtener la mínima varianza del estimador.

b) La afijación óptima con función de coste, como medio de asignar fracciones de muestreo y tasa de submuestreo a los estratos, de tal manera que el coste esperado sea mínimo para un valor determinado de la varianza del estimador.

Es conocido que la forma más sencilla de autoponderación, entre los elementos de una muestra, es la obtenida a través de una m.a.s, dado que cada elemento de la muestra pondera 1/n en la media. A pesar de la sencillez de las muestras autoponderadas, Kish (1965) indica varias razones para añadir ponderaciones desiguales en partes de la muestra; una de éstas es la de balancear con ponderaciones desiguales las diferencias en no respuesta entre partes de la muestra, si ésta está estratificada. Para el caso de una muestra no estratificada, sugiere la necesidad de realizar alguna suposición acerca de la aleatoridad, proponiendo particionar la muestra en un nùmero finito de subconjuntos aleatoriamente seleccionados y ponderar cada particiòn. Al utilizar el mismo estimador en cada partición, la varianza de la media ponderada aumenta y el sesgo se mantiene en la medida que lo sea el estimador empleado (Murthy y Sethi, 1961 y Kish, 1965).

Nosotros en el **capítulo III** proponemos, al igual que Kish, particionar la muestra en un nùmero finito de subconjuntos aleatoriamente seleccionados y ponderar cada particiòn, utilizando ternas de dos, de entre los estimadores muestrales estudiados en el segundo capítulo, definiendo el estimador media ponderada sobre el total de elementos de la nueva muestra y obteniendo diversas expresiones del estimador media ponderada, en función de los estimadores empleados en la particiòn.

Al utilizar distintos estimadores, nosotros hemos encontrado que, bajo ciertas condiciones, es posible no sólo disminuir la varianza sino encontrar un valor mínimo absoluto de ésta. Así mismo, al combinar diversos estimadores, sesgados e insesgados, disminuimos el sesgo del estimador media ponderada.

CAPITULO I

CAPITULO I

LA NO RESPUESTA EN LAS DISTINTAS FASES DE REALIZACION DE UNA ENCUESTA: EFECTOS Y TRATAMIENTO

1. Introducción

Como fuentes principales para la información estadística debemos considerar simultaneamente a los censos, muestras y registros administrativos. Los primeros proporcionan infraestructura y marcos para los segundos, los cuales no deben considerarse como una actividad independiente sino encuadrada dentro de la organización y planificación estadística. Y los terceros deben considerarse una fuente de información, en ocasiones imprescindible, para nuestro objetivo.

La calidad de la investigación estadística, encaminada a la obtención de datos para la satisfación de necesidades de información de la sociedad a través de la recogida de datos muestrales, censales o aprovechando datos administrativos, se logra extremando el cuidado en la realización de las diferentes etapas que componen una muestra o censo. Pese a ello, nos encontramos que, al seleccionar una muestra, algunas de las unidades que la componen no responden.

Hablaremos de no respuesta cuando no se obtienen observaciones, total o parcialmente, de las unidades seleccionadas para la muestra. La presencia de la no respuesta en la encuesta crea problemas en la estimación de los parámetros poblacionales, en los costes, tiempos y recursos y en la difusión de los resultados.

Nuestro objetivo en éste capítulo es doble. En una primera parte estudiamos los errores en las encuestas, las causas y efectos que produce la no respuesta sobre la estimación de un parámetro poblacional, tanto a nivel puntual como por intervalos, proponiendo, para estos últimos, soluciones que traten de evitar el sesgo, consustancial con la no respuesta. Así mismo, en una segunda parte estudiamos los principales procedimientos en las distintas fases del muestreo, para tratar la no respuesta.

La notación que vamos a emplear en el desarrollo de éste capítulo, y que ampliaremos en el siguiente, es :

A) Para la Población

Y = Característica a estudiar

N = Número de unidades

Y_N = Total poblacional

$\bar{Y}_N$ = Media poblacional

N_1 = Número de unidades que responden

Y_{N1} = Total de las unidades que responden

$\bar{Y}_{N1}$ = Media de las unidades que responden

N_2 = Número de unidades que no responden

Y_{N2} = Total de las unidades que no responden

$\bar{Y}_{N2}$ = Media de las unidades que no responden

$t_{N1} = N_1 / N$ = Tasa de respuesta

$t_{N2} = N_2 / N$ = Tasa de no respuesta

B) Para la muestra

n = Número de unidades

$f = n / N$ = fracción de muestreo

n_1 = Número de unidades que responden

n_2 = Número de unidades que no responden

$\overset{*}{Y}_N$ = Total muestral, estimador de Y_N

$\bar{Y}_n$ = Media muestral, estimador de $\bar{Y}_N$

Y_{n1} = Total de las unidades que responden

$\bar{Y}_{n1}$ = Media de las unidades que responden

Y_{n2} = Total de las unidades que no responden

$\bar{Y}_{n2}$ = Media de las unidades que no responden

$t_{n1} = n_1 / n$ = Tasa de respuesta

$t_{n2} = n_2 / n$ = Tasa de no respuesta

B = valor absoluto del sesgo

2. Errores en las encuestas . Causas de la no respuesta

Siempre que se empleen métodos de sondeo, las estimaciones presentarán ciertas discrepancias con los resultados correspondientes de un recuento completo, basado en las mismas definiciones y en las mismas técnicas de recogida de datos; este fenómeno es consustancial con los métodos de muestreo y no se puede evitar.

Son dos las fuentes de las discrepancias : los errores de muestreo y los ajenos al muestreo.

Los primeros provienen de que las estimaciones no se basan en todas las unidades que integran la población sino tan sólo en una parte de ellas. Los segundos se producen por el mal entendimiento de las preguntas, conceptos, definiciones o instrucciones, tanto por parte del respondiente como por parte de los individuos que intervienen en las distintas etapas del tratamiento de los datos.

Dentro de estos últimos se incluyen los errores producidos por respuestas intencionadamante erróneas por parte del respondiente, a los que denominaremos errores de observación, y los errores debidos a la ausencia de obtener datos de algunas partes de la población de la encuesta, a los que llamaremos errores de no observación.

Podemos distinguir tres causas que generan errores de no observación : La no cobertura, las respuestas inconsistentes y la no respuesta o falta de respuesta.

La no cobertura, llamada tambien de marcos incompletos, consiste en dejar sin incluir algunas unidades o secciones completas de la población definida para la encuesta, en el marco operacional de muestreo con que se está trabajando.

La no cobertura se refiere a los errores negativos producidos por las ausencias en la inclusión de elementos que deberían estar en la muestra , existiendo tambien el error positivo de sobrecobertura, que es cuando se agregan a la muestra elementos que no deberían estar.

Las respuestas inconsistentes son aquellas que, relacionadas con otro dato del mismo individuo, indican una respuesta incorrecta o en todo caso muy poco real. Plateck (1986), las define como imposibilidades lògicas o posibilidades altamente improbables.

La no respuesta consiste en la no obtención de observaciones, total (entrevista no realizada) o parcial (respuestas omitidas), de las unidades seleccionadas y designadas para la muestra.

Las causas por las que no se obtienen observaciones en el total de una unidad, son :

* Por que no están en el momento de la visita

* Por que rechazan la entrevista

* Por que no les es posible por incapacidad o imposibilidad

* Por que no se encuentran

* Por que han sido encuestados pero se ha perdido el cuestionario

Las causas por las que no se obtienen observaciones parcialmente en una unidad, son :

* Rechazo a la pregunta por motivos personales ,de intimidad, fiscales, etc..

* Pregunta mal formulada y que no se entiende

* Carecer de los conocimientos apropiados para contestar a la pregunta

La falta de información, para éste último tipo, la desglosamos entre el " no sabe " , el "no contesta" y los que aparecen en blanco o son ilegibles. Para algunos autores consultados , Lininger (1984) , Chevry (1967) y Cruz (1990), la contestación "no sabe" puede, o no, ser una información faltante, dado que para algunas preguntas referentes a opiniones y ciertos tipos de información, el "no sabe" es una respuesta tan válida o útil como cualquier otra, ya que refleja desconocimiento de la información a dar sobre la pregunta formulada. El "no contesta" o "ilegible" es, pués, la considerada como una falta total de información.

Con la falta de respuesta, total o parcial, pueden originarse algunos problemas bàsicos en las encuestas (Lininger, 1984 y Azorín y Sanchez-Crespo, 1986) y que son :

* Los sesgos, (entendiendo por sesgo el error que conduce a una diferencia entre el valor esperado de la poblaciòn y el valor real), que surgen dado que la informaciòn faltante no es una muestra aleatoria de los datos que se buscan y cuya magnitud es desconocida. En teoría, si la información faltante constituyera una muestra aleatoria, su omisión tendría pocas consecuencias prácticas, dado que la omisión de una parte aleatoria de una muestra de esa categoría deja una muestra más pequeña, pero siempre aleatoria.

* Cuando la información faltante se necesita como parte de una serie de información más amplia sobre el mismo entrevistado. Por ejemplo en la confección de algún índice, formado a partir de las respuestas dadas y que al desconocer una o varias de éstas, la medida de éste puede resultar incompleta o sesgada.

* En el análisis de etapas múltiples, esto es, cálculos que utilizan información sobre tres o más variables y en el que al faltar información de alguna de ellas no será posible clasificar al individuo en las tablas de entrada múltiple.

* Una disminuciòn en el tamaño de la muestra que disminuye la precisiòn y como consecuencia un incremento de la varianza muestral.

3. Efectos de la no respuesta

En el estudio de la no respuesta es conveniente pensar que la población está dividida en dos estratos, los que responden y los que no responden. Al seleccionar una muestra para estudiar una determinada característica de la población, nos encontraremos con un conjunto de unidades que responden y otras que no responden. Sobre éstas últimas la muestra no provee información y, en principio, no podemos suponer que las características del estrato de no respondientes son las mismas que las de los que responden.

Vamos a estudiar el efecto de la no respuesta sobre la estimación de la muestra, en cinco vertientes : estimación puntual de la media y del total , estimación de la varianza y razón poblacionales y estimaciòn por intervalos de la media.

En lo que sigue denotamos por Y_i el valor de la característica Y en la unidad i, responda o no ; por X_i el valor de la característica Y en la unidad i que responde. Suponemos, sin pérdida de generalidad, que las unidades están ordenadas y las primeras son las que han respondido.

3.1. En la estimación de la media poblacional

Definición 3.1.1 Media poblacional a estimar : $\bar{Y}_N = 1/N \cdot \sum_{i=1}^{N} Y_i$

Proposición 3.1.2 La media poblacional puede descomponerse en suma de dos términos, que dependen de las unidades que responde y no responden.

Demostración :

$$\bar{Y}_N = 1/N \cdot \sum_{i=1}^{N} Y_i = 1/N \cdot [\sum_{i=1}^{N_1} X_i + \sum_{i=N1+1}^{N} Y_i] = 1/N. [N_1 . \bar{Y}_{N1} + N_2 . \bar{Y}_{N2}] =$$

$$t_{N1} . \bar{Y}_{N1} + t_{N2} . \bar{Y}_{N2}$$

c.q.d.

Corolario 3.1.3 La media poblacional puede descomponerse en suma de varios tèrminos que dependen de las unidades que responden y de las que no responden, supuesto que haya varios tipos de no respuesta, en la forma:

$$\bar{Y}_N = t_{N1} . \bar{Y}_{N1} + \sum_{i=2}^{k} t_{N2i} . \bar{Y}_{N2i}$$

siendo k los tipos de no respuesta , $\bar{Y}_{N2i}$ la media y t_{N2i} la tasa de no respuesta para cada uno de estos grupos.

Demostraciòn : Consecuencia de la proposiciòn anterior

Proposición 3.1.4 Siendo $\bar{Y}_n$ un estimador muestral de $\bar{Y}_N$, tal que $E(\bar{Y}_n) = \bar{Y}_{N1}$, media poblacional de los que responden, la cantidad de sesgo y sesgo relativo (SR) al estimar $\bar{Y}_N$ por $\bar{Y}_n$, vienen determinados por:

$$\text{sesgo}(\bar{Y}_n) = t_{N2} [\bar{Y}_{N1} - \bar{Y}_{N2}] \qquad (1)$$

$$SR(\bar{Y}_n) = t_{N2} (1-W) / (1- t_{N2}(1-W)) \text{ , siendo } W = \bar{Y}_{N2}/\bar{Y}_{N1}$$

Demostración :

Dado que el sesgo viene definido por : $\text{sesgo} = E(\bar{Y}_n) - \bar{Y}_N = \bar{Y}_{N1} - \bar{Y}_N$

De acuerdo con la proposición 3.1.2, se tiene :

$$\text{sesgo} = \bar{Y}_{N1} - \bar{Y}_N = \bar{Y}_{N1} - [t_{N1} . \bar{Y}_{N1} + t_{N2} . \bar{Y}_{N2}] = t_{N2} [\bar{Y}_{N1} - \bar{Y}_{N2}]$$

Por otro lado, el sesgo relativo es el cociente :

$$SR(\bar{Y}_n) = t_{N2} . [\bar{Y}_{N1} - \bar{Y}_{N2}] / \bar{Y}_N$$

y de acuerdo con la proposiciòn 3.1.2, se sigue :

$$SR(\bar{Y}_n) = t_{N2} . [\bar{Y}_{N1} - \bar{Y}_{N2}] / (t_{N1}. \bar{Y}_{N1} + t_{N2}. \bar{Y}_{N2}) =$$

$$t_{N2}.(1-\bar{Y}_{N2}/\bar{Y}_{N1})/(t_{N1} + t_{N2}.\bar{Y}_{N2}/\bar{Y}_{N1}) =$$

$$t_{N2}.(1-W)/(1 - t_{N2} + t_{N2}.W) =$$

$$t_{N2} (1-W) / (1- t_{N2}(1-W))$$

c.q.d

3.1.1. Conclusiones

Conclusiòn 3.1.5 A la vista de lo demostrado y como conclusiòn, podemos decir que dado que la muestra, como hemos indicado, no provee información acerca de la media poblacional de los que no responden, el tamaño del sesgo nos es desconocido y, en consecuencia, hace imposible la asignación de intervalos de confianza útiles a la media poblacional a partir de los resultados de la muestra (Cochran, 1977).

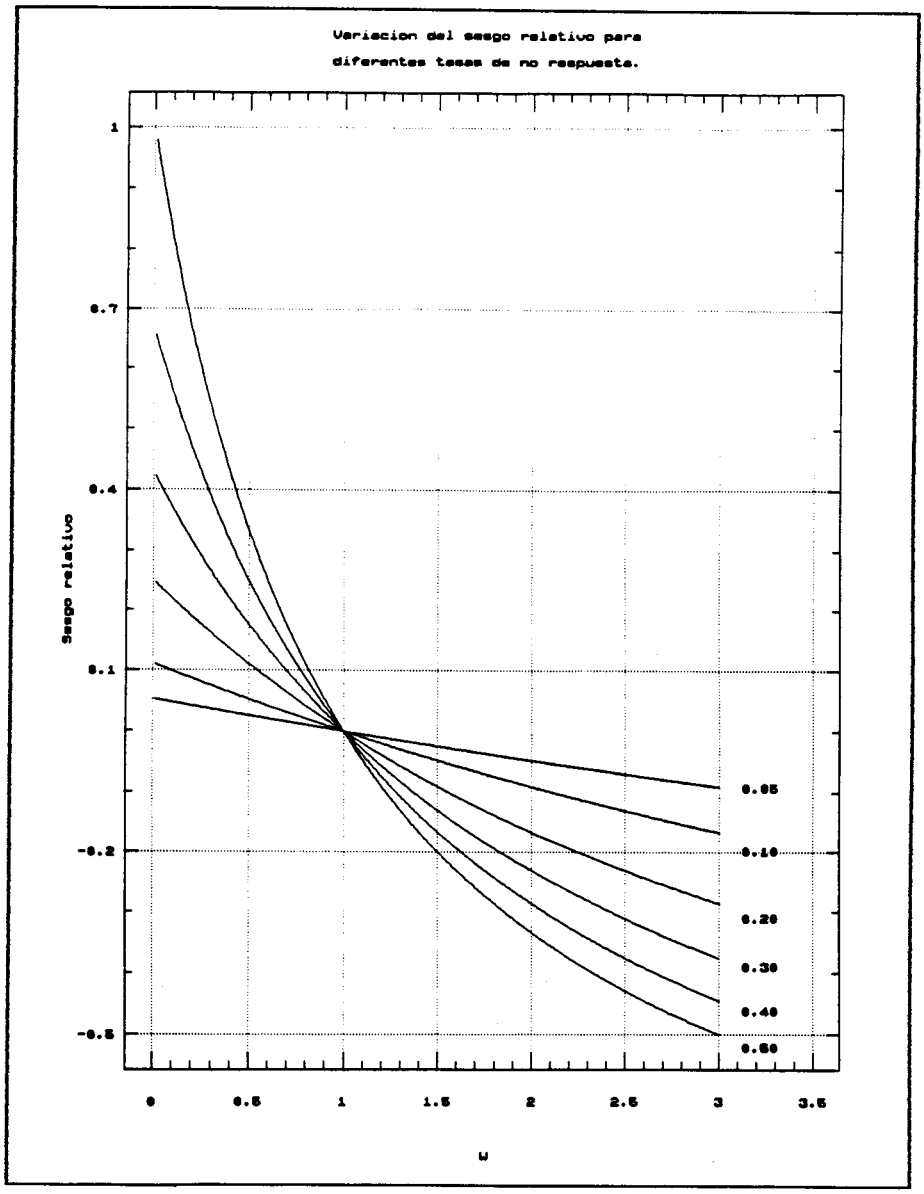
Para el caso en que estimemos que las respuestas de los que no responden son iguales a las de los que responden, y/o la tasa de no respondientes es prácticamente nula, el estimador propuesto es insesgado, dado que la expresión (1) de la proposición 3.1.4 se hará cero.

Por otro lado, si en el Gráfico I.1 representamos la expresiòn obtenida en la proposiciòn 3.1.4, referente al sesgo relativo del estimador de la media poblacional, esto es :

$$SR(\bar{Y}_n) = t_{N2}(1-W) / (1- t_{N2}(1-W))$$

Fijando t_{N2} (utilizamos tasas del 5%, 10%, 20%, 30%, 40% y 50%) y variando W en el intervalo (0,3) con incrementos de 0.2, obtenemos diferentes gráficas correspondientes a la expresiòn indicada anteriormente, segun la tasa de de no respuesta utilizada.

Gráfico I.1



Observamos que :

* Variando W, podemos pasar de una tasa a otra inferior, sin variar el sesgo.

* De igual forma, si conocemos a priori el valor de la tasa de no respuesta, podemos obtener la magnitud del sesgo relativo con la ayuda del gràfico, siempre que conozcamos por otras encuestas anàlogas el valor de W.

* Finalmente, si fijamos unos lìmites tolerables para que el sesgo quede dentro de ellos, es posible determinar los lìmites en la tasa de no respuesta necesario para tal fìn.

3.2. En la estimación de la varianza poblacional

Definición 3.2.6 Varianza ajustada o cuasivarianza :

a) Poblacional : $S_N^2 = 1/(N-1) \cdot \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2$

b) Poblacional de los que responden : $S_{N1}^2 = 1/(N_1-1) \cdot \sum_{i=1}^{N_1} (X_i - \bar{Y}_{N1})^2$

c) Poblacional de los que no responden : $S_{N2}^2 = 1/(N_2-1) \cdot \sum_{i=N1+1}^{N} (Y_i - \bar{Y}_{N2})^2$

Proposición 3.2.7 La varianza ajustada poblacional puede descomponerse en suma de cuatro términos, que dependen de las unidades que responde y no responden, en la forma :

$$S_N^2 = (N_1-1)/(N-1) \cdot S_{N1}^2 + N_1/(N-1) \cdot (\bar{Y}_{N1} - \bar{Y}_N)^2 + (N_2-1)/(N-1) \cdot S_{N2}^2 + N_2/(N-1) \cdot (\bar{Y}_{N2} - \bar{Y}_N)^2$$

Demostración :

De la definición 3.2.6 , se sigue :

$$(N-1) \cdot S_N^2 = \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2 = \sum_{i=1}^{N_1} (X_i - \bar{Y}_N)^2 + \sum_{i=N1+1}^{N} (Y_i - \bar{Y}_N)^2 =$$

$$\sum_{i=1}^{N_1} [(X_i - \bar{Y}_{N1}) + (\bar{Y}_{N1} - \bar{Y}_N)]^2 + \sum_{i=N1+1}^{N} [(Y_i - \bar{Y}_{N2}) + (\bar{Y}_{N2} - \bar{Y}_N)]^2 =$$

al desarrollar, los términos con doble producto son iguales a cero, dado que para el primer sumando, se tiene :

$$\Sigma\ (X_i-\bar{Y}_{N1}).(\bar{Y}_{N1}-\bar{Y}_N) = (\bar{Y}_{N1}-\bar{Y}_N).\Sigma\ (X_i-\bar{Y}_{N1}) = (\bar{Y}_{N1}-\bar{Y}_N).0$$

siendo para el otro sumando análogo, por lo que quedará :

$$\sum_{i=1}^{N_1} (X_i - \bar{Y}_{N1})^2 + \sum_{i=1}^{N_1} (\bar{Y}_{N1} - \bar{Y}_N)^2 + \sum_{i=N1+1}^{N} (Y_i - \bar{Y}_{N2})^2 + \sum_{i=N1+1}^{N} (\bar{Y}_{N2} - \bar{Y}_N)^2 =$$

y de acuerdo con la definición 3.2.6, se sigue :

$$(N_1-1)\ .\ S_{N1}^2 + N_1\ .\ (\bar{Y}_{N1} - \bar{Y}_N)^2 + (N_2-1)\ .\ S_{N2}^2 + N_2\ .\ (\bar{Y}_{N2} - \bar{Y}_N)^2$$

despejando S_N^2 de la igualdad, se obtiene la expresión propuesta.

c.q.d.

Corolario 3.2.8 Una descomposición de la varianza poblacional es

$VAR_N = t_{N1}\ .\ VAR_{N1} + t_{N2}\ .\ VAR_{N2} + t_{N1}\ .\ t_{N2}\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})^2$, siendo

VAR_N, VAR_{N1}, y VAR_{N2} las varianzas poblacional, de los que responden y de los que no responden, respectivamente.

Demostración :

Por la proposición 3.1.2 , deducimos :

$$\bar{Y}_{N1} - \bar{Y}_N = t_{N2}\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})$$

$$\bar{Y}_{N2} - \bar{Y}_N = -1.\ t_{N1}\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})$$

Elevando al cuadrado y multiplicando cada expresión por N_1 y N_2, respectivamente, y sumando miembro a miembro, obtenemos :

$$N_1\ .\ (\bar{Y}_{N1} - \bar{Y}_N)^2 + N_2\ .\ (\bar{Y}_{N2} - \bar{Y}_N)^2 =$$

$$N_1\ .\ t_{N2}^2\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})^2 + N_2.\ t_{N1}^2\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})^2 =$$

$$(\ N_1\ .\ t_{N2}^2 + N_2.\ t_{N1}^2\ )\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})^2 =$$

$$N_1.N_2.(\ N_1+N_2)/N^2\ .(\bar{Y}_{N1} - \bar{Y}_{N2})^2 =$$

$$N_1 \cdot N_2 \cdot 1/N \cdot (\bar{Y}_{N1} - \bar{Y}_{N2})^2 \qquad (1)$$

Utilizando la expresión obtenida en la proposición 3.2.7 y sutituyendo en ella la expresión (1), se sigue :

$$(N-1).S_N^2 = (N_1-1).S_{N1}^2 + (N_2-1).S_{N2}^2 + N_1.N_2.1/N.(\bar{Y}_{N1} - \bar{Y}_{N2})^2$$

Teniendo en cuenta la relación entre la varianza ajustada y la varianza, se tiene :

$$N \cdot VAR_N = N_1 \cdot VAR_{N1} + N_2 \cdot VAR_{N2} + N_1 \cdot N_2 \cdot 1/N \cdot (\bar{Y}_{N1} - \bar{Y}_{N2})^2$$

Finalmente, diviendo por N, obtenemos la expresión propuesta.

c.q.d.

Corolario 3.2.9 Siendo S_n^2 la varianza ajustada muestral, puede descomponerse en suma de cuatro términos, que dependen de las unidades que responden y no responden en la muestra, en la forma :

$$S_n^2 = (n_1-1)/(n-1) \cdot S_{n1}^2 + n_1/(n-1) \cdot (\bar{Y}_{n1} - \bar{Y}_n)^2 +$$
$$(n_2-1)/(n-1) \cdot S_{n2}^2 + n_2/(n-1) \cdot (\bar{Y}_{n2} - \bar{Y}_n)^2$$

Así mismo, se tiene :

$$VAR_n = t_{n1} \cdot VAR_{n1} + t_{n2} \cdot VAR_{n2} + t_{n1} \cdot t_{n2} \cdot (\bar{Y}_{n1} - \bar{Y}_{n2})^2$$

siendo, VAR_n, VAR_{n1}, y VAR_{n2} la varianza muestral, de los que responden y de los que no responden, respectivamente

Demostraciòn :

Anàloga a las proposiciones anteriores

c.q.d.

Proposición 3.2.10 Siendo S_n^2 un estimador muestral de S_N^2 , tal que $E(S_n^2) = S_{N1}^2$, varianza ajustada poblacional de los que responden, la cantidad de sesgo al estimar S_N^2 por S_n^2, viene determinado por :

$$(N_2-1)/(N-1). \ [S_{N1}^2 - S_{N2}^2] - N_1/(N-1) \cdot (\bar{Y}_{N1} - \bar{Y}_N)^2 -$$
$$N_2/(N-1) \cdot (\bar{Y}_{N2} - \bar{Y}_N)^2 + 1/(N-1) \ S_{N1}^2$$

o bien por :

$$(N_2-1)/(N-1).[S_{N1}^2 - S_{N2}^2] - N_1.N_2.1/N.1/(N-1).(\bar{Y}_{N1}-\bar{Y}_{N2})^2 + 1/(N-1) \ S_{N1}^2$$

demostración :

Teniendo en cuenta que : Sesgo = $E(S_n^2) - S_N^2 = S_{N1}^2 - S_N^2$ y utilizando la expresiòn de S_N^2 obtenida en la proposición 3.2.7 , se sigue que :

$$\text{Sesgo} = S_{N1}^2 - [\ (N_1-1)/(N-1)\ .\ S_{N1}^2 + N_1/(N-1)\ .\ (\bar{Y}_{N1} - \bar{Y}_N)^2 +$$

$$(N_2-1)/(N-1)\ .\ S_{N2}^2 + N_2/(N-1)\ .\ (\bar{Y}_{N2} - \bar{Y}_N)^2\ ] =$$

$$[\ 1\ - (N_1-1)/(N-1)\ ]\ .\ S_{N1}^2 - (N_2-1)/(N-1)\ .\ S_{N2}^2 -$$

$$N_1/(N-1)\ .\ (\bar{Y}_{N1} - \bar{Y}_N)^2 - N_2/(N-1)\ .\ (\bar{Y}_{N2} - \bar{Y}_N)^2 \qquad (1)$$

El primer término, queda reducido a $N_2/(N-1)\ .\ S_{N1}^2$. Sumando y restando a la expresión anterior el término $1/(N-1)\ .\ S_{N1}^2$, obtenemos la primera expresión propuesta.

Si sustituimos los dos últimos términos de la expresión anterior por la expresión (1) del corolario 3.2.8, se sigue la segunda expresión propuesta.

c.q.d.

3.2.1. Conclusiones

Como conclusiòn de èste apartado, consecuencia de la última proposición, podemos indicar que el estimador de la varianza ajustada, basado sólo en el estrato que responde, es insesgado cuando :

$$S_{N1}^2 = S_{N2}^2 = N_1/(N-1)\ .\ (\bar{Y}_{N1} - \bar{Y}_N)^2 + N_2/(N-1)\ .\ (\bar{Y}_{N2} - \bar{Y}_N)^2$$

o bien cuando :

$$S_{N1}^2 = S_{N2}^2 = N_1.\ N_2\ .\ 1/N\ .\ 1/(N-1)\ .\ (\bar{Y}_{N1} - \bar{Y}_{N2})^2$$

3.3. En la estimación del total poblacional

Definición 3.3.11 Total poblacional a estimar : $Y_N = \sum_{i=1}^{N} Y_i$

Proposición 3.3.12 El total poblacional puede descomponerse en suma de dos términos, que dependen de las unidades que responde y no responden.

Demostración :

$$Y_N = \sum_{i=1}^{N} Y_i = \sum_{i=1}^{N_1} X_i + \sum_{i=N1+1}^{N} Y_i = N_1.\ \bar{Y}_{N1} + N_2.\ \bar{Y}_{N2}$$

c.q.d.

Proposición 3.3.13 Siendo Y_n un estimador muestral de Y_N, tal que $E(Y_n) = Y_{N1}$, total poblacional de los que responden, la cantidad de sesgo y sesgo relativo (SR) al estimar Y_N por Y_n, vienen determinados por:

$$sesgo(Y_n) = -1 \,.N_2 \,.\, \bar{Y}_{N2}$$

$$SR(Y_n) = -1.\ t_{N2}\ W \,/\, (1-\ t_{N2}(1-W)) \text{ , siendo } W = \bar{Y}_{N2} \,/\, \bar{Y}_{N1}$$

Demostración :

Dado que el sego viene definido por :

$$sesgo = E(Y_n) - Y_N = Y_{N1} - Y_N = N_1.\ \bar{Y}_{N1} - Y_N$$

Y de acuerdo con la proposición anterior, se sigue :

$$sesgo = N_1.\ \bar{Y}_{N1} - [N_1.\ \bar{Y}_{N1} + N_2.\ \bar{Y}_{N2}] = -1.\ N_2\ .\ \bar{Y}_{N2}$$

Para el sesgo relativo, se tiene :

$$SR(Y_n) = -1.\ N_2\ \bar{Y}_{N2} \,/\, Y_N = -1.\ N\ .\ t_{N2}\ .\ \bar{Y}_{N2} \,/\, Y_N = -1\ t_{N2}\ .\ \bar{Y}_{N2} \,/\, \bar{Y}_N$$

y siguiendo la misma transformaciones que las indicadas en la última parte de la proposiciòn 3.1.4, se obtiene la expresiòn propuesta.

c.q.d.

3.3.1. Conclusiones

Conclusión 3.3.14 A la vista de lo demostrado y como conclusiòn, podemos decir que dado que la muestra, como hemos indicado, no provee información acerca de Y_{N2}, el tamaño del sesgo nos es desconocido y en consecuencia hace imposible la asignación de intervalos de confianza útiles a Y_N, a partir de los resultados de la muestra (Cochran, 1977).

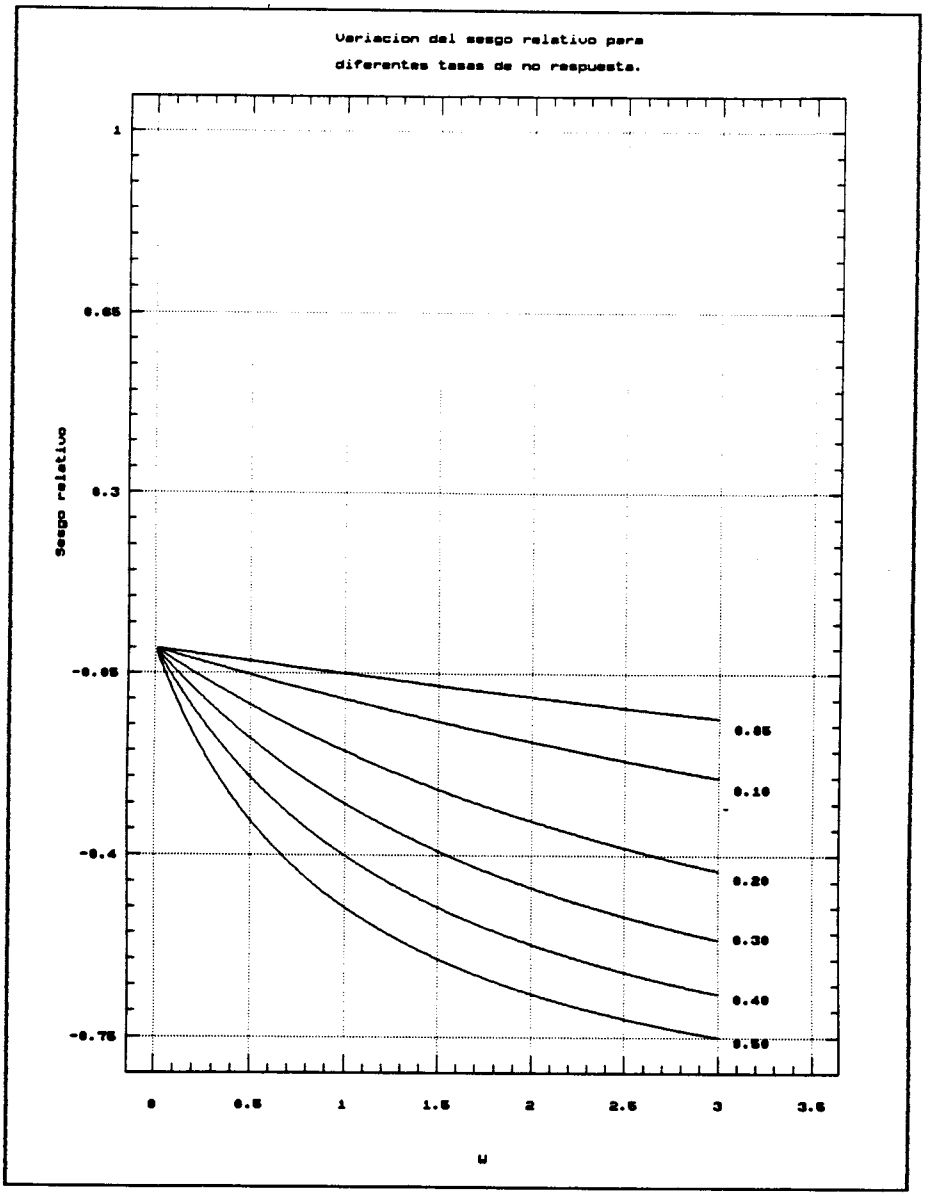
Para el caso en que estimemos que las respuestas de los que no responden son iguales a las de los que responden, el sesgo relativo adquiere el valor de la tasa de no respuesta

Por otro lado, si en el Gráfico I.2 representamos la expresiòn obtenida en la proposiciòn 3.3.13, referente al sesgo relativo del estimador del total poblacional :

$$SR(Y_n) = -1.\ t_{N2}\ W \,/\, (1-\ t_{N2}(1-W))$$

Fijando t_{N2} (utilizamos tasas del 5%, 10%, 20%, 30%, 40% y 50%) y variando W en el intervalo (0,3) con incrementos de 0.2, obtenemos diferentes gráficas correspondientes a la expresiòn antes indicada, según la tasa de de no respuesta utilizada.

Gráfico I.2



Observamos que :

* Variando W, podemos pasar de una tasa a otra inferior, sin variar el sesgo.

* De igual forma, si conocemos a priori el valor de la tasa de no respuesta, podemos obtener la magnitud del sesgo relativo con la ayuda del gràfico, siempre que conozcamos por otras encuestas anàlogas el valor de W.

* Finalmente, si fijamos un valor para que el sesgo, es posible determinar los lìmites en la tasa de no respuesta necesarios para tal fin.

3.3.2. Comparación entre los sesgos relativos de los estimadores de la media y del total poblacional

Proposicion 3.3.15 A igualdad de tasa de no respuesta, para cualquier valor de $W = \bar{Y}_{N2}/\bar{Y}_{N1} > 0$, se tiene que :

$$SR(\bar{Y}_n) > SR(Y_n) \qquad (1)$$

siendo $\bar{Y}_n$ e Y_n el estimador de la media y del total poblacional, respectivamente.

Demostración :

Teniendo en cuenta las relaciones obtenidas para los sesgos relativos en las proposiciones 3.1.4 y 3.3.13, para que sea cierta la desigualdad (1), se debe verificar :

$t_{N2}\ (1-W) \ /\ (\ 1-\ t_{N2}\ (1-W)\ ) > -1.\ t_{N2}.\ W\ /\ (\ 1-\ t_{N2}\ (1-W)\ )$ ==>

$t_{N2}\ (1-W)\ .\ (\ 1-\ t_{N2}\ (1-W)\ ) > -1.\ t_{N2}.\ W\ .\ (\ 1-\ t_{N2}\ (1-W)\ )$ ==>

$1-\ W\ -\ t_{N2}\ (1-W)^2 > -\ W\ +\ t_{N2}.\ (1-W)\ .\ W$ ==>

$1 > t_{N2}.\ (1-W)\ .(1-\ W\ +\ W\ )$ ==>

$1\ -\ W < 1/t_{N2}$ ==> $W > 1\ -\ (1/t_{N2})$ ==> $W > 0$

Dado que $1/t_{N2} > 0$, se sigue que $1 - 1/t_{N2} < 0$ y en definitiva la desigualdad (1) se verifica para $W > 0$.

c.q.d.

Proposiciòn 3.3.16 A igual valor de W>1, el sesgo relativo que obtenemos al estimar la media poblacional, con una tasa t_{N2}, es el mismo que el que obtenemos al estimar el total poblacional, con una tasa t_{N2}^{+}. Siendo la relaciòn entre una y otra :

$t_{N2}^{+} = t_{N2}\ .\ (1-W)\ /\ (t_{N2}.(1-W)\ -\ W)$, siendo $W = \bar{Y}_{N2}/\bar{Y}_{N1}$

Demostraciòn :

Igualando las expresiones que nos dan los sesgos relativos (proposiciones 3.1.4 y 3.3.13), se tiene

$$t_{N2}.(1-W)\ /\ (\ 1-\ t_{N2}.(1-W)\ ) = -1.t_{N2}^{+}.\ W\ /\ (\ 1-\ t_{N2}^{+}.\ (1-W)\ )$$

y trasformando, obtenemos :

$t_{N2}.(1-W) - t_{N2}.t_{N2}^{+}.(1-W)^2 = -1.t_{N2}^{+}.\ W + t_{N2}^{+}.t_{N2}.\ (1-W).W \Longrightarrow$

$t_{N2}\ (1-W) = t_{N2}\ (1-W)\ t_{N2}^{+} - t_{N2}^{+}\ .\ W$

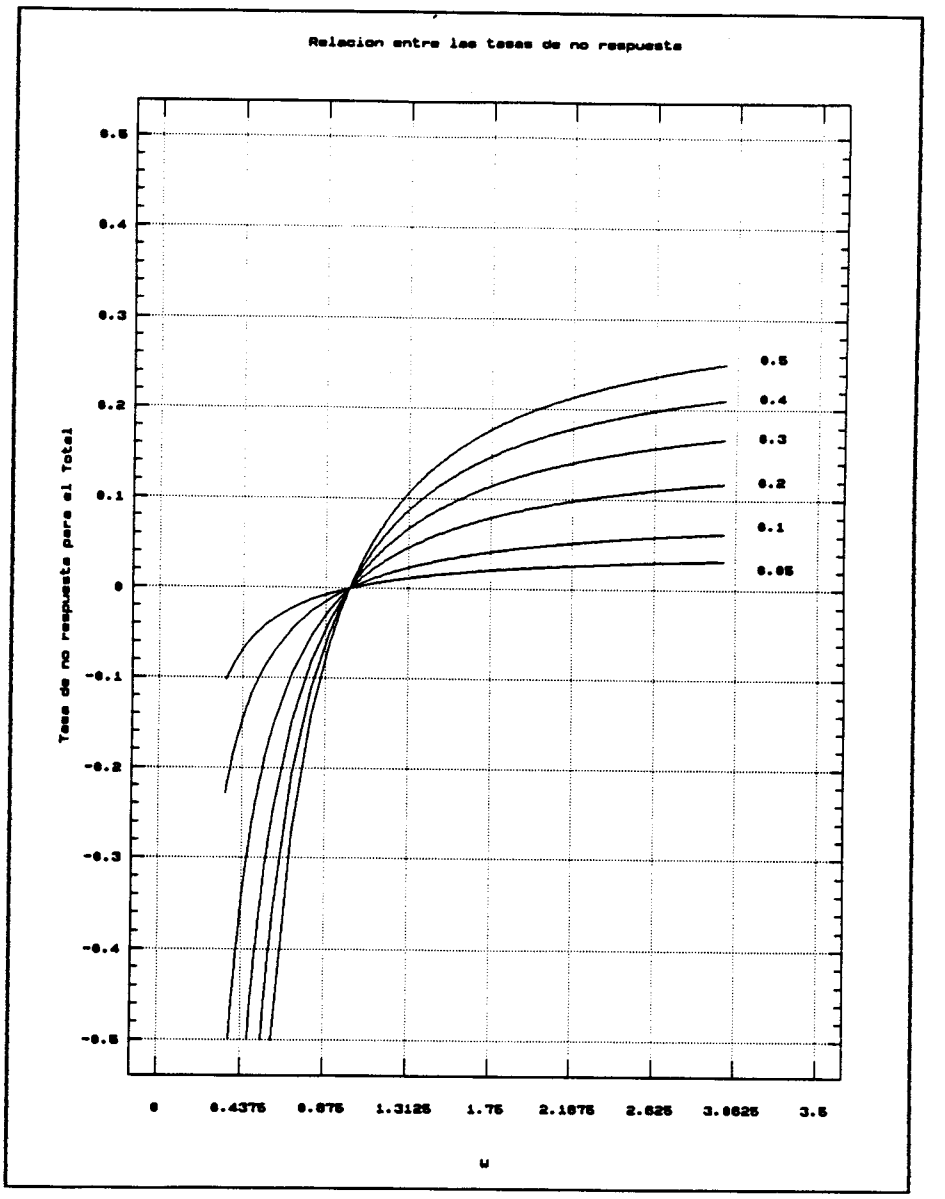
Despejando, obtenemos la expresión propuesta.

c.q.d.

3.3.3. Conclusiones

En el Gráfico I.3 hemos representado la expresiòn obtenida en la proposiciòn 3.3.16

Gráfico I.3



Fijando t_{N2} (utilizamos tasas del 5%, 10%, 20%, 30%, 40% y 50%) y variando W en el intervalo (0,3) con incrementos de 0.2, obtenemos diferentes curvas, según la tasa de no respuesta (t_{N2}) utilizada al estimar la media poblacional. En nuestro caso como la tasa de no respuesta tiene que ser igual o mayor que cero, estudiamos las funciones para valores de W>1.

De la proposición 3.3.15 concluimos que, frente a la no respuesta y con referencia al sesgo, es mejor estimar el total que la media poblacional siempre que se empleé la misma tasa de respuesta, dado que el sesgo relativo del estimador de la media es mayor que el del estimador del total.

Sin embargo, de la proposición 3.3.16 concluimos que, frente a la no respuesta y con referencia al sesgo, es lo mismo estimar el total que la media, siempre que se utilice, unas tasas de no respuesta distintas, en principio, una de la otra y que verifiquen la relación :

$$t_{N2}^{+} = t_{N2} \cdot (1-W) / (t_{N2}.(1-W) - W) \ , \quad \text{siendo } W = \bar{Y}_{N2}/\bar{Y}_{N1}$$

3.4. En la estimaciòn de razones

Con frecuencia encontramos situaciones en las que se cree que la tasa de Y con respecto a otra característica Z es menos variable que la misma Y. En éste caso sería mejor estimar R_N, la tasa de Y a Z en la población, a través de la muestra y después multiplicar por el total conocido de Z, para estimar el total de Y.

Definición 3.4.17 Siendo Y y Z dos características de una población, definimos razón de Y a Z en la población como el cociente :

$R_N = Y_N / Z_N$, siendo Y_N , Z_N los totales poblacionales de Y y Z

Proposición 3.4.18 Siendo R_n un estimador muestral de R_N, tal que $E(R_n) = R_{N1}$, razón poblacional de los que responden, la cantidad de sesgo al estimar R_N por R_n, viene determinado por :

$$\text{sesgo}(R_n) = C_{N2} [R_{N1} - R_{N2}] \ , \text{ donde } C_{N2} = Z_{N2} / Z_N$$

Demostración :

Dado que sesgo = $E(R_n) - R_N = R_{N1} - R_N =$

$$(Y_{N1} / Z_{N1}) - (Y_N / Z_N) =$$

$$(Y_{N1} / Z_{N1}) - (Y_{N1} + Y_{N2})/ Z_N =$$

$$[Y_{N1}(Z_N - Z_{N1}) - (Z_{N1}. Y_{N2})]/(Z_N . Z_{N1}) =$$

$$(Y_{N1}.Z_{N2} - Z_{N1}. Y_{N2})/(Z_N.Z_{N1}) =$$

$$(Y_{N1}.Z_{N2})/(Z_N.Z_{N1}) - (Z_{N1}. Y_{N2})/(Z_N.Z_{N1}) =$$

$$(Z_{N2} / Z_N) .[(Y_{N1} / Z_{N1}) - (Y_{N2} / Z_{N2})] \quad \text{c.q.d.}$$

3.5. En la estimación por intervalo

La estimación por intervalo comprende los del total y media poblacionales, tomando como referencia los estimadores muestrales respectivos. Corresponde al intervalo formulado en la siguiente forma :

$$P\ [\ |\bar{Y}_N - \bar{Y}_n| \leq \delta\ .\ S(\bar{Y}_n)\ ] = 1-\alpha = P_\alpha \qquad [1]$$

supuesto que la estimación fuera para la media poblacional $\bar{Y}_N$, el estimador media muestral, $\bar{Y}_n$, insesgado y el error máximo admisible, que representa la precisión mínima a exigir de los resultados, de tamaño igual a $\delta\ .\ S(\bar{Y}_n)$; siendo δ una constante que depende del nivel de significación α y $S(\bar{Y}_n)$ el error estándar de la distribución de muestreo del estimador muestral, definido como la raiz cuadrada de la varianza de éste.

Por consiguiente, de [1] deducimos que el intervalo de confianza con un nivel de significación α, para el parámetro media poblacional, tendrá la forma :

$$[\ \bar{Y}_n - \delta\ .\ S(\bar{Y}_n)\ ,\ \bar{Y}_n + \delta\ .\ S(\bar{Y}_n)\ ] \qquad (1)$$

Siendo δ un número real, que dependerá de la distribución muestral de la media de la muestra, de tal manera que, si el tamaño n de esta es grande, se puede aproximar por el cuantil $Z_{\alpha/2}$ de una distribución N(0,1).

Para el caso en que desconozcamos $VAR(\bar{Y}_n)$ y utilicemos su estimación $\overset{*}{VAR}(\bar{Y}_n)$, entonces los intervalos de confianza con un nivel de significación α tendrán la forma :

$$[\ \bar{Y}_n - t_{\alpha/2,n-1}\ .\ \overset{*}{S}(\bar{Y}_n)\ ,\ \bar{Y}_n + t_{\alpha/2,n-1}\ .\ \overset{*}{S}(\bar{Y}_n)\ ]$$

siendo $\overset{*}{S}(\bar{Y}_n) = (\overset{*}{VAR}(\bar{Y}_n))^{1/2}$ y δ un número real, que dependerá de la distribución muestral de la media de la muestra, de tal manera que, si el tamaño n de esta es grande, se puede aproximar por el cuantil $t_{\alpha/2,n-1}$ de una distribución t de Student con n-1 grados de libertad.

Para el caso en que el estimador media muestral fuera sesgado con sesgo absoluto B, los intervalos de confianza para la media poblacional corresponderán a intervalos formulados en la forma :

$$P\ [\ -a \leq \bar{Y}_N - \bar{Y}_n \leq a\ ] = \text{nivel de confianza}$$

siendo a una constante que depende de $S(\bar{Y}_n)$

Para concretar éste tipo de intervalos, dos son los caminos, según que el error máximo admisible dependa del error estándar o del error total, definido com la raiz cuadrada del error cuadrático medio. Ambos intervalos quedarán formulados en la forma :

$$P\ [\ |\bar{Y}_N - \bar{Y}_n| \leq \delta\ .S(\bar{Y}_n)\ ]\ =\ 1-\beta\ =\ P_\beta \qquad [2]$$

$$P\ [\ |\bar{Y}_N - \bar{Y}_n| \leq \delta\ .\ (\ ECM(\bar{Y}_n)\ )^{1/2}\ ]\ =\ 1-\Gamma\ =\ P_\Gamma \qquad [3]$$

con el mismo significado para δ y $S(\bar{Y}_n)$ que el indicado anteriormente.

En lo que sigue procederemos a concretar los intervalos de confianza [2] y [3], formulando los intervalos equivalentes para el estimador $\bar{Y}_n \pm B$, que es insesgado (Proposición 3.5.19). Obteniendo a partir de ellos las relaciones de orden existentes entre α, β y Γ. Posteriormente plantearemos en qué medida es posible sustituir [2] ó [3] por un intervalo [1]; esto es, en qué medida es posible enunciar intervalos de confianza como si no existiera ningún sesgo.

3.5.1. Intervalo de confianza para un estimador sesgado con un error máximo admisible dependiendo del error estándar

Denotemos por B el valor absoluto del sesgo : $B = |E(\bar{Y}_n) - \bar{Y}_N|$

Proposisición 3.5.19 Si $\bar{Y}_n$ es un estimador sesgado de $\bar{Y}_N$, un estimador insesgado será

a) $\bar{Y}_n - B$, si el sesgo es positivo

b) $\bar{Y}_n + B$, si el sesgo es negativo

Además, $VAR(\bar{Y}_n-B) = VAR(\bar{Y}_n+B) = VAR(\bar{Y}_n)$

Demostración :

a) Si el sesgo es positivo :

$$E(\bar{Y}_n) - \bar{Y}_N > 0 \Longrightarrow E(\bar{Y}_n) > \bar{Y}_N \Longrightarrow E(\bar{Y}_n) - B = \bar{Y}_N \Longrightarrow E(\bar{Y}_n - B) = \bar{Y}_N$$

b) Si el sesgo es negativo :

$$E(\bar{Y}_n) - \bar{Y}_N < 0 \Longrightarrow E(\bar{Y}_n) < \bar{Y}_N \Longrightarrow E(\bar{Y}_n) + B = \bar{Y}_N \Longrightarrow E(\bar{Y}_n + B) = \bar{Y}_N$$

La relación entre las varianzas es consecuencia de la linealidad.

c.q.d.

Proposición 3.5.20 Si $\bar{Y}_n$ es un estimador sesgado de $\bar{Y}_N$, el intervalo de confianza, con un nivel de significación β y un error máximo admisible de tamaño igual a $Z_{\alpha/2}.S(\bar{Y}_n)$, tiene la forma :

$$[\ \bar{Y}_n - Z_{\alpha/2} \cdot S(\bar{Y}_n)\ ,\ \bar{Y}_n + Z_{\alpha/2} \cdot S(\bar{Y}_n)\ ] \quad (2)$$

Así mismo, se verifica :

a) si el sesgo es positivo

$$P[-Z_{\alpha/2} - B/S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2} - B/S(\bar{Y}_n] = 1-\beta$$

b) si el sesgo es negativo

$$P[-Z_{\alpha/2} + B/S(\bar{Y}_n)] \leq Z \leq Z_{\alpha/2} + B/S(\bar{Y}_n)] = 1-\beta$$

siendo $Z \div N(0,1)$

Demostración :

La primera parte es consecuencia de la propia formulación del intervalo [2]; esto es :

$$P\ [\ |\bar{Y}_N - \bar{Y}_n| \leq \delta\ .S(\bar{Y}_n)\ ] = 1-\beta = P_\beta \quad ==>$$

$$P\ [\ -\delta.S(\bar{Y}_n) \leq \bar{Y}_N - \bar{Y}_n \leq \delta\ .S(\bar{Y}_n)\ ] = 1-\beta = P_\beta$$

tomando $\delta = Z_{\alpha/2}$, se obtiene la expresión (2) propuesta.

La segunda parte, la demostraremos considerando que el sesgo es negativo, siendo similar el otro supuesto. Partiendo de (2)

$$P[\ \bar{Y}_n - Z_{\alpha/2}\ .S(\bar{Y}_n) \leq \bar{Y}_N \leq \bar{Y}_n + Z_{\alpha/2}\ .S(\bar{Y}_n)\ ] = 1-\beta =$$

$$P[\ \bar{Y}_n + B - Z_{\alpha/2}\ .S(\bar{Y}_n) - B \leq \bar{Y}_N \leq \bar{Y}_n + B + Z_{\alpha/2}\ .S(\bar{Y}_n) - B\ ] =$$

$$P[\ -\ Z_{\alpha/2}\ .S(\bar{Y}_n)\ + B \leq (\bar{Y}_n+B) - \bar{Y}_N \leq\ Z_{\alpha/2}\ .S(\bar{Y}_n) + B\ ] =$$

dado que $\bar{Y}_n+B$ es un estimador insesgado y como consecuencia del teorema central del límite : $Z = [(\bar{Y}_n+B) - \bar{Y}_N]/S(\bar{Y}_n) \div N(0,1)$, se sigue :

$$P[\ -Z_{\alpha/2} + B/S(\bar{Y}_n) \leq\ Z\ \leq\ Z_{\alpha/2} + B/S(\bar{Y}_n)\ ] = 1-\beta$$

c.q.d.

A. <u>Variación de los extremos de la región de error para un estimador sesgado</u>

Una vez que tenemos el intervalo de confianza para el estimador sesgado con error máximo admisible dependiendo del error estándar, procedemos a definir la región de error y la de probabilidad asociada a éste.

Así mismo comprobamos, experimentalmente, que el nivel de significación no es el mismo que si el estimador fuera insesgado. Posteriormente concretaremos ésta última afirmación y la demostraremos.

Definición 3.5.21 La región de probabilidad y la región de error, constituida por las dos colas a derecha e izquierda, asociadas a un estimador insesgado o sesgado, son respectivamente :

Para $\bar{Y}_n$ insesgado :

$$[\ -Z_{\alpha/2}\ ,\ Z_{\alpha/2}\ ]\ ;\ Z \leq -Z_{\alpha/2} \quad y \quad Z \geq Z_{\alpha/2}$$

Para $\bar{Y}_n$ sesgado, con sesgo negativo :

$$[-Z_{\alpha/2} + B\ /\ S(\bar{Y}_n),\ Z_{\alpha/2} + B\ /\ S(\bar{Y}_n)\ ]$$

$$Z \leq -Z_{\alpha/2} + B\ /\ S(\bar{Y}_n) \quad y \quad Z \geq Z_{\alpha/2} + B\ /\ S(\bar{Y}_n)$$

Para $\bar{Y}_n$ sesgado, con sesgo positivo :

$$[\ -Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ ,\ Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ ]$$

$$Z \leq -Z_{\alpha/2} - B\ /\ S(\bar{Y}_n) \quad y \quad Z \geq Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)$$

De acuerdo con ésta definición, para el estimador sesgado y de acuerdo con la proposición 3.5.20 , su región de error tendrá una masa de probabilidad igual a ß y para el estimador insesgado, de acuerdo con lo indicado al principio de éste apartado, igual a α .

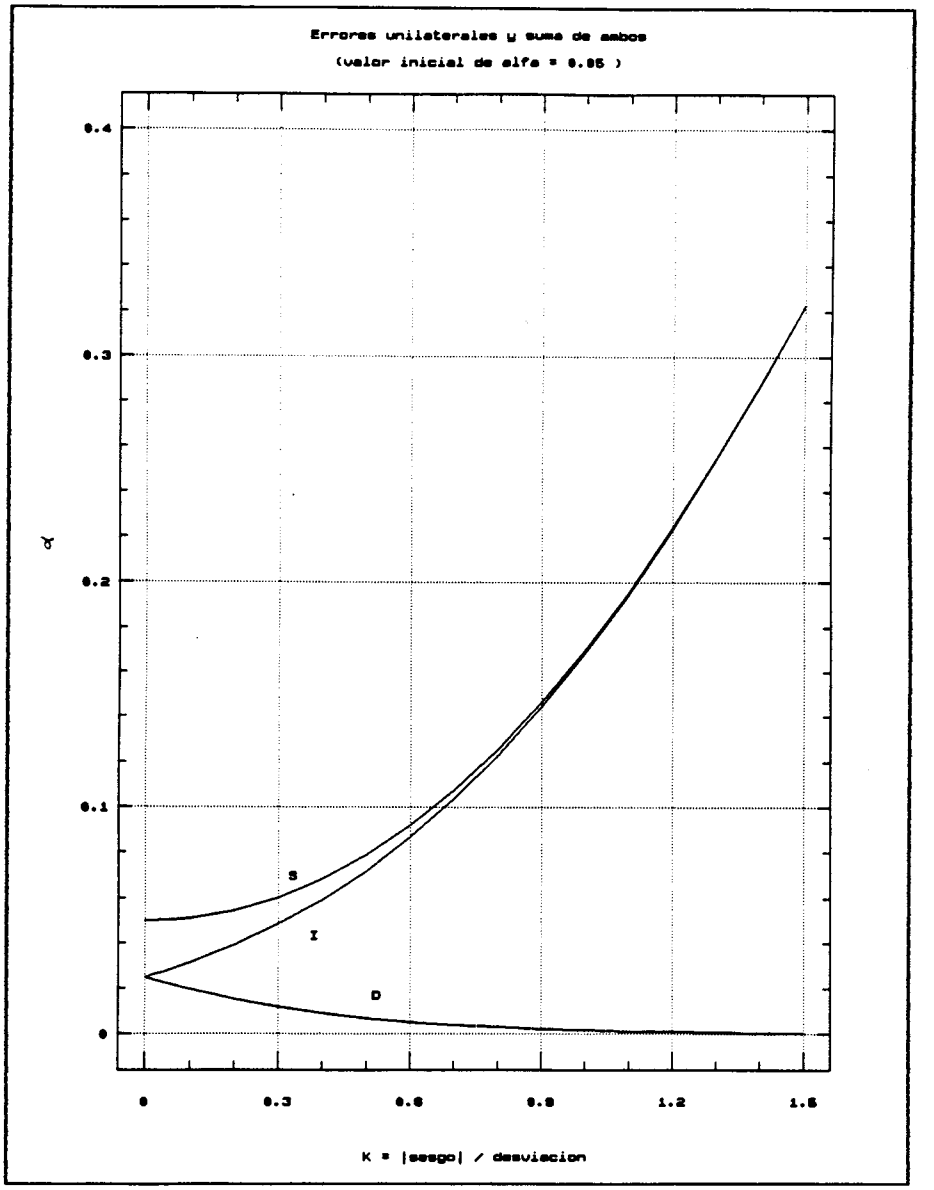
TABLA I.1

$B/S(\bar{Y}_n)$	Error a la izquierda	Error a la derecha	ß = Suma de los errores
0.00	0.0250	0.0250	0.0500
0.02	0.0262	0.0238	0.0500
0.04	0.0274	0.0228	0.0502
0.06	0.0287	0.0217	0.0504
0.08	0.0301	0.0207	0.0508
0.10	0.0314	0.0197	0.0511
0.20	0.0392	0.0154	0.0546
0.30	0.0485	0.0119	0.0604
0.40	0.0594	0.0091	0.0685
0.50	0.0721	0.0069	0.0790
0.60	0.0869	0.0052	0.0921
0.70	0.1038	0.0038	0.1076
0.80	0.1230	0.0029	0.1259
0.90	0.1446	0.0021	0.1467
1.00	0.1685	0.0015	0.1700
1.10	0.1949	0.0011	0.1960
1.20	0.2236	0.0008	0.2244
1.30	0.2546	0.0006	0.2552
1.40	0.2877	0.0004	0.2881
1.50	0.3228	0.0003	0.3231

Raj (1968) y posteriormente Cochran (1977), fijan un valor de $Z_{\alpha/2}$ y variando el cociente $B/S(\bar{Y}_n)$ obtienen experimentalmente que la región del error aumenta conforme aumentamos éste cociente. Cochran obtiene como regla de trabajo que el efecto del sesgo en la exactitud de un estimador es despreciable, si el cociente antes indicado es menor que 0.1, aunque indica que es posible llegar al valor 0.2; Raj, llega a la conclusión de que éste es menor que 0.1.

En la tabla I.1 se observa tal afirmación, considerando que el sesgo es negativo y que $Z_{\alpha/2}$ = 1.96 (nivel de confianza = 95%).

Gráfico I.4



En el Gráfico I.4 hemos representado los valores de la tabla I.1, y en el puede observarse que una región de error (cola izquierda) va aumentando y la otra (cola derecha) disminuye, con respecto a las regiones de error para $\bar{Y}_n$ insesgado de valor 0.025 y 0.025 . Lo generalizaremos en la Proposición 3.5.24.

El hecho de que la región del error aumente cuando $\bar{Y}_n$ es sesgado, hace que la región de probabilidad asociada disminuya, por lo que para los intervalos (2) su nivel de confianza 1-ß, es inferior a 1-α, nivel de confianza para el estimador insesgado. Lo generalizaremos en la Proposición 3.5.25 .

B. Generalización

En lo que sigue demostramos las consecuencia obtenidas experimentalmente

Definición 3.5.22 Definimos Φ (t), como la masa de probabilidad en el intervalo [0,t] tal que Φ (t) + 1/2 es la función de densidad de una variable aleatoria N(0,1).

Proposición 3.5.23 Fijado un nivel de confianza de 1-α para el estimador insesgado, se tiene que, para un sesgo $B/S(\bar{Y}_n)$, el nivel de confianza 1-ß para el estimador sesgado, con sesgo positivo o negativo, es :

$$P_{ß} = 1 - ß = \Phi(Z_{\alpha/2} - B/S(\bar{Y}_n)) + \Phi(Z_{\alpha/2} + B/S(\bar{Y}_n))$$

Demostración :

De la definición 3.5.21 y la proposición 3.5.20, la probabilidad para la región de probabilidad asociada a un estimador sesgado es

$$P [-Z_{\alpha/2} + B / S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2} + B / S(\bar{Y}_n)] = 1-ß \quad ó$$

$$P [-Z_{\alpha/2} - B / S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2} - B / S(\bar{Y}_n)] = 1-ß$$

Aplicando la definición 3.5.22 y teniendo en cuenta la simetría de la distribución normal, se sigue la expresión propuesta.

Proposición 3.5.24 Siendo B el valor absoluto del sesgo asociado al estimador $\bar{Y}_n$, las colas de la región de error (Definición 3.5.21) asociadas al estimador insesgado y las del estimador sesgado, son tales que :

a) Cola a la izquierda :

$$P [Z \leq -Z_{\alpha/2} + B / S(\bar{Y}_n)] \geq P [Z \leq -Z_{\alpha/2}] \text{ (sesgo negativo)}$$

$$P\ [\ Z \le -Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ ]\ \le\ P\ [\ Z \le -Z_{\alpha/2}\ ]\ \text{(sesgo positivo)}$$

b) Cola a la derecha :

$$P\ [\ Z \ge Z_{\alpha/2} + B\ /\ S(\bar{Y}_n)\ ] \le\ P\ [\ Z \ge Z_{\alpha/2}\ ]\ \text{(sesgo negativo)}$$

$$P\ [\ Z \ge Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ ] \ge\ P\ [\ Z \ge Z_{\alpha/2}\ ]\ \text{(sesgo positivo)}$$

Demostración :

(Planteamos la hipótesis del sesgo negativo, siendo la demostración análoga si la hipótesis es la de que el sesgo sea positivo)

a) Teniendo en cuenta la simetría de la distribución normal y la definición 3.5.22

$$P[\ Z \le -Z_{\alpha/2} + B\ /\ S(\bar{Y}_n)\ ]\ =\ P\ [\ Z \ge Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ ] =$$

$$1/2 - \Phi\ (\ Z_{\alpha/2}\ - B\ /\ S(\bar{Y}_n)\ )$$

Por otro lado :

$$P\ [\ Z \le -\ Z_{\alpha/2}\ ]\ = P\ [\ Z \ge Z_{\alpha/2}\ ]\ = 1/2 - \Phi\ (\ Z_{\alpha/2}\ )$$

Como :

$\Phi\ (\ Z_{\alpha/2}\ ) \ge\ \Phi\ (\ Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ )$, pues $Z_{\alpha/2}\ \ge Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)$

se sigue que :

$$1/2 - \Phi\ (\ Z_{\alpha/2}\ ) \le 1/2 - \Phi\ (\ Z_{\alpha/2} - B\ /\ S(\bar{Y}_n)\ )$$

b) es análoga

c.q.d.

Proposición 3.5.25 La probabilidad ß asociada a la región de error para el estimador sesgado es mayor que α que es la del estimador insesgado.

Demostración :

Teniendo en cuenta la definición 3.5.21, la proposición 3.5.20 y considerando como hipòtesis que el sesgo es negativo, demostramos la siguiente desigualdad :

$$P\ [-Z_{\alpha/2} + B/S(\bar{Y}_n) \le Z \le Z_{\alpha/2} + B/S(\bar{Y}_n)] \le P[-\ Z_{\alpha/2} \le Z \le Z_{\alpha/2}\ ]$$

que demuestra la proposición propuesta.

Utilizando la definición 3.5.22, podemos escribirla como :

$$\Phi(Z_{\alpha/2} - B/S(\bar{Y}_n)) + \Phi(Z_{\alpha/2} + B/S(\bar{Y}_n)) \leq 2\, \Phi(Z_{\alpha/2})$$

y cambiando los términos :

$$\Phi(Z_{\alpha/2} + B/S(\bar{Y}_n)) - \Phi(Z_{\alpha/2}) \leq \Phi(Z_{\alpha/2}) - \Phi(Z_{\alpha/2} - B/S(\bar{Y}_n))$$

Denotando por :

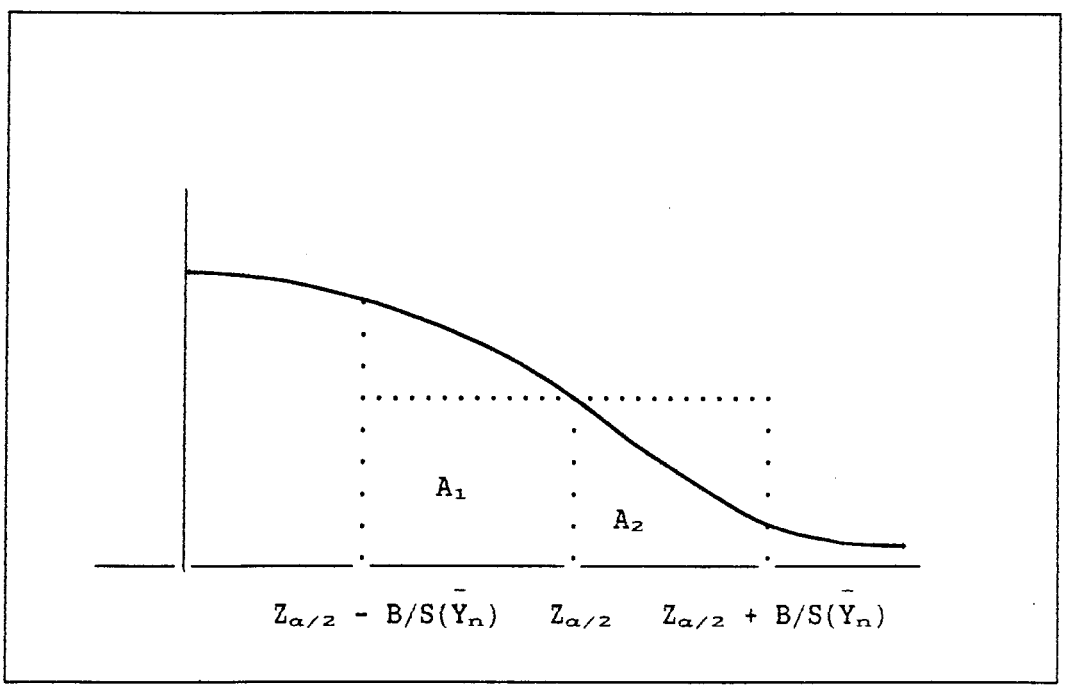
$$A_2 = \Phi(Z_{\alpha/2} + B/S(\bar{Y}_n)) - \Phi(Z_{\alpha/2}) \quad y$$

$$A_1 = \Phi(Z_{\alpha/2}) - \Phi(Z_{\alpha/2} - B/S(\bar{Y}_n))$$

Lo que hay que demostrar es que : $A_1 \geq A_2$

Dado que $Z_{\alpha/2}$, y $[Z_{\alpha/2} - B/S(\bar{Y}_n)$, $Z_{\alpha/2} + B/S(\bar{Y}_n)]$, están contenidos en él intervalo $[0, \infty)$ (se supone que $Z_{\alpha/2} \geq B/S(\bar{Y}_n)$), según se indica en el siguiente gráfico

Gráfico I.5



Se sigue :

$$A_1 = \int_{Z_{\alpha/2} - B/S(\bar{Y}_n)}^{Z_{\alpha/2}} f(t)\,.\,dt \geq f(Z_{\alpha/2})\,.\,[\,Z_{\alpha/2} - (Z_{\alpha/2} - B/S(\bar{Y}_n))\,] =$$

$$f(Z_{\alpha/2})\,.\,B/S(\bar{Y}_n) \geq \int_{Z_{\alpha/2}}^{Z_{\alpha/2} + B/S(\bar{Y}_n)} f(t)\,.\,dt = A_2$$

Es decir $A_1 \geq A_2$ c.q.d.

Una vez estudiado el intervalo de confianza para un estimador sesgado con un error máximo admisible que depende del error estándar, veamos el otro planteamiento [3] esbozado al principio de éste apartado utilizando el error total.

3.5.2. Intervalo de confianza para un estimador sesgado con un error máximo admisible dependiendo del error total

El intervalo de confianza, la región de error y el nivel de significación para estos intervalos, se contemplan en las dos proposiciones siguientes:

Proposición 3.5.26 Si $\bar{Y}_n$ es un estimador sesgado de $\bar{Y}_N$, el intervalo de confianza con un nivel de significación Γ y un error máximo admisible de tamaño igual a $Z_{\alpha/2}.(ECM(\bar{Y}_n))^{1/2}$, tiene la forma :

$$[\,\bar{Y}_n - Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2}.S(\bar{Y}_n)\ ,\ \bar{Y}_n + Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2}.S(\bar{Y}_n)\,] \quad (3)$$

Así mismo se verifica

a) si el sesgo es positivo :

$$P[-Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2} - B/S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2} - B/S(\bar{Y}_n)]$$

$$= 1 - \Gamma$$

b) si el sesgo es negativo :

$$P[-Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n)]$$

$$= 1 - \Gamma$$

Demostración :

La primera parte es consecuencia de la propia formulación del intervalo (3); esto es :

$$P\ [\ |\bar{Y}_N - \bar{Y}_n|\ \leq\ Z_{\alpha/2}\ .(ECM(\bar{Y}_n))^{1/2}\ ]\ =\ 1-\Gamma\ =\ P_\Gamma \quad (4)$$

Como $(ECM(\bar{Y}_n))^{1/2} = (\ S(\bar{Y}_n)^2 + B^2\ )^{1/2} = S(\bar{Y}_n)\ [1 + (B/S(\bar{Y}_n))^2]^{1/2}$

el error máximo admisible toma la forma :

$$Z_{\alpha/2}.\ (ECM(\bar{Y}_n))^{1/2} = [\ Z_{\alpha/2}.(\ 1 + (\ B/S(\bar{Y}_n)\ )^2\ )\ ^{1/2}\ ].S(\bar{Y}_n) = \delta.S(\bar{Y}_n)$$

que es la misma que la del error máximo admisible utilizando el error estàndar, con :

$$\delta = Z_{\alpha/2}\ .\ (\ 1 + (\ B/S(\bar{Y}_n)\ )^2\ )\ ^{1/2} \qquad (5)$$

Por lo que la expresión (4), tomará la forma

$$P\ [\ -\ \delta.S(\bar{Y}_n) \leq \bar{Y}_N - \bar{Y}_n \leq \delta\ .S(\bar{Y}_n)\ ]\ =\ 1-\Gamma\ =\ P_\Gamma$$

sustituyendo δ por el valor obtenido en (5), se obtiene la expresión (3) propuesta.

La segunda parte, la demostraremos considerando que el sesgo es negativo, siendo similar el otro supuesto. Partiendo de (3)

$$P[\bar{Y}_n - Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2} \leq \bar{Y}_N \leq \bar{Y}_n + Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2}]\ =$$

$$P[\bar{Y}_n+B-Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2}-B \leq \bar{Y}_N \leq \bar{Y}_n+B+Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2}-B]\ =$$

$$P[-Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2}+B \leq(\bar{Y}_n+B)-\bar{Y}_N \leq Z_{\alpha/2}.(1+(\ B/S(\bar{Y}_n)\ )^2\ )^{1/2}+B]\ =\ 1-\Gamma$$

dado que $\bar{Y}_n+B$ es un estimador insesgado y como consecuencia del teorema central del límite : $Z = [(\bar{Y}_n+B) - \bar{Y}_N]/S(\bar{Y}_n) \div N(0,1)$, se sigue :

$$P[-Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2}+B/S(\bar{Y}_n) \leq Z \leq Z_{\alpha/2}.(1+(B/S(\bar{Y}_n))^2)^{1/2}+B/S(\bar{Y}_n)]\ =\ 1-\Gamma$$

c.q.d.

Proposición 3.5.27 Si $\bar{Y}_n$ es un estimador sesgado de $\bar{Y}_N$, la región de error, constituida por dos colas a derecha e izquierda, asociada a un intervalo de confianza con un error máximo admisible de tamaño igual a $Z_{\alpha/2}.(\ ECM(\bar{Y}_N)\ )^{1/2}$, son :

a) si el sesgo es positivo :

$$Z \leq\ -\ Z_{\alpha/2}.(\ 1 + (\ B/S(\bar{Y}_n)\ )^2\ )^{1/2} - B/S(\bar{Y}_n) \qquad y$$

$$Z \geq\ Z_{\alpha/2}.(\ 1 + (\ B/S(\bar{Y}_n)\ )^2\ )^{1/2} - B/S(\bar{Y}_n)$$

b) si el sesgo es negativo :

$$Z \leq - Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n) \quad y$$

$$Z \geq Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n)$$

y el nivel de confianza para cualquiera de ellos de :

$$P_\Gamma = 1 - \Gamma = \Phi (Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} - B/S(\bar{Y}_n)) +$$

$$\Phi (Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n))$$

Demostración :

Consecuencia de la proposición anterior, teniendo en cuenta la simetría de la distribución normal y la definición 3.5.22

c.q.d.

El estudio de las regiones de error vamos a realizarlo, experimentalmente, para un caso particular de t y luego generalizar las conclusiones, al igual que en el apartado 3.5.1 .

En lo que sigue, el apartado A es teórico, el apartado B son conclusiones experimentales y el apartado C la generalización de éstas.

A. Estudio de los extremos de la región de error

Los extremos de la región de error contemplados en la proposición anterior no son funciones lineales, esto es, para un valor de $Z_{\alpha/2}$, aumentan o disminuyen en función del valor del cociente $B/S(\bar{Y}_n)$ y, en consecuencia, la región del error aumenta o disminuye.

Consideremos la región de error para el sesgo negativo. Si denotamos por Y_1 e Y_2 los valores que toma el extremo inferior y superior, respectivamente, de la región de error, estudiemos su variación.

De acuerdo con la proposición 3.5.27, escribimos :

$$Y_1 = - Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n) = - t .(1+K^2)^{1/2} + K$$

$$Y_2 = Z_{\alpha/2}.(1 + (B/S(\bar{Y}_n))^2)^{1/2} + B/S(\bar{Y}_n) = t .(1+K^2)^{1/2} + K$$

Fijamos un valor para $t = Z_{\alpha/2} > 1$ (que equivale a un $\alpha < 0.3174$) y variamos el valor de $K = B/S(\bar{Y}_n)$ ($K \geq 0$). Se tiene :

* La función $Y_1 = - t .(1+K^2)^{1/2} + K$,

1) presenta un máximo en el punto A ($1/(t^2-1)^{1/2}$, $-1.(t^2-1)^{1/2}$), que existe para valores de t > 1. Esta circunstancia nos hará limitar el estudio de estas funciones para valores mayores del valor de t antes indicado. Denotaremos por K_1 la abcisa del punto A.

2) corta al eje de ordenadas en el punto B (0, -t)

3) presenta otro punto, de igual ordenada que el anterior, de abcisa

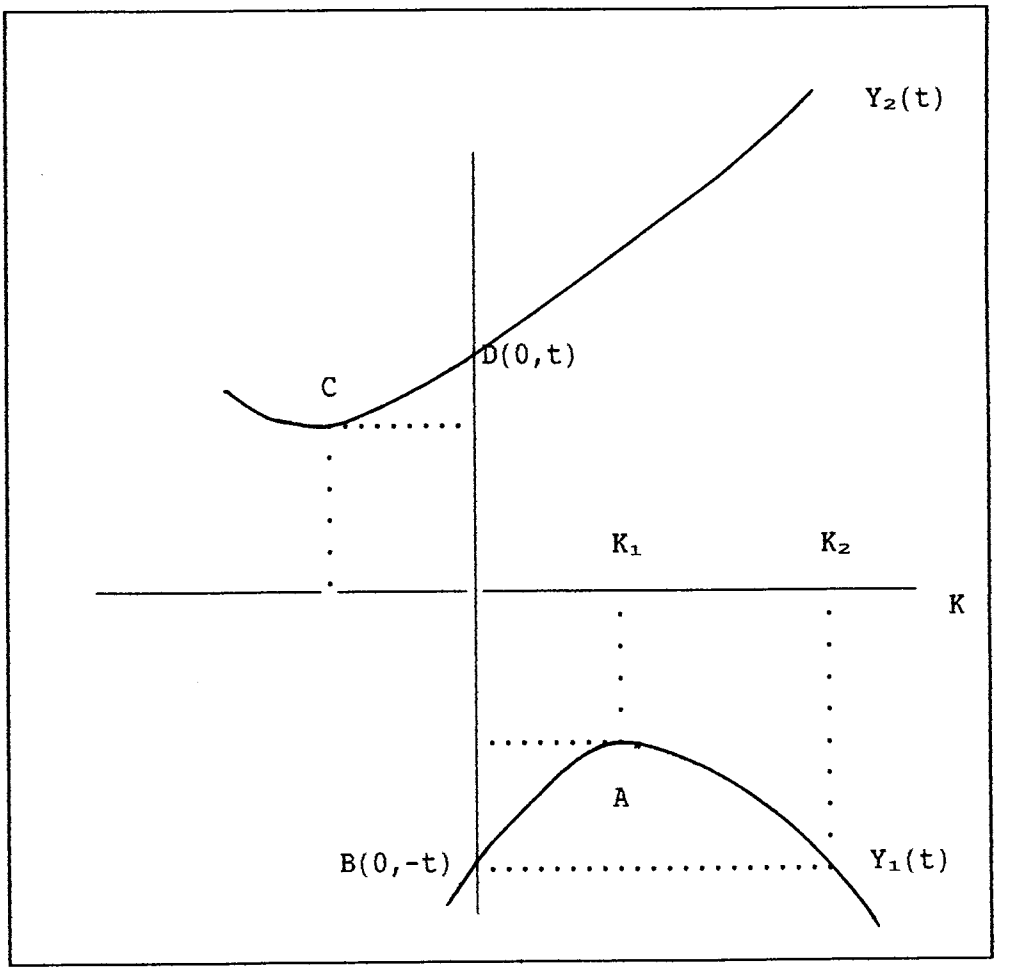
$$K_2 = 2t/t^2-1$$

4) para un valor de K, Si $t_1 < t_2 \Rightarrow Y_1(t_1) > Y_1(t_2)$, dado que :

$$t_1 < t_2 \Rightarrow -t_1 > -t_2 \Rightarrow -t_1.(1+K^2)^{1/2} > -t_2.(1+K^2)^{1/2} \Rightarrow$$

$$-t_1.(1+K^2)^{1/2} + K > -t_2.(1+K^2)^{1/2} + K$$

Gráfico I.6



A la vista de lo indicado y tomando como referencia el valor fijo $t = Z_{\alpha/2} > 1$ (en el Gràfico I.6, representamos las funciones que estamos estudiando, para una valor de t, arbitrario), podemos deducir que :

$$-t \leq Y_1 \leq -1.(t^2-1)^{1/2} \quad , \quad \text{para } 0 \leq K \leq K_1$$

$$-1.(t^2-1)^{1/2} \geq Y_1 \geq -t \quad , \quad \text{para } K_1 \leq K \leq K_2$$

$$Y_1 < -t \quad , \quad \text{para } K > K_2$$

Es decir:

La función Y_1 toma valores a la derecha de -t, cuando K varía entre el valor 0 y K_2, siendo el valor más alejado, el de la ordenada del punto para $K = K_1$.

La función Y_1 toma valores a la izquierda de -t, cuando K es mayor que K_2.

* La función $Y_2 = t.(1+K^2)^{1/2} + K$,

1) presenta un mínimo en el punto C $(-1/(t^2-1)^{1/2} , 1.(t^2-1)^{1/2})$

2) corta al eje de ordenadas en el punto D (0, t)

3) para un valor de K, Si $t_1 < t_2 \Longrightarrow Y_2(t_1) < Y_2(t_2)$, dado que :

$$t_1 < t_2 \Longrightarrow t_1.(1+K^2)^{1/2} < t_2.(1+K^2)^{1/2} \Longrightarrow$$

$$t_1.(1+K^2)^{1/2} + K < t_2.(1+K^2)^{1/2} + K$$

Tomando como referencia el valor fijo $t = Z_{\alpha/2} > 1$, se tiene :

$Y_1 > t$, para cualquier valor de K, en particular para

$K = K_1 = 1/(t^2-1)^{1/2}$ $\quad Y_2 = (t^2+1)/(t^2-1)^{1/2}$

$K = K_2 = 2t/(t^2-1)$ $\quad Y_2 = (t^3+3t)/(t^2-1)$

Por otro lado y comparando ambas funciones, Y_2 crece más deprisa que Y_1 (ver Gráfico I.6).

B. **Variación de los extremos de la región de error para un estimador sesgado**

Si denotamos por α, β y Γ las probabilidades de las regiones de error asociadas al intervalo de confianza para el estimador insesgado, sesgado con un error máximo admisible que depende del error estándar y sesgado con un error máximo admisible que depende del error total, respectivamente ¿ Cual es el comportamiento de estos errores, para un valor de α concreto?

En la tabla I.1 (pág. 21) se muestra el comportamiento del error α y β y su generalización en las proposiciones 3.5.24 y 3.5.25 .

Tabla I.2

($Z_{\alpha/2} = 1.96$; $K_1 = 0.59$; $K_2 = 1.379$)

$K=B/S(\bar{Y}_n)$	Y_1	Y_2	$I=P(Z\leq Y_1)$	$D=P(Z\leq Y_2)$	$\Gamma=I+D$
0.02	-1.94039	1.98039	0.026166	0.0238300	0.05000
0.04	-1.92157	2.00157	0.027330	0.0226660	0.05000
0.06	-1.90352	2.02352	0.028486	0.0215100	0.05000
0.08	-1.88626	2.04626	0.029630	0.0203650	0.05000
0.10	-1.86978	2.06978	0.030757	0.0192370	0.04990
0.20	-1.79882	2.19882	0.036024	0.0139460	0.04997
0.30	-1.74630	2.34630	0.040379	0.0094800	0.04986
0.40	-1.71098	2.51098	0.043542	0.0060200	0.04956
0.50	-1.69135	2.69135	0.045385	0.0035580	0.04894
0.60	-1.68573	2.88573	0.045924	0.0019525	0.04788
0.70	-1.69248	3.09248	0.045277	0.0009924	0.04627
0.80	-1.71002	3.31002	0.043631	0.0004664	0.04410
0.90	-1.73691	3.53691	0.041202	0.0002024	0.04140
1.00	-1.77186	3.77186	0.038209	0.0000810	0.03829
1.10	-1.81375	4.01375	0.034858	0.0000299	0.03489
1.20	-1.86162	4.26162	0.031328	0.0000101	0.03134
1.30	-1.91464	4.51464	0.027769	0.0000032	0.02777
1.40	-1.97211	4.77211	0.024298	0.0000009	0.02430

En la tabla I.2 se reflejan para el valor $t = Z_{\alpha/2} = 1.96$ ($\alpha = 0.05$), los valores del extremo inferior (Y_1), extremo superior (Y_2) de la zona de error, junto con el error a la izquierda (I) y derecha (D) de cada extremo antes indicado, respectivamente, y la suma (Γ) de ambos errores.

Para éste valor de $t = 1.96$, el error Γ es menor o igual que el error α; o lo que es igual, $P_\Gamma \geq P_\alpha$, siendo P_Γ y P_α las probabilidades asociadas a los intervalos de confianza del estimador sesgado y el insesgado.

¿ Ocurre ésto siempre ?. Veremos que, en general, el error Γ no es siempre menor que el error α para valores por debajo de $K = K_2$. Lo generalizaremos en la proposición 3.5.29.

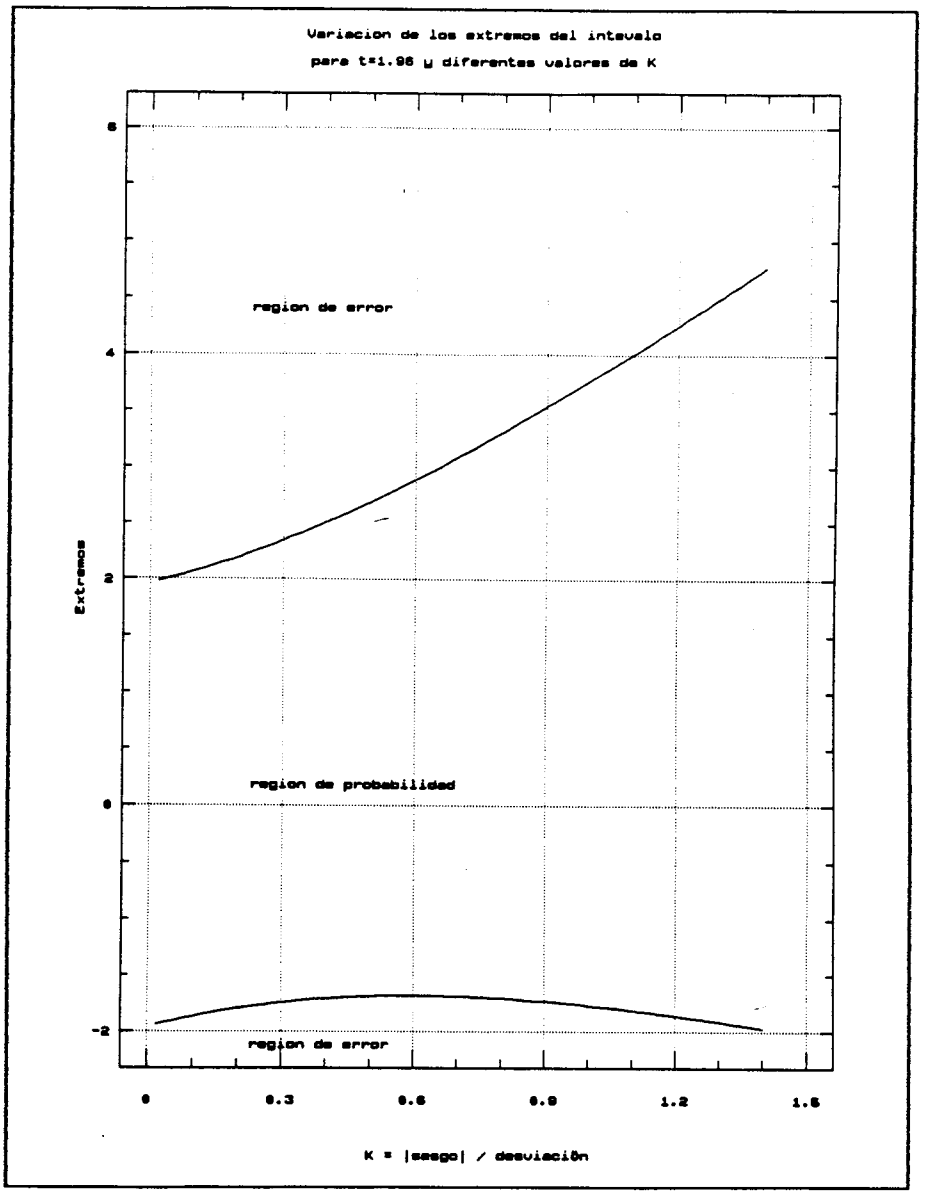
Por otro lado, si comparamos para los mismos valores de K, las tablas I.1 e I.2 en sus tres últimas columnas, observamos que el error Γ es inferior al error β. Lo generalizaremos en la proposición 3.5.30.

En el Gráfico I.7 se representan las dos primeras columna de la Tabla I.2 y observamos el comportamiento, indicado en el punto A de éste apartado, de las funciones de los extremos del intervalo :

* El extremo superior es mayor que el valor inicial $t = 1.96$

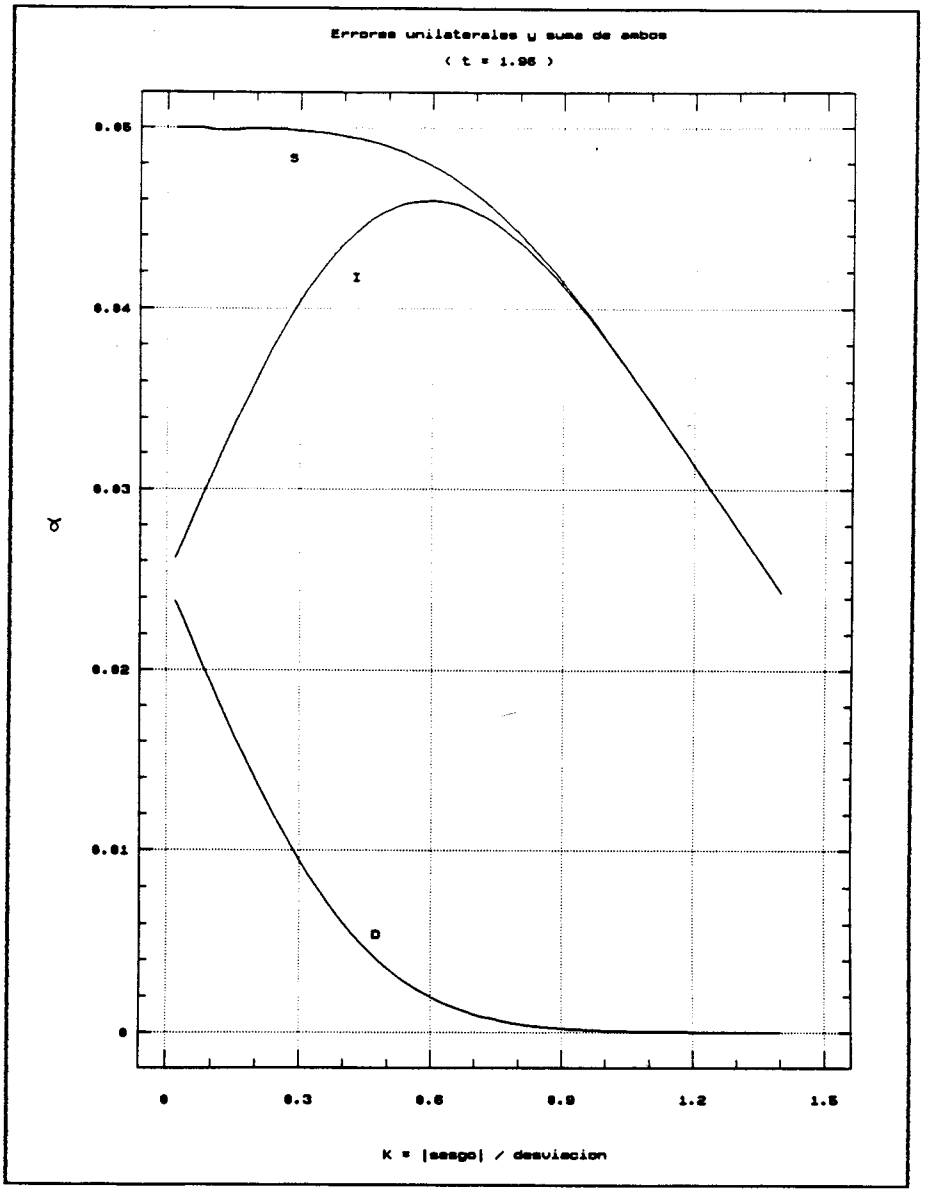
* El extremo inferior es mayor, con respecto al valor inicial t = -1.96, hasta el valor K_1 = 0.59, luego decrece y para K_2 = 1.379 toma de nuevo el valor - 1.96 y a partir de dicho valor de k, toma valores por debajo de t = -1.96

Gráfico I.7



Y en el gráfico I.8 se representan las tres últimas columnas de la tabla I.2. Pueden apreciarse diferencias de las que aparecen en el gráfico I.4.

Gráfico I.8



C. Generalización

Una vez estudiado el concepto de error asociado a un intervalo de confianza (Proposiciones 3.5.24 y 3.5.27 y principio del punto A), en lo que sigue demostramos de una forma genérica las consecuencias que en el punto B de éste apartado fueron obtenidas experimentalmente para un valor t = 1.96.

Proposición 3.5.28 Considerando que Y_1 e Y_2 representan los extremos de la región de error asociada a un intervalo de confianza con un error

máximo admisible de valor $Z_{\alpha/2}$. $(ECM(\bar{Y}_n))^{1/2}$ y que Z_1 y Z_2 representan los extremos de la región de error asociada a un intervalo de confianza con un error máximo admisible de valor $Z_{\alpha/2}$. $S(\bar{Y}_n)$, ambos para un estimador sesgado $\bar{Y}_n$, se tienen las siguientes desigualdades, según el valor de $K = B/S(\bar{Y}_n)$ y $Z_{\alpha/2} > 1$.

Para $0 \leq K \leq K_2$: $- Z_{\alpha/2} \leq Y_1 < Z_1 < Z_{\alpha/2} \leq Z_2 < Y_2$

Para $K > K_2$: $Y_1 < - Z_{\alpha/2} < Z_1 < Z_{\alpha/2} \leq Z_2 < Y_2$

Demostración :

Consideremos que el sesgo es negativo (la demostración es la misma para cuando el sesgo sea positivo).

De acuerdo con las proposiciones 3.5.20 y 3.5.27 podemos escribir, tomando $Z_{\alpha/2} = t$

$Z_1 = -t + K$ $\qquad$ $Z_2 = t + K$

$Y_1 = -t (1 + K^2)^{1/2} + K$ $\qquad$ $Y_2 = t (1 + K^2)^{1/2} + K$

El comportamiento de Z_1 y Z_2 es lineal creciente ($Z_1 < Z_2$, para cualquier valor de K) .

Z_1 crece más deprisa que Y_1, esto es :

$Z_1 > Y_1$ para cualquier valor de K, dado que, para que fuera cierta la desigualdad :

$-t + K < -t (1 + K^2)^{1/2} + K$, debería ser $(1 + K^2)^{1/2} < 0$.

Con un razonamiento análogo, podemos afirmar que $Z_2 < Y_2$.

Luego : $Y_1 < Z_1 < Z_2 < Y_2$ (1)

Por otro lado, de la propia definición de Z_1 y Z_2 y de ser $K > 0$ podemos deducir que para cualquier valor de K :

$-t < Z_1$ y $t < Z_2$ (2)

Finalmente, del estudio gráfico de las funciones Y_1 e Y_2 (punto A de éste apartado), se deduce para $t > 1$:

$$\begin{aligned} 0 \leq K \leq K_2 \quad &: \quad -t < Y_1 \quad y \quad t < Y_2 \\ K > K_2 \quad &: \quad -t > Y_1 \quad y \quad t < Y_2 \end{aligned} \qquad (3)$$

De (1), (2) y (3), se obtiene las desigualdades propuestas

c.q.d.

Proposición 3.5.29 Denotando por P_α, P_β, P_r las probabilidades asociadas a los intervalos de confianza correspondientes a un estimador insesgado la primera y para un estimador sesgado las dos últimas , con un error máximo admisible de valor $Z_{\alpha/2} \cdot S(\bar{Y}_n)$,para los dos primeros, y $Z_{\alpha/2} \cdot (ECM(\bar{Y}_n))^{1/2}$, para el último, y siendo $K = B/S(\bar{Y}_n)$ variable y $t = Z_{\alpha/2}$ un valor fijado ($t > 1$) , se tiene :

Para valores de $K \geq K_2$:

$$P_\beta \leq P_\alpha \leq P_r$$

Para valores de $K < K_2$:

$$P_\beta \leq P_\alpha \quad y \quad P_\beta \leq P_r$$

$$P_r \leq P_\alpha \quad ó \quad P_\alpha \leq P_r \quad ,$$ según el valor de t que estemos utilizando

Demostración :

Dado que $[-Z_{\alpha/2}, Z_{\alpha/2}]$, $[Z_1, Z_2]$, $[Y_1, Y_2]$ son las regiones asociadas a las probabilidades P_α, P_β, P_r, respectivamente (Proposiciones 3.5.23 y 3.5.27), se sigue de las proposiciones 3.5.25 y 3.5.28 :

$P_\beta \leq P_\alpha$, para cualquier valor de K

$P_\beta \leq P_r$, pues $[Z_1, Z_2]$ está contenido en $[Y_1, Y_2]$, para cualquier valor de K

$P_\alpha \leq P_r$, pues $[-Z_{\alpha/2}, Z_{\alpha/2}]$ está contenido en $[Y_1, Y_2]$, para valores de $K > K_2$

Finalmente para comprobar la dualidad $P_r \leq P_\alpha$ ó $P_\alpha \leq P_r$ para valores de $K < K_2$, podemos hacerlo experimentalmente :

En la Tabla I.3 comparamos los valores obtenidos para $t = 1.96$ y $t = 1.5$. Pueden comprobarse cada una de las dos posibilidades antes aludidas, para un mismo valor de K.

Por ejemplo : Para valores $0.02 \leq K \leq 0.80$, para $t = 1.96$, $P_r > P_\alpha$ y para $t = 1.5$ es $P_r < P_\alpha$

Tabla I.3

($Z_{\alpha/2}$ = 1.96; P_α = 0.95000) ($Z_{\alpha/2}$ = 1.5; P_α = 0.8663)

K	P_Γ	$P_\Gamma - P_\alpha$	P_Γ	$P_\Gamma - P_\alpha$
0.02	0.95000	0.000000	0.86639	- 0.0000000
0.04	0.95000	0.000000	0.86639	- 0.0000001
0.06	0.95000	0.000000	0.86639	- 0.0000003
0.08	0.95000	0.000001	0.86638	- 0.0000010
0.10	0.95001	0.000002	0.86638	- 0.0000024
0.20	0.95003	0.000026	0.86635	- 0.0000346
0.30	0.95014	0.000136	0.86623	- 0.0001514
0.40	0.95044	0.000434	0.86600	- 0.0003875
0.50	0.95106	0.001052	0.86567	- 0.0007116
0.60	0.95212	0.002120	0.86538	- 0.0010090
0.70	0.95373	0.003727	0.86528	- 0.0011040
0.80	0.95590	0.005899	0.86559	- 0.0007980
0.90	0.95860	0.008592	0.86646	0.0000790
1.00	0.96171	0.011706	0.86802	0.0016380
1.10	0.96511	0.015108	0.87031	0.0039230
1.20	0.96866	0.018657	0.87330	0.0069130
1.30	0.97223	0.022223	0.87693	0.0105420
1.40	0.97570	0.025696	0.88110	0.0147190

c.q.d.

Corolario 3.5.30 Denotando por α , β , Γ las probabilidades asociadas a las regiones de error de los intervalos de confianza correspondientes a un estimador insesgado, la primera, y para un estimador sesgado las dos últimas, con un error máximo admisible de valor

$Z_{\alpha/2}.S(\bar{Y}_n)$, para los dos primeros, y $Z_{\alpha/2}.(ECM(\bar{Y}_n))^{1/2}$, para el último, y siendo $K = B/S(\bar{Y}_n)$ variable y $t = Z_{\alpha/2}$ un valor fijado previamente ($t > 1$), se tiene :

Para valores de $K \geq K_2$:

$$\beta \geq \alpha \geq \Gamma$$

Para valores de $K < K_2$:

$\beta \geq \alpha$ y $\beta \geq \Gamma$, siendo $\Gamma \geq \alpha$ ó $\alpha \geq \Gamma$, según el valor de t que estemos utilizando

Demostración :

Consecuencia de la proposición anterior y de verificarse $\alpha + P_\alpha = 1$, $\beta + P_\beta = 1$ y $\Gamma + P_\Gamma = 1$ c.q.d.

3.5.3. Comparación del intervalo de confianza de un estimador sesgado con el de un estimador insesgado

Hasta aquí hemos visto el comportamiento de los extremos de las regiones de error para un estimador sesgado, con dos tipos diferentes de errores máximos admisibles. Siendo $t = Z_{\alpha/2}$ y $K = B / S(\bar{Y}_n)$, hemos visto:

a) Para un estimador insesgado :

$$P[\ \bar{Y}_n - t\ .\ S(\bar{Y}_n) \le \bar{Y}_N \le \bar{Y}_n + t\ .\ S(\bar{Y}_n)\ ] = P[-t \le Z \le t] =$$

$$2.\Phi(t) = P_\alpha$$

b) Para un estimador sesgado :

$$P[\ \bar{Y}_n - t.S(\bar{Y}_n) \le \bar{Y}_N \le \bar{Y}_n + t.S(\bar{Y}_n)\ ] = P[\ -t + K \le Z \le t + K\ ] =$$

$$\Phi(t - K) + \Phi(t + K) = P_\beta$$

ó bien

$$P[\bar{Y}_n - t.\ (1+K^2)^{1/2}.\ S(\bar{Y}_n) \le \bar{Y}_N \le \bar{Y}_n + t.(1+K^2)^{1/2}.\ S(\bar{Y}_n)] =$$

$$P[\ -t.(1+K^2)^{1/2} + K \le Z \le t.(1+K^2)^{1/2} + K\ ] =$$

$$\Phi(t.(1+K^2)^{1/2}-K) + \Phi(t.(1+K^2)^{1/2}+K) = \Phi\ (Y_1) + \Phi\ (Y_2) = P_\gamma$$

Nos podemos plantear la siguiente cuestión :

¿ Es posible que la probabilidad P_γ de que un estimador particular $\bar{Y}_n$ difiera del valor que se estima en más de $\delta.(ECM(\bar{Y}_n))^{1/2}$ se pueda interpretar de la misma manera que si el estimador fuera insesgado y se empleara $\delta.S(\bar{Y}_n)$ para fijar los límites de probabilidad ?

Los autores Hansen, Hurwitz y Madow (1953) y posteriormente Kish (1965) llegan, experimentalmente, a la conclusión de que tal cuestión es posible en tanto que el sesgo del estimador no sobrepase su error estándar y en tanto interese sólo la magnitud absoluta de los errores y no su dirección.

Nosotros, en lo que sigue, proponemos :

A) Definir y estudiar la variación relativa entre los extremos Y_1 e Y_2 de la región de error asociada al intervalo de confianza de probabilidad P_γ, respecto de los extremos -t y t de la región de error asociado al intervalo de confianza de probabilidad P_α, para cualquier tipo de sesgo (positivo o negativo). De tal forma que para un valor de

$K = B/S(\bar{Y}_n)$ y, fijada una variación relativa Θ para el extremo inferior, podamos determinar un intervalo de valores para t, dentro del cual la variación entre los extremos inferiores sea menor o igual que la variación fijada.

B) Dentro del intervalo de valores para t que obtengamos, estudiar la diferencia entre P_α y P_r y las condiciones que nos permitan conocer, si dentro de éste intervalo existe un valor de t en el que $P_\alpha = P_r$ o bien, fijado un valor Ω, si existe un intervalo de valores de t, dentro del cual la diferencia $P_r - P_\alpha$ es menor que Ω.

En definitiva y como conclusión, nuestro objetivo es obtener para cada valor de K, mayor o menor que la unidad, un intervalo de valores de t dentro del cual conozcamos, a partir de unas variaciones Θ fijadas previamente, el comportamiento de los extremos Y_1 e Y_2 respecto de los extremos -t y t así como de la probabilidad contenida entre ambos extremos. Este comportamiento es el mismo que el que siguen los extremos del intervalo de confianza para el estimador sesgado frente al insesgado y en consecuencia nos permitan dar respuesta, mas concreta, a la cuestión antes planteada.

A. Variación de los extremos del intervalo asociado a P_r, respecto de los del intervalo asociado a P_α

Hemos denotado por Y_1 e Y_2 los extremos asociados a la región de probabilidad P_r, que de acuerdo con la proposición 3.5.27 toman, para un estimador con el sesgo negativo, la siguiente expresión :

$$Y_1 = - t (1 + K^2)^{1/2} + K \qquad Y_2 = t (1 + K^2)^{1/2} + K$$

y en el que $t = Z_{\alpha/2}$, $(t>1)$ y $K = B/S(\bar{Y}_n)$

En la siguiente tabla I.4, indicamos para diferentes valores de t y de K, el valor del extremo inferior, extremo superior y sus variaciones relativas V_1 y V_2, que definiremos posteriormente (Definiciones 3.5.31, 3.5.32 y Proposición 3.5.33), así como las probabilidades P_r y P_α asociadas a los intervalos de probabilidad $[Y_1,Y_2]$ y $[-t,t]$, respectivamente y su diferencia.

Como puede observarse, para valores de $K < 1$, y distintos valores de t, la diferencia entre P_α y P_r es muy pequeña (los valores de t elegidos, corresponden a valores de α de 0.1336, 0.05 y 0.01, respectivamente).

Sin embargo, en algunos casos, la variación es muy grande, por lo que el intervalo [Y_1, Y_2] difiere a nuestro entender bastante del intervalo [-t, t]. Por ejemplo :

Para K = 0.9 la región de probabilidad para el estimador sesgado segun los valores indicados de t son [-1.12 , 2.92], [-1.74 , 3.54] y [-2.56 , 4.36], respectivamente, y el del estimador insesgado, segun los mismos

valores de t, [-1.5 , 1.5] , [-1.96 , 1.96] y [-2.57 , 2.57], respectivamente. Siendo sus variaciones para el extremo inferior de aproximadamente 0.255 , 0.114 y 0.005 respectivamente.

Obviamente al poder elegir uno de estos tres intervalos nos quedaríamos con el tercero ($\alpha = 0.01$), siempre que aceptáramos que la diferencia entre las probabilidades, que es de 0.005, es pequeña. Para los otros valores de t, la diferencia de probabilidad en uno de ellos es todavía más pequeña, pero su variación y el error asociado son más grandes.

Tabla I.4

K	Y_1	Y_2	V_1	V_2	P_r	P_α	$P_r - P_\alpha$
			(t = 1.5)				
0.1	-1.40748	1.60748	0.06168	0.07165	0.86638	0.86639	-0.000002
0.3	-1.26605	1.86605	0.15597	0.24403	0.86623	0.86639	-0.000151
0.5	-1.17705	2.17705	0.21530	0.45137	0.86567	0.86639	-0.000712
0.7	-1.13098	2.53098	0.24601	0.68732	0.86528	0.86639	-0.001104
0.9	-1.11804	2.91804	0.25464	0.94536	0.86646	0.86639	0.000079
1.1	-1.12991	3.32991	0.24673	1.21994	0.87031	0.86639	0.003923
1.3	-1.16018	3.76018	0.22654	1.50679	0.87693	0.86639	0.010542
1.5	-1.20416	4.20416	0.19722	1.80278	0.88572	0.86639	0.019338
			(t = 1.96)				
0.1	-1.86978	2.06978	0.04603	0.05601	0.95001	0.95000	0.000002
0.3	-1.74630	2.34630	0.10903	0.19709	0.95014	0.95000	0.000136
0.5	-1.69135	2.69135	0.13707	0.37314	0.95106	0.95000	0.001052
0.7	-1.69248	3.09248	0.13649	0.57780	0.95373	0.95000	0.003727
0.9	-1.73691	3.53691	0.11382	0.80455	0.95860	0.95000	0.008592
1.1	-1.81375	4.01375	0.07462	1.04783	0.96511	0.95000	0.015108
1.3	-1.91464	4.51464	0.02314	1.30339	0.97223	0.95000	0.022223
1.5	-2.03344	5.03344	-0.03747	1.56808	0.97900	0.95000	0.028992
			(t = 2.57)				
0.1	-2.48282	2.6828	0.03392	0.04390	0.98983	0.98983	0.000002
0.3	-2.38316	2.9832	0.07270	0.16076	0.98999	0.98983	0.000161
0.5	-2.37335	3.3733	0.07652	0.31259	0.99081	0.98983	0.000985
0.7	-2.43708	3.8371	0.05172	0.49303	0.99253	0.98983	0.002704
0.9	-2.55758	4.3576	0.00483	0.69556	0.99472	0.98983	0.004893
1.1	-2.72058	4.9206	-0.05859	0.91462	0.99674	0.98983	0.006911
1.3	-2.91511	5.5151	-0.13429	1.14596	0.99822	0.98983	0.008392
1.5	-3.13313	6.1331	-0.21912	1.38643	0.99914	0.98983	0.009305

Por otro lado, no es necesario que K < 1 para que la diferencia entre las probabilidades sea muy pequeña. Por ejemplo :

Para K = 1.1, la diferencia entre las probabilidades es de 0.003, 0.015 y 0.007, respectivamente, y su variación para el extremo inferior es de 0.25, 0.07 y -0.06 (la variación negativa indica, segùn veremos, que el extremo de la izquierda asociado a la región de probabilidad del estimador sesgado está a la izquierda del asociado a la región de probabilidad del estimador insesgado). Al poder elegir uno de estos tres intervalos, nos quedaremos con el tercero, siempre que aceptáramos que la diferencia entre las probabilidades (0.007) es pequeña

Definición 3.5.31 Definimos entre Y_1 y $-t$, t e Y_2 , las funciones :

$$d_1 = Y_1 - (-t) = -t\,[\,(1 + K^2)^{1/2} - 1\,] + K$$
$$d_2 = Y_2 - t = t\,[\,(1 + K^2)^{1/2} - 1\,] + K \qquad (1)$$

Las funciones d_1 y d_2 dependen de K y, para un valor concreto de t, tienen la gráfica que se indica en el Gráfico I.9. En ella observamos para valores de $K > 0$ (en lo que sigue, los valores K_1 y K_2, son los obtenidos en el punto A del apartado 3.5.2) que :

* d_1 en el intervalo $[\,0,\ K_2\,]$ toma el valor máximo en el punto de abcisa K_1 y de valor : $d_1^M = t - (t^2 - 1)^{1/2}$ y el valor cero en los puntos de abcisa $K = 0$ y $K = K_2$

* d_1 en el intervalo $K > K_2$ decrece indefinidamente. De éste intervalo, hay un punto de abcisa $K = K_3$ en el que su ordenada, en valor absoluto, es la misma que la que toma en el punto de abcisa $K = K_1$, y obtenido de la forma :

$$-t\,[\,(1 + K^2)^{1/2} - 1\,] + K = -d_1^M$$

siendo el resultado $K_3 = K_2\,[\,1 + (t^2 - t/K_1)^{1/2}\,] - K_1$

* d_2 crece indefinidamente. Las distancias correspondientes a los puntos de abcisa $K = K_1$ y $K = K_2$, son respectivamente :

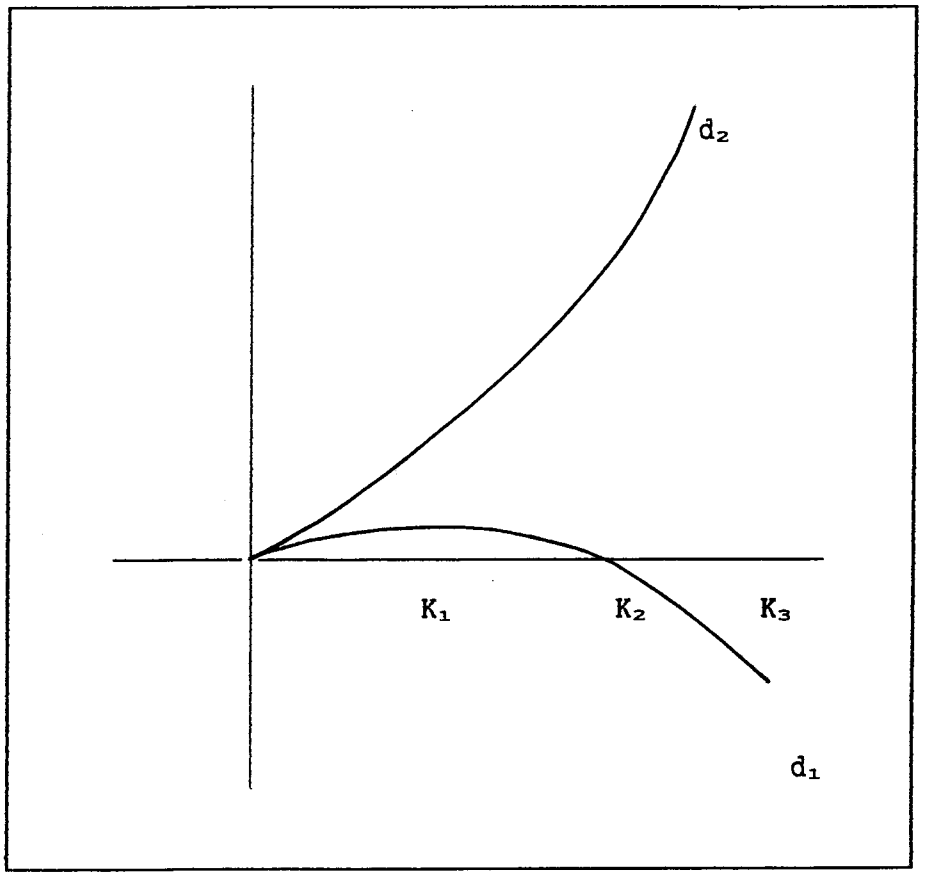
$$\frac{t^2 + 1}{t\,(t^2 - 1)^{1/2}} - 1 \quad \text{y} \quad \frac{4}{t^2 - 1}$$

Definición 3.5.32 Definimos variación relativa entre Y_1 y $-t$ y entre Y_2 y t, como los cocientes

$$V_1 = \frac{d_1}{t} \quad \text{y} \quad V_2 = \frac{d_2}{t}$$

De tal forma que la variación relativa del extremo inferior V_1 es positiva si el extremo Y_1 está a la derecha del extremo $-t$ ($Y_1 \geq -t$) y negativa si está a la izquierda ($Y_1 < -t$).

Gráfico I.9



Proposición 3.5.33 La variación relativa de los extremos de la región asociada a la probabilidad P_r, respecto de los de la región asociada a P_α, viene determinada por

Extremo inferior : $V_1 = K/t + 1 - (1 + K^2)^{1/2}$

Extremo superior : $V_2 = K/t - 1 + (1 + K^2)^{1/2}$

Demostración :

Utilizando las expresiones (1) de la definición 3.5.31 y la definición 3.5.32, se obtienen las expresiones propuestas

c.q.d.

Proposición 3.5.34 Fijado un valor de K, y siendo $t_1 < t_2$ dos valores para t, se tiene

$$V_1(t_1) > V_1(t_2) \quad y \quad V_2(t_1) > V_2(t_2)$$

siendo el recíproco tambien cierto. Además, existe un valor de t, t>1, para el que la variación en el extremo inferior es nula, siendo éste valor :

$$t = [1 + (1+K^2)^{1/2}]/K$$

Demostración :

De la proposición 3.5.33, escribimos para cada valor de t :

$$V_1(t_1) = K/t_1 + 1 - (1 + K^2)^{1/2} \qquad y$$

$$V_1(t_2) = K/t_2 + 1 - (1 + K^2)^{1/2}$$

Como $1/t_1 > 1/t_2$, se demuestra la desigualdad propuesta. De una forma análoga para V_2 y el recíproco.

En cuanto a la existencia del valor de $t > 1$, de la definición 3.5.32 y teniendo en cuenta que la función d_1 se anula en $K=0$ y $K=K_2$ (definición 3.5.31), bastará con encontrar el valor de t para el que K_2 coincida con el valor de K fijado; esto es :

$$K_2 = 2t/(t^2 - 1) = K \implies K.t^2 - 2t - K = 0 \implies t = \frac{1 \pm (1+K^2)^{1/2}}{K}$$

y nos quedamos con la solución propuesta $t = [1 + (1+K^2)^{1/2}]/K$, dado que la otra es menor que cero.

Además, la expresión propuesta es tal que : $[1 + (1+K^2)^{1/2}]/K > 1$ para cualquier $K > 0$, lo que demuestra que $t > 1$

c.q.d.

En el gráfico I.10 se representan las variaciones de los extremos inferiores de cada intervalo para diferentes valores de K y t (Tabla I.4, ampliada a todos los valores de $K \in [0.1 , 1.5]$ con incrementos de 0.1).

Proposición 3.5.35 Para cada valor de $t > 1$, una cota superior de la variación en el extremo inferior, dentro del intervalo $[0, K_2]$ y una cota superior del valor absoluto de la variación en el extremo inferior, dentro del intervalo $(K_2 , K_3]$ viene determinada por :

$$\Theta_t = 1 - (1 - 1/t^2)^{1/2}$$

siendo ésta más grande conforme disminuye t

Demostración :

Hemos visto, en la definición 3.5.31, que en el intervalo $[0, K_2]$ el valor máximo que toma la función d_1 es $d_1^M = t - (t^2 - 1)^{1/2}$ en el punto de abcisa $K = K_1$. Calculando la variación en éste punto, se tiene :

$$\Theta_t = V_1 (K = K_1) = t - (t^2 - 1)^{1/2} / t = 1 - (1 - 1/t^2)^{1/2}$$

teniendo en cuenta que $K_1 = 1/(t^2-1)^{1/2}$

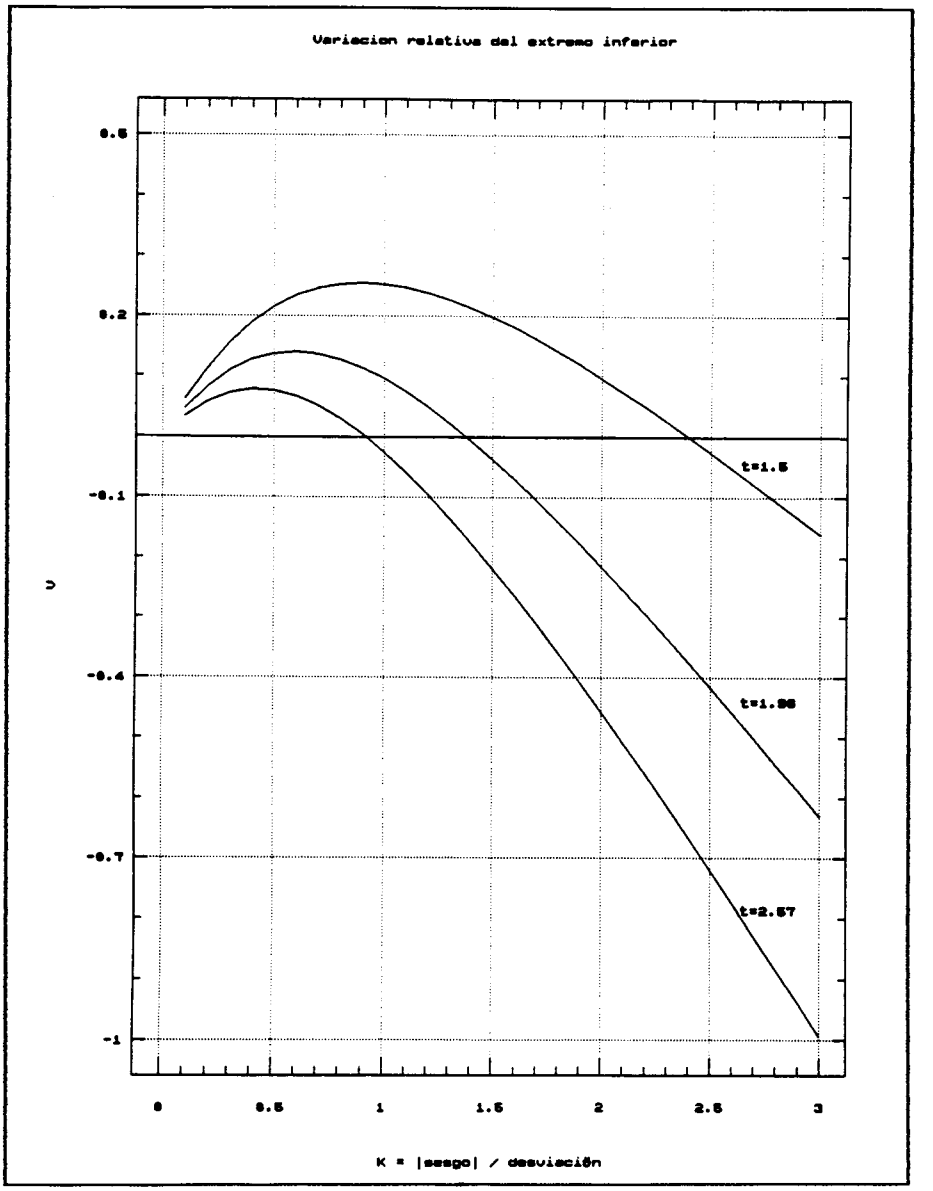
Por otro lado de la proposición 3.5.34 y para el valor de $K = K_1$, se tiene que :

Si $t_1 < t_2 \Rightarrow V_1(t_1) > V_1(t_2)$, eso es : $\Theta_{t1} > \Theta_t$

Finalmente, hemos visto en la definición 3.5.31, que en el punto $K=K_3$ el valor máximo que toma la función d_1 es $d_1^M = t - (t^2 - 1)^{1/2}$, que es el valor en el punto de abcisa $K = K_1$. Como la función d_1 es decreciente en el intervalo [K_2 , K_3] y estamos considerando su valor absoluto, d_1^M es la cota superior en dicho intervalo.

c.q.d.

Gráfico I.10



Proposición 3.5.36 Fijado un valor de K y una variación $\Theta > 0$, una c.n.y s. para que exista un valor de $t > 1$ para el que el valor absoluto de la variación del extremo inferior $V_1(t)$ sea igual a Θ es que :

a) Si la variación queremos que sea positiva : $\Theta < 1 + K - (1+K^2)^{1/2}$

b) Si la variación queremos que sea negativa : $\Theta < (1+K^2)^{1/2} - 1$

Demostración :

(==>)

De la proposición 3.5.33, $V_1(t) = K/t + 1 - (1+K^2)^{1/2}$

Para el caso a) : $V_1(t) = \Theta \Longrightarrow t = \dfrac{K}{(1+K^2)^{1/2} + \Theta - 1}$

La condición $t > 0 \Longrightarrow (1+K^2)^{1/2} + \Theta - 1 > 0 \Longrightarrow$

$$\Theta > 1 - (1+K^2)^{1/2} \qquad (1)$$

La condición $t > 1 \Longrightarrow K > (1+K^2)^{1/2} + \Theta - 1 \Longrightarrow$

$$\Theta < 1 + K - (1+K^2)^{1/2} \qquad (2)$$

de (1) y (2) $\Longrightarrow 1 - (1+K^2)^{1/2} < \Theta < 1 + K - (1+K^2)^{1/2}$ y al ser el extremo de la izquierda : $1 - (1+K^2)^{1/2} < 0$, se sigue que :

$$0 < \Theta < 1 + K - (1+K^2)^{1/2}$$

Para el caso b) : $V_1(t) = -\Theta \Longrightarrow t = \dfrac{K}{(1+K^2)^{1/2} - \Theta - 1}$

La condición $t > 0 \Longrightarrow (1+K^2)^{1/2} - \Theta - 1 > 0 \Longrightarrow$

$$\Theta < (1+K^2)^{1/2} - 1 \qquad (3)$$

La condición $t > 1 \Longrightarrow K > (1+K^2)^{1/2} - \Theta - 1 \Longrightarrow$

$$\Theta > (1+K^2)^{1/2} - (1+K) \qquad (4)$$

de (3) y (4) $\Longrightarrow (1+K^2)^{1/2} - (1+K) < \Theta < (1+K^2)^{1/2} - 1$ y al ser el extremo de la izquierda : $(1+K^2)^{1/2} - (1+K) < 0$, se sigue que :

$$0 < \Theta < (1+K^2)^{1/2} - 1$$

(<==)

Es trivial

c.q.d.

Para una acotación de la variación del extremo superior, dado que V_2 tiene por expresión la indicada en la proposición 3.5.33, podemos fijar una variación máxima Θ y determinar el intervalo $[0, K]$ en el que $V_2 \leq \Theta$. La resolución de ésta desigualdad conlleva una expresión complicada para el extremo K del intervalo; podemos plantearnos el problema con otro enfoque más fácil de desarrollar y ligando ésta variación con la del extremo inferior.

Así mismo, planteamos obtener una expresión para los extremos Y_1 e Y_2 en función de la variación relativa.

Definición 3.5.37 Definimos $V_2 - V_1$, como la diferencia entre la variación del extremo superior e inferior, que de acuerdo con las expresiones indicadas en la proposición 3.5.33, toma la expresión :

$$V_2 - V_1 = 2 [(1 + K^2)^{1/2} - 1] \qquad K>0$$

que es el de una función extrictamente creciente

Proposición 3.5.38 Si elegimos un valor Θ, tal que $V_2 - V_1 =< \Theta$, existe un intervalo $[0, K]$, dentro del cual se verifica ésta desigualdad. Además, si

$\Theta \leq \Theta_2$, el intervalo es tal que $K \leq K_2$

$\Theta_2 < \Theta \leq \Theta_3$, el intervalo es tal que $K_2 < K \leq K_3$

siendo Θ_2 y Θ_3 el valor de $V_2 - V_1$ en los puntos de abcisa K_2 y K_3, respectivamente.

Demostración :

De la definición de $V_2 - V_1$ y para un valor Θ, la desigualdad :

$V_2 - V_1 = 2 [(1 + K^2)^{1/2} - 1] \leq \Theta$, tiene por solución el intervalo:

$$K =< (\Theta(1 + \Theta/4))^{1/2}$$

Lo que demuestra la primera parte de ésta proposición. Por otro lado ésta función, en el intervalo $[0, K_3]$, toma los valores

$$V_2 - V_1 = \begin{cases} 0 & \text{para } K = 0 \\ V_2(K=K_1) - V_1(K=K_1) & \text{para } K = K_1 \\ V_2(K=K_2) = \Theta_2 & \text{para } K = K_2 \\ V_2(K=K_3) - V_1(K=K_1) = \Theta_3 & \text{para } K = K_3 \end{cases} \quad \text{pués } V_1(K=K_2) = 0$$

y en consecuencia, se verifica lo propuesto dado que la función es estrictamente creciente.

c.q.d.

Proposición 3.5.39 La variación del extremo superior es tal que :

$$V_2 = 2K/t - V_1$$

Demostración :

De la definición 3.5.37, y de la proposición 3.5.33, se tiene :

$V_2 = V_1 + 2 [(1 + K^2)^{1/2} - 1]$ y $V_1 = K/t + 1 - (1 + K^2)^{1/2}$

Multiplicando la segunda por 2 y sumando ambas igualdades, se sigue :

$V_2 + 2 V_1 = V_1 + 2K/t$, y en definitiva $V_2 = 2K/t - V_1$

c.q.d.

Proposición 3.5.40 Siendo $\Theta \leq \Theta_t$ la variación relativa del extremo inferior asociado a P_r, se tiene :

$$Y_1 = -t.(1-\Theta) \quad e \quad Y_2 = 2K + t.(1-\Theta) = 2K - Y_1$$

Demostración :

De la definición 3.5.32 y de la proposición 3.5.39, se sigue :

$$V_1 = \frac{Y_1 - (-t)}{t} = \Theta \implies Y_1 = -t.(1-\Theta)$$

$$V_2 = \frac{Y_2 - t}{t} = 2K/t - V_1 = 2K/t - \Theta \implies Y_2 = 2K + t.(1-\Theta) = 2K - Y_1$$

c.q.d.

Proposición 3.5.41 Para un valor de $K = B/S(\bar{Y}_n)$ y fijada una variación $\Theta > 0$, existe :

a) Un intervalo de valores de t : $[t_I \; t_S]$, dentro del cual la variación relativa del extremo inferior asociado a P_r con el del extremo inferior asociado a P_α es menor o igual que Θ, con la condición de ser
$\Theta < 1 + K - (1+ K^2)^{1/2}$.

b) Un intervalo $(t_S \; t_{S*}]$, complementario del anterior, en el que la variación relativa del extremo inferior asociado a P_r con el del extremo inferior asociado a P_α es mayor que $-\Theta$, con la condición de ser $\Theta < (1+K^2)^{1/2} - 1$

Demostración :

Siendo V_1 la variación del extremo inferior del intervalo asociado a P_r, se sigue (ver gráfico I.11) :

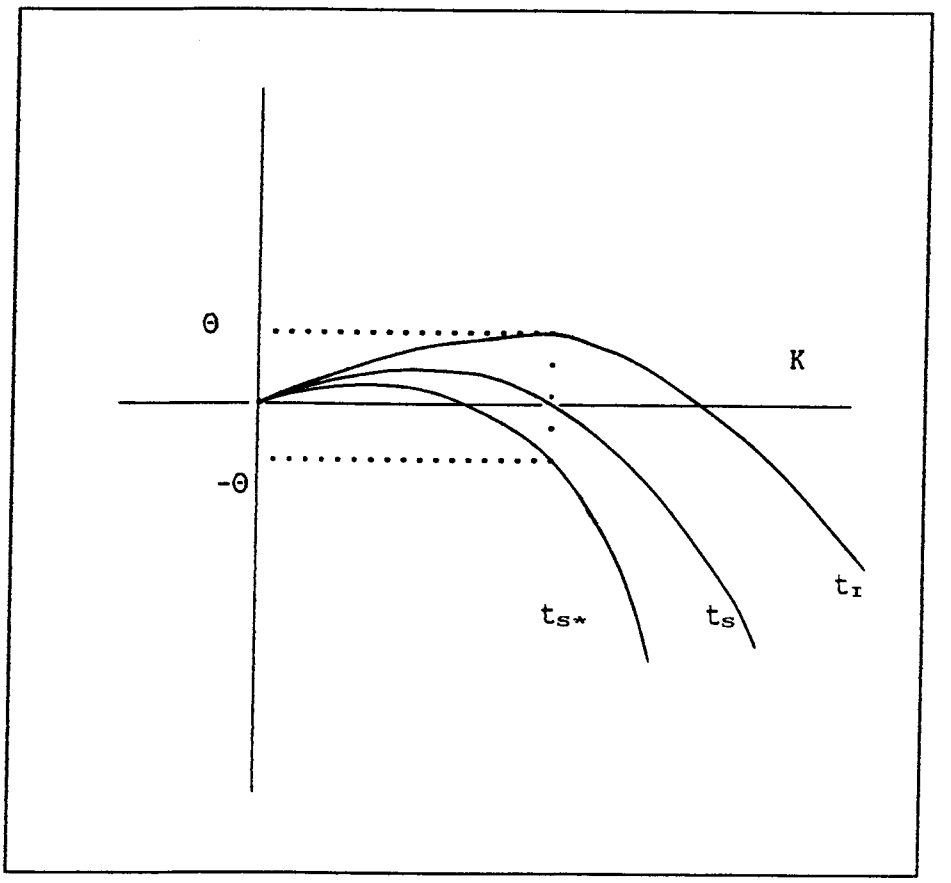
* De acuerdo con la proposición 3.5.34, 2ª parte, para el valor K, existe un valor para $t > 1$, en el que $V_1(t) = 0$, siendo éste valor :

$$t_s = [\ 1 + (\ 1 + K^2\ )^{1/2}\ ]\ /\ K$$

* De acuerdo con la proposición 3.5.34, 1ª parte, y de la consecuencia anterior se tiene para el valor de K que hemos fijado :

$$\begin{aligned} &\text{Para todo } t \le t_s \ : \ V_1(t) \ge V_1(t_s) = 0 \\ &\text{Para todo } t > t_s \ : \ V_1(t) < V_1(t_s) = 0 \end{aligned} \qquad (2)$$

Gráfico I.11



Si consideramos que la variación sea positiva :

* Elegimos los $t \le t_s$, pués $V_1(t) \ge 0$

* De la proposición 3.5.33, determinamos el valor de $t > 1$, tal que $\Theta = V_1(t)$, esto es :

$$\Theta = K/t + 1 - (\ 1 + K^2\ )^{1/2}$$

obteniendo el valor : $t_I = \dfrac{K}{(1 + K^2)^{1/2} - (1-\Theta)}$

Además, para que $t_I > 1$:

$$K / [(1 + K^2)^{1/2} - (1-\Theta)] > 1 \Longrightarrow \Theta < 1 + K - (1 + K^2)^{1/2}$$

Como $\Theta > 0$, se deduce que $t_I < t_S$ (recíproco de la proposición 3.5.34)

* De la proposición 3.5.33, determinamos para cada valor de $t \leq t_S$ su variación para el valor de K :

$$V_1(t) = K/t + 1 - (1 + K^2)^{1/2}$$

y eliminamos los valores de $t \leq t_S$ tales que $V_1(t) > \Theta$, que de acuerdo con la proposición 3.5.34, serán todos aquellos tales que : $t<t_I$

* En definitiva nos quedamos con el intervalo : $t_I \leq t \leq t_S$

Si consideramos que la variación sea negativa :

* Elegimos los $t > t_S$, pués $V_1(t) < 0$

* De la proposición 3.5.33, los valores de t tales que $V_1(t) \geq -\Theta$ son aquellos que verifican :

$$K/t + 1 - (1 + K^2)^{1/2} > -\Theta \Longrightarrow t \leq \frac{k}{(1 + K^2)^{1/2} - (1+\Theta)} = t_{S*}$$

Como $\Theta < (1 + K^2)^{1/2} - 1$, se deduce por la proposición 3.5.36, que $t_{S*} > 1$

* En definitiva nos quedamos con el intervalo : $t_S < t \leq t_{S*}$

c.q.d.

Con la determinación de los intervalos para t, contemplados en la proposición anterior y dado que algunas proposiciones planteadas anteriormente, en concreto las proposiciones 3.5.28, 3.5.29 y 3.5.30, hacen referencia a aspectos que vamos a utilizar a continuación, pero en las que se relacionan con valores de K y no de t, en la siguiente proposición relacionamos los intervalos de t con los intervalos de K. Estos últimos fueron estudiados en el punto A del apartado anterior.

Proposición 3.5.42 Fijado un valor de K, y siendo t tal que $t_I \leq t \leq t_S$ o bien $t_S < t \leq t_{S*}$, se tiene que :

$$K \in [\ 0, K_2(t)\ ] \quad \text{ó} \quad K > K_2(t) \quad , \text{ respectivamente}$$

Demostración :

Dado que t_s lo hemos construido con la condición de que $K_2(t_s) = K$ (proposición 3.5.41) y dado que :

$$t_1 \leq t \leq t_s \Longrightarrow K_2(t_1) \geq K_2(t) \geq K_2(t_s)$$

pués la función $K_2 = 2t/(t^2-1)$ es estrictamente decreciente.

Deducimos que : $K \leq K_2(t)$

Por otro lado, como $t_s < t \leq t_{s*} \Longrightarrow K_2(t_s) > K_2(t) \geq K_2(t_{s*})$

Deducimos que : $K > K_2(t)$

c.q.d.

Para finalizar éste punto, en la tabla I.5 damos los valores para t>1 de Y_1, Y_2, V_1, V_2 P_r, P_α y P_r-P_α, tomando K = 0.8, y unos resultados que siguen lo planteado en la proposición 3.5.41.

Tabla I.5

(K = 0.8)

t	Y_1	Y_2	V_1	V_2	P_r	P_α	$P_r - P_\alpha$
1.00	-0.48062	2.08062	0.51938	1.08062	0.66587	0.68269	-0.016815
1.10	-0.60869	2.20869	0.44665	1.00790	0.71504	0.72867	-0.013632
1.20	-0.73675	2.33675	0.38604	0.94729	0.75964	0.76986	-0.010224
1.30	-0.86481	2.46481	0.33476	0.89601	0.79957	0.80640	-0.006824
1.40	-0.99287	2.59287	0.29080	0.85205	0.83486	0.83849	-0.003631
1.50	-1.12094	2.72094	0.25271	0.81396	0.86559	0.86639	-0.000798
1.60	-1.24900	2.84900	0.21938	0.78062	0.89197	0.89040	0.001573
1.65	-1.31303	2.91303	0.20422	0.76547	0.90362	0.90106	0.002567
1.70	-1.37706	2.97706	0.18996	0.75121	0.91430	0.91087	0.003429
1.75	-1.44109	3.04109	0.17652	0.73777	0.92404	0.91988	0.004161
1.80	-1.50512	3.10512	0.16382	0.72507	0.93290	0.92814	0.004764
1.90	-1.63319	3.23319	0.14043	0.70168	0.94817	0.94257	0.005606
2.00	-1.76125	3.36125	0.11938	0.68062	0.96051	0.95450	0.006014
2.10	-1.88931	3.48931	0.10033	0.66158	0.97033	0.96427	0.006062
2.20	-2.01737	3.61737	0.08301	0.64426	0.97802	0.97219	0.005830
2.30	-2.14544	3.74544	0.06720	0.62845	0.98395	0.97855	0.005399
2.40	-2.27350	3.87350	0.05271	0.61396	0.98845	0.98360	0.004843
2.50	-2.40156	4.00156	0.03938	0.60062	0.99181	0.98758	0.004225
2.60	-2.52962	4.12962	0.02707	0.58832	0.99427	0.99068	0.003595
2.70	-2.65769	4.25769	0.01567	0.57692	0.99606	0.99307	0.002990
2.75	-2.72172	4.32172	0.01028	0.57153	0.99675	0.99404	0.002705
2.80	-2.78575	4.38575	0.00509	0.56634	0.99732	0.99489	0.002434
2.85	-2.84978	4.44978	0.00008	0.56133	0.99781	0.99563	0.002180
2.90	-2.91381	4.51381	-0.00476	0.55649	0.99821	0.99627	0.001943
2.95	-2.97784	4.57784	-0.00944	0.55181	0.99855	0.99682	0.001724
3.00	-3.04187	4.64187	-0.01396	0.54729	0.99882	0.99730	0.001523

Aplicando la proposición 3.5.41, obtenemos para el valor K = 0.8, los extremos de los intervalos donde la variación relativa del extremo inferior es, en valor absoluto, menor que $\Theta = 0.20$:

$$t_I = 1.66 \quad , \quad t_S = 2.85 \quad y \quad t_{S*} = 9.92$$

Aunque en la tabla sólo indicamos valores hasta t = 3, observese cómo en el intervalo, según los valores indicados anteriormente :

$t_I \leq t \leq t_S$ la variación es positiva y es tal que $V_1(t) < 0.20$

$t_S < t \leq t_{S*}$ la variación es negativa y es tal que $V_1 > - 0.20$

En definitiva, en el intervalo $t_I \leq t \leq t_{S*}$, se verifica que

$$|V_1| < 0.20$$

B. Variación entre P_r y P_α

Queremos medir la variación entre P_r y P_α, esto es, la diferencia : $P_r - P_\alpha$, para valores de K y valores de t > 1, dentro del intervalo donde la variación del extremo inferior sea , en valor absoluto, menor o igual que Θ. Con estas condiciones, planteamos y resolvemos las siguientes cuestiones :

¿ Cómo de pequeña es la diferencia entre P_r y P_α ?
¿ Existe algún valor para t, en el que $P_r = P_\alpha$?

En lo que sigue, dado que necesitaremos expresar el valor de Y_1 e Y_2 para un valor de t concreto, lo indicaremos como : $Y_1(t)$ e $Y_2(t)$

Así mismo, el cuadrado de dicho valor como : $Y_1(t)^2$ e $Y_2(t)^2$

Proposición 3.5.43 Siendo a y b números reales positivos (a < b); f(x) una función real, positiva, de variable real; F(x)+Q el conjunto de primitivas de la función f(x) y D una constante, tal que D = F(b) - F(a). Se tiene :

$$D / (2.b^{1/2}) < \int_{a^{1/2}}^{b^{1/2}} f(t^2)\, dt < D / (2.a^{1/2})$$

Demostración :

Tomando $t^2 = x$:

$$\int_{a^{1/2}}^{b^{1/2}} f(t^2)\, dt = \int_{a}^{b} f(x) \,.\, 1 / (2.x^{1/2})\, dx$$

y teniendo en cuenta que en el intervalo: $a < x < b$ se verifica que

$$1/a^{1/2} > 1/ a^{1/2} > 1/b^{1/2} \quad y$$

$$f(x) / (2.b^{1/2}) < f(x) / (2.x^{1/2}) < f(x) / (2.a^{1/2})$$

se sigue :

$$\int_a^b f(x).1/(2.b^{1/2})\, dx < \int_a^b f(x).1/(2.x^{1/2})\, dx < \int_a^b f(x).1/(2.a^{1/2})\, dx \quad (1)$$

Como $\int_a^b f(x)\, dx = F(b)-F(a) = D$, de (1) se sigue la expresión propuesta

c.q.d.

Corolario 3.5.44 Siendo $f(t^2) = 1/(2.\pi)^{1/2}.\ e^{-t^2/2}$ y siendo a y b dos números reales positivos y arbitrarios, se tiene :

$$I < \int_{a^{1/2}}^{b^{1/2}} f(t^2)\, dt < S$$

siendo :

$$I = 1/(2.b^{1/2}) \ .\ D$$

$$S = 1/(2.a^{1/2}) \ .\ D$$

$$D = 2/(2.\pi)^{1/2} \ .\ (e^{-a/2} - e^{-b/2})$$

Demostración :

Es consecuencia de la proposición anterior. En efecto, al ser : $t^2=x$, deducimos que $f(x)=1/(2.\pi)^{1/2}.e^{-x/2}$ y $F(x) = (-2)/(2.\pi)^{1/2}.e^{-x/2}$, con lo que $D = 2/(2.\pi)^{1/2} \ .\ (e^{-a/2} - e^{-b/2})$ y sustituyendo en la expresión (1) de la proposición 3.5.43, obtenemos la expresión propuesta.

c.q.d.

Proposición 3.5.45 La diferencia entre P_r y P_α , para un valor de t y de K, es tal que :

$$I < P_r - P_\alpha < S$$

siendo :

$$I = 1/(2.\pi)^{1/2} \left[\frac{e^{-t^2/2} - e^{-Y_2^2/2}}{Y_2} - \frac{e^{-t^2/2} - e^{-Y_1^2/2}}{Y_1} \right]$$

$$S = 1/(2.\pi)^{1/2} \left[\frac{e^{-t^2/2} - e^{-Y_2^2/2}}{t} + \frac{e^{-t^2/2} - e^{-Y_1^2/2}}{t} \right]$$

y en las que Y_1 e Y_2 son los extremos del intervalo asociado a la probabilidad P_r, para el valor de t y K correspondiente (punto A, apartado 3.5.2)

Demostración :

De la proposición 3.5.27 y de la definición 3.5.22, se tiene:

$P_r = \Phi (Y_2) + \Phi (-Y_1)$ y $P_\alpha = \Phi (t) + \Phi (-t) = 2 . \Phi (t)$

Luego la diferencia entre P_r y P_α, podemos expresarla como :

$P_r - P_\alpha = [\Phi (Y_2) - \Phi (t)] - [\Phi (t) - \Phi (-Y_1)]$

$$\text{Como : } \Phi (Y_2) - \Phi (t) = \int_t^{Y_2} 1/(2.\pi)^{1/2} . e^{-t^2/2} . dt \qquad (1)$$

$$\Phi (t) - \Phi (-Y_1) = \int_{-Y_1}^{t} 1/(2.\pi)^{1/2} . e^{-t^2/2} . dt \qquad (2)$$

Aplicamos el corolario 3.5.44 a (1) y a (2), dado que Y_2, t y $-Y_1$ son todos ellos mayores que cero :

* Para la primera expresión ($a = t^2$ y $b = Y_2^2$) :

$$A_1 < \Phi (Y_2) - \Phi (t) < A_2 \qquad (3)$$

$$\text{siendo : } A_1 = 1/(2.\pi)^{1/2} \left[\frac{e^{-t^2/2} - e^{-Y_2^2/2}}{Y_2} \right]$$

$$A_2 = 1/(2.\pi)^{1/2} \left[\frac{e^{-t^2/2} - e^{-Y_2^2/2}}{t} \right]$$

* Para la segunda expresión ($b = t^2$ y $a = (-Y_1)^2$) :

$$B_1 < \Phi(t) - \Phi(-Y_1) < B_2 \qquad (4)$$

siendo : $$B_1 = 1/(2.\pi)^{1/2} \left[\frac{e^{-(-Y_1)^2/2} - e^{-t^2/2}}{t} \right]$$

$$B_2 = 1/(2.\pi)^{1/2} \left[\frac{e^{-(-Y_1)^2/2} - e^{-t^2/2}}{-Y_1} \right]$$

De (3) y (4) , se sigue : $A_1 - B_2 < P_r - P_\alpha < A_2 - B_1$

Luego $I = A_1 - B_2$ y $S = A_2 - B_1$

c.q.d.

En la Tabla I.6, continuación de la Tabla I.5, indicamos los valores para t>1 de I(t), S(t) y de la diferencia entre P_r y P_α que tambien es una función de t, considerando el valor de K = 0.8

Tabla I.6

t	I(t)	S(t)	$P_r - P_\alpha$
1.00	-0.14177	0.082714	-0.016815
1.10	-0.10380	0.063112	-0.013632
1.20	-0.07724	0.048534	-0.010224
1.30	-0.05746	0.037794	-0.006824
1.40	-0.04223	0.029945	-0.003631
1.50	-0.03036	0.024229	-0.000798
1.60	-0.02110	0.020045	0.001573
1.65	-0.01729	0.018377	0.002567
1.70	-0.01395	0.016929	0.003429
1.75	-0.01105	0.015659	0.004161
1.80	-0.00855	0.014534	0.004764
1.90	-0.00456	0.012612	0.005606
2.00	-0.00173	0.010994	0.006014
2.10	0.00019	0.009573	0.006062
2.20	0.00139	0.008289	0.005830
2.30	0.00205	0.007111	0.005399
2.40	0.00234	0.006030	0.004843
2.50	0.00236	0.005045	0.004225
2.60	0.00221	0.004160	0.003595
2.70	0.00197	0.003379	0.002990
2.75	0.00183	0.003028	0.002705
2.80	0.00168	0.002703	0.002434
2.85	0.00154	0.002403	0.002180
2.90	0.00140	0.002128	0.001943
2.95	0.00126	0.001878	0.001724
3.00	0.00113	0.001650	0.001523

Si tomamos para Θ = 0.20 podemos comprobar que :

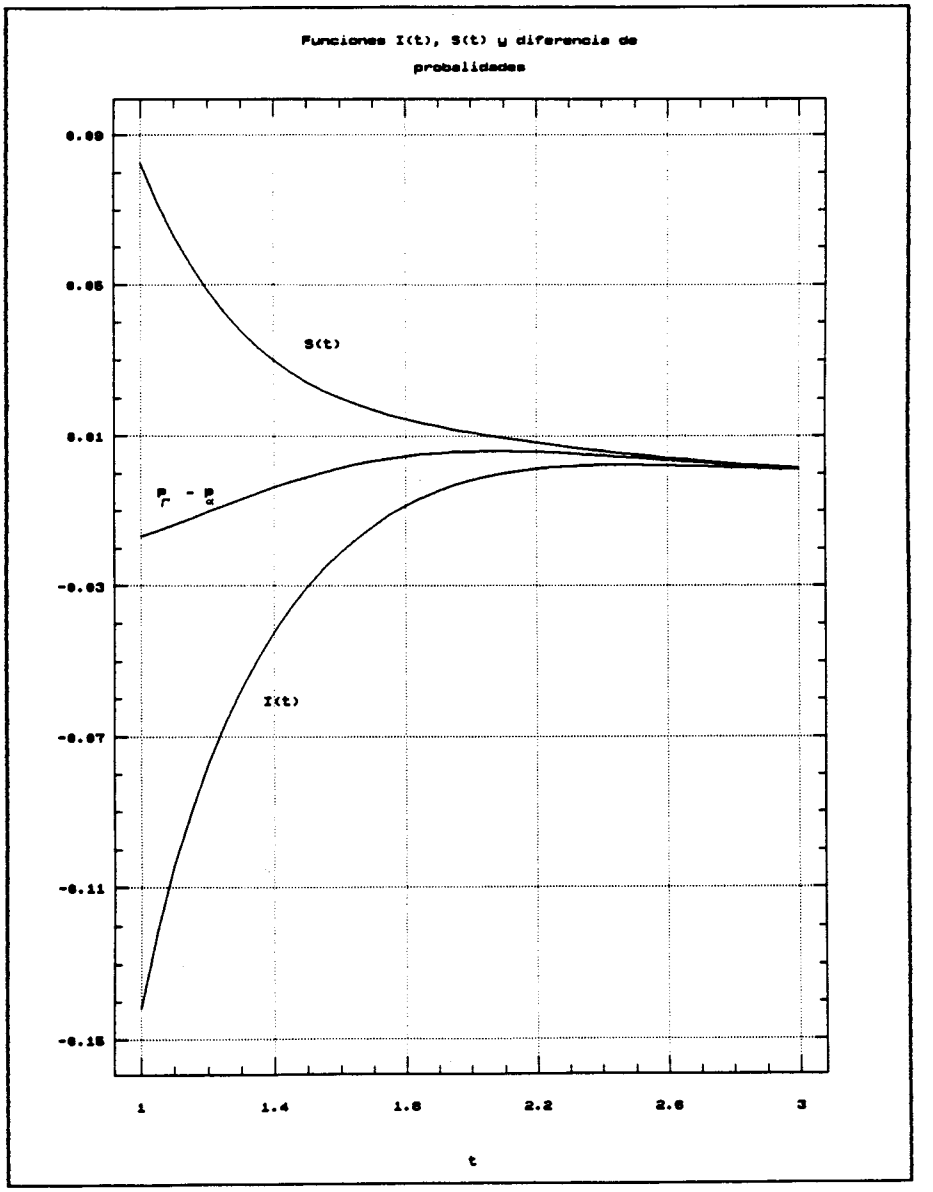
* Dentro del intervalo $[t_I, t_S]$ = [1.66 , 2.85] para valores cercanos al extremo superior, la diferencia $P_r - P_\alpha$ es cada vez mas pequeña.

* No hay un valor para t, en el que $P_r = P_\alpha$.

Por el contrario, para Θ = 0.25 , comprobamos que dentro del intervalo $[t_I, t_S]$ = [1.51 , 2.85], existe un valor para t (comprendido entre 1.51 y 1.60), en el que $P_r = P_\alpha$

En el gráfico I.12, se indica el comportamiento de estas tres funciones, tomando los valores de la tabla I.6 .

Gráfico I.12



Proposición 3.5.46 Siendo P_r y P_α las probabilidades asociadas a los intervalos $[Y_1,Y_2]$ y $[-t,t]$, respectivamente, fijado un valor de K, se tiene :

Para todo t : $t_I \leq t \leq t_S$: $P_r \leq P_\alpha$ o bien $P_\alpha \leq P_r$

Para todo t : $t_S \leq t \leq t_{S*}$: $P_r > P_\alpha$

Demostración :

Consecuencia de las proposiciones 3.5.29 y 3.5.42

c.q.d.

Con lo indicado en la tabla I.6, el gráfico I.12 y ésta última proposición, centramos el problema en el sentido de :

* En el intervalo $[t_I,t_S]$ puede existir un valor en el que $P_\alpha = P_r$ o un intervalo para valores de t, dentro del cual $|P_r - P_\alpha|$ sea menor que una cantidad positiva fijada.

* En el intervalo $[t_S,t_{S*}]$, al ser $P_r > P_\alpha$, es posible encontrar un intervalo para valores de t, dentro del cual $P_r - P_\alpha$ sea menor que una cantidad positiva fijada.

Dado que I(t) y S(t) son funciones que dependen de t, pero tambien del valor Y_1 e Y_2, que a su vez dependen de t y de K, no es posible conocer su comportamiento en el intervalo $[t_I, t_S]$ y $(t_S, t_{S*}]$, por lo que vamos a acotarlas en cada uno de estos intervalos.

Proposición 3.5.47 Para un valor de K fijo, y siendo I la cota inferior de $P_r - P_\alpha$, para cada valor de t, tal que $t_I < t < t_S$, se verifica :

$$I_1 < I < I_2$$

siendo :

$$I_1 = \left[e^{t_S^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_S^2 - Y_2(t_I)^2)/2}}{Y_2(t_S)} - \frac{1 - e^{(t_S^2 - Y_1(t_I)^2)/2}}{Y_1(t_S)} \right]$$

$$I_2 = \left[e^{t_I^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_I^2 - Y_2(t_S)^2)/2}}{Y_2(t_I)} - \frac{1 - e^{(t_I^2 - Y_1(t_S)^2)/2}}{Y_1(t_I)} \right]$$

Demostración :

Denotando por $y = \dfrac{e^{-t^2/2} - e^{-Y_2^2/2}}{Y_2} = \dfrac{1 - e^{(t^2 - Y_2^2)/2}}{Y_2 . e^{t^2/2}}$

$$z = -1.\ \frac{e^{-t^2/2} - e^{-Y_1^2/2}}{Y_1} = \frac{1 - e^{(t^2-Y_1^2)/2}}{-Y_1 \cdot e^{t^2/2}}$$

podemos expresar la cota inferior I de la diferencia $P_r - P_\alpha$ (proposición 3.5.45) en la forma :

$$I = 1/(2.\pi)^{1/2}\ (\ Y + Z\ )$$

Buscando una acotación para Z e Y, obtenemos nuestro propósito.

Acotación de Y :

Llamando $x = e^{(t^2-Y_2^2)/2}$, se sigue : $2.\mathrm{Ln}\ x + Y_2^2 = t^2$

Como $t_I^2 < t^2 < t_S^2$, se sigue : $(t_I^2 - Y_2^2)/2 < \mathrm{Ln}\ x < (t_S^2 - Y_2^2)/2$

Además del punto A, apartado 3.5.2, se obtiene que :

$$-[Y_2(t_I)]^2 > -[Y_2(t)]^2 > -[Y_2(t_S)]^2$$

Luego : $\mathrm{Ln}\ x < (t_S^2 - Y_2(t)^2)/2 < (t_S^2 - Y_2(t_I)^2)/2$, y

$$\mathrm{Ln}\ x > (t_I^2 - Y_2(t)^2)/2 > (t_I^2 - Y_2(t_S)^2)/2$$

y en consecuencia :

$$e^{(t_I^2 - Y_2(t_S)^2)/2} < x < e^{(t_S^2 - Y_2(t_I)^2)/2} \Longrightarrow$$

$$1 - e^{(t_I^2 - Y_2(t_S)^2)/2} > x > 1 - e^{(t_S^2 - Y_2(t_I)^2)/2} \qquad (1)$$

Por otro lado de : $t_I^2 < t^2 < t_S^2 \Longrightarrow e^{-t_I^2/2} > e^{-t^2/2} > e^{-t_S^2/2}$

y de ser : $Y_2(t_I) < Y_2(t) < Y_2(t_S) \Longrightarrow 1/Y_2(t_I) > 1/Y_2(t) > 1/Y_2(t_S)$

y en consecuencia podemos escribir :

$$\frac{1}{Y_2(t_I).e^{t_I^2/2}} > \frac{1}{Y_2(t).e^{t^2/2}} > \frac{1}{Y_2(t_S).e^{t_S^2/2}} \qquad (2)$$

Finalmente de (1) y (2), obtenemos :

$$\frac{1 - e^{(t_I^2 - Y_2(t_S)^2)/2}}{Y_2(t_I).e^{t_I^2/2}} > \frac{1 - e^{(t^2 - Y_2(t)^2)/2}}{Y_2(t).e^{t^2/2}} > \frac{1 - e^{(t_S^2 - Y_2(t_I)^2)/2}}{Y_2(t_S).e^{t_S^2/2}}$$

Esto es :

$$\frac{1 - e^{(t_I^2 - Y_2(t_S)^2)/2}}{Y_2(t_I).e^{t_I^2/2}} > Y > \frac{1 - e^{(t_S^2 - Y_2(t_I)^2)/2}}{Y_2(t_S).e^{t_S^2/2}} \qquad (3)$$

Acotación de Z :

Llamando $x = e^{(t^2-Y_1^2)/2}$, se sigue : $2.\mathrm{Ln}\ x + Y_1^2 = t^2$

Como $t_I^2 < t^2 < t_S^2$, se sigue : $(t_I^2 - Y_1^2)/2 < \mathrm{Ln}\ x < (t_S^2 - Y_1^2)/2$

Además del punto A, apartado 3.5.2, se obtiene que :

$$-[-Y_1(t_I)]^2 > -[-Y_1(t)]^2 > -[-Y_1(t_S)]^2$$

Luego : $\mathrm{Ln}\ x < (t_S^2 - Y_1(t)^2)/2 < (t_S^2 - Y_1(t_I)^2)/2$, y

$$\mathrm{Ln}\ x > (t_I^2 - Y_1(t)^2)/2 > (t_I^2 - Y_1(t_S)^2)/2$$

y en consecuencia :

$$e^{(t_I^2 - Y_1(t_S)^2)/2} < x < e^{(t_S^2 - Y_1(t_I)^2)/2} \Longrightarrow$$

$$1 - e^{(t_I^2 - Y_1(t_S)^2)/2} > x > 1 - e^{(t_S^2 - Y_1(t_I)^2)/2} \qquad (4)$$

Por otro lado de : $t_I^2 < t^2 < t_S^2 \Longrightarrow e^{-t_I^2/2} > e^{-t^2/2} > e^{-t_S^2/2}$

y de ser : $-Y_1(t_I) < -Y_1(t) < -Y_1(t_S) \Longrightarrow 1/[-Y_1(t_I)] > 1/[-Y_1(t)] > 1/[-Y_1(t_S)]$

y en consecuencia podemos escribir :

$$\frac{1}{-Y_1(t_I).e^{t_I^2/2}} > \frac{1}{-Y_1(t).e^{t^2/2}} > \frac{1}{-Y_1(t_S).e^{t_S^2/2}} \qquad (5)$$

Finalmente de (4) y (5), obtenemos :

$$\frac{1 - e^{(t_I^2 - Y_1(t_S)^2)/2}}{-Y_1(t_I).e^{t_I^2/2}} > \frac{1 - e^{(t^2 - Y_1(t)^2)/2}}{-Y_1(t).e^{t^2/2}} > \frac{1 - e^{(t_S^2 - Y_1(t_I)^2)/2}}{-Y_1(t_S).e^{t_S^2/2}}$$

Esto es :

$$\frac{1 - e^{(t_I^2 - Y_1(t_S)^2)/2}}{-Y_1(t_I).e^{t_I^2/2}} > Z > \frac{1 - e^{(t_S^2 - Y_1(t_I)^2)/2}}{-Y_1(t_S).e^{t_S^2/2}} \qquad (6)$$

Finalmemte, multplicando las expresiones (3) y (6) por $1/(2.\pi)^{1/2}$, y sumando ambas, obtenemos los términos I_1 e I_2 propuestos .

c.q.d.

Proposición 3.5.48 Para un valor de K fijo, y siendo S la cota superior de $P_r - P_\alpha$, para cada valor de t, tal que $t_I < t < t_S$, se verifica :

$$S_1 < S < S_2$$

siendo :

$$S_1 = [\, e^{t_S^2} . 2.\pi]^{-1/2} . \left[\frac{1 - e^{(t_S^2 - Y_2(t_I)^2)/2}}{t_S} + \frac{1 - e^{(t_S^2 - Y_1(t_I)^2)/2}}{t_S} \right]$$

$$S_2 = [\, e^{t_I^2} . 2.\pi]^{-1/2} . \left[\frac{1 - e^{(t_I^2 - Y_1(t_S)^2)/2}}{t_I} + \frac{1 - e^{(t_I^2 - Y_2(t_S)^2)/2}}{t_I} \right]$$

Demostración :

La demostración es análoga a la anterior, por lo que omitiremos algunos pasos intermedios, utilizados anteriormente.

Denotando por $Y = \dfrac{e^{-t^2/2} - e^{-Y_2^2/2}}{t} = \dfrac{1 - e^{(t^2 - Y_2^2)/2}}{t . e^{t^2/2}}$

$$Z = \frac{e^{-t^2/2} - e^{-Y_1^2/2}}{t} = \frac{1 - e^{(t^2 - Y_1^2)/2}}{t . e^{t^2/2}}$$

podemos expresar la cota inferior S de la diferencia $P_r - P_\alpha$ (proposición 3.5.45) en la forma :

$$I = 1/(2.\pi)^{1/2} \ (\ Y + Z \ )$$

Buscando una acotación para Z e Y, obtenemos nuestro propósito.

Acotación de Y :

Llamando $x = e^{(t^2-Y_2^2)/2}$, la expresión para Y es la misma que la de la expresión (1) de la proposición anterior, salvo para el denominador que sería :

$$\frac{1}{t_I.e^{t_I^2/2}} > \frac{1}{t.e^{t^2/2}} > \frac{1}{t_S.e^{t_S^2/2}}$$

en definitiva, quedaría en la forma :

$$\frac{1 - e^{(t_I^2 - Y_2(t_S)^2)/2}}{t_I.e^{t_I^2/2}} > Y > \frac{1 - e^{(t_S^2 - Y_2(t_I)^2)/2}}{t_S.e^{t_S^2/2}} \qquad (1)$$

Acotación de Z :

Llamando $x = e^{(t^2-Y_1^2)/2}$, la expresión para Z es la misma que la de la expresión (4) de la proposición anterior, salvo para el denominador que sería el mismo que para la expresión de Y. En definitiva, quedaría en la forma :

$$\frac{1 - e^{(t_I^2 - Y_1(t_S)^2)/2}}{t_I.e^{t_I^2/2}} > Z > \frac{1 - e^{(t_S^2 - Y_1(t_I)^2)/2}}{t_S.e^{t_S^2/2}} \qquad (2)$$

Finalmemte, multplicando las expresiones (1) y (2) por $1/(2.\pi)^{1/2}$, y sumando ambas obtenemos los términos S_1 y S_2 propuestos .

c.q.d.

Proposición 3.5.49 Para una valor de K y siendo I la cota inferior de la diferencia entre P_r y P_α, para cada valor de t, tal que $t_S < t < t_{S^*}$, se verifica :

$$I_1^* < I < I_2^* \qquad y \qquad S_1^* < S < S_2^*$$

siendo :

$$I_1^* = \left[e^{t_{s*}^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_{s*}^2 - Y_2(t_s)^2)/2}}{Y_2(t_{s*})} - \frac{1 - e^{(t_{s*}^2 - Y_1(t_s)^2)/2}}{Y_1(t_{s*})} \right]$$

$$I_2^* = \left[e^{t_s^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_s^2 - Y_2(t_{s*})^2)/2}}{Y_2(t_s)} - \frac{1 - e^{(t_s^2 - Y_1(t_{s*})^2)/2}}{Y_1(t_s)} \right]$$

$$S_1^* = \left[e^{t_{s*}^2} . 2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_{s*}^2 - Y_2(t_s)^2)/2}}{t_{s*}} + \frac{1 - e^{(t_{s*}^2 - Y_1(t_s)^2)/2}}{t_{s*}} \right]$$

$$S_2^* = \left[e^{t_s^2} . 2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_s^2 - Y_1(t_{s*})^2)/2}}{t_s} + \frac{1 - e^{(t_s^2 - Y_2(t_{s*})^2)/2}}{t_s} \right]$$

Demostración :

Igual que las proposiciones 3.5.47 y 3.5.49, cambiando el intervalo $[t_I, t_s]$ por el intervalo $(t_s, t_{s*}]$

c.q.d.

Proposición 3.5.50 Para valores de t comprendidos en los intervalos $[t_I, t_s]$ ó $(t_s, t_{s*}]$, la diferencia entre P_r y P_α está comprendida entre :

$I_1 < P_r - P_\alpha < S_2$ ó $I_1^* < P_r - P_\alpha < S_2^*$, respectivamente

Siendo para el intervalo $[t_I, t_s]$:

$$I_1 = \left[e^{t_s^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_s^2 - Y_2(t_I)^2)/2}}{Y_2(t_s)} - \frac{1 - e^{(t_s^2 - Y_1(t_I)^2)/2}}{Y_1(t_s)} \right]$$

$$S_2 = \left[e^{t_I^2} . 2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_I^2 - Y_1(t_s)^2)/2}}{t_I} + \frac{1 - e^{(t_I^2 - Y_2(t_s)^2)/2}}{t_I} \right]$$

y para el intervalo $(t_s, t_{s*}]$:

$$I_1^* = \left[e^{t_{s*}^2} .2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_{s*}^2 - Y_2(t_s)^2)/2}}{Y_2(t_{s*})} - \frac{1 - e^{(t_{s*}^2 - Y_1(t_s)^2)/2}}{Y_1(t_{s*})} \right]$$

$$S_2^* = \left[e^{t_s^2} . 2.\pi \right]^{-1/2} . \left[\frac{1 - e^{(t_s^2 - Y_1(t_{s*})^2)/2}}{t_s} + \frac{1 - e^{(t_s^2 - Y_2(t_{s*})^2)/2}}{t_s} \right]$$

Demostración :

Es consecuencia de las proposiciones 3.5.47, 3.5.48 y 3.5.49

c.q.d.

Proposición 3.5.51 Fijado un valor Ω, una condición necesaria para que exista al menos un valor perteneciente a uno de los intervalos $[t_I, t_s]$ ó $(t_s, t_{s*}]$, para el que la diferencia entre P_r y P_α sea menor o igual que Ω, es que :

$$I_1 < \Omega < S_2 \quad \text{ó} \quad I_1^* < \Omega < S_2^*$$

respectivamente. Siendo I_1 y S_2 ó I_1^* y $< S_2^*$ las cotas, de valor el indicado en la proposición 3.5.50.

Demostración :

Es consecuencia de la proposición anterior.

c.q.d.

3.5.4. Conclusiones

Al comienzo del apartado 3.5.3 hemos planteado la posibilidad de identificar un intervalo de confianza para un estimador sesgado con una probabilidad P_r con otro intervalo de confianza para un estimador insesgado con una probabilidad P_α, indicando una solución dada, experimentalmente, por los autores allí indicados.

Nosotros hemos demostrado a lo largo de éste apartado que es posible identificar estos intervalos, con unas características definidas, para cualquier valor de K, no siendo necesario que K<1, esto es que " el sesgo del estimador no sobrepase su error estándar ", por lo que hemos ampliado ésta solución.

Para finalizar y a título de conclusión, reflejamos en las Tabla I.7 para diversos valores de K y de Θ :

* Los intervalos de valores de t: $[t_I, t_S]$, dentro de los cuales la variación del extremo inferior es , en valor absoluto, menor que Θ (consecuencia de la Proposición 3.5.41).

* La cota superior e inferior dentro de cada intervalo de la diferencia entre las probabilidades P_r y P_α (consecuencia de las Proposiciones 3.5.50 y 3.5.51) .

Tabla I.7

K	Θ	t_I	t_S	I_1	S_2
0.2	0.05	2.86517	10.0990	-0.00235	-0.00726
0.2	0.10	1.66939	10.0990	-0.04220	0.09830
0.2	0.15	1.17783	10.0990	-0.09127	0.30971
0.2	0.20	0.90990	10.0990	-0.12544	0.54233
0.2	0.25	0.74128	10.0990	-0.14765	0.77197
0.4	0.05	3.14879	5.1926	-0.00100	-0.00411
0.4	0.10	2.25947	5.1926	-0.01263	0.01929
0.4	0.15	1.76186	5.1926	-0.03653	0.08539
0.4	0.20	1.44387	5.1926	-0.06325	0.18200
0.4	0.25	1.22312	5.1926	-0.08730	0.29358
0.6	0.05	2.77533	3.6103	-0.00313	0.00276
0.6	0.10	2.25403	3.6103	-0.01453	0.02379
0.6	0.15	1.89759	3.6103	-0.03247	0.06458
0.6	0.20	1.63849	3.6103	-0.05259	0.12155
0.6	0.25	1.44165	3.6103	-0.07190	0.18934
0.8	0.05	2.41966	2.8508	-0.00540	0.01339
0.8	0.10	2.10181	2.8508	-0.01905	0.03678
0.8	0.15	1.85777	2.8508	-0.03642	0.07092
0.8	0.20	1.66450	2.8508	-0.05480	0.11376
0.8	0.25	1.50766	2.8508	-0.07232	0.16300
1.0	0.05	2.15418	2.4142	-0.00509	0.02580
1.0	0.10	1.94472	2.4142	-0.02052	0.05019
1.0	0.15	1.77238	2.4142	-0.03807	0.08072
1.0	0.20	1.62810	2.4142	-0.05595	0.11620
1.0	0.25	1.50554	2.4142	-0.07292	0.15547
1.2	0.05	1.96062	2.1350	-0.00300	0.03854
1.2	0.10	1.81255	2.1350	-0.01944	0.06245
1.2	0.15	1.68528	2.1350	-0.03706	0.09000
1.2	0.20	1.57470	2.1350	-0.05465	0.12050
1.2	0.25	1.47774	2.1350	-0.07135	0.15333
1.4	0.05	1.81708	1.94319	-0.000542	0.05097
1.4	0.10	1.70635	1.94319	-0.017409	0.07361
1.4	0.15	1.60834	1.94319	-0.034925	0.09850
1.4	0.20	1.52097	1.94319	-0.052223	0.12524
1.4	0.25	1.44261	1.94319	-0.068657	0.15346

De la tabla observamos que para K = 1.4 encontramos un valor para t, dentro del intervalo [1.81, 1.94] para el que el extremo inferior del intervalo asociado a P_Γ difiere del asociado a P_α en Θ = 0.05 con una diferencia de probabilidad comprendida entre -0.000542 (que tomaremos valor cero, en lo que sigue) y 0.05.

En consecuencia :

a) Entre la región de probabilidad asociada al estimador sesgado que tiene por extremos (Proposición 3.5.27 y punto A del apartado 3.5.2):

$$Y_1 = -1.70t + 1.4 \qquad y \qquad Y_2 = 1.70t + 1.4$$

y la región de probabilidad asociada al estimador insesgado que tiene por extremos -t y t , hay una diferencia entre sus probabilidades asociadas comprendida entre 0 y 0.05.

Como ésta diferencia es pequeña, podemos considerar que los niveles de significación en los que se mueven los intervalos de confianza para el estimador sesgado y el insesgado son el mismo; esto es $\Gamma=\alpha$.

b) La variación relativa del extremo inferior Y_1 respecto de -t es de Θ = 0.05.

c) Como t = $Z_{\alpha/2}$ y teniendo en cuenta que , para K = 1.4 y Θ = 0.05, el valor de t está dentro del intervalo [1.81, 1.94], deducimos que el error en el que nos movemos es de :

$$0.0262 \leq \alpha \leq 0.0351$$

d) Como consecuencia, las regiones de probabilidad [-t,t] e $[Y_1,Y_2]$ se mueven en una probabilidad entre :

$$0.9649 \leq 1-\alpha \leq 0.9731 \qquad (1)$$

Como consecuencia de ésto último y de lo indicado en a) los extremos de variación para el nivel de confianza en los intervalos asociados al estimador sesgado e insesgado son los indicados en (1).

e) Identificamos pués, los intervalos de confianza, siendo K = 1.4

Así mismo, en la Tabla I.8, quedan reflejados, para diversos valores de K y de Θ :

* Los intervalos de valores de t: $(t_s,t_{s*}]$, dentro de los cuales la variación del extremo inferior es mayor que -Θ (consecuencia de la Proposición 3.5.41).

Tabla I.8

K	Θ	t_s	t_{s*}	I_1^*	S_2^*
0.6	0.05	3.6103	5.1639	-0.000100	0.000208
0.6	0.10	3.6103	9.0648	-0.000100	0.000327
0.6	0.15	3.6103	37.0590	-0.000000	0.000327
0.8	0.05	2.8508	3.4688	-0.001300	0.004392
0.8	0.10	2.8508	4.4291	-0.001400	0.004799
0.8	0.15	2.8508	6.1244	-0.001000	0.004811
0.8	0.20	2.8508	9.9225	-0.000600	0.004811
0.8	0.25	2.8508	26.1226	-0.000200	0.004811
1.0	0.05	2.4142	2.7456	-0.001900	0.015189
1.0	0.10	2.4142	3.1825	-0.004900	0.017542
1.0	0.15	2.4142	3.7848	-0.004800	0.017914
1.0	0.20	2.4142	4.6682	-0.003900	0.017928
1.0	0.25	2.4142	6.0896	-0.002800	0.017928
1.2	0.05	2.1350	2.3435	0.000200	0.029145
1.2	0.10	2.1350	2.5971	-0.006300	0.035081
1.2	0.15	2.1350	2.9123	-0.009300	0.037567
1.2	0.20	2.1350	3.3145	-0.009500	0.038186
1.2	0.25	2.1350	3.8455	-0.008400	0.038253
1.4	0.05	1.9432	2.0881	0.003452	0.043579
1.4	0.10	1.9432	2.2564	-0.004864	0.052715
1.4	0.15	1.9432	2.4541	-0.010380	0.058328
1.4	0.20	1.9432	2.6899	-0.013319	0.061034
1.4	0.25	1.9432	2.9758	-0.014099	0.061952

* Así mismo, la cota superior e inferior dentro de cada intervalo de la diferencia entre las probabilidades P_r y P_α (consecuencia de las Proposiciones 3.5.50 y 3.5.51).

Como puede observarse hay valores de K en los que no existe el intervalo $(t_s, t_{s*}]$, al no ser $\Theta < (1+K^2)^{1/2} - 1$.

Por ejemplo, para K = 2, para que exista t_{s*} es necesario que $\Theta <$ 0.019, por debajo de los valores que hemos propuesto (consecuencia de la Proposición 3.5.36).

4. Tratamiento de la no respuesta

4.1. Introducción

Ya hemos indicado en el apartado 2 de éste capítulo que la no respuesta consiste en que no se obtienen observaciones, total o parcialmente, de las unidades seleccionadas para la muestra.

Durante la realización de la encuesta debemos plantearnos la falta de respuesta desde dos puntos de vista diferentes, según la etapa en la que nos encontremos. El primero es ¿ Cómo evitar la no respuesta ? y el segundo es que aceptando que hay no respuesta ¿ cómo debemos tratarla ?

En lo que sigue realizamos una revisión bibliográfica de las técnicas más usuales para tratar la no respuesta, en cada una de las siguientes etapas : Planificación, Diseño, Realización, Procesamiento o Validación.

4.2. Fase de planificación

En ésta etapa se especifican los fines de estudio así como las condiciones, recursos y limitaciones, incluyendo dentro de ésta fase la infraestructura y el marco.

Como marco, en muestreo de poblaciones finitas, entendemos el listado de unidades de muestreo que constituyen la población o alguna otra forma de definir y delimitar dichas unidades.

Para ésta fase, Deming (1960) aconseja que se supriman del marco todas las unidades de muestreo de las que se sepa son inaccesibles o que van a causar dificultades no justificadas por los fines del estudio en relación con su costo. Sostiene que es preferible una eliminación (corte, poda, etc..) de unidades antes de proceder a la selección a tener que prescindir de ellas una vez que hayan salido seleccionadas.

4.3. Fase de diseño

Comprende el diseño muestral (despiece de la población original, definición de estratos y/o conglomerados, unidades de muestreo, estimadores a utilizar y tratamiento de la no respuesta), así como, cuestionario, procedimientos de recogida, evaluaciones, etc..

En ésta etapa la no respuesta puede evitarse o quedar más reducida. Los factores que inciden directa o indirectamente en ella, son estudiados a continuación.

4.3.1. Factores que tienen un efecto indirecto en la tasa de respuesta

* Marco muestral

* Tamaño muestral

* Estratificación

* Reparto de la muestra por estratos

* Procedimiento muestral dentro de los estratos

* Reparto de la muestra por etapas, especialmente conglomerados de la muestra

4.3.2. Factores que afectan directamente a la tasa de respuesta

* Listado de unidades para la selección muestral

* Tema y tipo de encuesta

* Procedimiento mediante el cual se realiza el cuestionario

* Longitud y complejidad del cuestionario

* Cuestiones controvertidas para encuestadores y encuestados

* Selección, desarrollo y control del equipo de campo incluyendo a los entrevistadores

* Tipo de área para la encuesta

* Manejo y coste de las encuestaciones repetidas y procedimientos de seguimiento

* Publicidad y medios

4.4. Fase de campo o de realización

Comprende la ejecución del programa y la recogida de datos. Es la última etapa en la que podemos evitar o disminuir la no respuesta

4.4.1. Mejoramiento de los procedimientos de entrevista

* Garantía del anonimato del entrevistado

* Motivación para la cooperación del entrevistado

* Concretar con el entrevistado la visista

4.4.2. Repetición de intentos para conseguir la información

A los intentos deliberados de obtener respuesta de las no respuestas, se les conoce por visitas adicionales, revisitas o encuestaciones repetidas (Call Backs) y comprenden tanto las visitas propiamente como el envío repetido del cuestionario a los que no han respondido en las encuestas por correo.

Deming (1953) formula un modelo que permite que informaciones referentes al tipo, características demográficas y socioeconòmicas y cooperación del entrevistado, incentivos del entrevistador y coste de las

visitas adicionales sean incorporados al modelo, siendo ésta información la que permite distribuir a los entrevistados en diferentes clases o grupos.

El modelo estima el número medio de entrevistas durante k llamadas, a partir del estimador muestral. Cochran (1977), determina el número óptimo de llamadas usando éste modelo.

Una vez que se conozca la tasa de respuesta, la decisión de realizar una o varias visitas adicionales , pudiendo variar su número en diversas partes de la muestra, se basa en varios supuestos :

* El número posible de nuevas respuestas debe ser suficiente alto para justificar ésta tarea y ésto depende de las causas por las que la encuesta no ha podido realizarse.

* El coste de tales visitas y el tiempo disponible

Para el primero de ellos, si es por ausencia, es razonable pensar que va a haber un porcentaje de personas que van a responder; si por el contrario ha sido por rechazo o incapacidad, es razonable pensar lo contrario.

Hay otro método que pretende evitar por completo las visitas adicionales mediante la recolección de primeras visitas solamente, que se ponderan con información acerca de la probabilidad de encontrar al entrevistado. Es el **esquema de Politz y Simmons** (1949) y consiste en lo siguiente :

Suponemos que todas las visitas se hacen en K períodos similares. Se le pregunta al entrevistado si habría estado disponible para la entrevista en cada uno de los K períodos.

Si es r el número de períodos afirmativos ($0 \leq r \leq K$) se toma la razón $r+1/K+1$ como estimación de la frecuencia con que está en casa durante las horas de entrevista.

La ponderación, W, es la inversa de dicha frecuencia, de tal forma que W es mayor cuanto menor es r.

Los resultados de la primera visita se ponen en K+1 grupos según el valor de r : 0, 1, 2,...,t, ..., K

Para el grupo t, sea n_t el número de entrevistas e $\bar{Y}_t$ la media del grupo. El estimador de Politz - Simmons para la media poblacional es :

$$\bar{Y}_{ps} = \frac{\sum_{t=0}^{K} n_t \cdot Y_t \cdot W_t}{\sum_{t=0}^{K} n_t W_t} \quad , \text{ con } \quad W_t = \frac{K+1}{r+1}$$

Se obtiene así una media sesgada, pero con menor sesgo que la media estándar y con varianza aumentada (Cochran) .

Por otro lado, Kish (1965), señala una serie de inconvenientes sobre éste procedimiento y concluye " Dudamos que, en la mayoría de los casos, éste esquema resulte válido, práctico y económico. De cualquier manera, si la situación de la encuesta permite una sola visita, los resultados ponderados, con sesgos y varianza incrementados, puede resultar preferible al sesgo de las primeras visitas sin ponderar " (Sic).

Hay un tercer procedimiento conocido por **reemplazo**, sugerido por Kish y Hess (1959), que propone agregar a las nuevas direcciones de la encuesta otras direcciones de no respuesta tomadas de una encuesta anterior con procedimientos de muestreo semejantes y convirtiendose, en consecuencia, en reemplazos de direcciones de no respuesta en la encuesta en que se trabaja.

Como condición previa, las no respuestas de la encuesta en que se trabaja deben ser semejantes a las no respuestas que actuan de reemplazos. El procedimiento es como sigue :

Sea N el total de direcciones para una encuesta y sea t_R la tasa de respuesta que se espera obtener despues de K visitas.

El total de respuestas previstas será : $N \cdot t_R$

Supongamos que tenemos un número de direcciones de no respuesta, al que denotaremos por n, de un estudio anterior y que la tasa de respuesta para estas direcciones y en ésta encuesta es de t_R', despues de K visitas.

El total de respuestas previstas será : $n \cdot t_R'$

Al reemplazar $n \cdot t_R'$ en las $N \cdot t_R$, el total de respuestas que necesitamos es de :

$$N \cdot t_R - n \cdot t_R'$$

Y el total de direcciones :

$$1 / t_R \cdot (N \cdot t_R - n \cdot t_R') = N - n \cdot t_R'/t_R$$

El área de incertidumbre debida a la no respuesta se reduce en alrededor de :

$$(1 - t_R). (1 - t_R')$$

4.4.3. Submuestreo de las no respuestas

Se le conoce por el método de Hansen y Hurwitz (1946) y parte de la idea de que se puede considerar a la población dividida en dos estratos : Los que responden y los que no.

Al seleccionar una muestra, los que responden representan una muestra aleatoria del primer estrato y los que no responden, del segundo. Se toma una submuestra de un tamaño conveniente y se recoge información mediante entrevistas personales a esas unidades. Después se toman en cuenta las dos muestras para obtener un estimador de la media poblacional.

4.4.4. Encuestación delegada

Si las instrucciones de la encuesta indican que se puede encuestar, en el caso de que no sea posible hacerlo a la persona indicada, a cualquier otro miembro de la familia que tenga una edad por encima de una fijada previamente, habremos reducido la posibilidad de que dicha unidad quede sin respuesta. No es aconsejable cuando la información es confidencial o personal.

4.4.5. Muestreo por cuotas

Las cuotas se establecen con base en características conocidas de la población en estudio, tales como vecindario, edad, sexo, nivel educacional, situación de empleo, alquiler o valor de la casa, etc...

El diseño de la encuesta sigue los principios del muestreo probabilístico, hasta llegar el momento de seleccionar las personas que han de ser entrevistadas.

Esta etapa final es dejada en manos de los entrevistadores y son ellos los que eligen a las personas a encuestar, dentro de un perfil previo (ésta forma de seleccionar dentro de las cuotas no es aleatoria, sino a juicio del entrevistador y constituye el punto débil de éste muestreo). Cochran (1977) lo describe como un muestreo estratificado con una selección mas o menos no aleatoria dentro de los estratos

El muestreo por cuotas no tiene no respuesta ya que el entrevistador no registra aquellos que intenta entrevistar sin conseguirlo, dado que el entrevistador tiene instrucciones de seguir muestreando hasta obtener la cuota necesaria de su grupo.

4.4.6 Técnicas de respuesta aleatorizada

Esta técnica fué desarrollada con el único fín de obtener información de confianza concerniente a cuestiones planteadas en las encuestas y en las que se prestan a no decir la verdad ó a no contestar.

El encuestado debe seleccionar, por medio de un mecanismo aleatorio, una de dos preguntas : una transcendente (opciòn 1) y otra intranscendente (opción 2), que el entrevistador le formula y contestar Si o No sin que el entrevistador sepa cual ha sido la pregunta seleccionada.

Considerando que ambas cuestiones son complementarias, se denota por π la verdadera proporción de la población que contesta a la pregunta transcendente (opción 1) y por p la probabilidad, conocida, de que el

mecanismo aleatorio elija la pregunta transcendente. La proporción de la población, que denotaremos por Θ, que contesta afirmativamente, viene expresada por:

$$\Theta = p.\pi + (1 - p).(1 - \pi)$$

Observese que la verdadera proporción de la población que contesta afirmativamente a la pregunta intranscendente es $1 - \pi$, al suponer las cuestiones complementarias.

Estimando Θ por la proporción de la población que contesta afirmativamente en la encuesta, que denotaremos por Θ^*, se tiene el siguiente modelo :

$$\Theta^* = p.\pi^* + (1 - p).(1 - \pi^*)$$

Despejando, se tiene : $\pi^* = (\Theta^* - 1 + p)/ (2p + 1)$

donde π^* es un estimador insegado de π (Warner, 1965)

El mecanismo aleatorio que hemos comentado anteriormente, puede ser por ejemplo (Enrich, 1983) una caja conteniendo proporciones conocidas de bolas blancas y negras. El entrevistado recibe la instrucción de seleccionar al azar una bola, sin que el entrevistador vea el color de ella. Si sale blanca, debe elejir la opción 1 y negra la opción 2.

Esta técnica descrita tiene muchas modificaciones (Ladoux, 1982); entre ellas podemos indicar las que parten de la hipótesis de que las dos cuestiones que se plantean (opciones 1 y 2) no son complementarias.

Bajo éste supuesto, Θ y p tienen el mismo significado de antes; a la verdadera proporción de la población, que contesta afirmativamente a la pregunta transcendente, la denotamos por π_1 y a la verdadera proporción de la población, que contesta afirmativamente a la pregunta intranscendente, la denotamos por π_2. Como las cuestiones no son complementarias, se tiene que $\pi_2 <> 1 - \pi_1$

El modelo, para éste supuesto quedaría :

$$\Theta = p.\pi_1 + (1 - p).(1 - \pi_2)$$

Si π_2 es desconocida es necesario dos muestras independientes para estimar π_1 y π_2.

4.5. Fase de procesamiento o validación

En esta fase se incluyen los aspectos informáticos y, en general, la consideración de la información estadística considerada como un todo integrado. Es la etapa en la que es necesario tratar las unidades con entradas inconsistentes y/o con no respuesta (ya sin posibilidad de evitarla), antes de proceder a la estimación de los parámetros poblacionales.

Los procedimientos empleados en ésta etapa son la depuración y/o la imputación.

Entendemos por depuración (Villan y Bravo, 1990 y Villar, 1992) en una encuesta el conjunto de técnicas que permiten corregir, mediante unas reglas y a partir de la información recogida en la encuesta y en ocasiones de otra adicional, las entradas inconsistentes, entendiendo por estas las imposibilidades lógicas o posibilidades altamente improbables (Platek, 1986). Con la depuración separamos los cuestionarios en aceptables y los que tienen no respuesta.

Las técnicas de depuración mas usuales, son :

. Sistemas generales de depuración, basadas en la Metodología de Fellegi & Oldt (1980)

. Método de estimación máximo - verosimil : Algoritmo E.M.

. Método general de depuración CIDAC

Una vez detectado el dato anómalo hay que suplirlo por otro con la imputación.

Antes de pasar a definir qué entendemos por imputación, damos una descripción breve de algunas palabras que se utilizan :

Por "**unidad**" entenderemos el conjunto de preguntas de la encuesta que deben ser contestadas por cada individuo de la muestra; también se le llama **registro**.

Cada pregunta de la **unidad** diremos que es una variable, que toma como valores las posibles respuestas de la pregunta; tambien se le llama **campo** o **item**. Esta variable puede ser cualitativa o cuantitativa.

A los registros completos se les denomina donantes y a los que tienen campos con no respuesta receptores.

A los campos que se utilizan para establecer la relación entre unos y otros, campos de control.

Entendemos por imputación como

. La asignación de datos a campos vacíos de un registro

. La asignación de datos a unidades vacías

. Reemplazamiento del dato de un campo por las de otro del mismo campo y distinto registro.

4.5.1 Efectos de la imputación en la estimación de algún parámetro poblacional

Vamos a estudiar el efecto de la imputación en el sesgo de no respuesta, cuando estimamos la media , el total y el de una razón poblacional; para los dos primeros estudiamos el sesgo relativo de no respuesta, demostrando que es el mismo para ambos estimadores y

determinando la pérdida relativa o disminución en el sesgo. Así mismo estudiaremos los efectos en la varianza muestral para uno de los parámetros antes indicados.

En lo que sigue Y_i representa las unidades de la muestra en las que hay unidades con no respuesta. Después de realizar una imputación consideramos que :

$$Y_i = \begin{cases} X_i & \text{para las unidades que responden} \\ X_i^* & \text{para las unidades que no respoden, pero se les ha imputado un valor} \end{cases}$$

A. En la estimación de la media poblacional

Consideremos $\bar{Y}_n^*$ un estimador de $\bar{Y}_N$, tal que :

$$E(\bar{Y}_n^*) = t_{N1} \cdot \bar{Y}_{N1} + t_{N2} \cdot \bar{Y}_{N2}^*$$

siendo $\bar{Y}_{N2}^*$ una estimación de $\bar{Y}_{N2}$ como consecuencia de una imputación. Teniendo en cuenta la proposición 3.1.2, se tiene :

$$\text{sesgo}(\bar{Y}_n^*) = E(\bar{Y}_n^*) - \bar{Y}_N = t_{N2} \, (\bar{Y}_{N2}^* - \bar{Y}_{N2}) \qquad (1)$$

Vamos a comparar éste sesgo con el obtenido antes de la imputación, considerando que :

$$E(\bar{Y}_n) = \bar{Y}_{N1}$$

Proposición 4.5.1 La imputación reduce en términos absolutos o a lo sumo deja invariante el sesgo de no respuesta en la estimación de la media poblacional.

Demostración :

Admitiendo la hipótesis de que $\bar{Y}_{N2}^*$ está más cerca de la media poblacional de los que no responden, que de los que responden, podemos escribir

$$|\bar{Y}_{N2}^* - \bar{Y}_{N2}| \leq |\bar{Y}_{N1} - \bar{Y}_{N2}| \qquad (2)$$

y según la proposición 3.1.4, como $\bar{Y}_n$ es el estimador antes de la imputación

$$\text{sesgo}(\bar{Y}_n) = t_{N2} \, (\bar{Y}_{N1} - \bar{Y}_{N2}) \qquad (3) \text{ , supuesto que } E(\bar{Y}_n) = \bar{Y}_{N1}$$

Comparando (1) y (3), teniendo en cuenta (2), podemos escribir

$$|\text{sesgo}(\bar{Y}_n^*)| \leq |\text{sesgo}(\bar{Y}_n)| \qquad (4)$$

Si no admitimos la hipótesis, es decir, suponemos que $\bar{Y}_{N2}^*$ está mas cerca de la media poblacional de los que responden, entonces tomaremos al imputar, como valor de $\bar{Y}_{N2}^*$, el valor $\bar{Y}_{N1}$ y, en consecuencia, la expresión (2) será una igualdad y los sesgos iguales

c.q.d

Corolario 4.5.2 Una cota superior del valor absoluto del sesgo después de la imputación en la estimación de la media poblacional, es :

$$t_{N2} \ (\bar{Y}_{N1} - \bar{Y}_{N2})$$

Demostración :

Se deduce de la proposición anterior, expresión (4) y (3)

c.q.d.

B. <u>En la estimación del total poblacional</u>

Consideremos Y_n^* un estimador de Y_N , tal que :

$E(Y_n^*) = N_1 \ . \ \bar{Y}_{N1} + N_2 \ . \ \bar{Y}_{N2}^*$, siendo $\bar{Y}_{N2}^*$ una estimación de $\bar{Y}_{N2}$ como consecuencia de una imputación.

Teniendo en cuenta la proposición 3.3.12, se tiene :

$$\text{sesgo}(Y_n^*) = E(Y_n^*) - Y_N = N_2 \ (\bar{Y}_{N2}^* - \bar{Y}_{N2}) \qquad (1)$$

Vamos a comparar éste sesgo con el obtenido antes de la imputación, considerando que $E(Y_n) = Y_{N1}$

Proposición 4.5.3 La imputación reduce en términos absolutos el sesgo de no respuesta en la estimación del total poblacional

Demostración :

Admitiendo la hipótesis de que $\bar{Y}_{N2}^*$ está más cerca de la media poblacional de los que no responden, podemos escribir

$$|\bar{Y}_{N2}^* - \bar{Y}_{N2}| \leq |\bar{Y}_{N2}| \qquad (2)$$

según la proposición 3.3.13, se tiene :

$$\text{sesgo}(Y_n) = -1. \ N_2 \ . \ \bar{Y}_{N2} \qquad (3) \ , \text{ supuesto que } E(Y_n) = Y_{N1}$$

Comparando (1) y (3), teniendo en cuenta (2), podemos escribir

$$|\text{sesgo}(\bar{Y}_n^*)| \leq |\text{sesgo}(\bar{Y}_n)| \qquad (4)$$

Si no admitimos la hipótesis, es decir, suponemos que Y_{N2}^* está mas cerca de la media poblacional de los que responden, entonces tomaremos al imputar, como valor de Y_{N2}^*, el valor Y_{N1} y, en consecuencia, la expresión (2) quedará en la forma :

$$|\bar{Y}_{N1} - \bar{Y}_{N2}| \leq |\bar{Y}_{N2}| \qquad (5)$$

y comparando (1) y (3), teniendo en cuenta (5), seguirá siendo válida la expresión (4)

c.q.d

Corolario 4.5.4 Una cota superior del valor absoluto de sesgo después de la imputación en la estimación del total poblacional es :

$$N_2 . \bar{Y}_{N2}$$

Demostración :

Es una consecuencia de la proposición , expresión (4) y (3)

c.q.d.

Proposición 4.5.5 El sesgo relativo de no respuesta, después de la imputación, es el mismo para el estimador de la media que del total poblacional e igual a la expresión

$$t_{N2}(W^* - W)/(1 - t_{N2}(1-W))$$

donde, $W^* = \bar{Y}_{N2}^*/\bar{Y}_{N1}$ y $W = \bar{Y}_{N2}/\bar{Y}_{N1}$

Demostración :

Hemos visto que los sesgos, después de la imputación, para la estimación de la media y del total (proposiciones 4.5.1 y 4.5.3) son

$$\text{sesgo}(\bar{Y}_n^*) = t_{N2} (\bar{Y}_{N2}^* - \bar{Y}_{N2})$$

$$\text{sesgo}(Y_n^*) = N_2 (\bar{Y}_{N2}^* - \bar{Y}_{N2})$$

Utilizando la expresión para $\bar{Y}_N$ e Y_N, obtenidas en las proposiciones 3.1.2 y 3.3.12, respectivamente, se sigue :

$$SR(\bar{Y}_n^*) = \text{sesgo}(\bar{Y}_n^*)/ \bar{Y}_N = t_{N2} (\bar{Y}_{N2}^* - \bar{Y}_{N2})/(t_{N1}.\bar{Y}_{N1} + t_{N2}.\bar{Y}_{N2})$$

$$SR(Y_n^*) = \text{sesgo}(Y_n^*)/ Y_N = N_2 (\bar{Y}_{N2}^* - \bar{Y}_{N2})/(N_1.\bar{Y}_{N1} + N_2.\bar{Y}_{N2}) =$$

(multplicando el numerador y el denominador por N)

$$= t_{N2} (\bar{Y}_{N2}^* - \bar{Y}_{N2})/(t_{N1}.\bar{Y}_{N1} + t_{N2}.\bar{Y}_{N2}) \qquad (1)$$

$$= SR(\bar{Y}_n^*)$$

Transformando la expresión (1); esto es, diviendo por $\bar{Y}_{N1}$, obtenemos la expresión propuesta.

c.q.d.

Proposición 4.5.6 La pèrdida relativa o disminución, en el sesgo despuès de la imputación, con respecto al de no respuesta, es :

Para el estimador $\bar{Y}_n^*$: $1 - \left| \dfrac{W^* - W}{1 - W} \right|$ (1)

Para el estimador Y_n^* : $1 - \left| \dfrac{W^* - W}{W} \right|$ (2)

donde, $W^* = \bar{Y}_{N2}^*/\bar{Y}_{N1}$ y $W = \bar{Y}_{N2}/\bar{Y}_{N1}$

Demostración :

Teniendo en cuenta las proposiciones 3.1.4 , 3.3.13 y lo indicado al principio de los puntos A y B, podemos escribir :

$$\frac{|\text{sesgo}(\bar{Y}_n)| - |\text{sesgo}(\bar{Y}_n^*)|}{|\text{sesgo}(\bar{Y}_n)|} = \frac{|t_{N2}(1 - W)| - |t_{N2}(W^* - W)|}{|t_{N2}(1 - W)|} =$$

$$= \frac{|1 - W| - |W^* - W|}{|1 - W|} = (1)$$

Para el otro estimador :

$$\frac{|\text{sesgo}(Y_n)| - |\text{sesgo}(Y_n^*)|}{|\text{sesgo}(Y_n)|} = \frac{|-1.\, t_{N2}\, W| - |t_{N2}(W^* - W)|}{|-1.\, t_{N2}\, W|} =$$

$$= \frac{|W| - |W^* - W|}{|W|} = (2)$$

c.q.d.

C. En la estimación de la razón poblacional

Consideremos R_n^* un estimador de R_N , tal que :

$E(R_n^*) = (Y_{N1} + Y_{N2}^*)/Z_N$, siendo Y_{N2}^* una estimación de Y_{N2} como consecuencia de una imputación y Z_N el total poblacional de la variable Z conocido.

Teniendo en cuenta la definición 3.4.17, se tiene :

$$sesgo(R_n^*) = E(R_n^*) - R_N = C_{N2}\ (R_{N2}^* - R_{N2}) \qquad (1)$$

Vamos a comparar éste sesgo con el obtenido antes de la imputación, considerando que $E(R_n) = R_{N1}$

Proposición 4.5.7 La imputación reduce, en términos absolutos o a lo sumo, deja invariante el sesgo de no respuesta en la estimación de la razón poblacional

Demostración :

Admitiendo la hipótesis de que R_{N2}^* está más cerca de la razòn poblacional de los que no responden que de los que responden, podemos escribir

$$|R_{N2}^* - R_{N2}| \leq |R_{N1} - R_{N2}| \qquad (2)$$

según la proposición 3.4.18, se tiene

$$sesgo(R_n) = C_{N2}\ (R_{N1} - R_{N2}) \quad (3)\ , \text{ siendo } E(R_n) = R_{N1}$$

Comparando (1) y (3), teniendo en cuenta (2), podeemos escribir

$$|sesgo(R_n^*)| \leq |sesgo(R_n)| \qquad (4)$$

Si no admitimos la hipótesis; es decir, suponemos que R_{N2}^* está mas cerca de la razòn poblacional de los que responden, entonces tomaremos al imputar, como valor de R_{N2}^*, el valor R_{N1} y, en consecuencia, la expresión (2) será una igualdad y los sesgos iguales.

c.q.d

Corolario 4.5.8 Una cota superior del valor absoluto del sesgo después de la imputación en la estimación de la razòn poblacional, es :

$$C_{N2}\ (R_{N1} - R_{N2})$$

Demostración :

Se deduce de la proposición anterior, expresión (4) y (3)

D. <u>En la varianza muestral</u>

En el corolario 3.2.9 establecimos la descomposición de la varianza ajustada muestral frente a la no respuesta. Si consideramos que hemos realizado una imputación, el estimador de la media poblacional tendrá una expresión del tipo :

$$\bar{Y}_n = n_1/n\ .\ \bar{Y}_{n1} + n_2/n\ .\ \bar{Y}_{n2}^* \quad (1) \quad , \text{ siendo } \bar{Y}_{n2}^* = \Sigma\ X_i^*/n_2$$

Veamos la expresión que toma la varianza ajustada muestral después de la referida imputación

Proposición 4.5.9 La varianza ajustada poblacional, despúes de una imputación, toma la expresión :

$$S_n^2 = 1/(n-1) \cdot [(n_1-1).S_{n1}^2 + n_1.\bar{Y}_{n1}^2 + n_2 \cdot A - n \cdot \bar{Y}_n^2]$$

$$\text{con} \quad A = \sum_{i=1}^{n_2} X_i^{*2} / n_2$$

Demostración :

La varianza ajustada muestral de los que no responden, S_{n2}^2, una vez realizada la imputación, tendrá por expresión :

$$S_{n2}^2 = 1/(n_2-1) \cdot \sum_{i=1}^{n_2} (X_i^* - \bar{Y}_{n2}^*)^2 =$$

$$1/(n_2-1) \cdot (\Sigma X_i^{*2} - 2 \cdot \Sigma X_i^* \cdot \bar{Y}_{n2}^* + \Sigma \bar{Y}_{n2}^{*2}) =$$

$$1/(n_2-1) \cdot (n_2 \cdot A - n_2 \cdot \bar{Y}_{n2}^{*2}) \qquad (2)$$

teniendo en cuenta que

$$A = \sum_{i=1}^{n_2} X_i^{*2} / n_2 \quad \text{y} \quad \bar{Y}_{n2}^* = \Sigma X_i^*/n_2$$

Por otro lado, de la expresión (1) deducimos que :

$$\bar{Y}_{n2}^* = n/n_2 \cdot \bar{Y}_n - n_1/n_2 \cdot \bar{Y}_{n1} \qquad (3)$$

Sustituyendo (2) en la expresión obtenida en el el corolario 3.2.9, se tiene :

$$(n-1). \ S_n^2 = (n_1-1).S_{n1}^2 + n_1 (\bar{Y}_{n1} - \bar{Y}_n)^2 + n_2 \cdot A - n_2 \cdot \bar{Y}_{n2}^{*2} +$$

$$n_2 \cdot (\bar{Y}_{n2}^* - \bar{Y}_n)^2 =$$

sustituyendo (3) en ésta expresión y simplificando se obtiene la expresión propuesta.

c.q.d.

Corolario 4.5.10 Una expresión simplificada de la varianza ajustada muestral, despues de una imputación, es :

$$S_n^2 = n_1/n(S_{n1}^2 + \bar{Y}_{n1}^2) + n_2/n \cdot A - \bar{Y}_n^2$$

Demostración :

Tomando como aproximación de n-1, n y de n_1-1, n_1 en la expresión obtenida anteriormente, se sigue la expresión propuesta.

c.q.d.

4.5.2 Métodos de clasificación de unidades, en muestras con no respuesta, antes de la imputación

En algunas de las técnicas de imputación que estudiaremos a continuación, las unidades son clasificadas en grupos, para después imputar dentro de cada grupo a las unidades con no respuesta. Estudiamos tres técnicas de clasificación : Postestratificación, Area de balances y celdas ponderadas.

Postestratificación

Dada una muestra muestra aleatoria simple de extensión n, procederemos a clasificar las unidades que la componen formando L estratos en la muestra. Los tamaños de los estratos, N_i, a nivel poblacional, pueden obtenerse de manera bastante exacta a partir de las estadísticas oficiales.

En lugar de la media muestral $\bar{Y}_n$, usamos la estimación : $\bar{Y}_n = \sum_{i=1}^{L} W_i.\bar{Y}_i$

donde Y_i es la media del estrato i y $W_i = N_i/N$.

Este método es casi tan preciso como el muestreo estratificado proporcional (Cochran, 1977), siempre que :

* La muestra sea razonablemente grande (Cochran, indica el número de 20 unidades o más por estrato).

* Los efectos de los errores en las ponderaciones W_i puedan ignorarse.

Para evaluar la eficiencia de la postestratificación, Thonsen y Siring (1983) obtiene una expresión del sesgo de no respuesta, en la forma :

$$\text{sesgo} = 1/\bar{t} \; . \; \sum_{}^{L} \bar{Y}_{Ni1}. W_i. (t_{ni1}-\bar{t}) + \sum_{}^{L} (\bar{Y}_{Ni1} - \bar{Y}_{Ni2}) W_i. (1-t_{ni1})$$

donde t_{ni1} es la tasa de respuesta en el estrato i y $\bar{t} = \sum_{}^{L} W_i.t_{ni1}$

además, Y_{Ni1} y Y_{Ni2} son las medias poblacionales de los que responden y no responden, respectivamente, en el estrato i con n_i unidades en él.

La primera componente del sesgo puede ser estimada por la muestra, pero no la segunda al ser desconocida la media muestral de los que no responden.

Areas de balance y clases de ponderaciones

Un área de balance es un área geográfica, que puede ser a su vez un estrato individual, un grupo de estratos, una provincia, una unidad muestral primaria ó un conglomerado (Oh y Scheuren,1983 y Plateck,1986).

Una clase de ponderacióm es una agrupación de unidades muestrales con unas características comunes (tipos de vivienda, grupos especiales de renta,...) sin tener en cuenta la localización geográfica ó el área en el marco de la muestra.

La definición de una u otra es preferible en la etapa de planificación antes que en la etapa de procesamiento, dado que eligiendo la primera etapa se pueden realizar ajustes agrupando áreas si es necesario, bien porque la tasa de no respuesta sea baja, bien porque la muestra sea demasiado pequeña.

El esquema para estimación después de la imputación, utilizando cualquiera de estas dos técnicas de clasificación, es el siguiente :

$Y = \sum_{b} Y_b$ estima el total de una característica

Para un área de balance b, se tiene :

$$Y_b = \sum_{1}^{n_{1b}} W_i \cdot X_i \cdot 1/\pi_i + \sum_{i+n_{1b}}^{n_{1b+u}} W_j \cdot X_j^* \cdot 1/\pi_j$$

π_i y π_j son las probabilidades de inclusión

X_i y X_j^* son unidades que contestan y las que son imputadas

n_{ib} unidades que contestan

u unidades que han sido imputadas, dentro de las que no contestan

4.5.3. Técnicas de imputación

Las técnicas para imputación de datos podemos diferenciarlas en dos tipos generales. En ambas, el objetivo es el de predecir el mejor valor a asignar a un campo de un registro

A. Las que utilizan los valores existentes en algún registro de la variable a imputar o relacionados con ellos :

A.1. Sustitución

Con ésta técnica se pretende sustituir la no respuesta por un valor conocido de las unidades o relacionado con ellas. Comprende tres subtécnicas :

A.1.1. Sustitución por la media

Sustituye las unidades con no respuesta por la media de las unidades que responden (Bailar et al. 1978).

A.1.2. Sustitución por duplicación

Sustituye las unidades con no respuesta por un subconjunto, aleatoriamente seleccionado, de unidades muestrales con respuesta. La técnica es la siguiente :

Si n_1 y n_2 son el número de unidades que responden y no responden, respectivamente, en la muestra, los diferentes subconjuntos que podemos formar para sustituir las unidades que no responden por unidades que responden es, para el caso en que $n_1 \geq n_2$: $C\ n_1\ ,\ n_2$ donde C indica las combinaciones.

Para el caso en que $n_1 < n_2$, cada unidad de los que responden deberá ser usada el menor número posible de veces, de la siguiente forma :

Se determina K, tal que :

$n_2 = K\ .\ n_1 + m$ K es un número entero y $m < n_1$

con lo cual :

$n_1 - m$ unidades que responden se utilizarán K veces y
m unidades que respondem se utilizarán K+1 veces

es decir : $n_2 = K\ .\ (\ n_1 - m\ ) + (\ K + 1\ ).\ m$

A.1.3. Sustitución utilizando variables auxiliares

Se sustituye las unidades con no respuesta por datos históricos (que pueden ser del més anterior) o por datos que provienen de fuentes externas : Administrativos, de otras encuestas, del censo.. etc.

A.2. Métodos de regresión

Asignan a los campos con datos faltantes el valor de predicción (valor medio condicionado a los datos del registro) de modelos del tipo :

$$X_j = a + b_1.X_1 + b_2.X_2 + b_3.X_3 + \ldots\ldots + b_k.X_k + \varepsilon \qquad (1)$$

onde X_j es la variable dependiente con campos a imputar y las variables restantes X_i (i<>j) son los donantes.

Debe existir una fuerte correlación entre estas variables y la variable dependiente. En general el modelo (1) da lugar a un conjunto muy extenso de procedimientos de imputación que dependen de :

* El subconjunto de registros a los que se aplique el modelo

* El tipo de regresores en el modelo

* Los supuestos sobre la distribución y los parámetros del término aleatorio ε

En particular si al modelo (1) le añadimos una perturbación ε_i aleatoria, es decir :

$$X_j = a + b_1.X_1 + b_2.X_2 + b_3.X_3 + ... + b_k.X_k + \varepsilon_i$$

Obtenemos el modelo de regresión aleatorio.

Los residuos ε_i se calculan por uno de estos métodos :

* Se obtiene una muestra aleatoria de tamaño n_2 = nº de individuos que no responden en la muestra de extensión n, entre el total de individuos n_1 = nº de individuos que responden y se suman a los valores estimados.

* Se obtiene aleatoriamente una muestra de tamaño n_2, de una distribución N(0, S^2) donde S^2 es la varianza residual entre los n_1

A.3. Método de Buck

Es un método de estimación que consiste en :

1) Estimar el vector de medias con las n_1 observaciones completas

2) Estimar el valor de la variable a imputar en la observación i, volviendo al paso 1)

Añade, además, un término de corrección para los términos de la matriz de varianzas y covarianzas con objeto de obtener estimaciones insesgadas de éstos (Villar y Bravo, 1990).

B. Las que utilizan un conjunto de campos - variables de control- de la encuesta. En todas ellas hay un registro donante :

B.1. Imputación determinista o deductiva

Se aplica en situaciones en las que las respuestas que faltan se pueden deducir del resto de la información proveniente del conjunto de datos

B.2. Fichero caliente (Hot - Deck)

Consiste en sustituir un valor que falta por otro existente en la muestra. La elección no es aleatoria entre todas las unidades de la muestra, sino que previamente se realiza una clasificación de estas por grupos disjuntos, de forma que sean lo más homogéneas posibles dentro de cada grupo.

Una vez que estamos dentro del grupo, la asignación puede hacerse :

* Secuencial : El donante recibe el valor del registro anterior o posterior. Esto exige facilitar valores iniciales para cubrir la eventualidad de que el primero o el último del grupo sea receptor y no tenga donante.

* Aleatorio : Se elige al azar de entre los donantes del grupo

Dado que con estos supuestos estamos admitiendo que los valores faltantes siguen la misma distribución que los que responden dentro de su grupo, las variables de clasificación deben ser muy restrictivas y estar correladas con los valores que faltan y con los que contestan.

Hay una variación a éste método, **Hot - Deck modificado**, consistente en que la clasificación de las unidades muestrales en grupos se realiza mediante un considerable número de variables. La asignación del donante se hace en base a que el receptor coincide con él en todas las variables de clasificación. De no encontrarse ninguno, se eliminan algunas variables en orden de menor importancia para de ésta forma conseguir un donante a un nivel inferior.

Con éste método reducimos el sesgo de no respuesta al estimar un parámetro poblacional; en efecto al clasificar las unidades muestrales en H grupos e imputar dentro de cada grupo las unidades que no responden, el grupo es más homogéneo. Para cada grupo tendremos :

$\bar{Y}_i$ es la media muestral del grupo, que estima la media poblacional del grupo, denotada por μ_i . Sea w_i la proporción de unidades de la muestra que caen en el grupo, tal que $E(w_i) = W_i$, proporción de unidades en la población.

El sesgo para éste grupo, será : $E(\bar{Y}_i) - \mu_i = B_i$

Para el conjunto de grupos tendremos :

$\bar{Y} = \Sigma\ Y_i.w_i$ que será el estimador de la media poblacional μ, tal que $\mu = \Sigma\ \mu_i.W_i$

El sesgo del estimador será :

$$E(\bar{Y}) - \mu = E(\ \Sigma\ \bar{Y}_i.w_i\ ) - \Sigma\ \mu_i.W_i = \Sigma\ (\ E(\bar{Y}_i) - \mu_i\ ).W_i = \Sigma\ W_i.B_i$$

Como B_i se reduce, al ser el grupo homogéneo, también se reduce el del conjunto.

B.3. Donante

Está indicado cuando intervienen variables cuantitativas. En éste procedimiento se define una función distancia que mide el grado de proximidad entre cada posible registro donante y receptor. Una vez seleccionado el donante se imputa, en bloque, los valores del donante en los campos a imputar del receptor.

B.4. Imputación múltiple

Con la imputación, si P es un parámetro poblacional, obtenemos un estimador p de éste, siendo VAR(p) su varianza. Si repetimos el proceso un número determinado de veces I, obtenemos en cada ocasión los estimadores :

$$(\ p_i \ , VAR(p_i) \ )$$

La imputación múltiple (Rubin, 1986), contempla que el estimador de P, después de I imputaciones es :

$p^* = \sum_{i=1}^{I} p_i/I$, siendo su varianza VAR(p^*), la suma de estos dos términos :

* La media de las varianzas : $\sum_{i=1}^{I} VAR(p_i)/I$

* La varianza de los p_i, esto es : $\sum_{i=1}^{I} (p_i - p^*)^2/(I-1)$

CAPITULO II

CAPITULO II

ESTIMADORES POBLACIONALES, CON EL MARCO NO DEPURADO, A TRAVES DE UNA MUESTRA CON NO RESPUESTA

1. Introducción

En un sentido amplio la finalidad de una encuesta por muestreo es obtener información para satisfacer una necesidad definida; sin embargo, frecuentemente, el interés se centra en cuatro características del universo de la población bajo estudio. Estas son : Población total, media de la población, proporción y razón de la población. Las poblaciones que se consideran son finitas en el sentido de que el número de unidades que contiene es limitado.

El objetivo pues de una encuesta por muestreo es el de obtener una estimación, puntual o por intervalos, de algunos de los parámetros poblacionales (dejando al margen el interés en estos últimos años por estimar la forma o tipo de distribución de origen, utilizando técnicas no paramétricas, basadas en los procedimientos de núcleo Kernel).

El problema surge cuando, al seleccionar una muestra, algunas de las unidades no responden para una característica en particular. No es correcto basar los resultados de la encuesta únicamente en las unidades que respondieron, dado que quienes no lo hicieron son diferentes de los otros.

Dentro de los diferentes tipos de muestreo y considerando que hay elementos de la muestra que no responden, nosotros en éste capítulo hemos elegido tres de ellos :

Para el primero, **muestreo sin reemplazamiento de tipo aleatorio simple**, proponemos un conjunto de estimadores directos, cada uno con una hipótesis que lo caracteriza, para la media y el total, tanto de la población como de la población de los que responden y los comparamos en cuanto a precisión y exactitud (eficiencia y acuracidad, respectivamente).

En la elección de estos estimadores han influido diversos factores, siendo el principal de estos el considerar las diferentes situaciones en las que nos podemos encontrar, en muestras con unidades que no responden y su correspondiente tratamiento para determinar un estimador : Sustitución por cero; sustitución con una variable auxiliar, eliminación, eliminación pero considerando el tamaño de la

Con los otros dos métodos de muestreo elegidos, **muestreo estratificado y muestro de conglomerados bietápicos,** planteamos las mismas hipótesis que en los estimadores anteriores para cada uno de los estratos/conglomerados y proponemos estimadores directos tanto para la media y el total del conjunto poblacional como de la población de los que responden.

Así mismo, proponemos una generalización del estimador de Hansen y Hurwitz para estos dos últimos tipos de muestreo. Además, para el muestreo estratificado y con éste estimador, proponemos diversas consideraciones según que la estratificación sea proporcionada o desproporcionada, estudiando en éste último caso dos cuestiones:

a) La afijación óptima, como medio de asignar fracciones de muestreo a los estratos, con el objetivo de obtener la mínima varianza del estimador.

b) La afijación óptima con función de coste, como medio de asignar fracciones de muestreo y tasa de submuestreo a los estratos, de tal manera que el coste esperado sea mínimo para un valor determinado de la varianza del estimador.

La notación que vamos a emplear en el desarrollo de éste capítulo es, a parte de la indicada en el capítulo anterior, la que sigue :

A) Para un estrato / conglomerado "i"

N_i = Número de unidades, tal que $N = \sum_i N_i$

Y_{Ni} = Total poblacional, tal que $Y_N = \sum_i Y_{Ni}$

$\bar{Y}_{Ni} = Y_{Ni} / N_i$ = media poblacional

$W_{Ni} = N_i / N$ = ponderación

N_{i1} = Número de unidades que responden

N_{i2} = Número de unidades que no responden

Y_{Ni1} = Total poblacional de las unidades que responden

Y_{Ni2} = Total poblacional de las unidades que no responden

$\bar{Y}_{Ni1}$ = Media poblacional de las unidades que responden

$\bar{Y}_{Ni1}$ = Media poblacional de las unidades que no responden

$t_{Ni1} = N_{i1} / N_i$ = Tasa de respuesta

$t_{Ni1} = N_{i1} / N_i$ = Tasa de respuesta

$t_{Ni2} = N_{i2} / N_i$ = Tasa de no respuesta

B) Para la muestra del estrato / conglomerado "i"

n_i = Número de unidades de la muestra

$f_i = n_i / N_i$ = Fracción de muestreo

n_{i1} = Número de unidades de la muestra que responden

n_{i2} = Número de unidades de la muestra que no responden

$u_i = n_{i2} / k_i$ = Total de unidades de la submuestra para obtener información de los que no responden

k_i = Tasa de submuestra

Y_{ni1} = Total muestral de las unidades que responden

Y_{ni2} = Total muestral de las unidades que no responden

$\bar{Y}_{ni1}$ = Media muestral de las unidades que responden

$\bar{Y}_{ni1}$ = Media muestral de las unidades que no responden

$t_{ni1} = n_{i1} / n_i$ = Tasa de respuesta

$t_{ni2} = n_{i2} / n_i$ = Tasa de no respuesta

Por otro lado, si denotamos por Y una característica a estudiar en una población con N individuos y extraemos una muestra de extensión n sin reemplazamiento, de tipo aleatorio simple (m.a.s.), se sigue :

La media muestral $\bar{Y}_n$ es un estimador insesgado de la media poblacional $\bar{Y}_N$, siendo la varianza de éste estimador :

$$VAR(\bar{Y}_n) = 1/n \cdot (1 - n/N) \cdot S_N^2 = 1/n \cdot (1 - f) \cdot S_N^2$$

f = n/N es la fracción de muestreo y S_N^2 la varianza ajustada o cuasi varianza poblacional, cuya expresión es :

$$S_N^2 = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2}{N - 1}$$

La varianza ajustada muestral tiene por expresión :

$$S_n^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2}{n - 1}$$

y es un estimador insesgado de S_N^2 en éste tipo de muestreo

por lo que un estimador insesgado de $VAR(\bar{Y}_n)$ es :

$$\overset{*}{VAR}(\bar{Y}_n) = 1/n \, . \, (1 - f) \, . \, S_n^2$$

De forma análoga $\overset{*}{Y}_N = N.\bar{Y}_n$ es un estimador insesgado del total poblacional Y_N, siendo la varianza de éste estimador :

$$\overset{*}{VAR}(Y_N) = N^2 . \, VAR(\bar{Y}_n)$$

y una estimación insesgada de ésta :

$$\overset{*}{VAR}(\overset{*}{Y}_N) = N^2/n \, . \, (1 - f) . \, S_n^2$$

2. Estimadores directos o expandidos de la media y del total de una población finita, a través de una muestra aleatoria simple con no respuesta.

En lo que sigue en éste apartado estudiamos y proponemos diversos estimadores de la media y del total.

Con el fín de concretar las hipótesis de partida para cada estimador, los hemos denotado con letras mayúsculas A, B, C..., de modo que podamos identificarlos para un uso posterior.

Consideramos que el marco muestral, como conjunto de unidades de muestreo a partir de las cuales seleccionamos la muestra, está no depurado.

La depuración del marco conllevaría la eliminación de todos los individuos que no responden en el total de éste, siendo necesario conocer tales unidades.

2.1. Estimador A :

Si fijamos la fracción de muestreo f, el tamaño n de la muestra con no respuesta permanece fijo, por lo que el resultado es una muestra con igual probabilidad para cada elemento. Procederemos a asignar a cada unidad Y_i de la población el siguiente valor :

$$Y_i = \begin{cases} X_i & \text{Si la unidad responde} \\ 0 & \text{Si la unidad no responde} \end{cases}$$

Definición 2.1.1 Estimador directo o expandido de

a) la media

$$A{:}\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n}$$

b) el total : $A{:}\overset{*}{Y}_N = N.\bar{Y}_n$

que en lo que sigue, los expresaremos por $\bar{Y}_n$ e $\overset{*}{Y}_N$, respectivamente

2.1.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez / insesgadez de cada uno de ellos.

Proposición 2.1.2 $\bar{Y}_n$ es un estimador insesgado de $\bar{Y}_N$ y sesgado para $\bar{Y}_{N1}$, siendo el sesgo igual a :

$$-1.t_{N2}.\ \bar{Y}_{N1}$$

Demostración :

$$E\ (\ \bar{Y}_n\ ) = \bar{Y}_N \quad (\ m.a.s.\ ) \qquad (1)$$

Por otro lado :

$$\bar{Y}_n = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i=1}^{n1} X_i}{n} = n_1 / n\ .\ \bar{Y}_{n1} = t_{n1}.\bar{Y}_{n1} \qquad (2)$$

Como $E(\bar{Y}_{n1}) = \bar{Y}_{N1}$ y $E(t_{n1}) = t_{N1}$ y (2) ==> $E(\bar{Y}_n) = t_{N1}\ \bar{Y}_{N1}$

que es otra forma de expresar (1)

El sesgo del estimador para el parámetro poblacional $\bar{Y}_{N1}$ viene definido por :

$$sesgo(\bar{Y}_n) = E(\bar{Y}_n) - \bar{Y}_{N1} = -1.\ \bar{Y}_{N1}\ (\ 1 - t_{N1}\ ) = -1.t_{N2}.\ \bar{Y}_{N1}\ ;$$

es decir, es directamente proporcional a la tasa de no respuesta poblacional

c.q.d.

Proposición 2.1.3 $\overset{*}{Y}_N$ es un estimador insesgado de Y_N y de Y_{N1}

Demostración :

$$E(\overset{*}{Y}_N) = E(N.\bar{Y}_n) = N \; . \; E(\bar{Y}_n) = N.\bar{Y}_N = Y_N \quad (\text{proposición 2.1.2})$$

Utilizando la misma proposición, deducimos :

$$E(\overset{*}{Y}_N) = E(N.\bar{Y}_n) = N \; . \; E(\bar{Y}_n) = N. \; N_1/N. \; \bar{Y}_{N1} = N_1. \; \bar{Y}_{N1} = Y_{N1}$$

c.q.d.

2.1.2. Varianza de los estimadores y sus estimaciones

Damos una expresión de la varianza de cada estimador y de una estimación de ésta. Proponemos despues una descomposición de la varianza en suma de términos que dependen sólo de las unidades que responden.

Proposición 2.1.4 Las varianzas de $\bar{Y}_n$ e $\overset{*}{Y}_N$, son respectivamente

$$VAR \; (\bar{Y}_n) = 1/n \; . \; (1 - f). \; S_N^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR} \; (\bar{Y}_n) = 1/n \; . \; (1 - f). \; S_n^2$$

$$VAR \; (\overset{*}{Y}_N) = N^2/n \; . \; (1 - f). \; S_N^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR} \; (\overset{*}{Y}_N) = N^2/n \; . \; (1 - f). \; S_n^2$$

Demostración :

Respecto de la varianza del estimador $\bar{Y}_n$, como la muestra de extensión n es la de una m.a.s., tendrá la varianza de ésta según lo indicado al comienzo de éste apartado.

Respecto de la varianza de $\overset{*}{Y}_N$, se tiene :

$$VAR \; (\overset{*}{Y}_N) = VAR \; (\; N.\bar{Y}_n \;) = N^2. \; VAR \; (\bar{Y}_n)$$

$$\overset{*}{VAR} \; (\overset{*}{Y}_N) = \overset{*}{VAR} \; (\; N.\bar{Y}_n \;) = N^2. \; \overset{*}{VAR} \; (\bar{Y}_n)$$

y de la proposición 2.1.2, se deducen las expresiones antes indicadas.

c.q.d.

Definición 2.1.5 La varianza de las unidades que responden en la muestra y de las unidades que responden en el total de la población, así como las varianzas ajustadas respectivas, vienen expresadas por :

$$VAR_{n1} = 1/n_1 \; . \; \sum_{i=1}^{n1} (\; X_i - \bar{Y}_{n1} \;)^2 =$$

$$1/n_1 \left(\sum_{i=1}^{n1} X_i^2 - 2.\ \bar{Y}_{n1}.\ \sum_{i=1}^{n1} X_i + \sum_{i=1}^{n1} \bar{Y}_{n1}^2 \right) =$$

dado que $\bar{Y}_{n1} = 1/n_1 .\ \sum_{i=1}^{n1} X_i$, se sigue

$$1/n_1 \left(\sum_{i=1}^{n1} X_i^2 - n_1 .\ \bar{Y}_{n1}^2 \right) =$$

$$1/n_1 \left(\sum_{i=1}^{n1} X_i^2 \right) - \bar{Y}_{n1}^2$$

$$VAR_{N1} = 1/N_1 .\ \sum_{i=1}^{N1} (X_i - \bar{Y}_{N1})^2 = \ldots.(\text{como el anterior})$$

$$= 1/N_1 \left(\sum_{i=1}^{N1} X_i^2 \right) - \bar{Y}_{N1}^2$$

Así mismo, se tiene que :

$$S_{N1}^2 = N_1/(N_1-1) .\ VAR_{N1} \quad y \quad S_{n1}^2 = n_1/(n_1-1) .\ VAR_{n1}$$

varianza ajustada del total de unidades que responden en la población y en la muestra, respectivamente.

Utilizamos estas expresiones para descomponer la varianza de los estimadores $\overset{*}{\bar{Y}}_N$ e $\bar{Y}_n$ en suma de dos términos que dependen sólo de las unidades que responden, tal y como sugiere Kish (1965)

Proposición 2.1.6 Para el estimador $\bar{Y}_n$, su varianza podemos expresarla de una de estas dos formas :

$$VAR (\bar{Y}_n) = (1-f) .\ 1/n .\ N_1/(N-1) .\ (VAR_{N1} + t_{N2}\ \bar{Y}_{N1}^2)$$

ó (1)

$$VAR(\bar{Y}_n) \div (1-f) .\ 1/n .\ t_{N1} .\ (S_{N1}^2 + t_{N2} .\ \bar{Y}_{N1}^2)$$

y su estimación por :

$$\overset{*}{VAR} (\bar{Y}_n) = (1-f) .\ 1/n .\ n_1/(n-1) .\ (VAR_{n1} + t_{n2}\ \bar{Y}_{n1}^2)$$

ó (2)

$$\overset{*}{VAR}(\bar{Y}_n) \div (1-f).\ 1/n.\ t_{n1} .\ (S_{n1}^2 + t_{n2} .\ \bar{Y}_{n1}^2)$$

Demostración :

Descomponemos S_N^2, en la forma :

$$(N-1).S_N^2 = \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2 =$$

$$\sum_{i=1}^{N} Y_i^2 - 2.\ \bar{Y}_N.\ \sum^{N} Y_i + \sum^{N} \bar{Y}_N^2 \quad =$$

dado que $\bar{Y}_N = 1/N.\ \sum_{i=1}^{N} Y_i$, se sigue

$$\sum_{i=1}^{N} Y_i^2 - N\ .\ \bar{Y}_N^2 \quad =$$

$$(\sum_{i=1}^{N} Y_i^2 - N_1\ .\ \bar{Y}_{N1}^2) + (N_1.\ \bar{Y}_{N1}^2 - N\ .\ \bar{Y}_N^2) \quad =$$

Como $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N1} X_i^2$

el primer término es VAR_{N1} segun la definición 2.1.5 ; por otro lado,

$\bar{Y}_N = 1/N.\ \sum_{i=1}^{N} Y_i = N_1/N.\ \bar{Y}_{N1}$, se sigue

$$N_1.\ VAR_{N1} + N_1.\ \bar{Y}_{N1}^2\ .\ (1 - N_1/N) =$$

$$N_1.(VAR_{N1} + t_{N2}.\ \bar{Y}_{N1}^2) =$$

tenemos pués que

$$S_N^2 = 1/(N-1).\ N_1.(VAR_{N1} + t_{N2}.\ \bar{Y}_{N1}^2) \quad (3)$$

(Para otra demostración más inmediata, puede utilizarse el corolario 3.2.8 del capítulo I, teniendo en cuenta la situación en la que estamos)

Para la segunda descomposición de (1) utilizamos la proposición 3.2.7 del capítulo I, teniendo en cuenta que :

$$\bar{Y}_{N2} = 0,\ \bar{Y}_N = N_1/N\ .\ \bar{Y}_{N1} \quad y \quad S_{N2}^2 = 0$$

por lo que obtenemos la expresión para la varianza ajustada poblacional de :

$$S_N^2 = (N_1 - N)/(N-1)\ S_{N1}^2 + N_1\ /(N-1).(1 - N_1/N)\ \bar{Y}_{N1}^2 \quad (4)$$

teniendo en cuenta que $N_1-1 \div N$ y $N-1 \div N$ y la definición de tasa de respuesta y no respuesta poblacional, se obtiene de (4):

$$S_N^2 \div t_{N1} \cdot S_{N1}^2 + t_{N1} \cdot t_{N2} \cdot \bar{Y}_{N1}^2 \qquad (5)$$

Finalmente, tomando como referencia la expresión de $VAR(\bar{Y}_n)$ en la proposición 2.1.2 y sustituyendo en ella las expresiones (3) o (5), obtenemos la descomposición de la varianza propuesta en (1)

La descomposición de $\overset{*}{VAR}(\bar{Y}_n)$ es análoga a la anterior

c.q.d.

Proposición 2.1.7 Para el estimador $\overset{*}{Y}_N$, su varianza podemos expresarla como :

$$VAR(\overset{*}{Y}_N) = (1-f).\ N^2/n\ .\ N_1/(N-1)\ .\ (VAR_{N1} + t_{N2}\ \bar{Y}_{N1}^2)$$

ó

$$VAR(\overset{*}{Y}_n) \div (1-f).\ N^2/n.\ t_{N1}\ .\ (S_{N1}^2 + t_{N2}\ .\ \bar{Y}_{N1}^2)$$

y su estimación por :

$$\overset{*}{VAR}(\overset{*}{Y}_N) = (1-f).\ N^2/n\ .\ n_1/(n-1)\ .\ (VAR_{n1} + t_{n2}\ \bar{Y}_{n1}^2)$$

ó

$$\overset{*}{VAR}(\bar{Y}_n) \div (1-f).\ 1/n.\ t_{n1}\ .\ (S_{n1}^2 + t_{n2}\ .\ \bar{Y}_{n1}^2)$$

Demostración :

Es consecuencia de la proposición 2.1.4 y de la descomposición de S_N^2 y S_n^2 en la proposición 2.1.6

c.q.d.

2.2. Estimador B :

Fijamos la extensión de la muestra en n unidades y sustituimos las unidades que no contestan Y_i en la muestra, por otras elegidas al azar de entre las que se saben que contestan X_i, a través de una variable auxiliar (se conoce N_1).

Definición 2.2.8 Estimador directo o expandido de

a) la media

$$B{:}\bar{Y}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

b) el total : $B:\overset{*}{Y}_{N1} = N_1 \cdot \bar{Y}_n$

que en lo que sigue, los expresaremos por $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, respectivamente

2.2.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez ó insesgadez de cada uno de ellos.

Proposición 2.2.9 $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$ son estimadores insesgados de la media y del total de la población de los que responden. Son sesgados respecto de la media y total poblacional, siendo el sesgo del primer estimador igual a :

$$t_{N2} \left(\bar{Y}_{N1} - \bar{Y}_{N2} \right)$$

Demostración :

$$E(\bar{Y}_n) = 1/n \sum_{i=1}^{n} E(X_i) = 1/n \cdot n \cdot \sum_{i=1}^{N1} X_i \cdot 1/N_1 = \bar{Y}_{N1}$$

$$E(\overset{*}{Y}_{N1}) = N_1 \cdot E(\bar{Y}_n) = N_1 \cdot \bar{Y}_{N1} = Y_{N1}$$

El sesgo del estimador, para la media poblacional, viene expresado por:

sesgo $(\bar{Y}_n) = \bar{Y}_{N1} - \bar{Y}_N$, y de acuerdo con la expresión obtenida en la proposición 3.1.2 del capítulo I para la media poblacional, se tiene :

$$\text{sesgo } (\bar{Y}_n) = t_{N2} \left(\bar{Y}_{N1} - \bar{Y}_{N2} \right)$$

c.q.d.

2.2.2. Varianza de los estimadores y sus estimaciones

Damos una expresión de la varianza de cada estimador y de una estimación de ésta. Dado que la expresión de las varianzas depende sólo de las unidades que responden, no procede una descomposición como la indicada en las proposiciones 2.1.6 y 2.1.7 sino tan sólo relacionar la varianza ajustada S_{N1}^2 con VAR_{N1} y S_{n1}^2 con VAR_{n1} a través de las relaciones usuales.

Proposición 2.2.10 Las varianzas de $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, son respectivamente

$$VAR\ (\bar{Y}_n) = 1/n \cdot (1 - n/N_1) \cdot S_{N1}^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR}\ (\bar{Y}_n) = 1/n \cdot (1 - n/N_1) \cdot S_{n1}^2$$

$$VAR\ (\overset{*}{Y}_{N1}) = N_1^2/n\ .\ (1 - n/N_1).\ S_{N1}^2 \qquad \text{estimada por :}$$

$$VAR\ (\overset{*}{Y}_{N1}) = N_1^2/n\ .\ (1 - n/N_1).\ S_{n1}^2$$

Demostración :

$$VAR\ (\bar{Y}_n) = E[\bar{Y}_n - E(\bar{Y}_n)]^2 = E\ [\bar{Y}_n - \bar{Y}_{N1}]^2 =$$

$$E\ [1/n\ .\ \sum_{i=1}^{n} X_i - n/n.\bar{Y}_{N1}]^2 =$$

$$1/n^2\ .\ E[\ \sum_{i=1}^{n}(X_i - \bar{Y}_{N1})\ ]^2$$

Si $Z_i = X_i - \bar{Y}_{N1}$, tendremos :

$$VAR\ (\bar{Y}_n) = 1/n^2\ .\ E[\ \sum_{i=1}^{n} Z_i\ ]^2 = 1/n^2\ E\ (\sum_{i=1}^{n} Z_i^2 + 2 \sum_{i=1}^{n} \sum_{j>i} Z_i\ .\ Z_j\ ) =$$

$$1/n^2\ [\ E \sum_{i=1}^{n} Z_i^2 + 2\ .\ C_{n,2}\ .\ E(Z_i.\ Z_j)\ ] \qquad (1)$$

donde $C_{n,2}$, representa las combinaciones de n unidades tomadas de dos en dos. Como :

$$E\ (\ Z_i^2\ ) = 1/N_1\ .\ \sum_{i=1}^{N1} Z_i^2 \Longrightarrow E\ (\ \sum_{i=1}^{n} Z_i^2\ ) = n/N_1\ .\ \sum_{i=1}^{N1} Z_i^2 \qquad (2)$$

$$E\ (Z_i.Z_j) = E_1\ (Z_i.E_2(\ Z_j\ )) \qquad (3)$$

siendo $E_2(\ Z_j\ ) = E\ (Z_j\ /\ Z_i)$ (Z_j / Z_i es Z_j condicionada a Z_i) y

$$E_2(\ Z_j\ ) = 1/(N_1-1)\ .\ \sum_{i<>j}^{N1} Z_j = -1/(N_1-1).\ Zi\ ,\ \text{dado que}\ Z_i = X_i - \bar{Y}_{N1}$$

y $\sum_{i=1}^{N1} Z_i = \sum_{i=1}^{N1} X_i - N_1.Y_{N1} = 0$ y en consecuencia : $\sum_{i=1}^{N1} Z_i = 0 = \sum_{i<>j} Z_i + Z_i$

Podemos expresar (3) , en la forma :

$$E\ (\ Z_i\ .\ Z_j) = E_1\ (\ -1/(N_1-1)\ .\ Z_i^2\ ) = -1/(N_1-1)\ .\ 1/N_1.\ \sum_{i=1}^{N1} Z_i^2 \qquad (4)$$

Así pues, sustituyendo (2) y (4) en (1) :

$$VAR\ (\bar{Y}_n) = 1/n^2.[\ n/N_1\ .\ \sum_{i=1}^{N1} Z_i^2 + n.(n-1)\ .\ -1/(N_1-1)\ .\ 1/N_1.\ \sum_{i=1}^{N1} Z_i^2]$$

$$= 1/n.\ 1/N_1\ .\ [\ 1-\ (n-1)/(N_1-1)\ ].\ \sum_{i=1}^{N1} Z_i^2$$

$$= 1/n.\ 1/N_1\ .\ (N_1\ -\ n)/(N_1-1).\ \sum_{i=1}^{N1} Z_i^2$$

$$= 1/n.\ (1\ -\ n\ /N_1)\ .\ \sum_{i=1}^{N1} Z_i^2/(N_1-1)$$

$$= 1/n.\ (1\ -\ n\ /N_1)\ .\ S_{N1}^2 \qquad \text{c.q.d.}$$

Para obtener la estimación de $VAR\ (\bar{Y}_n)$, sólo hay que tener en cuenta que $E(S_{n1}^2) = S_{N1}^2$

Respecto de la varianza de $\overset{*}{Y}_{N1}$, se tiene :

$$VAR\ (\overset{*}{Y}_{N1}) = VAR\ (\ N_1.\bar{Y}_n\ ) = N_1^2.\ VAR\ (\bar{Y}_n)$$

$$\overset{*}{VAR}\ (\overset{*}{Y}_{N1}) = \overset{*}{VAR}\ (\ N_1.\bar{Y}_n\ ) = N_1^2.\ \overset{*}{VAR}\ (\bar{Y}_n)$$

y de la primera parte de ésta proposición se deducen las expresiones antes indicadas.

2.3. Estimador C :

Fijando en n la extensión de la muestra, eliminamos de ésta las unidades que no contestan (Y_i), reduciendo a n_1 el número de unidades muestrales, que son las que contestan (X_i) (se conoce N_1).

Definición 2.3.11 Estimador directo o expandido de

a) la media

$$C:\bar{Y}_n = \frac{\sum_{i=1}^{n1} X_i}{n_1}$$

b) el total : $C:\overset{*}{Y}_{N1} = N_1\ .\ \bar{Y}_n$

que en lo que sigue, los expresaremos por $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, respectivamente

2.3.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez / insesgadez de cada uno de ellos.

Proposición 2.3.12 $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$ son estimadores insesgados de la media y del total de la población de los que responden. Son sesgados respecto de la media y total poblacional.

Demostración :

Análoga a la proposición 2.2.9

c.q.d.

2.3.2. Varianza de los estimadores y sus estimaciones

Damos una expresión de la varianza de cada estimador y, de una estimación de ésta. Como la expresión de las varianzas depende sólo de las unidades que responden, no procede una descomposición como la indicada en las proposiciones 2.1.6 y 2.1.7 sino tan sólo relacionar la varianza ajustada S_{N1}^2 con VAR_{N1} y S_{n1}^2 con VAR_{n1} a través de las relaciones usuales.

Proposición 2.3.13 Las varianzas de $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, son respectivamente

$$VAR\ (\bar{Y}_n) = 1/n_1\ .\ (1 - n_1/N_1\ ).\ S_{N1}^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR}\ (\bar{Y}_n) = 1/n_1\ .\ (1 - n_1/N_1\ ).\ S_{n1}^2$$

$$VAR\ (\overset{*}{Y}_{N1}) = N_1^2/n_1\ .\ (1 - n_1/N_1).\ S_{N1}^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR}\ (\overset{*}{Y}_{N1}) = N_1^2/n_1\ .\ (1 - n_1/N_1).\ S_{n1}^2$$

Demostración :

La primera parte es análoga a la indicada en la proposición 2.2.10, cambiando n por n_1

Para la segunda parte, las expresiones indicadas se obtienen al utilizar las relaciones entre las varianzas de ambos estimadores.

c.q.d.

2.4. Estimador D :

Nosotros proponemos que fijando en n la extensión de la muestra, eliminamos de ésta las unidades que no contestan, pero teniendo en cuenta que la muestra es de extensión n (se conoce N_1).

Definición 2.4.14 Estimador directo o expandido de

a) la media (denotamos por X_i las unidades que contestan)

$$D:\bar{Y}_n = \frac{\sum_{i=1}^{n1} X_i}{n}$$

b) el total : $D:\overset{*}{Y}_{N1} = N_1 \cdot \bar{Y}_n$

en lo que sigue los expresaremos por $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, respectivamente

2.4.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez ó insesgadez de cada uno de ellos.

Proposición 2.4.15 $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$ son estimadores sesgados de la media y del total de la población de los que responden y de la media y total poblacional; siendo el sesgo del primero, a la hora de estimar la media de los que responden y la media poblacional, respectivamente

$$-1 \cdot t_{n2} \cdot \bar{Y}_{N1} \qquad y \qquad (t_{n1} - t_{N1}).\bar{Y}_{N1} - t_{N2} \cdot \bar{Y}_{N2}$$

Demostración :

$$E(\bar{Y}_n) = 1/n \sum_{i=1}^{n1} E(X_i) = 1/n \cdot n_1 \cdot \sum_{i=1}^{N1} X_i \cdot 1/N_1 = n_1/n. \bar{Y}_{N1} = t_{n1}. \bar{Y}_{N1}$$

$$E(\overset{*}{Y}_{N1}) = N_1. E(\bar{Y}_n) = t_{n1}. N_1. \bar{Y}_{N1} = t_{n1}. Y_{N1}$$

El sesgo del estimador de la media de los que responden, viene definido por :

$$sesgo(\bar{Y}_n) = E(\bar{Y}_n) - \bar{Y}_{N1} = -1 \cdot (1 - n_1/n) \cdot \bar{Y}_{N1} = -1 \cdot t_{n2} \cdot \bar{Y}_{N1} ;$$

es decir, es directamente proporcional a la tasa de no respuesta muestral t_{n2}

El sesgo del estimador de la media poblacional, viene definido por :

$$sesgo(\bar{Y}_n) = E(\bar{Y}_n) - \bar{Y}_N = t_{n1}.\bar{Y}_{N1} - \bar{Y}_N = (t_{n1} - t_{N1}).\bar{Y}_{N1} - t_{N2} \cdot \bar{Y}_{N2}$$

teniendo en cuenta la proposición 3.1.2 del capítulo I .

2.4.2. Varianza de los estimadores y sus estimaciones

Damos una expresión de la varianza de cada estimador y, de una estimación de ésta. Como la expresión de las varianzas depende sólo de las unidades que responden, no procede una descomposición como la indicada en las proposiciones 2.1.6 y 2.1.7 sino tan sólo relacionar la varianza ajustada S_{N1}^2 con VAR_{N1} y S_{n1}^2 con VAR_{n1} a través de las relaciones usuales.

Proposición 2.4.16 Las varianzas de $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, son respectivamente :

$$VAR\ (\bar{Y}_n) = 1/n\ .\ (1 - n_1/N_1\ ).\ n_1/n.\ S_{N1}^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR}\ (\bar{Y}_n) = 1/n\ .\ (1 - n_1/N_1\ ).\ n_1/n.\ S_{n1}^2$$

$$VAR\ (\overset{*}{Y}_{N1}) = N_1^2/n\ .\ (1 - n_1/N_1).\ n_1/n.\ S_{N1}^2 \qquad \text{estimada por :}$$

$$\overset{*}{VAR}\ (\overset{*}{Y}_{N1}) = N_1^2/n\ .\ (1 - n_1/N_1).\ n_1/n.\ S_{n1}^2$$

Demostración :

$$VAR\ (\bar{Y}_n) = E[\bar{Y}_n - E(\bar{Y}_n)]^2 = E\ [\bar{Y}_n - n_1/n.\bar{Y}_{N1}]^2 =$$

$$E\ [1/n\ .\ \sum_{i=1}^{n1} X_i - n_1/n.\bar{Y}_{N1}]^2 =$$

$$1/n^2\ .\ E[\ \sum_{i=1}^{n1}(X_i - \bar{Y}_{N1})\ ]^2$$

Si $Z_i = X_i - \bar{Y}_{N1}$, tendremos :

$$VAR\ (\bar{Y}_n) = 1/n^2\ .\ E[\ \sum_{i=1}^{n1} Z_i\ ]^2 = 1/n^2\ E\ (\sum_{i=1}^{n1} Z_i^2 + 2 \sum_{i=1\ j>i}^{n1} \Sigma\ Z_i\ .\ Z_j\ ) =$$

$$1/n^2\ [\ E \sum_{i=1}^{n1} Z_i^2 + 2.\ C_{n1,2}\ .\ E(Z_i.\ Z_j)\ ] \qquad (1)$$

donde $C_{n1,2}$ representa las combinaciones de n_1 unidades tomadas de dos en dos. Como :

$$E\ (\ Z_i^2\ ) = 1/N_1\ .\ \sum_{i=1}^{N1} Z_i^2$$

$E\ (Z_i.Z_j) = E_1\ (Z_i.E_2(\ Z_j))$, siendo $E_2\ (Z_j) = E\ (Z_j\ /\ Z_i)$ y

$E_2(\ Z_j\ ) = 1/(N_1-1)\ .\ \sum_{i<>j} Z_j = -1/(N_1-1).\ Zi$, dado que $Z_i = X_i - \bar{Y}_{N1}$

y $\sum_{i=1}^{N1} Z_i = \sum_{i=1}^{N1} X_i - N_1.Y_{N1} = 0$ y en consecuencia : $\sum_{i=1}^{N1} Z_i = 0 = \sum_{i<>j} Z_i + Z_i$

deducimos que :

$$E\ (\ Z_i.Z_j) = E_1\ (\ -1/(N_1-1)\ .\ Z_i^2\ ) = -1/(N_1-1)\ .\ 1/N_1.\ \sum_{i=1}^{N1} Z_i^2$$

Así pues, sustituyendo en (1) :

$$VAR\ (\bar{Y}_n) = 1/n^2.[\ n_1/N_1\ .\ \sum_{i=1}^{N1} Z_i^2 + n_1.(n_1-1)\ .\ -1/(N_1-1)\ .\ 1/N_1.\ \sum_{i=1}^{N1} Z_i^2]$$

$$= n_1/n^2.\ 1/N_1\ .\ [\ 1-\ (n_1-1)/(N_1-1)\ ].\ \sum_{i=1}^{N1} Z_i^2$$

$$= n_1/n^2.\ 1/N_1\ .\ (N_1 - n_1)/(N_1-1).\ \sum_{i=1}^{N1} Z_i^2$$

$$= n_1/n^2.\ (1 - n_1\ /N_1)\ .\ \sum_{i=1}^{N1} Z_i^2/(N_1-1)$$

$$= n_1/n^2.\ (1 - n_1\ /N_1)\ .\ S_{N1}^2$$

Para obtener la estimación de $VAR\ (\bar{Y}_n)$, sólo hay que tener en cuenta que $E(S_{n1}^2) = S_{N1}^2$. Para la segunda parte, las expresiones indicadas se obtienen al utilizar las relaciones entre las varianzas de ambos estimadores.

c.q.d.

2.5. Estimador E :

En una muestra de extensión n, nos proponemos sustituir las unidades que no contestan Y_i por el valor medio de las que contestan X_i (Bailar et al. 1978). Como consecuencia de ésta imputación el total muestral será :

$$Y_n = \sum_{i=1}^{n1} X_i + n_2\ .\ \bar{Y}_{n1} = n_1\ .\ \bar{Y}_{n1} + n_2\ .\ \bar{Y}_{n1} = n\ .\ \bar{Y}_{n1}$$

Luego $\bar{Y}_n = Y_n/n = \bar{Y}_{n1}$ y, en consecuencia, damos la siguiente definición.

Definición 2.5.17 Estimador directo o expandido de

a) la media

$$E{:}\bar{Y}_n = \frac{\sum_{i=1}^{n1} X_i}{n_1}$$

b) el total : $E{:}Y_{N1} = N_1 \cdot Y_n$

que en lo que sigue, los expresaremos por $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, respectivamente

2.5.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez / insesgadez de cada uno de ellos.

Proposición 2.5.18 $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$ son estimadores insesgados de la media Y_{N1} y del total Y_{N1} de la población de los que responden y sesgados de la media y total poblacional

Demostración :

Como la proposición 2.3.12

2.5.2. Varianza de los estimadores y sus estimaciones

Veamos una expresión de las varianzas de ambos estimadores y luego una relación entre la varianza muestral despues de imputar los datos con la varianza muestral de sólo los elementos que responden.

Proposición 2.5.19 Las varianzas de $\bar{Y}_n$ e $\overset{*}{Y}_{N1}$, son respectivamente

$$VAR\ (\bar{Y}_n) = 1/n_1 \cdot (1 - n_1 / N_1). S_{N1}^2 \qquad \text{estimada por :}$$
$$\overset{*}{VAR}\ (\bar{Y}_n) = 1/n_1 \cdot (1 - n_1 / N_1). S_{n1}^2$$

$$VAR\ (\overset{*}{Y}_{N1}) = N_1^2/n_1 \cdot (1 - f). S_{N1}^2 \qquad \text{estimada por :}$$
$$\overset{*}{VAR}\ (\overset{*}{Y}_{N1}) = N_1^2/n_1 \cdot (1 - f). S_{n1}^2$$

Demostración :

$$VAR\ (\bar{Y}_n) = E[\bar{Y}_n - E(\bar{Y}_n)]^2 = E\ [\bar{Y}_n - \bar{Y}_{N1}]^2 = E\ [\bar{Y}_{n1} - \bar{Y}_{N1}]^2 = VAR\ (\bar{Y}_{n1})$$

y de acuerdo con la proposición 2.3.13, se tiene lo indicado.

Proposición 2.5.20 La varianza muestral se infraestima por el factor t_{n1} al imputar con la media los campos en blanco. Así mismo, la varianza ajustada se infraestima por el factor $(n_1-1)/(n-1)$

Demostración :

$$VAR_n = 1/n \cdot \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2 = 1/n \cdot \sum_{i=1}^{n} (Y_i - \bar{Y}_{n1})^2 \quad \text{pues} \quad \bar{Y}_n = \bar{Y}_{n1}$$

$$= 1/n \cdot \left[\sum_{i=1}^{n_1} (X_i - \bar{Y}_{n1})^2 + \sum_{i=1+n1}^{n} (Y_i - \bar{Y}_{n1})^2 \right]$$

$$= 1/n \cdot \sum_{i=1}^{n_1} (X_i - \bar{Y}_{n1})^2 \text{ , pues } \sum_{i=1+n1}^{n} (Y_i - \bar{Y}_{n1})^2 = 0 \text{ al sustituir } Y_i = \bar{Y}_{n1}$$

$$= n_1/n \cdot 1/n_1 \cdot \sum_{i=1}^{n_1} (X_i - \bar{Y}_{n1})^2 = n_1/n \cdot VAR_{n1}$$

Por otro lado :

$$S_n^2 = n/(n-1) \cdot VAR_n = n/(n-1) \cdot n_1/n \cdot VAR_{n1}$$

$$= n/(n-1) \cdot n_1/n \cdot (n_1-1)/n_1 \cdot S_{n1}^2$$

$$= (n_1-1)/(n-1) \cdot S_{n1}^2$$

que demuestra la relación

2.6. Estimador de Hansen y Hurwitz (estimador F) :

De entre las unidades que no responden n_2 procedemos a realizar un submuestreo, obteniendo información de una muestra de extensión :

$$u = n_2 / k$$

siendo k la tasa de submuestra

Teorema 2.6.21 Un estimador insesgado de la media poblacional es :

$$F:\bar{Y}_n = \frac{n_1 . \bar{Y}_{n1} + n_2 . \bar{Y}_u}{n}$$

que en lo que sigue, lo expresaremos por $\bar{Y}_n$ y siendo :

$\bar{Y}_{n1}$ la media muestral de los que responden

$\bar{Y}_u$ la media de la submuestra

Con una varianza :

$$VAR\ (\ \bar{Y}_n\ ) = \frac{k-1}{n}\ t_{N2}\ S_{N2}^2 + \frac{1}{n}\left(1 - \frac{n}{N}\right) S_N^2$$

siendo, S_{N2}^2 y S_N^2 las varianzas ajustadas para el total de unidades N_2 y N respectivamente :

$$S_{N2}^2 = 1/(N_2-1) \cdot \sum_{i=1}^{N2} (Y_i - \bar{Y}_{N2})^2 \; ; \quad S_N^2 = 1/(N-1) \cdot \sum_{i=1}^{N} (Y_i - \bar{Y}_N)^2$$

y t_{N2} la tasa poblacional de no respuesta

Demostración :

Hansen y Hurwitz (1946) contempla una muestra aleatoria simple como marco de sus resultados y demuestra éste teorema.

Proposición 2.6.22 Un estimador de VAR $(\bar{Y}_n)$ es

$$\overset{*}{VAR}\,(\bar{Y}_n) = \frac{k-1}{n}\, t_{n2}\, S_{n2}^2 + \frac{1}{n}\left(1 - \frac{n}{N}\right) S_n^2$$

siendo :

$$S_{n2}^2 = 1/(n_2-1) \cdot \sum_{i=1}^{n2} (Y_i - \bar{Y}_{n2})^2 \; ; \quad S_n^2 = 1/(n-1) \cdot \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$$

y t_{n2} la tasa muestral de no respuesta

Demostración :

Dado que $E(\, t_{n2} \cdot S_{n2}^2\,) = t_{N2} \cdot S_{N2}^2$ y $E(\, S_n^2\,) = S_N^2$ se sigue

$$E\,(\,\overset{*}{VAR}\,(\bar{Y}_n)\,) = VAR\,(\bar{Y}_n)$$

c.q.d.

Cochran y Rao, proponen un estimador para la varianza del estimador de la media poblacional de Hansen y Hurwitz, del que sólo indicamos su expresión

Proposición 2.6.23 Un estimador de VAR $(\bar{Y}_n)$ que sólo depende de los elementos muestrales es

$$\begin{aligned}\overset{*}{VAR}\,(\bar{Y}_n) = {} & \frac{(N-n)\,(n_1-1)}{N\,(n-1)} \cdot t_{n1} \cdot S_{n1}^2/n_1 + \\ & + \frac{(N-1)\,(n_2-1) - (n-1)\,(u-1)}{N\,(n-1)} \cdot t_{n2} \cdot S_u^2/u + \\ & + \frac{(N-n)}{N\,(n-1)} \cdot \left[\, t_{n1}.(\bar{Y}_{n1} - \bar{Y}_n)^2 + t_{n2}.(\bar{Y}_u - \bar{Y}_n)^2 \,\right]\end{aligned}$$

donde $S_u^2 = 1/(u-1) \cdot \sum_{i=1}^{u} (Y_i - \bar{Y}_u)^2$

2.7. Estimador de duplicación (estimador G) :

Consideramos que n_1, número de unidades que responden en la muestra de extensión n, es tal que $n_1 \geq n/2$. Como consecuencia, $n_1 > n_2$ y las unidades que responden serán utilizadas una sola vez y además no todas. En concreto :

Número de unidades duplicadas de entre las n_1 : $n_2 = n - n_1$

Número de unidades no duplicadas de entre las n_1 : $n_1 - n_2 = 2\, n_1 - n$

Como ya indicamos en el apartado 4.5 del capítulo I, el número de subconjuntos aleatoriamente seleccionados de unidades que responden para ser duplicados es de : $C\ n_1,\ n_2$ (C = combinaciones). Nosotros, una vez elegido éste, procederemos a ordenarlo, escribiendo en primer lugar las que no son duplicadas y luego las que son. De acuerdo con lo indicado anteriormente, tendremos :

$X_1, X_2, \ldots\ldots, X_{2n1-n},$ $\quad$ $X_{2n1-n+1}, X_2, \ldots\ldots, X_{n1}$

(no duplicadas) $\quad$ (duplicadas)

Definición 2.7.24 El estimador $\bar{Y}_n$ de la media poblacional, antes, durante y después de de la imputación, quedará (X_i e Y_i, representan las unidades que responden y las que no responden. X_i^*, las unidades imputadas ; en nuestro caso, duplicadas) como sigue :

$$\bar{Y}_n = 1/n \left[\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i \right]$$

$$G{:}\bar{Y}_n = 1/n \left[\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} X_i^* \right]$$

$$G{:}\bar{Y}_n = 1/n \left[\sum_{i=1}^{2n_1-n} X_i + 2. \sum_{i=2n_1-n+1}^{n_1} X_i \right] = 1/n \left[\sum_{1}^{n_1} X_i + \sum_{2n_1-n+1}^{n_1} X_i \right]$$

que en lo que sigue, lo expresaremos por $\bar{Y}_n$

2.7.1 Sesgo

El estimador propuesto estima a la media de la población y de los que responden. Estudiamos la sesgadez / insesgadez de éste.

Proposición 2.7.25 $\bar{Y}_n$ es un estimador insesgado de la media de los que responden y sesgado de la media poblacional, siendo el sesgo igual a:

$$t_{N2} \; (\; \bar{Y}_{N1} - \bar{Y}_{N2} \;)$$

Demostración :

$$E(\bar{Y}_n) = 1/n \; E \; [\sum_{i=1}^{n_1} X_i + \sum_{i=2n_1-n+1}^{n_1} X_i \;] = 1/n \; [\; n_1 \; E(X_i) + (n-n_1) \; E(X_i) \;] =$$

$$= E(X_i) = 1/N_1 \sum_{1}^{N} X_i = \bar{Y}_{N1}$$

Lo que demuestra la primera parte de la proposición. Para determinar el sesgo se sigue lo mismo que en la proposición 2.2.9

c.q.d.

2.7.2. Varianza del estimador y su estimación

Veamos una expresión de la varianza del estimador y de una estimación de ésta

Proposición 2.7.26 La varianza de $\bar{Y}_n$ y su estimación toma la expresión siguiente

$$VAR \; (\bar{Y}_n) = 1/n. \; [\; (1 - n \; / \; N_1 \;) + 2 \; . \; (\; 1 - n_1 \; / \; n \;)] \; . \; S_{N1}^2$$

estimada por :

$$\overset{*}{VAR} \; (\bar{Y}_n) = 1/n. \; [\; (1 - n \; / \; N_1 \;) + 2 \; . \; (\; 1 - n_1 \; / \; n \;)] \; . \; S_{n1}^2$$

Demostración :

$$VAR \; (\bar{Y}_n) \quad = \quad E[\bar{Y}_n \; - \; E(\bar{Y}_n)]^2 \quad = \quad E \; [\bar{Y}_n - \bar{Y}_{N1}]^2$$

$$= \; 1/n^2 \; E \; [\sum_{i=1}^{n_1} (X_i - \bar{Y}_{N1}) + \sum_{j=2n_1-n+1}^{n_1} (X_j - \bar{Y}_{N1}) \;]^2 \; =$$

$$= 1/n^2 . E \; [(\Sigma_i[X_i-\bar{Y}_{N1}])^2 + 2 \; (\Sigma_i[X_i-\bar{Y}_{N1}]).(\Sigma_j[X_j-\bar{Y}_{N1}]) + (\Sigma_j[X_j-\bar{Y}_{N1}])^2]$$

Llamando a $Z_i = X_i - \bar{Y}_{N1}$, podemos expresarla en la forma :

$$VAR\ (\bar{Y}_n) = 1/n^2 \cdot E[\ (\Sigma_i Z_i)^2 + 2 \Sigma_i \Sigma_j Z_i \cdot Z_j + (\Sigma_j Z_j)^2\ ] \quad (1)$$

Hagamos un inciso y veamos qué forma tiene la expresión $E\ [\ (\Sigma_i^k Z_i)^2\ ]$,

que en (1) aparece en dos términos (suponemos que las k unidades, genéricas del nivel muestral, pasan al nivel poblacional a N_1) :

$$E\ [\ (\Sigma_i^k Z_i)^2\ ] = [\ E \Sigma_i^k Z_i^2 + 2\ C_{k,2} \cdot E(Z_i \cdot Z_j)\ ] \quad (2)$$

donde $C_{k,2}$, representa las combinaciones de k unidades tomadas de dos en dos. Como :

$$E\ (\ Z_i^2\ ) = 1/N_1 \cdot \Sigma_{i=1}^{N1} Z_i^2 \quad (3)$$

$E\ (Z_i.Z_j) = E_1\ (Z_i \cdot E_2(\ Z_j\ )\ )$, siendo $E_2(\ Z_j\ ) = E\ (Z_j\ /\ Z_i)$ y

$E_2(\ Z_j\ ) = 1/(N_1-1) \cdot \Sigma_{i<>j} Z_j = -1/(N_1-1).\ Zi$, dado que $Z_i = X_i - \bar{Y}_{N1}$

y $\Sigma_{i=1}^{N1} Z_i = \Sigma_{i=1}^{N1} X_i - N_1.Y_{N1} = 0$ y en consecuencia : $\Sigma_{i=1}^{N1} Z_i = 0 = \Sigma_{i<>j} Z_i + Z_i$

deducimos que :

$$E\ (\ Z_i.Z_j) = E_1\ (\ -1/(N_1-1) \cdot Z_i^2\ ) = -1/(N_1-1) \cdot 1/N_1 \cdot \Sigma_{i=1}^{N1} Z_i^2 \quad (4)$$

Así pues la expresión (2) quedará en la forma :

$$E\ [\ (\Sigma Z_i)^2\ ] = [\ k/N_1 \cdot \Sigma_{i=1}^{N1} Z_i^2 + k.(k-1) \cdot -1/(N_1-1) \cdot 1/N_1 \cdot \Sigma_{i=1}^{N1} Z_i^2]$$

$$= k/N_1 \cdot [\ 1- (k-1)/(N_1-1)\ ]. \Sigma_{i=1}^{N1} Z_i^2$$

$$= k/N_1 \cdot (N_1 - k)/(N_1-1). \Sigma_{i=1}^{N1} Z_i^2$$

$$= k \cdot (1 - k/N_1) \cdot \sum_{i=1}^{N1} Z_i^2/(N_1-1)$$

$$= k \cdot (1 - k/N_1) \cdot S_{N1}^2 \quad (5)$$

Volviendo a la expresión (1) y teniendo en cuenta la expresión (5), se tiene que :

El primer término vale ($k = n_1$) : $n_1 \cdot (1 - n_1/N_1).S_{N1}^2$

El último término vale ($k = n-n_1$) : $(n-n_1) \cdot (1 - (n-n_1)/N_1).S_{N1}^2$ (6)

Veamos el valor del segundo término de (1) (la variación de i y j es la indicada en la definición 2.7.24) :

$$E\left[\sum_i \sum_j Z_i \cdot Z_j \right] = E\left[\left(\sum_{i=1}^{2n_1-n} Z_i + \sum_{k=2n_1-n+1}^{n_1} Z_k \right) \cdot \sum_{j=2n_1-n+1}^{n_1} Z_j \right]$$

$$= \sum_i \sum_j E(Z_i . Z_j) + \sum_{k<>j} \sum E(Z_k \cdot Z_j) + \sum_k E(Z_k^2) =$$

$$(2n_1-n).(n-n_1).E(Z_i . Z_j) + (n-n_1).(n-n_1-1).E(Z_k \cdot Z_j) + (n-n_1).E(Z_k^2) =$$

[teniendo en cuenta las expresiones (3) y (4) y que

$1/(N_1-1) \Sigma Z_i^2 = S_{N1}^2$, se sigue]

$$= (2n_1-n).(n-n_1).(-1/N_1).S_{N1}^2 + (n-n_1).(n-n_1-1).(-1/N_1).S_{N1}^2$$

$$+ (n-n_1).(N_1-1)./N_1.S_{N1}^2$$

$$= (n-n_1).(1-n_1/N_1).S_{N1}^2 \quad (7)$$

Sustituyendo, finalmente (6) y (7) en la expresión (1), se tiene :

$$VAR(\bar{Y}_n) = 1/n^2 [n_1 \cdot (1 - n_1/N_1).S_{N1}^2 + 2. (n-n_1).(1-n_1/N_1).S_{N1}^2$$

$$+ (n-n_1) \cdot (1 - (n-n_1)/N_1).S_{N1}^2]$$

$$= .... = 1/n. [(1 - n / N_1) + 2 \cdot (1 - n_1 / n)] \cdot S_{N1}^2$$

c.q.d.

Para obtener la estimación de $VAR(\bar{Y}_n)$, sólo hay que tener en cuenta que $E(S_{n1}^2) = S_{N1}^2$

3. Comparación entre estos estimadores

Las características de cada estimador estudiado anteriormente, en cuanto a estimadores de la media poblacional y de la media de los que responden, y sesgos respectivos, quedan reflejadas en la siguiente tabla.

Tabla II.1

Estimador $\bar{Y}_n$	Varianza del estimador	I/S	Sesgo (1)	I/S	Sesgo (2)
A: $1/n \sum_{i=1}^{n} Y_i$	$1/n.(1-n/N).S_N^2$	I	0	I	$-1.t_{N2}.\bar{Y}_{N1}$
B: $1/n \sum_{i=1}^{n} X_i$	$1/n.(1-n/N_1).S_{N1}^2$	S	$t_{N2}(\bar{Y}_{N1}-\bar{Y}_{N2})$	I	0
C: $1/n_1 \sum_{i=1}^{n1} X_i$	$1/n_1 .(1-n_1/N_1).S_{N1}^2$	S	$t_{N2}(\bar{Y}_{N1}-\bar{Y}_{N2})$	I	0
D: $1/n \sum_{i=1}^{n1} X_i$	$n_1/n^2.(1-n_1/N_1).S_{N1}^2$	S	$t_{n1}.\bar{Y}_{N1}-\bar{Y}_N$	S	$-1.t_{n2}.\bar{Y}_{N1}$
E: Como el estimador C					
F: $\frac{n_1.Y_{n1}+n_2.Y_u}{n}$ ($u=n_2/k$)	$(k-1)/n.t_{N2}.S_{N2}^2$ + $1/n.(1-n/N).S_N^2$	I	0	-	-
G: $1/n(\sum_{i=1}^{n_1} X_i+\sum_{i=1}^{n_2} X_i^*)$	$[(1/n(1-n/N_1) + 2/n(1-n_1/n)].S_{N1}^2$	S	$t_{N2}(\bar{Y}_{N1}-\bar{Y}_{N2})$	I	0

(1) : Respecto de la media poblacional (2): Respecto de la media de los que responden.

Al estimar un parámetro poblacional a través del parámetro muestral, los errores de muestreo los medimos a través de la varianza del estimador, o de su error estándar. La varianza nos dará una idea de la precisión (Cochran, 1977) del estimador, en el sentido de lo cerca que se encuentra el estimador muestral del valor esperado de éste.

Diremos que un estimador es tanto más preciso cuanto menor sea su error estándar (Azorín y Sánchez-Crespo, 1986) o más eficiente cuanto menor sea su varianza (Cramer (1960). Para medir lo cerca que se encuentra el estimador muestral del valor poblacional real; es decir, para medir la exactitud (Cochran, 1977) o acuracidad (Azorín y Sánchez-Crespo, 1986) lo haremos a través del error cuadrático medio.

Cuando se desea comparar dos estimadores insesgados, puede medirse la eficiencia entre ellos. Cuando se desea comparar dos estimadores, no siendo ambos insesgados, es mejor medir la acuracidad (Cochran, 1977). Oh y Sheuren (1983) miden la pérdida relativa de eficiencia entre dos estimadores, concepto que nosotros utilizaremos para medir la que existe entre dos de los estimadores estudiados.

En lo que sigue y con respecto a los estimadores antes indicados, vamos a estudiar:

a) La relación de orden entre las varianzas de los estimadores sesgados y la eficiencia entre los insesgados.

b) La acuracidad entre el estimador sesgado con menor varianza (estimador D) y el más eficiente de los estimadores insesgados. Demostraremos que bajo ciertas condiciones el primero es más acurado que el segundo. Así mismo, la acuracidad entre el grupo de estimadores sesgados.

c) Los intervalos de confianza para los estimadores sesgados, considerando un error máximo admisible que depende del error total. Veremos bajo qué condiciones el intervalo asociado al estimador D, es el que mejor se aproxima al intervalo asociado al estimador insesgado.

3.1. Varianza

Proposición 3.1.1 Para una muestra de extensión n y siendo t_{n1} la tasa de respuesta de la muestra, se tiene :

$$\text{Si } t_{n1} > 0.5 : VAR(D:\bar{Y}_n) < VAR(B:\bar{Y}_n) < VAR(C:\bar{Y}_n) < VAR(G:\bar{Y}_n)$$

$$\text{Si } t_{n1} \leq 0.5 : VAR(D:\bar{Y}_n) < VAR(B:\bar{Y}_n) < VAR(G:\bar{Y}_n) \leq VAR(C:\bar{Y}_n)$$

Demostración :

1) Entre el estimador $B:\bar{Y}n$ y el estimador $C:\bar{Y}n$

Como $n_1 < n \Rightarrow 1 - n/N_1 < 1 - n_1/N_1$, multiplicando por $1/n \Rightarrow$

$$1/n\,(1 - n/N_1) < 1/n\,(1 - n_1/N_1) < 1/n_1\,(1 - n_1/N_1) \qquad (1)$$

multiplicando los dos miembros de la desigualdad (1) por S_{N1}^2, se tiene :

$$VAR(B:\bar{Y}_n) < VAR(C:\bar{Y}_n)$$

2) Entre el estimador D:$\bar{Y}_n$ y el estimador C:$\bar{Y}_n$

Como $n_1 < n \Rightarrow n_1/n^2 < n/n^2$, multiplicando por $1-n_1/N_1 > 0 \Rightarrow$

$n_1/n^2\ (1-n_1/N_1) < n/n^2(1-n_1/N_1) = 1/n\ (1 - n_1/N_1) < 1/n_1\ (1 - n_1/N_1)$ (2)

multiplicando los dos miembros de la desigualdad (2) por S_{N1}^2, se tiene :

$$VAR(D:\bar{Y}_n) < VAR(C:\bar{Y}_n)$$

3) Entre el estimador C:$\bar{Y}_n$ y el estimador G:$\bar{Y}_n$

La desigualdad $1/n\ (1-n/N_1) < 1/n_1\ (1-n_1/N_1)$ es cierta, según (1).

Podemos plantearnos bajo qué condiciones se verifica

$$1/n\ (1 - n/N_1) + 2/n\ (1 - n_1/n) \leq 1/n_1\ (1 - n_1/N_1) \quad ?$$

Desarrollando y simplificando , se tiene :

$$3/n - 2\ n_1/n^2 \leq 1/n_1$$

Teniendo en cuenta que $t_{n1} = n_1/n$, la desigualdad anterior podemos expresarla en la forma :

$$2\ t_{n1}^2 - 3\ t_{n1} + 1 \geq 0$$

La solución de ésta inecuación es $t_{n1} \geq 1$ ó $t_{n1} \leq 1/2$

Luego :

$$VAR(G:\bar{Y}_n) \leq VAR(C:\bar{Y}_n) \quad si \quad t_{n1} \leq 1/2$$

(el otro intervalo, no es posible, dado que $t_{n1} \leq 1$)

$$VAR(C:\bar{Y}_n) < VAR(G:\bar{Y}_n) \quad si \quad t_{n1} > 1/2$$

4) Entre el estimador D:$\bar{Y}_n$ y el estimador B:$\bar{Y}_n$

Hemos demostrado en (2) que : $n_1/n^2\ (1-n_1/N_1) < 1/n\ (1-n_1/N_1)$

y en (1) que : $1/n\ (1 - n/N_1) < 1/n_1\ (1 - n_1/N_1)$

Podemos plantearnos bajo qué condiciones se verifica

$$n_1/n^2\ (1-n_1/N_1) < 1/n\ (1 - n/N_1) \quad ?$$

Desarrollando y simplificando , se tiene : $N_1\ n_1 - n_1^2 < n\ N_1 - n^2$

Teniendo en cuenta que $t_{n1} = n_1/n$, la desigualdad anterior podemos expresarla en la forma :

$$t_{n1}^2 - N_1/n\ t_{n1} + N_1/n - 1 > 0$$

La solución de ésta inecuación es $t_{n1} > N_1/n-1$ ó $t_{n1} < 1$

De las dos soluciones, sólo la segunda es posible, pues $N_1/n-1 > 1$

Luego : $VAR(D:\bar{Y}_n) < VAR(B:\bar{Y}_n)$

Finalmente, de los apartados 1), 3) y 4) se deducen las desigualdades propuestas.

c.q.d.

Proposición 3.1.2 Para tasas de respuesta superiores al 50%, la pérdida relativa entre la varianza del estimador sesgado $G:\bar{Y}_n$ y el estimador sesgado $D:\bar{Y}_n$ depende del cociente $F_1=N_1/n$ y de la tasa de respuesta t_{n1} , siendo ésta de valor :

$$t_{n2}/t_{n1}\ .\ \left[\ 3 + \frac{t_{n1} - t_{n2}}{F_1 - t_{n1}}\ \right]$$

Y siendo las variaciones del resto de los estimadores sesgados, con respecto al estimador sesgado más eficiente, inferiores a ese valor.

Demostración :

La variación relativa entre las varianzas de ambos estimadores es :

$$\frac{VAR(G:\bar{Y}_n) - VAR(D:\bar{Y}_n)}{VAR(D:\bar{Y}_n)} = \frac{1/n(1-n/N_1) + 2/n(1-n_1/n) - n_1/n^2\ (1- n_1/N_1)}{n_1/n^2\ (1- n_1/N_1)}$$

$$- \ .... \ - \ \frac{t_{n1}^2 - 3\ N_1/n\ t_{n1} + 3\ N_1/n - 1}{- t_{n1}^2 + N_1/n\ t_{n1}}$$

$$= \ .... \ = \ t_{n2}/t_{n1}\ .\ \left[\ 3 + \frac{t_{n1} - t_{n2}}{F_1 - t_{n1}}\ \right] \qquad (1)$$

que es la expresión propuesta.

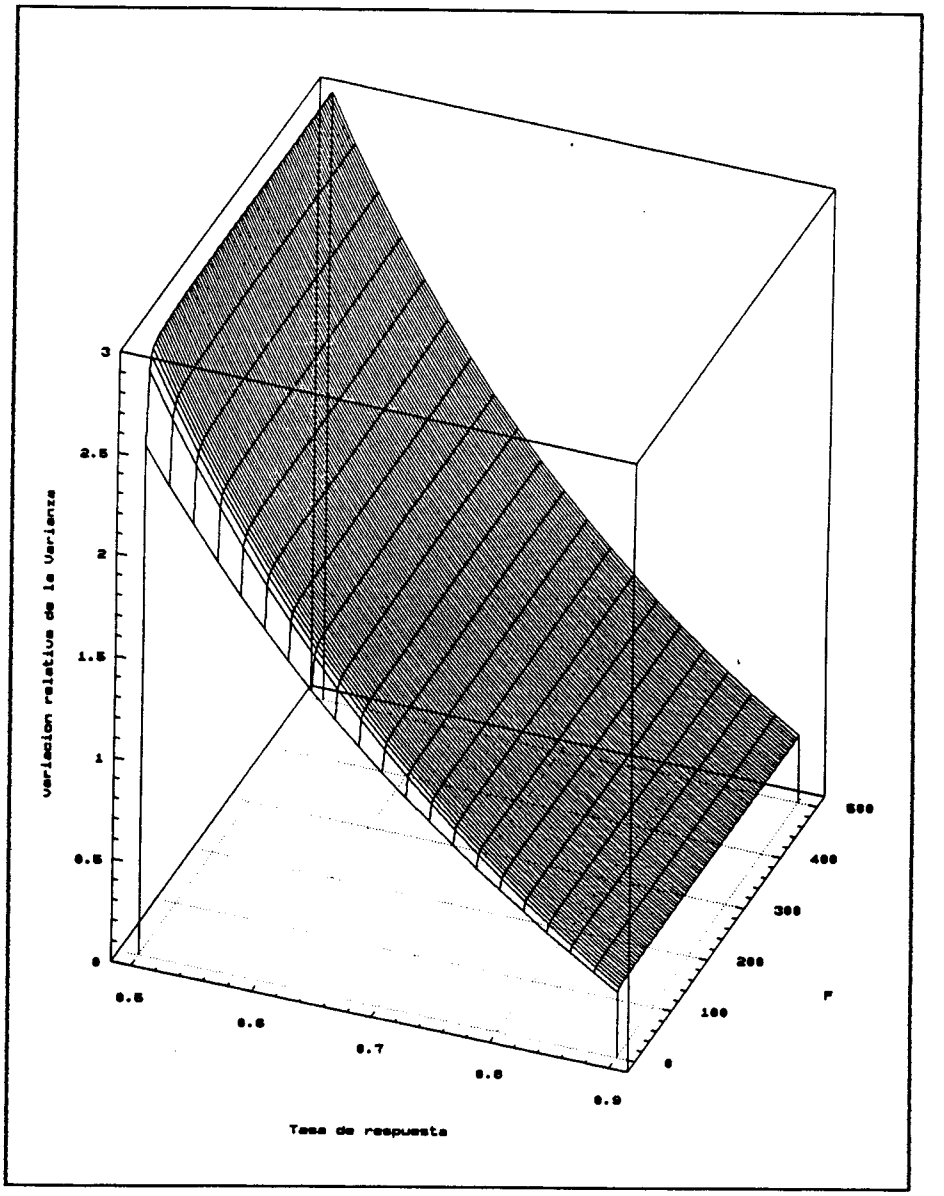
Para cualquier otro estimador sesgado S (B,C ó E), como de la proposición anterior, se tiene :

$$VAR(S:\bar{Y}_n) < VAR(G:\bar{Y}_n) \Longrightarrow VAR(S:\bar{Y}_n) - VAR(D:\bar{Y}_n) < VAR(G:\bar{Y}_n) - VAR(D:\bar{Y}_n)$$

Luego :
$$\frac{VAR(S:\bar{Y}_n) - VAR(D:\bar{Y}_n)}{VAR(D:\bar{Y}_n)} < \frac{VAR(G:\bar{Y}_n) - VAR(D:\bar{Y}_n)}{VAR(D:\bar{Y}_n)} = (1)$$

c.q.d.

Gráfico II.1



Como puede apreciarse en el Gráfico II.1, la pérdida relativa entre las varianzas de los estimadores G y D disminuye conforme aumentamos la tasa de respuesta y, dentro de una concreta, aumenta ligeramente al disminuir F_1.

En la Tabla II.2 reflejamos la pérdida relativa entre la varianza de los estimadores G y D, utilizados para la obtención del gráfico anterior.

Tabla II.2

F_1	$t_{n1}=0.60$	$t_{n1}=0.70$	$t_{n1}=0.80$	$t_{n1}=0.90$
500	2.0002	1.2860	0.7503	0.3335
100	2.0013	1.2874	0.7515	0.3342
50	2.0026	1.2892	0.7515	0.3342
5	2.0303	1.3256	0.7857	0.3550
2.5	2.0701	1.3809	0.8382	0.3888

Proposición 3.1.3 El estimador insesgado $A:\bar{Y}_n$ es más eficiente que el estimador $F:\bar{Y}_n$

Demostración :

Es trivial, dado que la varianza del primero es uno de los dos términos de la varianza del segundo (Tabla II.1).

c.q.d.

3.2. Acuracidad y sesgo

En lo que sigue vamos a demostrar que, en ocasiones, el estimador $D:\bar{Y}_n$, con menor varianza de entre los estimadores sesgados (proposición 3.1.1) es más exacto que el estimador $A:\bar{Y}_n$, el más eficiente de todos los estimadores insesgados (Proposición 3.1.3) y por consiguiente preferible a éste (Cansado, 1983). Así mismo demostramos que es el más acurado, bajo ciertas condiciones, del grupo de los estimadores sesgados estudiados.

Proposición 3.2.4 El estimador sesgado $D:\bar{Y}_n$ es más acurado que el estimador insesgado $A:\bar{Y}_n$, siempre que la tasa de respuesta muestral sea superior al 50% e igual a la tasa de respuesta poblacional y la razón, supuesto que se conozca, de $\bar{Y}_{N2} / \bar{Y}_{N1}$, sea inferior a $1/(n)^{1/2}$.

Demostración :

La varianza del estimador $A:\bar{Y}_n$, utilizando la proposición 2.1.6 , tiene por expresión :

$$VAR(A:\bar{Y}_n) = 1/n\ (1-n/N)\ N_1/N\ S_{N1}^2 + 1/n\ (1-n/N)\ N_1/N\ (1 - N_1/N)\ \bar{Y}_{N1}^2$$

$$= 1/n\ (1-n/N)\ N_1/N\ S_{N1}^2 + 1/n\ (1-n/N)\ t_{N1}\ t_{N2}\ \bar{Y}_{N1}^2$$

$$= 1/n\ (1-n/N)\ N_1/N\ S_{N1}^2 + 1/n\ \ t_{N1}\ t_{N2}\ \bar{Y}_{N1}^2 \qquad (1)$$

Para el estimador $D:\bar{Y}_n$, su varianza es :

$$VAR(D:\bar{Y}_n) = n_1/n^2\ (1-n_1/N_1\ )\ \ S_{N1}^2 \qquad (2)$$

y su error cuadrático medio :

$$ECM(D:\bar{Y}_n) = VAR(D:\bar{Y}_n) + (t_{n1}\ \bar{Y}_{N1} - \bar{Y}_N)^2 \qquad (3)$$

Suponiendo que $t_{n1} = t_{Nn1}$ y teniendo en cuenta la proposición 3.1.2 del capítulo I, se tiene :

$$t_{n1}.\bar{Y}_{N1} - \bar{Y}_N = t_{N1}.\bar{Y}_{N1} - \bar{Y}_N = t_{N1}.\bar{Y}_{N1} - (t_{N1}.\bar{Y}_{N1} + t_{N2}.\bar{Y}_{N2}) =$$

$$-1.t_{N2}.\ \bar{Y}_{N2}$$

Y la expresión (3), quedará en la forma :

$$ECM(D:\bar{Y}_n) = \ VAR(D:\bar{Y}_n) + t_{N2}^2\ \bar{Y}_{N2}^2 \qquad (4)$$

Admitiendo que $t_{N1} > 0.50$, o lo que es igual, que $t_{N1} > t_{N2}$ se sigue de (2) y (4) :

$$ECM(D:\bar{Y}_n) \leq \ n_1/n^2\ (1-n_1/N_1) + t_{N1}\ t_{N2}\ \bar{Y}_{N2}^2 \qquad (5)$$

Si comparamos el segundo término de (1) y (5) vemos que :

$t_{N1}\ t_{N2}\ \bar{Y}_{N2}^2 < 1/n\ t_{N1}\ t_{N2}\ \bar{Y}_{N1}^2$, siempre que $\bar{Y}_{N2}\ /\ \bar{Y}_{N1} < 1/(n)^{1/2}$

Y podemos plantearnos la siguiente cuestión :

¿ Bajo qué condiciones, se verifica

$$n_1/n^2\ (1-n_1/N_1) < 1/n\ (1-n/N)\ N_1/N \quad ? \qquad (6)$$

Desarrollando y simplificando, teniendo en cuenta que $t_{n1} = n_1/n$, la desigualdad anterior podemos expresarla en la forma :

$$t_{n1}^2 - N_1/n\ t_{n1} + N_1^2/\ N^2\ (N/n - 1 > 0$$

cuya solución es $t_{n1} < N_1/N < 1$ ó $t_{n1} > N_1/n - N_1/N > 1$

La primera de las soluciones nos indica que la inecuación (6) es siempre posible.

c.q.d

Proposición 3.2.5 El estimador $D:\bar{Y}_n$ es, de los estimadores sesgados, el que presenta menor sesgo

Demostración :

Si consideramos todos los estimadores sesgados que hemos estudiado en el apartado anterior, de las proposiciones 2.2.9, 2.3.12 y 2.7.24, se tiene :

$$\text{sesgo}(B{:}\bar{Y}_n) = \text{sesgo}(C{:}\bar{Y}_n) = \text{sesgo}(G{:}\bar{Y}_n) = \bar{Y}_{N1} - \bar{Y}_N \quad (1)$$

y de la proposición 2.4.15 :

$$\text{sesgo}(D{:}\bar{Y}_n) = t_{n1}.\bar{Y}_{N1} - \bar{Y}_N \quad (2)$$

Dado que $t_{n1} < 1 \Longrightarrow t_{n1}.\bar{Y}_{N1} < \bar{Y}_{N1}$ al ser $\bar{Y}_{N1} > 0$, y deducimos que

$$t_{n1}.\bar{Y}_{N1} - \bar{Y}_N < \bar{Y}_{N1} - \bar{Y}_N \Longrightarrow (2) < (1)$$

c.q.d.

Proposición 3.2.6 El estimador $D{:}\bar{Y}_n$ es, de los estimadores sesgados, el más acurado.

Demostración :

Dado que la acuracidad la medimos por el error cuadrático medio y éste viene expresado por la suma de la varianza y el cuadrado del sesgo, se deduce de las proposiciones 3.1.1 y 3.2.5 que es cierta la proposición.

3.3. Intervalos de confianza

Si fijamos un nivel de significación α para el estimador insesgado $\bar{Y}_n$, podemos plantearnos cual sería el comportamiento de los intervalos de confianza asociados a los diferentes estimadores sesgados que podemos obtener de una misma muestra de extensión n, tomando en consideración que el error máximo admisible, para la obtención de estos, dependa del error total.

El estudio de la varianza que nos clasifica cada uno de estos estimadores lo hemos realizado en una proposición anterior, así como el sesgo en su conjunto demostrando que el estimador D para la media poblacional es el que presenta menor sesgo de los sesgados. Ahora bién teniendo en cuenta que, en lo que sigue, vamos a trabajar con la expresión genérica

$$K = |\text{sesgo}| \ / \ S(\bar{Y}_n)$$

Al hacer la consideración de que el sesgo sea negativo y tomar el valor absoluto, el resultado es que éste puede ser mayor que el valor absoluto del sesgo asociado a los otros estimadores, por lo que tendremos en cuenta ésta cuestión.

Definición 3.3.7 Entenderemos por K_D, K_B, K_C y K_G la razón entre el valor absoluto del sesgo asociado al estimador correspondiente de la media poblacional y su error estándar.

Proposición 3.3.8 Para una misma muestra, las razones K_D, K_B, K_C y K_G asociados a los estimadores sesgados, verifican :

Para tasas de respuesta > 50% :

$$K_G < K_C < K_B < K_D \quad \text{ó} \quad K_G < K_C < K_B$$

según que $|\text{sesgo}(D:\bar{Y}_n)| < |\text{sesgo}(S:\bar{Y}_n)|$ ó no

Para tasas de respuesta ≤ 50% :

$$K_C \leq K_G < K_B < K_D \quad \text{ó} \quad K_C \leq K_G < K_B$$

según que $|\text{sesgo}(D:\bar{Y}_n)| < |\text{sesgo}(S:\bar{Y}_n)|$ ó no

y en donde S simboliza a cualquier otro de los estimadores sesgados.

Demostración :

De la proposición 3.2.5 vimos que el sesgo $(D:\bar{Y}_n)$ < sesgo $(B:\bar{Y}_n)$ = sesgo $(C:\bar{Y}_n)$ = sesgo $(G:\bar{Y}_n)$

De la proposición 3.1.1 vimos la relación entre las varianzas de estos estimadores

De una y otra se sigue las desigualdades propuestas junto con la definición 3.3.7

c.q.d.

Teniendo en cuenta lo indicado anteriormente respecto de la posibilidad de que el sesgo del estimador D para la media poblacional sea o no menor, en valor absoluto, que el de los otros, vamos a considerar en lo que sigue que es cierta la posibilidad antes indicada; en consecuencia, admitimos que :

$$|\text{sesgo}\ (D:\bar{Y}_n)| < |\text{sesgo}\ (S:\bar{Y}_n)| \quad S = B, C, D$$

Así mismo consideramos que la tasa de respuesta es superior o igual del 50%.

A. <u>Extremos de las regiones de error asociados a los estimadores</u>

Del punto A, apartado 3.5.2, capítulo I, deducimos que para un valor de $t = Z_{\alpha/2} > 1$:

* Los extremos inferiores $Y_1(t)$ son tales que :

$$Y_1(t) = -t.(1+K^2)^{1/2} + K \quad < 0$$

El comportamiento de $Y_1(t)$ para cualquier $K > 0$ es arbitrario, pese a la relación establecida para los valores de K de cada estimador sesgado en la proposición 3.3.7, excepto en los siguientes supuestos :

1) $K_D < K_1$, siendo $K_1 = 1/(t^2-1)^{1/2}$

Dado que al ser creciente $Y_1(t)$ en el intervalo $[0,K_1]$ y teniendo en cuenta que $K_G < K_C < K_B < K_D$, se verifica :

$$Y_1^G < Y_1^C < Y_1^B < Y_1^D \quad (1)$$

2) $K_G > K_1$

Dado que al ser decreciente $Y_1(t)$ para $K > K_1$ y, teniendo en cuenta que $K_G < K_C < K_B < K_D$, se verifica :

$$Y_1^G > Y_1^C > Y_1^B > Y_1^D \quad (2)$$

* Los extremos superiores $Y_2(t)$ son tales que :

$$Y_2(t) = t.(1+K^2)^{1/2} + K \quad > 0$$

El comportamiento del extremo superior para cualquier $K > 0$ es el de una función creciente; esto es :

$$Y_2^G < Y_2^C < Y_2^B < Y_2^D \quad (3)$$

dado que : $K_G < K_C < K_B < K_D$

Como conclusión, podemos indicar una condición para que los intervalos asociados a cada estimador estén encajados.

Proposición 3.3.9 Fijado un valor de $t > 1$ y siempre que $K_G > K_1$, la región de probabilidad asociada al estimador G, está contenida en la del estimador C; la de ésta en la de B y, finalmente, la de B en la de D.

Demostración :

Consecuencia de (2) y (3)

c.q.d.

B. <u>Variación relativa para el extremo inferior</u>

Pretendemos medir cuán de cerca se encuentra el extremo inferior de cada uno de las regiones de probabilidad, antes aludidas, con la región [-t,t] asociada al estimador insesgado.

Utilizaremos, para ello, las definiciones 3.5.31 y 3.5.32 del capítulo I, referentes a la variación relativa $V_1(t)$.

Proposición 3.3.10 Fijado un valor de $t > 1$ y siempre que $K_D < K_1$, se verifica :

$$V_1^G < V_1^C < V_1^B < V_1^D$$

Demostración :

Como $V_1(t) = \dfrac{d_1(t)}{t} = \dfrac{Y_1 + t}{t}$ y dado que de (1) $Y_1^G < Y_1^C < Y_1^B < Y_1^D$

pués $K_D < K_1$ y se sigue la relación propuesta.

c.q.d.

Proposición 3.3.11 Fijado un valor de $t>1$ y siempre que $K_G > K_1$ y $K_D < K_2$, se verifica :

$$V_1^G > V_1^C > V_1^B > V_1^D$$

siendo $K_2 = 2t/(t^2-1)$ y $K_1 = 1/(t^2-1)^{1/2}$

Demostración :

Como $V_1(t) = \dfrac{d_1(t)}{t} = \dfrac{Y_1 + t}{t}$ y dado que de (2) $Y_1^G > Y_1^C > Y_1^B > Y_1^D$

pues $K_G > K_1$

Como la variación relativa, a la izquierda de -t es negativa (definición 3.5.32, cap. I), imponemos la condición $K_D < K_2$, lo que nos asegura que todos los extremos tienen variación positiva

c.q.d.

C. <u>Probabilidad asociada</u>

De la proposición 3.5.27, capítulo I, se tiene :

$$P_r = \Phi (-Y_1(t)) + \Phi (-Y_2(t)) \quad (4)$$

Teniendo en cuenta lo indicado en el punto A de éste apartado, se sigue

Proposición 3.3.12 Fijado un valor de $t>1$ y siempre que $K_G > K_1$, se verifica :

$$P_r^G < P_r^C < P_r^B < P_r^D$$

siendo $K_1 = 1/(t^2-1)^{1/2}$

Demostración :

De la expresión (2) , se sigue :

$$-Y_1^G < -Y_1^C < -Y_1^B < -Y_1^D \quad ==>$$

$$\Phi(-Y_1^G) < \Phi(-Y_1^C) < \Phi(-Y_1^B) < \Phi(-Y_1^D) \qquad (5)$$

De la expresión (3), se sigue :

$$\Phi(Y_2^G) < \Phi(Y_2^C) < \Phi(Y_2^B) < \Phi(Y_2^D) \quad (6)$$

De (5) y (6) y teniendo en cuenta (4), obtenemos la relación propuesta.

c.q.d.

Podemos, después de todo lo visto, plantearnos la siguiente cuestión :

De los cuatro intervalos ¿ Cual es en el que la probabilidad asociada presenta una diferencia más pequeña, en valor absoluto, al compararla con la del estimador insesgado, cuya probabilidad es $P_\alpha = 1 - \alpha$?

Para los supuestos $K_G > K_1$ ó $K_D < K_1$ nada podemos plantear dado que en el intervalo $[0,K_2]$ sabemos que P_α es mayor o menor que P_r (Proposición 3.5.29, capítulo I).

Proposición 3.3.13 Fijado un valor de t y siempre que $K_G > K_2$, se verifica :

$$P_r^G - P_\alpha < P_r^C - P_\alpha < P_r^B - P_\alpha < P_r^D - P_\alpha$$

siendo $K_2 = 2t/(t^2-1)$

Demostración :

Si suponemos $K_G > K_2$, de acuerdo con la proposición 3.5.29 del capítulo I, sabemos que $P_\alpha \leq P_r$, para cualquier valor de K.

Teniendo en cuenta la proposición anterior y lo indicado anteriormente, se deduce la relación propuesta.

c.q.d.

4. Estimadores de la media y del total de una población finita, o de una parte de ella, con el marco muestral no depurado, a través de una muestra estratificada con no respuesta

En términos generales, en el muestreo estratificado debemos realizar los siguientes pasos :

a) La Población completa, constituida por todas las unidades que pueden ser muestreadas, se divide en subpoblaciones a las que llamaremos estratos.

b) Dentro de cada estrato se selecciona una muestra apartir del conjunto de unidades que constituyen éste

c) De la muestra obtenida en cada estrato, determinamos una media del conjunto de unidades muestrales, estimador de la media poblacional del estrato. Esta media, ponderarada adecuadamente, nos determina un estimador de la media de toda la Población.

d) El mismo procedimiento seguiremos para determinar la varianza de la media muestral de cada estrato y determinar la varianza de la media de toda la Población.

Supongamos que seleccionamos una muestra estratificada aleatoria de extensión n, de H estratos, con n_i elementos de los N_i del estrato i-esimo, para estudiar una característica Y de la población . Denotamos con Y_{ij} el valor de dicha característica en el estrato i del elemento j.

Si los estratos están limpios, en el sentido de que todos responden, existe una fórmula estándar de estratificación a aplicar a estos estratos.

El problema surge cuando al seleccionar una muestra de cada estrato, algunas de las unidades no responden para una característica en particular. No es correcto basar los resultados de la encuesta únicamente en las unidades que respondieron en cada estrato, dado que quienes no lo hicieron son diferentes de los otros.

Dado que el proceso de selección y de estimación se lleva a cabo separada e independientemente dentro de cada estrato, podemos determinar un estimador de la media poblacional y su varianza, a partir de un estimador de la media y su varianza de cada estrato.

Nosotros vamos a considerar el problema de la no respuesta, en el sentido de que el conjunto de los que no responden "corta" a todos los estratos, encontrandonos con N_{i1} y N_{i2} unidades de N_i y n_{i1} y n_{i2} unidades de n_i que responden / no responden, respectivamente. En ambos casos $N_i = N_{i1} + N_{i2}$ y $n_i = n_{i1} + n_{i2}$. Además denotaremos por N y M el total de unidades de la población y de los que responden.

En lo que sigue vamos a estudiar estimadores de la media y del total tanto de la población total como de la población de los que responden, utilizando para ello los estimadores insesgados que para m.a.s. sin reemplazamiento hemos visto en el segundo y tercer apartado de èste capítulo.

4.1. Estimador A :

Consideremos que N_{i1} y N_{i2} son ambas dsconocidas para cualquier estrato. Si fijamos la fracción de muestreo f_i , el tamaño n_i de la muestra permanece fijo, siendo el resultado una muestra con igual probabilidad para cada elemento. Procederemos a asignar a cada unidad Y_{ij} de la población el siguiente valor :

$$Y_{ij} = \begin{cases} X_{ij} & \text{Si la unidad j responde en el estrato i-ésimo} \\ 0 & \text{Si la unidad no responde} \end{cases}$$

Definición 4.1.1 Estimador directo o expandido para el estrato i, de

a) la media

$$\bar{Y}_{ni} = \frac{\sum_{i=1}^{ni} Y_{ij}}{n_i}$$

b) el total : $\overset{*}{Y}_{Ni} = N_i \cdot \bar{Y}_{ni}$

Para estos estimadores es válido lo indicado en el apartado 2 de éste capítulo, proposiciones 2.1.2 a 2.1.7, considerando que la media y el total poblacional es el de cada estrato. Por otro lado, si consideramos los valores definidos para Y_{ij}, se tendrá para cada estrato i :

$$\bar{Y}_{ni} = \frac{\sum_{j=1}^{ni1} X_{ij}}{n_i} = n_{i1}/n_i \cdot \bar{Y}_{ni1} = t_{ni1} \cdot \bar{Y}_{ni1}$$

Para el conjunto de los estratos, definimos :

$Y_M = \sum_{i=1}^{H} \sum_{j=1}^{n_{ij}} X_{ij} = \sum_{i=1}^{H} Y_{Ni1}$, es el total poblacional de los que responden

$\bar{Y}_M = Y_M / M$, la media poblacional de los que responden, siendo $M = \sum_{i=1}^{H} N_{i1}$

$$Y_N = \sum_{i=1}^{H} \sum_{j=1}^{n_i} Y_{ij} = \sum_{i=1}^{H} Y_{Ni} \text{ , es el total poblacional}$$

$$\bar{Y}_N = Y_N / N \text{ , la media poblacional}$$

Definición 4.1.2 Estimador directo o expandido del conjunto de la muestra, de :

a) el total

$$A{:}\overset{*}{Y}_N = \sum_{i=1}^{H} N_i . \bar{Y}_{ni}$$

b) la media

$$A{:}\bar{Y}_n = \frac{\overset{*}{Y}_N}{\sum_{i=1}^{H} N_i . t_{ni1}}$$

que en lo que sigue los expresaremos por $\overset{*}{Y}_N$ e $\bar{Y}_n$, respectivamente

Proposición 4.1.3 De la definición 4.1.1 y 4.1.2, podemos expresar $\bar{Y}_n$ en la forma

$$\bar{Y}_n = \frac{\overset{*}{Y}_N}{\sum_{i=1}^{H} N_i . t_{ni1}} = \frac{\sum_{i=1}^{H} N_i . \bar{Y}_{ni}}{\sum_{i=1}^{H} N_i . t_{ni1}} = \frac{\sum_{i=1}^{H} N_i . t_{ni1} . \bar{Y}_{ni1}}{\sum_{i=1}^{H} N_i . t_{ni1}} = \sum_{i=1}^{H} W_i . \bar{Y}_{ni1}$$

siendo $W_i = \dfrac{N_i . t_{ni1}}{\sum_{i=1}^{H} N_i . t_{ni1}}$ y tal que $\sum_{i=1}^{H} W_i = 1$

c.q.d.

4.1.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez / insesgadez de cada uno de ellos

Proposición 4.1.4 $\overset{*}{Y}_N$ es un estimador insesgado de Y_N y de Y_M, total de la población y de la población que responden, respectivamente

Demostración :

Utilizando la definición 4.1.1 y teniendo en cuenta la proposición 2.1.2, se sigue :

$$E(\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i. \; E(\bar{Y}_{ni}) = \sum_{i=1}^{H} N_i. \; \bar{Y}_{Ni} = \sum_{i=1}^{H} Y_{Ni} = Y_N$$

dado que $E(\bar{Y}_{ni}) = \bar{Y}_{Ni}$.

Por otro lado :

$$E(\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i. \; E(\bar{Y}_{ni}) = \sum_{i=1}^{H} N_i.N_{i1}/N_i. \; \bar{Y}_{Ni} = \sum_{i=1}^{H} N_{i1}. \; \bar{Y}_{Ni} = \sum_{i=1}^{H} Y_{Ni1} = Y_M$$

dado que $E(\bar{Y}_{ni}) = t_{Ni1}. \; \bar{Y}_{Ni}$

c.q.d

Proposición 4.1.5 $\bar{Y}_n$ es un estimador insesgado de $\bar{Y}_M$

Demostración :

A partir de la proposición 4.1.3, se tiene

$$E(\bar{Y}_n) = \sum_{i=1}^{H} E(W_i). \; E(\bar{Y}_{ni1}) = \sum_{i=1}^{H} E(W_i). \; \bar{Y}_{Ni1} \quad \text{(proposición 2.1.2)}$$

Por otro lado $E(W_i) = \dfrac{N_i.t_{Ni1}}{\sum_{i=1}^{H} N_i.t_{Ni1}}$, dado que $E(t_{ni1}) = t_{Ni1}$

Luego : $$E(\bar{Y}_n) = \frac{\sum_{i=1}^{H} N_{i1}. \; Y_{Ni1}}{\sum_{i=1}^{H} N_{i1}} = Y_M/M = \bar{Y}_M \quad \text{(definición 4.1.1)}$$

c.q.d.

4.1.2 Varianza de los estimadores y sus estimaciones

Damos una expresión de la varianza de cada estimador y de una estimación de ésta. Proponemos despues una descomposición de la varianza en suma de términos que dependen sólo de las unidades que responden en cada estrato.

Proposición 4.1.6 La varianza de $\overset{*}{Y}_N$, toma la expresión

$$VAR\ (\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i^2 \ .\ 1/n_i .\ (1-n_i/N_i)\ .\ S_{Ni}^2$$ siendo su estimación

$$\overset{*}{VAR}\ (\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i^2 \ .\ 1/n_i .\ (1-n_i/N_i)\ .\ S_{ni}^2$$

Demostración :

$$VAR\ (\overset{*}{Y}_N) = VAR\ (\sum_{i=1}^{H} N_i .\ \bar{Y}_{ni}) = \sum_{i=1}^{H} N_i^2\ .\ VAR\ (\bar{Y}_{ni}) =$$

$$\sum_{i=1}^{H} N_i^2 \ .\ 1/n_i .\ (1-n_i/N_i)\ .\ S_{Ni}^2 \quad (\text{ proposición 2.1.4 })$$

c.q.d.

Las expresiones anteriores podemos descomponerlas en suma de términos que dependen sólo de las unidades que responden

Definición 4.1.7 La varianza de las unidades que responden en la muestra y de las unidades que responden en el total de la población de cada estrato, vienen expresadas por :

$$VAR_{ni1} = 1/n_{i1}\ .\ \sum_{j=1}^{n_{i1}} (\ X_{ij} - \bar{Y}_{ni1}\ )^2 = \ldots = 1/n_{i1}\ (\ \sum_{j=1}^{n_{i1}} X_{ij}^2\ ) - \bar{Y}_{ni1}^2$$

$$VAR_{Ni1} = 1/N_{i1}\ .\ \sum_{j=1}^{N_{i1}} (\ X_{ij} - \bar{Y}_{Ni1}\ )^2 = \ldots =$$

$$= 1/N_{i1}\ (\ \sum_{j=1}^{N_{i1}} X_{ij}^2\ ) - \bar{Y}_{Ni1}^2$$

Consecuencia de aplicar la definición 2.1.5, a cada estrato

Proposición 4.1.8 Para el estimador $\overset{*}{Y}_N$ su varianza podemos expresarla en la forma :

$$VAR\ (\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i^2 .1/n_i .(1-n_i/N_i).N_{i1}/(N_i-1).(VAR_{Ni1} + t_{Ni2}.\bar{Y}_{Ni1}^2)$$

y su estimación por :

$$\overset{*}{VAR}\ (\overset{*}{Y}_N) = \sum_{i=1}^{H} N_i^2 .1/(n_i-1).(1-n_i/N_i).t_{ni1}.(VAR_{ni1} + t_{ni2}.\bar{Y}_{ni1}^2)$$

Demostración :

Es consecuencia de la Proposición 4.1.6 y de la descomposición de S_{Ni}^2 y S_{ni}^2, siguiendo el desarrollo de la proposición 2.1.7 con la definición 4.1.7

c.q.d.

Para estimar la varianza del estimador de la media, Kish (1965) da la siguiente expresión :

Proposición 4.1.9 Un estimador de la varianza de $\bar{Y}_n$ es

$$\overset{*}{VAR}(\bar{Y}_n) = \sum_{i=1}^{H} w_i^2 . 1/(n_i-1) . (1-n_i/N_i) . n_i/n_{i1} . [VAR_{ni1} + t_{ni2} . (\bar{Y}_n - \bar{Y}_{ni1})^2]$$

En ésta expresión aparece VAR_{ni1} que es la varianza por elemento de los que responden dentro del estrato n_i. Podemos obtener facilmente una expresión en la que dicho término se sustituye por la varianza alrededor de $\bar{Y}_n$, con lo que el cálculo de $\overset{*}{VAR}(\bar{Y}_n)$ se realizaría como si no hubiera habido estratificación. Veamos ésta expresión :

Definición 4.1.10 Para cada estrato, la varianza de las unidades que responden en la muestra y de las unidades que responden en el total de la población del estrato, tomadas ambas alrededor de la media de la muestra, vienen expresadas por :

$$\overset{*}{VAR}_{ni1} = 1/n_{i1} \cdot \sum_{j=1}^{n_{i1}} (X_{ij} - \bar{Y}_n)^2$$

$$\overset{*}{VAR}_{Ni1} = 1/N_{i1} \cdot \sum_{j=1}^{N_{i1}} (X_{ij} - \bar{Y}_n)^2$$

Proposición 4.1.11 La relación entre $\overset{*}{VAR}_{ni1}$ y VAR_{ni1} es

$$\overset{*}{VAR}_{ni1} = VAR_{ni1} + (\bar{Y}_n - \bar{Y}_{ni1})^2$$

Demostración :

$$VAR_{ni1} = 1/n_{i1} \cdot \sum_{j=1}^{n_{i1}} (X_{ij} - \bar{Y}_{ni1})^2 =$$

$$1/n_{i1} \cdot \sum_{j=1}^{n_{i1}} (X_{ij} - \bar{Y}_n + \bar{Y}_n - \bar{Y}_{ni1})^2 =$$

$$1/n_{i1} . [\sum_{j=1}^{n_{i1}} (X_{ij} - \bar{Y}_n)^2 + 2 . \sum_{j=1}^{n_{i1}} (X_{ij} - \bar{Y}_n)(\bar{Y}_n - \bar{Y}_{ni1}) + \sum_{j=1}^{n_{i1}} (\bar{Y}_n - \bar{Y}_{ni1})^2]$$

El segundo término al descomponerlo en suma de cuatro términos, tiene por valor :

$$-2.\ n_{i1}\ .(\bar{Y}_n - \bar{Y}_{ni1})^2$$

por lo que la expresión quedará reducida a :

$$VAR_{ni1} = 1/n_{i1}\ .[\ \sum_{j=1}^{n_{i1}} (\ X_{ij} - \bar{Y}_n\ )^2 - n_{i1}\ .(\bar{Y}_n - \bar{Y}_{ni1})^2] =$$

$$\overset{*}{VAR}_{ni1} - (\bar{Y}_n - \bar{Y}_{ni1})^2$$

c.q.d.

Proposición 4.1.12 Una estimación de la varianza de $\bar{Y}_n$ tiene por expresión :

$$\overset{*}{VAR}\ (\bar{Y}_n) = \sum_{i=1}^{H} W_i^2.1/(n_i-1).(1-n_i/N_i).n_i/n_{i1}.[\overset{*}{VAR}_{ni1} - t_{ni1}.(\bar{Y}_{ni1}-\bar{Y}_n)^2]$$

Demostración :

Consecuencia de la proposición 4.1.9 y 4.1.11

c.q.d.

4.2. Estimador F

A partir del estimador de Hansen y Hurwitz para una m.a.s. sin reemplazamiento, vamos a obtener un estimador de la media poblacional. Se propone una postestratificación, dentro de cada estrato, separando los que responden de los que no lo hacen. Con el fín de obtener información de éste segundo substrato, se toma una submuestra de un tamaño conveniente y se recoge información de ella.

En lo que sigue :

1) Definimos un estimador de la media poblacional y hacemos diversas consideraciones según que la estratificación sea desproporcionada o proporcionada, obteniendo en éste último caso un estimador insesgado de la varianza del estimador.

2) Determinamos la afijación óptima para el número de individuos que hace mínima la varianza del estimador o viceversa.

3) Determinamos la afijación óptima para el número de individuos y la tasa de submuestreo que, para un valor dado de la varianza hacen que la función de costo esperado sea mínima.

Proposición 4.2.13 Para cada estrato i del que obtenemos una muestra aleatoria simple de n_i unidades, de las que n_{i1} son unidades que responden y n_{i2} unidades que no responden, un estimador insesgado de la media poblacional para éste estrato, es :

$$\bar{Y}_{ni} = \frac{n_{i1}\ \bar{Y}_{n_{i1}} + n_{i2}\ \bar{Y}_{u_i}}{n_i}$$

siendo :

$\bar{Y}_{n_{i1}}$ la media muestral de los que responden

u_i una muestra de n_{i2} / k_i unidades obtenida de entre los que no responden

$\bar{Y}_{u_i}$ la media de la muestra correspondiente

Con una varianza :

$$VAR\ (\ \bar{Y}_{ni}\ ) = \frac{k_i - 1}{n_i}\ t_{Ni2}\ \frac{S^2_{N_{i2}}}{N_{i2}} + \frac{1}{n_i}\left(1 - \frac{n_i}{N_i}\right)\frac{S^2_{N_i}}{N_i}$$

siendo :

$S^2_{N_{i2}}$ y $S^2_{N_i}$ las varianzas ajustadas para el total de unidades N_{i2} y N_i

Demostración :

Consecuencia de aplicar a cada estrato los resultados del teorema 2.6.21

c.q.d.

4.2.1 Sesgo y varianza

El estimador propuesto estima a la media de la población. Estudiamos la sesgadez / insesgadez, así como una expresión para su varianza

Proposición 4.2.14 Para un muestreo con H estratos, un estimador insesgado de la media poblacional es :

$$F{:}\bar{Y}_n = \sum_{i=1}^{H} W_{Ni}\ \bar{Y}_{ni}$$

que en lo que sigue lo expresaremos por $\bar{Y}_n$

siendo su varianza :

$$VAR\ (\ \bar{Y}_n\ ) = \sum_{i=1}^{H} W_{Ni}^2\ \ VAR\ (\ \bar{Y}_{ni}\ )$$

donde W_{Ni} son las ponderaciones por estrato ($\sum_{i=1}^{H} W_{Ni} = 1$)

Demostración :

De la proposición 4.2.13, obtenemos :

$E (\bar{Y}_{ni}) = \bar{Y}_{Ni}$ $\quad i= 1,..., H$, siendo $\bar{Y}_{Ni}$ la media poblacional del estrato i

Luego :

$$E (\bar{Y}_{n}) = E (\sum_{i=1}^{H} W_{Ni} \bar{Y}_{ni}) = \sum_{i=1}^{H} W_{Ni} E (\bar{Y}_{ni}) = \sum_{i=1}^{H} W_{Ni} \bar{Y}_{Ni} = \bar{Y}_{N}$$

Por otro lado :

$$VAR (\bar{Y}_{n}) = VAR(\sum_{i=1}^{H} W_{Ni} \bar{Y}_{ni}) = \sum_{i=1}^{H} VAR(W_{Ni} \bar{Y}_{ni}) = \sum_{i=1}^{H} W_{Ni}^2 VAR(\bar{Y}_{ni})$$

dada la independencia entre los estratos

c.q.d.

Corolario 4.2.15 Una expresión para la varianza de éste estimador es

$$VAR (\bar{Y}_{n}) = \sum_{i=1}^{H} W_{Ni}^2 \frac{k_i - 1}{n_i} t_{Ni2} \frac{S^2}{N_{i2}} + \sum_{i=1}^{H} W_{Ni}^2 \frac{1}{n_i} (1 - \frac{n_i}{N_i}) \frac{S^2}{N_i}$$

Demostración :

Consecuencia de aplicar a la proposición 4.2.14 los resultados de la proposición 4.2.13

c.q.d.

Corolario 4.2.16 Una expresión para la varianza del estimador es :

$$VAR (\bar{Y}_{n}) = \sum_{i=1}^{H} W_{Ni}^2 \frac{k_i - 1}{n_i} t_{Ni2} \frac{S^2}{N_{i2}} + \sum_{i=1}^{H} W_{Ni}^2 \frac{1}{n_i} (1-f_i) \frac{S^2}{N_i}$$

Consecuencia de sustituir en la expresión anterior n_i/N_i por f_i, siendo f_i la fracción de muestreo del estrato i

c.q.d.

De las dos expresiones de la varianza obtenida en los corolarios anteriores, observamos un primer término que representa la contribución a la varianza del estimador, debida al hecho de haber establecido contacto con una fracción, para cada estrato, de los que no responden.

En los estratos en los que la muestra responda en la totalidad de sus unidades, la tasa de submuestra k_i es la unidad y la contribución es cero.

Corolario 4.2.17 Una expresión para la varianza del estimador es :

$$VAR\ (\ \bar{Y}_n\ ) = \sum_{i=1}^{H} \frac{1}{f_i}\ A_i - \sum_{i=1}^{H} B_i$$

donde :

$$A_i = \frac{W_{Ni}}{N}\ [\ t_{Ni2}\ (\ k_i - 1\ )\ S^2_{N_{i2}} + S^2_{N_i}\ ]$$

$$B_i = \frac{W_{Ni}}{N}\ S^2_{N_i}$$

Demostración :

Teniendo en cuenta que $W_{Ni} = N_i / N$ y que $n_i = N_i . f_i$, sustituyendo en la expresión de la varianza obtenida en el Corolario 4.2.16 y agrupando los términos en los que aparece la fracción de muestreo f_i, obtenemos la expresión indicada.

c.q.d.

4.2.2. Estimador de la varianza

Nuestro siguiente objetivo es determinar un estimador de $VAR\ (\ \bar{Y}_n\ )$ siendo necesario para ello un marco de estratificación de muestras proporcionales, con el fín de obtener una estimación de las ponderaciones W_{Ni} para cada estrato.

Estudiamos tambien el caso de afijación igual entre las muestras de cada estrato, dentro de la proporcionalidad de muestras.

Proposición 4.2.18 Para una estratificación con muestras autoponderadas, la varianza del estimador toma la siguiente expresión :

$$VAR\ (\ \bar{Y}_n\ ) = \frac{1}{n} \sum_{i=1}^{H} W_{Ni}\ (\ k_i - 1\ )\ t_{Ni2}\ S^2_{N_{i2}} + \frac{1 - f}{n} \sum_{i=1}^{H} W_{Ni}\ S^2_{N_i}$$

demostración :

Como el muestreo es proporcional, se tiene que :

$$f = \frac{n}{N} = f_i = \frac{n_i}{N_i} \qquad i = 1, \ldots, H$$

siendo f y f_i las fracciones de muestreo del total de la muestra y de la muestra en el estrato i, respectivamente

Como $W_{Ni} = N_i / N$, junto con la igualdad anterior, deducimos que

$$W_{Ni} = \frac{N_i}{N} = \frac{n_i}{n} \qquad i = 1,\ldots, H \qquad (1)$$

es decir : $E(n_i / n) = W_{Ni}$

Además :

$$1 - f_i = 1 - f \qquad i = 1,\ldots, H$$

Como consecuencia de todo esto, el segundo término de la expresion de la varianza del estimador, obtenida en el Corolario 4.2.16, podemos escribirla en la forma:

$$\sum_{i=1}^{H} W_{Ni}^2 \frac{1}{n_i} (1 - f_i) S_{N_i}^2 = \frac{1 - f}{n} \sum_{i=1}^{H} W_{Ni} S_{N_i}^2$$

dado que W_{Ni} / n_i es igual a 1 / n, por (1)

LLevando estos resultados a $VAR (\bar{Y}_n)$, obtenemos la expresión indicada.

c.q.d.

Corolario 4.2.19 Para una estratificación con muestras autoponderadas, una expresión para la varianza del estimador de la media es :

$$VAR (\bar{Y}_n) = \frac{1}{f} \sum_{i=1}^{H} A_i - \sum_{i=1}^{H} B_i$$

donde :

$$A_i = \frac{W_{Ni}}{N} [t_{Ni2} (k_i - 1) S_{N_{i2}}^2 + S_{N_i}^2]$$

$$B_i = \frac{W_{Ni}}{N} S_{N_i}^2$$

Demostración :

Teniendo en cuenta que $f_i = f$ y sustituyendo en la expresión de la varianza obtenida en el Corolario 4.2.17, obtenemos la expresión indicada.

c.q.d.

Proposición 4.2.20 Para una estratificación con muestras autoponderadas, un estimador insesgado de la varianza del estimador de la media toma la siguiente expresión :

$$VAR^{*}(\bar{Y}_n) = \frac{1}{n}\sum_{i=1}^{H} W_{ni}\,(k_i - 1)\,t_{ni2}\,S^2_{n_{i2}} + \frac{1-f}{n}\sum_{i=1}^{H} W_{ni}\,S^2_{n_i}$$

donde :

$W_{ni} = n_i / n$ tal que $E(W_{ni}) = W_{Ni}$, por ser muestras autoponderadas

$t_{ni2} = n_{i2} / n_i$, tal que $E(t_{ni2}) = t_{Ni2}$ (m.a.s.)

$S^2_{n_{i2}}$ y $S^2_{n_i}$ los estimadores insesgados de $S^2_{N_{i2}}$ y $S^2_{N_i}$, respectivamente

Demostración :

$$E[VAR^{*}(\bar{Y}_n)] = \frac{1}{n}E\left[\sum_{i=1}^{H} W_{ni}.(k_i-1).W_{ni2}.S^2\right] + \frac{1-f}{n}.E\left[\sum_{i=1}^{H} W_{ni}.S^2_{ni}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{H} E(W_{ni}).(k_i-1).E(W_{ni2}).E(S^2_{n_{i2}}) + \frac{1-f}{n}.\sum_{i=1}^{H} E(W_{ni}).E(S^2_{n_i})$$

$$= VAR(\bar{Y}_n)$$

c.q.d.

Proposición 4.2.21 Para una estratificación con muestras autoponderadas y en el que la afijación es igual en todos los estratos, la varianza del estimador de la media toma la siguiente expresión :

$$VAR(\bar{Y}_n) = \frac{1}{n.H}\sum_{i=1}^{H} (k_i - 1)\,t_{Ni2}\,S^2_{N_{i2}} + \frac{1-f}{n.H}\sum_{i=1}^{H} S^2_{N_i}$$

Demostración :

Dado que la afijación es igual en todos los estratos : $f_i = cte = c$, siendo $c = 1 / H$

Como consecuencia $n_i = n/H$, donde $n > 2H$, para que cada muestra contenga al menos dos unidades y poder calcular la varianza de cada estrato, se tiene que :

$$W_{Ni} = n_i / n = 1 / H$$

Sustituyendo este último resultado en la expresión de $VAR(\bar{Y}_n)$, obtenida en la Proposición 4.2.18, obtenemos la expresión anteriormente indicada.

c.q.d.

Proposición 4.2.22 Para una estratificación con muestras autoponderadas y en el que la afijación es igual en todos los estratos, un estimador insesgado de la varianza del estimador toma la siguiente expresión :

$$VAR^{*}(\bar{Y}_n) = \frac{1}{n.H} \sum_{i=1}^{H} (k_i - 1) t_{ni2} \frac{S^2}{n_{i2}} + \frac{1 - f}{n.H} \sum_{i=1}^{H} \frac{S^2}{n_i}$$

donde :

$t_{ni2} = n_{i2} / n_i$ tal que $E(t_{ni2}) = t_{Ni2}$

$S^2_{n_{i2}}$ y $S^2_{n_i}$ los estimadores insesgados de $S^2_{N_{i2}}$ y $S^2_{N_i}$, respectivamente

Demostración :

Como la Proposición 4.2.20

c.q.d.

4.2.3. Afijación óptima

Con el muestreo proporcional no podemos aumentar la precisión de la media poblacional. Para ésto necesitamos usar razones de muestreo diferentes para los diversos estratos (muestreo desproporcionado), de manera que obtengamos la mínima varianza para la media Poblacional o para un estimador de ésta, por unidad de costo.

¿ Cual es la afijación óptima que hace posible tal objetivo ?

Para dar respuesta a la cuestión planteada podemos tomar dos caminos, en ambos considerando que las tasas k_i de submuestreo estén fijadas :

* Concretar el total n de elementos de la muestra, dandole un valor n_o
* Concretar el valor de la varianza y determinar el menor valor de n para conseguir tal valor

Proposición 4.2.23 Siendo n_o el total de unidades a muestrear por estratificación, el valor de la razón de muestreo f_i para cada estrato, que hace mínima la varianza del estimador, es :

$$f_i = K \left[(k_i - 1) t_{Ni2} \frac{S^2}{N_{i2}} + \frac{S^2}{N_i} \right]^{1/2} \qquad i = 1,\ldots,H$$

donde K es una constante fija de valor :

$$K = \frac{n_o}{N . \sum_{i=1}^{H} (A_i N_i)^{1/2}}$$

siendo :

$$A_i = \frac{W_{Ni}}{N} \left[t_{Ni2} (k_i - 1) S^2_{N_{i2}} + S^2_{N_i} \right]$$

Demostración :

Tenemos que encontrar los valores f_i que hacen mínima VAR ($\bar{Y}_n$) , con la condición :

$$n_o = \sum_{i=1}^{H} n_i = \sum_{i=1}^{H} f_i N_i \qquad (1)$$

Para ello definimos la función lagrangiana :

$$F = VAR (\bar{Y}_n) - t \left[n_o - \sum_{i=1}^{H} f_i N_i \right]$$

Siendo VAR ($\bar{Y}_n$) la expresión de la varianza del estimador obtenida en el Corolario 4.2.17

La derivada de la función lagrangiana respecto de cada variable f_i , deberá ser igual a cero; esto es :

$$F'_{f_i} = \frac{-1. A_i}{f_i^2} + t N_i = 0 \qquad i= 1,....,H$$

De donde obtenemos que $f_i = \left[\frac{1}{t} . \frac{A_i}{N_i} \right]^{1/2}$ (2)

Sustituyendo en (1), obtenemos :

$$t^{1/2} = \frac{\sum_{i=1}^{H} (A_i N_i)^{1/2}}{n_o}$$

y finalmente de (2) la expresión indicada.

c.q.d.

Proposición 4.2.24 Siendo V_o el valor de la varianza del estimador, el menor valor de n , total de la muestra, se consigue si la razón de muestreo f_i para cada estrato es :

$$f_i = K_1 \left[(k_i - 1) t_{Ni2} S^2_{N_{i2}} + S^2_{N_i} \right]^{1/2} \qquad i = 1,...,H$$

donde K_1 es una constante fija de valor :

$$K_i = \frac{\sum_{i=1}^{H} (A_i N_i)^{1/2}}{V_o + \sum_{i=1}^{H} B_i}$$

siendo :

$$A_i = \frac{W_{Ni}}{N} \left[t_{Ni2} (k_i - 1) S^2_{N_{i2}} + S^2_{N_i} \right]$$

$$B_i = \frac{W_{Ni}}{N} S^2_{N_i}$$

Demostración :

Tenemos que encontrar los valores f_i que hacen :

$$n = \sum_{i=1}^{H} n_i = \sum_{i=1}^{H} f_i N_i$$

Con la condición :

$$V_o = VAR (\bar{Y}_n) \qquad (1)$$

Siendo VAR ($\bar{Y}_n$) la expresión de la varianza del estimador obtenida en el Corolario 4.2.17

Para ello definimos la función lagrangiana :

$$F = [n - \sum_{i=1}^{H} f_i N_i] + t [V_o - \sum_{i=1}^{H} \frac{1}{f_i} A_i - \Sigma B_i]$$

La derivada de la función lagrangiana respecto de cada variable f_i , deberá ser igual a cero, esto es :

$$F'_{f_i} = \frac{t A_i}{f_i^2} - N_i = 0 \qquad i= 1,....,H$$

De donde obtenemos que $\frac{1}{f_i} = [\frac{1}{t} . \frac{N_i}{A_i}]^{1/2}$ (2)

Sustituyendo en (1), obtenemos : $t^{1/2} = \frac{\sum_{i=1}^{H} (A_i N_i)^{1/2}}{V_o + \sum_{i=1}^{H} B_i}$

y finalmente de (2) la expresión indicada.

c.q.d.

Corolario 4.2.25 Para el supuesto en que todas las unidades de la muestra de cada estrato respondieran, el valor numérico de la razón de muestreo de cada estrato es proporcional a la desviación típica ajustada del estrato, tanto si fijamos un valor para el total de la muestra como para la varianza del estimador de la media Poblacional.

Demostración :

Si todas las unidades responden $k_i = 1$, para cada estrato i

Como consecuencia :

$$A_i = \frac{N_i}{N} \cdot \frac{S^2}{N_i}$$

Sustituyendo en cada expresión de f_i, obtenida en las propiedades 4.2.23 y 4.2.24, se tiene :

$$f_i = K \left[\frac{S^2}{N_i} \right]^{1/2} = K \frac{S}{N_i} \qquad \text{(fijando el valor de } n = n_o \text{)}$$

$$f_i = K_1 \left[\frac{S^2}{N_i} \right]^{1/2} = K_1 \frac{S}{N_i} \qquad \text{(fijando el valor de VAR } (\bar{Y}_n) = V_o \text{)}$$

c.q.d.

4.2.4. Afijación óptima con función de coste

En éste apartado pretendemos escoger el tamaño de la fracción de muestreo f_i y la tasa de submuestreo k_i, para cada estrato, de tal manera que el coste esperado sea mínimo para un valor determinado V_o de la varianza del estimador.

Para ello, consideramos que :

C_o : Coste de cada cuestionario

C_1 : Coste de procesar cada cuestionario que respondió

C_2 : Coste de procesar cada cuestionario en el submuestreo

n : Total de cuestionarios : $n = \sum_{i=1}^{H} n_i$

r : Total de cuestionarios que respondieron : $r = \sum_{i=1}^{H} n_{i1}$

u : Total de cuestionarios en el submuestreo : $u = \sum_{i=1}^{H} u_i$

El coste total sería $C' = C_o\, n + C_1\, r + C_2\, u$ y el coste esperado de la encuesta será : $C = E(C')$

Como :

$$E(r) = \sum_{i=1}^{H} E(n_{i1}) = \sum_{i=1}^{H} E(n_i\ t_{ni1}) = \sum_{i=1}^{H} E_1(t_{ni1}\ E_2(n_{i1})) = \sum_{i=1}^{H} t_{Ni1}\ n_i$$

$$E(u) = \sum_{i=1}^{H} E(u_i) = \sum_{i=1}^{H} E(n_{i2}/k_i) = \sum_{i=1}^{H} 1/k_i\ n_i\ t_{Ni2}$$

La expresión del coste esperado quedará en la forma :

$$C = C_o\ n + C_1 \sum_{i=1}^{H} n_i\ t_{Ni1} + C_2 \sum_{i=1}^{H} 1/k_i\ n_i\ t_{Ni2}$$

Proposición 4.2.26 Los valores de f_i y k_i que hacen mínimo el coste esperado, para un valor fijado V_o de la varianza del estimador, son :

$$k_i^2 = \frac{C_2\ [\ S^2_{N_i} - t_{Ni2}\ S^2_{N_{i2}}\ ]}{S^2_{N_{i2}}\ [\ C_o + C_1\ t_{Ni1}\ ]}$$

$$f_i = \frac{\sum_{i=1}^{H} a_i\ A_i}{V_o + \sum_{i=1}^{H} B_i} \cdot \frac{1}{a_i}$$

siendo :

$$A_i = \frac{W_{Ni}}{N}\ [\ t_{Ni2}\ (\ k_i - 1\ )\ S^2_{N_{i2}} + S^2_{N_i}\ ]$$

$$B_i = \frac{W_{Ni}}{N}\ S^2_{N_i}$$

$$a_i = \frac{N\ (\ C_2\ )^{1/2}}{k_i\ S_{N_{i2}}}$$

Demostración :

Tenemos que hacer mínima la función de coste; esto es :

$$C = C_o \sum f_i\ N_i + C_1 \sum_{i=1}^{H} f_i\ N_i\ t_{Ni1} + C_2 \sum_{i=1}^{H} 1/k_i\ f_i\ N_i\ t_{Ni2}$$

dado que : $n_i = f_i\ .\ N_i \quad (\ i = 1,\ldots,H)$

Con la condición de ser : $V_o = VAR\ (\ \bar{Y}_n\ )$, definida por la expresión obtenida en el corolario 4.2.17

La función lagrangiana es :

$$F = [C_o \sum f_i N_i + C_1 \sum_{i=1}^{H} f_i N_i t_{Ni1} + C_2 \sum_{i=1}^{H} 1/k_i f_i N_i t_{Ni2}] +$$

$$t \left[\sum_{i=1}^{H} \frac{1}{f_i} A_i - \sum_{i=1}^{H} B_i \right]$$

Derivando e igualando a cero (i = 1,...,H) :

$$F'_{fi} = - C_o N_i - C_1 N_i t_{Ni1} - C_2 N_i/k_i t_{Ni2} + t\ 1/f_i^2 A_i = 0$$

$$F'_{k_i} = C_2 f_i/k_i^2 N_i t_{Ni2} - t\ 1/f_i\ W_{Ni}/N\ t_{Ni2} S^2_{N_{i2}} = 0 \qquad (1)$$

Pasando cada término en t al segundo miembro de cada ecuación y diviendo miembro a miembro cada par de ecuaciones obtenemos :

$$\frac{C_o N_i + C_1 N_i t_{Ni1} + C_2 N_i/k_i t_{Ni2}}{C_2\ 1/k_i^2 N_i t_{Ni2}} = \frac{A_i}{W_{Ni}/N\ t_{Ni2} S^2_{N_{i2}}}$$

Multiplicando en cruz y despejando k_i^2, obtenemos el valor indicado.

Para calcular f_i procederemos de la siguiente forma :

En cada ecuación (1), después de simplificar ($W_{Ni} = N_i/N$), obtenemos:

$$f_i^2 = \frac{t\ S^2_{N_{i2}}\ k_i^2}{N^2\ C_2} \Rightarrow \quad f_i = \frac{t^{1/2}}{a_i} \quad ; \quad \text{siendo} \quad a_i = \frac{N\ (C_2)^{1/2}}{S_{N_{i2}}\ k_i}$$

Sustituyendo en la expresión :

$$V_o = \sum_{i=1}^{H} \frac{1}{f_i} A_i - \sum_{i=1}^{H} B_i$$

Obtenemos :

$$t^{1/2} = \frac{\sum_{i=1}^{H} a_i A_i}{V_o + \sum_{i=1}^{H} B_i}$$

que sustituida en la expresión de $f_i = t^{1/2}/a_i$ nos da la expresión propuesta.

c.q.d.

5. Estimadores de la media y del total de una población finita o de una parte de ella con el marco muestral no depurado, a través de una muestra de conglomerados bietápicos con no respuesta

Una de las operaciones más delicadas en el diseño muestral es el de la formación de la lista de unidades a muestrear que, junto con la información complementaria, se conoce con el nombre de marco.

En la práctica no suele disponerse de una lista actualizada de las unidades y aun conociéndola, si tal lista está muy dispersa por todo el territorio, el coste de realizar tal muestreo será prohitivo, aún cuando ésta situación sea deseable desde el punto de vista de aumentar la precisión y en consecuencia de reducir los errores de muestreo.

En general, es necesario recurrir a una muestra de grupos de unidades a los que denominamos conglomerados. En el conglomerado de dos etapas, dentro de cada conglomerado elegimos una muestra de unidades que constituye la segunda etapa.

Consideramos, en lo que sigue, a la población distribuida en N conglomerados, de los cuales seleccionamos una muestra de extensión n. El número de unidades del conglomerado i-ésimo es N_i, de las cuales seleccionamos una m.a.s. sin reemplazamiento de extensión n_i.

El problema surge, cuando al seleccionar una muestra dentro de cada conglomerado, algunas de las unidades no responden para una característica en particular. No es correcto basar los resultados de la encuesta únicamente en las unidades que respondieron en cada conglomerado, dado que quienes no lo hicieron son diferentes de los otros.

Nosotros vamos a considerar el problema de la no respuesta, en el sentido de que el conjunto de los que no responden "corta" a todos los conglomerados, encontrandonos con N_{i1} y N_{i2} unidades de N_i y n_{i1} y n_{i2} unidades de n_i que responden / no responden, respectivamente. En ambos casos $N_i = N_{i1} + N_{i2}$ y $n_i = n_{i1} + n_{i2}$.

En lo que sigue vamos a estudiar estimadores de la media y del total tanto de la población total como de la población de los que responden, utilizando para ello estimadores insesgados que para una m.a.s. hemos visto en el segundo y tercer apartado de éste capítulo.

5.1. Estimador A :

Consideremos que N_{i1} y N_{i2} son ambas dsconocidas para cualquier conglomerado. Si fijamos la fracción de muestreo f_i , el tamaño n_i de la muestra permanece fijo, siendo el resultado una muestra con igual probabilidad para cada elemento. Procederemos a asignar a cada unidad Y_{ij} de la población el siguiente valor :

$$Y_{ij} = \begin{cases} X_{ij} & \text{Si la unidad j del conglomerado i responde} \\ 0 & \text{Si la unidad no responde} \end{cases}$$

Definición 5.1.1 Estimador directo o expandido de

a) la media : $\bar{Y}_n = 1/n \cdot \sum_{i=1}^{n} \bar{Y}_{ni}$

b) el total : $\overset{*}{Y}_N = N/n \cdot \sum_{i=1}^{n} N_i \cdot \bar{Y}_{ni} = N/n \cdot \sum_{i=1}^{n} \overset{*}{Y}_{Ni}$, si convenimos en

denotar por $\overset{*}{Y}_{Ni} = N_i \cdot \bar{Y}_{ni}$, siendo $\overset{*}{Y}_{Ni}$ e $\bar{Y}_{ni}$ los estimadores del total y de la media en el conglomerado i-ésimo, cuya definición y propiedades son anàlogas a las indicadas en las proposiciones 2.1.2 a 2.1.7, considerando que la media y el total poblacional es el de cada conglomerado.

Así mismo, denotamos por Y_{Ni} e $\bar{Y}_{Ni}$ el total y la media poblacional del conglomerado i-ésimo y por Y_{Ni1} e $\bar{Y}_{Ni1}$ el total y la media poblacional de los que responden en el conglomerado i-ésimo.

Por $Y_N = \sum_{i=1}^{N} Y_{Ni}$ e $\bar{Y}_N = Y_N/N$ el total y la media de la población distribuida en N conglomerados.

Por $Y_M = \sum_{i=1}^{N} Y_{Ni1}$ e $\bar{Y}_M = Y_M/N$ el total y la media de los que responden en la población distribuida en N conglomerados.

5.1.1 Sesgo

Los estimadores propuestos estiman a la media y total de la población y de los que responden. Estudiamos la sesgadez ó insesgadez de cada uno de ellos.

Proposición 5.1.2 $\bar{Y}_n$ es un estimador insesgado de $\bar{Y}_N$ y sesgado de Y_M, media del conjunto poblacional y de la población de los que responden, respectivamente.

Demostración :

De la proposición 2.1.2 se tiene, para cualquier i :

$$E(\bar{Y}_{ni}) = \bar{Y}_{Ni} \quad (1) \qquad \text{y} \quad E(\bar{Y}_{ni}) = t_{Ni1} \cdot \bar{Y}_{Ni1} \quad (2)$$

Utilizando (1) :

$$E(\bar{Y}_n) = 1/n \cdot E_1\Big(\sum_{i=1}^{n} E_2(\bar{Y}_{ni})\Big) = 1/n \cdot E_1\Big(\sum_{i=1}^{n} Y_{Ni}\Big) =$$

$$1/n.\sum_{i=1}^{n} E_1(\bar{Y}_{Ni}) = 1/n.n.\sum_{i=1}^{N} 1/N \cdot \bar{Y}_{Ni} = \bar{Y}_N$$

que nos confirma la insesgadez.

Utilizando (2) :

$$E(\bar{Y}_n) = 1/n \cdot E_1\Big(\sum_{i=1}^{n} E_2(\bar{Y}_{ni})\Big) = 1/n \cdot E_1\Big(\sum_{i=1}^{n} t_{Ni1}.\bar{Y}_{Ni1}\Big) =$$

$$1/n.\sum_{i=1}^{n}[\,t_{Ni1}.E_1(\bar{Y}_{Ni1})] = 1/n.\sum_{i=1}^{n}[t_{Ni1}.\sum_{i=1}^{N} 1/N\,\bar{Y}_{Ni1}\,] =$$

$$1/n.\Big(\sum_{i=1}^{n} t_{Ni1}\Big)\sum_{i=1}^{N} 1/N \cdot \bar{Y}_{Ni1} = 1/n.\Big(\sum_{i=1}^{n} t_{Ni1}\Big).\bar{Y}_M$$

que nos confirma la sesgadez.

c.q.d.

Proposición 5.1.3 $\hat{Y}_N^*$ es un estimador insesgado de Y_N y de Y_M, total de la población y de la población que responden, respectivamente

Demostración :

Para cualquier i, se tiene : $E(\hat{Y}_{Ni}^*) = N_i \cdot E(\bar{Y}_{ni})$, consecuencia de la Definición 2.1.1; además utilizando la Definición 5.1.1 y (1) ó (2) de la proposición 5.1.2, se sigue :

Para Y_N :

$$E(\hat{Y}_N^*) = N/n \cdot E\Big(\sum_{i=1}^{n} N_i \cdot \bar{Y}_{ni}\Big) = N/n.\,E_1\Big(\sum_{i=1}^{n} N_i.\,E_2(\bar{Y}_{ni})\Big) =$$

$$N/n.\,E_1\Big(\sum_{i=1}^{n} N_i.\,\bar{Y}_{Ni}\Big) = N/n\,.\sum_{i=1}^{n} E_1(Y_{Ni}) = N.\sum_{i=1}^{N} 1/N.\,Y_{Ni} = \sum_{i=1}^{N} Y_{Ni} = Y_N$$

Para Y_M :

$$E(\hat{Y}_{Ni}^*) = N_i \cdot N_{i1} / N_i \cdot \bar{Y}_{Ni1} = N_{i1} \cdot \bar{Y}_{Ni1} = Y_{Ni1}$$

Luego :

$$E\,(\overset{*}{Y}_N) = N/n.\ E_1\ (\sum_{i=1}^{n} E_2\,(\overset{*}{Y}_{Ni})\ ) =$$

$$= N/n.\ \sum_{i=1}^{n} E_1\,(Y_{Ni1}) = N/n\ .\ n\ .\ 1/N\ .\ \sum_{i=1}^{N} Y_{Ni1} = Y_M$$

c.q.d.

5.1.2. Varianza de los estimadores y sus estimaciones

En lo que sigue hay que tener en cuenta la varianza ajustada o cuasi varianza poblacional entre los conglomerados y dentro de cada conglomerado. La primera representada por la variabilidad de la media de cada conglomerado respecto de la media poblacional (S_N^2) y la segunda representada por la variabilidad de las unidades de un conglomerado respecto de la media poblacional asociada a éste (S_{Ni}^2). Así pues :

Definición 5.1.4 Definimos varianza ajustada poblacional entre conglomerados y dentro de cada conglomerado, respectivamente

$$\overset{*}{S}_N^2 = 1/(N-1).\ \sum_{i=1}^{N} (\ \bar{Y}_{Ni} - \bar{Y}_N\ )^2 \quad , \quad y$$

$$S_{Ni}^2 = 1/(N_i-1).\ \sum_{j=1}^{Ni} (\ Y_{ij} - \bar{Y}_{Ni}\ )^2$$

Proposición 5.1.5 La varianza de $\bar{Y}_n$, toma la expresión

$$VAR\ (\ \bar{Y}_n\ ) = 1/n.(1-n/N).\overset{*}{S}_N^2 + 1/n.\ 1/N.\ \sum_{i=1}^{N} 1/n_i.\ (1-n_i/N_i).S_{Ni}^2$$

Demostración :

Por el teorema de Madow (Raj, 1968 y Azorín y Sánchez-Crespo, 1986) :

$$VAR\ (\ \bar{Y}_n\ ) = E_1.V_2\ (\bar{Y}_n) + V_1.E_2\ (\bar{Y}_n)$$

El cálculo de cada uno de los términos, es como sigue :

Como $E_2\ (\bar{Y}_n) = E_2\ (\ 1/n.\ \sum_{i=1}^{n} \bar{Y}_{ni}) = 1/n.\ \sum_{i=1}^{n} E_2\ (\bar{Y}_{ni}) = 1/n.\ \sum_{i=1}^{n} \bar{Y}_{Ni}$, se

tiene que :

$$V_1.E_2\ (\bar{Y}_n) = V_1\ (\ 1/n.\ \sum_{i=1}^{n} \bar{Y}_{Ni}\ ) = V_1\ (\text{ media de una m.a.s. })$$

$$= 1/n \cdot (1-n/N) \cdot \overset{*}{S}_N^2$$

Por otro lado :

$$V_2(\bar{Y}_n) = V_2\left(1/n \cdot \sum_{i=1}^{n} \bar{Y}_{ni}\right) = 1/n^2 \cdot \sum_{i=1}^{n} V_2(\bar{Y}_{ni})$$

$$= 1/n^2 \cdot \sum_{i=1}^{n} 1/n_i \cdot (1-n_i/N_i) \cdot S_{Ni}^2 \quad \text{(proposición 2.1.2)}$$

Finalmente :

$$E_1.V_2(\bar{Y}_n) = E_1\left[1/n^2 \cdot \sum_{i=1}^{n} 1/n_i \cdot (1-n_i/N_i) \cdot S_{Ni}^2\right] =$$

$$1/n^2 \cdot \sum_{i=1}^{n} E_1\left[1/n_i \cdot (1-n_i/N_i) \cdot S_{Ni}^2\right] =$$

$$1/n^2 \cdot n \cdot 1/N \cdot \sum_{i=1}^{N} 1/n_i \cdot (1-n_i/N_i) \cdot S_{Ni}^2 =$$

$$1/n \cdot 1/N \cdot \sum_{i=1}^{N} 1/n_i \cdot (1-n_i/N_i) \cdot S_{Ni}^2$$

La suma de ambas expresiones nos dá la descomposición de $VAR(\bar{Y}_n)$ propuesta.

c.q.d.

Podemos descomponer la varianza asignada a cada conglomerado de la muestra en suma de dos términos que dependen sólo de las unidades que responden

Proposición 5.1.6 La varianza de $\bar{Y}_n$, toma la expresión

$$VAR(\bar{Y}_n) =$$

$$1/n.(1-n/N).\overset{*}{S}_N^2$$

$$+$$

$$1/n \cdot 1/N \cdot \sum_{i=1}^{N} 1/n_i \cdot N_{i1}/(N_i-1).(1-n_i/N_i).[VAR_{Ni1} + t_{Ni2}.Y_{Ni1}^2]$$

Demostración : Es consecuencia de la proposición 2.1.6 que nos permite descomponer S_{Ni}^2, para cada i, en suma de dos términos que dependen sólo de las unidades que responden y de la proposición anterior.

c.q.d.

Definición 5.1.7 Definimos varianza ajustada muestral entre conglomerados y dentro de cada conglomerado :

$$\overset{*}{S}_n^2 = 1/(n-1).\ \overset{n}{\underset{i=1}{\Sigma}} \left(\bar{Y}_{ni} - \bar{Y}_n \right)^2 , \quad \text{y}$$

$$S_{ni}^2 = 1/(n_i-1).\ \overset{n}{\underset{j=1}{\Sigma}} \left(Y_{ij} - \bar{Y}_{ni} \right)^2$$

Proposición 5.1.8 Un estimador de la varianza de $\bar{Y}_n$ toma la expresión

$$\overset{*}{VAR}\left(\bar{Y}_n \right) = 1/n.(1-n/N).\overset{*}{S}_n^2 + 1/n.\ 1/N.\ \overset{N}{\underset{i=1}{\Sigma}}\ 1/n_i.\ (1-n_i/N_i).S_{ni}^2$$

Demostración :

En lo que sigue descomponemos $\overset{*}{S}_n^2$ y calculamos $E\ (\overset{*}{S}_n^2)$, a partir de la cual obtenemos la expresión que queremos demostrar, esto es :

$$E\ [\ \overset{*}{VAR}\ (\bar{Y}_n)\ ] = VAR\ (\bar{Y}_n)$$

Utilizamos en algunos pasos de la demostración la expresión práctica del cálculo de la varianza de una variable :

$$VAR\ (X) = E(X^2) - [E(X)]^2 \quad \text{o bien} \quad E(X^2) = VAR\ (X) + [E(X)]^2 \quad (1)$$

* Descomposición de $\overset{*}{S}_n^2$ (definición 5.1.7)

$$(n-1).\overset{*}{S}_n^2 = \overset{n}{\underset{i=1}{\Sigma}} \left(\bar{Y}_{ni} - \bar{Y}_n \right)^2 = ... = \overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}^2 - n\ \bar{Y}_n^2$$

Como $\bar{Y}_n = 1/n\ .\ \overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}$, se deduce :

$$(n-1).\overset{*}{S}_n^2 = \overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}^2 - 1/n\ .\ \left(\overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}\right)^2$$

* Esperanza de $\overset{*}{S}_n^2$:

$$E\ [(n-1).\overset{*}{S}_n^2] = E\ [\ \overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}^2\ ] - E\ [\ 1/n\ .\ \left(\overset{n}{\underset{i=1}{\Sigma}}\ \bar{Y}_{ni}\right)^2\ ] \qquad (2)$$

veamos la descomposición de cada uno de los términos de (2) :

$$** \quad E\left[\sum_{i=1}^{n} \bar{Y}_{ni}^2\right] = E_1\left(\sum_{i=1}^{n} E_2\,(\bar{Y}_{ni}^2)\right) = \qquad \text{aplicando (1)}$$

$$E_1\left(\sum_{i=1}^{n} [V_2(\bar{Y}_{ni}) + (E_2\,(\bar{Y}_{ni}))^2]\right) =$$

$$E_1\left(\sum_{i=1}^{n} [V_2(\bar{Y}_{ni}) + \bar{Y}_{Ni}^2]\right) =$$

$$\sum_{i=1}^{n} E_1\,(V_2(\bar{Y}_{ni})) + \sum_{i=1}^{n} E_1(\bar{Y}_{Ni})^2 =$$

$$n/N.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n/N.\sum_{i=1}^{N} \bar{Y}_{Ni}^2 \qquad (3)$$

[Hacemos notar que de acuerdo con la proposición 2.1.4, se tiene :

$$V_2(\bar{Y}_{ni}) = \text{VAR (media de una m.a.s)} = 1/n_i.(1-n_i/N_i).S_{Ni}^2] \quad (4)$$

$$** \quad E\left[1/n\,.\,(\sum_{i=1}^{n} \bar{Y}_{ni})^2\right] = 1/n\,.\,E_1\left[E_2\left(\sum_{i=1}^{n} \bar{Y}_{ni}\right)^2\right] = \text{aplicando (1)}$$

$$1/n\,.\,E_1\left[V_2\left(\sum_{i=1}^{n} \bar{Y}_{ni}\right) + (E_2\left(\sum^{n} \bar{Y}_{ni}\right))^2\right]=$$

$$1/n\,.\,E_1\left[\sum_{i=1}^{n} V_2(\bar{Y}_{ni}) + (\sum_{i=1}^{n} \bar{Y}_{Ni})^2\right] =$$

$$1/n\,.\,\sum_{i=1}^{n} E_1(\,V_2(\bar{Y}_{ni})\,) + 1/n\,.\,E_1(\sum_{i=1}^{n} \bar{Y}_{Ni})^2 =$$

$$1/n\,.\,n/N.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + 1/n\,.\,E_1(\sum_{i=1}^{n} \bar{Y}_{Ni})^2 =$$

aplicando (1) al segundo térnimo de ésta última expresión :

$$1/n\,.\,n/N.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + 1/n\,[V_1(\sum_{i=1}^{n} \bar{Y}_{Ni}) + (E_1(\sum_{i=1}^{n} \bar{Y}_{Ni}))^2] =$$

$$1/n\,.\,n/N.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n.\,V_1(\,1/n\,.\sum_{i=1}^{n} \bar{Y}_{Ni}) + 1/n.\,(\sum_{i=1}^{n} E_1\,(\bar{Y}_{Ni}))^2 =$$

$$1/n \;.\; n/N.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n.\; V_1(\; 1/n \;.\; \sum_{i=1}^{n} \bar{Y}_{Ni}) + 1/n.\; (n/N.\; \sum_{i=1}^{N} \bar{Y}_{Ni})^2 \;=$$

$$1/n \;.\; n/N.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n.\; V_1(\; 1/n \;.\; \sum_{i=1}^{n} \bar{Y}_{Ni}) + n\; \bar{Y}_N^2 \qquad (5)$$

[Hacemos notar que según la proposición 2.1.2 , se tiene :

$$V_1\; (1/n.\; \sum_{i=1}^{n} \bar{Y}_{Ni}) = \text{VAR (media de una m.a.s.)} = 1/n.\; (1 - n/N).S_N^{2\,*} \;] \qquad (6)$$

Sustituyendo en (2) las expresiones (3) y (5), se sigue :

$$E\;[(n-1).\overset{*}{S}_n^2] = [\; n/N.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n/N.\; \sum_{i=1}^{N} \bar{Y}_{Ni}^2 \;]$$

$$-$$

$$[1/n \;.\; n/N.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n.\; V_1(\; 1/n \;.\; \sum_{i=1}^{n} \bar{Y}_{Ni}) + n\; \bar{Y}_N^2 \;] =$$

Agrupando tèrminos y teniendo en cuenta (6) :

$$(n-1)/N \;.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + n/N.\; \sum_{i=1}^{N} \bar{Y}_{Ni}^2 - (1 - n/N) \;.\; \overset{*}{S}_N^2 - n\; \bar{Y}_N^2 =$$

$$(n-1)/N \;.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) - (1 - n/N) \;.\; \overset{*}{S}_N^2 + n/N\; [\; \sum_{i=1}^{N} \bar{Y}_{Ni}^2 - N \;.\; \bar{Y}_N^2 \;] =$$

Aplicando la definición 5.1.4 al tercer sumando de la expresión anterior, se tiene :

$$(n-1)/N \;.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) - (1 - n/N) \;.\; \overset{*}{S}_N^2 + n/N.\; (N-1) \;.\; \overset{*}{S}_N^2 =$$

$$(n-1)/N \;.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + (n-1) \;.\; \overset{*}{S}_N^2 =$$

$$(n-1)\; [\; 1/N \;.\; \sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + \overset{*}{S}_N^2 \;]$$

En definitiva, llegamos a la expresión :

$$E\,[(n-1).\overset{*}{S}_n^2] = (n-1)\,[\,1/N\,.\,\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) \;+\; \overset{*}{S}_N^2\,]$$

Simplificando, nos queda :

$$E\,[\overset{*}{S}_n^2] = 1/N\,.\,\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) \;+\; \overset{*}{S}_N^2 \qquad (7)$$

La expresión (7) es equivalente a :

$$E\,[1/n.(1-n/N).\overset{*}{S}_n^2] = 1/N.1/n.(1-n/N\,).\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + 1/n.(1-n/N).\,\overset{*}{S}_N^2 =$$

$$1/N\,.\,1/n\,.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) - 1/N^2\,.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) + \;1/n.\,(1-\,n/N).\;\overset{*}{S}_N^2$$

pasando al primer miembro el segundo término del segundo miembro, teniendo en cuenta que éste se transforma como sigue :

$$-\,1/N^2\,.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) \;=\; -\,1/N\,.\,1/n\;E\,(\sum_{i=1}^{n} V_2(\bar{Y}_{ni})\,)$$

queda :

$$E\,[\;1/n.\,(1\,-\,n/N\,)\,.\,\overset{*}{S}_n^2] + 1/N\,.\,1/n\;E\,(\sum_{i=1}^{n} V_2(\bar{Y}_{ni})\,) =$$

$$1/N\,.\,1/n\,.\sum_{i=1}^{N} V_2(\bar{Y}_{ni}) \;+\; 1/n.\,(1-\,n/N).\;\overset{*}{S}_N^2$$

Teniendo en cuenta (4) y la proposición 5.1.5, el segundo término de la igualdad anterior es $VAR(\bar{Y}_n)$. Por lo que escribimos :

$$VAR\,(\,\bar{Y}_n\,) = E\,[\;1/n.\,(1\,-\,n/N\,)\,.\,\overset{*}{S}_n^2 + 1/N\,.\,1/n\sum_{i=1}^{n} V_2(\bar{Y}_{ni})\,]$$

c.q.d.

5.2. Estimador F

A partir del estimador de Hansen y Hurwitz para una m.a.s. sin reemplazamiento, vamos a obtener un estimador de la media poblacional. Se propone una postestratificación, dentro de cada conglomerado de la

muestra, separando los que responden de los que no lo hacen. Con el fín de obtener información de éste segundo substrato, se toma una submuestra de un tamaño conveniente y se recoge información de ella.

Proposición 5.2.9 Para cada conglomerado i de la muestra de extensión n, del que obtenemos una muestra aleatoria simple de n_i unidades, de las que n_{i1} son unidades que responden y n_{i2} unidades que no responden, un estimador insesgado de la media poblacional para el caracter Y, es :

$$\bar{Y}_{ni} = \frac{n_{i1}\ \bar{Y}_{n_{i1}} + n_{i2}\ \bar{Y}_{u_i}}{n_i}$$

siendo :

$\bar{Y}_{n_{i1}}$ la media muestral de los respondientes

u_i el tamaño de una muestra de n_{i2} / k_i unidades obtenida de entre los que no responden

$\bar{Y}_{u_i}$ la media de la muestra correspondiente

Con una varianza :

$$VAR\ (\ \bar{Y}_{ni}\ ) = \frac{k_i - 1}{n_i}\ t_{Ni2}\ \frac{S^2}{N_{i2}} + \frac{1}{n_i}\left(1 - \frac{1}{N_i}\right)\frac{S^2}{N_i}$$

siendo :

$S^2_{N_{i2}}$ y $S^2_{N_i}$ las varianzas ajustadas para el total de unidades N_{i2} y N_i

Demostración :

Consecuencia de aplicar a cada conglomerado de la muestra el teorema 2.6.21

Definición 5.2.10 Estimador directo o expandido de la media

$$F{:}\bar{Y}_n = 1/n \cdot \sum_{i=1}^{n} \bar{Y}_{ni}$$

Siendo $\bar{Y}_{ni}$ la media muestral del conglomerado i, segun la proposición anterior.

5.2.1. Sesgo

EL estimador propuesto estima a la media la población y de los que responden. Estudiamos la sesgadez / insesgadez para cada uno de ellos

Proposición 5.2.11 $\bar{Y}_n$ es un estimador insesgado de $\bar{Y}_N$ y sesgado de $\bar{Y}_M$

Demostración :

$$E(\bar{Y}_n) = E_1\, E_2\, (\bar{Y}_n) = E_1[\ 1/n\ .\ \sum_{i=1}^{n} E_2\ (\bar{Y}_{ni})\ ] = E_1[\ 1/n\ .\ \sum_{i=1}^{n} \bar{Y}_{Ni})\ ] =$$

$$1/n.\ \sum_{i=1}^{n} E_1(\bar{Y}_{Ni}) = 1/n.\ n/N.\ \sum_{i=1}^{N} \bar{Y}_{Ni} =$$

$$1/N\ .\ \sum_{i=1}^{N} \bar{Y}_{Ni} = \bar{Y}_N$$

La demostración de la sesgadez de Y_M es análoga a la indicada en la proposición 5.1.2, siguiendo el procedimiento anterior.

5.2.2. Varianza

Obtenemos una expresión de la varianza del estimador $\bar{Y}_n$, descompuesta en tres términos: El primero corresponde a la varianza entre los conglomerados, el segundo a la varianza entre las unidades que respondieron dentro de cada estrato (Definición 5.1.4) y el tercero a las que responden en la segunda etapa (submuestreo).

Proposición 5.2.12 La varianza del estimador $\bar{Y}_n$ toma la siguiente expresión

$$VAR\ (\ \bar{Y}_n\ ) = 1/n\ .\ (\ 1 - n/N\ ).\ \frac{S^{*2}}{N} +$$

$$1/n.1/N.\ \sum_{i=1}^{N} \frac{k_i - 1}{n_i}\ t_{Ni2}\ S^2_{Ni2} + 1/n.\ 1/N.\ \sum_{i=1}^{N} \frac{1}{n_i}\ (\ 1 - \frac{1}{N_i}\ )\ \frac{S^2}{N_i}$$

Demostración :

Aplicando el teorema de Madow(Raj,1968 y Azorín y Sánchez-Crespo,1986)

$$VAR\ (\ \bar{Y}_n\ ) = E_1\ V_2\ (\ \bar{Y}_n\ ) + V_1\ E_2\ (\ \bar{Y}_n\ )$$

En la demostración de la proposición 5.1.5, vimos que :

$$V_1 E_2 (\bar{Y}_n) = 1/n . (1 - n/N). \overset{*}{S}_N^2 \qquad (1)$$

Por otro lado :

$$V_2 (\bar{Y}_n) = V_2 (1/n . \sum_{i=1}^{n} \bar{Y}_{ni}) = 1/n^2 . \sum_{i=1}^{n} V_2 (\bar{Y}_{ni}) =$$

$$1/n^2 . \sum_{i=1}^{n} [\frac{k_i - 1}{n_i} t_{Ni2} S^2_{N_{i2}} + \frac{1}{n_i} (1 - \frac{1}{N_i}) S^2_{N_i}] = 1/n^2 . \Sigma [*]$$

Luego :

$$E_1 V_2 (\bar{Y}_n) = 1/n^2 . \sum_{i=1}^{n} E_1 [*] = 1/n^2 . n . 1/N . \sum_{i=1}^{N} E_1 [*] \qquad (2)$$

Sumando (1) y (2), obtenemos la expresión propuesta

c.q.d.

Proposición 5.2.13 Un estimador de la varianza de $\bar{Y}_n$ toma la expresión

$$\overset{*}{VAR} (\bar{Y}_N) = 1/n.(1-n/N).\overset{*}{S}_n^2 + 1/n. 1/N. \sum_{i=1}^{n} 1/n_i. (1-n_i/N_i).S_{ni}^2 +$$

$$1/n. 1/N. \sum_{i=1}^{n} (K_i-1) /n_i. t_{ni2} . S_{ni2}^2$$

Demostración :

Es análoga a la realizada para la proposición 5.1.8, con la única diferencia de que donde ponemos $V_2(\bar{Y}_{ni})$, tomará por valor la expresión demostrada por Hansen y Hurwitz (teorema 2.6.21).

c.q.d.

CAPITULO III

CAPITULO III

ESTIMADORES POBLACIONALES OBTENIDOS POR PONDERACION ENTRE DOS O MAS ESTIMADORES, CON EL MARCO NO DEPURADO, A TRAVES DE UNA MUESTRA CON NO RESPUESTA

1. Introducción

Es conocido que con el m.a.s. es la forma más sencilla de autoponderación, dado que cada elemento de la muestra pondera 1/n en la media. A pesar de la sencillez de las muestras autoponderadas, Kish (1965) indica varias razones para añadir ponderaciones desiguales en partes de la muestra; una de éstas es la de balancear con ponderaciones desiguales las diferencias en no respuesta entre partes de la muestra, si ésta está estratificada. Para el caso de una muestra no estratificada, sugiere la necesidad de realizar alguna suposición acerca de la aleatoridad, proponiendo particionar la muestra en un nùmero finito de subconjuntos aleatoriamente seleccionados y ponderar cada partición.

Nosotros proponemos en lo que sigue particionar la muestra en un nùmero finito de subconjuntos aleatoriamente seleccionados, ponderar cada partición, utilizar para cada subconjunto uno de los estimadores propuestos en el capítulo anterior y definir el estimador media ponderada sobre el total de elementos de la nueva muestra. Para el caso en que la partición sea de dos subconjuntos, utilizaremos ternas de dos de entre los estimadores muestrales estudiados en el primer capítulo, obteniendo diversas expresiones del estimador media ponderada, en función de los estimadores empleados en la partición; dado que conocemos su clasificación según su varianza (apartado 3 del capítulo anterior), la combinación de estos estimadores la haremos en función de ésta.

Por otro lado, si para cada partición utilizamos el mismo estimador (Kish 1965 y Murthy y Sethi, 1961) la varianza de la media ponderada aumenta y el sesgo se mantiene en la medida que lo sea el estimador utilizado. Al utilizar distintos estimadores, nosotros hemos encontrado que, bajo ciertas condiciones, es posible no sólo disminuir la varianza sino encontrar un valor mínimo absoluto de ésta. Así mismo, al combinar diversos estimadores, sesgados e insesgados, disminuimos el sesgo del estimador media ponderada.

La notación que vamos a emplear en el desarrollo de éste capítulo es la utilizada en los capítulos anteriores.

2. Estimador media ponderada en una partición de m subconjuntos. Sesgo y varianza de éste estimador

Definición 2.1 Para una m.a.s de extensión n con no respuesta, dividida en m subconjuntos de unidades aleatoriamente seleccionadas, tales que para cada subconjunto expresamos por :

p_i , a su proporciòn en el total de la muestra ($\sum_{i=1}^{m} p_i = 1$)

$n_i = n.p_i$, $\bar{Y}_{pi}$, $Y_{pi} = n_i . \bar{Y}_{pi}$ y $W_i>0$, el número de unidades, media, total y ponderaciòn, respectivamente.

Se tiene que :

* En las condiciones anteriores, diremos que los m subconjuntos constituyen una partición de la muestra.

* El nùmero de unidades en la nueva muestra es :

$$k = \sum_{i=1}^{m} n_i . W_i = \sum_{i=1}^{m} n . p_i . W_i = n. \sum_{i=1}^{m} p_i . W_i$$

* La media en la nueva muestra, que la llamaremos media ponderada con referencia a la muestra de extensiòn n, es :

$$\bar{Y}_{k/n} = 1/k . \sum_{i=1}^{m} Y_{pi} . W_i = 1/k . \sum_{i=1}^{m} \bar{Y}_{pi} . n_i . W_i = \frac{\sum \bar{Y}_{pi} . p_i . W_i}{\sum p_i . W_i} \quad (i = 1 \ldots m)$$

Proposiciòn 2.2 La media ponderada $\bar{Y}_{k/n}$ es un estimador sesgado o insesgado de $\bar{Y}_N$, segun el estimador empleado en cada particiòn

Demostración :

De la definición 2.1. se sigue :

$$E(\bar{Y}_{k/n}) = \frac{\sum E(\bar{Y}_{pi}) . p_i . W_i}{\sum p_i . W_i} \quad (i = 1 \ldots m) \quad (1)$$

Si suponemos que $E(\bar{Y}_{pi}) = \bar{Y}_N$, para cualquier subconjunto, se sigue de (1) :

$$E(\bar{Y}_{k/n}) = \frac{\sum_{m} p_i . W_i}{\sum_{m} p_i . W_i} . \bar{Y}_N = \bar{Y}_N \quad (2)$$

y el estimador media ponderada es insesgado.

Si suponemos que $E(\bar{Y}_{pi}) = \bar{Y}_{N1}$, para cualquier subconjunto, se sigue de (1) :

$$E(\bar{Y}_{k/n}) = \frac{\sum_{m} p_i.W_i}{\sum_{m} p_i.W_i} . \bar{Y}_{N1} = \bar{Y}_{N1} \qquad (3)$$

y el estimador media ponderada tiene el mismo sesgo que el de los estimadores empleados en cada subconjunto.

Finalmente, para el caso en que los estimadores sean sesgados e insesgados, obtendremos un sesgo para el estimador media ponderada, intermedio entre un tipo (2) y otro (3). Concretaremos éste caso posteriormente (Proposición 3.1.2)

Proposición 2.3 La varianza del estimador $\bar{Y}_{k/n}$ viene determinada por la expresión :

$$VAR\ (\bar{Y}_{k/n}) = \frac{\Sigma\ p_i^2 .\ W_i^2 \cdot\ VAR(\bar{Y}_{pi})}{(\ \Sigma\ p_i.W_i\ )^2} \quad (i = 1 \ldots m\ )$$

Demostración :

Teniendo en cuenta la proposición anterior y las propiedades de la varianza, se sigue

$$VAR\ (\bar{Y}_{k/n}) = VAR\ \left[\frac{\Sigma\ \bar{Y}_{pi}.\ p_i.W_i}{\Sigma\ p_i.W_i}\right] = \frac{1}{(\Sigma\ p_i.W_i)^2} .\ VAR\ [\Sigma\ \bar{Y}_{pi}.\ p_i.W_i]$$

$$= \frac{\Sigma\ VAR(\bar{Y}_{pi}.\ p_i.W_i)}{(\Sigma\ p_i.W_i)^2} = \frac{\Sigma\ VAR(\bar{Y}_{pi}).p_i^2.\ W_i^2}{(\Sigma\ p_i.W_i)^2} \qquad (1)$$

c.q.d.

La expresión indicada para la varianza del estimador media ponderada en la proposición anterior es genérica, en tanto que no concretamos el tipo de estimador empleado para cada subconjunto de la partición. Veamos un caso particular, tomando como ejemplo el estimador $A:\bar{Y}_{pi}$ (apartado 2.1 del capítulo II) para cada subconjunto. Demostramos el aumento de la varianza, aludida en el apartado anterior, al emplear el mismo

estimador en todos los subconjuntos.

En lo que sigue, para expresar el tipo de estimador empleado en el subconjunto de extensión n_i, con una proporción p_i y una ponderación W_i, utilizaremos la misma notación empleada para la muestra de extensión n en el capítulo anterior; en consecuencia, teniendo en cuenta que $n_i = n.p_i$ escribiremos

$$A{:}\bar{Y}_{ni} \quad \text{ó} \quad A{:}\bar{Y}_{pi}$$

Proposición 2.4 La varianza del estimador $\bar{Y}_{n/k}$, utilizando el estimador $A{:}\bar{Y}_{pi}$ para cada subconjunto de la partición de la muestra de extensión n, es :

$$VAR\ (\bar{Y}_{k/n}) = \left[\frac{\Sigma\ p_i .\ W_i^2}{(\Sigma\ p_i.W_i)^2} - n/N\ .\ \frac{\Sigma\ p_i^2\ .\ W_i^2}{(\Sigma\ p_i.W_i)^2} \right].\ 1/n\ .\ S_N^2$$

Demostración :

De la proposición 2.1.4 , cap. II se tiene, para una muestra de extensión n_i :

$$VAR\ (\ A{:}\bar{Y}_{pi}\ ) = 1/n_i\ (\ 1 - n_i/N\ )\ S_N^2$$

y dado que $n_i = n.p_i$, se sigue :

$$VAR\ (\ A{:}\bar{Y}_{pi}\ ) = 1/(n.p_i)\ (\ 1 - (n.p_i)\ /\ N\ )\ S_N^2 \quad (1)$$

Por otro lado la expresión de la varianza, segun la proposición 2.3 toma la forma :

$$VAR\ (\bar{Y}_{k/n}) = \frac{\overset{m}{\Sigma}\ p_i^2.\ W_i^2 \cdot\ VAR(A{:}\bar{Y}_{pi})}{(\ \Sigma\ p_i.W_i\ )^2}$$

y sustituyendo (1) en ésta expresión, quedará

$$\frac{1}{(\ \Sigma\ p_i.W_i\ )^2} .\Sigma\ p_i^2.\ W_i^2 \cdot\ 1/(n.p_i)\ (\ 1 - (n.p_i)\ /\ N\ )\ S_N^2 =$$

$$\left[\frac{1}{(\ \Sigma\ p_i.W_i\ )^2}.\ \Sigma\ p_i.\ W_i^2 - \frac{1}{(\ \Sigma\ p_i.W_i\ )^2}\ .\ \Sigma\ p_i^2.\ W_i^2.n/N \right].1/n.S_N^2$$

que es la expresión propuesta

c.q.d.

Proposición 2.5 La varianza del estimador $\bar{Y}_{k/n}$, utilizando el estimador $A{:}\bar{Y}_{pi}$ para cada subconjunto de la partición de la muestra de extensión n, aumenta con respecto a la varianza del estimador $A{:}\bar{Y}_n$. Esto es :

$$VAR\ (\bar{Y}_{k/n}) > VAR\ (A{:}\bar{Y}_n)$$

Demostración :

Si descomponemos cada tèrmino de la proposición anterior en la forma que sigue :

$$\frac{\Sigma\ p_i.\ W_i^2}{(\Sigma\ p_i.W_i)^2} = 1 + \frac{\Sigma\ p_i.\ W_i^2 - (\Sigma\ p_i.\ W_i)^2}{(\Sigma\ p_i.\ W_i)^2} \qquad (i = 1 \ldots m)$$

$$= 1 + \frac{\sum_{i=1}^{m} p_i.\ W_i^2 - \sum_{i=1}^{m} p_i^2.\ W_i^2 - 2 \sum_{i<>j}^{m} p_i.\ p_j.\ W_i.\ W_j}{(\Sigma\ p_i.\ W_i)^2}$$

$$= 1 + [A]$$

$$\text{siendo } [A] = \frac{\sum_{i=1}^{m} W_i^2\ p_i(1-p_i) - 2 \sum_{i<>j}^{m} p_i.\ p_j.\ W_i.\ W_j}{(\Sigma\ p_i.\ W_i)^2} \qquad (1)$$

$$\frac{\Sigma\ p_i^2.\ W_i^2}{(\Sigma\ p_i.W_i)^2} = 1 + \frac{\Sigma\ p_i^2\ .\ W_i^2 - (\Sigma\ p_i.\ W_i)^2}{(\Sigma\ p_i.\ W_i)^2}$$

$$= 1 + \frac{\sum_{i=1}^{m} p_i^2\ .\ W_i^2 - \sum_{i=1}^{m} p_i^2.\ W_i^2 - 2 \sum_{i<>j}^{m} p_i.\ p_j.\ W_i.\ W_j}{(\Sigma\ p_i.\ W_i)^2}$$

$$= 1 - [B]$$

$$\text{siendo } [B] = \frac{2 \sum_{i<>j}^{m} p_i . p_j . W_i . W_j}{(\Sigma\ p_i . W_i)^2} \qquad (2)$$

Luego :

$$VAR\ (\bar{Y}_{k/n}) = 1/n\ (1 + [A] - n/N\ (\ 1 - [B]\ ))\ S_N^2$$

$$= 1/n\ (\ 1 - n/N\ )\ S_N^2 \ + 1/n\ ([A] + n/N\ [B]\ )\ S_N^2$$

$$- VAR\ (A:\bar{Y}_n) + 1/n\ ([A] + n/N\ [B]\ )\ S_N^2$$

c.q.d.

La demostración es válida para cualquiera de los estimadores estudiados en el capítulo II, dado que la descomposición que hacemos de la varianza en las expresiones (1) y (2) nos sirve para cualquier expresión de la varianza.

3. Estimador media ponderada en una partición de 2 subconjuntos

Pretendemos utilizar las ponderaciones con el fín de obtener un estimador, al que hemos denominado media ponderada, cuyo sesgo sea igual o menor y la varianza menor que la del estimador que podamos emplear en la muestra. La forma de conseguirlo es combinar varios estimadores, uno para cada partición.

En éste apartado reducimos la partición a dos subconjuntos, con el fín de obtener relaciones entre las ponderaciones que nos permitan conseguir tal objetivo. Hay dos caminos a seguir : Trabajar con el supuesto de que las ponderaciones se consideran consecutivas (Kish, 1965), esto es, W_i+1 y W_i o bien con la razón de ponderación, definida por $W = W_p / W_{i-p}$, $W>0$. Nosotros proponemos el segundo camino.

En lo que sigue consideramos una muestra de extensión n y la muestra ponderada de extensión k, en los tèrminos definidos en el apartado anterior de éste capìtulo (Definición 2.1).

3.1. Media ponderada y sesgo

Proposición 3.1.1 Para una m.a.s. de extensión n con no respuesta, dividida en dos subconjuntos aleatoriamente seleccionados, tales que p y 1-p son sus proporciones en el conjunto de unidades de la muestra, W_p y W_{i-p}, $\bar{Y}_p$ e $\bar{Y}_{i-p}$ las ponderaciones y las medias respectivas de cada uno de los subconjuntos y $W = W_p / W_{i-p}$ la razón entre ambas ponderaciones se tiene :

$$\bar{Y}_{k/n} = \frac{W.p.\bar{Y}_p + (1-p).\bar{Y}_{i-p}}{W.p + 1 - p}$$

$$VAR(\bar{Y}_{k/n}) = \frac{W^2 . p^2 . VAR(\bar{Y}_p) + (1-p)^2 . VAR(\bar{Y}_{1-p})}{(W.p + 1 - p)^2}$$

Demostración :

Utilizando la definición 2.1 y la proposición 2.3, se tiene :

$$\bar{Y}_{k/n} = \frac{W_p.p.\bar{Y}_p + W_{1-p} . (1-p). \bar{Y}_{1-p}}{W_p .p + W_{1-p} . (1 - p)}$$

$$VAR(\bar{Y}_{k/n}) = \frac{W_p^2 . p^2 . VAR(\bar{Y}_p) + W_{1-p}^2 . (1-p)^2 . VAR(\bar{Y}_{1-p})}{[W_p .p + W_{1-p} . (1 - p)]^2}$$

Dividiendo los dos miembros de cada expresión por W_{1-p} y teniendo en cuenta que $W = W_p / W_{1-p}$ se obtiene las expresiones propuestas

c.q.d.

En el apartado anterior vimos que el estimador media ponderada era insesgado o sesgado según el tipo particular de estimadores empleados en la partición (Proposición 2.2), sin concretar más. Vamos a ver que, al ponderar un estimador sesgado con otro insesgado o siendo los dos sesgados con distinto sesgo, el sesgo de la media ponderada es menor que el del estimador sesgado, utilizandolo para el conjunto de la muestra .

Proposición 3.1.2 Siendo $S:\bar{Y}_p$ la media asociada a un estimador sesgado, denotado por S, en la partición de la muestra de extensión n, de proporción p y ponderación W_p y tal que $E(S:\bar{Y}_p) = \delta.\bar{Y}_{N1}$ y siendo $I:\bar{Y}_{1-p}$ la media asociada a un estimador insesgado, denotado por I, en la partición de la muestra de extensión n, de proporción 1-p y ponderación W_{1-p} y tal que $E(I:\bar{Y}_{1-p}) = \bar{Y}_N$, si denotamos por $W = W_p / W_{1-p}$, el sesgo del estimador media ponderada (δ es una constante), es :

$$sesgo(\bar{Y}_{k/n}) = \frac{W. p}{W.p + 1 - p} . sesgo (S:\bar{Y}_n) \quad (1)$$

y es tal que :

$$sesgo(\bar{Y}_{k/n}) < sesgo(S:\bar{Y}_n)$$

Demostración :

De acuerdo con la proposición 3.1.1, la media ponderada para ésta partición es :

$$\bar{Y}_{k/n} = \frac{W.p.\bar{Y}_p + (1-p).\bar{Y}_{1-p}}{W.p + 1 - p}$$

y utilizando las hipótesis planteadas, su esperanza queda en la forma

$$E(\bar{Y}_{k/n}) = \frac{W.p.\delta.\bar{Y}_{N1} + (1-p).\bar{Y}_N}{W.p + 1 - p}$$

y el sesgo :

$$\text{sesgo}(\bar{Y}_{k/n}) = E(\bar{Y}_{k/n}) - \bar{Y}_N = \frac{W.p.\delta.\bar{Y}_{N1}}{W.p + 1 - p} + \left(\frac{1 - p}{W.p + 1 - p} - 1\right)\bar{Y}_N$$

$$= \frac{W.p.\delta.\bar{Y}_{N1}}{W.p + 1 - p} - \frac{W.p.\bar{Y}_N}{W.p + 1 - p}$$

$$= \frac{W.p}{W.p + 1 - p}\,(\delta.\bar{Y}_{N1} - \bar{Y}_N) \qquad (2)$$

Por otro lado :

$$\text{sesgo}\,(S:\bar{Y}_n) = E(S:\bar{Y}_n) - \bar{Y}_N = \delta.\bar{Y}_{N1} - \bar{Y}_N \qquad (3)$$

De (2) y (3) se deduce (1), lo que demuestra la primera parte de la proposición.

Si el cociente entre los sesgos (1) / (3) : $\frac{W.\,p}{W.p + 1 - p}$ es < 1

habremos demostrado la segunda parte, esto es que :

$$\text{sesgo}(\bar{Y}_{k/n}) < \text{sesgo}(S:\bar{Y}_n)$$

En efecto, supongamos que :

$$\frac{W.\,p}{W.p + 1 - p} > 1 \quad ==> \quad 0 > 1 - p \quad ==> \quad p > 1$$

lo que no es posible.

c.q.d.

Proposición 3.1.3 Fijado un valor para p, el valor de W que reduce el sesgo a la razón $r = a/b$ ($a<b$), es igual a :

$$W = \frac{(1-p).a}{(b-a).p} = \frac{(1-p).r}{p.\,(1-r)}$$

Demostración :

Utilizando la proposición anterior, se tiene que

$\frac{W.\,p}{W.p + 1 - p} = \frac{a}{b}$ y despejando W, obtenemos la primera expresión

Como $a = b.r$ y $b - a = b.(1-r)$, obtenemos la otra expresión

Proposición 3.1.4 A igual valor de W, el sesgo de la media ponderada aumenta conforme la proporción p aumenta. Así mismo, para un valor de p, el sesgo aumenta conforme la razón de ponderaciones, W, aumenta

Demostración :

Para dos proporciones p_1 y p_2, los sesgos respectivos son (proposición 3.1.2)

$$\frac{W.\,p_1}{W.p_1 + 1 - p_1} . \left(\delta.\bar{Y}_{N1} - \bar{Y}_N \right) \qquad (1)$$

y

$$\frac{W.\,p_2}{W.p_2 + 1 - p_2} . \left(\delta.\bar{Y}_{N1} - \bar{Y}_N \right) \qquad (2)$$

Si consideramos $p_1 < p_2$, deducimos que : $1-p_2 < 1-p_1$ y $W.p_1 < W.p_2$, dado que $W>0$.

Como todos los tèrminos de ambas desigualdades son positivos, deducimos que :

$$(1-p_2).W.p_1 < (1-p_1).W.p_2 \quad ==> (1-p_2) \,/\, W.p_2 < (1-p_1) \,/\, W.p_1$$

Luego : $1 + [\ (1-p_2) \,/\, W.p_2\ ] < 1 + [\ (1-p_1) \,/\, W.p_1\ ]$ (3)

Como el inverso de cada término de (3) es el primer factor que aparece en las expresiones (2) y (1), deducimos que (2) > (1)

Para la segunda cuestión, como la función :

$$Y = \frac{W.\ p}{W.p + 1 - p} \quad ==> \quad dY/dW = \frac{p.(1-p)}{(W.p + 1 - p)^2} > 0$$

es creciente para cualquier valor de W, en particular para W>0.

c.q.d.

Proposición 3.1.5 Siendo $S{:}\bar{Y}_p$ la media asociada a un estimador sesgado, denotado por S, en la partición de la muestra de extensión n, de proporción p y ponderación W_p y tal que $E(S{:}\bar{Y}_p) = \bar{Y}_{N1}$ y siendo $S_1{:}\bar{Y}_{1-p}$ la media asociada a otro estimador sesgado, denotado por S_1, en la partición de la muestra de extensión n, de proporción 1-p y ponderación W_{1-p} y tal que $E(S_1{:}\bar{Y}_{1-p}) = \delta.\bar{Y}_{N1}$ (δ es una constante, $\delta < 1$), si denotamos por $W = W_p / W_{1-p}$, el sesgo del estimador media ponderada es :

$$sesgo(\bar{Y}_{k/n}) = sesgo(S{:}\bar{Y}_n) - \frac{(1 - p).(1 - \delta)}{W.p + 1 - p} . \bar{Y}_{N1} \qquad (1)$$

y es tal que :

$$sesgo(\bar{Y}_{k/n}) < sesgo(S{:}\bar{Y}_n) \qquad (2)$$

Demostración :

De acuerdo con la proposición 3.1.1, la media ponderada para ésta partición es :

$$\bar{Y}_{k/n} = \frac{W.p.\bar{Y}_p + (1-p).\ \bar{Y}_{1-p}}{W.p + 1 - p}$$

y utilizando las hipótesis planteadas, su esperanza queda en la forma

$$E(\bar{Y}_{k/n}) = \frac{W.p.\bar{Y}_{N1} + (1-p).\delta.\ \bar{Y}_{N1}}{W.p + 1 - p}$$

y el sesgo :

$$sesgo(\bar{Y}_{k/n}) = E(\bar{Y}_{k/n}) - \bar{Y}_N = \left[\frac{W.p + 1 - p}{W.p + 1 - p} + \frac{(1-p)(\delta-1)}{W.p + 1 - p} \right].\bar{Y}_{N1} - \bar{Y}_N \qquad (3)$$

Como el sesgo del estimador S es

$$Sesgo(S:\bar{Y}_n) = E(\bar{Y}_n) - \bar{Y}_N = \bar{Y}_{N1} - \bar{Y}_N \qquad (4)$$

De (3) y (4), se sigue la expresión (1). Por otro lado, de (1) escribimos :

$$sesgo(S:\bar{Y}_n) - sesgo(\bar{Y}_{n/k}) = \frac{(1-p).(1-\delta)}{W.p + 1 - p} . \bar{Y}_{N1}$$

y al ser la expresión : $\dfrac{(1-p).(1-\delta)}{W.p + 1 - p} > 0$, para $0 < \delta < 1$

se sigue la expresión (2) propuesta

c.q.d.

Proposición 3.1.6 Para un $\delta>1$, a igual valor de W, el sesgo de la media ponderada aumenta conforme la proporción p disminuye. Así mismo, para un valor de p, el sesgo aumenta conforme la razón de ponderaciones, W, disminuye.

Demostración :

De la proposición 3.1.5, tenemos

$$sesgo(S:\bar{Y}_n) - sesgo(\bar{Y}_{k/n}) = \frac{(1-p).(1-\delta)}{W.p + 1 - p} . \bar{Y}_{N1}$$

Para dos proporciones p_1 y p_2, la diferencia de sesgos será

$$\frac{1 - p_1}{W.p_1 + 1 - p_1} .(1-\delta).\bar{Y}_{N1} \qquad (1)$$

$$\frac{1 - p_2}{W.p_2 + 1 - p_2} .(1-\delta).\bar{Y}_{N1} \qquad (2)$$

De $p_1< p_2$ ==> $1-p_2 < 1-p_1$ y $W.p_1 < W.p_2$, dado que $W>0$.

Como todos los tèrminos de ambas desigualdades son positivos, deducimos que :

$$(1-p_2).W.p_1 < (1-p_1).W.p_2 \;\Longrightarrow\; W.p_1/(1-p_1) < W.p_2/(1-p_2) \;\Longrightarrow$$

$$[W.p_1/(1-p_1)]+1 < [W.p_2/(1-p_2)]+1 \qquad (3)$$

Como el inverso de cada término de (3) es el primer factor que aparece en las expresiones (1) y (2) y el resto de los factores son iguales en ambas, deducimos que (1) > (2). En definitiva :

$$[\text{sesgo}(S:\bar{Y}_n)]_{p=p1} = \text{sesgo}(\bar{Y}_{k/n})+(1) > \text{sesgo}(\bar{Y}_{k/n})+(2) = [\text{sesgo}(S:\bar{Y}_n)]_{p=p2}$$

La demostración de la segunda cuestión es análoga a la indicada en la proposición 3.1.4

c.q.d.

En la proposición 3.1.1 dimos una expresión para la $VAR(\bar{Y}_{k/n})$, que es una función de W y p. Si fijamos el valor de p, la función varía según los valores de W.

Desde un punto de vista genérico vamos a estudiar ésta función considerando W como la única variable independiente y obtendremos las siguientes conclusiones que utilizaremos más adelante

Proposición 3.1.7 Denotando por $a_1 = p^2.VAR(\bar{Y}_p)$; $a_2 = (1-p)^2.VAR(\bar{Y}_{1-p})$ y $a_3 = 1 - p$ y para un valor fijo de p, tal que $0 < p < 1$, la función :

$$VAR(\bar{Y}_{k/n}) = Y = \frac{W^2.a_1 + a_2}{(W.p + a_3)^2} \qquad (1)$$

presenta un mínimo absoluto para el valor de W :

$$W_M = \frac{1-p}{p} \cdot \frac{VAR(\bar{Y}_{1-p})}{VAR(\bar{Y}_p)} \qquad (2)$$

Cuyo valor es :

$$Y_M = [\ 1/\ VAR(\bar{Y}_p) + 1/\ VAR(\bar{Y}_{1-p})\ ]^{-1} \qquad (3)$$

o bien

$$Y_M = VAR(\bar{Y}_{1-p}) \cdot \frac{1 - p}{1 + p.(W_M-1)} \qquad (4)$$

Demostración :

Observamos que tanto a_1, como a_2 y a_3 son términos positivos. La primera derivada de la función Y tiene por expresión :

$$\frac{dY}{dW} = \frac{2.a_1.a_3.W - 2.p.a_2}{(W.p + a_3)^3} = 0 \Longrightarrow W_M = \frac{p.a_2}{a_1.a_3} \qquad (5)$$

El valor $W_M > 0$, dado que los tèrminos que intervienen son todos positivos.

La segunda derivada, tiene por expresión :

$$\frac{d^2Y}{dW} = \frac{2.a_1.a_3.(W.p + a_3) - 3.p.(2.W.a_1.a_3 - 2.p.a_2)}{(W.p + a_3)^4}$$

al sustituir $W = W_M$, se tiene que : $2.W_M.a_1.a_3 - 2.p.a_2 = 0$, dado que es el numerador de la expresión para la primera derivada particularizada, del cual hemos obtenido W_M.

Particularizando y simplificando, nos queda :

$$\left[\frac{d^2Y}{dW^2}\right]_{W=W_M} = \frac{2.a_1.a_3}{(W_M.p + a_3)^3} > 0$$

Lo que nos confirma que para $W = W_M$, la función presenta un valor mínimo. Sustituyendo en (5) el valor de cada tèrmino, obtenemos el valor propuesto en (2).

Para determinar el valor mínimo, tendremos que particularizar la expresión (1)

$$[VAR(\bar{Y}_{k/n})]_{W=Wm} = Y_M \qquad (6)$$

Sustituyendo en (6), la expresión (5), tenemos :

$$Y_M = \frac{W_M^2.a_1 + a_2}{(W_M.p + a_3)^2} = \frac{p^2.a_1.a_2^2 + a_2.(a_1.a_3)^2}{(p^2.a_2 + a_1.a_3^2)^2} = \frac{a_1 \,.\, a_2}{p^2.a_2 + a_1.a_3^2} =$$

$$= \frac{1}{\frac{p^2}{a_1} + \frac{(1-p)^2}{a_2}} = \left[\frac{p^2}{a_1} + \frac{(1-p)^2}{a_2}\right]^{-1}$$

Sustituyendo en la última expresión obtenida, el valor de a_1 y a_2, se obtiene la expresión (3).

Para obtener la expresión (4), procederemos de la siguiente forma :

$$Y_M = \frac{W_M^2.a_1 + a_2}{(W_M.p + a_3)^2} = \frac{W_M^2.a_1 + a_2}{W_M.p + a_3} \,.\, \frac{1}{W_M.p + 1 - p} \qquad (7)$$

El primer factor, se transforma en la forma :

$$\frac{W_M^2.a_1 + a_2}{W_M.p + a_3} = \frac{[p^2.a_1.a_2^2 + a_2.(a_1.a_3)^2].a_1.a_3}{(p^2.a_2 + a_1.a_3^2).(a_1.a_3)^2} = \frac{a_2}{a_3}$$

Sustituyendo en (7), no queda la expresión propuesta (4)

Por otro lado, la función presenta una única asíntota de ecuación :

$$Y = \frac{p^2.VAR(\bar{Y}_p)}{p^2} = VAR(\bar{Y}_p) \quad (\text{asíntota horizontal })$$

Además es positiva, para cualquier valor de W (su gráfica está situada en la parte superior del eje de abcisas).

c.q.d.

3.2. Ponderación entre estimadores sesgados

Hemos visto en el primer apartado que la varianza del estimador media ponderada, cuando utilizamos en cada partición el mismo estimador, aumenta con respecto al estimador de la muestra, que se supone es el mismo que el utilizado en la partición.

Nosotros, en lo que sigue y en el subapartado siguiente, vamos a ver que combinando dos estimadores es posible disminuir la varianza del estimador media ponderada e incluso que ésta alcance un valor mínimo absoluto.

En particular, vemos dos casos generales según que los dos estimadores sean sesgados, objetivo de éste subapartado, o bien uno sesgado y otro insesgado, objetivo del siguiente subapartado.

Para el primer caso, combinaremos los estimadores sesgados B, C, D y G en función del orden que establecimos entre ellos, en cuanto a su varianza (Proposición 3.1.2, capítulo II), eligiendo tres de éstas combinaciones.

Para el segundo caso, combinaremos el de menor varianza de los sesgados, estimador D, con el mas eficiente de los insesgados, estimador A (Proposiciónes 3.1.2 y 3.1.4, capítulo II).

En ambos casos, proponemos dos valores de la proporción p, a los que denotaremos por p_o y p_1 , dentro de los cuales, existe un intervalo de extremo inferior cero, para la razón de ponderaciones W y en el que la varianza de la media muestral ponderada es mas pequeña que si empleáramos, a nivel muestral, el estimador cuya proporción es p.

3.2.1. Estimador C+B

Proposición 3.2.6 Para una m.a.s. de extensión n, con no respuesta, dividida en dos subconjuntos aleatoriamente seleccionados, tales que p y 1-p son sus proporciones en el conjunto de unidades de la muestra, W_p y W_{1-p}, $C{:}\bar{Y}_p$ y $B{:}\bar{Y}_{1-p}$, las ponderaciones y las medias respectivas de cada uno de los subconjuntos y siendo $W = W_p / W_{1-p}$ la razón entre ambas ponderaciones, se tiene

1) La media ponderada de la muestra, utilizando los estimadores C y B, tiene por expresión :

$$C{+}B{:}\bar{Y}_{k/n} = \frac{W.p.(C{:}\bar{Y}_p) + (1-p).(B{:}\bar{Y}_{1-p})}{W.p + 1 - p}$$

que en lo que sigue, la expresaremos por $\bar{Y}_{k/n}$

2) La media ponderada es un estimador sesgado y es tal que :

$$\text{sesgo}(\bar{Y}_{k/n}) = t_{N2} \cdot (\bar{Y}_{N1} - \bar{Y}_{N2}) = \text{sesgo}\ (C{:}\bar{Y}_n) \qquad (1)$$

3) La varianza de la media ponderada tiene por expresión :

$$\text{VAR}(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ \text{VAR}(C{:}\bar{Y}_p) + (1-p)^2\ .\ \text{VAR}(B{:}\bar{Y}_{1-p})}{(W.p + 1 - p)^2}$$

y al ser n.p y n.(1-p) las unidades de la muestra empleadas en cada partición, se tiene :

$$\text{VAR}\ (B{:}\bar{Y}_{1-p}) = \frac{1}{n(1-p)} \cdot \left[\ 1 - \frac{n(1-p)}{N_1}\ \right] \cdot S_{N1}^2$$

$$\text{VAR}\ (C{:}\bar{Y}_p) = \frac{1}{n_1.p} \cdot \left[\ 1 - \frac{n_1.p}{N_1}\ \right] \cdot S_{N1}^2$$

Demostración :

La primera y tercera parte son consecuencia de la proposición 3.1.1. Para ésta última hay que tener en cuenta, además, las proposiciones 2.2.10 y 2.3.13 del capítulo II.

La segunda parte es consecuencia de la proposición 2.2 . Para demostrar la igualdad (1) propuesta, procederemos de la siguiente forma :

* Dado que $E(C:\overline{Y}_p) = \overline{Y}_{N1}$ y $E(B:\overline{Y}_{1-p}) = \overline{Y}_{N1}$ (proposiciones 2.3.12 y 2.2.9, capítulo II)

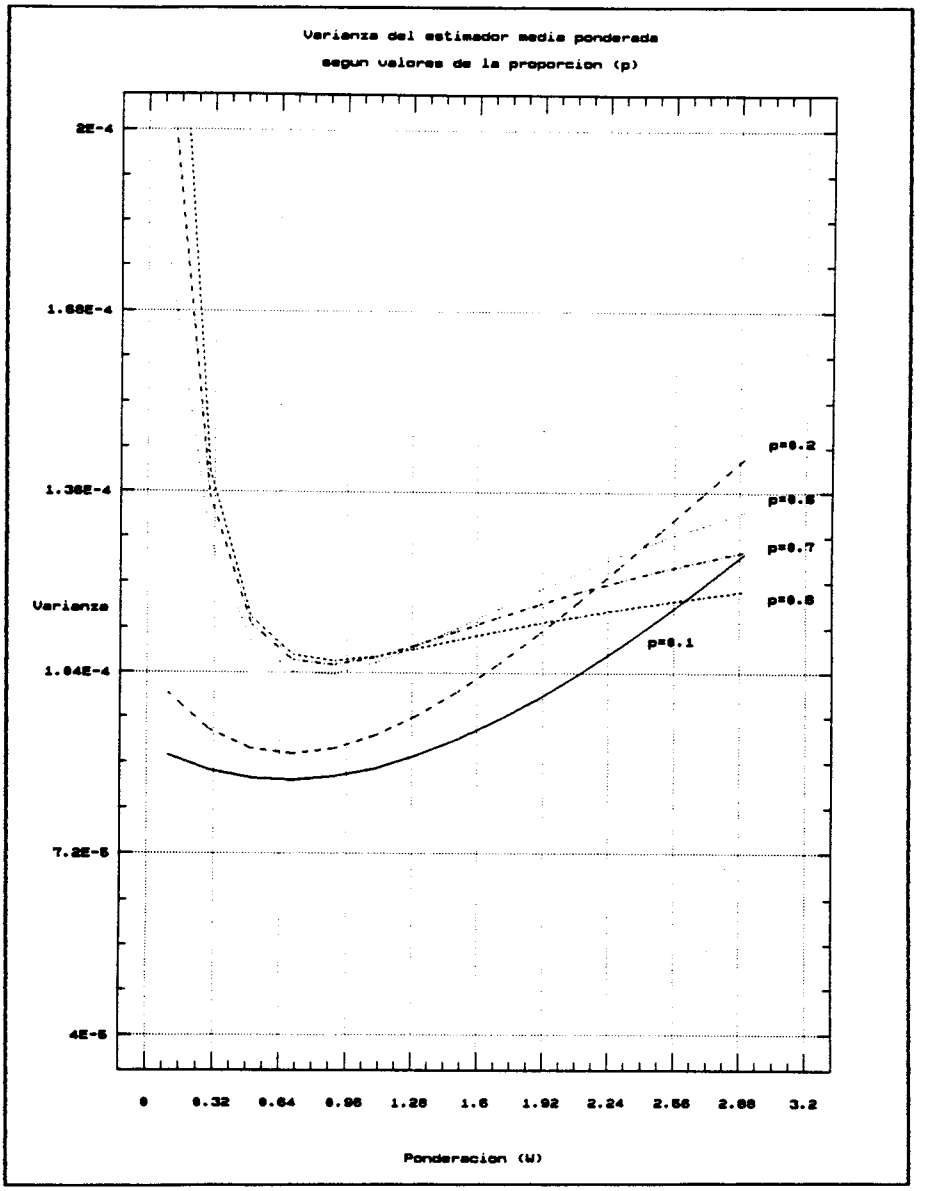
* Dado que $E(\overline{Y}_{k/n}) = \overline{Y}_{N1}$ (proposición 2.2)

* Los tres estimadores lo son de la media poblacional $\overline{Y}_N$

Se sigue que sus sesgos son iguales. El valor del sesgo es consecuencia de la proposición 2.3.12, capítulo II

c.q.d.

Gráfico III.1



En el gráfico III.1 indicamos para diferentes valores de p y de W, el comportamiento de la varianza de la media ponderada al utilizar una partición de dos subconjuntos de proporción p y 1-p y los estimadores C y B, para cada uno de ellos, respectivamente.

La tasa de respuesta es fija y la consideramos igual a 0.80. Así mismo consideramos la extensión de la muestra n = 10.000 y el total de unidades que responden N_1=50.000, por lo que F_1 = 5

El valor de la varianza del estimador muestral $C:\bar{Y}_n$, salvo el factor S_{N1}^2 , es de $1.05*10^{-4}$, obtenido de la expresión :

$$VAR\ (C:\bar{Y}_n) = \frac{1}{n_1} \cdot \left[1 - \frac{n_1}{N_1} \right] \cdot S_{N1}^2 \quad \text{(proposición 2.3.13)}$$

En el referido gráfico podemos observar (los datos que se indican, están obtenidos de la Tabla III.1 que se verá posteriormente) que :

* Hay valores de $p>p_o$ = 0.67935 en los que el valor de la varianza de la media ponderada está por encima del valor de la varianza del estimador muestral C, para cualquier valor de W.

* Hay valores de $p<p_o$ en los que el valor de la varianza de la media ponderada está por debajo del valor de la varianza del estimador muestral C, para un intervalo de valores de W que depende del valor empleado de p.

Así, por ejemplo, para p=0.5 el intervalo de valores para W es de [0, 1.16496]

* Dentro de éste intervalo de valores de W, hay uno para el que la varianza del estimador de la media pondera toma el valor mínimo.

Así, en el ejemplo de antes, para W_M = 0.78261 la varianza es mínima y de valor : $7.8537*10^{-5}$, salvo el factor S_{N1}^2 .

En la siguiente proposición generalizamos todo esto.

Proposición 3.2.7 Si ponderamos el estimador $C:\bar{Y}_p$ con el estimador $B:\bar{Y}_{1-p}$ en una partición de dos subconjuntos, existe una proporción p_o y p_1 tal que para cualquier proporción p, con la condición :

$$0< p \leq \min(1,p_o)$$

1) Tiene asociado un intervalo para los valores de W, de extremos :

$$0 \text{ y } a + [t_{n2}.(1/p - 1/p_o)]^{1/2} \quad \text{si } p \geq p_1$$

o bien

$a - [t_{n2}.(1/p - 1/p_o)]^{1/2}$ y $a + [t_{n2}.(1/p - 1/p_o)]^{1/2}$ si $p < p_1$

dentro de los cuales y, para cualquier valor de la razón W, se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (C:\bar{Y}_n)$$

2) Dentro del intervalo [0 , a] existe un valor de W :

$$W_M = \frac{t_{n1}.(F_1 - 1 + p)}{F_1 - p.t_{n1}}$$

para el que la varianza del estimador de la media ponderada toma el valor mínimo absoluto :

$$VAR(\bar{Y}_{k/n}) = \frac{F_1 - 1 + p}{F_1.n.[1+p.(W_M-1)]} \cdot S_{N1}^2$$

Siendo $F_1 = N_1 / n$, $a = 1 - t_{n1}/F_1$, $p_o = t_{n2}/(1-a^2)$, $p_1 = t_{n2}$

Demostración :

1) Utilizando la proposición 3.2.6, apartado 2 y la proposición 2.3.13 del capítulo II, escribimos :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ VAR(C:\bar{Y}_p) + (1-p)^2\ .\ VAR(B:\bar{Y}_{1-p})}{(W.p + 1 - p)^2} \qquad (1)$$

$$VAR\ (B:\bar{Y}_{1-p}) = \frac{1}{n(1-p)} \cdot \left[1 - \frac{n(1-p)}{N_1} \right] \cdot S_{N1}^2$$

$$VAR\ (C:\bar{Y}_p) = \frac{1}{n_1.p} \cdot \left[1 - \frac{n_1.p}{N_1} \right] \cdot S_{N1}^2 \qquad (2)$$

$$VAR\ (C:\bar{Y}_n) = \frac{1}{n_1} \cdot \left[1 - \frac{n_1}{N_1} \right] \cdot S_{N1}^2$$

Pretendemos demostrar bajo qué condiciones se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (C:\bar{Y}_n) \qquad (3)$$

Sustituyendo en la desigualdad (3) las expresiones (1) y (2) :

$$W^2.p^2.\ \frac{1}{n_1.p}.\left(1 - \frac{n_1.p}{N_1}\right) + (1-p)^2.\ \frac{1}{n(1-p)}.\left(1 - \frac{n(1-p)}{N_1}\right) < (W.p+1-p)^2.\frac{1}{n_1}.\left(1 - \frac{n_1}{N_1}\right)$$

Transformando la desigualdad :

$$W^2.p.(1/n_1-p/N_1) + (1-p).(1/n-(1-p)/N_1) - W^2.p^2.(1/n_1-1/N_1)$$

$$- 2.W.p.(1-p).(1/n_1-1/N_1) - (1-p)^2.(1/n_1-1/N_1) \quad < 0$$

y agrupando tèrminos, se tiene, despues de multiplicar por n_1 :

$$W^2.p.(1-p) - 2.W.p.(1-p).(1 - \frac{n}{N_1}.t_{n1}) + (1-p).t_{n1} - (1-p)^2 < 0$$

y simplificando :

$$W^2.p. - 2.W.p.(1 - \frac{n}{N_1}.t_{n1}) + t_{n1} - (1-p) < 0 \qquad (1)$$

Como $p>0$, la solución de la inecuación (1) se encuentra en el intervalo comprendido entre las soluciones de la ecuación :

$$W^2.p. - 2.W.p.(1 - \frac{n}{N_1}.t_{n1}) + t_{n1} - (1-p) = 0 \qquad (2)$$

El discriminante de ésta ecuación, tomando $a = 1 - (n/N_1).t_{n1}$, es :

$$4.p^2.(1-(n/N_1).t_{n1})^2 - 4.p.(t_{n1}+p-1) =$$

$$4.p.(p.a^2 - t_{n1} + 1 - p)$$

y la ecuación (2) tiene solución, siempre que se verifique :

$$4.p.(p.a^2 - t_{n1} + 1 - p) \geq 0$$

esto es, para valores de p tales que:

$$p \leq (1-t_{n1})/(1-a^2) = t_{n2}/(1-a^2) = p_o$$

El valor p_o es > 0, dado que $t_{n2} > 0$ y $1-a^2 > 0$ al ser $a < 1$

El intervalo pues, solución de la inecuación (1) , para valores $p \leq p_o$

$$[\ a - (t_{n2}/p - 1 + a^2)^{1/2} , a - (t_{n2}/p - 1 + a^2)^{1/2} \ ]$$

que de acuerdo con el valor obtenido para p_o, toma la expresión :

$$[\ a - (t_{n2}.(1/p - 1/p_o))^{1/2} , a + (t_{n2}.(1/p - 1/p_o))^{1/2} \ ]$$

Por otro lado, observamos que para $p = p_o$, el intervalo se reduce al punto $W = a$. Además, el valor de p a partir del cual el extremo inferior es nulo o positivo es :

$$a - (t_{n2}.(1/p - 1/p_o))^{1/2} \geq 0 \Rightarrow 1/p \leq 1/p_o + a^2/t_{n2} = 1/t_{n2}$$

esto es para valores de p tales que :

$$p \geq t_{n2} = p_1$$

Como p_o puede ser mayor que 1, restringimos todas las soluciones a valores de p, tales que : $0 < p \leq 1$

c.q.d.

2) Por la proposición 3.1.7 y para cualquier valor de p, existe un valor para W, para el cual la varianza de la media ponderada tiene un valor mínimo. Para el caso en el que estamos estos valores son :

$$W_M = \frac{1-p}{p} \cdot \frac{VAR(B:\bar{Y}_{1-p})}{VAR(C:\bar{Y}_p)} \qquad (4)$$

$$VAR(\bar{Y}_{k/n}) = VAR(B:\bar{Y}_{1-p}) \cdot \frac{1-p}{1 + p.(W_M-1)}$$

Teniendo en cuenta las expresiónes (2) , sustituyendo en (4) y simplificando, se tienen las expresiones propuestas al principio de la proposición. Demostramos que W_M está contenido en el intervalo [0 , a] subintervalo del obtenido en el apartado anterior de ésta proposición; para ello :

* Veamos que $W_M \geq 0$ para cualquier $p > 0$:

Como $F_1 = N_1/n > 1$ y $p.t_{n1} < F_1 \Rightarrow F_1 - p.t_{n1} > 0$, y por consiguiente para que :

$$W_M = \frac{t_{n1}.(F_1 - 1 + p)}{F_1 - p.t_{n1}} \geq 0 \Rightarrow F_1 - 1 + p \geq 0 \Rightarrow p \geq 1-F_1$$

y al ser $F_1 = N_1/n > 1$, se sigue que para todos los $p > 0$ es válida la desigualdad.

* Veamos que $W_M \leq a$, siendo $a = 1 - t_{n1}/F_1$, para cualquier $p \leq p_o$:

$$W_M = \frac{t_{n1}.(F_1 - 1 + p)}{F_1 - p.t_{n1}} = 1 - \frac{F_1.t_{n2} + t_{n1}.(1-2p)}{F_1 - p.t_{n1}}$$

Luego $W_M \leq 1 - t_{n1}/F_1$, es lo mismo que

$$\frac{F_1.t_{n2} + t_{n1}.(1-2p)}{F_1 - p.t_{n1}} \geq t_{n1}/F_1 \Rightarrow$$

$$F_1^2.t_{n2} + F_1.t_{n1} - 2.p.F_1.t_{n1} \geq F_1.t_{n1} - p.t_{n1}^2 \quad ==>$$

$$p \leq \frac{F_1^2.t_{n2}}{(2F_1 - t_{n1}).t_{n1}} = p_o = \frac{t_{n2}}{(2 - t_{n1}/F_1).t_{n1}/F_1} = \frac{t_{n2}}{(1+a).(1-a)}$$

c.q.d.

Como aplicación, consideremos una partición de dos subconjuntos de proporción p y 1-p y los estimadores C y B, para cada uno de ellas, respectivamente . Siguiendo la proposición 3.2.7, indicamos en la tabla III.1 para un valor N_1 = 50000 , n = 1000 (F_1=5) y t_{n1} = 0.70 , 0.80 y 0.90

Tabla III.1

t_{n1}	p_o	p_1	VAR(C:$\bar{Y}_n$) (A)	p	ES	W_M	VAR($\bar{Y}_{k/n}$) (B)	(B)/(A)
0.70	1.15207	0.30	.00012286	0.30	1.72000	0.62839	.00008328	0.67790
0.70	1.15207	0.30	.00012286	0.35	1.63249	0.64038	.00008351	0.67974
0.70	1.15207	0.30	.00012286	0.40	1.56771	0.65254	.00008362	0.68064
0.70	1.15207	0.30	.00012286	0.45	1.49739	0.66489	.00008361	0.68053
0.70	1.15207	0.30	.00012286	0.50	1.44275	0.67742	.00008346	0.67934
0.70	1.15207	0.30	.00012286	0.55	1.39390	0.69014	.00008317	0.67700
0.70	1.15207	0.30	.00012286	0.60	1.34948	0.70306	.000082742	0.67348
0.70	1.15207	0.30	.00012286	0.65	1.27008	0.71617	.000082157	0.66872
0.70	1.15207	0.30	.00012286	0.70	1.23363	0.72949	.000081417	0.66270
0.70	1.15207	0.30	.00012286	0.75	1.19852	0.74302	.000080519	0.65539
0.70	1.15207	0.30	.00012286	0.80	1.16420	0.75676	.000079463	0.64679
0.70	1.15207	0.30	.00012286	0.85	1.13006	0.77072	.000078250	0.63692
0.70	1.15207	0.30	.00012286	0.90	1.09534	0.78490	.000076884	0.62580
0.70	1.15207	0.30	.00012286	0.95	0.98566	0.79931	.000075370	0.61348
0.80	0.67935	0.20	.00010500	0.20	1.68000	0.69421	.000080951	0.77096
0.80	0.67935	0.20	.00010500	0.25	1.55106	0.70833	.000080899	0.77047
0.80	0.67935	0.20	.00010500	0.30	1.45014	0.72269	.000080715	0.76871
0.80	0.67935	0.20	.00010500	0.35	1.36634	0.73729	.000080392	0.76564
0.80	0.67935	0.20	.00010500	0.40	1.29343	0.75214	.000079924	0.76118
0.80	0.67935	0.20	.00010500	0.45	1.22736	0.76724	.000079307	0.75530
0.80	0.67935	0.20	.00010500	0.50	1.16496	0.78261	.000078537	0.74797
0.80	0.67935	0.20	.00010500	0.55	1.10313	0.79825	.000077612	0.73916
0.80	0.67935	0.20	.00010500	0.60	1.03732	0.81416	.000076534	0.72889
0.80	0.67935	0.20	.00010500	0.65	0.95529	0.83036	.000075304	0.71718
0.90	0.30525	0.10	.00009111	0.10	1.64000	0.75153	.000079987	0.87791
0.90	0.30525	0.10	.00009111	0.15	1.40229	0.76773	.000079780	0.87563
0.90	0.30525	0.10	.00009111	0.20	1.23521	0.78423	.000079428	0.87177
0.90	0.30525	0.10	.00009111	0.25	1.08907	0.80105	.000078926	0.86626
0.90	0.30525	0.10	.00009111	0.30	0.89572	0.81818	.000078269	0.85905

* los valores p_o y p_1

* La varianza del estimador $C:\bar{Y}_n$, salvo el factor S_{N1}^2

* Para diferentes valores de p comprendidos entre p_o y p_1 el intervalo de valores para W : [0, ES], dentro del cual :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (C:\bar{Y}_n)$$

donde $ES = a + [t_{n2}.(1/p - 1/p_o)]^{1/2}$, con la notación indicada en la proposición 3.2.7

* Así mismo el valor W_M , la varianza de la media ponderada para dicho valor salvo el factor S_{N1}^2 y la razón entre ambas varianzas .

Observamos que para una tasa de respuesta del 70% y una proporción p = 0.50 y tomando como razón de ponderaciones el valor W = 0.67742 la varianza muestral se reduce, utilizando el estimador media ponderada, en un 32% respecto de si utilizáramos el estimador C en toda la muestra.

3.2.2. Estimador G+B

Proposición 3.2.8 Para una m.a.s. de extensión n, con no respuesta, dividida en dos subconjuntos aleatoriamente seleccionados, tales que p y 1-p son sus proporciones en el conjunto de unidades de la muestra, W_p y W_{1-p}, $G:\bar{Y}_p$ y $B:\bar{Y}_{1-p}$, las ponderaciones y las medias respectivas de cada uno de los subconjuntos y siendo $W = W_p / W_{1-p}$ la razón entre ambas ponderaciones, se tiene

1) La media ponderada de la muestra, utilizando los estimadores G y B, tiene por expresión :

$$G+B:\bar{Y}_{k/n} = \frac{W.p.(G:\bar{Y}_p) + (1-p).(B:\bar{Y}_{1-p})}{W.p + 1 - p}$$

que en lo que sigue, la expresaremos por $\bar{Y}_{k/n}$

2) La media ponderada es un estimador sesgado y es tal que :

$$sesgo(\bar{Y}_{k/n}) = t_{N2} \cdot (\bar{Y}_{N1} - \bar{Y}_{N2}) = sesgo\ (G:\bar{Y}_n) \qquad (1)$$

3) La varianza de la media ponderada tiene por expresión :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ VAR(G:\bar{Y}_p) + (1-p)^2\ .\ VAR(B:\bar{Y}_{1-p})}{(W.p + 1 - p)^2}$$

y al ser n.p y n.(1-p) las unidades de la muestra empleadas en cada partición, se tiene :

$$VAR\ (B:\bar{Y}_{1-p}) = \frac{1}{n(1-p)} \cdot \left[1 - \frac{n(1-p)}{N_1} \right] \cdot S_{N1}^2$$

$$VAR\ (G:\bar{Y}_p) = \left[\frac{1}{n.p} \cdot \left(1 - \frac{n.p}{N_1}\right) + \frac{2}{n.p} \cdot \left(1 - \frac{n_1.p}{n.p}\right) \right] \cdot S_{N1}^2$$

Demostración :

La primera y tercera parte, son consecuencia de la proposición 3.1.1. Para ésta última parte hay que tener en cuenta, además, las proposiciones 2.2.10 y 2.7.26 del capítulo II.

La segunda parte es consecuencia de la proposición 2.2 . Para demostrar la igualdad (1) propuesta, procederemos de la siguiente forma :

* Dado que $E(G:\bar{Y}_p) = \bar{Y}_{N1}$ y $E(B:\bar{Y}_{1-p}) = \bar{Y}_{N1}$ (proposiciones 2.7.25 y 2.2.9, capítulo II)

* Dado que $E(\bar{Y}_{k/n}) = \bar{Y}_{N1}$ (proposición 2.2)

* Los tres estimadores lo son de la media poblacional $\bar{Y}_N$

Se sigue que sus sesgos son iguales. El valor del sesgo es consecuencia de la proposición 2.7.25, capítulo II

c.q.d.

Proposición 3.2.9 Si ponderamos el estimador $G:\bar{Y}_p$ con el estimador $B:\bar{Y}_{1-p}$ en una partición de dos subconjuntos, existe una proporción p_o y p_1 tal que para cualquier proporción p, con la condición :

$$0 < p \leq \min(1, p_o)$$

1) Tiene asociado un intervalo para los valores de W, de extremos:

$$0 \quad y \quad b/a + [\ 2.t_{n2}.\ 1/a.\ (1/p - 1/p_o)\ ]^{1/2}\ ,\ si\ p \geq \min(1,p_1)$$

o bien

$$b/a-[\ 2.t_{n2}.\ 1/a.\ (1/p - 1/p_o)\ ]^{1/2} \ y\ b/a + [\ 2.t_{n2}.\ 1/a.\ (1/p - 1/p_o)\ ]^{1/2}$$

para el caso en que $p < \min(1,p_1)$

dentro de los cuales, y para cualquier valor de la razón W, se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (G:\bar{Y}_n)$$

2) Dentro del intervalo [0 , b/a] , existe un valor de W :

$$W_M = \frac{F_1 - 1 + p}{a.F_1 - p}$$

para el que la varianza del estimador de la media ponderada toma el valor mínimo absoluto :

$$VAR(\bar{Y}_{k/n}) = \frac{F_1 - 1 + p}{F_1.n.[1+p.(W_M-1)]} \cdot S_{N1}^2$$

Siendo $F_1 = N_1/n$, $a = 3-2t_{n1}$, $b = a - n/N_1$, $p_o = (2.a.t_{n2}.F_1)/(a+b)$ y $p_1 = 2.t_{n2}/a$

Demostración :

Utilizando la proposición 3.1, teniendo en cuenta los estimadores que vamos a utilizar y las proposiciones 2.2.10 y 2.7.25 del capítulo II, escribimos :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.p^2.VAR(G:\bar{Y}_p) + (1-p)^2 . VAR(B:\bar{Y}_{1-p})}{(W.p + 1 - p)^2} \qquad (1)$$

$$VAR\ (B:\bar{Y}_{1-p}) = \frac{1}{n(1-p)} . \left[1 - \frac{n(1-p)}{N_1} \right] . S_{N1}^2$$

$$VAR\ (G:\bar{Y}_p) = \left[\frac{1}{n.p}.\left(1 - \frac{n.p}{N_1}\right) + \frac{2}{n.p}.\left(1- \frac{n_1.p}{n.p}\right) \right] . S_{N1}^2 \qquad (2)$$

$$VAR\ (G:\bar{Y}_n) = \left[\frac{1}{n}.\left(1 - \frac{n}{N_1}\right) + \frac{2}{n}.\left(1- \frac{n_1}{n}\right) \right] . S_{N1}^2$$

Pretendemos demostrar bajo qué condiciones se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (G:\bar{Y}_n) \qquad (3)$$

Sustituyendo en la desigualdad (3), las expresiones (1) y (2) :

$$W^2.p^2.\left[\frac{1}{n.p}.\left(1 - \frac{n.p}{N_1}\right) + \frac{2}{n.p}.\left(1- \frac{n_1}{n}\right) \right] + (1-p)^2.\frac{1}{n(1-p)}.\left(1 - \frac{n(1-p)}{N_1}\right) <$$

$$(W.p+1-p)^2.\left[\frac{1}{n}.\left(1- \frac{n}{N_1}\right) + \frac{2}{n}.\left(1- \frac{n_1}{n}\right) \right]$$

Transformando la desigualdad :

$$W^2.p.(1/n-p/N_1+2/n.(1-n_1/n)) + (1-p).(1/n-(1-p)/N_1)$$

$$- W^2.p^2.(1/n-1/N_1+2/n.(1-n_1/n)) - 2.W.p.(1-p).(1/n-1/N_1+2/n(1-n_1/n))$$

$$- (1-p)^2.\ (1/n-1/N_1+2/n.(1-n_1/n)) < 0$$

y agrupando tèrminos :

$$W^2.p/n.(1-p).(3-2n_1/n) - 2.W.p.(1-p)(3/n-1/N_1-2n_1/n^2) +$$

$$(1-p)/n.(3p-2+2(1-p).n_1/n) < 0$$

dividiendo por 1-p , multiplicando por n y llamando $a = 3 - 2t_{n1}$, se sigue :

$$W^2.p.a - 2.W.p.(\ a - n/N_1) + p.a - 2.t_{n2} < 0 \qquad (1)$$

Como $p.a > 0$ ($p > 0$ y $a > 0$), la solución de la inecuación (1) se encuentra en el intervalo comprendido entre las soluciones de la ecuación

$$W^2.p.a - 2.W.p.(\ a - n/N_1) + p.a - 2.t_{n2} = 0 \qquad (2)$$

El discriminante de ésta ecuación, tomando $b = a - n/N_1$, es :

$$4.p^2.b^2 - 4.p.a.(p.a - 2.t_{n2}) =$$

$$4.p.[p.(b^2-a^2) + 2.p.a.t_{n2}] =$$

$$4.p.(-p.n/N_1.(a+b) + 2.a.t_{n2})$$

La ecuación (2) tiene solución siempre que se verifique :

$$4.p.(-p.n/N_1.(a+b)+2.a.t_{n2}) \geq 0$$

esto es, para valores de p tales que:

$$p \leq 2.a.t_{n2}.N_1 / n.(a+b) = p_o = 2.a.t_{n2}.F_1 / (a+b)$$

Como $a+b > 0$ ($a = 3-2t_{n1} = 1+2t_{n2} > 1 \Rightarrow 2a > n/N_1$, pues $n/N_1 < 1 \Rightarrow 2a - n/N_1 = a+b > 1$) y el resto de los elementos son positivos, se sigue que $p_o > 0$

El intervalo pues, solución de la inecuación (1) , para valores $p \leq p_o$ tiene de extremos :

$$b/a - (1/(p.a^2).(-p.n/N_1.(a+b)+2.a.t_{n2}))^{1/2} \quad y$$

$$b/a + (1/(p.a^2).(-p.n/N_1.(a+b)+2.a.t_{n2}))^{1/2}$$

que de acuerdo con el valor obtenido para p_o, toma la expresión :

$$[\ b/a - (\ 2.t_{n2}.1/a.(1/p - 1/p_o)\ )^{1/2}\ ,\ b/a + (\ 2.t_{n2}.1/a.(1/p - 1/p_o))^{1/2}\ ]$$

Por otro lado, observamos que para $p = p_o$, el intervalo se reduce al punto $W = b/a$. Además, el valor de p a partir del cual el extremo inferior es nulo o positivo es :

$$b/a - (\ 2.t_{n2}.1/a.(1/p - 1/p_o)\ )^{1/2} \geq 0 \quad ==> \quad 1/p \leq 1/p_o + b^2/\ 2.a.t_{n2} \quad ==>$$

$1/p \geq a\ /\ 2.\ t_{n2}$, dado que $1/p_o + b^2/\ 2.a.t_{n2} = a\ /\ 2.\ t_{n2}$

esto es para de p, tales que :

$$p \geq 2.t_{n2}/a = p_1$$

Como p_o y p_1 puede ser mayor que 1, restringimos todas las soluciones a valores de p, tales que : $0 < p \leq 1$

2) Por la proposición 3.1.7 y para cualquier valor de p, existe un valor para W para el cual la varianza de la media ponderada tiene un valor mínimo. Para el caso en el que estamos estos valores son :

$$W_M = \frac{1-p}{p} \cdot \frac{VAR(B:\bar{Y}_{1-p})}{VAR(G:\bar{Y}_p)} \qquad (4)$$

$$VAR(\bar{Y}_{k/n}) = VAR(B:\bar{Y}_{1-p}) \cdot \frac{1-p}{1 + p.(W_M-1)}$$

Teniendo en cuenta la expresió (2), sustituyendo en (4) y simplificando, se tienen las expresiones propuestas al principio de la proposición.

Demostramos que W_M está contenido en el intervalo [0 ,b/a] subintervalo del obtenido en el apartado anterior de ésta proposición; para ello :

* Veamos que $W_M \geq 0$ para cualquier $p > 0$:

Como $a = 3 - 2t_{n1} = 1 + 2\ t_{n2} > 1$, $F_1 = N_1/n > 1$ y $p < 1$ ==> $a.F_1 - p > 0$, y por consiguiente para que :

$$W_M = \frac{F_1 - 1 + p}{a.F_1 - p} \geq 0 \quad ==> \quad F_1 - 1 + p \geq 0 \quad ==> p \geq 1-F_1$$

y al ser $F_1 = N_1/n > 1$, se sigue que para todos los $p > 0$ es válida la desigualdad.

* Veamos que $W_M \leq b/a$, siendo $a = 3 - 2\ t_{n1}$ y $b = a - 1/F_1$, para cualquier $p \leq p_o$:

$$W_M = \frac{F_1 - 1 + p}{a.F_1 - p} = 1 - \frac{2.F_1 - 2.p + 1 - 2.t_{n1}.F_1}{a.F_1 - p}$$

Luego $W_M \leq b/a = 1 - 1/(a.F_1)$ es lo mismo que

$$\frac{2.F_1 - 2.p + 1 - 2.t_{n1}.F_1}{a.F_1 - p} \geq 1/(a.F_1) \quad ==>$$

$$2.a.F_1^2 - 2.p.a.F_1 + a.F_1 - 2.a.t_{n1}.F_1^2 \geq a.F_1 - p \quad ==>$$

$$p \leq \frac{2.a.F_1^2 - 2.a.t_{n1}.F_1^2}{2.a.F_1 - 1} = p_o = \frac{2.a.F_1.t_{n2}}{2.a - 1/F_1} = \frac{2.a.F_1.t_{n2}}{a + b}$$

dado que $a + b = a + a - 1/F_1 = 2.a - 1/F_1$

c.q.d.

3.2.3. Estimador B+D

Proposición 3.2.10 Para una m.a.s. de extensión n, con no respuesta, dividida en dos subconjuntos aleatoriamente seleccionados, tales que p y 1-p son sus proporciones en el conjunto de unidades de la muestra, W_p y W_{1-p}, $B:\bar{Y}_p$ y $D:\bar{Y}_{1-p}$, las ponderaciones y las medias respectivas de cada uno de los subconjuntos y siendo $W = W_p / W_{1-p}$ la razón entre ambas ponderaciones, se tiene

1) La media ponderada de la muestra, utilizando los estimadores B y D, tiene por expresión :

$$B+D:\bar{Y}_{k/n} = \frac{W.p.(B:\bar{Y}_p) + (1-p).(D:\bar{Y}_{1-p})}{W.p + 1 - p}$$

que en lo que sigue, la expresaremos por $\bar{Y}_{k/n}$

2) La media ponderada es un estimador sesgado y es tal que :

a)

$$\text{sesgo}(\bar{Y}_{k/n}) = \text{sesgo}(B{:}\bar{Y}_n) - \frac{(1-p).t_{n2}}{W.p + 1 - p} . \bar{Y}_{N1} \quad (1)$$

y

b)

$$\text{sesgo}(\bar{Y}_{k/n}) < \text{sesgo}(B{:}\bar{Y}_n) \quad (2)$$

3) La varianza de la media ponderada tiene por expresión :

$$\text{VAR}(\bar{Y}_{k/n}) = \frac{W^2.\, p^2.\, \text{VAR}(D{:}\bar{Y}_p) + (1-p)^2 \,.\, \text{VAR}(B{:}\bar{Y}_{1-p})}{(W.p + 1 - p)^2}$$

y al ser n.p y n.(1-p) las unidades de la muestra empleadas en cada partición, se tiene :

$$\text{VAR}\,(D{:}\bar{Y}_{1-p}) = \frac{n_1.(1-p)}{[n.(1-p)]^2} . \left[1 - \frac{n_1.(1-p)}{N_1} \right] . S_{N1}^2$$

$$\text{VAR}\,(B{:}\bar{Y}_p) = \frac{1}{n.p} . \left[1 - \frac{n.p}{N_1} \right] . S_{N1}^2$$

Demostración :

La primera y tercera parte son consecuencia de la proposición 3.1.1. Para ésta última parte hay que tener en cuenta, además, las proposiciones 2.2.10 y 2.4.16 del capítulo II.

La segunda parte es consecuencia de la proposición 3.1.5, tomando el valor $\delta = t_{n1}$, dado que :

$E(D{:}\bar{Y}_p) = t_{n1}.\bar{Y}_{N1}$ y $E(B{:}\bar{Y}_{1-p}) = \bar{Y}_{N1}$ (proposiciones 2.4.15 y 2.2.9, capítulo II)

c.q.d.

Proposición 3.2.11 Si ponderamos el estimador $B{:}\bar{Y}_p$ con el estimador $D{:}\bar{Y}_{1-p}$ en una partición de dos subconjuntos, existe una proporción p_o y p_1 tal que para cualquier proporción p, con la condición :

$$0 < p \leq \min(1, p_o)$$

1) Tiene asociado un intervalo para valores de W, de extremos:

$$0 \quad y \quad a + [\ t_{n2}.b.(1/p - 1/p_o)\ ]^{1/2}\ ,\ \text{siendo}\ p \geq \min(1,p_1)$$

o bien

$$a - [\ t_{n2}.b.(1/p - 1/p_o)\ ]^{1/2} \quad y \quad a + [\ t_{n2}.b.(1/p - 1/p_o)\ ]^{1/2}$$

siendo $p < \min(1,p_1)$

dentro de los cuales y para cualquier valor de la razón W, se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (B{:}\bar{Y}_n)$$

2) Dentro del intervalo [0 , a] , existe un valor de W :

$$W_M = \frac{F_1 - t_{n1}.(1 - p)}{F_1 - p}\ .\ t_{n1}$$

para el que la varianza del estimador de la media ponderada toma el valor mínimo absoluto :

$$VAR(\bar{Y}_{k/n}) = t_{n1}.\ \frac{F_1 - t_{n1}.(1 - p)}{F_1.n.[1+p.(W_M-1)]}\ .\ S_{N1}^2$$

Siendo $F_1 = N_1/n$, $a = 1 - n/N_1$, $b = a - n/N_1.t_{n1}$

$$p_o = \frac{b.t_{n2}.F_1}{a + t_{n1}^2}\ ,\ p_1 = \frac{b.t_{n2}.F_1}{t_{n1}^2 + a.F_1}$$

Demostración :

Utilizando la proposición 3.1, teniendo en cuenta los estimadores que vamos a utilizar y las proposiciones 2.2.10 y 2.4.16 del capítulo II, escribimos :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ VAR(B{:}\bar{Y}_p) + (1-p)^2\ .\ VAR(D{:}\bar{Y}_{1-p})}{(W.p + 1 - p)^2} \qquad (1)$$

$$VAR\ (D{:}\bar{Y}_{1-p}) = \frac{n_1.(1-p)}{[n.(1-p)]^2}\ .\ [\ 1 - \frac{n_1.(1-p)}{N_1}\ ]\ .\ S_{N1}^2$$

$$VAR\ (B{:}\bar{Y}_p) = \frac{1}{n.p} \cdot \left[\ 1 - \frac{n.p}{N_1}\ \right] \cdot S_{N1}^2 \qquad (2)$$

$$VAR\ (B{:}\bar{Y}_n) = \frac{1}{n} \cdot \left[\ 1 - \frac{n}{N_1}\ \right] \cdot S_{N1}^2$$

Pretendemos demostrar bajo qué condiciones se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (B{:}\bar{Y}_n) \qquad (3)$$

Sustituyendo en la desigualdad (3), las expresiones (1) y (2) :

$$W^2.p^2.\frac{1}{n.p}.\left(1 - \frac{n.p}{N_1}\right) + (1-p)^2.\frac{n_1}{n^2.(1-p)}.\left(1 - \frac{n_1.(1-p)}{N_1}\right) < (W.p+1-p)^2.\frac{1}{n}.\left(1 - \frac{n}{N_1}\right)$$

Transformando la desigualdad :

$$W^2.p.(1/n-p/N_1) + n_1/n^2.\ (1-p).(1-(n_1.(1-p)/N_1)) - W^2.p^2.(1/n-1/N_1)$$

$$- 2.W.p.(1-p).(1/n-1/N_1) - (1-p)^2.(1/n-1/N_1) < 0$$

y agrupando tèrminos, se tiene :

$$W^2.p.(1-p).1/n - 2.W.p.(1-p).1/n.(1-n/N_1)$$

$$+ (1-p).1/n.[t_{n1}-t_{n1}^2.n/N_1.(1-p)-(1-p).(1-n/N_1)] < 0$$

multiplicando por n y simplificando :

$$W^2.p - 2.W.p.(1-n/N_1) + t_{n1} - t_{n1}^2.n/N_1.(1-p) - (1-p).(1-n/N_1) < 0$$

denotando por $a = 1 - n/N_1$ y $c = t_{n1}^2.\ n/N_1$, se tiene :

$$W^2.p - 2.W.p.a + t_{n1} - (1-p).(a+c) < 0 \qquad (1)$$

Como $p > 0$, la solución de la inecuación (1) se encuentra en el intervalo comprendido entre las soluciones de la ecuación :

$$W^2.p - 2.W.p.a + t_{n1} - (1-p).(a+c) = 0 \qquad (2)$$

El discriminante de ésta ecuación, después de simplificar, y tomando $b = a - (n/N_1).t_{n1}$ es :

$$4.p.[\ p.a^2 - t_{n1} + (a + c).(1 - p)\ ] =$$

$$4.p.[\ -p.n/N_1.(a + t_{n1}^2) + a - t_{n1} + n/N_1.\ t_{n1}^2\ ] =$$

$$(a - t_{n1} + n/N_1 . t_{n1}^2 = t_{n2} - n/N_1.(1+t_{n1}).t_{n2} = b.t_{n2})$$

$$4.p.[-p.n/N_1.(a+t_{n1}^2) + b.t_{n2}]$$

La ecuación (2) tiene solución siempre que se verifique :

$$4.p.[-p.n/N_1.(a+t_{n1}^2) + b.t_{n2}] \geq 0$$

esto es, para valores de p tales que:

$$p \leq b.t_{n2}.N_1 / n.(a+t_{n1}^2) = p_o = b.t_{n2}.F_1 / (a+t_{n1}^2)$$

El valor p_o es > 0, dado que todos los tèrminos son positivos.

El intervalo pues, solución de la inecuación (1) , para valores $p \leq p_o$ tiene de extremos :

$$a - (1/p.(-p.n/N_1.(a+t_{n1}^2)+b.t_{n2}))^{1/2}$$

y

$$a + (1/p(-p.n/N_1.(a+t_{n1}^2)+ b.t_{n2}))^{1/2}]$$

que de acuerdo con el valor obtenido para p_o, toma la expresión :

$$[a - (b.t_{n2}.(1/p - 1/p_o))^{1/2} , a + (b.t_{n2}.(1/p - 1/p_o))^{1/2}]$$

Por otro lado, observamos que para $p = p_o$, el intervalo se reduce al punto W = a. Además, el valor de p a partir del cual el extremo inferior es nulo o positivo es :

$$a - (b.t_{n2}.(1/p - 1/p_o))^{1/2} \geq 0 \Longrightarrow 1/p \leq 1/p_o + a^2/b.t_{n2}$$

esto es para valores de p, tales que :

$$p \geq b.t_{n2}/(a + (n/N_1).t_{n1}^2) = p_1 = b.t_{n2}.F_1 / (a.F_1 + t_{n1}^2)$$

dado que p_o y p_1 pueden ser mayor que 1, restringimos todas las soluciones a valores de p, tales que : $0 < p \leq 1$

c.q.d.

2) Por la proposición 3.1.7 y para cualquier valor de p, existe un valor para W, para el cual la varianza de la media ponderada tiene un valor mínimo. Para el caso en el que estamos estos valores son :

$$W_M = \frac{1-p}{p} \cdot \frac{VAR(D:\bar{Y}_{1-p})}{VAR(B:\bar{Y}_p)} \qquad (4)$$

$$VAR(\bar{Y}_{k/n}) = VAR(D:\bar{Y}_{1-p}) \cdot \frac{1 - p}{1 + p.(W_M-1)}$$

Teniendo en cuenta las expresiónes (2) , sustituyendo en (4) y simplificando se tienen las expresiones propuestas al principio de la proposición. Demostramos que W_M está contenido en el intervalo [0 , a] subintervalo del obtenido en el apartado anterior de ésta proposición; para ello :

* Veamos que $W_M \geq 0$ para cualquier $p > 0$:

Como $F_1 = N_1/n > 1$ y $p < 1$ ==> $F_1 - p > 0$, y por consiguiente para que :

$$W_M = t_{n1}.\frac{F_1 - t_{n1}.(1 - p)}{F_1 - p} \geq 0 \quad ==> \quad F_1 \geq t_{n1}.(1 - p) \quad ==>$$

$$F_1/t_{n1} \geq 1 - p \; ==> \; p \geq 1 - F_1/t_{n1}$$

y al ser $1-F_1/ t_{n1} < 0$, se sigue que para todos los $p > 0$ es válida la desigualdad.

* Veamos que $W_M \leq a$, siendo $a = 1 - 1/ F_1$, para cualquier $p \leq p_o$:

$$W_M = \frac{t_{n1}.(F_1 - t_{n1}.(1 - p)}{F_1 - p} = 1 - \frac{F_1 - p - F_1.t_{n1} + t_{n1}^2(1-p)}{F_1 - p}$$

Luego $W_M \leq a = 1 - 1/F_1$ es lo mismo que

$$\frac{F_1 - p - F_1.t_{n1} + t_{n1}^2(1-p)}{F_1 - p} \geq 1/F_1 \quad ==>$$

$$F_1^2 - p.F_1 - F_1^2.t_{n1} + t_{n1}^2(1-p).F_1 \geq F_1 - p \quad ==>$$

$$p \leq \frac{F_1^2 - F_1^2.t_{n1} + F_1.t_{n1}^2 - F_1}{F_1 + F_1.t_{n1}^2 - 1} =$$

$$\frac{F_1 - F_1.t_{n1} + t_{n1}^2 - 1}{1 + t_{n1}^2 - (1/F_1)} =$$

$$\frac{F_1(1 - t_{n1}) - (1 - t_{n1}^2)}{a + t_{n1}^2} =$$

$$\frac{t_{n2}\,(\,F_1 - 1 - t_{n1}\,)}{a + t_{n1}^2} = \frac{t_{n2}\,.\,F_1\,.b}{a + t_{n1}^2} = p_o$$

c.q.d.

Proposición 3.2.12 Si ponderamos el estimador $B{:}\bar{Y}_p$ con el estimador $D{:}\bar{Y}_{1-p}$ en una partición de dos subconjuntos, se verifica para una tasa de no respuesta dada, t_{n2} que :

1) Fijado un valor de p , el sesgo del estimador media ponderada crece al disminuir W

2) A igual valor de W , el sesgo del estimador media ponderada crece al disminuir p

3) Siendo W_M la razón de ponderaciones para un valor de p , en la que la varianza de la media ponderada toma el valor mínimo absoluto, una cota inferior de ésta es :

$$\frac{(1 - p).t_{n2}}{1 - p/F_1}\,.\bar{Y}_{N1}\,, \quad \text{siempre que} \quad 0 < p \le \min(1,p_o) \qquad (1)$$

Siendo $F_1=N_1/n_1$, y p_1 y p_o los valores indicados en la proposición 3.2.11

Demostración :

La diferencia entre los sesgos asociados al estimador $B{:}\bar{Y}_n$ y a la media ponderada $\bar{Y}_{k/n}$, es :

$$\text{sesgo}(B{:}\bar{Y}_n) - \text{sesgo}\,(\bar{Y}_{k/n}) = \frac{(1 - p).t_{n2}}{W.p + 1 - p}\,.\,\bar{Y}_{N1} \qquad (2)$$

Aplicando a (2) la proposición 3.1.6, deducimos el apartado 1) y 2).

El apartado 3), es consecuencia de la proposición 3.2.11 en cuanto a existencia del valor W_M.

La expresión (1) se obtiene al sustituir W por $a = 1 - 1/F_1$, dado que $W_M < a$

c.q.d.

3.3. Ponderación entre un estimador sesgado y otro insesgado

3.3.1. Estimador D+A

Proposición 3.2.13 Para una m.a.s. de extensión n, con no respuesta, dividida en dos subconjuntos aleatoriamente seleccionados, tales que p y 1-p son sus proporciones en el conjunto de unidades de la muestra, W_p y W_{1-p}, $D{:}\bar{Y}_p$ y $A{:}\bar{Y}_{1-p}$, las ponderaciones y las medias respectivas de cada uno de los subconjuntos y siendo $W = W_p / W_{1-p}$ la razón entre ambas ponderaciones, se tiene

1) La media ponderada de la muestra, utilizando los estimadores B y D, tiene por expresión :

$$D{+}A{:}\bar{Y}_{k/n} = \frac{W.p.(D{:}\bar{Y}_p) + (1-p).(A{:}\bar{Y}_{1-p})}{W.p + 1 - p}$$

que en lo que sigue, la expresaremos por $\bar{Y}_{k/n}$

2) La media ponderada es un estimador sesgado y es tal que :

a)
$$sesgo(\bar{Y}_{k/n}) = \frac{W.\ p}{W.p + 1 - p} \ .\ sesgo(D{:}\bar{Y}_n)$$

y

b)
$$sesgo(\bar{Y}_{k/n}) < sesgo(D{:}\bar{Y}_n)$$

3) La varianza de la media ponderada tiene por expresión :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ VAR(D{:}\bar{Y}_p) + (1-p)^2\ .\ VAR(A{:}\bar{Y}_{1-p})}{(W.p + 1 - p)^2}$$

y al ser n.p y n.(1-p) las unidades de la muestra empleadas en cada partición se tiene :

$$VAR\ (D{:}\bar{Y}_{1-p}) = \frac{n_1.(1-p)}{[n.(1-p)]^2} \ .\ \left[\ 1 - \frac{n_1.(1-p)}{N_1}\ \right] \ .\ S_{N1}{}^2$$

$$VAR\ (A{:}\bar{Y}_{1-p}) = \frac{1}{n.(1-p)} \cdot \left[\ 1 - \frac{n.(1-p)}{N}\ \right] \cdot S_N^2$$

Demostración :

La primera y tercera parte son consecuencia de la proposición 3.1.1. Para ésta última parte hay que tener en cuenta, además, las proposiciones 2.4.16 y 2.1.4 del capítulo II.

La segunda parte es consecuencia de la proposición 3.1.2, dado que :

$E(D{:}\bar{Y}_p) = t_{n1}.\bar{Y}_{N1}$ y $E(A{:}\bar{Y}_{1-p}) = \bar{Y}_N$ (proposiciones 2.4.15 y 2.1.2, capítulo II)

c.q.d.

Proposición 3.2.14 Si ponderamos el estimador $D{:}\bar{Y}_p$ con el estimador $A{:}\bar{Y}_{1-p}$ en una partición de dos subconjuntos, para una muestra en la que la tasa de respuesta sea mayor que $R^2.b/a$, existe una proporción p_o y p_1 tal que para cualquier proporción p, con la condición :

$$0 < p \leq \min(1,p_o)$$

1) Tiene asociado un intervalo de extremos :

0 y $a + [\ 1/t_{n1}.(a.t_{n1} - b.R^2)\ .(1/p - 1/p_o)\ ]^{1/2}$, siendo $p \geq \min(1,p_1)$

o bien

$a - [\ 1/t_{n1}.(a.t_{n1} - b.R^2)\ .(1/p - 1/p_o)\ ]^{1/2}$ y

$a + [\ 1/t_{n1}.(a.t_{n1} - b.R^2)\ .(1/p - 1/p_o)\ ]^{1/2}$

siendo $p < \min(1,p_1)$

dentro de los cuales, y para cualquier valor de la razón W, se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (D{:}\bar{Y}_n)$$

2) Dentro del intervalo [0 , a] existe un valor de W :

$$W_M = \frac{F - 1 + p}{t_{n1}.(F_1 - p.t_{n1})} \cdot \frac{F_1}{F} . R^2$$

para el que la varianza del estimador de la media ponderada toma el valor mínimo absoluto :

$$VAR(\bar{Y}_{k/n}) = t_{n1} \cdot \frac{F - 1 + p}{F.n.[1+p.(W_M-1)]} \cdot S_N^2$$

Siendo $F = N/n$, $F_1 = N_1/n$, $a = 1 - n_1/N_1$, $b = 1 - n/N$, $R^2 = S_N^2 / S_{N1}^2$

$$p_o = \frac{a.t_{n1} - b.R^2}{(1-b).R^2+(a-a^2).t_{n1}}, \quad p_1 = - \frac{a.t_{n1} - b.R^2}{a.\ t_{n1} + R^2.n/N}$$

Demostración :

Utilizando la proposición 3.1, teniendo en cuenta los estimadores que vamos a utilizar y las proposiciones 2.1.4 y 2.4.16 del capítulo II escribimos :

$$VAR(\bar{Y}_{k/n}) = \frac{W^2.\ p^2.\ VAR(D:\bar{Y}_p) + (1-p)^2\ .\ VAR(A:\bar{Y}_{1-p})}{(W.p + 1 - p)^2} \qquad (1)$$

$$VAR\ (D:\bar{Y}_p) = \frac{n_1.p}{(n.p)^2} \cdot \left[\ 1 - \frac{n_1.p}{N_1}\ \right] \cdot S_{N1}^2$$

$$VAR\ (A:\bar{Y}_{1-p}) = \frac{1}{n.(1-p)} \cdot \left[\ 1 - \frac{n.(1-p)}{N}\ \right] \cdot S_N^2 \qquad (2)$$

$$VAR\ (D:\bar{Y}_n) = \frac{n_1}{n^2} \cdot \left[\ 1 - \frac{n_1}{N_1}\ \right] \cdot S_{N1}^2$$

Pretendemos demostrar bajo qué condiciones se verifica que :

$$VAR\ (\bar{Y}_{k/n}) < VAR\ (D:\bar{Y}_n) \qquad (3)$$

Sustituyendo en la desigualdad (3), las expresiones (1) y (2) y dividiendo la desigualdad resultante por S_{N1}^2 :

$$W^2.p\ .\ \frac{n_1}{n^2}.\left(1 - \frac{n_1.p}{N_1}\right)+(1-p)/n\ .\left(1 - \frac{n.(1-p)}{N}\right).R^2 < (W.p+1-p)^2.\frac{n_1}{n^2}.\left(1- \frac{n_1}{N_1}\right)$$

agrupando tèrminos, se tiene :

$$W^2.p.(1-p).n_1/n^2 - 2.W.p.(1-p).n_1/n^2.(1-n_1/N_1)$$

$$+ (1-p).\ 1/n.\ [R^2.(1 - n.(1-p)/N) - (1-p).\ n_1/n\ .(1-n_1/N_1)] < 0$$

multiplicando por n y simplificando :

$$W^2.p.t_{n1} - 2.W.p.t_{n1}.(1-n_1/N_1) + R^2.(1 - n.(1-p)/N) - (1-p).t_{n1}.(1-n_1/N_1) < 0$$

denotando por $a = 1 - n_1/N_1$ y $b = 1 - n/N$, se tiene :

$$W^2.p.t_{n1} - 2.W.p.a.t_{n1} + R^2.(b + p.n/N) - (1-p).a.t_{n1} < 0 \qquad (1)$$

Como $p.t_{n1}>0$, la solución de la inecuación (1) se encuentra en el intervalo comprendido entre las soluciones de la ecuación :

$$W^2.p.t_{n1} - 2.W.p.a.t_{n1} + R^2.(b + p.n/N) - (1-p).a.t_{n1} = 0 \qquad (2)$$

El discriminante de ésta ecuación, despues de agrupar, es :

$$4.p.t_{n1}\ [\ -p.(\ R^2.n/N + t_{n1}.(a - a^2)\ ) + a.t_{n1} - R^2.b\ ] = 4.p.t_{n1}.D$$

denotando por $D = -p.(\ R^2.n/N + t_{n1}.(a - a^2)\ ) + a.t_{n1} - R^2.b$

La ecuación (2) tiene solución, siempre que se verifique : $D \geq 0$; esto es, para valores de p tales que :

$$p \leq \frac{a.t_{n1} - R^2.b}{R^2.n/N + t_{n1}.(a - a^2)} = p_o$$

El valor p_o es > 0, siempre que el término : $a.t_{n1} - R^2.b > 0$. dado que todos los tèrminos del denominador son positivos, al ser $a < 1$.

Es decir, para tasas de respuesta, tales que :

$$t_{n1} > R^2.b/a$$

El intervalo, solución de la inecuación (1), para valores $p \leq p_o$ es :

$$[\ a - (D/(p.t_{n1}))^{1/2}\ ,\ a + (D/(p.t_{n1}))^{1/2}\ ]$$

que de acuerdo con el valor obtenido para p_o, toma la expresión :

$$a \pm [\ 1/t_{n1}.(a.t_{n1} - b.R^2)\ .(1/p - 1/p_o)\ ]^{1/2}$$

Por otro lado, observamos que para $p = p_o$, el intervalo se reduce al punto $W = a$.

Además, el valor de p a partir del cual el extremo inferior es nulo o positivo , es :

$$a - [\ 1/t_{n1}.(a.t_{n1} - b.R^2)\ .(1/p - 1/p_o)\ ]^{1/2} \geq 0$$

Denotando por $A = a.t_{n1} - b.R^2$, se sigue que :

$$1/p \leq 1/p_o + a^2.t_{n1}/A = \frac{R^2.n/N + t_{n1}.(a - a^2) + t_{n1}.\ a^2}{a.t_{n1} - b.R^2}$$

$$\text{esto es para } p \geq \frac{a.t_{n1} - b.R^2}{R^2.n/N + t_{n1}.a} = p_1$$

dado que p_o y p_1 pueden ser mayor que 1, restringimos todas las soluciones a valores de p, tales que : $0 < p \leq 1$

2) Por la proposición 3.1.7 y para cualquier valor de p, existe un valor para W, para el cual la varianza de la media ponderada tiene un valor mínimo. Para el caso en el que estamos, estos valores son :

$$W_M = \frac{1-p}{p} \cdot \frac{VAR(A:\bar{Y}_{1-p})}{VAR(D:\bar{Y}_p)} \qquad (4)$$

$$VAR(\bar{Y}_{k/n}) = VAR(A:\bar{Y}_{1-p}) \cdot \frac{1 - p}{1 + p.(W_M-1)}$$

Teniendo en cuenta las expresiónes (2) , sustituyendo en (4) y simplificando, se tienen las expresiones propuestas al principio de la proposición.

Demostramos que W_M está contenido en el intervalo [0 , a] subintervalo del obtenido en el apartado anterior de ésta proposición; para ello :

* Veamos que $W_M \geq 0$ para cualquier $p > 0$:

Como $F_1 = N_1/n > 1$ y $p.t_{n1} < F_1$ ==> $F_1 - p.t_{n1} > 0$, y por consiguiente para que :

$$W_M = \frac{F_1 - 1 + p}{t_{n1}.(F_1 - p.t_{n1})} \cdot \frac{F_1}{F} \cdot R^2 \geq 0 \quad ==> F_1 \geq 1 - p \quad ==>$$

$$p \geq 1 - F_1$$

y al ser $1 - F_1 < 0$, se sigue que para todos los $p > 0$ es válida la desigualdad.

* Veamos que $W_M \leq a$, siendo $a = 1-n_1/N_1 = 1-n_1/n \cdot n/N_1 = 1 - t_{n1}/F_1$, para cualquier $p \leq p_o$:

$$W_M = \frac{F_1 - 1 + p}{t_{n1}.(F_1 - p.t_{n1})} \cdot \frac{F_1}{F} \cdot R^2 =$$

$$= 1 - \frac{F.F_1.t_{n1} - F.p.t_{n1}^2 - R^2.F_1.F + F_1.R^2 - F_1.R^2.p}{F.t_{n1}.(F_1 - p.t_{n1})}$$

Luego $W_M \leq a = 1 - t_{n1}/F_1$ es lo mismo que

$$\frac{F.F_1.t_{n1} - F.p.t_{n1}^2 - R^2.F_1.F + F_1.R^2 - F_1.R^2.p}{F.t_{n1}.(F_1 - p.t_{n1})} \geq t_{n1}/F_1 \quad ==>$$

$$F.F_1^2.t_{n1} - F.p.F_1.t_{n1}^2 - R^2.F_1^2.F + F_1^2.R^2 - F_1^2.R^2.p \geq$$

$$F.F_1.t_{n1}^2 - F.p.t_{n1}^3$$

y de ésta desigualdad, dividiendo por $F.F_1^2$ y teniendo en cuenta que $b = 1 - n/N = 1 - 1/F$, obtenemos :

$$p \leq \frac{t_{n1} - R^2 + R^2/F - t_{n1}^2/F}{t_{n1}^2/F + R^2/F - t_{n1}^3/F_1} = \frac{t_{n1}\,(\,1 - t_{n1}/F\,) - R^2\,(\,1 - 1/F)}{a.\,t_{n1}/F.\,t_{n1} + R^2/F} =$$

$$\frac{a.t_{n1} - R^2.b}{(1-b).R^2 + a(1-a).t_{n1}} = p_o \text{ , dado que } 1 - b = 1/F \text{ y } 1 - a = t_{n1}/F$$

c.q.d.

Proposición 3.2.15 Si ponderamos el estimador $D{:}\bar{Y}_p$ con el estimador $A{:}\bar{Y}_{1-p}$ en una partición de dos subconjuntos, la razón entre los sesgos asociados a la media ponderada $\bar{Y}_{k/n}$ y al estimador $D{:}\bar{Y}_n$ es :

$$\text{sesgo}(\bar{Y}_{k/n}) \;/\; \text{sesgo}\,(D{:}\bar{Y}_n) = \frac{W.\,p}{W.p + 1 - p}$$

y se verifica :

1) Fijado un valor de p, la razón crece al crecer W

2) A igual valor de W, la razón crece al crecer p

3) Para una muestra en la que la tasa de respuesta sea mayor que $R^2.b/a$ y siendo W_M la razón de ponderaciones, para un valor de p , en la que la varianza de la media ponderada toma el valor mínimo absoluto, una cota superior de ésta es :

$$\frac{p - p/F_1}{1 - p/F_1} \quad \text{siempre que} \quad 0 < p \leq \min(1, p_o) \qquad (1)$$

Siendo $F_1=N_1/n_1$, y a, b, R^2, p_1 y p_o los valores indicados en la proposición 3.2.14

Demostración :

La razón entre sesgos es consecuencia de la proposición 3.1.2

El apartado 1) y 2) es consecuencia de la proposición 3.1.4

El apartado 3) es consecuencia de la proposición 3.2.14 en cuanto a existencia del valor W_M y de ser $W_M < a$

Por otro lado, como la razón entre sesgos es una función creciente para W (proposición 3.1.4) , se sigue :

$$[\text{sesgo}(\bar{Y}_{k/n})/\text{sesgo}\ (D:\bar{Y}_n)]_{(W=W_M)} < [\text{sesgo}(\bar{Y}_{k/n})/\text{sesgo}\ (D:\bar{Y}_n)]_{(W=a)}$$

Tomando la parte de la derecha de la desigualdad el valor indicado en (1)

c.q.d.

BIBLIOGRAFIA Y SOFTWARE

BIBLIOGRAFIA Y SOFTWARE

Los títulos de los artículos y los libros constituidos por capítulos y escritos por diferentes autores, los hemos indicado entre comillas. Para los segundos, se añade el título del libro dentro del cual está contenido.

Para las revistas se han utilizado las siguientes abreviaturas, cuando su título es muy extenso :

AER : Agricultural Economics Research
AS : The American Statistician
BISI : Bulletin of the International Statistical Institute
EE : Estadística Española
EUSTAT : Instituto Vasco de Estadística
ISR : International Statistical Review
JASA : Journal of the American Statistical Association
JRSS : Journal of the Royal Statistical Society
JSPI : Journal of Statistical Planning and Inference
QCAS : Quality Control and Applied Statistics
REIS : Revista de Estudios Sociales
SMR : Sociological Methodes and Research
SSM : Survey Samples and Measurement
TEIO : Trabajos de Estadística e Investigación Operación

El fondo del material bibliográfico que señalo y utilizado en la elaboración de ésta Tesis, ha sido

Para los libros

Departamento de Matemática Aplicada, Estadística, Econometría e Investigación Operativa y Facultad de Ciencias Empresariales (ETEA). Universidad de Córdoba.

Para las revistas

Unidad de Economía Agraria, Departamento de Matemática Aplicada, Estadística, Econometría e Investigación Operativa y Facultad de Ciencias Empresariales (ETEA). Universidad de Córdoba

Centro de Información y Documentación Científica (C.I.N.D.O.C.), peteneciente al Consejo Superior de Investigaciones Científicas. Madrid

Instituto de Estadística de Andalucía. Junta de Andalucía. Sevilla

En lo que sigue, señalamos el material bibliográfico antes aludido y a continuación el Software.

BIBLIOGRAFIA

American Statistical Asociation (1990) Proceding of the Section on Survey Research Method. Edit. ASA

Aparicio, P. (1988) "Estimación de los errores muestrales mediante el método de los conglomerados últimos". Edit. REIS, nº44, 145-164

Azorín F. y Sanchez-Crespo, J.M. (1986), Métodos y aplicaciones del muestreo. Edit. Alianza Universidad Textos

Azorín F. (1984), Aspectos y aplicaciones en el muestreo. Edit. EUSTAT. Vitoria

Bailar, B.A., Bailey L. y Corby, C. (1978), "A comparison of some adjustment and weighting procedures for survey data". Edit. SSM,----, 175-198

Basaulto, J. y Murgui, S. (1982), "Diseño muestral óptimo en el caso de no respuesta". Edit. TEIO, Vol. 33, nº 2, 3-15

Bickel, P.J. (1993), "Nonparametric inference under biased sampling from a finite population". Edit. QCAS, nº2, Mar-Abr

Bouza, C.N. (1987), "Evaluación de reglas de submuestreo para la estimación de una diferencia en el caso de la no respuesta". Edit TEIO, Vol. 2, nº 2, 33-40

Cansado, E. (1983), Muestreo y aplicaciones. Edit. EUSTAT. Vitoria

Caridad, J.M. (1985), Diseños muestrales en poblaciones finitas. Edit. Servicio de Publicaciones de la Universidad de Córdoba.

Cochran, W.G. (1977), Sampling Techniques. Edit. John Wiley & Sons. New York

Cramer, H. (1960), Métodos Matemáticos de Estadística. Edit. Aguilar. Madrid.

Cruz, P. (1990), "Del no sabe al no contesta : Un lugar de encuentro para diversas respuestas ". Edit REIS, nº 52, 139-156

Chao, A y Shen-Ming, L. (1993), "Estimating the number of clases via sample coverage". Edit QCAS, Vol. 38, nº 1, Enero-Febrero

Chaudhuri, A. (1988), Unified Theory and Strategies of Survey Sampling. Edit. North-Holland. Amsterdan

Chaudhuri, A. (1988), "Optimality of Sampling Strategies". Handbook of statistics, vol. 6, 427-447. Edit.North-Holland. Amsterdan

Chevry, G. (1967), Práctica de las encuestas estadísticas. Edit. Ariel

Deming, W. (1953), "On a probability mechanism to attain an economic balance between the result and error of non-response and the bias of non-response". Edit. JASA, nº48, 743-772

Deming, W. (1960), Sample Desing in Business Research. Edit. John Wiley & Sons. New York

Durbin, J. (1954), "Nonresponse and call-backs in surveys". Edit. BISI, nº 34/2, 72-86

Emrich, L. (1983), "Randomized response techniques". Incomplete data in sample surveys, Vol. 2, 73-79. Edit. Academic Press

Fellegi, I y Oldt, D. (1980), "Un enfoque sistemático de la edición e imputación automáticas". Edit. EE, nº88, 33-80

Ford, B.L. (1983), "An overview of Hot-deck procedures". Incomplete data in sample surveys, Vol. 2, 185-206. Edit. Academic Press

Frankel, L.R. y Dutka, S. (1983), "Survey desing in anticipation of nonresponse and imputation". Incomplete data in sample surveys, Vol. 3, 69-83. Edit. Academic Press

Gad, N. (1990), Evaluation of questionnaire desing effects. Edit. EUSTAT. Vitoria

Gourieroux, C. (1981), Theorie des Sondages. Edit. Económica. París

Granquist, L. (1991), Macro-Editing Methods for rationalizing the editing of quantitative data. Edit. EUSTAT. Vitoria

Grosbras, J.M. (1987), Methodes Statistiques des Sondages. Edit. Económica. París

Hahn, G.J. , Meeker, W.Q. (1991), Statistical Intervals (A guide for practitioners). Edit. John Wiley & Sons. New York

Hansen, M.H., Hurwitz, W.N. (1946), "The problem of nonresponse in sample surveys". Edit. JASA, nº41, 517-529

Hansen, M.H., Hurwitz, W.N. y Madow, W.G. (1953), Sample Survey Methods and Theory (Vol I y II). Edit. John Wiley & Sons. New York

Hansen, M.H., Hurwitz, W.N. y Madow, W.G. (1961), "Measurement errors in censuses and surveys". Edit. BISI, 38/2, 359-374

Hawkins, D.F. (1975), "Estimation of nonresponse bias". Edit. SMR, nº3, 461-488

Herzog, T.N. y Rubin, D.B., (1983), "Using Multiple imputations to handle nonresponse". Incomplete data in sample surveys, Vol. 2, 210-245. Edit. Academic Press

Horvitz, D.G. y Thompson, D.J. (1952), "A generalization of sampling whithout replacement from a finite universe". Edit. JASA, nº 47, 663-685

Houseman, E.E. (1953), "Estatistical treatment of the nonresponse problem". Edit. AER, nº 5, 12-18

King, B.F. (1983), "Quota sampling". Incomplete data in sample surveys, Vol. 2, 63-71. Edit. Academic Press

Kish, L. y Hess, I. (1959) "A replacement procedure for reducing the bias of nonresponse". Edit. AS, nº 13/4, 17-19

Kish, L. (1965), Survey Sampling. Edit John Wiley & Sons. New York

Kish, L. (1986), Operaciones estadísticas por muestreo . Edit. EUSTAT. Vitoria.

Kun, H. (1993), "The estimations of stratun means vector with randon sample sizes". Edit. JSPI, nº 37, 43-49

Ladoux, M. (1982), Modelos de respuesta aleatorizada. Edit. Instituto Nacional de Estadística

Levy, P.S. (1991), Sampling of Populations : Methods and Applications. Edit. John Wiley & Sons. New York

Lininger, L. (1984), La encuesta por muestreo : Teoría y práctica. Edit. CECSA

Madow, W.G., Nisselson, H. y Olkin, I. (1983), "Problems of incomplete data". Incomplete data in sample surveys, Vol. 1, 15-28. Edit. Academic Press

Martinez A, Rodriguez C. y Gutierrez R. (1993), Inferencia Estadística: Un enfoque clásico. Edit. Pirámide

Martón, A. (1988), Sampling and non-sampling errors in surveys. Edit. EUSTAT. Vitoria

Miras, J. (1976), "Estimación de errores de muestreo. Método de los grupos aleatorios. Edit. EE, Julio-Diciembre

Murthy, M.N. y Sethi, V.K. (1961), "Randomized rounded-off multipliers in sampling theory". Edit. JASA, nº 56, 328-334

Oh, H.L. y Scheuren, F.J. (1983), "Weighting Adjustment for unit nonresponse". Incomplete data in sample surveys, Vol. 2, 143-183. Edit. Academic Press

Passeron, J.C. (1982) "Los silencios : Contribución a la interpretación de las no-respuestas en las encuestas de opinión". Edit. REIS, nº 17, 83-

Pfeffermann, D. (1993), "The role of sampling weights when Modeling Survey y Data". Edit. ISR, nº 61/2, 317-337

Plateck, R. y Gray, G.B. (1983), "Imputation Methodology". Incomplete data in sample surveys, Vol. 2, 255-294. Edit. Academic Press

Plateck, R. (1986), Metodología y tratamiento de la no-respuesta. Edit. EUSTAT. Vitoria

Politz, A.N. y Simmons, W.R. (1949), "An attempt to get the 'not at homes' into the sample without call-backs". Edit. JASA, nº 44, 9-31

Rao, C.R. (1983), Inferencia estadística lineal. Edit. EUSTAT. Vitoria

Rao, J.N.K. (1988), "Variance estimation in sample surveys". Handbook of statistics, vol. 6, 427-447. Edit.North-Holland. Amsterdan

Rao, J.N.K. (1973), "On double sampling for stratification and analytical surveys". Edit. BiométriKa, nº 60, 125-133

Rao, J.N.K. (1983), "Comparison of Domains in the Presence of nonresponse". Incomplete data in sample surveys, Vol. 3, 215-226. Edit. Academic Press

Rao, P.S.R.S. (1983), "Callbacks, Follow-Up, and Repeated Telephone Calls". Incomplete data in sample surveys, Vol. 2, 33-43. Edit. Academic Press

Rao, P.S.R.S. (1983), "Randomization Approach". Incomplete data in sample surveys, Vol. 2, 97-105. Edit. Academic Press

Raj, D. (1958), "On the relative accuracy of some sampling techniques". Edit. JASA, nº 53, 98-101

Raj, D. (1968), Sampling Theory. Edit. Mc.Graw-Hill. New York

Raj, D. (1972), The design of sample Surveys. Edit. Mc.Graw-Hill. New York

Ríos, S. (1967), Métodos Estadísticos. Edit. Ediciones del Castillo. Madrid

Rodriguez, J., Ferreres M.L. y Nuñez, A. (1991), "Inferencia Estadística, niveles de precisión y diseño muestral". Edit REIS, nº 54, 139-162

Rodriguez-Pongo, P. y Villar,I. (1985), "Propuesta de una regla para simplificar el proceso de obtención del conjunto completo en la metodología de Fellegi y Holt". Edit EE, nº 106, 27-44

Rubin, D.B. (1986), Multiple imputation for Nonresponse in Surveys. Edit. John Wiley & Sons. New York

Ruiz, M. (1986), "Sesgo de no respuesta en el intento n". Edit. EE, nº 112-113, 75-78

Ruiz, M. (1988), "Estimación insesgada con observaciones erradas y no respuesta". Edit. TEIO, Vol. 3, nº 1, 71-80

Ruiz, M. y Santos, J. (1989), "Estrategias intermedias de muestreo". Edit. EE, Vol. 31, nº 121, 227-235

Sanchez, M. y López L. (1985), "Estimación de registros desconocidos en series de datos". Edit. TEIO, Vol.36, nº3, 259-268

Sanchez-Crespo, J.L. (1976), Muestreo de Poblaciones finitas aplicado al diseño de encuestas. Edit. Instituto Nacional de Estadística. Madrid.

Sanchez-Crespo, J.L. y Gabeiras J.M. (1987), "Un esquema mixto de muestreo con probabilidades desiguales". Edit. EE, nº 115, 5-39

Särndad, C. (1992) Model Assisted Survey Sampling. Edit. Springer-Verlag. New York

Sande, I.G. (1983), "Hot-Deck imputation Procedures". Incomplete data in sample surveys, Vol. 3, 339-349. Edit. Academic Press

Sirken, M. (1983), "Handling Missing Data by Network Sampling". Incomplete data in sample surveys, Vol. 2, 81-89. Edit. Academic Press

Thomsen, I. y Siring, E. (1983), "On the causes and effects nonresponse : Norwegian Experiences". Incomplete data in sample surveys, Vol. 3, 25-59. Edit. Academic Press

Valliant, R. (1993), "Poststratification and Conditional Variance Estimation. Edit. JASA, Vol.88, nº 421

Veres, E. (1988) "Algoritmo para la partición aleatoria de una población finita, en subconjuntos de igual tamaño y de elementos consecutivos". Edit EE, Vol. 30, nº 118, 233-251

Villán, I. y Bravo, M.S. (1990), Procedimientos de depuración de datos estadísticos. Edit. EUSTAT. Vitoria.

Villar, I. (1992), "Análisis de reglas de depuración de datos". Edit. EE, Vol.34, nº 129, 151-171

Warner, S.L. (1965), "Randomized response : A survey technique for eliminating evasive answer bias". Edit. JASA, nº 60, 63-69

Woodruff, R.S. (1952), "Confidence intervals for medians and other position measures". Edit. JASA, nº 47, 635-646

Yañez de Diego, I. (1989), Teoría de muestras. Universidad de Educación a Distancia. Madrid

Yates, F. y Grundy, P.M. (1953) , "Selection without replacement from within strata with probability proportional to size". Edit. JRSS, nº 15, 235-261

ZarKovich, S.S. (1967) Los métodos de muestreo y los censos. Edit. Organización Naciones Unidades (FAO)

SOFTWARE

Para la elaboración de los gráficos hemos utilizado STATGRAPHICS V.7, instalado en la Unidad de Economia Agraria (Universidad de Córdoba).

Para la elaboración de las tablas y demás cálculos hemos utilizado SAS, instalado en el Centro de Cálculo (Universidad de Córdoba); habiendo conectado a través del Centro de Cálculo Científico ubicado en la Facultad de Veterinaria de la referida Universidad.