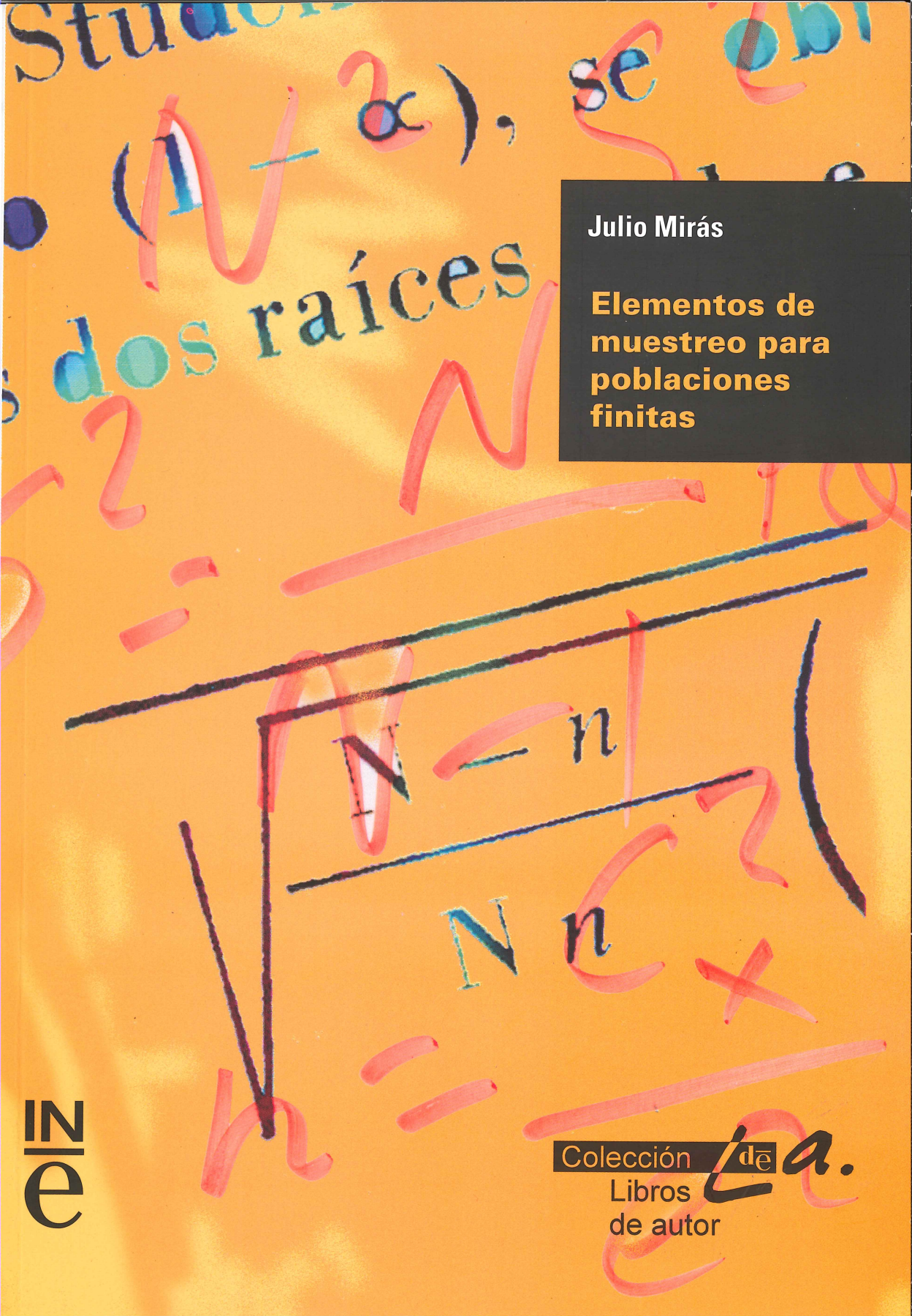


Julio Mirás

Elementos de muestreo para poblaciones finitas

Colección de la. Libros de autor

INE

Julio Mirás

Elementos de muestreo para poblaciones finitas

INE
INSTITUTO NACIONAL DE ESTADISTICA

Colección de Libros de autor

Madrid, 2000

Nota:
La reimpresión de esta obra se ha realizado utilizando el material original publicado en su primera edición.

Ficha editorial

Título: Elementos de muestreo para poblaciones finitas

Número INE: 91.98-85-501 (0.00)

Depósito Legal: M-25519-1985

ISBN: 84-260-1198-5

Tarifa: 7

INE ARTES GRÁFICAS
Paseo de la Castellana, 183. Madrid

PRESENTACION

Me es grato presentar ELEMENTOS DE MUESTREO PARA POBLACIONES FINITAS de cuyo autor, el estadístico facultativo D. Julio Mirás, me consta su sólida formación teórica y amplia experiencia en la elaboración de estadísticas. Estoy convencido de que su libro será de gran utilidad, no sólo para los estadísticos que aspiran a ingresar como funcionarios del Instituto sino también para todos los que se inicien en la teoría y técnica del muestreo.

De otra parte, con esta edición el Instituto Nacional de Estadística continúa su tarea de ofrecer a los estudiosos y al público interesado, no solamente las estadísticas que elabora sino también monografías acerca de las técnicas y métodos más importantes.

Deseo que este libro cumpla sus objetivos, contribuyendo a mantener y aumentar el nivel alcanzado por la estadística española en su ya larga trayectoria, y agradezco de antemano la acogida de todos ustedes.

LUIS RUIZ-MAYA PEREZ
Director General

A Mercedes

INDICE

Páginas

INTRODUCCION XIII

I. MUESTREO Y ESTIMACION EN POBLACIONES FINITAS. CONCEPTOS GENERALES

I.1. Concepto de procedimiento de muestreo aleatorio 1
I.2. Concepto de procedimiento de estimación 2
I.3. Muestreo con y sin reposición 6
I.4. Distribuciones marginales de las componentes de la muestra 6
I.5. Muestras no ordenadas. Vector aleatorio $\varepsilon(s)$ 8
I.6. Selección de muestras con reposición. Modelo polinomial 9
I.7. Selección sin reposición 12
Algunos procedimientos de selección sin reposición y probabilidades desiguales 16
I.8. Formación de estimadores 23
I.9. Propiedades de los estimadores 26
Intervalo de confianza cuando el estimador es insesgado. 29
Intervalo de confianza cuando el estimador es sesgado .. 31
Hipótesis de normalidad 33
I.10. Ejemplos 35

Páginas

II. MUESTREO ALEATORIO CON PROBABILIDADES IGUALES

II.1. Estimadores de la media y el total 47
II.2. Eficiencia del muestreo sin reposición 51
II.3 Estimación de las varianzas 51
II.4. Estimación de la proporción y el total de clase. Varianza de los estimadores. Estimación de las varianzas 53
II.5. Intervalos de confianza para medias y totales 55
II.6 Tamaño de la muestra para una precisión dada 56
II.7 Esperanza y varianza de s 59
II.8. Estimador de la razón. Sesgo 62
II.9. Estimador de la razón. Varianza 68
Estimación de la varianza de $\hat{R}$ 70
Intervalos de confianza para $\hat{R}$ 71
II.10. Estimadores de razón para la media y el total 72
Condiciones en que es más conveniente el uso de estimadores de razón 73
II.11. Estimación de medias y totales en subpoblaciones 75
II.12. Estimación de la diferencia de medias 82
II.13. Ejemplos 87

III. MUESTREO UNIETAPICO DE CONGLOMERADOS

III.1. Concepto y notaciones 97
III.2 Medida de la homogeneidad de los conglomerados 98
III.3 Estimación de la media y el total 102
III.4. Estimación de la proporción y el total de clase 103
III.5 Eficiencia del muestreo por conglomerados 104
III.6. Estimación de las varianzas, de la eficiencia y del coeficiente δ 106
III.7 Tamaño óptimo de la muestra y de los conglomerados .. 111
III.8. Muestreo con reposición de conglomerados con probabilidades desiguales 117
Probabilidades proporcionales al tamaño 121
Casos particulares 123

III.9. Muestreo sin reposición de conglomerados con probabilidades desiguales. 124
Estimación insesgada de la varianza. 127
III.10. Ejemplos 130

IV. MUESTREO ESTRATIFICADO

IV.1. Estratificación de la población 141
IV.2. Estimación de la media, el total, la proporción y el total de clase 144
IV.3. Tamaño de la muestra. Afijación óptima 147
IV.4. Eficiencia del muestreo estratificado 156
IV.5. Estimación insesgada de σ^2 y S^2 a partir de una muestra estratificada 159
IV.6. Estimadores de razón en muestreo estratificado 163
El estimador separado 164
El estimador combinado 166
IV.7. Unidades que entran con certeza en la muestra 169
IV.8. Ejemplos 174

V. MUESTREO BIETAPICO

V.1. Cálculo de esperanzas y varianzas 187
V.2. Formación de estimadores lineales insesgados. Muestreo con reposición 190
V.3. Formación de estimadores lineales insesgados. Muestreo sin reposición 192
V.4. Muestras autoponderadas 193
V.5. Varianza de estimadores lineales en muestreo bietápico . 196
V.5.1. Muestreo con reposición 196
Casos particulares 198
V.5.2. Muestreo sin reposición 199
V.6. Método de los conglomerados últimos para la estimación de varianzas 201
V.7. Procedimiento general para el cálculo de varianzas. Teorema I de Durbin 204

Páginas

V.8. Procedimiento general para la estimación de varianzas en muestreo sin reposición. Teorema II de Durbin 208
V.9. Tamaño óptimo de una muestra bietápica 211
V.10. Ejemplos 215

VI. METODOS ESPECIALES PARA LA ESTIMACION DE VARIANZAS

VI.1. Método de los grupos aleatorios 225
Grupos aleatorios de distinto tamaño 228
Precisión del método 229
VI.2. Método de los conglomerados últimos 231
VI.3. Método de semimuestras reiteradas 232
Precisión del método 233
Forma de aumentar la precisión 234
VI.4. Aplicación al muestreo estratificado 237
Método de los conglomerados últimos 238
Método de semimuestras reiteradas 239
VI.5. Diseños con dos unidades muestrales por estrato 240
Técnica por separado 240
Técnica combinada 241
Semimuestras equilibradas 241
VI.6. Ejemplos 243

VII. ALGUNAS TECNICAS UTILES

VII.1. Muestreo sistemático 247
VII.1.1. Análisis del muestreo sistemático. Estimación de la media 249
Estimación de la varianza 252
VII.2. Muestreo en dos ocasiones 253
VII.2.1. Estimación del cambio 254
Ganancia en precisión con el solapamiento 255
Estimador lineal de mínima varianza 255
VII.2.2. Estimación de la segunda ocasión 258

Páginas

VII.3. Muestreo en dos fases 260
VII.3.1. Consideración del coste 260
VII.3.2. Muestreo doble para estratificación 262
Afijación proporcional 267
Afijación y tamaños de muestra óptimos 268
Estimación de la varianza 271
VII.3.3. Muestreo doble para estimadores de razón 274
Tamaños óptimos de muestra 277
Estimación de la varianza 278
VII.3.4. Muestreo doble para estimadores de regresión lineal ... 279
Estimador óptimo 282
Estimador por diferencia 283
VII.3.5. Muestreo doble para probabilidades proporcionales al tamaño 283
VII.4. Ejemplos 286

VIII. ERRORES AJENOS AL MUESTREO. DEFECTOS DEL MARCO.

VIII.1. El problema de las unidades vacías o extrañas 293
VIII.1.1. Notaciones y relaciones básicas 295
VIII.1.2. Estimación del total cuando se muestrea en el marco depurado 297
VIII.1.3. Estimación del total cuando se muestrea en el marco no depurado 297
VIII.1.4. Resumen y conclusiones 303
VIII.2. El problema de las unidades repetidas 305
VIII.2.1. Planteamiento formal del problema 305
VIII.2.2. Muestreo en el marco no depurado para estimación del total 307
VIII.2.3. Resumen y conclusiones 311
VIII.2.4. Estimación del número de unidades distintas en el marco 312
VIII.2.5. Estimación de la media en un marco con unidades repetidas 313
VIII.3. Ejemplos 315

Páginas

IX. ERRORES AJENOS AL MUESTREO. NO RESPUESTA. EFECTO DEL ENTREVISTADOR.

IX.1. El problema de la no respuesta 319
IX.I.1. Efectos de la no respuesta 321
IX.1.2. La sustitución de unidades que faltan 323
IX.1.3. Algunas técnicas para la no respuesta 324
Método de Hansen y Hurwitz 324
Modelo de Deming 329
Método de Politz y Simmons 333
IX.1.4. No respuesta parcial 334
IX.1.5. Modelos de respuesta aleatorizada 336
Modelo de Warner 336
Modelo con preguntas no relacionadas 338
Modelo para variables contínuas 340
IX.2. Efecto del entrevistador 341
IX.2.1. Modelo para el efecto del entrevistador 342
IX.2.2. Estimación y contraste del efecto entrevistador 345

X. MODELOS DEL ERROR TOTAL EN CENSOS Y ENCUESTAS

X.1. Fundamento de los modelos 350
X.2. El error total en un censo 351
X.3. El error total en una encuesta 353
X.4. Aplicación del modelo a una proporción 356
X.5. Experimentación y estimación 357
Estimación de la varianza de respuesta simple 358
Estimación de la varianza total de respuesta 358
X.6. Ejemplos 361

BIBLIOGRAFIA 363

INTRODUCCION

La toma de decisiones en el ámbito de las actividades económicas y sociales, tanto del sector privado como del sector público, así como en la investigación científica de las mismas, se viene haciendo a partir de informaciones de índole cuantitativa acerca de campos de estudio cada vez más variados y con mayor intensidad por lo que se refiere a las variables en cuestión y a las relaciones entre ellas. Las consecuencias de tales decisiones implican generalmente beneficios o costes económicos, o sociales; en consecuencia, la demanda de datos suficientemente precisos y oportunos para en primer lugar, poder tomar alguna decisión racional y en segundo lugar, para que tal decisión no resulte desacertada, ha venido incrementándose continuamente.

A menudo nos encontramos con necesidades de datos acerca de colectivos finitos como son por ejemplo los habitantes, las familias, las unidades productivas, pero cuyo número de unidades es muy elevado. La realización de investigaciones censales implica fuertes inversiones y requiere plazos largos de tiempo para su preparación y ejecución, por lo que su práctica, reducida a unas pocas características estructurales, ha quedado a cargo de las Administraciones públicas con periodicidad decenal o quinquenal.

Las técnicas de muestreo para poblaciones finitas se han desarrollado para satisfacer estas demandas, permitiendo obtener información de forma más rápida y más barata, así como, mediante encuestas sucesivas, suplir la falta de información que se tendría en los períodos intercensales. Por estas razones se incluye algún curso de muestreo en las

especialidades de escuelas y facultades dedicadas a ciencias aplicadas y se exige su conocimiento a quienes aspiran a ingresar en los cuerpos especiales del Instituto Nacional de Estadística. A ellos se dirige este libro que pretende servir como instrumento de iniciación; no se trata pues de un manual de recetas ni presenta técnicas excesivamente sofisticadas o nuevos campos de investigación teórica. No obstante he procurado mostrar con cierto rigor, de forma accesible a los no especialistas que dispongan de algunos conocimientos previos acerca de cálculo de probabilidades y estimación, los métodos básicos para el muestreo de poblaciones finitas.

La estadística matemática ordinaria cuando trata el problema de estimación supone muestreo aleatorio simple de poblaciones infinitas, en consecuencia la distribución de una muestra de tamaño n es igual al producto de n funciones de distribución iguales a la población base; sin embargo en poblaciones finitas se practican técnicas como p.e. la selección sin reposición y la estratificación, bajo las cuales no se cumple la condición de muestreo aleatorio simple (en el sentido de que cualquier muestra es posible y está constituida por n variables aleatorias independientes e idénticamente distribuidas). Con esto se pretende entre otros objetivos, reducir o restringir el conjunto de posibles muestras de modo que las consideradas menos deseables (menos representativas) no tengan posibilidad de ser seleccionadas, o tengan menor probabilidad. Por otra parte, el tipo de problemas que nos encontramos en poblaciones finitas no se plantea bajo la forma de estimación de un parámetro θ que determina una única distribución dentro de una familia $F(x\ ;\theta)$ conocida; lo usual es no hacer hipótesis acerca de la distribución, en este sentido paramétrico, sino que se desea la estimación de características descriptivas (como la media, total, varianza, etc.) o la distibución de frecuencias correspondiente a una partición en intervalos, o clases disjuntas, del campo de valores de la variable.

Estas peculiaridades han dado lugar al desarrollo de una disciplina estadística dedicada específicamente al diseño de procedimientos de muestreo y estimación para poblaciones finitas; no se trata de una teoría de la estimación diferente sino que contempla la finitud del colectivo en estudio, la posibilidad de utilizar distintos procedimientos de selección de la muestra y la ausencia de hipótesis paramétricas acerca de la distribución base.

Es obvio que como todo instrumento, la estimación mediante muestras tiene sus limitaciones. El hecho de que muestras diferentes obtenidas por el mismo procedimiento, den lugar a estimaciones diferentes cada una con una determinada probabilidad, nos conduce a un cierto grado de incertidumbre que trata de reducirse mediante diseños óptimos y de cuantificarse con el conocimiento de la distribución del estimador, o al menos de su valor esperado y varianza. Por otra parte una muestra diseñada para dar estimaciones de nivel nacional, puede no ser aceptable para proporcionar estimaciones a nivel territorial inferior; otra situación se presenta cuando se desea estimar características relativas a un colectivo que es una pequeña subpoblación ya que el número de unidades muestrales puede ser tan pequeño que no permita estimaciones fiables.

Para terminar esta breve introducción, deseo expresar mis votos en favor de quienes se inicien en la técnica del muestreo, muchas veces considerada árida y tediosa por algunos estudiantes. Mi agradecimiento al profesor D. Luis Ruiz-Maya, actualmente director del Instituto, sin cuyo estímulo personal y benévola acogida a mi trabajo, este libro permanecería inédito.

Julio Mirás
Madrid, 1985

I. MUESTREO Y ESTIMACION EN POBLACIONES FINITAS, CONCEPTOS GENERALES

I.1. CONCEPTO DE PROCEDIMIENTO DE MUESTREO ALEATORIO

Designamos por $\mathcal{A} = \{A_1, A_2, ..., A_i, ..., A_N\}$ al conjunto de N unidades que constituyen la *población finita* objeto de estudio; cada entero i, de 1 a N, identifica a una única unidad de la población.

Una sucesión ordenada de n elementos de $\mathcal{A}$:

$$s_j = \{A_{j1}, A_{j2}, ..., A_{jr}, ..., A_{jn}\}$$

donde A_{jr} denota la unidad que ocupa el lugar r en la sucesión, se denomina *muestra de tamaño* n. El conjunto de todas las N^n muestras de tamaño n, que se pueden formar con los N elementos de $\mathcal{A}$, se denomina *espacio muestral* y lo designamos por $\mathcal{S}$.

Un proceso mediante el que se selecciona una de estas muestras, de modo que cada una tenga una determinada probabilidad de ser elegida decimos que es un *procedimiento de muestreo aleatorio* (PMA).

El mecanismo aleatorio empleado por el investigador para seleccionar la muestra, debe definir en $\mathcal{S}$ una función de probabilidad $\mathcal{P}$, tal que:

$$\mathcal{P}: \quad \Pr(s_j) \geq 0, \forall_j \; ; \; \sum_{\mathcal{S}} \Pr(s_j) = 1$$

En la práctica, podemos imaginar el procedimiento constituido por n extracciones aleatorias en $\mathcal{A}$, de modo que el cálculo de la probabilidad de una determinada muestra es:

$$\Pr(s_j) = \Pr(A_{j1}).\ \Pr(A_{j2}/A_{j1}) \ldots \Pr(A_{jn}/A_{j1}, A_{j2}, \ldots A_{jn-1})$$

Con la notación $\{\mathcal{A}, \mathcal{S}, \mathcal{P}\}$ expresamos que un PMA definido en $\mathcal{A}$, establece en el espacio muestral $\mathcal{S}$ la ley de probabilidad $\mathcal{P}$. Puede ocurrir que aplicando un determinado PMA, no todas las muestras $s_j \in \mathcal{S}$ tengan posibilidad de ser seleccionadas; cuando sea necesario precisar este hecho, denotaremos por $\mathcal{S}^*$ al subconjunto de $\mathcal{S}$ constituido por todas las muestras posibles; esto es, aquéllas tales que $\Pr(s_j) > 0$.

No vamos a tomar en consideración, dentro de la teoría elemental para poblaciones finitas, los procedimientos de muestreo aleatorio en los que el tamaño de la muestra, n, no está prefijado de antemano. No obstante, hacemos notar que en la práctica se presentan situaciones donde es necesario considerar modelos con n variable.

I.2. CONCEPTO DE PROCEDIMIENTO DE ESTIMACION.

El problema de estimación en poblaciones finitas, en esencia, puede plantearse de la siguiente forma:

Está definida una característica $\mathcal{X}$ en $\mathcal{A}$, que toma el valor numérico X_i en la unidad A_i, y deseamos determinar el valor numérico, θ, de una cierta función de los N valores X_i, por ejemplo el total $\Sigma\ X_i$, o la media $1/N\ \Sigma\ X_i$. Para este propósito seleccionamos una muestra, s, y utilizando la información que nos proporciona la observación de sus unidades, atribuimos a θ un valor $\hat{\theta}\ (s)$.

La función $\hat{\theta}$ que asocia a cada muestra, s, el valor numérico $\hat{\theta}(s)$, se denomina *estimador* y $\hat{\theta}(s)$ es la estimación particular de θ dada s.

Estando definido un procedimiento de muestreo aleatorio $\{\mathcal{A}, \mathcal{S}, \mathcal{T}\}$ un estimador es una variable aleatoria, definida como una función $\hat{\theta}$, de $\mathcal{S}^*$ en $\mathcal{R}$ cuya ley de probabilidad viene dada por

$$\Pr(\hat{\theta} = t) = \sum_{\mathcal{C}} \Pr(s_j)$$

siendo $\mathcal{C}$ el subconjunto de $\mathcal{S}^*$ constituido por todas las muestras para las cuales $\hat{\theta} = t$; esto es: $\mathcal{C} = \{s_j / \hat{\theta}(s_j) = t\}$

El conjunto constituído por un procedimiento de muestreo aleatorio $\{\mathcal{A}, \mathcal{S}, \mathcal{T}\}$ y un estimador $\hat{\theta}$ para θ, que está definido en toda muestra tal que $\Pr(s_j) > 0$, decimos que es un *procedimiento de estimación* y lo representamos por $\{\mathcal{A}, \mathcal{S}, \mathcal{T}; \hat{\theta}, \theta\}$.

Para establecer un procedimiento de estimación debemos de considerar dos operaciones, que aún estando relacionadas, están netamente diferenciadas; una de ellas consiste en diseñar el procedimieno de muestreo y la otra consiste en formar el estimador. En cuanto al procedimiende muestreo se dice que es *no restringido* cuando todas las muestras son posibles. En la práctica se utilizan procedimientos restringidos (en los que no todas las muestras son posibles) para facilitar la tarea de selección, para conseguir muestras más representativas, para reducir costes, etc. y en definitiva para diseñar procedimientos más eficientes, en el sentido del equilibrio entre coste y exactitud de la estimación. Las técnicas de estratificación, formación de conglomerados, selección sistemática, selección sin reposición, etc. dan origen a procedimientos de este tipo en los cuales las muestras consideradas no deseables, por ser menos representativas o más costosas, no tienen posibilidad de selección. Cuando el procedimiento es no restringido y todas las muestras son equiprobables, suele decirse que se trata de un diseño simple y en caso contrario que se trata de un diseño complejo, especialmente cuando se combinan varias de las técnicas citadas.

La segunda operación consiste en diseñar el estimador; esto es, en establecer la función $\hat{\theta}$ que para un procedimiento de muestreo

dado permita calcular una estimación, $\hat{\theta}\ (s)$ para θ, cualquiera que sea la muestra, s, seleccionada. Como veremos más adelante, las propiedades de $\hat{\theta}$ dependen de su distribución de probabilidad, y ésta depende del procedimiento de muestreo; por tanto, la formación de estimadores, en el conjunto del diseño, no es una operación independiente del procedimiento de muestreo que se adopte.

El uso de información previa es imprescindible para realizar diseños complejos y formar los estimadores, cuando se desea con los recursos dados establecer un procedimiento de estimación lo más eficiente posible.

En la práctica, generalmente, se desean estimaciones simultáneas relativas a diferentes características $\mathfrak{X}$, $\mathfrak{Y}$, $\mathfrak{Z}$, etc. definidas en las unidades de la población. Un diseño eficiente para una de ellas puede no serlo para otras; por esta razón, se plantea en estos casos un problema de objetivos múltiples para el que habrá de buscarse una solución de compromiso; en algún caso puede ocurrir que para algunas características las necesidades de un diseño óptimo sean tan dispares, respecto de las restantes, que habrá de reducirse el objetivo prescindiendo de aquéllas y estableciendo si es necesario un procedimiento por separado.

Para cada problema concreto, el diseño del procedimiento de estimación es un trabajo que requiere un cierto grado de experiencia e imaginación. La teoría sirve de guía para conducir el razonamiento pero no puede ofrecer en un manual la amplia gama de soluciones que puede aportar, para cada situación particular, la imaginación creadora del investigador.

Conviene hacer alguna observación acerca del problema de estimación en poblaciones finitas. La teoría clásica de estimación en poblaciones infinitas supone que la distribución de $\mathfrak{X}$, en la población base $\mathcal{A}$, pertenece a una familia conocida y queda totalmente determinada por uno o varios parámetros, cuya estimación se hace mediante la observación de una muestra aleatoria simple y un estimador convenientemente elegido. En la teoría de poblaciones finitas no se hacen hipótesis acerca de la distribución base; el problema generalmente consiste en estimar la distribución de frecuencias de $\mathfrak{X}$, o simplemente alguna caracterís-

tica particular como la media o el total (que no caracterizan totalmente a la distribución poblacional, excepto en el caso de poblaciones dicotómicas $X_i = 1$ ó 0). Es obvio que si se dispone de información a priori acerca de la población, podremos mejorar el procedimiento de selección de muestras y el diseño del estimador, para obtener estimaciones más acuradas; ejemplos son la estratificación y la selección con probabilidades no iguales.

Si los valores X_i constituyen un conjunto suficientemente denso en un intervalo de $\mathcal{R}$ (para lo cual es necesario, aunque no suficiente, que N sea grande) de modo que se les puede atribuir una distribución contínua conocida, estamos en condiciones de comprobar si los estimadores gozan de propiedades deseables; así por ejemplo, si es Normal, también lo serán las distribuciones de sumas muestrales y en cualquier otro caso de distribución paramétrica conocida, sus momentos nos permitirán caracterizar la distribución del estimador con suficiente aproximación (media, varianza, coeficientes de asimetría y curtosis, etc.).

Una diferencia a tener muy en cuenta es que la teoría de poblaciones infinitas supone siempre muestreo aleatorio simple, de forma que la función de densidad es el producto de n funciones de densidad, iguales a la de la población base y en consecuencia, la función característica de una suma de n observaciones es el producto de n funciones características, iguales también a la de base. Este hecho facilita grandemente el estudio cuando se parte de una familia paramétrica conocida. En poblaciones finitas se muestrea sobre un conjunto de unidades físicas; esto puede hacerse de muchas formas y es en este punto donde el investigador hace uso de su ingenio y conocimiento a priori para diseñar un procedimiento que origine un espacio muestral con propiedades deseables; esto es, que esté constituído por muestras lo más parecidas a la población base, para lo cual el diseño debe atribuir probabilidades nulas o muy pequeñas a las muestras no deseables y probabilidades altas a las más representativas. Esto produce complicaciones en el cálculo de probabilidades de la distribución del estimador, por lo que usualmente nos conformamos con disponer del conocimiento de la esperanza y la varianza y acaso de momentos de tercer y cuarto orden.

I.3. MUESTREO CON Y SIN REPOSICION.

Decimos que un procedimiento aleatorio de muestreo es *sin reposición* cuando todas las muestras que tienen algún elemento repetido son sucesos imposibles; el procedimiento no concede posibilidad de selección a estas muestras, ello implica que su probabilidad es cero.

Así pues, siendo $\mathcal{S}^* \subset \mathcal{S}$ el conjunto de muestras posibles, tales que $\forall\, s_j \in \mathcal{S}^*$ es $\Pr(s_j) > 0$, y $\forall\, s_j \notin \mathcal{S}^*$ es $P_r(s_j) = 0$, cuando el procedimiento es sin reposición toda muestra $s_j \in \mathcal{S}^*$ está formada por n unidades distintas.

Aunque en ocasiones es útil para el razonamiento teórico tener en cuenta la ordenación de las unidades de la muestra, en la práctica no suele utilizarse la información proporcionada por el orden en que han sido seleccionadas; por esta razón cuando sea conveniente designaremos por Ω al conjunto de las $\binom{N}{n}$ combinaciones, cada una denotada por ω_j, que resultan de considerar como una única muestra a cada clase de $n!$ muestras de $\mathcal{S}$ formadas por los mismos elementos. Si queremos considerar el orden de extracción designamos por Ω' al conjunto de $\binom{N}{n}\, n!$ muestras ordenadas y por ω_j' a cada una de ellas.

Decimos que un procedimiento de muestreo aleatorio es *con reposición* cuando existe alguna muestra $s_j \in \mathcal{S}^*$, esto es tal que $\Pr(s_j) > 0$, que tiene entre sus n elementos alguno repetido.

Pueden existir diferentes modelos, atendiendo a que todas las unidades de $\mathcal{A}$ puedan repetirse en la muestra igual número de veces, o que unas puedan repetirse más veces que otras, o incluso que algunas no puedan repetirse y sí las demás. En cualquier caso, de acuerdo con la definición, si existe al menos una muestra con probabilidad no nula, que tiene algún elemento repetido, decimos que el PMA es Con Reposición.

I.4. DISTRIBUCIONES MARGINALES DE LAS COMPONENTES DE LA MUESTRA.

La distribución conjunta de las componentes de la muestra, viene dada por la ley de probabilidad $\mathcal{P}$, que define la probabilidad de cada una de las muestras posibles. La distribución marginal de la componente r^a

de la muestra, vendrá dada por la probabilidad que tiene cada unidad $A_i \in \mathcal{A}$ de ocupar el lugar r.

Si designamos por A_r^* la componente r^a de la muestra genérica s, su distribución marginal, $\mathfrak{T}(r)$, viene dada por:

$$\mathfrak{T}(r): \Pr(A_r^* = A_i) = P_{ir}\ ;\ i = 1,2,\ldots,N$$

donde P_{ir} es la suma de las probabilidades de todas las muestras, s_j, que contienen a A_i en el lugar r:

$$P_{ir} = \sum_{\mathcal{R}} \Pr(s_j);\quad \mathcal{R} = \{s_j \in \bar{\mathbb{S}} / A_i = A_{jr}\}$$

Respecto de las distribuciones marginales, es interesante considerar dos propiedades importantes:

1. Equidistribución.

   Todas las componentes de la muestra tienen la misma distribución de probabilidad. Esto es,

$$\mathfrak{T}_{(1)} \equiv \mathfrak{T}_{(2)} \equiv \ldots \equiv \mathfrak{T}_{(r)} \equiv \ldots \equiv \mathfrak{T}_{(n)}$$

2. Independencia.

   Las distribuciones marginales de las componentes de la muestra son independientes. Esto es, para todo r y todo i, P_{ir} es independiente de cuales sean las unidades de $\mathcal{A}$ que ocupan cualquier otro lugar de la muestra.

Estas dos propiedades de equidistribución e independencia de las marginales, se exigen en la teoría ordinaria de poblaciones infinitas, pero en poblaciones finitas se utilizan procedimientos que pueden no cumplir alguna, o ninguna. En particular, los procedimientos Sin reposición no cumplen la independencia, puesto que P_{ir} depende de si A_i ocupa, o no, otro lugar distinto al r. Esta particularidad no es exclusiva de los modelos Sin reposición, puesto que también existen procedimientos Con reposición que no verifican la citada propiedad. En ambos casos, con y sin reposición, existen procedimientos que tampoco cumplen la propiedad de equidistribución.

I.5. MUESTRAS NO ORDENADAS. VECTOR ALEATORIO $\varepsilon(s)$.

Como ya se ha mencionado, la información proporcionada por el orden de las unidades en la muestra, no suele utilizarse para obtener estimaciones acerca de las características globales de la población; en estos casos, para la formación de estimadores y para el cálculo de sus momentos, es suficiente el conocimiento de cuales unidades constituyen la muestra y, en el muestreo con reposición, cuantas veces se repiten en élla. Esta información viene dada por el vector aleatorio N–dimensional:

$$\varepsilon(s) = (\varepsilon_1, \varepsilon_2, \ldots \varepsilon_i, \ldots, \varepsilon_N); \sum_1^N \varepsilon_i = n$$

donde ε_i = N° de veces que A_i está en la muestra s.

Si el procedimiento de muestreo es Con reposición; en general, ε_i puede tomar un valor del conjunto $\{0, 1, 2, \ldots, n\}$, con la condición $\sum_1^N \varepsilon_i = n$, dada por el tamaño de la muestra. Si el procedimiento es Sin reposición, cada ε_i toma uno de los valores $\{1, 0\}$, según que A_i esté presente o no, en la muestra.

El vector $\varepsilon(s)$ es un instrumento muy útil para dar forma explícita a los estimadores lineales que dependen de la identificación de las unidades muestrales; esto es, aquéllos constituídos por tantos sumandos como unidades distintas hay en la muestra, en los que intervienen coeficientes, K_i, definidos para cada unidad en particular. Esta clase de estimadores puede escribirse en la forma:

$$\hat{\theta} = \sum_{i=1}^N K_i X_i \varepsilon_i$$

El problema del cálculo de los momentos del estimador $\hat{\theta}$, en particular la esperanza y la varianza, se reduce al cálculo previo de los momentos $E(\varepsilon_i)$; $E(\varepsilon_i^2)$, $E(\varepsilon_i \varepsilon_j)$, $i \neq j$; del vector N–dimensional $\varepsilon(s)$. Para ello, establecido el procedimiento de muestreo, debe-

mos determinar el modelo probabilístico que le corresponde; esto es, determinar la distribución de $\varepsilon(s)$, mediante la cual obtendremos los momentos necesarios. Operando de esta forma:

$$E(\hat{\theta}) = E\left(\sum_1^N K_i X_i \varepsilon_i\right) = \sum_1^N K_i X_i E(\varepsilon_i)$$

$$V(\hat{\theta}) = V\left(\sum_1^N K_i X_i \varepsilon_i\right) = \sum_1^N K_i^2 X_i^2 V(\varepsilon_i) +$$

$$+ \sum_{i \neq j}^N K_i K_j X_i X_j \operatorname{Cov}(\varepsilon_i, \varepsilon_j)$$

donde

$$V(\varepsilon_i) = E(\varepsilon_i^2) - E(\varepsilon_i)^2$$

$$\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) - E(\varepsilon_i) E(\varepsilon_j), i \neq j$$

I.6. SELECCION DE MUESTRAS CON REPOSICION. MODELO POLINOMIAL.

El modelo probabilístico que responde a la selección de muestras con reposición, mediante n extracciones independientes y con ley de probabilidad constante, es el polinomial. Veamos la forma de practicarlo:

Sea M_i un entero positivo asociado a la unidad A_i que denominamos *tamaño* de A_i. En la práctica M_i representa el número de unidades menores que contiene (p.e. cuando los A_i son conglomerados) o bien es una media del peso o importancia que concedemos a la probabilidad de selección de A_i. Siendo:

$$M = \sum_1^N M_i \; ; \qquad P_i = M_i/M \; ; \qquad \left(\sum_1^N P_i = 1\right)$$

el siguiente método permite seleccionar muestras con reposición de modo que en cada extracción, las unidades de la población tienen una probabilidad P_i proporcional a su tamaño. El intervalo de los números

enteros $\{1, M\}$ se divide en N intervalos I_i conteniendo M_i enteros consecutivos:

$$I_1 = \{1, M_1\}; I_2 = \{1 + M_1, M_1 + M_2\}; \ldots\ldots; I_N =$$

$$= \left\{1 + \sum_{1}^{N-1} M_h, M\right\}$$

a continuación se elige un entero $\delta \in \{1, M\}$, con probabilidades iguales (utilizando p.e. una tabla de números aleatorios) y seleccionamos como primera unidad de la muestra aquella A_i tal que $\delta \in I_i$. Repitiendo este proceso n veces obtenemos una muestra de tamaño n, de modo que para cualquiera de las n extracciones se cumple

$$Pr(A_i) = Pr(\delta \in I_i) = M_i/M = P_i$$

El método de Lahiri puede abreviar este proceso; consiste en lo siguiente: siendo M_o un entero mayor o igual que todos los M_i, se elige un par de números aleatorios (i, j) tales que

$$1 \leqslant i \leqslant N ; \qquad 1 \leqslant j \leqslant M_o$$

si $j \leqslant M_i$ la unidad seleccionada es A_i, en caso contrario se repite la selección del par (i, j) tantas veces como sea necesario. La probabilidad de A_i es proporcional a M_i; en efecto:

Sea β la probabilidad de que el par (i, j) no dé lugar a la selección de una unidad:

$$Pr(j > M_i) = \beta$$

entonces la probabilidad de seleccionar A_i exactamente en el $r°$ intento es:

$$\beta^{r-1} \frac{1}{N} \cdot \frac{M_i}{M_o}$$

por tanto la probabilidad de seleccionar A_i es:

$$\Pr(A_i) = \sum_{r=1}^{\infty} \beta^{r-1} \frac{M_i}{N M_o} = \frac{M_i}{N M_o} (1 + \beta + \beta^2 + \dots) = \frac{M_i}{N M_o} \frac{1}{1-\beta}$$

resultando proporcional a M_i. Repitiendo n veces seleccionamos una muestra de tamaño n.

Puesto que las n extracciones son independientes y se realizan con la misma ley de probabilidad en $\mathcal{A}$, cualquiera de los dos procedimientos descritos responde al modelo Polinomial cuya función de probabilidad es:

$$\Pr(\varepsilon_1 = r_1, \dots, \varepsilon_h = r_h, \dots, \varepsilon_N = r_N) =$$

$$= \frac{n!}{r_1!\, r_2! \dots r_h! \dots r_N!} P_1^{r_1} P_2^{r_2} \dots P_h^{r_h} \dots P_N^{r_N}$$

con las condiciones

$$\sum_1^N P_i = 1 \; ; \quad \sum_1^N r_i = n$$

Su función característica,

$$\varphi(t_1, \dots, t_h, \dots t_N) = \left(\sum_1^N P_h \, e^{it_h} \right)^n$$

permite calcular fácilmente esperanzas, varianzas y covarianzas. En particular,

$$E(\varepsilon_i) = n P_i \; ; \quad E(\varepsilon^2) = n(n-1) P^2 + n P; \quad E(\varepsilon_i \varepsilon_j) = n(n-1) P_i P_j$$

$$V(\varepsilon_i) = n P_i (1 - P_i); \quad \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = -n P_i P_j, \; (i \neq j)$$

Si el muestreo es con probabilidades iguales, esto es $P_i = 1/N$, resulta:

$$E(\varepsilon_i) = n/N \; ; \quad V(\varepsilon_i) = n/N\,(1 - 1/N) \; ; \quad Cov(\varepsilon_i, \varepsilon_j) = -n/N^2$$

I.7. SELECCION SIN REPOSICION.

Consideremos en primer lugar el procedimiento básico que consiste en la selección irrestricta de muestras en Ω', con probabilidades iguales. Cualquier procedimiento en el que todas las muestras, $\omega_i' \in \Omega'$, (ordenadas y formadas por n unidades distintas) tengan la misma probabilidad de selección responde a este modelo.

Una técnica práctica consiste en efectuar n extracciones sucesivas en $\mathcal{A}$, atribuyendo en cada una de ellas, probabilidades iguales a las unidades no seleccionadas en las anteriores:

$$\Pr(A_r^* = A_i) = \begin{cases} \dfrac{1}{N-(r-1)} & \text{Si } A_i \text{ no ha sido seleccionada en} \\ & \text{las } (r-1) \text{ extracciones anteriores.} \\ 0 & \text{En el caso contrario.} \end{cases}$$

Otra técnica equivalente a la anterior consiste en efectuar extracciones en $\mathcal{A}$, con probabilidades iguales para todas las unidades, rechazando las unidades repetidas, hasta obtener una muestra de n unidades distintas. También es equivalente a extraer tantas muestras con reposición y probabilidades iguales, como sea necesario, hasta obtener una que no tenga unidades repetidas (rechazando las anteriores).

La probabilidad de una determinada muestra en Ω' es:

$$\Pr(\omega_j') = 1/N.\; 1/N\text{-}1\,, \ldots, 1/N - (r-1), \ldots 1/N - (n-1) =$$

$$= \frac{(N-n)!}{N!} = \frac{1}{\binom{N}{n} n!} = \frac{1}{\text{N° de muestras en } \Omega'}$$

La probabilidad de una determinada muestra $\omega_j \in \Omega$; esto es, de una cualquiera de las $n!$ muestras $\omega_j' \in \Omega'$, formadas por las mismas unidades que ω_j , por ser equiprobables, es:

$$\Pr(\omega_j) = \frac{n!}{\binom{N}{n} n!} = \frac{1}{\binom{N}{n}}$$

La probabilidad de que una determinada unidad $A_i \in \mathcal{A}$, aparezca en el lugar r° de la muestra, utilizando la equiprobabilidad de las muestras, es:

$$P_{ir} = \frac{\text{N° muestras con } (A_r^* = A_i)}{\text{N° total de muestras } \omega_j'} = \frac{\binom{N-1}{n-1} (n-1)!}{\binom{N}{n} n!} = \frac{1}{N}$$

por tanto para todo i y todo r es $P_{ir} = 1/N$; en consecuencia del procedimiento cumple la propiedad de equidistribución de las marginales.

La probabilidad de que una determinada unidad, A_i , aparezca en la muestra (en cualquier lugar), se denota por π_i y es:

$$\pi_i = \frac{\text{N° muestras con } A_i \in \omega_j}{\text{N° total de muestras } \omega_j} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

esta probabilidad también puede obtenerse como

$$\pi_i = \sum_{r=1}^{n} P_{ir} = n \cdot 1/N = n/N .$$

La probabilidad de que un determinado par de unidades distintas (A_i, A_j) aparezca en la muestra, se denota por π_{ij} y es:

$$\pi_{ij} = \frac{\text{N° muestras con } (A_i, A_j) \in \omega_j}{\text{N° total de muestras } \omega_j} = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

Estas probabilidades nos permiten obtener los momentos del vector aleatorio $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_i, \ldots, \varepsilon_N)$.

Teniendo en cuenta que cada ε_i toma el valor uno o cero, según que A_i esté presente, o no, en la muestra:

$$E(\varepsilon_i) = 1 \cdot \Pr(\varepsilon_i = 1) + 0 \cdot \Pr(\varepsilon_i = 0) = 1 \cdot \pi_i = \pi_i$$

$$E(\varepsilon^2) = 1^2 \cdot \Pr(\varepsilon_i = 1) = \pi_i$$

$$V(\varepsilon_i) = E(\varepsilon^2) - E(\varepsilon_i)^2 = \pi_i - \pi_i^2 = \pi_i \ (1 - \pi_i)$$

Teniendo en cuenta que el producto $\varepsilon_i \cdot \varepsilon_j$, para $i \neq j$, toma el valor 1 si A_i y A_j están presentes en la muestra y el valor cero en los demás casos,

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) - E(\varepsilon_i) E(\varepsilon_j) =$$

$$= 1 \cdot \pi_{ij} - \pi_i \pi_j = \pi_{ij} - \pi_i \pi_j$$

Sustituyendo π_i y π_{ij} por sus valores en el caso estudiado de muestreo con probabilidades iguales, se obtiene:

$$E(\varepsilon_i) = n/N, \quad E(\varepsilon_i^2) = n/N$$

$$V(\varepsilon_i) = n/N \ (1 - n/N)$$

$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \frac{n(n-1)}{N(N-1)} = -\frac{(N-n)n}{(N-1)N^2}$$

PROPIEDADES GENERALES DE LAS PROBABILIDADES DE SELECCION EN MUESTREO SIN REPOSICION.

Cualquiera que sea el procedimiento de muestreo sin reposición, las probabilidades $\pi_i = \Pr(A_i \in \text{muestra})$; $\pi_{ij} = \Pr(A_i, A_j \in \text{muestra})$, verifican las siguientes propiedades:

1. $$\sum_{i=1}^{N} \pi_i = n$$

Puesto que $$\sum_{1}^{N} \pi_i = E\left(\sum_{1}^{N} \varepsilon_i\right) = E(n) = n$$

2. $$\sum_{j(j\neq i)}^{N} \pi_{ij} = (n-1)\,\pi_i$$

Puesto que $$\sum_{j(j\neq i)}^{N} \pi_{ij} = E\left(\sum_{j(j\neq i)}^{N} \varepsilon_i\,\varepsilon_j\right) =$$

$$= E\left(\varepsilon_i\left(\sum_{1}^{N} \varepsilon_j - \varepsilon_i\right)\right) = E(\varepsilon_i\,(n-\varepsilon_i)) = (n-1)\,\pi_i$$

3. $$\sum_{j(j\neq i)} (\pi_i\,\pi_j - \pi_{ij}) = \pi_i\,(1-\pi_i)$$

Puesto que $$\sum_{j(j\neq i)} \pi_i\,\pi_j = \pi_i\left(\sum_{1}^{N} \pi_j - \pi_i\right) = \pi_i\,(n-\pi_i)$$

y teniendo en cuenta 2.

ALGUNOS PROCEDIMIENTOS DE SELECCION SIN REPOSICION Y PROBABILIDADES DESIGUALES

Hemos visto el procedimiento básico de selección sin reposición, en el que todas las muestras son equiprobables y cada uno de los N elementos de la población tiene igual probabilidad de estar presente en la muestra. En ciertas situaciones puede ser conveniente el uso de procedimientos de muestreo que concedan diferentes probabilidades de inclusión en la muestra a las unidades de la población. En particular interesan aquellos según los cuales la probabilidad π_i es proporcional al tamaño, M_i de la unidad A_i .

En primer lugar, veamos qué ocurre en el modelo polinomial si no se repone la unidad seleccionada en cada extracción; esto equivale en el esquema de urna a eliminar las M_i bolas de color i que representan a la unidad A_i, cuando ésta ha sido extraida. Consideremos para simplificar, el caso $n = 2$

$$P_{i1} = \frac{M_i}{M} = P_i$$

$$P_{i2} = \sum_{j \neq i} P_j \cdot Pr\,(A_{i2} / A_{j1}) = \sum_{j \neq i} P_j \frac{M_i}{M - M_j} = \sum_{j \neq i} P_j \frac{P_i}{1 - P_j}$$

y siendo

$$A = \sum_{1}^{N} \frac{P_i}{1 - P_i} \quad , \text{ resulta}$$

$$P_{i2} = P_i \left(A - \frac{P_i}{1 - P_i} \right)$$

Las probabilidades de inclusión en la muestra, son

$$\pi_i = P_{i1} + P_{i2} = P_i \left(1 + A - \frac{P_i}{1 - P_i}\right); \qquad (1)$$

$$\pi_{ij} = \Pr(A_{i1}, A_{j2}) + \Pr(A_{j1}, A_{i2}) = P_i P_j \left(\frac{1}{1-P_i} + \frac{1}{1-P_j}\right)$$

Yates y Grundy (1953) estudian este modelo y presentan un ejemplo, para el caso $N = 4$ y $n = 2$, con los siguientes datos:

$$N = 4; \quad n = 2; \quad P_i = 0{,}1; \quad 0{,}2; \quad 0{,}3; \quad 0{,}4$$

Aplicando las fórmulas P_{i1} y P_{i2} para $i = 1, 2, 3, 4$ se obtienen las distribuciones marginales que se muestran en el siguiente cuadro

PROBABILIDADES P_{i1} y P_{i2}

Componentes	A_1	A_2	A_3	A_4
Primera	0,1	0,2	0,3	0,4
Segunda	0,13452	0,24127	0,30833	0,31587

donde se observa que las distribuciones marginales de la primera y segunda componentes no son iguales. En el siguiente cuadro aparecen las probabilidades de inclusión en la muestra de cada una de las cuatro unidades:

A_i	π_i	nP_i
A_1	0,23452	0,2
A_2	0,44127	0,4
A_3	0,60833	0,6
A_4	0,71587	0,8

se observa que no son proporcionales al tamaño de la unidad, en cuyo caso serían las dadas en la columna $n\,P_i$.

A primera vista, la forma más inmediata de resolver el problema de conseguir π_i proporcionales a unos valores dados, consiste en determinar el vector P_i $(i = 1, 2, \ldots, N)$ a partir de las ecuaciones (1) igualando π_i a los valores deseados.

Yates y Grundy y Narain (1951) dan métodos para despejar los P_i , pero originan pesados cálculos para valores del tamaño muestral superiores a 2 y resultan poco útiles para valores incluso moderados de dicho tamaño. Fellegi (1963), Van Beek y Vermetten (1967) y Brewer (1967) han estudiado este problema mediante el uso de procedimientos iterativos, si bien el proceso de cálculo se complica a medida que crece el tamaño n.

Un procedimiento propuesto por Ikeda, consigue probabilidades π_i proporcionales a unas M_i , ó P_i , dadas; consiste en efectuar la primera extracción con probabilidades P_i y sin reposición, efectuar todas las demás, hasta la n-ésima, con probabilidades iguales. Con este procedimiento resulta:

π_i = P_r (A_i en la primera extracción) + P_r (A_i en una de las $(n-1)$ siguientes y no en la primera) =

$$= P_i + (1 - P_i)\,\frac{n-1}{N-1} \quad ; \text{ de donde resulta}$$

$$\pi_i = \frac{(N-n)\,P_i + (n-1)}{N-1} \quad ; \quad P_i = \frac{N-1}{N-n}\,\pi_i - \frac{n-1}{N-n}$$

En esta última fórmula se observa que el procedimiento tiene alguna restricción ya que las P_i deben ser todas positivas. En la práctica esto hace que el procedimiento no es aplicable cuando existe gran diferencia entre el mayor y el menor M_i. Así por ejemplo, no podría aplicarse al caso estudiado antes ($N = 4$, $n = 2$, $M_i = 1, 2, 3, 4$) ya que

para A_1 resultaría:

$$P_1 = \frac{3}{2} \cdot 2 \cdot \frac{1}{10} - \frac{1}{2} = \frac{3}{10} - \frac{1}{2} \text{ negativa.}$$

La probablidad de inclusión en la muestra, de un par de unidades (A_i, A_j) es:

$$\pi_{ij} = P_i \frac{n-1}{N-1} + P_j \frac{n-1}{N-1} + (1 - P_i - P_j) \frac{(n-1)(n-2)}{(N-1)(N-2)}$$

simplificando resulta:

$$\pi_{ij} = \frac{n-1}{N-1}\left(\frac{N-n}{N-2}(P_i + P_j) + \frac{n-2}{N-2}\right)$$

Una importante propiedad de este procedimiento es que la probabilidad de selección de cada una de las $\binom{N}{n}$ muestras distintas no ordenadas, es proporcional a la suma de los tamaños de las n unidades que la componen. Consecuencia de esto es que la razón muestral $R = \bar{x} / \bar{y}$ resulta insesgado para la razón poblacional cuando se toman las P_{i1} porporcionales a los valores Y_i de la variable auxiliar. Sirva como ejercicio la prueba de estas dos afirmaciones.

Una generalización de este procedimiento ha sido propuesta por Midzuno (1952). Consiste en efectuar m extracciones sin reposición y con probabilidades iguales; en la extracción $(m + 1)$ se asignan probabilidades

$$P_i + \sum_{r-1}^{m} \frac{P_r}{N - m}$$

donde P_r corresponde a la unidad extraïda en r-ésimo lugar $(1 \leqslant r \leqslant m)$ y por último las $n - (m + 1)$ restantes extracciones se efectúan también sin reposición y probabilidades iguales. Para $m = 0$, coincide con el estudiado antes.

Otros procedimientos han sido propuestos y estudiados por diversos autores. Brewer (1963) para $n = 2$ presenta el siguiente:

Siendo $P_i = M_i / M$, tales que $0 < P_i < 1/2$ la primera extracción se realiza asignando probabilidades proporcionales a:

$$\frac{P_i\ (1 - P_i)}{(1 - 2P_i)} \quad \text{para la unidad } A_i$$

y la segunda, asignando probabilidades proporcionales $(a\ P_i)$ a las $(N - 1)$ unidades no extraídas en la primera; esto es:

$$\frac{P_j}{1 - P_i} \quad \text{para la unidad } A_j, \text{ siendo } A_i \text{ la primera extraída.}$$

Siendo C la constante de proporcionalidad para las probabilidades de la primera extracción, las distribuciones marginales son:

$$P_{i1} = \frac{C\,P_i\,(1 - P_i)}{1 - 2P_i} \quad ; \qquad 1^a \text{ Componente}$$

$$P_{i2} = C\,P_i \sum_{j \neq i}^{N} \frac{P_j}{1 - 2P_j} \quad ; \; 2^a \text{ Componente}$$

de donde resulta que no son idénticamente distribuídas y que π_i es proporcional a P_i; en efecto:

$$\pi_i = P_{i1} + P_{i2} = C\,(A + 1)\,P_i$$

siendo $A = \sum_{i=1}^{N} \frac{P_j}{1 - 2P_j}$, de donde resulta

$$C\,(A + 1) = C\,T\,E. \text{ para unas } P_i \text{ dadas.}$$

La probabilidad π_{ij} para $j \neq i$, que en este modelo por ser $n = 2$, coincide con la probabilidad de la muestra formada por las unidades A_i, A_j en cualquier orden es:

$$\pi_{ij} = \frac{C\,P_i\,(1-P_i)}{1-2P_i} \cdot \frac{P_j}{1-P_i} + \frac{C\,P_j\,(1-P_j)}{1-2P_j} \cdot \frac{P_i}{1-2P_j}$$

de donde simplificando se obtiene,

$$\pi_{ij} = C\,P_i\,P_j \left(\frac{1}{1-2P_i} + \frac{1}{1-2P_j} \right)$$

La probabilidad π_i , también puede obtenerse en este caso como:

$$\pi_i = \sum_{j \neq i} \pi_{ij}$$

Otro modelo de esta clase ha sido propuesto y estudiado por Durbin. Cumpliendo P_i las mismas condiciones que en el anterior se extrae la primera unidad de la muestra con probabilidad P_i y la segunda, sin reposición de la extraida en primer lugar, con probabilidad proporcional a:

$$P_j \left(\frac{1}{1-2P_i} + \frac{1}{1-2P_j} \right)$$

Con este procedimiento se obtienen probabilidades de inclusión en la muestra, proporcionales a P_i . A diferencia del procedimiento de Brewer en éste, las distribuciones marginales de las dos componenetes de la muestra son iguales. Es obvio, por otra parte, que en todos estos modelos sin reposición las distribuciones de las componentes de la muestra no son independientes puesto que las probabilidades de la segunda componente dependen del resultado de la primera extracción.

El procedimiento que describimos a continuación proporciona muestras de tamaño n, sin reposición, de modo que las probabilidades

de inclusión, π_i, resultan proporcionales al tamaño M_i de la correspondiente unidad. Siendo $M = \Sigma M_i$, debe cumplirse la condición $(M_i/M) < 1/n$ para todo $i = 1, 2, \ldots,$ N. Se numeran aleatoriamente las unidades de la población $A_1, A_2, \ldots A_i, \ldots, A_N$; el intervalo de los números enteros $\{1, 2, \ldots, M\}$ se divide en N intervalos I_i de longitud M_i, tal como se hizo en el modelo polinomial, y se elige aleatoriamente un entero K comprendido en $\{1, 2, \ldots, M/n\}$, supuesto M/n entero. La muestra seleccionada está formada por las unidades cuyo intervalo I_i contenga a uno de los números:

$$K;\ K + M/n;\ K + 2M/n;\ \ldots;\ K + (n - 1)\, M/n$$

Es claro que la condición $(M_i/M) < 1/n$ asegura que ninguna unidad puede aparecer dos veces (muestreo sin reposición) ya que la sucesión de aleatorios "salta" más que la longitud de cualquier intervalo; esto es $M/n > M_i$. La probabilidad π_i es obvio que resulta proporcional a M_i, pero el cálculo de π_{ij} es prácticamente inabordable; Hartley y Rao (1962) han obtenido una fórmula aproximada. Un ejemplo aparece en el siguiente cuadro:

Unidades	Tamaño M_i	Tamaño Acumulado	Intervalos I_i
A_1	20	20	(1 – 20)
A_2	8	28	(21 – 28)
A_3	15	43	(29 – 43)
A_4	10	53	(44 – 53)
A_5	25	78	(54 – 78)
A_6	12	90	(79 – 90)

Si para seleccionar una muestra de tamaño $n = 3$, por ser $M/n = 30$, el K elegido entre 1 y 30 resulta ser K = 18, la sucesión aleatoria es: 18, 48, 78 por tanto los intervalos son el 1, el 4 y el 5; la muestra será (A_1, A_4, A_5).

I.8. FORMACION DE ESTIMADORES.

La formación de un estimador es una operación que debe considerarse conjuntamente con la de establecer el Procedimiento de muestreo aleatorio, en el contexto del problema de diseñar un Procedimiento de estimación. Hemos dicho que dada una muestra, s, el estimador $\hat{\theta}$ recoge una información obtenida de aquélla, mediante la que se obtiene una estimación particular $\hat{\theta}$ (s); esta información puede estar constituída por:

1. Los n valores X_i, o los n pares (X_i, Y_i) si se trata de observaciones bidimensionales, correspondientes a las unidades de la muestra, que la teoría ordinaria supone son observados sin error.

2. Informaciones relativas a cada una de las unidades muestrales, como son:

   a) Cuál es la unidad; esto es, su identificación en la población.

   b) El lugar que ocupa en la muestra seleccionada.

   c) Cuál es la muestra particular en que aparece la unidad.

Por otra parte, puede ocurrir que antes de efectuar el muestreo se dispone de otras informaciones, como son:

3. a) Información acerca de características de la población (totales, varianzas, etc) relativas a variables $\mathcal{Y}$, $\mathcal{Z}$, etc. distintas de las $\mathcal{X}$ que se desea estimar.

3.b) Información relativa a las N unidades $A_i \in \mathcal{A}$, como puede ser el conocimiento de sus tamaños, usualmente denotados por M_i y que representan la importancia o peso que cada A_i tiene en $\mathcal{A}$, con respecto al problema de estimación planteada.

El tipo más simple de estimador solamente utiliza la información del grupo 1, como por ejemplo cuando se toma la media de los n valores muestrales para estimar $\bar{X}$ en muestreo con probabilidades iguales; para formar el estimador no es necesario en estos casos, disponer de informaciones de los grupos 2 y 3. Generalmente el procedimiento consiste en actuar por analogía; así para estimar la media $\bar{X}$, la varianza σ^2, o la razón $R = \bar{X}/\bar{Y}$ se forman con las observaciones la media, la varianza o la razón muestral, respectivamente. No siempre este método da lugar a estimadores insesgados; s_1^2 y $\hat{R} = \bar{x}/\bar{y}$ por ejemplo, no lo son, aunque serán consistentes y en algunos casos puede corregirse su sesgo multiplicándolos por una constante convenientemente elegida, como ocurre con $(n/n - 1)\, s_1^2$ que es insesgado para σ^2.

Otro tipo de estimador se construye utilizando información del grupo 1 y alguna otra información previa al proceso de muestreo del grupo 3. Así por ejemplo, si se dispone del total, Y, de una característica $\mathcal{Y}$ correlacionada con $\mathcal{X}$ el estimador de razón para el total X:

$$\hat{X}_R = \hat{R}\, Y = \frac{\bar{x}}{\bar{y}}\, Y \; ; \; \bar{x} \text{ e } \bar{y} \text{ son medias muestrales.}$$

puede ser más acurado que el estimador usual

$$\hat{X} = (N/n)\, x \qquad ; \qquad x = \text{total muestral.}$$

que utiliza solamente información del grupo 1.

El estimador "de regresión lineal" para $\bar{X}$:

$$\hat{\bar{X}}_{RL} = \bar{x} + B(\bar{Y} - \bar{y})$$

que utiliza el conocimiento previo de $\bar{Y}$ es insesgado para $\bar{X}$, cuando $\bar{x}$ e $\bar{y}$ lo son para $\bar{X}$ e $\bar{Y}$, siendo B una constante que cuanto más próxima esté del coeficiente de regresión lineal de $\mathcal{X}$ sobre $\mathcal{Y}$, mayor precisión proporcionará al estimador.

Un tipo más complejo y de especial importancia en la práctica de poblaciones finitas, está constituído por estimadores que además de la informción del grupo 1, utilizan la del grupo 2 (a); se trata de estimadores lineales de la forma:

$$\hat{\theta} = \sum_{i=1}^{n^*} C_i \, X_i$$

donde la suma se extiende a las n^* unidades distintas de la muestra ($n^* \leqslant n$, en caso de muestreo con reposición) y C_i depende de cual es la unidad A_i a que corresponde el i° sumando de $\hat{\theta}$. La utilidad de esta clase de estimadores reside en el hecho de que, si se pueden identificar las unidades muestrales, puede introducirse en $\hat{\theta}$, a través de C_i, alguna información previa disponible acerca de ellas; en particular, puede conseguirse que $\hat{\theta}$ sea insesgado para procedimientos de muestreo en $\mathcal{A}$ con probabilidades desiguales. La presentación de una teoría para esta clase de estimadores, en muestreo sin reposición y probabilidades desiguales, se debe a Horvitz y Thompson (1952).

Si junto con la información del grupo 1, se utiliza la del grupo 2(b), tenemos un tipo de estimador que puede expresarse de la forma,

$$\hat{\theta} = \sum_{j=1}^{n} K_j \, X_i$$

donde el coeficiente K_j que se aplica al valor observado, X_i, en la unidad muestral A_i, depende del lugar del orden j, de 1 a n, que dicha unidad ocupa en la muestra. Esto implica la necesidad de conocer el orden de extracción de las unidades muestrales.

Mencionemos, por último, otro tipo de estimador que utiliza la información del grupo 2(c). Puede expresarse de la forma,

$$\hat{\theta}\,(s) = K\,(s) \sum_{i=1}^{n} X_i$$

donde K (s) es un coeficiente que se aplica a la muestra particular que se ha seleccionado. Esto implica que para cada posible muestra, se define un K(s) distinto y por tanto, es preciso conocer cual ha sido la muestra observada.

Es obvio que pueden formarse tipos de estimadores más complicados combinando informaciones de varios grupos; así cuando los coeficientes dependen de cual es la undiad i^a y del lugar de orden que ocupa en la muestra.

Excepto la razón muestral, todos los estimadores mencionados pertenecen a la clase de estimadores lineales; esto es, formados por n sumandos, tantos como unidades muestrales, cada uno de los cuales es el producto de un coeficiente dado, por un valor observado en la correspondiente unidad muestral.

A modo de curiosidad, mencionemos que para R también puede darse un estimador lineal de la forma:

$$\hat{R}' = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{Y_i}$$

que equivale a la media de las razones individuales en los n elementos de la muestra.

Conviene hacer notar que tanto $\hat{R}$ como $\hat{R}'$ sí son lineales para un conjunto dado de $(Y_1, Y_2, \ldots Y_i, \ldots Y_n)$.

I.9 PROPIEDADES DE LOS ESTIMADORES.

Hemos visto que un estimador $\hat{\theta}$ es una variable aleatoria unidimensional, por tanto sus propiedades vendrán dadas por su distribución de probabilidad en el cuerpo de los números reales. Particularmente nos interesan su esperanza matemática y su varianza, cuyo cálculo puede efectuarse de la siguiente forma:

$$E(\hat{\theta}) = \sum_{S} \hat{\theta}(s_j) \cdot Pr(s_j); \text{ o bien}$$

$$E(\hat{\theta}) = \sum_{\mathcal{R}} t \cdot Pr(\hat{\theta} = t)$$

y designando por $\bar{\theta}$ a la E ($\hat{\theta}$), la varianza y la desviación típica vienen dadas por

$$V(\hat{\theta}) = E(\hat{\theta} - \bar{\theta})^2 = E(\hat{\theta}^2) - E(\bar{\theta})^2$$

$$\sigma(\hat{\theta}) = \sqrt{V(\hat{\theta})}$$

Si $\hat{\theta}$ es un estimador para θ, se denomina *sesgo* a la diferencia:

$$B(\hat{\theta}) = E(\hat{\theta}) - \theta$$

El estimador se dice que es *insesgado* cuando $B(\hat{\theta}) = 0$, o bien $E(\hat{\theta}) = \theta$.

En todo estimador es deseable la propiedad de ser insesgado que cumple cuando su esperanza matemática coincide con el valor a estimar, pero también es deseable que su dispersión alrededor de θ sea pequeña. Se dice que un estimador es más preciso que otro cuando tiene menor varianza (menor desviación típica). Estas dos propiedades, precisión y ausencia de sesgo, son independientes en el sentido de que un estimador insesgado puede ser más o menos preciso que otro sesgado. Para comparar dos estimadores insesgados se utiliza su desviación típica, que también se denomina *error de muestreo,* y se dice que es más eficiente el más preciso. Para comparar estimadores, alguno de los cuales es sesgado, se utiliza una media conjunta de sesgo y la precisión; se trata del *error cuadrático medio,* que se define como la esperanza matemática del cuadrado de la diferencia entre la estimación y el valor a estimar, esto es:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

El ECM puede descomponerse, como sigue, en dos sumandos; uno es la varianza y otro el cuadrado del sesgo:

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = E(\hat{\theta} - E(\hat{\theta}) + B(\theta))^2 =$$

$$= E(\hat{\theta} - \bar{\theta})^2 + E(B(\theta)^2) + 2E(B(\hat{\theta})(\hat{\theta} - \bar{\theta})) = V(\hat{\theta}) + B(\hat{\theta})^2$$

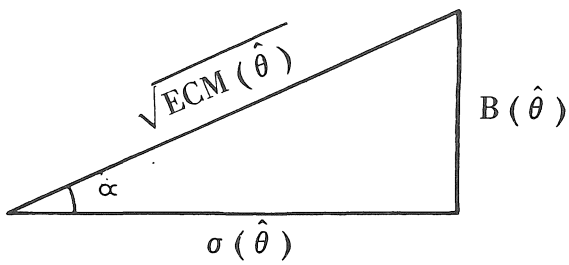
por ser

$$E(B(\hat{\theta})\,(\hat{\theta} - \overline{\theta}) = B(\hat{\theta})\,.\,E(\hat{\theta} - \overline{\theta}) = B(\hat{\theta})\,.\,0 = 0$$

Si $\hat{\theta}$ es insesgado su ECM coincide con la varianza.

Se habla de *acuracidad* para expresar la propiedad de un estimador cuyo ECM es pequeño; en otras palabras, un estimador se dice que es más acurado que otro cuando su ECM es menor.

Podemos representar gráficamente la $\sqrt{ECM\,(\hat{\theta})}$ por la hipotenusa de un triángulo rectángulo cuyos catetos son la desviación típica $\sigma\,(\hat{\theta}) = \sqrt{V\,(\hat{\theta})}$ y el sesgo $B\,(\hat{\theta})$.



La contribución del sesgo a la $\sqrt{ECM(\hat{\theta})}$ puede medirse por la relación $B\,/\sigma$, que representa la tg. trigonométrica del ángulo $\propto$; cuanto menor sea ésta, mayor tiende a ser la contribución de $\sigma(\hat{\theta})$ a la $\sqrt{ECM(\hat{\theta})}$. Como elemento de referencia, observemos que para un sesgo del orden del 10 por ciento de la desviación típica, ésta es del orden del 99,5 por ciento de la $\sqrt{ECM(\hat{\theta})}$.

La presencia de un sesgo pequeño, o nulo, es obviamente una propiedad deseable para un estimador; no obstante, si la reducción del sesgo va acompañada de un aumento de la desviación típica, puede resultar que en conjunto el estimador tiene mayor ECM; o lo que es lo mismo, es menos acurado.

Conviene aclarar una cuestión terminológica; la denominación de "error de muestreo" a la desviación típica de un estimador, $\sigma\,(\hat{\theta})$, no debe inducir a confusión entre el error puntual:

$$e\,(\hat{\theta}) = \hat{\theta} - \theta$$

diferencia entre una estimación particular y el valor que se desea estimar y el error de muestreo, que es la esperanza matemática de e^2 $(\hat{\theta})$. El error de muestreo nos informa de la precisión del estimador pero no del error puntual presente en una estimación particular; aquél puede, en general, calcularse pero éste será siempre desconocido.

Otra propiedad deseable de los estimadores, cuando no son insesgados, en la *consistencia* que se cumple cuando el sesgo B $(\hat{\theta})$ tiende a cero al aumentar el tamaño de la muestra. Si es posible obtener de forma explícita, exacta o aproximada, las relaciones que ligan B $(\hat{\theta})$ y σ $(\hat{\theta})$ con el tamaño muestral, cabe la posibilidad de comprobar fácilmente la consistencia de $\hat{\theta}$ y a continuación obtener el tamaño muestral que garantice una relación B $(\hat{\theta})/\sigma$ $(\hat{\theta})$ suficientemente pequeña, para cumplir los requisitos de acuracidad.

INTERVALO DE CONFIANZA CUANDO EL ESTIMADOR ES INSESGADO.

Admitiendo que la distribución del estimador $\hat{\theta}$ es Normal, con media $E(\hat{\theta}) = \theta$ y varianza σ^2 $(\hat{\theta})$, para un α dado $(0 < \alpha < 1)$ se determina en las tablas de la distribución N(0,1) el valor t_α, tal que

$$\Pr\{-t_\alpha < N(0,1) < t_\alpha\} = 1 - \alpha$$

y con este valor t_α, el intervalo para θ correspondiente a un coeficiente de confianza $(1 - \alpha)$, es:

$$(\hat{\theta} - t_\alpha \sigma(\hat{\theta}),\ \hat{\theta} + t_\alpha \sigma(\hat{\theta}))$$

En efecto: $(\hat{\theta} - \theta)/\sigma(\hat{\theta})$ es N(0,1), por tanto

$$\Pr\left\{\left|\frac{\hat{\theta} - \theta}{\sigma(\hat{\theta})}\right| < t_\alpha\right\} = 1 - \alpha$$

de donde se obtiene que el intervalo aleatorio $(\hat{\theta} \pm t_\alpha \sigma(\hat{\theta}))$ cubre a θ con probabilidad $(1 - \alpha)$.

Si no se dispone del conocimiento previo de $\sigma(\hat{\theta})$ se utiliza una estimación insesgada $\hat{\sigma}(\hat{\theta})$ procedente de la misma muestra que nos proporciona la estimación $\hat{\theta}$ para θ; en este caso no podemos asegurar que el intervalo cumpla exactamente los requisitos probabilísticos enunciados; no obstante, en la mayoría de las situaciones $\hat{\theta}$ tomará la forma de una suma de variables normales (estamos en la hipótesis de normalidad) y analizando el caso particular podremos recurrir a la distribución t-Student que nos dará una solución exacta o, al menos suficientemente aproximada.

Si la hipótesis de normalidad es muy arriesgada, y ni siquiera puede admitirse a título aproximado, puede hacerse uso de la siguiente desigualdad estudiada en Cálculo de Probabilidades. válida para toda variable aleatoria con media y varianza finitas:

$$\Pr\left\{ |\hat{\theta} - \theta| < K \right\} \geqslant 1 - \frac{\sigma^2(\hat{\theta})}{K^2}$$

que nos permite construir, en cualquier caso, el intervalo:

$$\left\{ \hat{\theta} - \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}}, \hat{\theta} + \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}} \right\}$$

obtenido haciendo $\sigma^2(\hat{\theta})/K^2 = \alpha$; esto es, cumpliendo:

$$\Pr\left\{ \hat{\theta} - \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}} < \theta < \hat{\theta} + \frac{\sigma(\hat{\theta})}{\sqrt{\alpha}} \right\} \geqslant 1 - \alpha$$

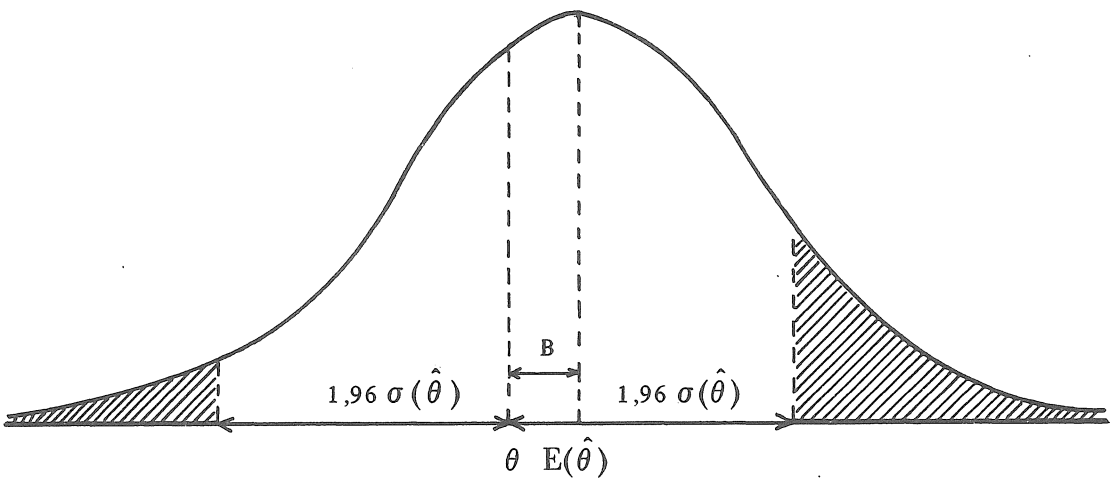
Como es sabido, los intervalos así construídos, para un mismo valor de α , son más amplios que los obtenidos bajo la hipótesis de normalidad, o bajo cualquier otra que suponga el conocimiento de la distribución de $\hat{\theta}$; por esta razón, cabe pensar que aún utilizando una estimación de $\sigma^2(\hat{\theta})$ en vez de $\sigma^2(\hat{\theta})$, seguirán conservando la validez de sus afirmaciones probabilísticas.

INTERVALO DE CONFIANZA CUANDO EL ESTIMADOR ES SESGADO.

Si $\hat{\theta}$ tiene un sesgo $B = E(\hat{\theta}) - \theta$, la probabilidad de confianza sufre una perturbación de modo que para un t_α dado por la distribución N(0,1), la probabilidad de confianza no es $(1 - \alpha)$ sino menor. Supongamos $t_\alpha = 1{,}96$ y que $\hat{\theta}$ tiene sesgo positivo; entonces la distribución de $\hat{\theta}$ sufre una traslación hacia la derecha de magnitud B, por tanto, como se indica en la figura, $\Pr\{(\hat{\theta}-\theta) > 1{,}96\,\sigma(\hat{\theta})\}$ no es $\alpha/2$ sino mayor y $\Pr\{(\hat{\theta}-\theta) < 1{,}96\,\sigma(\hat{\theta})\}$ tampoco es $\alpha/2$ sino menor. Como por otra parte, de todos los intervalos de igual amplitud, el centrado es el que recoge mayor probabilidad resulta que en conjunto, por no ser centrado el intervalo cuando $B \neq 0$, resulta que $\Pr\{|\hat{\theta}-\theta| < 1{,}96\,\sigma(\hat{\theta})\}$ disminuye. Si el sesgo es negativo ocurre lo mismo, si bien sería mayor que $\alpha/2$ la probabilidad en la cola izquierda y menor que $\alpha/2$ en la derecha. La disminución de la probabilidad de confianza depende de la relación $|B|/\sigma(\hat{\theta})$; en efecto, llevando el origen a $E(\hat{\theta})$ para calcular las probabilidades tenemos, en la cola derecha:

$$\int_{1{,}96\sigma(\hat{\theta})-B}^{\infty} N(0,\sigma(\hat{\theta})) = \int_{1{,}96-\frac{B}{\sigma(\hat{\theta})}}^{\infty} N(0,1) \quad \text{en vez de} \quad \int_{1{,}96}^{\infty} N(0,1)$$

y, análogamente en la cola izquierda. Por tanto, resulta que la alteración



de la probabilidad de confianza depende de $|B|/\sigma(\hat{\theta})$, de modo que cuanto mayor es esta razón más se reduce aquella probabilidad. La tabla adjunta muestra para $t_\alpha = 1{,}96$ la probabilidad de confianza que corresponde a diversos valores de $|B|/\sigma(\hat{\theta})$ con B positivo. Se observa que cuando la relación del sesgo al error de muestreo es del 10 por ciento, o menor, la disminución de la probabilidad de confianza no es importante. Esto nos da una regla práctica para admitir el uso de un estimador sesgado si tenemos acotado su sesgo respecto al error de muestreo. Debe tenerse en cuenta que la probabilidad de error por exceso aumenta si B es positivo y disminuye la probabilidad de error por defecto; si B es negativo, ocurre lo contrario. Asi p.e. en la tabla observamos que con $B > 0$,

$$\Pr\left\{(\hat{\theta} - \theta) > 1{,}96\,\sigma(\hat{\theta})\right\} = 0{,}0314 \quad \text{para } B/\sigma(\hat{\theta}) = 0{,}1$$

en vez de 0,025 si $\hat{\theta}$ fuera insesgado.

PROBABILIDAD DE CONFIANZA Y PROBABILIDAD DE ERROR PARA $t_\alpha = 1{,}96$, EN FUNCION DE $|B|/\sigma(\hat{\theta})$

$\frac{\lvert B \rvert}{\sigma(\hat{\theta})}$	Probabilidad de confianza	Probabilidad de error $>1{,}96\,\sigma(\hat{\theta})$	Probabilidad de error $<11{,}96\,\sigma(\hat{\theta})$
0,02	0,9500	0,0262	0,0238
0,04	0,9498	0,0274	0,0228
0,06	0,9496	0,0287	0,0217
0,08	0,9492	0,0301	0,0207
0,10	0,9489	0,0314	0,0197
0,20	0,0454	0,0392	0,0154
0,40	0,9315	0,0594	0,0091

HIPOTESIS DE NORMALIDAD.

En poblaciones finitas no cabe hablar, en rigor, de que tengan distribución normal, ya que ésta es contĭnua; no obstante admitimos este supuesto cuando la distribución de frecuencias se ajusta a las correspondientes frecuencias obtenidas de la distribución normal.

Si el estimador está formado por una suma, o combinación lineal, de variables cuya población base es normal, sabemos que también el estimador tiene distribución normal en el muestreo.

Si la distribución de la población base no es normal, se ha demostrado que en condiciones muy generales, la distribución de estimadores del tipo citado converge a la normal cuando tiende a infinito el tamaño de la muestra. Por esta razón, en estos casos, puede admitirse que el estimador se distribuye normalmente aunque proceda de observaciones efectuadas en una población base no normal. Sin embargo para admitir la proximidad de la distribución del estimador a la normal, debe tenerse en cuenta que cuanto más alejada esté la población base de dicha distribución, mayor será el tamaño de muestra requerido; en otras palabras, si la convergencia es "lenta" podemos correr el riesgo de estar trabajando con un tamaño de muestra menor del que es necesario para aceptar razonablemente la hipótesis de normalidad del estimador.

En la práctica de poblaciones finitas es habitual encontrarse con poblaciones normales o, al menos con simetría en su distribución de frecuencias por lo que la hipótesis de normalidad de los estimadores del tipo citado, será razonable incluso para tamaños de muestra moderados. Pero también son muy abundantes las poblaciones muy asimétricas, con una fuerte concentración de la frecuencia en valores pequeños y moderados de la variable y una marcada cola hacia la derecha que corresponde a frecuencias bajas de valores altos y muy altos. Así, por ejemplo, la distribución de las explotaciones agrarias según la variable: superficie de las tierras; o las empresas de construcción respecto de la variable: número de empleados; o la distribución de los municipios respecto de la variable: número de habitantes; etc.

Por estas razones la hipótesis de normalidad de los estimadores, debe analizarse antes de construir intervalos de confianza tomando t_α de las tablas N (0,1).

Cochran de la siguiente regla práctica para poblaciones base con asimetría marcadamente positiva (cola hacia la derecha); consiste en tomar:

$$n > 25\ G_1^2$$

donde $G_1 = \mu_3 / \sigma^3$ es el coeficiente de asimetría de Fisher de la población base.

En poblaciones infinitas o muestreo con reposición, se demuestra (véase Cramer) que los momentos centrales de orden 3 y 4 de la distribución de la media muestral, son:

$$\mu_3(\bar{x}) = \frac{\mu_3}{n^2} \quad ; \quad \mu_4(\bar{x}) = \frac{\mu_4 + 3(n-1)\sigma^4}{n^3}$$

siendo μ_3 , μ_4 los de la población base. Entonces, siendo:

$$G_1 = \frac{\mu_3}{\sigma^3} \qquad \text{y} \qquad G_2 = \frac{\mu_4}{\sigma^4} - 3$$

los coeficientes de asimetría y curtosis de la población base, para la media muestral de tamaño n resulta:

$$G_1(\bar{x}) = \frac{\mu_3(\bar{x})}{\sigma^3(\bar{x})} = \frac{G_1}{\sqrt{n}}$$

$$G_2(\bar{x}) = \frac{\mu_4(\bar{x})}{\sigma^4(\bar{x})} = \frac{G_2}{n}$$

por tanto se observa que $G_1(\bar{x})$ tiende a cero al ritmo que aumenta $\sqrt{n}$ y $G_2(\bar{x})$ lo hace según aumenta n. Para la distribución normal, por cumplir $\mu_3 = 0$, $\mu_4 = 3\sigma^4$ se verifica que G_1 y G_2 son cero; por tanto una medida empírica de proximidad a la normal de la distribución de la media muestral viene dada por $G_1(\bar{x})$ y $G_2(\bar{x})$ que deberán ser próximos a cero. Es claro que cuanto mayor sean G_1 y G_2 (población base) mayor tamaño de muestra es necesario para que $G_1(\bar{x})$ y $G_2(\bar{x})$ se acerquen al valor cero.

I.10. EJEMPLOS.

Ejemplo I.1.

Para la población $\mathcal{A} = \{A_1, A_2, A_3\}$ consideremos dos Procedimientos de selección de muestras de tamaño 2: Se extraen al azar dos bolas (*(a)* Con reposición *(b)* Sin reposición) de una urna que contiene sesis bolas, tres con el n° 1, dos con el n° 2 y una con el n° 1.

El espacio muestral $\mathcal{S}$ está constituido por $3^2 = 9$, muestras cuyas probabilidades de selección $\mathcal{P}(a)$ y $\mathcal{P}(b)$ aparecen en el siguiente cuadro:

Espacio muestral $\mathcal{S}$	Probabilidad	
	$\mathcal{P}(a)$	$\mathcal{P}(b)$
(A_1, A_1)	9/36	6/30
(A_1, A_2)	6/36	6/30
(A_1, A_3)	3/36	3/30
(A_2, A_1)	6/36	6/30
(A_2, A_2)	4/36	2/30
(A_2, A_3)	2/36	2/30
(A_3, A_1)	3/36	3/30
(A_3, A_2)	2/36	2/30
(A_3, A_3)	1/36	cero

Ambos procedimientos de muestreo en $\mathcal{A}$, son Con reposición puesto que existen muestras con elementos repetidos que tienen probabilidad no nula. En el procedimiento *(b)* una muestra de $\mathcal{S}$ es imposible, por tanto su probabilidad es cero.

Ejemplo I.2.

Determine el alumno, la distribución de probabilidad de la variable aleatoria

$$T(s) = \text{N}^{\text{o}} \text{ de unidades distintas en } s.$$

en los dos casos *(a)* y *(b)* ejemplo I.1 y calcule su esperanza y varianza.

Ejemplo I.3.

Para la población $\mathcal{A} = \{A_1, A_2, A_3, A_4, A_5\}$ consideremos el siguiente procedimiento de selección de muestras de tamaño 3: De una urna con tres bolas numeradas del 1 al 3 se extraen al azar, sin reposición, dos bolas; a continuación, de otra urna con dos bolas numeradas con el 4 y el 5, se extrae una de éstas.

El conjunto $\mathcal{S}^*$ de muestras posibles se muestra en el cuadro adjunto. Todas tienen la misma probabilidad de selección, que es 1/12. Puede observarse inmediatamente que se trata de un procedimiento Sin reposición puesto que ninguna muestra con unidades repetidas es posible. Por otra parte, de las $5^3 = 75$ muestras de $\mathcal{S}$ solamente son posibles 12; así por ejemplo, (A_3, A_4, A_5) no es posible pues del subconjunto $\{A_4, A_5\}$ solamente se selecciona una unidad; tampoco es posible

Espacio $\mathcal{S}^*$ de muestras posibles	
(A_1, A_2, A_4)	(A_2, A_3, A_4)
(A_1, A_2, A_5)	(A_2, A_3, A_5)
(A_1, A_3, A_4)	(A_3, A_1, A_4)
(A_1, A_3, A_5)	(A_3, A_1, A_5)
(A_2, A_1, A_4)	(A_3, A_2, A_4)
(A_2, A_1, A_5)	(A_3, A_2, A_5)

(A_4, A_1, A_2) puesto que la primera extracción se realiza en el subconjunto $\{A_1, A_2, A_3\}$.

Ejemplo I.4.

Determine el alumno, el conjunto de muestras no ordenadas que constituyen el espacio Ω en el ejemplo I.3.; así como la distribución de probabilidad de la variable aleatoria:

$T(s)$ = Suma de los subíndices de las unidades de s.

Ejemplo I.5.

Para una población constituida por 12 unidades A_i, numeradas de 1 a 12, se considera el siguiente procedimiento de muestreo: Se selecciona un entero, al azar, en el conjunto $\{1, 2, 3, 4\}$ y siendo δ este número, se forma la muestra $(A_\delta, A_{\delta+4}, A_{\delta+8})$. Siendo $X_i = i$, se pide la distribución, esperanza y varianza del estimador.

$\hat{\theta}(\omega)$ = Máximo valor de x_i en ω

Muestras posibles en Ω	$Pr(\omega)$	$\hat{\theta}(\omega)$	$\theta^2(\omega)$
(A_1, A_5, A_9)	1/4	9	81
(A_2, A_6, A_{10})	1/4	10	100
(A_3, A_7, A_{11})	1/4	11	121
(A_4, A_8, A_{12})	1/4	12	144
TOTAL ...	1	42	446

La distribución de $\hat{\theta}\ (\omega)$ y los cálculos necesarios se presentan en el cuadro adjunto. El procedimiento es restringido puesto que de las $\binom{12}{3} = 220$ muestras de Ω , solamente son posibles 4. La esperanza y varianza de $\hat{\theta}\ (\omega)$ son:

$$E(\hat{\theta}) = \Sigma\ \hat{\theta}\ (\omega)\ Pr(\omega) = 42/4 = 21/2$$

$$V(\hat{\theta}) = \Sigma\ \hat{\theta}^2\ (\omega)\ Pr(\omega) - E\ (\hat{\theta})^2 = 446/4 - \frac{21^2}{4} = 5/4$$

Ejemplo I.6.

Con respecto al ejercicio anterior, determine el alumno, el ECM de $\hat{\theta}\ (\omega)$ como estimador de θ = Máximo valor de i en la población. Calcule el ECM del estimador

$\hat{\theta}'\ (\omega) = \dfrac{2\ \Sigma\ X_i}{n} - 1$, y verifique si, para estimar θ, es más acurado que $\hat{\theta}\ (\omega)$.

Ejemplo I.7.

Siendo M_i los tamaños de las unidades de una población finita, consideremos el siguiente procedimiento de muestreo: Se extrae la primera unidad con probabilidad proporcional al tamaño y las $(n-1)$ siguientes, sin reposición, con probabilidades iguales. Probar que la probabilidad de cada muestra $\omega \in \Omega$, es proporcional a la suma de los tamaños de sus unidades.

La probabilidad de una determinada muestra formada por n unidades distintas $\omega = (A_1, A_2, \ldots, A_n)$ es:

$$Pr\ (\omega) = \sum_{j=1}^{n} Pr\ (A_j \text{ la } 1^a)\ Pr\left(\begin{matrix}\text{Restantes unidades en} \\ (n-1)\text{ extracciones}\end{matrix} \Big/ A_j \text{ en la } 1^a\right) =$$

$$= \sum_{j=1}^{n} \frac{M_j}{M} \cdot \frac{1}{\binom{N-1}{n-1}} = \frac{1}{M\binom{N-1}{n-1}} \sum_{j=1}^{n} M_j = K\left(\sum_{j=1}^{n} M_j\right)$$

Ejemplo I.8.

Compruebe que la suma en Ω , de las probabilidades Pr (ω) del ejemplo anterior, es 1.

Ejemplo I.9.

En una población de N = 10 unidades, se encuentran éstas formando 4 subconjuntos, siendo los valores de una característica $\mathfrak{X}$, en cada unidad, dados por la tabla adjunta:

$A_{(1)}$	$A_{(2)}$	$A_{(3)}$	$A_{(4)}$
1; 2; 3	4; 6	9; 11	2; 2; 5

Se considera un procedimiento de muestreo que consiste en elegir con probabilidades iguales un subconjunto $A_{(i)}$. Calcúlese sobre todas las posibles muestras, la varianza de la media muestral y compruebe que es igual a:

$$V(\bar{x}) = E(V(\bar{x} \mid n) + V(E(\bar{x} \mid n))$$

siendo n el tamaño de la muestra. Calcule el sesgo y la varianza de $\bar{x}$ como estimador de $\overline{X}$.

Solución:

$X = 4{,}5$; $E(\bar{x}) = 5$; $B(\bar{x}) = 0{,}5$; $V(\bar{x}) = 9{,}5$

Ejemplo I.10.

Por experiencias anteriores se sabe que un estimador insesgado de θ , tiene un error de muestreo del 1,8 % . Si una muestra proporciona la estimación $\hat{\theta} = 12.508$, determínese el intervalo de confianza para θ, con $\alpha = 0{,}05$, admitiendo la hipótesis de normalidad.

Solución:

$$\{12.508 \pm 1{,}95 \,.\, 0{,}018 \,.\, 12.508\} \equiv \{12.067,\ 12.949\}$$

Ejemplo I.11.

Para el procedimiento de muestreo del ejemplo I.7., calcule la probabilidad π_i de que A_i aparezca en la muestra.

Solución:

$$\pi_i = \Pr(A_i \text{ en } 1^a \text{ extracc.}) + \Pr(A_i \text{ en extracción} \neq \text{de la } 1^a)$$

Entonces, si es $P_i = M_i / M$:

$$\pi_i = P_i + (1 - P_i)\frac{n-1}{N-1} = \frac{P_i(N-n) + (n-1)}{(N-1)}$$

Ejemplo I.12.

Para el procedimiento de muestreo del ejemplo I.7., calcule la probabilidad π_{ij} de que el par de unidades $(A_i,\ A_j)$ aparezca en la muestra.

Solución:

$$\pi_{ij} = \frac{(P_i + P_j)(N-n)(n-1) + (n-1)(n-2)}{(N-1)\ N-2)}$$

Ejemplo I.13.

En los siguientes casos se desea seleccionar un número aleatorio, Z, con probabilidades iguales para todos los del conjunto de referencia. Compruébese si se consigne o no este propósito.

A. Un número Z del conjunto de enteros [1−200]. Método: Se selecciona un número aleatorio, δ, en el conjunto [1−100] y otro número aleatorio, i, en el conjunto [1;2]; si $i = 1$ se toma $Z = \delta$ y si $i = 2$, $Z = 100 + \delta$.

B. Un número Z del conjunto de enteros [1−120]. Método: Se selecciona un número aleatorio, i, en el conjunto [1;2]; si $i = 1$ se elige Z al azar en el conjunto [1 − 100] y si $i = 2$ en el conjunto [101 − 120].

C. Un número Z del conjunto de enteros [1 − 17]. Método: Se selecciona un número aleatorio, δ, del conjunto [1 − 100]; se toma Z = resto de δ (módulo 20) si pertenece al conjunto pedido, en otro caso se repite la operación tantas veces como sea necesario.

D. Un número Z del conjunto de enteros [1 − 46]. Método: Se selecciona un número aleatorio, δ, del conjunto [0; 1; 2; 3; 4] y otro, γ, del conjunto [0; 1; . . . ; 9] si $\delta \neq 4$ y del conjunto [0; 1; . . . ; 6] si $\delta = 4$; por último se toma δ = decenas y γ = unidades de Z.

Solución:

A : Sí ; B : No ; C : Sí ; D : No

Ejemplo I.14.

Consideremos el siguiente procedimiento de muestreo aleatorio sobre una población con tres unidades: $\{A_1, A_2, B\}$:

En una primera extracción se asignan probabilidades iguales a las tres unidades; si se extrae una A no se repone y se efectúa una segunda extracción con probabilidades iguales para las otras dos, y si la extraída es B se repone, asignándole en la segunda extracción doble probabilidad que a cada una de las otras dos.

Fórmese la distribución de probabilidad en el espacio muestral. Calcúlense las distribuciones merginales y las condicionadas.

Solución:

Distribuciones marginales:

1ª Componente	2ª Componente
$Pr(A_1) = 1/3$	$Pr(A_1) = 1/4$
$Pr(A_2) = 1/3$	$Pr(A_2) = 1/4$
$Pr(B) = 1/3$	$Pr(B) = 1/2$

Probabilidad condicional (2^a / 1^a)

	A_1	A_2	B
A_1	0	1/2	1/2
A_2	1/2	0	1/2
B	1/4	1/4	1/2

Las distribuciones marginales no son independientes, ni están igualmente distribuidas.

Ejemplo I.15.

Calcúlese la esperanza matemática de ε para cada una de las tres unidades del ejemplo I.14.

Solución: $E(\varepsilon) = (7/12,\ 7/12,\ 10/12)$

Ejemplo I.16.

Para la situación del ejemplo I.14, se define en la población:

$$X(A_1) = X(A_2) = 1; \qquad X(B) = 0$$

Siendo $T = \sum_{1}^{N=3} X_i$ el número de unidades de tipo A en la población, consideremos el estimador $\hat{T} = K \sum_{i=1}^{n=2} X_i$. Calcúlese el sesgo, la varianza de $\hat{T}$, y el valor de K para que sea insesgado.

Solución:

$$B(\hat{T}) = \frac{7K - 12}{6} \quad ; \qquad V(\hat{T}) = \frac{17}{36} K^2 \quad ; \qquad \hat{T} \text{ es insesgado para T, con } K = 12/7.$$

Ejemplo I.17.

Consideremos en una población de cuatro unidades $\mathcal{A} = \{A_1, A_2, A_3, A_4\}$ definida $\mathfrak{X}$ de la siguiente forma: $X_i = 2i$. Supongamos que se realiza un muestreo aleatorio de tamaño $n = 2$, tal que la probabilidad de cada muestra viene dada por:

$$Pr(A_i, A_j) = 0: \quad \text{para} \quad i = j$$

$$Pr(A_i, A_j) = Pr(A_j, A_i) = \frac{i + j}{60} \quad ; \quad \text{para} \quad i \neq j$$

Calcule las distribuciones marginales de las dos componentes de la muestra; el sesgo y la varianza de la clase de estimadores $\hat{\bar{X}} = K_1 X_1 + K_2 X_2$ para $\bar{X}$. Determine K_1 y K_2 de modo que $\hat{\bar{X}}$ sea insesgado y de mínima varianza.

Solución:

Por ser $Pr(A_i, A_j) = Pr(A_j, A_i)$; para $i \neq j$, las dos componentes tienen la misma distribución:

$$Pr(A_i) = \frac{10 + 2i}{60} \quad ; \quad \text{pra} \quad i = 1, 2, 3, 4.$$

Calculemos $E(X_1)$; $E(X_1^2)$; $E(X_1 X_2)$. Por ser igualmente distribuidas las dos componentes:

$$E(X_1) = E(X_2) = \sum_1^4 2i \left(\frac{10 + 2i}{60}\right) = \frac{16}{3}$$

$$E(X_1^2) = E(X_2^2) = \sum_1^4 4i^2 \left(\frac{10 + 2i}{60}\right) = \frac{100}{3}$$

$$E(X_1 \cdot X_2) = \sum_{i \neq j} (2i \cdot 2_j) \left(\frac{i + j}{60}\right) = \frac{80}{3}$$

Puesto que la media poblacional es $\bar{X} = \frac{\sum_1^4 2i}{4} = 5$, el sesgo de $\hat{\bar{X}}$ es:

$$B(\hat{\bar{X}}) = E(K_1 X_1 + K_2 X_2) - 5 = \frac{16(K_1 + K_2) - 15}{3}$$

La varianza de $\hat{\bar{X}}$ es:

$$V(\hat{\bar{X}}) = V(K_1 X_1 + K_2 X_2) = (K_1^2 + K_2^2) V(X_1) + 2K_1 K_2 \, Cov(X_1, X_2) =$$

$$= \frac{44}{9} (K_1^2 + K_2^2) - \frac{32}{9} K_1 K_2$$

La condición necesaria de varianza mínima, derivando respecto de K_1 y K_2, es: $K_1 = K_2$ y para que sea insesgado la condición es $K_1 + K_2 = 15/16$; por tanto: $K_1 = K_2 = 15/32$.

Ejemplo I.18.

Sean $\mathcal{A}$ y $\mathcal{B}$ dos conjuntos finitos con N y N′. unidades cada uno, tales que $\mathcal{A} \subset \mathcal{B}$; $N < N'$. Supongamos que $\mathcal{T}$ es un procedimiento de selección tal que todas las unidades de $\mathcal{B}$ tienen la misma probabilidad; consideremos ahora el siguiente procedimiento para seleccionar una unidad de $\mathcal{A}$. Se aplica $\mathcal{T}$ a $\mathcal{B}$, si la unidad seleccionada pertenece a $\mathcal{A}$ se acepta, en caso contrario se repite el intento tantas veces como sea necesario hasta obtener una unidad de $\mathcal{A}$.

Calcúlese el número esperado de intentos necesario y pruébese que todas las unidades de $\mathcal{A}$ tienen igual probabilidad.

Solución:

Sea β la probabilidad de que en un intento se obtenga una unidad no perteneciente a $\mathcal{A}$, entonces la probabilidad de obtener una unidad de $\mathcal{A}$ en el r° intento es:

$t_r = \beta^{r-1} \quad (1-\beta)$. Se comprueba inmediatamente que por ser $\beta = \dfrac{N' - N}{N'} < 1$, el proceso termina con la selección de una unidad de A:

$$\Pr(\mathcal{A}) = \sum_{r=1}^{\infty} t_r = (1-\beta)\left\{1 + \beta + \beta^2 + \ldots\right\} = \frac{1-\beta}{1-\beta} = 1$$

El número esperado de intentos es:

$$\sum_{r=1}^{\infty} r\, t_r = (1-\beta)\left\{1 + 2\beta + 3\beta^2 + \ldots\right\} =$$

$$= (1-\beta)\,\frac{d}{d\beta}\left\{\beta + \beta^2 + \beta^3 + \ldots\right\} =$$

$$= (1-\beta)\,\frac{d}{d\beta}\left(\frac{\beta}{1-\beta}\right) = \frac{1}{1-\beta} = \frac{N'}{N}$$

La probabilidad de obtener una unidad de $\mathcal{A}$, sea A_i, es:

$$\sum_{r=1}^{\infty} \Pr(\text{Sea obtenida } A_i \text{ en el r° intento}) = \sum_{r=1} \beta^{r-1} \cdot \frac{1}{N'} =$$

$$= \frac{1}{N'}\,(1 + \beta + \beta^2 + \ldots) = \frac{1}{N'} \cdot \frac{1}{(1-\beta)} = \frac{1}{N}\,.$$

II. MUESTREO ALEATORIO CON PROBABILIDADES IGUALES

Suponemos en este capítulo que todas las muestras formadas por cualesquiera n elementos de la población son posibles y todas tienen la misma probabilidad. Si el procedimiento es sin reposición solamente son posibles las muestras formadas por n elementos distintos entre sí; en consecuencia las distribuciones marginales son iguales pero no independientes; por ello suele denominarse Irrestrictamente Aleatorio. Si el procedimiento es con reposición cualquier sucesión de n elementos de la población, todos distintos o no, es una muestra posible y se denomina Aleatorio Simple; las distribuciones marginales son iguales e independientes. En cualquier caso puede emplearse la denominación de muestreo aleatorio simple, haciendo mención de si es con o sin reposición para que no haya lugar a duda.

II.1. ESTIMADORES INSESGADOS DE LA MEDIA $\bar{X}$ Y EL TOTAL X.

En muestreo aleatorio con probabilidades iguales, sea con o sin reposición, la media muestral:

$$\bar{x} = \frac{1}{n} \sum_{1}^{n} X_i$$

Es un estimador insesgado para la media $\bar{X}$; en efecto: teniendo en cuenta que $E(\varepsilon_i) = n/N$ para todo i ; cap.I.7.

$$E(\bar{x}) = E\left\{\frac{1}{n} \sum_{1}^{N} X_i \varepsilon_i\right\} = \left\{\frac{1}{n} \sum_{1}^{N} X_i E(\varepsilon_i)\right\} = \frac{X}{n} \cdot \frac{n}{N} = \bar{X}$$

Para el total, $X = N\bar{X}$, el estimador insesgado es,

$$\hat{X} = N\bar{x} = \frac{N}{n}\, x\,; \qquad x = \text{total muestral}$$

por ser $E(\hat{X}) = NE(\bar{x}) = N\bar{X} = X$.

Aunque sea redundante vamos a dar otras dos formas de obtener $E(\bar{x})$, para ilustrar otros métodos de cálculo de esperanzas. El primero de ellos se basa en que, como hemos visto (Cap. I.7), las n componentes marginales de la muestra tienen la misma distribución y ésta es:

$$\Pr(A^*_r = A_i) = 1/N\,; \quad \text{para todo } i,\ r$$

por tanto:

$$E(\bar{x}) = E\left(1/n \sum_{j=1}^{n} X_j\right) = 1/n \sum_{j=i}^{n} E(X_j) = 1/n \,.\, n\, E(X_j) =$$

$$= \sum_{i=1}^{N} X_i \,.\, 1/N = \bar{X}$$

El segundo procedimiento consiste en efectuar el razonamiento en el espacio muestral Ω, donde todas las muestras son equiprobables.

$$E(\bar{x}(\omega)) = \sum_{\omega\in\Omega} \bar{x}(\omega)\Pr(\omega) = 1/n \sum_{\omega\in\Omega} x(\omega)\Pr(\omega) =$$

$= 1/n \Pr(\omega) \sum_{\omega\in\Omega} x(\omega)$ y siendo $\binom{N-1}{n-1}$ el número de muestras en que una determinada unidad está presente, resulta:

$$E(\bar{x}) = 1/n \,.\, 1/\binom{N}{n} \sum_{i=1}^{N} X_i \binom{N-1}{n-1} = \frac{\binom{N-1}{n-1}}{n\binom{N}{n}} \sum_{i=1}^{N} X_i = (1/N)X = \bar{X}$$

VARIANZA DE $\bar{x}$ Y DE X.

Teniendo en cuenta los valores de V (ε_i) y Cov (ε_i , ε_j) calculados en I.6., en muestreo con reposición es:

$$V(\bar{x}) = \frac{\sigma^2}{n}$$

En efecto:

$$V(\bar{x}) = V(1/n \sum_1^N X_i \varepsilon_i) = 1/n^2 \left[\sum_1^N X_i^2 V(\varepsilon_i) + \sum_{i \neq j}^N X_i X_j \operatorname{Cov}(\varepsilon_i, \varepsilon_j) \right]$$

$$= 1/n^2 \left[\left(\sum_1^N X_i^2 \right) \frac{n(N-1)}{N^2} - \left(\sum_{i \neq j}^N X_i X_j \right) \frac{n}{N^2} \right] =$$

$$= 1/nN^2 \left[\left(\sum_1^N X_i^2 \right) (N-1) - \sum_{i \neq j}^N X_i X_j \right]$$

utilizando ahora la identidad,

$$\left(\sum_1^N X_i \right)^2 = X^2 = \sum_1^N X_i^2 + \sum_{i \neq j}^N X_i X_j$$

se obtiene:

$$V(\bar{x}) = \frac{1}{n N^2} \left[\left(\sum_1^N X_i^2 \right) (N-1) - X^2 + \sum_1^N X_i^2 \right] =$$

$$= \frac{1}{n N^2} \left[\left(\sum_1^N X^2 \right) N - X_i^2 \right] = \frac{1}{n} \left(\sum_1^N X_i^2 / N - \bar{X}^2 \right) = (1/n)\, \sigma^2 .$$

Teniendo en cuenta los valores de V(ε_i) y Cov (ε_i , ε_j) obtenidos en I.7, para muestreo sin reposición se obtiene:

$V(\bar{x}) = \frac{N-n}{N-1} \frac{\sigma^2}{n}$, que denotando por S^2 la cuasivarianza,

$S^2 = (N / N-1)\ \sigma^2$, admite la forma usual,

$$V(\bar{x}) = \frac{N-n}{N}\ \frac{S^2}{n} \quad \text{, o bien} \quad (1-f)\ \frac{S^2}{n}$$

donde $f = n / N$, se denomina *fracción de muestreo.*

En efecto:

$$V(\bar{x}) = V(1/n \sum_1^N X_i \varepsilon_i) = 1/n^2 \left[\sum_1^N X^2\ V(\varepsilon_i) + \sum_{i \neq j}^N X_i X_j\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) \right] =$$

$$= 1/n^2 \left[\left(\sum_1^N X_i^2 \right) \frac{n\ (N-n)}{N^2} - \left(\sum_{i \neq j}^N X_i X_j \right) \frac{n\ (N-n)}{N^2\ (N-1)} \right] =$$

$$= \frac{(N-n)}{n^2\ N^2\ (N-1)} \left[\left(\sum_1^N X_i^2 \right) n\ (N-1) - \left(\sum_{i \neq j}^N X_i X_j \right) n \right] =$$

$$= \frac{(N-n)}{n^2\ N^2\ (N-1)} \left[\left(\sum_1^N X_i^2 \right) n\ (N-1) + \left(\sum_1^N X_i^2 \right) n - n X^2 \right] =$$

$$= \frac{N-n}{n^2\ N^2\ (N-1)} \left[\left(\sum_1^N X_i^2 \right) n\ N - n\ X^2 \right] =$$

$$= \frac{N-n}{n\ (N-1)} \left(\sum_1^N X_i^2 / N - \bar{X}_i^2 \right) = \frac{N-n}{N-1}\ \frac{\sigma^2}{n} = \frac{N-n}{N}\ \frac{S^2}{n}.$$

Para el estimador del total, teniendo en cuenta que $V(\hat{X}) = V(N\bar{x}) = N^2\ V(\bar{x})$, se obtiene:

$V(\hat{X}) = N^2 \frac{\sigma^2}{n}$; $V(\hat{X}) = N(N-n) \frac{S^2}{n}$, en muestreo con o sin reposición, respectivamente.

II.2. EFICIENCIA DEL MUESTREO SIN REPOSICION

El muestreo sin reposición restringe la selección a muestras formadas por elementos distintos; intuitivamente se deduce que será siempre más eficiente que el muestreo con reposición ya que en este caso existen muestras que teniendo el mismo tamaño presentan información redundante producida por las unidades repetidas. La eficiencia puede medirse comparando la varianza de uno cualquiera de los estimadores $\bar{x}$ ó $\hat{X}$; de acuerdo con los resultados anteriores, podemos escribir:

$$V_{SR} = \frac{N-n}{N-1} V_{CR}$$

por tanto, siempre que $n \geqslant 2$, la varianza en muestreo sin reposición es menor que en muestreo con reposición. En la práctica, si puede admitirse la aproximación $N - 1 \simeq N$, escribimos:

$$V_{SR} = (1-f) V_{CR}$$

donde $(1-f)$ recibe el nombre de *Factor de corrección para poblaciones finitas.* Obviamente la reducción de la varianza es tanto mayor cuanto mayor es la fracción de muestreo $f = n/N$. Más adelante, al hablar del tamaño muestral, insistiremos sobre esta cuestión.

II.3. ESTIMACION DE LAS VARIANZAS

Para establecer un estimador insesgado de las varianzas de $\bar{x}$ y de $\hat{X}$, es suficiente con encontrar un estimador insesgado para σ^2 ó S^2. Si razonamos por analogía y tomamos la varianza muestral s_1^2, en muestreo con reposición se obtiene:

$$E(s_1^2) = E\left\{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{x})^2\right\} = \frac{n-1}{n} \sigma^2$$

y en muestreo sin reposición:

$$E(s_1^2) = \frac{n-1}{n} S^2$$

por tanto, la cuasivarianza muestral

$$s^2 = \frac{n}{n-1} s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

es un estimador insesgado para σ^2 en muestreo con reposición y para S^2 en muestreo sin reposición. En efecto:

$$E(s_1^2) = E\left\{\frac{1}{n} \sum_{1}^{n} (X_i - \bar{x})^2\right\} = \frac{1}{n} E\left\{\sum_{1}^{n} \left[(X_i - \bar{X}) - (\bar{x} - \bar{X})\right]^2\right\} =$$

$$= \frac{1}{n} E\left\{\sum_{1}^{n} (X_i - \bar{X})^2 + \sum_{1}^{n} (\bar{x} - \bar{X})^2 - 2(\bar{x} - \bar{X}) \sum_{1}^{n} (\bar{x}_i - \bar{X})\right\} =$$

$$= \frac{1}{n} \cdot n\, E(X_i - \bar{X})^2 + \frac{1}{n} \cdot n\, E(\bar{x} - \bar{X})^2 - \frac{2n}{n} E(\bar{x} - \bar{X})^2; \text{ por ser}$$

$$\sum_{1}^{n} (X_i - \bar{X}) = n\,(\bar{x} - \bar{X}); \text{ continuemos:}$$

$$= \sigma^2 + V(\bar{x}) - 2V(\bar{x}) = \sigma^2 - V(\bar{x}), \text{ en muestreo con reposición:}$$

$$E(s_1^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2.$$

y en muestreo sin reposición:

$$E(s_1^2) = \sigma^2 - \frac{N-n}{N} \frac{S^2}{n} = \left(\frac{N-1}{N} - \frac{N-n}{Nn}\right) S^2 =$$

$$= \frac{(N-1)\,n - N + n}{Nn} S^2 = \frac{(N-1)\,N}{Nn} = \frac{n-1}{n} S^2$$

Estos resultados nos permiten dar:

$$s^2 = \frac{n}{n-1}\, s_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2 \quad \text{(Cuasivarianza muestral)}$$

como estimador insesgado de σ^2 , en muestreo con reposición y de S^2 en muestreo sin reposición; por tanto:

$$\hat{V}_{CR}(\bar{x}) = \frac{s^2}{n} \; ; \quad \hat{V}_{SR}(\bar{x}) = \frac{N-n}{N}\, \frac{s^2}{n} \, ,$$

son estimadores insesgados de

$\hat{V}(\bar{x})$. Análogamente, $\hat{V}_{CR}(\hat{X}) = N^2 \frac{s^2}{n}$; $\hat{V}_{SR}(\hat{X}) = N(N-n)\frac{s^2}{n}$,

son estimadores insesgados de $\hat{V}(X)$.

II.4. ESTIMACION DE LA PROPORCION P, Y EL TOTAL DE CLASE C. VARIANZA DE LOS ESTIMADORES. ESTIMACION DE LAS VARIANZAS.

Si es C el número de unidades de $\mathcal{A}$ que poseen un cierto atributo; esto es, pertenecen a una determinada clase $\mathcal{C}$ y los (N−C) restantes no poseen el atributo, C se denomina *total de clase* y P = C/N, *proporción.* Denotando por Q = (1−P) la proporción complementaria y teniendo en cuenta que P es la media de los valores

$$X_i = \begin{cases} 1 \text{ si } A_i \in \mathcal{C} \\ 0 \text{ si } A_i \notin \mathcal{C} \end{cases}$$

y C es el total; $C = \sum_{i=1}^{N} X_i$, resulta que la Varianza es $\sigma^2 = PQ$; en efecto:

$$\sigma^2 = \frac{1}{N} \sum_{1}^{N} X_i^2 - \bar{X}^2 = \frac{1}{N} \sum_{1}^{N} X_i^2 - P^2 = P - P^2 = P(1-P) = PQ.$$

por ser $X_i^2 = X_i$ para todo i. La Cuasivarianza es $S^2 = \frac{N}{N-1} PQ$.

Si en la muestra, denotamos por p la proporción y por c el total de clase, la varianza y cuasivarianza muestrales serán,

$$s_1^2 = pq : \qquad s^2 = \frac{n}{n-1} pq : \qquad p = \frac{c}{n}$$

Teniendo en cuenta estas notaciones y de acuerdo con los resultados obtenidos en II.2., para la estimación de $\overline{X}$ y X, ahora identificados con P y C respectivamente, podemos afirmar:

1) p y $\hat{C} = Np = \frac{N}{n} c$, son estimadores insesgados para P y C, respectivamente.

2) $V_{CR}(p) = \frac{PQ}{n}$; $V_{SR}(p) = \frac{N-n}{N-1} \frac{PQ}{n}$

3) $V_{CR}(\hat{C}) = \frac{N^2 PQ}{n}$; $V_{CR}(\hat{C}) = \frac{N^2 (N-n)}{N-1} \frac{PQ}{n}$

4) $s^2 = \frac{n}{n-1} pq$ es insesgado para PQ en muestreo con reposición y para $\frac{N}{N-1}$ PQ en muestreo sin reposición; por tanto los estimadores insesgados de $V(p)$ y $V(\hat{C})$ son:

$$\hat{V}_{CR}(p) = \frac{pq}{n-1} ; \qquad \hat{V}_{SR}(p) = \frac{N-n}{N} \frac{pq}{n-1}$$

$$\hat{V}_{CR}(\hat{C}) = \frac{N^2 pq}{n-1} : \qquad \hat{V}_{SR}(\hat{C}) = N(N-n) \frac{pq}{n-1}$$

Una forma alternativa de obtener los mismos resultados, en muestreo sin reposición, consiste en utilizar la distribución Hipergeométrica:

$$\Pr(c = r) = \frac{\binom{C}{r}\binom{N-C}{n-r}}{\binom{N}{n}} \quad ; \; r = 0, 1, 2, \ldots, n$$

cuya esperanza y varianza son:

$$E(c) = n\,P \; ; \qquad V(c) = \frac{N-n}{N-1}\; n\,PQ$$

En muestreo con reposición el problema corresponde al modelo Binomial:

$$\Pr(c = r) = \binom{n}{r} P^r Q^{n-r} \; ; \quad r = 0, 1, 2, \ldots, n$$

cuya esperanza y varianza son:

$$E(c) = nP \; ; \qquad V(c) = nPQ$$

de donde, teniendo en cuenta que $p = c/n$ y $\hat{C} = Np$, se obtienen los mismos resultados anteriores.

II.5. INTERVALOS DE CONFIANZA PARA MEDIAS Y TOTALES

Si admitimos la hipótesis de normalidad para la distribución de $\bar{x}$; $(\bar{x} - \bar{X}) / \sigma(\bar{x})$ se distribuye según la N(0, 1), por tanto para un α, $0 < \alpha < 1$ prefijado, el intervalo correspondiente a la probabilidad de confianza $(1 - \alpha)$ viene dado por los límites:

$$\left(\bar{x} - t_\alpha\, \sigma(\bar{x}) \; ; \; \bar{x} + t_\alpha\, \sigma(\bar{x})\right)$$

donde t_α se determina en las tablas de la N(0, 1) de modo que:

$$\Pr\left\{ |\, N(0, 1)\, | < t_\alpha \right\} = 1 - \alpha$$

Algunos valores aparecen en el siguiente cuadro:

$1 - \alpha$	0,50	0,682	0,95	0,954	0,99	0,997	0,999
t_α	0,68	1,00	1,96	2,00	2,58	3,00	3,29

Esto nos asegura que el 100 ($1 - \alpha$)% de los intervalos aleatorios que pueden construirse, uno de cada muestra, cubren el valor $\overline{X}$ de la media poblacional que se desea estimar. Después de obtenida una muestra particular, no cabe hablar de probabilidad de que el intervalo construido cubra el valor $\overline{X}$, por esta razón a $(1 - \alpha)$ en vez de probabilidad de confianza, se le denomina coeficiente de confianza.

Para el total X, por ser $\hat{X} = N \overline{x}$ y $\sigma (\hat{X}) = N \sigma (\overline{x})$, el intervalo se calcula de la misma forma multiplicando por N los límites del obtenido para $\overline{x}$.

En cualquier caso ha de tenerse en cuenta el factor de finitud $(1 - f)$ si el muestreo es sin reposición (supuesto que f no sea prácticamente cero).

Si la varianza poblacional σ_x^2 no es conocida y admitimos la hipótesis de normalidad de la población base, puede sustituirse para el cálculo de los límites del intervalo de confianza, σ_x por s_x (cuasivarianza muestral) a condición de tomar t_α en las tablas de la t–Student con $(n - 1)$ grados de libertad. Este procedimiento se justifica inmediatamente ya que:

$\dfrac{(\overline{x} - \overline{X})}{s / \sqrt{n}}$ se distribuye según la $t_{(n-1)}$ y por tanto se puede escribir:

$$\Pr \left\{ | \overline{x} - \overline{X} | < t_{(n-1),\alpha} \ \sigma(\overline{x}) \right\} = 1 - \alpha$$

En cualquier caso, si la distribución de la población base no es normal, la distribución de $\overline{x}$ tenderá a la normal pero debe estudiarse su estado de convergencia según el tamaño de la muestra y la discrepancia de la población base a la normal (En el capítulo I se han hecho algunas consideraciones sobre este particular y más adelante (II.7) se hacen acerca de la estimación de s_x).

II.6. TAMAÑO DE LA MUESTRA PARA UNA PRECISION DADA

Si se fija el error de muestreo, sea E, que se desea para el estimador,

despejamos n de la ecuación $V(\bar{\theta}) = E^2$. Operando de esta forma, para estimación de la media se obtiene:

$$n = \frac{\sigma^2}{E^2} \quad ; \qquad n = \frac{S^2}{E^2 + S^2/N}$$

en muestreo con o sin reposición, respectivamente. Para estimación del total:

$$n = \frac{N^2\ \sigma^2}{E^2} \quad ; \quad n = \frac{N^2\ S^2}{E^2 + NS^2}$$

Para estimar una proporción P o el total de clase C = NP, se obtiene n de las fórmulas anteriores teniendo en cuenta que:

$$\sigma^2 = PQ \ ; \qquad S^2 = \frac{N}{N-1}\ PQ$$

Si se fija el error de muestreo relativo, sea e, denotando por $C_x = \sigma/\bar{X}$, o por $C_{1x} = S/\bar{X}$, al coeficiente de variación poblacional, se obtiene de las fórmulas anteriores:

$$n = \frac{C_x^2}{e^2} \quad ; \qquad n = \frac{C_{1x}^2}{e^2 + C_{1x}/N}$$

en muestreo con o sin reposición respectivamente, siendo válidas estas fórmulas para estimación de la media y del total; teniendo en cuenta que será $e = \sigma(\bar{x})/\bar{X}$ en un caso y $e = \sigma(\hat{X})/X$ en el otro.

Para estimación de P ó de C, se sustituye C_x^2 y C_{1x}^2 por:

$$C_x^2 = \frac{Q}{P} = \frac{1-P}{P} \quad ; \quad C_{1x} = \frac{N}{N-1}\ \frac{Q}{P} = \frac{N\,(1-P)}{(N-1)P}$$

Si se fija un coeficiente de confianza $(1 - \alpha)$ para un error absoluto no superior a un valor dado K; esto es, se fija que:

$$\Pr \left\{ |\bar{x} - \bar{X}| \leqslant K \right\} = 1 - \alpha$$

el tamaño de la muestra se obtiene de la ecuación $t_\alpha \ \sigma(\bar{x}) = K$, o bien de $t_\alpha^2 \ V(\bar{x}) = K^2$, donde t_α es el coeficiente que en la distribución de $(\bar{x} - \bar{X})/\sigma(\bar{x})$, que será $N(0, 1)$ si $\bar{x}$ es Normal, corresponde al intervalo centrado de probabilidad $(1 - \alpha)$; de esta forma al intervalo centrado de semiamplitud $K = t_\alpha \ \sigma(\bar{x})$ corresponde en la distribución de $\bar{x}$ una probabilidad $(1 - \alpha)$. Debe tenerse en cuenta lo dicho en el capítulo I en cuanto a la validez de la hipótesis de normalidad. Operando de esta forma, se obtiene para estimación de la media:

$$n = \frac{t_\alpha^2 \ \sigma^2}{K^2} \quad ; \qquad n = \frac{S^2}{K^2/t_\alpha^2 + S^2/N}$$

en muestreo con y sin reposición respectivamente. Para estimación del total:

$$n = \frac{N^2 \ t_\alpha^2 \ \sigma^2}{N^2} \quad ; \quad n = \frac{N^2 \ S^2}{K^2/t^2 + S^2/N}$$

Aunque en los puntos anteriores hemos dado las fórmulas para muestreo con y sin reposición, la forma práctica de actuar consiste en determinar el tamaño de muestra que correspondería al muestreo con reposición y siendo éste n', el tamaño n para muestreo sin reposición es:

$$n = \frac{n'}{1 + (n' - 1)/N}$$

obtenido igualando las fórmulas de las varianzas con y sin reposición. Si designamos $f' = n'/N$ se obtiene:

$$n = \frac{n'}{(1 + f') - 1/N}$$

donde se observa que para un N (tamaño de la población) dado, la reducción del tamaño muestral es tanto mayor, al emplear muestreo sin reposición, cuanto mayor es f'.

Desde otro punto de vista, recordando lo dicho en cap. II.2, la varianza en muestreo sin reposición para un tamaño de muestra n, es el

$$100\ \frac{N(1-f)}{N-1}\ \%\ <\ 100\ \%$$

de la varianza en muestreo con reposición.

II.7. ESPERANZA Y VARIANZA DE s

Hemos visto en II.1 que s^2 es un estimador insesgado de σ^2 en muestreo con reposición y de S^2 en muestreo sin reposición y por ello se utiliza habitualmente s como estimación de σ ó de S. Es pues conveniente tener alguna idea del comportamiento de s en el muestreo.

La varianza de s^2 en muestreo sin reposición viene dada por una fórmula complicada (puede verse en Sukhatme. 1953) que no resulta manejable; por ello utilizaremos para $V(s^2)$ la que resulta en muestreo con reposición que viene dada por:

$$V(s^2) = \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}$$

donde μ_4 es el momento central de orden cuarto de la población; si ésta es Normal se cumple $\mu_4 = 3\sigma^4$ y en este caso es:

$$V(s^2)_N = \frac{2\sigma^4}{n-1}\ ; \qquad \text{(población normal)}$$

Si expresamos $V(s^2)$ en la forma:

$$V(s^2) = \frac{2\sigma^4}{n-1}\left(1 + \frac{n-1}{2n}\ G_2\right)$$

donde $G_2 = (\mu_4 / \sigma^4) - 3$, es el coeficiente de curtosis de Fisher, se observa que el término $\frac{n-1}{2n}\ G_2$ es el factor que multiplica la varian-

za de s^2 cuando la población no es normal ya que en este caso $G_2 = 0$. Si la distribución de la población no es normal $V(s^2)$ aumenta o disminuye según que G_2 sea positivo o negativo; esto es, que sea más o menos dispersa que la normal. Conviene tener en cuenta que si bien $V(s^2)_N$ disminuye con el tamaño de la muestra, el factor que representa el efecto de no normalidad es prácticamente constante, de modo que si $G_2 > 0$ el incremento porcentual de $V(s^2)$ es independiente del tamaño de la muestra.

La esperanza matemática de s puede obtenerse de la siguiente forma:

Para muestras de tamaño n , por ser $E(s^2) = \sigma^2$, podemos escribir,

$$s^2 = \sigma^2 + e \; ; \qquad E(e) = 0 \; ; \qquad V(e) = V(s^2)$$

y de aquí:

$$s = \sigma \sqrt{1 + e/\sigma^2}$$

donde desarrollando en serie de Mc–Laurin el factor $\sqrt{1 + e/\sigma^2}$ como función de e/σ^2, y despreciando potencias de orden superior a e^2, dado que cuando n aumenta, e disminuye con respecto a σ^2 en términos de probabilidad por ser s^2 estimador insesgado de σ^2, se obtiene:

$$s = \sigma \left(1 + \frac{1}{2} \frac{e}{\sigma^2} - \frac{1}{8} \left(\frac{e}{\sigma^2}\right)^2 + \cdots\right)$$

y tomando esperanza matemática:

$$E(s) = \sigma \left(1 - \frac{V(s^2)}{8\,\sigma^4}\right)$$

Sustituyendo $V(s^2)$ en $E(s)$, se obtiene que el sesgo relativo de s como estimador de σ es:

$$\frac{B(s)}{\sigma} = \frac{1}{4(n-1)} \left(1 + \frac{n-1}{2n} G_2\right)$$

Si la población es normal, $G_2 = 0$, resulta que $(B(s)/\sigma) = 1/4\,(n-1)$, por tanto disminuye con el tamaño de la muestra y no será importante para valores incluso moderados de n. Si la población no es normal, también tiende a cero cuando n crece pero debe tenerse en cuenta el valor de G_2 ya que para reducir $B(s)/\sigma$ al mismo nivel, hará falta un tamaño de muestra al menos $1/2\,G_2$ veces mayor.

La varianza de s puede obtenerse de la siguiente forma, utilizando la fórmula obtenida para $E(s)$:

$$V(s) = E(s^2) - E(s)^2 = \sigma^2 - \sigma^2 \left(1 - \frac{V(s^2)}{8\,\sigma^4}\right)^2 =$$

$$= \sigma^2 - \sigma^2 \left(1 + \frac{V(s^2)^2}{64\,\sigma^8} - \frac{2\,V(s^2)}{8\,\sigma^4}\right) =$$

$$= \left(\frac{2\,V(s^2)}{8\,\sigma^2} - \frac{V(s^2)^2}{64\,\sigma^4}\right);$$ de donde tomando el

primer término del paréntesis, por ser el segundo del orden de 1/17 del cuadrado del primero, resulta:

$$V(s) = \frac{V(s^2)}{4\,\sigma^2}$$, y sustituyendo $V(s^2)$:

$$V(s) = \frac{\sigma^2}{2\,(n-1)} \left(1 + \frac{n-1}{2\,n}\, G_2\right);$$ donde:

$$V(s)_N = \frac{\sigma^2}{2\,(n-1)}$$ es la varianza de s, en caso de población normal.

El incremento que experimenta debido al efecto de no normalidad es análogo al estudiado para el sesgo.

El tamaño de muestra necesario para estimar σ mediante s, puede

determinarse utilizando las fórmulas obtenidas, según los requerimientos de exactitud. Bajo hipótesis de normalidad, se ha construido la tabla adjunta en la que observa que el sesgo puede olvidarse ya que solamente se aprecia su influencia en las milésimas del E.C.M.; sin embargo el error de muestreo relativo de s, no desciende al 10% hasta un tamaño de muestra $n = 50$, si bien $n = 40$ puede ser una buena solución práctica ya que el incremento de precisión, al pasar de $n = 40$ a 50, puede no compensar los costes.

ERROR CUADRATICO MEDIO RELATIVO DE s COMO ESTIMADOR DE σ , BAJO HIPOTESIS DE NORMALIDAD

Tamaño de la muestra	$\frac{B(s)}{\sigma}$	$\frac{\sqrt{V(s)}}{\sigma}$	$\frac{\sqrt{E.C.M.}}{\sigma}$
10	0,0278	0,2357	0,2373
20	0,0132	0,1622	0,1627
30	0,0086	0,1313	0,1316
40	0,0064	0,1132	0,1134
50	0,0051	0,1010	0,1011

Si la población base no es normal, deberá tenerse una estimación de G_2 para introducir las correcciones en los cálculos dadas por las fórmulas estudiadas. Debe tenerse en cuenta el efecto de no normalidad ya que por ejemplo, para un valor moderado de G_2, $G_2 = 4$, el error de muestreo relativo de s, con $n = 40$, pasa a ser 13, 11% que equivale al que se obtiene con $n = 30$ para una distribución normal.

II.8. ESTIMADOR DE LA RAZON. SESGO

Supongamos que están definidas en todas las unidades de $\mathcal{A}$ dos características $\mathfrak{X}$ e $\mathfrak{Y}$ que toman el par de valores (X_i, Y_i) en la unidad A_i. Se denomina *razón* al cociente entre las medias o los totales:

$$R = \frac{\overline{X}}{\overline{Y}} = \frac{X}{Y}$$

En este punto vamos a estudiar el comportamiento de la *razón muestral*

$$\hat{R} = \frac{\bar{x}}{\bar{y}} = \frac{x}{y}$$

como estimador de R, formado por analogía. En primer lugar observemos que se trata de un estimador no lineal, formado por la razón de dos variables aleatorias. En general $\hat{R}$ es un estimador sesgado para R; veamos como se puede calcular su sesgo:

Calculemos Cov $(\hat{R}, \bar{y})$

$$\text{Cov}\ (\hat{R}, \bar{y}) = E(\hat{R}\,\bar{y}) - E(\hat{R})\ E(\bar{y}) = \bar{X} - E(\hat{R})\ \bar{Y}$$

y dividiendo por $\bar{Y}$,

$$\frac{\text{Cov}\ (\hat{R}, \bar{y})}{\bar{Y}} = R - E(\hat{R})$$, por tanto el sesgo es:

$$B(\hat{R}) = E(\hat{R}) - R = \frac{-\text{Cov}\ (\hat{R}, \bar{y})}{\bar{Y}}$$

Se deduce que el sesgo de $\hat{R}$ es cero si y solo si, $\hat{R}$ e $\bar{y}$ son variables incorreladas en el muestreo; esto significa que la recta de regresión de $\hat{R}$ sobre $\bar{y}$ es una horizontal de ordenada R; o bien que para cada valor de $\bar{y}$ fijo el valor esperado de $\hat{R}$ es igual a R más una desviación $e(\bar{y})$ cuya esperanza (al variar $\bar{y}$) es cero.

Si denotamos por $\rho\ (\hat{R}, \bar{y})$ al coeficiente de correlación lineal entre $\hat{R}$ e $\bar{y}$, podemos escribir:

Cov $(\hat{R}, \bar{y}) = -\rho\ (\hat{R}, \bar{y})\ \sigma\ (\hat{R})\ \sigma\ (\bar{y})$, de donde resulta, para el sesgo de $\hat{R}$:

$$B\ (\hat{R}) = -\rho(\hat{R}, \bar{y})\ \ \sigma(\hat{R})\ \frac{\sigma\ (\bar{y})}{\bar{Y}}$$

y en consecuencia, el valor absoluto del sesgo en relación a la desviación

típica:

$$\frac{|B(\hat{R})|}{\sigma(\hat{R})} = |\rho(\hat{R}, \bar{y})| \cdot C_{\bar{y}} \; ; \quad \frac{|B(\hat{R})|}{\sigma(\hat{R})} \leqslant C_{\bar{y}}$$

y por ser $|\rho| \leqslant 1$, se obtiene que es menor o igual que el coeficiente de variación de $\bar{y}$, por tanto

$$\frac{|B(\hat{R})|}{\sigma(\hat{R})} \leqslant \frac{1}{\sqrt{n}} C_y$$

de donde se deduce que el sesgo relativo de $\hat{R}$ será tanto menor cuanto menor sea el coeficiente de variación de los valores poblacionales Y_i y cuanto mayor sea el tamaño de la muestra.

Veamos ahora un procedimiento para obtener una expresión explícita del sesgo, con aproximación hasta términos de orden $1/n$, en función de características poblaciones. Se pretende con ello disponer de una fórmula que permite el análisis del sesgo en relación con características de la población que puedan ser estimadas con una muestra, puesto que su expresión exacta depende de Cov $(\hat{R}, \bar{y})$ para cuya estimación se necesitarían varias muestras.

Sean u y v definidas como sigue:

$$u = \frac{\bar{x} - \bar{X}}{\bar{X}} \; ; \qquad v = \frac{\bar{y} - \bar{Y}}{\bar{Y}} \; , \text{ entonces}$$

$\bar{x} = \bar{X}(1 + u)$; $\bar{y} = Y(1 + v)$ y el sesgo puede escribirse:

$$B(\hat{R}) = E(\hat{R} - R) = E\left\{ R\left(\frac{1 + u}{1 + v} - 1 \right) \right\} = RE\left\{ \frac{1 + u}{1 + v} - 1 \right\}$$

si ahora desarrollamos en serie $1/1 + v$, para lo cual es necesario que $|v| < 1$ (esto es lícito si para el tamaño de muestra, en valor absoluto

$\bar{y}$ se diferencia de $\bar{Y}$ menos del 100%), tenemos:

$$B(\hat{R}) = R\,E(1+n)\,(1-v+v^2\ldots) - 1) = R\,E(-v+v^2+n-nv+nv^2\ldots)$$

y por ser $E(u) = E(v) = 0$, despreciando términos que resultarían en el sesgo de orden $1/n^2$ o superior:

$$B(\hat{R}) = R\left\{E(v^2) - E(uv)\right\} = R(C^2_{\bar{y}} - C_{\bar{x}\,\bar{y}})$$

y de aquí, teniendo en cuenta la definición de u y v:

$$B(\hat{R}) = \frac{R}{n}(C_y^2 - C_{xy})$$

En muestreo sin reposición tendríamos:

$$\hat{B}(R) = \frac{N-n}{N-1} \cdot \frac{R}{n}(C_y^2 - C_{xy})$$

Esta fórmula aproximada del sesgo de $\hat{R}$ puede ponerse de la siguiente forma, en función del coeficiente de regresión lineal de $\mathfrak{X}$ sobre $\mathfrak{Y}$:

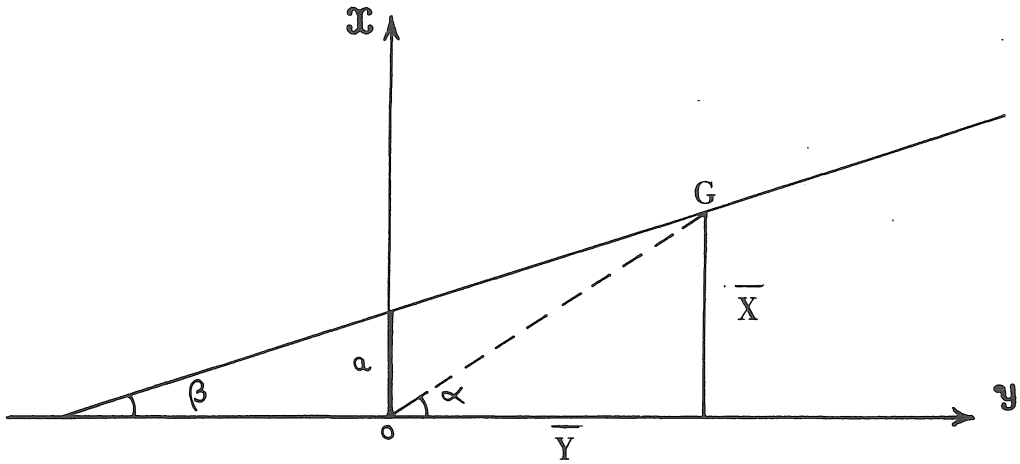
$$B(\hat{R}) = \frac{C_y^2}{n}(R - b)$$

En efecto: el coeficiente de regresión lineal, es

$$b = \frac{C_{xy}\,\bar{X}\,\bar{Y}}{C_y^2\,\bar{Y}^2} = R\,\frac{C_{xy}}{C_y^2}\;;\ \text{y } B(\hat{R}) \text{ puede escribirse como:}$$

$$B(\hat{R}) = \frac{C_y^2}{n}\left(R - \frac{R\,C_{xy}}{C_y^2}\right) = \frac{C_y^2}{n}(R-b)$$

Se obtiene de este resultado que aún no cumpliendose el modelo lineal si la recta de regresión para por el origen, es $R = b$, en consecuencia el sesgo de $\hat{R}$ resulta ser cero con la aproximación dada.



En la figura observamos que por ser $b = \text{tg}\ \beta$, el sesgo puede ponerse en función de la ordenada en el origen "a" de la recta de regresión de $\mathfrak{X}$ sobre $\mathfrak{Y}$ de la siguiente forma:

$$B(\hat{R}) = \frac{C_y^2}{n}(R - b) = \frac{C_y^2}{n}\left(R - \frac{\overline{X} - a}{\overline{Y}}\right)$$

$$B(\hat{R}) = \frac{C_y^2}{n}\frac{a}{\overline{Y}}\ ; \quad \frac{B(\hat{R})}{R} = \frac{C_y^2}{n}\frac{a}{\overline{X}}$$

donde se observa como ya sabemos que el sesgo aproximado es cero si $a = o$, pero también esta aproximación nos da una información de cómo disminuye con la relación $a/\overline{Y}$, o bien con la relación $a/\overline{X}$ en términos de sesgo relativo.

En muestreo sin reposición las fórmulas anteriores para $B(\hat{R})$ vendrían multiplicadas por el factor de finitud $(N-n)/(N-1)$.

Veamos ahora que si (X_i, Y_i) verifica el modelo de regresión lineal y la recta pasa por el origen, $\hat{R}$ es insesgado.

En efecto, en estas condiciones se tiene:

$X_i = b\,Y_i + a + \varepsilon(Y_i)$ donde "b" es el coeficiente de regresión, "a" la ordenada en el origen y $\varepsilon(Y_i)$ es la diferencia entre el valor X_i y el dado por la regresión que cumple:

$$E\left(\varepsilon(Y_i) / Y_i\right) = 0 \; ; \qquad V\left(\varepsilon(Y_i)\right) = \sigma_i^2$$

Si la recta de regresión pasa por el origen, $a = 0$, por tanto $R = b$; esto es, la razón es igual a la pendiente de la recta de regresión; es inmediato ver esto:

$$\overline{X} = \frac{1}{N} \sum \left(b\,Y_i + \varepsilon(Y_i)\right) = b\,\overline{Y} \quad \text{por tanto} \quad \frac{\overline{X}}{\overline{Y}} = b \; .$$

y para ver que es insesgado caculemos la esperanza matemática de:

$$\hat{R} = \frac{b\,\overline{y} + \overline{\varepsilon}}{\overline{y}} = \frac{\overline{X}}{\overline{Y}} + \frac{\overline{\varepsilon}(\overline{y})}{\overline{y}} \; ;$$

$$E(\hat{R}) = R + E\left(\frac{\overline{\varepsilon}(\overline{y})}{\overline{y}}\right) = R + E\left\{E\left(\frac{\overline{\varepsilon}}{\overline{y}} \mid \overline{y}\right)\right\} = R + 0 = R$$

puesto que por ser $E(\varepsilon_i) = 0$ para cada Y_i ; también la media muestral de ε_i, para cada conjunto muestral de n valores Y_i fijos, es cero.

Ilustremos lo dicho con ejemplo sencillo. Si los valores poblacionales vienen dados por la tabla:

Y_i (A_i)	1	2	2	2	3	3	4	4
X_i (A_i)	2	3	4	5	4	8	7	9

se cumple el modelo lineal $X_i = 2\ Y_i + \varepsilon_i$; las desviaciones ε_i vienen dadas por la tabla:

Y_i	1	2	3	4
X_i	0	−1; 0; 1	−2; 2	−1: 1

donde observamos que se cumple $(\bar{\varepsilon}_i \mid Y_i) = 0$ para todo Y_i y lo mismo ocurre para la media $\bar{\varepsilon}$ de las muestras con n valores Y_i fijos.

II.9. ESTIMADOR DE LA RAZON. SU VARIANZA.

En rigor el cálculo que vamos a efectuar es una aproximación del Error Cuadrático Medio de $\hat{R}$; no obstante, si su sesgo es cero obtenemos la varianza y por otra parte si acotado el sesgo resulta ser pequeño, prácticamente la varianza coincidirá con el ECM; así por ejemplo si el sesgo es un 10 % de la desviación típica de $\hat{R}$, la varianza es 0,9901 del ECM. Con esta advertencia, escribimos:

$$V(\hat{R}) = E\left(\frac{\bar{x}}{\bar{y}} - R\right)^2 = E\left(\frac{\bar{x} - R\bar{y}}{\bar{y}}\right)^2$$ si ahora sustituimos $\bar{y}$ por

$\bar{Y}$ en el denominador, tenemos con u, v ya definidas (Pág. 64):

$$\frac{1}{\bar{Y}^2} E\,(\bar{X}u - R\bar{Y}v)^2 = \frac{1}{\bar{Y}^2} E(\bar{X}^2 u^2 + R^2 \bar{Y}^2 v^2 - 2R\bar{X}\bar{Y}uv) =$$

$$= \frac{1}{\bar{Y}^2} (\sigma^2_{\bar{x}} + R^2 \sigma^2_{\bar{x}} - 2R\,\sigma_{\bar{x}\bar{y}}); \text{ por tanto:}$$

$$V(\hat{R}) = \frac{1}{n\,\bar{Y}^2} (\sigma_x^2 + R^2 \sigma_y^2 - 2R\,\sigma_{xy}); \text{ o también}$$

$$V(\hat{R}) \frac{1}{n \bar{Y}^2} (\sigma_x^2 + R^2 \sigma_y^2 - 2 R \rho \sigma_x \sigma_y)$$

siendo ρ el coeficiente de correlación lineal entre $\mathfrak{X}$ e $\mathfrak{Y}$. La varianza relativa, dividiendo por R^2, resulta:

$$\frac{V(\hat{R})}{R^2} = \frac{1}{n} (C_x^2 + C_y^2 - 2 C_{xy})$$

La exactitud de la aproximación obtenida puede estudiarse desarrollando en serie $1 / 1 + \nu$ en la expresión:

$$V(\hat{R}) = E\left(\frac{\bar{x} - R\bar{y}}{\bar{y}}\right)^2 = E\left(\frac{\bar{x} - R\bar{y}}{\bar{Y}(1+\nu)}\right)^2 =$$

$$= \frac{1}{\bar{Y}^2} E\left\{(\bar{x} - R\bar{y})(1 - \nu + \nu^2 \ldots)\right\}^2$$ y quedándonos con el primer término, obtenemos la aproximación calculada antes. Los términos que se desprecian son del orden de $1/n^2$ o mayores. Sukhatme (1954) ha estudiado el problema y obtiene los términos de orden $1/n^2$ en caso de distribución normal bivariante para ($\mathfrak{X}$, $\mathfrak{Y}$).

Para muestreo sin reposición, teniendo en cuenta el factor de finitud, se obtiene inmediatamente:

$$V(\hat{R}) = \frac{(1-f)}{n \bar{Y}^2} (S_x^2 + R^2 S_y^2 - 2 R S_{xy})$$

$$V(\hat{R}) = \frac{(1-f)}{n \bar{Y}^2} (S_x^2 + R^2 S_y^2 - 2 \rho S_x S_y)$$

ESTIMACION DE LA VARIANZA DE $\hat{R}$

Si definimos para cada unidad A_i de la población: $Z_i = X_i - RY_i$ se tiene que $\bar{Z} = \bar{X} - R\,\bar{Y} = 0$, por tanto la varianza de Z_i en la población es,

$$\sigma_z^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - RY_i)^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(X_i - \bar{X}) - R\,(Y_i - \bar{Y}) \right]^2$$

o bien $\sigma_z^2 = \sigma_x^2 + R^2\,\sigma_y^2 - 2R\,\sigma_{xy}$

y en consecuencia se cumple que, para muestreo con reposición, es:

$$V(\hat{R}) = \frac{1}{n\,\bar{Y}^2}\,\sigma_z^2$$

por lo tanto una estimación insesgada de σ_z^2 vendría dada por la cuasi-varianza muestral de los valores Z_i, por tanto

$$\frac{1}{n\,(n-1)\,\bar{Y}^2} \sum_{i=1}^{n} \left[(X_i - \bar{x}) - R\,(Y_i - \bar{y}) \right]^2 =$$

$$= \frac{1}{n\,\bar{Y}^2}\,(s_x^2 + R^2\,s_y^2 - 2R\,s_{xy}) \text{ sería una estimación insesgada de}$$

$V(\hat{R})$ pero aparece el valor de R desconocido; en esta situación la solución natural consiste en sustituir R por su estimación, resultando:

$$\hat{V}(\hat{R}) = \frac{1}{n(n-1)\,\bar{Y}^2} \sum_{1}^{n} (X_i - \hat{R}\,Y_i)^2$$

Para el cálculo práctico pueden ser útiles las formas de $\hat{V}(\hat{R})$:

$$\hat{V}(\hat{R}) = \frac{1}{n\,\bar{Y}^2}\,(s_x^2 + \hat{R}^2\,s_y^2 - 2\hat{R}\,s_{xy})$$

$$\hat{V}(\hat{R}) = \frac{1}{n(n-1)\,\bar{Y}^2}\left(\sum_1^n X_i^2 + \hat{R}^2 \sum_1^n Y_i^2 - 2\hat{R}\sum_1^n X_i Y_i\right)$$

En muestreo sin reposición las anteriores fórmulas deberán multiplicarse por el factor de finitud $(1-f) = (N-n)/N$.

INTERVALOS DE CONFIANZA PARA R

Del sesgo de $\hat{R}$ hemos estudiado en qué condiciones es nulo. Además hemos visto que en cualquier caso es del orden de $1/n$ del coeficiente de variación de $\bar{y}$, respecto a su desviación típica, por lo que incluso para valores moderados de n, su contribución al Error cuadrático medio puede ser despreciable si $C_{\bar{y}}$ es suficientemente pequeño. Se trata pues de un estimador consistente del que no disponemos fórmulas exactas para su sesgo y varianza, sino aproximaciones válidas cuando el tamaño de la muestra proporcione coeficientes de variación $C_{\bar{x}}$, $C_{\bar{y}}$ suficientemente pequeños.

En cuanto a la distribución de $\hat{R}$, bajo condiciones no muy exigentes respecto de la población base, tiende a la Normal; en muestras de tamaño moderado suele presentar cierta asimetría hacia la derecha para las poblaciones que usualmente son objeto de estudio. Una regla práctica consiste en admitir la aproximación normal para $\hat{R}$ cuando es $n \geqslant 30$ y $C_{\bar{x}}$, $C_{\bar{y}}$ ambos menores del 10 %. En este caso el intervalo de confianza se obtiene de la forma explicada en Cap. I.9.

Otro procedimiento consiste en suponer que $(\bar{x}, \bar{y})$ sigue una Normal bivariante de modo que:

$$\bar{Z} = \bar{x} - R\,\bar{y}$$

sigue una distribución Normal de media cero y varianza:

$$V(\bar{Z}) = \frac{1}{n}\,(\sigma_x^2 + R^2\,\sigma_y^2 - 2\,R\,\sigma_{xy})\,,$$

$$V(\bar{Z}) = \frac{N-n}{Nn}\,(S_x^2 + R^2\,S_y^2 - 2\,R\,S_{xy})$$

según se trate de muestreo con o sin reposición. Entonces:

$$\bar{Z}^* = \frac{\bar{x} - R\,\bar{y}}{\sqrt{V(\bar{Z})}}$$

sigue una normal con media cero y desviación típica 1. Si sustituímos S_x^2 , S_y^2 , S_{xy} por sus estimaciones muestrales s_x^2 , s_y^2 , s_{xy} , $\bar{Z}^*$ seguirá una t–Student. En estas condiciones, para un coeficiente de confianza dado $(1 - \alpha)$, se obtienen los límites de confianza para R despejando las dos raíces de la ecuación:

$$\frac{\bar{x} - R\,\bar{y}}{\sqrt{\frac{N-n}{N\,n}\left(s_x^2 + R^2\,s_y^2 - 2\,R\,s_{xy}\right)}} = \pm\, t_\alpha$$

II.10. ESTIMADORES DE RAZON PARA LA MEDIA Y EL TOTAL

Puesto que $R = \bar{X}/\bar{Y}$, la relación $\bar{X} = R\bar{Y}$ sugiere que si $\bar{Y}$ es conocida y en las n unidades de la muestra observamos $\bar{x}$ e $\bar{y}$, podemos formar para la media el estimador de razón:

$$\hat{\bar{X}}_R = \hat{R}\,\bar{Y} \quad ; \qquad \hat{R} = \bar{x}/\bar{y} = x/y$$

y para el total:

$$\hat{X}_R = \hat{R}\;Y$$

El comportamiento de estos estimadores se deduce inmediatamente de las propiedades de $\hat{R}$ como estimador de R, estudiadas anteriormente. Así pues, para $\hat{\bar{X}}_R$ tendremos:

$$E(\hat{\bar{X}}_R) = \bar{Y}\;E\;(\hat{R})$$

$$V(\hat{\bar{X}}_R) = \bar{Y}^2\;V\;(\hat{R})$$

y para $\hat{X}_R$:

$$E(\hat{X}_R) = Y\ E\ (\hat{R}) = N\ E\ (\hat{\bar{X}}_R)$$

$$V(\hat{X}_R) = Y^2\ V(R^2) = N^2\ V(\hat{\bar{X}}_R)$$

pòr lo tanto el estudio del sesgo, la varianza y la distribución aproximada de estos estimadores se reduce al estudio, ya efectuado, de $\hat{R}$. Lo mismo puede decirse de la construcción de intervalos de confianza para $\bar{X}$ ó X mediante $\hat{\bar{X}}_R$ ó $\hat{X}_R$.

El sesgo y la varianza relativos de $\hat{\bar{X}}_R$ y $\hat{X}_R$ coinciden con los de R, ya que:

$$\frac{B(\hat{\bar{X}}_R)}{\bar{X}} = \frac{\bar{Y}\,B(\hat{R})}{\bar{X}} = \frac{B(\hat{R})}{R}$$

$$\frac{V(\hat{\bar{X}}_R)}{\bar{X}^2} = \frac{\bar{Y}^2\,V(\hat{R})}{\bar{X}^2} = \frac{V(\hat{R})}{R^2}$$

y lo mismo ocurre con $\hat{X}_R$.

CONDICIONES EN QUE ES MAS CONVENIENTE EL USO DE ESTIMADORES DE RAZON

Veamos en primer lugar cuando el estimador de razón para la media es mejor que la media muestral de las X_i observadas. Para que sea

$V(\hat{\bar{X}}_R) < V(\bar{x})$, debe cumplirse:

$$\frac{(1-f)}{n}\ (S_x^2 + R^2\ S_y^2 - 2\ R\ S_x\ S_y) < \frac{(1-f)}{n}\ S_x^2$$

de donde simplificando, se obtiene:

$$R^2 S_y^2 < 2 \rho S_x S_y : \text{esto es} \quad \rho > \frac{1}{2} \frac{C_y}{C_x}$$

De este resultado se deduce que el estimador de razón, puede ser más o menos acurado que la media $\bar{x}$ dependiendo de los valores del coeficiente de correlación lineal ρ y de la relación entre los coeficientes de variación de $\mathfrak{X}$ y de $\mathfrak{Y}$. No debemos olvidar que $V(\hat{\bar{X}}_R)$ es una fórmula aproximada y que $\hat{\bar{X}}_R$ es sesgado, por tanto la aplicación de este resultado debe tener en cuenta lo dicho anteriormente acerca del tamaño muestral y la recta de regresión, para garantizar que estamos trabajando con una buena aproximación del Error cuadrático medio de $\hat{R}$.

Un resultado más preciso es el siguiente (David y Neyman, 1938). En muestreo con probabilidades iguales el estimador de razón de $\bar{X}$, es el mejor estimador lineal insesgado, para Y_i dadas, si se cumplen las condiciones siguientes:

1) La línea de regresión de $\mathfrak{X}$ sobre $\mathfrak{Y}$ es una recta pasando por el origen.

$$X_i = B Y_i + e_i \text{ para cada } Y_i \text{ con } E(e_i) = 0$$

2) La varianza de X_i alrededor de esta línea es proporcional a Y_i.

$$V(e_i) = \lambda Y_i$$

Es importante destacar la segunda condición; exige que la varianza de X_i para Y_i fijo, aumente proporcionalmente a medida que lo hace Y_i. Puede ocurrir que $V(X_i | Y_i) = V(e_i)$ no siga este comportamiento aún cumpliéndose la primera condición; en este caso se pierde la optimidad del estimador de razón. Así por ejemplo si $V(e_i) = \lambda Y_i^2$ se encuentra que el estimador óptimo es:

$$\left(\frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{Y_i} \right) \cdot \bar{Y}$$

que toma como multiplicador de $\bar{Y}$, la media de las razones en vez de la razón muestral.

II.11. ESTIMACION DE MEDIAS Y TOTALES EN SUBPOBLACIONES

En ocasiones se desea estimar la media de una característica referida a una subpoblación $\mathcal{A}_o \subset \mathcal{A}$, constituida por $N_o < N$ unidades, cuando se ha extraído una muestra aleatoria de tamaño n sobre toda la población $\mathcal{A}$. Para comodidad en las notaciones vamos a definir en cada unidad A_i las variables (X_i, Y_i) de la siguiente forma:

$$X_i = \begin{cases} X_i & \text{si} \quad A_i \in \mathcal{A}_o \\ 0 & \text{si} \quad A_i \notin \mathcal{A}_o \end{cases} \qquad Y_i = \begin{cases} 1 & \text{si} \quad A_i \in \mathcal{A}_o \\ 0 & \text{si} \quad A_i \notin \mathcal{A}_o \end{cases}$$

de manera que la media condicionada a $\mathcal{A}_o$, puede expresarse como la razón de X e Y:

$$\bar{X}_o = \frac{X_o}{N_o} = \frac{\sum_1^N X_i}{\sum_1^N Y_i}$$

teniendo en cuenta que los totales en $\mathcal{A}_o$ y en $\mathcal{A}$ coinciden, ya que hacemos cero X_i para cada A_i no perteneciente a $\mathcal{A}_o$. Denotaremos por P la proporción de unidades en la población que pertenecen a la subpoblación que nos interesa y por Q, la proporción complementaria:

$$P = \frac{N_o}{N} \; ; \; Q = 1 - P.$$

SE DESCONOCE N_o

Si no es conocido el número N_o de unidades en $\mathcal{A}_o$, la media muestral de las n_o unidades que han caido en la subpoblación:

$$\bar{x}_o = \frac{x_o}{n_o} = \frac{\sum_1^n X_i}{\sum_1^n Y_i}$$

es el estimador de $\bar{X}_o$ que nos proponemos. Presenta un primer problema y es que para aquellas muestras en que sea $n_o = 0$ (Todas las unidades han caido fuera de $\mathcal{A}_o$) no está definido; supongamos entonces que el procedimiento de muestreo es tal que todas las posibles muestras tienen al menos un elemento de $\mathcal{A}_o$. Esto puede conseguirse rechazando la muestra si $n_o = 0$ y extrayendo otra hasta obtener una en la cual $n_o \geqslant 1$. Si no es este el procedimiento empleado y se hace uso del estimador cuando n_o no es cero, en rigor no es éste un procedimiento de estimación ya que el estimador no está definido para algunas muestras; no obstante los resultados teóricos que vamos a obtener se aplican con suficiente aproximación si para los tamaños N, N_o y n la probabilidad de que sea $n_o = 0$ es muy pequeña.

En resúmen, el modelo supone la extracción de muestras de tamaño n en $\mathcal{A}$ de modo que al menos una unidad pertenece a $\mathcal{A}_o$; por lo tanto el estimador debería denotarse por $(\bar{x}_o \mid n_o \geqslant 1)$, pero para simplificar la notación usaremos la habitual.

En estas condiciones $\bar{x}_o$ es insesgado para $\bar{X}_o$. En efecto, n_o puede tomar los valores 1, 2, ... , K ; K = Min. (N_o , n) y para cualquier n_o fijo, la esperanza de $\bar{x}_o$ es $\bar{X}_o$, ya que la media de una muestra es insesgada para la media de la población cualquiera que sea su tamaño, y como por otra parte, la suma de las probabilidades $Pr(n_o)$ para $n_o \geqslant 1$ es 1 ya que hemos condicionado el procedimiento a que sea $n_o \geqslant 1$, resulta:

$$E(\bar{x}_o) = E\left(E(\bar{x}_o \mid n_o)\right) = E(\bar{X}_o) = \bar{X}_o$$

Vemos ahora el cálculo de la varianza de $\bar{x}_o$:

$$V(\bar{x}_o) = E\ V(\bar{x}_o \mid n_o) + V\ E(\bar{x}_o \mid n_o)$$

$= E \; V(\bar{x}_o / n_o)$ por ser cero el segundo sumando ya que la esperanza de $\bar{x}_o$ condicionada por n_o es constante ($= \bar{X}_o$). En muestreo con reposición para un n_o dado, es:

$$V(x_o / n_o) = \frac{\sigma^2_{ox}}{n_o}$$, donde σ^2_{ox} representa la varianza de los N_o valores X_i de la subpoblación $\mathcal{A}_o$, y es:

$$V(\bar{x}_o / n_o) = (1 - f_o) \frac{S^2_{ox}}{n_o}$$ en muestreo sin reposición.

El procedimiento de estimación recomendado consiste en despreciar la variabilidad de n_o y actuar como si se hubiese tomado una muestra de tamaño n_o en $\mathcal{A}_o$. Entonces se admite:

$$V(\bar{x}_o)_{CR} = \frac{\sigma^2_{ox}}{n_o} \quad ; \quad V(\bar{x}_o)_{SR} = (1 - f) \frac{S^2_{ox}}{n_o}$$

y como estimadores de la varianza:

$$\hat{V}(\bar{x}_o)_{CR} = \frac{s^2_{ox}}{n_o} = \frac{1}{n_o(n_o - 1)} \sum_{1}^{n_o} (X_i - \bar{x}_o)^2$$

$$\hat{V}(\bar{x}_o)_{SR} = (1 - f_o) \frac{s^2_{ox}}{n_o} = \frac{(1 - f_o)}{n_o(n_o - 1)} \sum_{1}^{n_o} (X_i - \bar{x}_o)^2$$

poniendo en esta última $f = n / N$ como estimación insesgada de $f_o = n_o / N_o$.

Esta solución puede estar lejos de la realidad si el tamaño de la muestra n y la proporción de unidades de $\mathcal{A}_o$ en la población, son pequeñas. El estimador es insesgado pero su varianza es mayor que la admitida antes, si nP es pequeño. Vemos entonces en qué condiciones puede aceptarse el procedimiento recomendado.

Si definimos:

$$v = \frac{n_o - \bar{n}_o}{\bar{n}_o}, \quad \text{siendo} \quad \bar{n}_o = E(\bar{n}_o) \text{ el valor esperado de unidades}$$

de $\mathcal{A}_o$ en la muestra, podemos escribir la varianza de $\bar{x}_o$ de la siguiente forma:

$$V(\bar{x}_o)_{CR} = \sigma^2_{ox} \; E\left(\frac{1}{n_o}\right) = \sigma^2_{ox} \; E\left(\frac{1}{\bar{n}_o\,(1+v)}\right) =$$

$$= \frac{\sigma^2_{ox}}{\bar{n}_o} E(1 - v + v^2 - v^3 + \ldots) \text{ y dado que}$$

$$E(v) = 0; \; E(v^2) = \frac{1}{\bar{n}^2_o} \; V(n_o) \text{ despreciando términos a partir de}$$

$E(v^3)$, tenemos:

$$V(\bar{x}_o)_{CR} \simeq \frac{\sigma^2_{ox}}{\bar{n}_o}\left(1 + \frac{V(n_o)}{\bar{n}^2_o}\right)$$

Veamos ahora cómo se puede calcular el segundo término entre paréntesis: Siendo $P = N_o / N$ la proporción de unidades de la subpoblación $\mathcal{A}_o$, se tiene:

$$Pr(n_o = 0) = Q^n \; ; \quad Pr(n_o \geqslant 1) = 1 - Q^n, \text{ siendo } Q = 1 - P.$$

$$\bar{n}_o = E(n_o / n_o \geqslant 1) = \frac{n\,P - 0\,.\,Pr\;(n_o = 0)}{1 - Q^n} = \frac{n\,P}{1 - Q^n}$$

$$V(n_o / n_o \geqslant 1) = \frac{n\,P\,Q - \bar{n}^2\,Q^n}{1 - Q^n} \text{ sustituyendo en } V(\bar{x}_o)_{CR}$$

se obtiene:

$$V(\bar{x}_o)_{CR} = \frac{\sigma^2_{ox}}{\bar{n}_o} \left\{ 1 + \frac{Q\,(1 - Q^n)}{n\,P} - \frac{Q^n}{1 - Q^n} \right\}$$

Para valores altos de n y no excesivamente pequeños de P, resulta que Q^n (probabilidad de $n_o = 0$) es prácticamente cero, y queda:

$$V(\bar{x})_{CR} = \frac{\sigma^2_{ox}}{\bar{n}_o} \left(1 + \frac{Q}{n\,P} \right) ; V(\bar{x})_{SR} \simeq (1-f)\ \frac{\sigma^2_{ox}}{\bar{n}_o} \left(1 + \frac{(1-f)\,Q}{n\,P} \right)$$

En resumen, el incremento porcentual de la varianza debido a la variabilidad del número de unidades de la subpoblación que caen en la muestra, decrece con relativa rapidez con $n\,P$. Las fórmulas obtenidas permiten en cada caso particular, tomar una decisión en cuanto a la validez de la aproximación σ^2_{ox}/n_o; como ilustración mencionemos que para $n\,P = 20$, con $P = 0,\ 1$, el incremento de la varianza de $\bar{x}_o$ es un 4,5 %. Así pues, el procedimiento y la fórmula condicional de la varianza requieren un valor $n\,P$ no pequeño.

Para estimar el total X_o en la subpoblación el estimador usual:

$$\hat{X}_o = \frac{N}{n}\, x_o = \frac{N}{n}\, x$$

no presenta el problema de $\bar{x}_o$ en cuanto a que siempre está definido ya que no interviene n_o en su denominador y si $n_o = 0$, es $x_o = 0$ y por tanto $\hat{X}_o$ toma el valor definido cero. Es también un estimador insesgado, como se comprueba inmediatamente:

$$E(\hat{X}_o) = \frac{N}{n}\, E\left(\sum_1^n X_i \right) = \frac{N}{n}\, E\left(\sum_1^N X_i\ \varepsilon_i \right) =$$

$$= \frac{N}{n} \sum_1^N X_i\ \frac{n}{N} = \sum_1^N X_i = X = X_o$$

dando, como ya dijimos, el valor $X_i = 0$ a las unidades no pertenecientes a $\mathcal{A}_o$.

Su varianza se obtiene también de forma inmediata, basándose en teoría ya conocida,

$$V(\hat{X}_o)_{CR} = \frac{N^2}{n} \sigma_x^2 \quad ; \quad C(\hat{X}_o) = \frac{N^2 (1-f)}{n} S_x^2$$

donde es conveniente observar, a diferencia del estimador condicional de la media, que interviene la varianza o cuasivarianza de las N unidades de la población (con $X_i = 0$ si $A_i \notin \mathcal{A}_o$).

Estimadores insesgados de la varianza, son:

$$\hat{V}(\hat{X}_o)_{CR} = \frac{N^2}{n(n-1)} \sum_1^n \left(X_i - \frac{x_o}{n} \right)^2 = \frac{N^2}{n(n-1)} \left\{ \sum_{i=1}^{n_o} X_i^2 - \frac{x_o^2}{n} \right\}$$

$$\hat{V}(\hat{X}_o)_{SR} = \frac{N^2 (1-f)}{n(n-1)} \sum_1^n \left(X_i - \frac{x_o}{n} \right)^2 = \frac{N^2 (1-f)}{n(n-1)} \left\{ \sum_{i=1}^{n_o} X_i^2 - \frac{x_o^2}{n} \right\}$$

donde debemos observar que la suma se extiende a las n unidades de la muestra completa y las diferencias cuadráticas lo son respecto de x_o/n y no de $\bar{x}_o = x_o/n_o$.

En resumen el estimador del total $\hat{X}_o$ es insesgado y su varianza viene dada de forma exacta por las fórmulas anteriores. Esto no significa que el procedimiento sea equivalente, en precisión, a extraer nP unidades muestrales en $\mathcal{A}_o$; en efecto, si es V_2 la varianza de $\hat{X}_o$ según el procedimiento estudiado y V_1 la varianza si se hubiera efectuado directamente el muestreo en $\mathcal{A}_o$ (con $n_o = n$P), se tiene:

$$V_1 = \frac{N^2}{n} P \sigma_{ox}^2 \quad ; \quad V_2 = \frac{N^2}{n} (P \sigma_{ox}^2 + PQ \bar{X}_o^2)$$

Si comparamos por cociente estas varianzas, se obtiene:

$$V_2 = V_1 \left(1 + \frac{Q}{C_{ox}^2}\right) \text{ donde } C_{ox}^2 = \frac{\sigma_{ox}}{\bar{X}_o} \text{ es el coeficiente de}$$

variación de $\mathfrak{X}$ dentro de la subpoblación; por tanto la pérdida de precisión será tanto mayor cuanto menor sea el tamaño relativo de la subpoblación ($P = 1 - Q$) y cuanto menor sea C_{ox}^2 . Otra forma de efectuar la comparación, quizá más práctica es por diferencia:

por ser $V_2 = V_1 + \frac{Q\, X_o^2}{n\, P}$, en términos de varianza relativa, resulta:

$$\frac{V_2}{X_o^2} = \frac{V_1}{X_o^2} + \frac{Q}{n\ P}$$

donde se observa que el incremento de varianza relativa será pequeño si es nP suficientemente grande; por el contrario si P es pequeña y n no es lo suficientemente grande, puede notarse el incremento, así por ejemplo para $P = 0{,}1$ y $n = 100$, la varianza relativa del estimador aumenta en la cantidad 0,09. El análisis ha sido efectuado mediante la siguiente descomposición σ_x^2 :

$$\sigma_x^2 = \frac{1}{N} \sum_{1}^{N_o} (X_i - \bar{X})^2 + \frac{1}{N} \sum_{N_o}^{N} (0 - \bar{X})^2 =$$

$$= \frac{1}{N} \sum_{1}^{N_o} (X_i - \bar{X}_o)^2 + \frac{1}{N} \sum_{1}^{N_o} (\bar{X}_o - \bar{X})^2 - Q\bar{X}^2$$

y teniendo en cuenta que $\bar{X}_o P = \bar{X}$, resulta:

$$\sigma_x^2 = P\, \sigma_{ox} + PQ\, \bar{X}_o^2$$

SE CONOCE N_o

Si N_o es conocido el estimador estudiado para la media en $\mathcal{A}_o$, $\bar{X}_o$; puede formarse evitando su indefinición cuando $n_o = 0$ y obtenerse su varianza exacta. En efecto, por ser $\hat{X}_o = \frac{N}{n} x_o$ estimador insesgado de X_o, al conocer N_o, aunque el muestreo se efectúe sobre $\mathcal{A}$, resulta que:

$$\hat{\bar{X}}_o = \frac{N x_o}{N_o n}$$

es insesgado para la media condicional $\bar{X}_o$. Su varianza viene dada por:

$$V(\hat{\bar{X}}_o)_{CR} = \frac{1}{N_o^2} V(\hat{X}_o)_{CR} = \frac{\sigma_x^2}{n P^2}$$

$$V(\hat{\bar{X}})_{SR} = \frac{(1-f)}{n P^2} S_x^2$$

cuyo estimador insesgado, según sea el muestreo con o sin reposición, es:

$$\hat{V}(\hat{\bar{X}})_{CR} = \frac{1}{P^2 n (n-1)} \sum_1^n \left(X_i - \frac{x_o}{n} \right)^2$$

$$\hat{V}(\hat{\bar{X}})_{SR} = \frac{(1-f)}{P^2 n (n-1)} \sum_1^n \left(X_i - \frac{x_o}{n} \right)^2$$

II.12. ESTIMACION DE LA DIFERENCIA DE MEDIAS

Si en cada unidad de la población está definido el par de valores (X_i, Y_i) correspondientes a dos características $\mathfrak{X}$ e $\mathfrak{Y}$, la diferencia de

medias viene dada por:

$$\bar{D} = \bar{X} - \bar{Y} = \frac{1}{N} \sum_{1}^{N} (X_i - Y_i)$$; entonces si denotamos por $Z_i = X_i - Y_i$, la diferencia en cada unidad, la varianza de Z_i viene dada por:

$$\sigma_z^2 = \frac{1}{N} \sum_{1}^{N} (Z_i - \bar{D})^2 = \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}$$

Utilizando, como siempre, letras minúsculas para las características muestrales, se obtiene que la diferencia de medias en la muestra:

$$\bar{d} = \bar{x} - \bar{y} = \frac{1}{n} \sum_{1}^{n} (X_i - Y_i) = \frac{1}{n} \sum_{1}^{n} Z_i$$

es un estimador insesgado de $\bar{D}$. Es consecuencia inmediata de que $E(\bar{x}) = \bar{X}$ y $E(\bar{y}) = \bar{Y}$.

La varianza de "$\bar{d}$" se obtiene también de forma inmediata:

$$V(\bar{d})_{CR} = V\left(\frac{1}{n} \sum_{1}^{n} Z_i\right) = \frac{1}{n^2} V(Z_i) = \frac{\sigma_z^2}{n}$$

en muestreo con reposición y:

$$V(\bar{d})_{SR} = (1-f) \frac{S_z^2}{n}$$ en muestreo sin reposición.

Teniendo en cuenta la descomposición de σ_z^2, resulta:

$$V(\bar{d})_{CR} = \frac{1}{n} (\sigma_x^2 + \sigma_y^2 - 2\sigma_{xy})$$

$$V(\bar{d})_{SR} = \frac{(1-f)}{n} (S_x^2 + S_y^2 - 2S_{xy})$$

Se observa que si $\mathfrak{X}$ e $\mathfrak{Y}$ están incorrelacionadas, $\sigma_{xy} = 0$, quedando la varianza de "$\bar{d}$" como $1/n$ de la suma de las varianzas de $\mathfrak{X}$ e $\mathfrak{Y}$. Si la correlación es positiva, la varianza de "$\bar{d}$" es menor y si la correlación es negativa la varianza es mayor.

La varianza de "$\bar{d}$" se estima de la forma habitual:

$$\hat{V}(\bar{d})_{CR} = \frac{1}{n} s_z^2 \quad ; \qquad \hat{V}(\bar{d})_{SR} = \frac{(1-f)}{n} s_z^2$$

son insesgados para $V(\bar{d})$, siendo:

$$s_z^2 = \frac{1}{n-1} \sum_1^n (Z_i - \bar{d})^2 = s_x^2 + s_y^2 - 2 s_{xy}$$

El problema de la estimación de la diferencia de totales (X – Y), se resuelve inmediatamente multiplicando el estimador de la diferencia de medias por N y su varianza por N^2.

Un caso muy interesante y que tiene diversas aplicaciones consiste en la estimación de la diferencia de dos proporciones. Conviene considerar la situación general en que las clases no sean disjuntas. Sean pues $\mathfrak{C}_x$ y $\mathfrak{C}_y$ dos clases de unidades en la población, con proporciones P_x y P_y que tienen N_o elementos comunes.

Si definimos las variables:

$$X_i = \begin{cases} 1 \text{ si } & A_i \in \mathfrak{C}_x \\ 0 \text{ si } & A_i \notin \mathfrak{C}_x \end{cases} \qquad Y_i = \begin{cases} 1 \text{ si } & A_i \in \mathfrak{C}_y \\ 0 \text{ si } & A_i \notin \mathfrak{C}_y \end{cases}$$

la diferencia de proporciones es la diferencia de medias $\bar{X} - \bar{Y}$:

$$P_x - P_y = \frac{\sum_1^N X_i}{N} - \frac{\sum_1^N Y_i}{N}$$

por lo tanto $\sigma_x^2 = P_x Q_x$; $\sigma_y^2 = P_y Q_y$. Veamos ahora el cálculo de la covarianza σ_{xy} : Teniendo en cuenta que el producto $X_i Y_i$ es 1 en la parte común a las dos clases y cero en el resto, resulta:

$$\sigma_{xy} = \frac{1}{N} \sum_{1}^{N} X_i Y_i - \bar{X}\bar{Y} = \frac{N_o}{N} - P_x P_y = P_o - P_x P_y$$

siendo P_o la proporción en la población de elementos comunes a las dos clases.

De acuerdo con los resultados obtenidos para la estimación de la diferencia de medias, podemos afirmar, que:

$p_x - p_y$ es un estimador insesgado de $P_x - P_y$, cuya varianza es:

$$V_{CR}(p_x - p_y) = \frac{1}{n} \left\{ P_x Q_x + P_y Q_y - 2(P_o - P_x P_y) \right\}$$

$$V_{SR}(p_x - p_y) = \frac{N-n}{(N-1)n} \left\{ P_x Q_x + P_y Q_y - 2(P_o - P_x P_y) \right\}$$

según sea con o sin reposición el el procedimiento de muestreo. En este segundo caso también se puede escribir:

$$V_{SR}(p_x - p_y) = \frac{(1-f)}{n} \cdot \frac{N}{N-1} \left\{ P_x P_x + P_y Q_y - 2(P_o - P_x P_y) \right\}$$

La estimación insesgada de estas varianzas viene dada por:

$$\hat{V}_{CR}(p_x - p_y) = \frac{1}{(n-1)} \left\{ p_x q_x + p_y q_y - 2(P_o - p_x p_y) \right\}$$

$$\hat{V}_{SR}(p_x - p_y) = \frac{(1-f)}{(n-1)} \left\{ p_x q_x + p_y q_y - 2(P_o - p_x p_y) \right\}$$

donde, si P_o no es conocido puede sustituirse por la proporción muestral p_o unidades comunes a ambas clases.

El caso más corriente en la práctica corresponde a clases $\mathcal{C}_x, \mathcal{C}_y$ disjuntas; por ejemplo, votantes a dos alternativas distintas en unas elecciones. En esta situación N_o es cero y por tanto también $P_o = 0$, y en consecuencia

$$\sigma_{xy} = - P_x P_y$$

de donde resulta para la varianza:

$$V_{CR}(p_x - p_y) = \frac{1}{n} (P_x Q_x + P_y Q_y + 2 P_x P_y)$$

$$V_{SR}(p_x - p_y) = \frac{(1-f)}{n} \frac{N}{(N-1)} (P_x Q_x + P_y Q_y + 2P_x P_y)$$

y su estimación insesgada:

$$\hat{V}_{CR}(p_x - p_y) = \frac{1}{(n-1)} (p_x q_x + p_y q_y + 2p_x p_y)$$

$$\hat{V}_{SR}(p_x - p_y) = \frac{(1-f)}{(n-1)} (p_x q_x + p_y q_y + 2 p_x p_y)$$

II.13. EJEMPLOS

Ejemplo II.1

De una población con N = 100 unidades se ha extraido una muestra sin reposición, de tamaño $n = 8$. Con los siguientes datos muestrales; estimar el total y su error de muestreo. Datos: 25, 32, 28, 35, 26, 34, 30, 28.

Solución:

$$\hat{X} = 2975 \; ; \qquad \hat{\sigma}(\hat{X}) = 124$$

Ejemplo II.2.

Basándose en la muestra del ejemplo II.1, determínese el tamaño de muestra necesario para que el error de muestreo sea 50.

Solución:

Tomando $s^2 = 13{,}3568$ como estimación de S^2, se obtiene, $n = 35$.

Ejemplo II.3.

Determínese el tamaño de muestra, en función de N y P, para estimar el total de clase con un error de muestreo relativo del 5 % . Aplicación al caso P = 0,26 ; N = 2000. (Supóngase $N - 1 \simeq N$ y muestreo sin reposición).

Solución:

$$n = \frac{NQ}{0{,}0025\ NP + Q} \; ; \text{ aplicación: } n = 726$$

Ejemplo II.4.

Determínese el tamaño de la muestra necesario para estimar el total X, con un error de muestreo relativo del 5% , en caso de muestreo con y sin reposición. Datos N = 500 ; C_x = 0,60.

Solución:

Con reposición n' = 144 ; Sin reposición n = 112

Ejemplo II.5.

Una muestra sin reposición de tamaño n = 100, procedente de una población de N = 2000 unidades presenta los siguientes datos:

$$\sum_{1}^{100} X_i = 273000 \; ; \quad \sum_{1}^{100} X_i^2 = 7532500$$

Fórmese el intervalo de confianza para el total X, con α = 0,05, admitiendo normalidad para la distribución del estimador.

Solución:

$$X : \{543.532 \; ; \; 548.511\}$$

Ejemplo II.6.

Una muestra aleatoria sin reposición, de tamaño n = 50, procedente de una población $\mathcal{A}$ de N = 750 unidades, presenta los siguientes datos:

$$\sum_{1}^{50} X_i = 454 \; ; \quad \sum_{1}^{50} X_i^2 = 4.306$$

de esta muestra, n_o = 20 unidades pertenecen a una subpoblación $\mathcal{A}_o$ y sus datos son:

$$\sum_{1}^{20} X_i = 172 \; ; \quad \sum_{1}^{20} X_i^2 = 1.536$$

Estímese la media y el total de la población $\mathcal{A}$ y de la subpoblación $\mathcal{A}_o$, así como el error de muestreo.

Solución:

Para la población:

$$\left.\begin{array}{ll} \hat{\bar{X}} = 9{,}08 & \hat{\sigma}(\hat{\bar{X}}) = 0{,}193 \\ \hat{X} = 6.810: & \hat{\sigma}(\hat{X}) = 145 \end{array}\right\} \simeq 2.1\%$$

Para la subpoblación $\mathcal{A}_o$:

$$\hat{\bar{X}}_o = 8{,}60\ ; \qquad \hat{\sigma}(\hat{\bar{X}}_o) = 0{,}374\ ; \simeq 4{,}3\%$$

$$\hat{X}_o = 2.580; \qquad \hat{\sigma}(\hat{X}_o) = 450\ ; \simeq 17{,}4\%$$

Ejemplo II.7.

Sabemos que la varianza de una variable $X_i = 1$ con Pr (1) = P y $X_i = 0$ con Pr (0) = 1 − P, es $\sigma^2 = P(1 - P)$. Si $\hat{P}$ es un estimador insesgado de P, pruébese que $\hat{\sigma}^2 = \hat{P}(1 - \hat{P})$ es un estimador sesgado de σ^2 con sesgo negativo.

Solución:

$$B(\hat{\sigma}^2) = - V(\hat{P})$$

Ejemplo II.8.

En una población de N = 500 unidades está definida una característica bidimensional (X_i, Y_i). Una muestra aleatoria, sin reposición, de tamaño $n = 80$ presenta los datos del cuadro adjunto.

ΣX_i	ΣY_i	$\Sigma X_i Y_i$	ΣX_i^2	ΣY_i^2
190	420	1045	512	2284

Estime el sesgo y el error de muestreo de $\hat{R}$.

Solución:

$s_x^2 = 0{,}76899$; $s_y^2 = 1$; $s_{xy} = 0{,}60127$; $\hat{R} = 0{,}45238$

$\hat{B}(\hat{R}) = -567.10^{-7}$. Un 1,3% de $\hat{R}$.

$\hat{\sigma}(\hat{R}) = 0{,}0128$. Un 2,8% de R.

Ejemplo II.9.

De una característica bidimensional (X_i, Y_i) definida en una población de N = 750 unidades, se conocen los siguientes datos:

$$\bar{X} = 8\ ;\ \bar{Y} = 10\ ;\ \sigma_y^2 = 2500\ ;\ b = 0{,}6$$

Determine, en muestreo sin reposición, a partir de qué tamaño de la muestra el sesgo relativo de $\hat{R}$ es menor del 10%

Solución:

$n \geqslant 58$

Ejemplo II.10.

De una población con N = 1000 unidades, una muestra aleatoria de tamaño $n = 200$ presenta $n_1 = 110$ unidades pertenecientes a una clase $\mathcal{C}_1$ y $n_2 = 80$ pertenecientes a otra clase $\mathcal{C}_2$ que es disjunta de $\mathcal{C}_1$. Estime el error de muestreo relativo de la diferencia de proporciones, en muestreo sin reposición.

Solución: 28,8%

Ejemplo II.11.

Determine el tamaño de la muestra necesario para estimar el total, de modo que el error máximo admisible sea del 1% con un coeficiente de confianza del 68% . Datos: $N = 1800$; $S^2 = 16$; $\overline{X} = 42$.

Solución:

Suponiendo distribución normal para el estimador, al coeficiente $(1 - \alpha) = 0{,}68$ corresponde $\lambda = 1$, por tanto $\sigma(\hat{X}) = 0{,}01\ X$; se obtiene $n \geqslant 87$.

Ejemplo II.12.

En una ciudad que tiene $N = 15000$ viviendas, se ha tomado una muestra aleatoria, sin reposición y probabilidades iguales, de tamaño $n = 600$. En cada vivienda se han observado dos variables, X_i = N° de personas e Y_i = N° de habitaciones, obteniendo los siguientes datos muestrales:

$$\Sigma\ X_i = 2946\ ; \quad \Sigma\ X_i^2 = 18694$$

$$\Sigma\ Y_i = 2150\ ; \quad \Sigma\ Y_i^2 = 10997\ ; \quad \Sigma\ X_i\, Y_i = 12800$$

1°) Estime el número medio de personas por vivienda, dando un intervalo de confianza. 2°) Sabiendo que el número total de personas en la ciudad es 74000, estime el número total de habitaciones en la ciudad mediante un estimador de razón. 3°) Obtenga un intervalo de confianza para la razón N° total de habitaciones / N° total de personas.

Solución:

N° medio de personas por vivienda = 4,91.

Intervalo de confianza (4,71– 5,11) para $\lambda = 2$.

$\hat{Y} = 0{,}730 \cdot 74000 = 53280$ habitaciones.

$\hat{R} = 0{,}730$ habitaciones por persona.

Intervalo de confianza (0,707 – 0,753) para $\lambda = 2$.

Ejemplo II.13.

Se ha seleccionado una muestra aleatoria de 1000 niños en edad escolar de una población total de 100.000 niños en dicha situación. En una entrevista con sus padres se les preguntó: 1°) Cúantos niños tenían en edad escolar y 2°) Si la familia había salido de vacaciones con los hijos. Los datos obtenidos son:

N° de niños en edad escolar en la familia	N° de niños en la muestra	Proporción de respuestas sí a las vacaciones
1	210	0,80
2	280	0,60
3	270	0,40
4	240	0,20

1°) Estime la proporción de niños en edad escolar que salieron de vacaciones con su familia. 2°) La proporción de familias con niños en edad escolar que salieron de vacaciones.

Solución:

1°) 0,492 ; 2°) 0,60

Ejemplo II.14

Una muestra piloto de tamaño $n = 32$, presenta la siguiente tabla de frecuencias:

Y_i \ X_i	92	93	95	100	102	103
1,80	1	2	1	–	–	–
2,10	–	2	1	2	1	–
2,20	–	–	2	3	–	–
2,25	–	–	1	2	4	–
2,40	–	–	–	2	2	1
2,42	–	–	–	–	3	2

Admitiendo que el total Y es conocido, cúal de los estimadores $\hat{\bar{X}}_R$, $\bar{x}$ parece ser más preciso para estimar $\bar{X}$.

Solución: $\bar{x}$.

Ejemplo II.15

Una población finita consta de N elementos y se extrae una muestra aleatoria de tamaño n, supuesto que ambos N y n son números pares, para estimar $\bar{X}$. Supongamos ahora que los N elementos se agrupan formando N/2 pares y se extrae una muestra aleatoria de $n/2$ pares. 1°) Dar un estimador insesgado de $\bar{X}$ y la fórmula general de su varianza. 2°) Dar una regla práctica para formar los pares de modo que produzca la menor varianza; aplicar al caso en que los valores poblacionales X_i son: 2,4,4,5,6,6,6,8,10,10,12,15,18,20, comparando el resultado con el muestreo sin la formación de pares.

Solución:

Designemos N/2 = M y $n/2 = m$; sea también $X_i^* = X_{i1} + X_{i2}$ para $i = 1,2,\ldots,$ M la suma de los valores de i° par formado en la población y su media $\bar{X}_M = \frac{1}{M} \sum X_i^*$. El estimador insesgado de $\bar{X}$ es

$$\hat{\bar{X}}^* = \frac{1}{n} \sum_1^m X^* = \frac{1}{2m} \sum_1^m X_i^*$$

y siendo $\sigma^{*2} = 1/M \sum (X^* - \bar{X}_M)^2$, su varianza en muestreo con reposición es:

$$V(\hat{\bar{X}}^*) = \frac{\sigma^{*2}}{4m}$$

y ahora teniendo en cuenta que $\bar{X}_M = (1/M)\, X = (2/N)\, X = 2\bar{X}$,

resulta:

$$V(\hat{\bar{X}}^*) = \frac{\sigma^2 + 1/N \sum_1^M (X_{i1} - \bar{X})(X_{i2} - \bar{X})}{2\ m}$$

$$= V(\bar{x}) + \frac{1}{n\ N} \sum_1^M (X_{i1} - \bar{X})(X_{i2} - \bar{X})$$

en consecuencia el muestreo por pares será tanto más preciso, que el muestreo por unidades, cuanto más negativo sea el término suma de productos cruzados. Para ello los pares deben formarse con unidades cuya correlación lineal sea lo más grande posible en valor y en signo negativo. Una regla práctica consiste en, ordenados los X_i de menor a mayor, emparejar el primero con el último, el segundo con el penúltimo y así sucesivamente. Esto hace que los X^* se concentren cerca de X_M haciendo σ^{*2} pequeña. En el ejemplo numérico propuesto, para los pares: (2,20); (4,18); (4,15); (5,12); (6,10); (6,10); (6,8)

se obtiene $V(\hat{\bar{X}}^*) = V(\bar{x}) - (900/n\ N)$; $V(\bar{x}) = 108{,}4/n$; $V(\hat{\bar{X}}^*) = 44{,}1/n$ por tanto la segunda es un 40,68% de la primera.

Ejemplo II.16.

Estimador de regresión lineal. En la situación del ejemplo II.14, supongamos que es conocida $\bar{Y} = 2{,}15$ y que la población es de tamaño grande de modo que se puede suponer que el muestreo es con reposición. Calcule la estimación

$$\hat{\bar{X}}_{RL} = \bar{x} + K(\bar{Y} - \bar{y}) \text{ con } K = 0{,}065$$

y compruebe que es más preciso este estimador que los estudiados en el citado ejemplo.

Solución:

$\hat{\bar{X}}_{RL}$ es insesgado puesto que

$$E(\hat{\bar{X}}_{RL}) = E(\bar{x}) + KE(\bar{Y} - \bar{y}) = E(\bar{x}), \text{ por ser } E(\bar{y}) = \bar{Y}.$$

Su varianza es:

$$V(\hat{\bar{X}}_{RL}) = \frac{1}{n} (\sigma_x^2 + K_y^2 \sigma^2 - 2K \sigma_{xy})$$

resultado que se obtiene inmediatamente considerando la cantidad $Z_i = X_i + K(\bar{Y} - Y_i)$ definida en cada unidad de la población, de modo que su media muestral es precisamente $\hat{\bar{X}}_{RL}$. Su estimación insesgada es:

$$\hat{V}(\hat{\bar{X}}_{RL}) = \frac{1}{n} (s_x^2 + K_y^2 s_y^2 - 2K s_{xy})$$

La aplicación numérica da:

$$\hat{\bar{X}}_{RL} = 98{,}962 \quad ; \quad \hat{V}(\hat{\bar{X}}_{RL}) = 0{,}4218$$

III. MUESTREO UNIETAPICO DE CONGLOMERADOS

III.1. CONCEPTO Y NOTACIONES

Consideremos una población finita $\mathcal{A}$, constituida por M unidades elementales que, de alguna forma natural o artificial, están agrupadas en N unidades mayores, U_i, que se denominan *conglomerados* o Unidades Primarias (UP).

El muestreo por conglomerados sin submuestreo, consiste en la selección de muestras constituídas por todas las unidades elementales de n unidades primarias elegidas aleatoriamente.

Esta técnica es menos eficiente que la selección irrestricta de unidades elementales; esto es debido a que, en la mayoría de las poblaciones, las unidades de un mismo conglomerado presentan cierto grado de similitud u homogeneidad y en consecuencia, las unidades elementales de la muestra se reparten menos uniformemente sobre la población. Normalmente la eficiencia disminuye al aumentar el tamaño de los conglomerados; así, por ejemplo, una muestra de 1.000 unidades pertenecientes a 10 conglomerados de 100 unidades cada uno, se reparte menos uniformemente que si se toman 100 conglomerados de 10 unidades cada uno. A pesar de este inconveniente, es útil en poblaciones muy numerosas y tiene la ventaja de que reduce los costes de observación, especialmente si deben efectuarse visitas a las unidades seleccionadas. Por otra parte, cuando la información--marco no permite individualizar las unidades elementales, se recurre a unidades mayores donde aquellas están

ubicadas; este es el caso de las divisiones territoriales creadas por necesidades administrativas, o áreas geográficas creadas por el investigador. De esta forma solamente es necesario formar listados de las unidades pertenecientes a los conglomerados de la muestra, lo cual es otro factor que reduce costes y tiempo de preparación de los trabajos.

Suponemos que todos los conglomerados son de igual tamaño, $\overline{M}$, de modo que $M = N\overline{M}$ es el número total de unidades. Siendo X_{ij} el valor numérico definido en la j^a unidad del i^o conglomerado utilizaremos las siguientes notaciones:

$$X_i = \sum_j X_{ij} \; ; \quad \text{Total del i}^o \text{ conglomerado}$$

$$X = \sum_i X_i = \sum_i \sum_j X_{ij} \; ; \quad \text{Total general}$$

$\overline{\overline{X}}$, $\overline{X}_i$, $\overline{X}$; Media general, Media del i^o conglomerado y Media entre totales de conglomerados, respectivamente.

$$\sigma^2 = \frac{1}{NM} \sum_i \sum_j (X_{ij} - \overline{\overline{X}})^2 \; ; \quad \text{Varianza}$$

$$\sigma_i^2 = \frac{1}{\overline{M}} \sum_j (X_{ij} - \overline{X}_i)^2 \; ; \quad \text{Varianza dentro del i}^o \text{ conglomerado}$$

$$\sigma_c^2 = \frac{1}{N} \sum_i (X_i - \overline{X})^2 \; ; \text{ Varianza entre totales de conglomerados}$$

Las correspondientes cuasivarianzas se definen de la forma usual, con las notaciones S^2, S_i^2, S_c^2.

III.2 MEDIDA DE LA HOMOGENEIDAD DE LOS CONGLOMERADOS

La varianza σ^2 admite la descomposición: $\sigma^2 = \sigma_b^2 + \sigma_w^2$;

donde:

$$\sigma_b^2 = \frac{1}{N} \sum_{i=1}^{N} (\bar{X}_i - \bar{\bar{X}})^2 = \frac{1}{\bar{M}^2} \sigma_c^2$$; se denomina Varianza "entre" conglomerados y

$$\sigma_w^2 = \frac{1}{N\bar{M}} \sum_{i=1}^{N} \sum_{j=1}^{\bar{M}} (X_{ij} - \bar{X}_i)^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \quad ;$$

se denomina Varianza "dentro" de conglomerados. En efecto:

$$\sigma^2 = \frac{1}{N\bar{M}} \sum_{i}^{N} \sum_{j}^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2 = \frac{1}{N\bar{M}} \sum_{i}^{N} \sum_{j}^{\bar{M}} \left[(X_{ij} - \bar{X}_i) + (\bar{X}_i - \bar{\bar{X}}) \right]^2$$

$$= \frac{1}{N\bar{M}} \sum_{i}^{N} \sum_{j}^{\bar{M}} (X_{ij} - \bar{X}_i)^2 + \frac{1}{N\bar{M}} \sum_{i}^{N} \sum_{j}^{\bar{M}} (\bar{X}_i - \bar{\bar{X}})^2 ;$$

por ser nulo el tercer sumando; operando resulta:

$$\sigma^2 = \sigma_w^2 + \frac{1}{N\bar{M}} . \bar{M} \sum_{i}^{N} (\bar{X}_i - \bar{\bar{X}})^2 = \sigma_w^2 + \sigma_b^2$$

Cuanto más homogéneas sean las unidades dentro de los conglomerados, menor será la componente "dentro", σ_w^2 , y mayor la componente "entre" σ_b^2

Para medir la homogeneidad de los conglomerados, con respecto a una característica $\mathfrak{X}$ que toma valores X_{ij} en las unidades de la población, se define el siguiente *coeficiente de homogeneidad de conglomerados:*

$$\delta = \frac{\sum_{i=1}^{N} \sum_{j \neq k}^{\bar{M}} (X_{ij} - \bar{\bar{X}}) (X_{ik} - \bar{\bar{X}})}{N\bar{M} (\bar{M} - 1) \sigma^2}$$

donde la suma del numerador se extiende a los $N\bar{M}\ (\bar{M}-1)$ pares de unidades distintas dentro de los N conglomerados. Teniendo en cuenta que

$$\sigma^2 = \frac{N\bar{M}-1}{N\bar{M}} S^2, \text{ admite la forma:}$$

$$\delta = \frac{\sum_{i=1}^{N} \sum_{j \neq k}^{\bar{M}} (X_{ij} - \bar{\bar{X}})(X_{ik} - \bar{\bar{X}})}{(\bar{M}-1)(N\bar{M}-1)\ S^2}$$

El coeficiente δ es tanto mayor cuanto más homogéneos son los conglomerados, y es tanto menor cuanto más heterogéneos, pudiendo tomar valores negativos en situaciones de gran heterogeneidad. Veamos como ejemplo dos situaciones extremas en una población con $N = 3$, $\bar{M} = 3$, cuyos valores X_{ij} en el caso (a) son:

$$U_1 : \{1,1,1,\}\ ; \qquad U_2 : \{2,2,2\}\ ; \qquad U_3 : \{3,3,3\}$$

y en el caso (b):

$$U_1 : \{1,2,3\}\ ; \qquad U_2 : \{1,2,3\}\ ; \qquad U_3 : \{1,2,3\}$$

Haciendo los cálculos necesarios se obtiene δ (a) $= 1$; δ (b) $= -0{,}5$, lo cual expresa que para una misma población de unidades elementales, la distribución de éstas en conglomerados homogéneos o heterogéneos da lugar a un valor de δ mucho mayor en el primer caso (habida cuenta de la diferencia de signo), pudiendo ser, como en este ejemplo, negativo en el segundo caso. Es inmediato ver que en esta última situación, bastaría una muestra de tamaño $n = 1$, para tener representada a toda la población de unidades elementales.

En poblaciones reales δ suele tomar valores positivos, en cuyo caso el muestreo por conglomerados produce un incremento de la varianza de los estimadores, respecto al muestreo irrestricto de unidades elementales, que como veremos más adelante puede expresarse en función de $\bar{M}$ y δ .

Veamos como puede obtenerse la relación que existe entre σ^2, σ_b^2 y δ :

$$\sigma_b^2 = \frac{1}{N} \sum_{i=1}^{N} (\bar{X}_i - \bar{\bar{X}})^2 = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\bar{M}} \sum_{j=1}^{\bar{M}} (X_{ij} - \bar{\bar{X}}) \right]^2 =$$

$$= \frac{1}{N\bar{M}^2} \sum_{i=1}^{N} \left[\sum_{j=1}^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2 + \sum_{j \neq k}^{\bar{M}} (X_{ij} - \bar{\bar{X}}) (X_{ik} - \bar{\bar{X}}) \right] =$$

$$= \frac{1}{N\bar{M}^2} \cdot N\bar{M} \sigma^2 + \frac{1}{N\bar{M}^2} \cdot N\bar{M}(\bar{M} - 1) \sigma^2 \cdot \delta =$$

$$= \frac{\sigma^2}{\bar{M}} + \frac{(\bar{M} - 1) \sigma^2 \delta}{\bar{M}} \quad ; \text{tenemos por tanto:}$$

$$\sigma_b^2 = \frac{\sigma^2}{\bar{M}} (1 + (\bar{M} - 1) \delta)$$

Otra relación interesante resulta de la anterior teniendo en cuenta que $\sigma^2 = \sigma_b^2 + \sigma_w^2$; tal relación es:

$$\delta = \frac{\sigma_b^2 - \sigma_w^2 / (\bar{M} - 1)}{\sigma^2}$$

y nos dice lo siguiente:

1. El valor máximo para δ es 1 y se obtiene cuando la varianza "dentro" es cero (homogeneidad total), en cuyo caso $\sigma_b^2 = \sigma^2$. Así ocurre en el caso (a) del ejemplo visto anteriormente.

2. El valor mínimo para δ es $-1 / \bar{M}-1$ y se obtiene cuando la varianza "entre" es cero (heterogeneidad dentro de conglomerados), en cuyo caso $\sigma_w^2 = \sigma^2$. Así ocurre en el caso (b) del ejemplo mencionado.

Estos extremos se obtienen, el máximo cuando todos los conglomerados están formados por unidades con valores X_{ij} iguales entre sí; $X_{ij} = \bar{X}_i$ para todo j ; y el mínimo cuando todos los conglomerados tienen la misma composición.

III.3. ESTIMACION DE LA MEDIA $\bar{\bar{X}}$ Y DEL TOTAL X.

En muestreo unietápico de conglomerados con probabilidades iguales, sea con o sin reposición, la media de las n $\bar{M}$ unidades simples de la muestra es un estimador insesgado para $\bar{\bar{X}}$. En efecto,

$$\bar{\bar{x}} = \frac{1}{n\,\bar{M}} \sum_{i=1}^{n} X_i \text{ puede escribirse como,}$$

$$\bar{\bar{x}} = \frac{1}{\bar{M}}\,\bar{x}, \text{ siendo } \bar{x} \text{ la media muestral de los totales}$$

de conglomerados, y de acuerdo con la teoría vista en el Cap. II,

$$E(\bar{\bar{x}}) = \frac{1}{\bar{M}}\,E(\bar{x}) = \frac{1}{\bar{M}}\,\bar{X} = \bar{\bar{X}}\,.$$

La varianza de $\bar{x}$ es:

$$V_{CR}(\bar{x}) = \frac{\sigma_c^2}{n}\,; \qquad V_{SR}(\bar{x}) = \frac{N-n}{N}\,\frac{S_c^2}{n}$$

por tanto la varianza de $\bar{\bar{x}}$, es:

$$V_{CR}(\bar{\bar{x}}) = \frac{\sigma_c^2}{n\bar{M}^2} \; ; \qquad V_{SR}(\bar{\bar{x}}) = \frac{N-n}{N\bar{M}^2} \; \frac{S_c^2}{n}$$

según se trate de muestreo con o sin reposición.

Teniendo en cuenta que $\sigma_c^2 = \bar{M}^2 \; \sigma_b^2$, en muestreo con reposición, se obtiene:

$$V_{CR}(\bar{\bar{x}}) = \frac{\sigma_b^2}{n}$$

y teniendo en cuenta que $S_c^2 = \bar{M}^2 \; S_b^2$, en muestreo sin reposición:

$$V_{SR}(x) = \frac{N-n}{N} \; \frac{S_b^2}{n} = (1-f)\frac{S_b^2}{n}$$

Para el total $X = N \bar{M} \bar{\bar{X}}$, se obtiene inmediatamente que $\hat{X} = N \bar{M} \bar{\bar{x}} = (N/n) \cdot x$, siendo x = total muestral, es un estimador insesgado cuya varianza es $V(\hat{X}) = N^2 \bar{M}^2 V(\bar{\bar{x}})$.

III.4. ESTIMACION DE LA PROPORCION P Y DEL TOTAL DE CLASE C.

Teniendo en cuenta las siguientes notaciones:

$p = \frac{c}{n\bar{M}} = \frac{\sum_{i=1}^{n} C_i}{n\bar{M}}$, es la proporción muestral siendo C_i el total de clase del i° conglomerado y c el total de clase en la muestra; $P = C/N\bar{M}$ es la proporción en la población y $P_i = C_i/\bar{M}$ es la proporción dentro del i° conglomerado, se tiene:

$$\sigma^2 = PQ\ ;\quad \sigma_i^2 = P_i\,Q_i\ ;\quad \bar{\bar{X}}_i = P_i\ ;\quad \bar{\bar{\bar{X}}} = P$$

$$\sigma_b^2 = \frac{1}{N}\sum_{i=1}^{N}(P_i - P)^2 = \frac{1}{N}\left\{\sum_{i=1}^{N} P_i^2 - NP^2\right\}$$

De acuerdo con los resultados obtenidos para la estimación de $\bar{\bar{\bar{X}}}$, que se identifica con P definiendo X_{ij} como uno o cero según que la unidad A_{ij} pertenezca o no a la clase; podemos afirmar que p es un estimador insesgado de P cuya varianza, en muestreo con reposición, viene dada por:

$$V_{CR}(p) = \frac{\sigma_b^2}{n} = \frac{1}{nN}\sum_{i=1}^{N}(P_i - P)^2$$

y en muestreo sin reposición:

$$V_{SR}(p) = (1-f)\frac{S_b^2}{n} = \frac{(1-f)}{n(N-1)}\sum_{i=1}^{N}(P_i - P)^2$$

Para el total de clase C, $C = N\,\bar{M}\,P$, se tiene $\hat{C} = N\,\bar{M}\,p$ es un estimador insesgado, cuya varianza es $V(\hat{C}) = N^2\,\bar{M}^2\,V(p)$.

III.5 EFICIENCIA DEL MUESTREO POR CONGLOMERADOS

De los resultados obtenidos anteriormente para muestreo con reposición, tenemos:

$$V_{CR}(\bar{\bar{x}}) = \frac{\sigma_b^2}{n}\ ;\qquad \sigma_b^2 = \frac{\sigma^2}{\bar{M}}\,(1 + (\bar{M}-1)\,\delta)$$

por tanto:

$$V_{CR}(\bar{\bar{x}}) = \frac{\sigma^2}{n\,\bar{M}}\,(1 + (\bar{M}-1)\,\delta)$$

y para muestreo sin reposición, por ser

$$\frac{N}{N-1}\, S_b^2 = \frac{(N\bar{M}-1)}{N\bar{M}^2}\, S^2\, (1 + (\bar{M}-1)\, \delta\,)$$

podemos escribir:

$$V_{SR}(\bar{\bar{x}}) = \frac{(N-1)\,(N\bar{M}-1)}{N^2\,\bar{M}}\, \frac{N-n}{N}\, \frac{S^2}{n\,\bar{M}}\, (1 + (\bar{M}-1)\, \delta\,)$$

de donde, admitiendo la aproximación $N-1 \simeq N$, $N\bar{M}-1 \simeq N\bar{M}$, el primer factor es próximo a la unidad, resulta:

$$V_{SR}(\bar{\bar{x}}) = (1-f)\, \frac{S^2}{n\,\bar{M}}\, (1 + (\bar{M}-1)\, \delta\,)$$

En consecuencia la varianza en muestreo por conglomerados, V_{MC}, puede escribirse en relación con la varianza, V_{MI}, en muestreo irrestricto del mismo número $n\,\bar{M}$ de unidades elementales, de la siguiente forma (independiente de que se trate de muestreo con o sin reposición y de que se trate del estimador de la media o del total):

$$V_{MC} = V_{MI}\, (1 + (\bar{M}-1)\, \delta\,)$$

En esta fórmula se observa que si el coeficiente de homogeneidad es $\delta = 0$, el muestreo por conglomerados es igualmente eficiente que el muestreo por unidades; es más eficiente si $\delta < 0$ y es menos eficiente si $\delta > 0$. El incremento de la varianza, respecto al muestreo por unidades, viene dado por:

$$\frac{V_{MC} - V_{MI}}{V_{MI}} = (\bar{M}-1)\, \delta$$

y expresando V_{MC} como porcentaje de V_{MI}:

$$100\,(1 + (\bar{M}-1)\, \delta\,)\%$$

En la práctica la pérdida de eficiencia puede ser relativamente elevada incluso para valores moderados de $\bar{M}$ y δ ; así por ejemplo, para $\bar{M} = 101$ y $\delta = 0{,}02$, la varianza se triplica, lo que significa que el error de muestreo se multiplica por $\sqrt{3}$; esto es, aumenta en un 73,2 %.

La eficiencia del muestreo por conglomerados puede estudiarse directamente, mediante la relación entre las varianzas V_{MC} (muestreo por conglomerados) y V_{MI} (muestreo por unidades) con igual tamaño de muestra $n\,\bar{M}$; en muestreo con reposición, resulta:

$$\left(\frac{V_{MC}}{V_{MI}}\right)_{CR} = \frac{\bar{M}\,\sigma_b^2}{\sigma^2} \quad ; \quad \text{(Con reposición)}$$

donde se observa que si $\bar{M}\,\sigma_b^2 = \sigma^2$ es igualmente eficiente, si $\bar{M}\,\sigma_b^2 < \sigma^2$ es más eficiente y si $\bar{M}\,\sigma_b^2 > \sigma^2$ es menos eficiente. En muestreo sin reposición la eficiencia viene dada por:

$$\left(\frac{V_{MC}}{V_{MI}}\right)_{SR} = \frac{\bar{M}\,S_b^2}{S^2} = \frac{N\,\bar{M} - 1}{\bar{M}\,(N-1)} \left(\frac{V_{MC}}{V_{MI}}\right)_{CR}$$

III.6 ESTIMACION DE LAS VARIANZAS, DE LA EFICIENCIA, Y DEL COEFICIENTE δ

El problema de la estimación de la varianza de los estimadores, en muestreo unietápico de conglomerados, se reduce al de estimar la varianza o cuasivarianza "entre", σ_b^2 , ó S_b^2 . Teniendo en cuenta los resultados obtenidos en II.3, se deduce inmediatamente que:

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{X}_i - \bar{\bar{x}})^2 = \frac{1}{\bar{M}^2\,(n-1)} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

es un estimador insesgado de σ_b^2 en muestreo con reposición y de S_b^2 en muestreo sin reposición; por tanto:

$$\hat{V}_{CR}(\bar{\bar{x}}) = \frac{s_b^2}{n} \quad ; \qquad \hat{V}_{SR}(\bar{\bar{x}}) = (1-f)\frac{s_b^2}{n}$$

son estimadores insesgados de $V(\bar{\bar{x}})$ en uno y otro caso respectivamente. Estimadores insesgados de $V(\hat{X})$ se obtienen de los anteriores multiplicando por $N^2 \bar{M}^2$, por ser $\hat{X} = N \bar{M} \bar{\bar{x}}$.

Si se trata de la estimación de una proporción P, se tiene:

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (P_i - p)^2 = \frac{1}{n-1} \left\{ \left(\sum_{i=1}^{n} P_i^2 \right) - n\, p^2 \right\}$$

por tanto:

$$\hat{V}_{CR}(p) = \frac{1}{n(n-1)} \left\{ \left(\sum_{i=1}^{n} P_i^2 \right) - n\, p^2 \right\}$$

$$\hat{V}_{SR}(p) = \frac{(1-f)}{n(n-1)} \left\{ \left(\sum_{i=1}^{n} P_i^2 \right) - n\, p^2 \right\}$$

son estimadores insesgados de $V(p)$ en muestreo con y sin reposición respectivamente. Estimadores insesgado de $V(\hat{C})$ se obtienen de los anteriores multiplicando por $N^2 \bar{M}^2$, por ser $\hat{C} = N \bar{M} p$.

También se deduce inmediatamente de los resultados obtenidos en II.1, que por ser σ_w^2 la media de las varianzas "dentro" de los conglomerados:

$$s_{1w}^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 = \frac{1}{n \bar{M}} \sum_{i=1}^{n} \sum_{j=1}^{\bar{M}} (X_{ij} - \bar{X}_i)^2$$

es un estimador insesgado de σ_w^2, sea con o sin reposición el procedi-

miento de muestreo. Por otra parte s_b^2 es un estimador insesgado para σ_b^2 , en muestreo con reposición, por tanto:

$$\hat{\sigma}^2 = s_b^2 + s_{1w}^2$$

es un estimador insesgado para σ^2 .

En muestreo sin reposición por ser, $S^2 = \frac{N\bar{M}}{N\bar{M}-1}\sigma^2$ y s_b^2 insesgado para $S_b^2 = \frac{N}{N-1}\sigma_b^2$, resulta que

$$\hat{S}^2 = \frac{\bar{M}(N-1)\, s_b^2 + N\bar{M}\, s_{1w}^2}{N\bar{M}-1} \quad , \text{ o bien}$$

$$\hat{S}^2 = \frac{\bar{M}(N-1)\, s_b^2 + N(\bar{M}-1)\, s_w^2}{N\bar{M}-1}$$

es un estimador insesgado para S^2 .

Si se trata del problema de estimación de una proporción P, s_{1w}^2 puede ponerse en la forma:

$$s_{1w}^2 = \frac{1}{n}\sum_{i=1}^{n} P_i Q_i = \frac{1}{n}\sum_{i=1}^{n} P_i (1 - Q_i)$$

por tanto, en muestreo con reposición:

$$\hat{\sigma}^2 = \hat{PQ} = \frac{1}{n-1}\left\{\left(\sum_{i=1}^{n} P_i^2\right) - n\,p^2\right\} + \frac{1}{n}\sum_{i=1}^{n} P_i\,(1-P_i) \;;$$

es un estimador insesgado de $\sigma^2 = PQ$; simplificando:

$$\hat{\sigma}^2 = \hat{PQ} = \frac{\sum_{i=1}^{n} P_i^2}{n(n-1)} + \frac{n\,pq - p}{(n-1)}$$

Observemos que $\frac{n}{(n-1)}\,pq$ es un estimador insesgado de PQ en muestreo irrestricto por unidades, pero en muestreo por conglomerados $\frac{n}{(n-1)}\,pq$ como estimador de PQ, tiene un sesgo positivo de cuantía:

$$\frac{1}{N(n-1)} \sum_{1}^{N} P_i\,Q_i = \frac{1}{(n-1)}\,\sigma_w^2$$

En efecto:

$$E\left(\frac{n\,p\,q}{n-1}\right) - E\left(\frac{p}{n-1}\right) + E\left(\frac{\sum_{i=1}^{n} P_i^2}{n\,(n-1)}\right) = PQ\,;$$

entonces el sesgo es,

$$E\left(\frac{n\,p\,q}{n-1}\right) - PQ = E\left(\frac{p}{n-1}\right) - E\left(\frac{\sum_{i=1}^{n} P_i^2}{n\,(n-1)}\right) =$$

$$= \frac{P}{n-1} - \frac{n\sum_{1}^{N} P_i^2}{Nn\,(n-1)} = \frac{\sum_{1}^{N} (P_i - P_i^2)}{N\,(n-1)} = \frac{\sum_{1}^{N} P_i\,(1-P_i)}{N\,(n-1)}$$

En muestreo sin reposición el estimador insesgado de $S^2 = \frac{N}{N-1}\,PQ$,

teniendo en cuenta los resultados ya obtenidos para $\hat{S}^2$, resulta ser:

$$\hat{S}^2 = \left(\frac{N\hat{PQ}}{N-1}\right) = \frac{\bar{M}(N-1)}{N\bar{M}-1} \cdot \frac{1}{(n-1)} \left\{ \sum_{i=1}^{n} P_i^2 - n\,p^2 \right\} +$$

$$+ \frac{N\bar{M}}{N\bar{M}-1} \cdot \frac{1}{n} \sum_{i=1}^{n} P_i (1-P_i)$$

La eficiencia del muestreo por conglomerados puede estimarse mediante la razón:

$$\frac{\bar{M}\ s_{1w}^2}{s_b^2 + s_{1w}^2}$$, en muestreo con reposición.

$$\frac{\bar{M}(N\bar{M}-1)\ s_w^2}{\bar{M}(N-1)\ s_b^2 + N(\bar{M}-1)\ s_w^2}$$; en muestreo sin reposición,

o bien $$\frac{N\bar{M}-1}{\bar{M}(N-1)} \cdot \frac{\bar{M}\ s_{1w}^2}{s_b^2 + s_{1w}^2}$$

Para el coeficiente de homogeneidad de conglomerados puede formarse el siguiente estimador de razón, formado por estimaciones insesgadas del numerador y denominador:

$$\hat{\delta} = \frac{s_b^2 - s_{1w}^2 \,/\, (\bar{M}-1)}{s_b^2 + s_{1w}^2}$$, en muestreo con reposición

$$\hat{\delta} = \frac{\frac{N-1}{N} s_b^2 - s_{1w}^2 \; / (\bar{M} - 1)}{\frac{(N-1)}{N} s_b^2 + s_{1w}^2}$$, en muestreo sin reposición.

o bien $$\hat{\delta} = \frac{\frac{N-1}{N} s_b^2 - \bar{M} s_w^2 \; / (\bar{M} - 1)^2}{\frac{N-1}{N} s_b^2 + \bar{M} s_w^2 / (\bar{M} - 1)^2}$$

III.7 TAMAÑO OPTIMO DE LA MUESTRA Y DE LOS CONGLOMERADOS

Si la composición y el tamaño de los conglomerados están prefijados, el cálculo del tamaño de muestra, n, necesario para una precisión dada es inmediato. Si para estimar la media se fija una varianza V_o, tenemos:

$$V(\bar{\bar{x}}) = \frac{\sigma_b^2}{n} = V_o \quad ; \quad \text{de donde resulta } n = \frac{\sigma_b^2}{V_o}$$

en muestreo con reposición, y

$$V(\bar{\bar{x}}) = \frac{N-n}{N-1} \frac{\sigma^2}{n} = V_o \quad : \quad n = \frac{n'}{1 + (n'-1)/N}$$

en muestreo sin reposición (siendo n' el tamaño que corresponde a la misma precisión en el caso con reposición).

Si no está prefijado el tamaño de los conglomerados, se plantea un problema en el que interviene n y $\bar{M}$ pero este último (el tamaño de los conglomerados) no aparece explícitamente en $V(\bar{\bar{x}})$. Real-

mente debe esperarse que al variar $\overline{M}$, también variarán σ_b^2 y σ_w^2 manteniendo la relación constante: $\sigma^2 = \sigma_b^2 + \sigma_w^2$. En los puntos límites se tiene:

$$\text{Para } \overline{M} = 1 \; ; \quad \sigma_w^2 = 0 \; ; \quad \sigma_b^2 = \sigma^2$$

$$\text{Para } \overline{M} = M \; ; \quad \sigma_w^2 = \sigma^2 \; ; \quad \sigma_b^2 = 0$$

siendo M el número total de unidades elementales en la población. Entre estos límites, al aumentar $\overline{M}$ debe esperarse que σ_b^2 disminuya. por tanto el problema consiste en determinar una relación de σ_b^2 como función monótona decreciente de $\overline{M}$. Para ello debemos recurrir a la experiencia sobre la población objeto de estudio efectuando estimaciones para diversos tamaños, por ejemplo $1/4\ M_o$, $1/2\ M_o$, $\overline{M}_o$, etc. A continuación debemos de proponer una función de ajuste $\sigma_b^2 = \sigma_b^2 . (\overline{M})$. Los dos tipos usualmente propuestos son del tipo:

$$\sigma_b^2 = \sigma^2 \ \overline{M}^{\alpha} \; ; \qquad -1 < \alpha < 0. \quad \text{(F. Smith, 1938)}$$

$$\sigma_w^2 = A \, \overline{M}^{g} \; ; \qquad g > 0. \qquad \text{(Jessen 1942)}$$

esta última implica $\sigma_b^2 = \sigma^2 - A \, \overline{M}^{g}$.

La primera de ellas puede justificarse mediante el siguiente razonamiento: si tomamos muestra de $\overline{M}$ unidades, la varianza de la media será $\sigma^2 \ \overline{M}^{-1}$ y dado que la varianza entre las medias de conglomerados (σ_b^2) de tamaño $\overline{M}$, será mayor, es razonable afectar a $\overline{M}$ de un exponente α mayor que -1.

La segunda función se estudió como una mejora de la primera; necesita al menos dos experiencias para estimar sus dos parámetros A y g, aunque teniendo en cuenta que para $\overline{M} = M$ debe ser $\sigma_w^2 = \sigma^2$, introduciendo esta condición bastaría con una sola experiencia.

A pesar de que ambos tipos de función admiten algunas objeciones, en la práctica nos movemos en un determinado campo de variación de $\bar{M}$ y por ello pueden dar resultados suficientemente ajustados.

Cualquiera que sea la función que se ha decidido ajustar, el problema se reduce a determinar n y $\bar{M}$ con alguna condición de optimalidad. El criterio usual consiste en minimizar la varianza para un coste dado y en este momento se plantea la forma de la función de coste; las más usuales son:

$$C = c_1\, n\, \bar{M} + c_2\, n \qquad \text{o bien} \qquad C = c_1\, n\, \bar{M} + c_2\, \sqrt{n}$$

la segunda de las cuales se justifica por consideraciones teórico–empíricas que conducen a que el coste de viajar entre n puntos es proporcional a $\sqrt{n}$. En cada caso debe estudiarse detenidamente qué función de coste refleja mejor la situación.

El problema es por tanto:

$$\text{Mínimo de } V = \frac{\sigma_b^2\,(\bar{M})}{n} \quad \text{con } C(n, \bar{M}) = C_o$$

Una solución consiste en utilizar las actuales posibilidades de cálculo automático: se da un valor a $\bar{M}$ en la ecuación de costo y con el valor de n obtenido, junto con el $\bar{M}$ dado, se determina la varianza; después de un cierto número de repeticiones de este proceso de tanteo se llega al par n, $\bar{M}$ que hace V mínima para el Coste C_o.

Puede intentarse la solución analítica mediante el método de los multiplicadores de Lagrange. Supongamos el problema:

$$\text{Mínima } V = \frac{\sigma_b^2\,(\bar{M})}{n} \quad \text{con } c_1\, n\, \bar{M} + c_2\, \sqrt{n} = C_o$$

Entonces, las derivadas de:

$$\phi = V + \lambda\,(c_1\, n\, \bar{M} + c_2\, \sqrt{n} - C_o)$$

respecto de n, $\bar{M}$ y λ igualadas a cero, teniendo en cuenta que $\frac{\partial V}{\partial n} = \frac{-V}{n}$, nos dan el sistema de ecuaciones:

$$\begin{cases} \dfrac{-V}{n} + \lambda\left(c_1 \bar{M} + \dfrac{1}{2} c_2 \dfrac{1}{\sqrt{n}}\right) = 0 \\ \dfrac{\partial V}{\partial \bar{M}} + \lambda c_1 n = 0 \\ n c_1 \bar{M} + c_2 \sqrt{n} = C_o \end{cases}$$

Eliminando λ entre las dos primeras:

$$\frac{-\partial V}{\partial M} \cdot \frac{\bar{M}}{V} = \frac{1}{1 + \dfrac{c_2}{2 c_1 \bar{M} \sqrt{n}}}$$

y despejando la raiz positiva para $\sqrt{n}$, de la ecuación de condición:

$$\sqrt{n} = \frac{-c_2 + \sqrt{c_2^2 + 4 c_1 C_o \bar{M}}}{2 c_1 \bar{M}}$$ y sustituyendo en la anterior:

$$\frac{-\partial V}{\partial \bar{M}} \frac{\bar{M}}{V} = \frac{-c_2 + \sqrt{c_2^2 + 4 c_1 C_o \bar{M}}}{\sqrt{c_2^2 + 4 c_1 C_o \bar{M}}}$$; o también

$$\frac{-\partial V}{\partial \bar{M}} \frac{\bar{M}}{V} = 1 - \frac{1}{\sqrt{1 + \frac{4 c_1 C_o \bar{M}}{c_2^2}}}$$

Para resolver esta ecuación en $\bar{M}$, debe sustituirse el primer miembro según la función de ajuste utilizada para σ^2 . Si es $\sigma_b^2 = \sigma_b^2 \bar{M}^{\alpha}$, se obtiene:

$$-\alpha = 1 - \frac{1}{\sqrt{1 + \frac{4 c_1 C_o \bar{M}}{c_2^2}}}$$

de donde puede despejarse explícitamente $\bar{M}$ y después obtener n de la ecuación de condición:

$$\bar{M} = \frac{1 - (1 + \alpha)^2}{(1 + \alpha)^2} \frac{c_2^2}{4 c_1 C_o} \quad ; \quad n = \frac{(1 + \alpha)}{(2 + \alpha)} \frac{2 C_o}{c_2}$$

Se observa que el tamaño óptimo de los conglomerados disminuye si aumenta c_1, si aumenta C_o o si disminuye c_2^2. El número de conglomerados en la muestra aumenta si lo hace C_o y disminuye si lo hace c_2.

Si se utiliza la función de ajuste:

$\sigma_w^2 = A \bar{M}^g \rightarrow \sigma_b^2 = \sigma^2 - A \bar{M}^g$, se obtiene la ecuación:

$$\frac{A g \bar{M}^g}{\sigma^2 - A \bar{M}^g} = 1 - \frac{1}{\sqrt{1 + \frac{4 c_1 C_o \bar{M}}{C_2^2}}}$$

que puede resolverse en $\bar{M}$ por tanteos y después calcular n en la ecuación de condición.

Con la función de coste $C_o = c_1 \; n \; \bar{M} + c_2 \; n$, se obtiene para $\bar{M}$ la ecuación:

$$\frac{-\partial V}{\partial \bar{M}} \; \frac{\bar{M}}{V} = \frac{1}{1 + \dfrac{c_2}{c_1 \; \bar{M}}}$$

por lo tanto si la ecuación de ajuste es $\sigma_b^2 = \sigma^2 \; \bar{M}^{\alpha}$, la solución óptima viene dada por:

$$\bar{M} = \frac{-\alpha}{1 + \alpha} \; \frac{c_2}{c_1} \quad ; \quad n = (1 + \alpha) \; \frac{C_o}{c_2}$$

y si la ecuación de ajuste es $\sigma_b^2 = \sigma^2 - A \; \bar{M}^g$, debe resolverse por tanteos la ecuación para $\bar{M}$:

$$\frac{A \, g \, \bar{M}^g}{\sigma^2 - A \; \bar{M}^g} = \frac{1}{1 + \dfrac{c_2}{c_1 \; \bar{M}}}$$

y obteniendo después $n = \dfrac{C_o}{c_1 \; \bar{M} + c_2}$.

Para este problema de tamaño óptimo de la muestra, debe tenerse en cuenta que tanto la función de ajuste para σ_b^2 como la función de coste, deben estudiarse detenidamente de forma que se decida utilizar aquellas que reflejen con suficiente aproximación la situación real de la encuesta. Observemos que también pueden hacerse los mismos ajustes para las cuasivarianzas, con el fin de estudiar el caso de muestreo sin

reposición, pero debe tenerse en cuenta que en el método de los multiplicadores se ha supuesto

$$\frac{-\partial V}{\partial n} = \frac{-V}{n}$$

, lo cual implica aproximar S_b^2 de modo que V sea inversamente proporcional a n (como ocurre en muestreo con reposición).

III.8. MUESTREO CON REPOSICION DE CONGLOMERADOS CON PROBABILIDADES DESIGUALES.

Supongamos ahora que de la población de N unidades primarias $U_1, U_2, \ldots, U_i, \ldots, U_N$ se extrae una muestra de tamaño n, con reposición, de modo que en todas y cada una de las extracciones, U_i tiene una probabilidad P_i. Siendo:

$$0 \leqslant P_i \leqslant 1 \quad \text{y} \quad \sum_1^N P_i = 1$$

este procedimiento responde al modelo polinomial estudiado en cap. I.6. Vamos a considerar estimadores lineales de la forma:

$$\hat{X} = \sum_1^n C_i X_i$$

ya descritos en cap. I.8, que son funciones lineales de los valores observados (totales de conglomerados) en las unidades de la muestra, $\mathfrak{X}(U_i) = X_i$. Los coeficientes C_i están definidos para cada una de las N unidades primarias, si bien en $\hat{X}$ solamente intervienen los correspondientes a las unidades muestrales. Una forma más correcta de escribir $\hat{X}$ es:

$$\hat{X} = \sum_1^n C(U_i)\ \mathfrak{X}(U_i)$$

entendiendo que si una U_i está repetida r_i veces en la muestra, tendremos

r_i sumandos iguales. Otra forma más conveniente es:

$$\hat{X} = \sum_{1}^{N} C_i X_i \varepsilon_i$$

donde C_i, X_i son constantes y ε_i es el vector aleatorio que expresa el número de veces que U_i está en la muestra. En Cap. I.6 hemos calculado sus momentos:

$$E(\varepsilon_i) = nP_i \; ; \quad E(\varepsilon_i^2) = n(n-1)P_i^2 + nP_i$$

$$E(\varepsilon_i \varepsilon_j) = n(n-1)P_i P_j \; ; \; \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = -n P_i P_j \; ; \; i \neq j$$

$$V(\varepsilon_i) = n P_i (1 - P_i)$$

Las fórmulas generales de la esperanza y la varianza de $\hat{X}$, son:

$$E(\hat{X}) = \sum_{1}^{N} C_i X_i n P_i \qquad (1)$$

$$V(\hat{X}) = n \left\{ \sum_{1}^{N} C_i^2 X_i^2 P_i - \left(\sum_{1}^{N} C_i X_i P_i \right)^2 \right\} \qquad (2)$$

En efecto:

$$E(\hat{X}) = E\left(\sum_{1}^{N} C_i X_i \varepsilon_i \right) = \sum_{1}^{N} C_i X_i E(\varepsilon_i) = \sum_{1}^{N} C_i X_i nP_i$$

para la varianza,

$$V(\hat{X}) = V\left(\sum_{1}^{N} C_i X_i \varepsilon_i \right) =$$

$$= \sum_{1}^{N} C_i^2 X_i^2 V(\varepsilon_i) + \sum_{i \neq j} C_i C_j X_i X_j \mathrm{Cov}(\varepsilon_i, \varepsilon_j) =$$

$$= \sum_{1}^{N} C_i^2 X_i^2 nP_i (1 - P_i) - \sum_{i \neq j} C_i C_j X_i X_j nP_i P_j =$$

$$= n \left\{ \sum_1^N C_i^2 X_i^2 P_i - \sum_1^N C_i^2 X_i^2 P_i^2 - \sum_{i \neq j} C_i C_j X_i X_j P_i P_j \right\}$$

y teniendo en cuenta que,

$\left(\sum_1^N C_i X_i P_i \right)^2 = \sum_1^N C_i^2 X_i^2 P_i^2 + \sum_{i \neq j} C_i C_j X_i X_j P_i P_j$ se obtiene el resultado propuesto.

Supongamos ahora que se desea estimar el total $X = \sum_1^N X_i$.

La condición para que $\hat{X}$ sea insesgado es que los coeficientes C_i tomen el valor $C_i = 1/nP_i$; en efecto:

$E(\hat{X}) = \sum_1^N C_i X_i nP_i = \sum_1^N X_i$, de donde $C_i nP_i = 1$, para todo i, por tanto $C_i = 1/nP_i$. Según esto el estimador insesgado del total X, toma la forma:

$$\hat{X} = \sum_1^n \frac{X_i}{n P_i} \tag{3}$$

y su varianza, teniendo en cuenta la fórmula general (2), será:

$$\hat{V}(\hat{X}) = \frac{1}{n} \left\{ \sum_1^N \frac{X_i^2}{P_i} - X^2 \right\} \tag{4}$$

que también puede escribirse como sigue:

$$\hat{V}(\hat{X}) = \frac{1}{n} \sum_1^N P_i \left(\frac{X_i}{P_i} - X \right)^2 \tag{5}$$

El estimador insesgado de $V(\hat{X})$ es:

$$\hat{V}(\hat{X}) = \frac{1}{n(n-1)} \sum_{1}^{n} \left(\frac{X_i}{P_i} - \hat{X} \right)^2 \qquad (6)$$

En efecto, desarrollando el cuadrado de cada sumando:

$$\sum_{1}^{n} \left(\frac{X_i}{P_i} - \hat{X} \right)^2 = \sum_{1}^{n} \left(\frac{X_i}{P_i^2} + \hat{X}^2 - 2\hat{X} \frac{X_i}{P_i} \right) = \sum_{1}^{n} \frac{X_i}{P_i} - n\hat{X}^2$$

podemos escribir,

$$E(\hat{V}(\hat{X})) = E\left\{ \frac{1}{n(n-1)} \left(\sum_{1}^{N} \frac{X_i^2}{P_i^2} \varepsilon_i - n\hat{X}^2 \right) \right\}$$

y ahora teniendo en cuenta que $V(\hat{X}) = E(\hat{X})^2 - X^2$ y la fórmula (5) de $V(\hat{X})$, se obtiene:

$$E(\hat{V}) = \frac{1}{n(n-1)} \left\{ \sum_{1}^{N} \frac{X_i^2}{P_i^2} E(\varepsilon_i) - n\left(V(\hat{X}) + X^2 \right) \right\} =$$

$$= \frac{1}{n-1} \sum_{1}^{N} \frac{X_i^2}{P_i} - \frac{1}{n(n-1)} \sum_{1}^{N} \frac{X_i^2}{P_i} + \frac{1}{n(n-1)} X^2 - \frac{1}{(n-1)} X^2$$

y simplificando, resulta:

$$E(\hat{V}) = \frac{1}{(n-1)} - \frac{1}{n(n-1)} \sum_{1}^{N} \frac{X_i^2}{P_i} - \left(\frac{1}{(n-1)} - \frac{1}{n(n-1)} \right) X^2 =$$

$$= \frac{1}{n} \left\{ \sum_{1}^{N} \frac{X_i^2}{P_i} - X^2 \right\}$$

La teoría general de muestreo con reposición y probabilidades desiguales fué presentada por Hansen y Hurwitz (1943); en su nombre suelen denominarse a los estimadores aquí estudiados, estimadores H-H.

PROBABILIDADES PROPORCIONALES AL TAMAÑO

Si es M_i el número de unidades que contiene el conglomerado U_i, es interesante el caso en que el procedimiento de muestreo origina probabilidades P_i proporcionales al tamaño M_i; esto es $P_i = M_i/M$. En esta situación el estimador del total admite la forma:

$$\hat{X} = \frac{M}{n} \sum_{1}^{n} \frac{X_i}{M_i}$$

con varianza:

$$V(\hat{X}) = \frac{M}{n}\left\{\sum_{1}^{N} \frac{X_i^2}{M_i} - \frac{X^2}{M}\right\} = \frac{M}{n}\left\{\sum_{1}^{N} M_i (\overline{X}_i - \overline{\overline{X}})^2\right\} \quad (7)$$

cuya estimación insesgada es:

$$\hat{V}(\hat{X}) = \frac{M^2}{n(n-1)} \sum_{1}^{n} (\overline{X}_i - \overline{\overline{x}})^2$$

donde $\overline{X}_i = X_i/M_i$ es la media del i° conglomerado y $\overline{\overline{x}}$ es la media de los n valores $\overline{X}_i$ de la muestra. Esta fórmula se obtiene inmediatamente de (6) haciendo $P_i = M_i/M$.

En la práctica, la asignación de probabilidades desiguales no siempre se hace tomándolas proporcionales al tamaño de los conglomerados, medido éste por el número M_i de unidades menores que contiene. En ocasiones esta medida se hace en relación a otra característica (Y_i) conocida de las unidades primarias; por ejemplo si X_i es la producción de una empresa, Y_i puede ser el número de empleados; si X_i es el valor añadido de una explotación agraria, Y_i puede ser la superficie de cultivo,

etc. En estas condiciones se recomienda tomar aquella característica $\mathcal{Y}$ tal que la relación entre X_i e Y_i sea para todo i, lo más próxima a la proporcionalidad directa; esto se deduce del hecho de que si para todo i es $Y_i = KX_i$ la varianza del estimador se anula; de $Y_i = KX_i$ resulta $P_i = X_i/X$ y por tanto:

$$V(\hat{X}) = \frac{1}{n}\left\{ \sum_1^N \frac{XX_i^2}{X_i} \quad X^2 \right\} = \frac{1}{n}\left\{ \left(\sum_1^N X_i\right)^2 - \left(\sum_1^N X_i\right)^2 \right\} = 0$$

El problema de estimación de la media, $\bar{X}$, de totales de conglomerados, se resuelve inmediatamente ya que por ser:

$\bar{X} = \frac{1}{N} X$, resulta $\hat{\bar{X}} = \frac{1}{N} \hat{X}$ estimador insesgado para $\bar{X}$ y su varianza será:

$$V(\bar{X}) = \frac{1}{N^2} V(\hat{X})$$

El problema de estimación de la media general, $\bar{\bar{X}}$, esto es, la media por unidad elemental supuesto que M_i es el número de ellas en U_i, también se resuelve inmediatamente; por ser

$$\bar{\bar{X}} = \frac{1}{M} \sum_1^N X_i = \frac{1}{M} X$$

resulta que $\hat{\bar{\bar{X}}} = \frac{1}{M} \hat{X}$ es insesgado para $\bar{\bar{X}}$ y su varianza es

$$V(\hat{\bar{\bar{X}}}) = \frac{1}{M^2} V(\hat{X})$$

Para la estimación de la proporción y el total de clase; X_{ij} toma el valor 1 ó 0 según que la unidad j^a del i^o conglomerado pertenezca, o no, a la clase $\mathcal{C}$, entonces podemos identificar:

C	=	(total de clase en la población)	=	X
C_i	=	(total de clase en U_i)	=	$\bar{X}_i$
P_i	=	(proporción en U_i)	=	$\bar{\bar{X}}_i$
P	=	(proporción en la población)	=	$\bar{\bar{X}}$

y obtenemos,

$$\hat{C} = \frac{M}{n} \sum_{1}^{n} \frac{C_i}{M_i} = \frac{M}{n} \sum_{1}^{n} P_i$$

$$\hat{P} = \frac{1}{n} \sum_{1}^{n} \frac{C_i}{M_i} = \frac{1}{n} \sum_{1}^{n} P_i$$

como estimadores insesgados de C y P respectivamente, cuyas varianzas son:

$$V(\hat{C}) = \frac{M}{n} \left\{ \sum_{1}^{N} \frac{C_i^2}{M_i} - \frac{C_i^2}{M} \right\} = \frac{M}{n} \sum_{1}^{N} M_i (P_i - P)^2$$

$$\hat{V}(\hat{P}) = \frac{1}{M^2} \; V(\hat{C})$$

y su estimador insesgado:

$$\hat{V}(\hat{C}) = \frac{M^2}{n(n-1)} \sum_{1}^{n} (P_i - p)^2 \; ; \quad V(\hat{P}) = \frac{1}{M^2} V(\hat{C})$$

donde $p = \bar{\bar{x}} = \frac{1}{n} \sum_{1}^{n} P_i$.

CASOS PARTICULARES

Si todos los conglomerados son del mismo tamaño $P_i = M_i/M =$ CTE; $P_i = \bar{M}/M$; estamos en el caso de muestreo con probabilidades iguales y sustituyendo M_i por $\bar{M}$ y M por $N\bar{M}$ en las fórmulas, se obtienen los resultados estudiados anteriormente. Si además de ser de igual tamaño, los conglomerados contienen una sola unidad, estamos en el

caso de la teoría básica estudiada en el capítulo II. Sustituyendo M por N y $\overline{M}$ por 1, se obtienen las fórmulas deducidas en el citado capítulo. Sirvan estas particularizaciones como ejercicio.

III.9. MUESTREO SIN REPOSICION DE CONGLOMERADOS CON PROBABILIDADES DESIGUALES

Supongamos ahora que la muestra de n unidades primarias se selecciona mediante algún procedimiento aleatorio sin reposición, tal que la probabilidad de que U_i aparezca en la muestra es π_i, y la probabilidad de que el par de unidades (U_i, U_j) aparezcan en cualquier orden en la muestra es π_{ij}. En el capítulo I se han estudiado las propiedades de π_i y π_{ij}, así como los momentos del vector aleatorio $(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ que nos indica cúales unidades aparecen en la muestra, tomando el correspondiente ε_i el valor 1, ó 0 en caso contrario; así pues, sabemos que:

$$E(\varepsilon_i) = \pi_i \; ; \quad E(\varepsilon_i \, \varepsilon_j) = \pi_{ij}$$

$$V(\varepsilon_i) = \pi_i (1 - \pi_i); \quad Cov(\varepsilon_i, \varepsilon_j) = \pi_{ij} - \pi_i \pi_j$$

La teoría de estimadores lineales para muestreo sin reposición con probabilidades desiguales, fué presentada por Horvitz y Thompson (1952), de aquí la denominación de estimadores H - T.

La condición para que el estimador

$$\hat{X} = \sum_1^n C_i \, X_i = \sum_1^N (C_i \, X_i) \, \varepsilon_i$$

sea insesgado para el total X, es que $C_i = 1/\pi_i$ para todo i. En efecto:

$$E(\hat{X}) = \sum_1^N C_i \, X_i \, E(\varepsilon_i) = \sum_1^N C_i \, X_i \, \pi_i \; ;$$

de donde igualando a $X = \sum_{1}^{N} X_i$, se obtiene $C_i = 1/\pi_i$. Por tanto la forma del estimador lineal insesgado del total es:

$$\hat{X} = \sum_{1}^{n} \frac{X_i}{\pi_i} = \sum_{1}^{N} \frac{X_i}{\pi_i} \varepsilon_i$$

esto significa que $\hat{X}$ es la suma de los valores observados en las unidades de la muestra, divididos por la probabilidad que tiene la correspondiente unidad de aparecer en la muestra. En el caso de muestreo con probabilidades iguales, sabemos que $\pi_i = n/N$ y en consecuencia se obtiene el estimador, ya conocido, $\hat{X} = (N/n) \sum_{1}^{n} X_i$.

La varianza de $\hat{X}$ puede expresarse en la forma:

$$V(\hat{X}) = \sum_{1}^{N} \left(\frac{1 - \pi_i}{\pi_i} \right) X_i^2 + \sum_{i \neq j}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) X_i X_j \qquad (1)$$

En efecto:

$$V(\hat{X}) = V\left(\sum_{1}^{N} \frac{X_i}{\pi_i} \varepsilon_i \right) = \sum_{1}^{N} \frac{X_i^2}{\pi_i^2} V(\varepsilon_i) +$$

$$+ \sum_{i \neq j}^{N} \frac{X_i X_j}{\pi_i \pi_j} \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = \sum_{1}^{N} \frac{X_i^2}{\pi_i^2} \pi_i (1 - \pi_i) +$$

$$+ \sum_{i \neq j}^{N} \frac{X_i X_j}{\pi_i \pi_j} (\pi_{ij} - \pi_i \pi_j)$$

de donde simplificando se obtiene el resultado.

Teniendo en cuenta que

$$X^2 = \left(\sum_1^N X_i\right)^2 = \sum_1^N X^2 + \sum_{i \neq j}^N X_i X_j$$, la fórmula (1) de

la varianza también puede expresarse como:

$$V(\hat{X}) = \sum_1^N \frac{X_i^2}{\pi_i} + \sum_{i \neq j}^N \frac{\pi_{ij}}{\pi_i \pi_j} X_i X_j - X^2 \qquad (2)$$

Según hemos visto anteriormente, en muestreo con reposición si las probabilidades de selección de cada unidad, en cada extracción, P_i, son exactamente proporcionales a los valores X_i, la varianza del estimador $\hat{X}$ se anula. En muestreo sin reposición también ocurre un hecho análogo; si las probabilidades π_i son exactamente proporcionales a X_i, la varianza de $\hat{X}$ se anula. Si en la fórmula (2) hacemos $\pi_i = n \frac{X_i}{X}$, resulta:

$$V(\hat{X}) = \sum_1^N \frac{X_i X}{n} + \sum_{i \neq j}^N \frac{\pi_{ij} X^2}{n^2} - X^2 =$$

$$= \frac{X^2}{n} + \frac{X^2}{n^2} \sum_{i \neq j}^N \pi_{ij} - X^2$$, y teniendo en cuenta (véase cap. I.7)

que $\sum_{i \neq j} \pi_{ij} = \sum_i (n-1)\pi_i = n(n-1)$ resulta, sustituyendo este resultado, que $V(\hat{X}) = 0$. También puede deducirse inmediatamente lo anterior, escribiendo la varianza de la forma:

$$V(\hat{X}) = \frac{1}{2} \sum_{i \neq j}^N (\pi_i \pi_j - \pi_{ij}) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_i} \right)^2 \qquad (3)$$

ya que todos los términos cuadráticos son nulos cuando, para todo i

es π_i proporcional a X_i. Para obtener la fórmula (3) partimos de (1) teniendo en cuenta la relación, estudiada en Cap. I.7:

$$\sum_{i,j\neq i} (\pi_i \ \pi_j - \pi_{ij}) = \pi_i(1-\pi_i); \text{ ahora despejamos } (1-\pi_i)$$

y lo sustituimos en (1):

$$V(\hat{X}) = \sum_{i=1}^{N} \sum_{j\neq i}^{N} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_i^2} \right) X_i^2 - \sum_{i=1}^{N} \sum_{j\neq i}^{N} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_i \pi_j} \right) X_i X_j =$$

$$= \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{X_i^2}{\pi_i^2} + \frac{X_j^2}{\pi_j^2} \right) - \sum_{i=1}^{N} \sum_{j>i}^{N} 2(\pi_i \pi_j - \pi_{ij}) \frac{X_i \ X_j}{\pi_i \ \pi_j} =$$

$$= \sum_{i=1}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j} \right)^2 =$$

$$= \frac{1}{2} \sum_{i\neq j}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_i} \right)^2$$

ESTIMACION INSESGADA DE LA VARIANZA

Si nos detenemos en la fórmula (1) de la varianza, se deduce inmediatamente el estimador insesgado:

$$\hat{V}(\hat{X}) = \sum_{1}^{n} \frac{1-\pi_i}{\pi_i^2} X_i^2 + \sum_{i\neq j}^{n} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij} \pi_i \pi_j} X_i X_j \qquad (4)$$

puesto que $E(\varepsilon_i) = \pi_i$; $E(\varepsilon_{ij}) = \pi_{ij}$ donde ε_{ij} toma el valor

1 ó 0 según que el par $(A_i\ ,\ A_j)$ esté presente, o no, en la muestra.

Otra expresión equivalente puede obtenerse de (2) razonando de la siguiente forma:

$$\hat{V}(\hat{X}) = \hat{X}^2 - B$$

donde B representa un estimador insesgado de X^2 y $\hat{X}^2$ un estimador insesgado de los dos primeros sumandos de (2). En efecto:

$$E(\hat{X}^2) = E\left(\sum_1^n \frac{X_i}{\pi_i}\right)^2 = E\left(\sum_1^n \frac{X_i^2}{\pi_i^2} + \sum_{i\neq j}^n \frac{X_i X_i}{\pi_i \pi_j}\right) =$$

$$= \sum_1^N \frac{X_i^2}{\pi_i} + \sum_{i\neq j}^N \frac{X_i X_j}{\pi_i \pi_j} \pi_{ij}$$

y por otra parte B es:

$$B = \sum_1^n \frac{X_i^2}{\pi_i} + \sum_{i\neq j}^n \frac{X_i X_j}{\pi_{ij}} \quad \text{, puesto que}$$

$$E(B) = \sum_1^N X_i^2 + \sum_{i\neq j}^N X_j X_j = \left(\sum_1^N X_i\right)^2 = X^2$$

por tanto el estimador insesgado de $V(\hat{X})$ resulta también:

$$\hat{V}(\hat{X}) = \hat{X}^2 - \sum_1^n \frac{X_i^2}{\pi_i} - \sum_{i\neq j}^n \frac{X_i X_i}{\pi_{ij}} \qquad (5)$$

Es obvio que también puede obtenerse (5) de (4) sin más que un breve desarrollo y conveniente agrupación de los sumandos.

Una estimación insesgada de la varianza distinta de la proporcionada por (4) ó (5), ha sido propuesta por Yates y Grundy (1953); si nos fijamos en la fórmula (3) de $V(\hat{X})$ se deduce inmediatamente que:

$$\hat{V}(\hat{X}) = \frac{1}{2} \sum_{i \neq j}^{n} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j} \right)^2, \text{ o bien}$$

$$\hat{V}(\hat{X}) = \sum_{i>j}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j} \right)^2 \qquad (6)$$

es insesgado para $V(\hat{X})$; recordemos que $E(\varepsilon_i \varepsilon_j) = \pi_{ij}$.

III.10. EJEMPLOS

Ejemplo III.1

Suponiendo que en una comunidad de 300.000 personas la proporción de desempleados es 0,08 y la población se encuentra en N = 600 conglomerados de $\overline{M}$ = 500 personas cada uno, determínese el tamaño de una muestra de conglomerados, sin reposición, que garantice la estimación del total de desempleados con un error de muestreo del 5 %. Se dispone de la siguiente estimación de σ_w^2 = 0,0720.

Solución:

n' = 100 (con reposición) ; n = 75 (sin reposición).

Ejemplo III.2

Si una muestra del tamaño calculado en el ejercicio anterior, proporciona la estimación $\hat{X}$ = 265000, suponiendo distribución normal para $\hat{X}$. ¿Es admisible la hipótesis P = 0,08 con un coeficiente de confianza del 95 % ?.

Solución:

El intervalo de confianza de $I_{0,95}$ = (21458; 26542) bajo la hipótesis P = 0,08; dado que 26500 ϵ $I_{0,95}$, se acepta.

Ejemplo III.3

En una encuesta continua efectuada por conglomerados de tamaño $\overline{M}$ = 200, se ha estimado que la relación σ_b^2/σ^2 vale 0,02451. El tamaño de la muestra que se viene utilizando es n = 80, con un coste total, dado por la función C = 200 n $\overline{M}$ + 25000 n, de 5.200.000 pts. Se desea reducir el coste a 5.000.000 pts. Determinense los tamaños óptimos y el incremento del error de muestreo de la media por unidad.

Solución:

Admitiendo la función de ajuste $\sigma_b^2 = \sigma^2 \bar{M}^{\alpha}$, se obtiene $n = 60$; $\bar{M} = 292$ y un incremento del error de muestreo de $\bar{\bar{x}}$, del 1,14 % .

Ejemplo III.4

Determínese el coeficiente de homogeneidad de conglomerados, en la situación del ejercicio III.1., y el incremento relativo del error de muestreo respecto al muestreo por unidades.

Solución:

$\hat{\delta} = 0{,}01978$; el error de muestreo es 329 % del que se obtendría en muestreo por unidades; es decir, 3,29 veces superior.

Ejemplo III.5

Determínese, en la situación del ejemplo III.3., el coeficiente de homogeneidad de conglomerados en función de $\bar{M}$, y calcúlese para $\bar{M} = 292$ el incremento del error de muestreo por conglomerados, respecto al muestreo por unidades.

Solución:

De $\sigma_b^2 = \sigma^2 (1 + (\bar{M} - 1)\delta)$ resulta $\delta = \dfrac{M^{\alpha + 1} - 1}{\bar{M} - 1}$.

El error de muestreo por conglomerados sería el 234, 3 % respecto a muestreo por unidades.

Ejemplo III.6

De una población con 5000 unidades distribuidas en conglomerados de $\bar{M} = 10$ unidades cada uno, se conoce el dato $\sigma_b^2 = 20000$. Uno de los conglomerados, sea A_g , presenta una media mayor que los demás

del orden de $\bar{X}_g = 8\ \bar{\bar{X}}$, y se decide tomar una muestra aleatoria sin reposición de tamaño (n − 1) entre los restantes y añadir fijo el grande. Se pide: Formar el estimador insesgado para la media $\bar{\bar{X}}$, calcular su varianza en función de n y $\bar{\bar{X}}$ y comprobar si existe más precisión con este procedimiento que tomando una muestra del mismo tamaño sobre toda la población, en el caso $\bar{\bar{X}} = 200$, $n = 60$.

Solución:

Denotemos por σ^2_{ob} la varianza "entre", correspondiente a los (N − 1) conglomerados después de eliminar A_g y por $\hat{\bar{\bar{X}}}^*$ el estimador insesgado de X. Se tiene:

$$\sigma^2_{ob} = \frac{N}{N-1}\left(\sigma^2_b - \frac{49\,\bar{\bar{X}}^2}{N}\right); \qquad \hat{\bar{\bar{X}}}^* = \frac{1}{N}\left(\bar{X}_g + (N-1)\,\bar{\bar{x}}_o\right)$$

$$V\left(\hat{\bar{\bar{X}}}^*\right) = \frac{(N-1)^2\,(N-n)\,\sigma^2_{ob}}{N^2\,(N-2)\,(n-1)} \quad . \text{Para } \bar{\bar{X}} = 200,\ n = 60$$

resulta: $V(\bar{\bar{x}}) = 294$; $V\left(\hat{\bar{\bar{X}}}^*\right) = 240$; por tanto si se toma A_g fijo en la muestra, se reduce el error de muestreo al 90,4 % ; esto es, disminuye un 9,6 % .

Ejemplo III.7

De una población que tiene 30.000 unidades en 200 conglomerados de tamaño iguales, se ha tomado una muestra aleatoria de $n = 8$ sin reposición, que presenta las siguientes proporciones de unidades de una clase:

P_i : 0,05; 0,04; 0,05; 0,03; 0,03; 0,06; 0,05; 0,04

Estímese el total C en la población y el error de muestreo relativo.

Solución:

$\hat{C} = 1313$; error de muestreo 8,4 %

Ejemplo III.8

Estímese el coeficiente δ de homogeneidad de conglomerados y la varianza por unidad, PQ, con los datos del ejercicio III.7.

Solución:

$$s^2_{1w} = 0{,}0394875; \qquad s_b = 0{,}0001125$$

$$\hat{PQ} = 0{,}0396; \qquad \hat{\delta} = 0{,}0003865$$

Ejemplo III.9

Teniendo en cuenta las fórmulas para π_i y π_{ij} estudiadas en el cap. I.7, compruebe que la fórmula general de la varianza, aplicada al caso de probabilidades iguales y muestreo sin reposición coincide con la obtenida en el capítulo II.

Ejemplo III.10

Supongamos una población de N = 5 unidades primarias, en la que X_i toma los valores:

3; 3; 4; 6; 8 para $i = 1, 2, 3, 4, 5$ respectivamente.

Se toma una muestra de tamaño $n = 2$, asignando en la primera extracción probabilidades proporcionales a 10; 16; 16; 25; 33 y también

en la segunda, prescindiendo de la unidad seleccionada en primer lugar.

1°. Calcule π_{ij} para todo para $i \neq j$. 2°. Compruebe que $\sum_{i=1}^{5} \pi_i = 2$.

3°. Obtenga la distribución del estimador insesgado del total.

Solución:

1°. $\pi_{12} = 0{,}03683$ $\pi_{23} = 0{,}06095$ $\pi_{34} = 0{,}10095$

$\pi_{13} = 0{,}03683$ $\pi_{24} = 0{,}10095$ $\pi_{35} = 0{,}14166$

$\pi_{14} = 0{,}06111$ $\pi_{25} = 0{,}14166$ $\pi_{45} = 0{,}23313$

$\pi_{15} = 0{,}08592$

2°. $\pi_1 = 0{,}22069$

$\pi_2 = 0{,}34039$

$\pi_3 = 0{,}34039$ $\sum \pi_i = 1{,}99998$

$\pi_4 = 0{,}49614$

$\pi_5 = 0{,}60237$

3°. Tenemos 10 posibles muestras de tamaño $n = 2$, con las probabilidades calculadas en 1°. cuyas correspondientes estimaciones son:

$\hat{X}_{12} = 22{,}41$ $\hat{X}_{23} = 20{,}56$ $\hat{X}_{34} = 23{,}84$

$\hat{X}_{13} = 25{,}34$ $\hat{X}_{24} = 20{,}91$ $\hat{X}_{35} = 25{,}03$

$\hat{X}_{14} = 25{,}69$ $\hat{X}_{25} = 22{,}09$ $\hat{X}_{45} = 25{,}37$

$\hat{X}_{15} = 26{,}87$

Ejemplo III.11

Si en la situación del ejemplo III.9, la muestra seleccionada es la formada por el par (A_1, A_4), calcule la estimación de $V(\hat{X})$.

Solución:

El estimador de Yates y Grundy, nos da:

$\hat{V}(\hat{X}) = 0{,}08115$, por tanto el error de muestreo relativo, estimado por la muestra dada es:

$$\frac{\sqrt{\hat{V}(\hat{X})}}{\hat{X}} \times 100 = 3{,}68\,\%$$

Ejemplo III.12

Resuelva las cuestiones del ejercicio III.10, suponiendo que la primera extracción se hace con probabilidades proporcionales a los números dados y la segunda, sin reposición, con probabilidades iguales.

Solución:

1°)

$\pi_{12} = 0{,}065$ $\quad \pi_{23} = 0{,}08$ $\quad \pi_{34} = 0{,}1025$

$\pi_{13} = 0{,}065$ $\quad \pi_{24} = 0{,}1025$ $\quad \pi_{35} = 0{,}1225$

$\pi_{14} = 0{,}0875$ $\quad \pi_{25} = 0{,}1225$ $\quad \pi_{45} = 0{,}145$

$\pi_{15} = 0{,}1075$

2°)

$\pi_1 = 0{,}325$

$\pi_2 = 0{,}37$

$\pi_3 = 0{,}37$

$\pi_4 = 0{,}4375$

$\pi_5 = 0{,}4975$

$\Big\}$ $\sum_{1}^{5} \pi_i = 2$

3°) Tenemos también 10 posibles muestras de tamaño $n = 2$, con las probabilidades calculadas en 1°), cuyas correspondientes estimaciones son:

$$\hat{X}_{12} = 17{,}34 \qquad \hat{X}_{23} = 18{,}92 \qquad \hat{X}_{34} = 24{,}53$$

$$\hat{X}_{13} = 20{,}04 \qquad \hat{X}_{24} = 21{,}82 \qquad \hat{X}_{35} = 26{,}90$$

$$\hat{X}_{14} = 22{,}95 \qquad \hat{X}_{25} = 24{,}19 \qquad \hat{X}_{45} = 29{,}80$$

$$\hat{X}_{15} = 25{,}31$$

Ejemplo III.13

Si en el ejemplo III.12, la muestra seleccionada es la formada por el par (A_1 , A_4), calcule la estimación de $V(\hat{X})$.

Solución:

El estimador de Yates y Grundy, nos da:

$\hat{V}(\hat{X}) = 6{,}28185$; por tanto el error de muestreo relativo, estimado por la citada muestra es: 10,44 %.

Ejemplo III.14

Se extrae una muestra sin reposición y con probabilidades iguales de tamaño $n = 2$, de una población que consta de tres unidades con valores X_1, X_2, X_3. Para estimar la media $\overline{X}$ se considera el estimador definido en cada muestra de la siguiente forma:

$$\left.\begin{aligned} T(S_1) &= \alpha X_1 + \gamma X_2 \\ T(S_2) &= \beta X_1 + \gamma X_3 \\ T(S_3) &= \gamma X_2 + \gamma X_3 \end{aligned}\right\}$$

1°) Determine los coeficientes α, β, γ con la condición de que T sea insesgado.

2°) Compruebe que dentro de la clase de estimadores lineales insesgados estudiada en el punto 1°), no existe uno de varianza mínima que sea independiente de los valores poblacionales X_i .

Ejemplo III.15

De una población finita de tamaño N, se extrae una muestra, con reposición y probabilidades iguales, de tamaño $n = 3$.

1°) Calcular la probabilidad de que la muestra esté formada por tres elementos iguales (γ_1), por dos iguales y uno distinto (γ_2) ó por tres elementos distintos (γ_3).

2°) Si es $\bar{x}'$ la media muestral de los elementos distintos, probar que su varianza es:

$$V(\bar{x}) = \frac{(2N-1)(N-1)}{6N^2} S^2$$

3°) Compruebe que si $\bar{x}$ es la media de todos los elementos de la muestra, $V(\bar{x}') < V(\bar{x})$.

Solución:

$$\gamma_1 = \frac{1}{N^2} \; ; \quad \gamma_2 = \frac{3(N-1)}{N^2} \; ; \quad \gamma_3 = \frac{(N-1)(N-2)}{N^2}$$

con estos datos es suficiente para completar el ejercicio.

Ejemplo III.16

De una población con N = 20 unidades se ha extraido una muestra

de tamaño $n = 3$. Los valores observados en los elementos de la muestra son $X_7 = 30$; $X_{16} = 40$; $X_{23} = 35$.

Estime la media de la población y la varianza del estimador, en los casos:

a) Muestreo con reposición y probabilidades iguales

b) Muestreo sin reposición y probabilidades iguales

c) Muestreo con reposición y probabilidades desiguales, siendo:

$$P_7 = 0{,}04; \qquad P_{16} = 0{,}06; \qquad P_{23} = 0{,}05$$

Solución:

a) $\bar{X} = 35;$ $\qquad \hat{V}(\hat{\bar{X}}) = 8{,}33$

b) $\hat{\bar{X}} = 35;$ $\qquad \hat{V}(\hat{\bar{X}}) = 7{,}08$

c) $\hat{\bar{X}} = 35{,}61;$ $\qquad \hat{V}(\hat{\bar{X}}) = 1{,}49$

Ejemplo III.17

Una población consta de $M = 40.000$ unidades distribuidas en $N = 400$ conglomerados de $\bar{M} = 100$ unidades cada uno. Una muestra aleatoria con probabilidades iguales, sin reposición, de tamaño $n = 25$ conglomerados presenta los siguientes datos:

Total de unidades pertenecientes a una clase C	Número de conglomerados de la muestra
12	2
17	3
23	9
33	5
36	6

1°) Estime el total de unidades de la población pertenecientes a la clase $\mathcal{C}$, y el error de muestreo relativo.

2°) Estime la eficiencia de este diseño, respecto del muestreo por unidades.

Solución:

1°) $\hat{C} = 10608$; Error de muestreo 5,88 %

2°) Error de muestreo por unidades 3,22 % . Por tanto el error de muestreo por conglomerados es el 182,61 % del error de muestreo por unidades.

Ejemplo III.18

Es una población finita que consta de N = 750 unidades, la proporción que pertenece a una clase $\mathcal{C}$ es P. Consideremos dos formas de agrupación de las unidades en N conglomerados de igual tamaño $\bar{M}$.

Agrupación I : N = 15 y $\bar{M}$ = 50

Agrupación II : N = 10 y $\bar{M}$ = 75

Supongamos ahora que el total de unidades pertenecientes a C en cada conglomerado es:

Agrupación I	*Agrupación II*
22, 12, 23, 11, 25	28, 36, 35, 25, 26
27, 14, 20, 23, 10	31, 37, 24, 34, 24
26, 21, 16, 24, 26	

1°) Calcule la descomposición de la varianza total en sus componentes entre y dentro, en cada caso.

2°) Calcule el coeficiente de homogeneidad de conglomerados, en cada caso.

3°) Si en el caso I se toma una muestra de 6 conglomerados, cuantos habrán de tomarse en el caso II para tener muestras de igual tamaño total de unidades. En esta situación, calcule la eficiencia del diseño II respecto del I y respecto del muestreo por unidades.

Solución:

1°) $0{,}24 = 0{,}2272 + 0{,}0128$; $\sigma^2 = \sigma_w^2 + \sigma_b^2$ } I

$0{,}24 = 0{,}2357 + 0{,}0043$; $\sigma^2 = \sigma_w^2 + \sigma_b^2$ } II

2°) $\delta_I = 0{,}0342$; $\delta_{II} = 0{,}0046$

3°) $1 + (\bar{M}-1)\delta_I = 2{,}6758$; $1 + (\bar{M}-1)\delta_{II} = 1{,}3404$

por tanto el error de muestreo en el caso II es un 115,8 % del error de muestreo por unidades y en el caso I es el 163,6 %. El diseño II respecto del I da un error de muestreo del 70,8 %.

IV. MUESTREO ESTRATIFICADO

IV.1. ESTRATIFICACION DE LA POBLACION. MUESTREO ESTRATIFICADO

Hemos visto en el capítulo II que la precisión de los estimadores depende del tamaño de la muestra y de la varianza en la población de la característica en estudio. Si el muestreo es por conglomerados, hemos visto en el capítulo III que también depende del grado de homogeneidad de éstos; para simplificar, tenemos en muestreo con reposición:

$$V(\bar{x}) = \frac{\sigma^2}{\bar{M}\, n} \quad (1 + (\bar{M} - 1)\, \delta)$$

fórmula que resume éstas ideas. El tamaño de la muestra es $\bar{M}n$, la varianza poblacional es σ^2, $\bar{M}$ el tamaño de los conglomerados y δ su coeficiente de homogeneidad. Para $\bar{M} = 1$, se convierte en la conocida fórmula $V(\bar{x}) = \sigma^2/n$ de muestreo aleatorio por unidades simples, con reposición. En la prática deben formarse los conglomerados tanto en su composición interna, como en su tamaño, de modo que la pérdida de precisión debida al factor $(\bar{M} - 1)\, \delta$ (cuando $\delta > 0$) sea lo menor posible, teniendo en cuenta los costes y las posibilidades materiales de formación de conglomerados.

En este capítulo estudiaremos una técnica que además de ser útil para dar estimaciones independientes acerca de subconjuntos de la

población, puede producir notables ganancias de precisión. La varianza σ^2 es muy sensible a valores extremos de la población, aunque sean poco frecuentes; por ello el muestreo alcanza su mayor eficacia cuando se emplea para estimar características de poblaciones constituidas por unidades similares. Partiendo de esta idea, la estratificación de la población consiste en agrupar sus unidades (que pueden ser conglomerados) en un cierto número, L, de clases disjuntas denominadas *estratos*, constituidas por unidades similares: con esto se persigue que la varianza en cada uno de ellos sea pequeña.

El *muestreo estratificado* consiste en seleccionar L muestras independientes, cada una de un estrato. El procedimiento de muestreo puede ser distinto de un estrato a otro, pero debe ser efectuado independiente. Generalmente no se usan procedimientos de selección distintos de un estrato a otro, con una interesante excepción que consiste en que de los L estratos, uno de ellos está constituido por las unidades que deseamos entren con certeza en la muestra; en dicho estrato no se muestrea sino que se observan todas sus unidades.

El procedimiento obliga a que entren en las n unidades de la muestra determinado número $(n_1, n_2, \ldots, n_L)$ de unidades de cada estrato; el muestreo estratificado es pues un procedimiento de muestreo restricto (o restringido). De las $\binom{N}{n}$ posibles muestras de tamaño n pasamos a tener un número menor, dado por el producto.

$$\binom{N_1}{n_1}\binom{N_2}{n_2}\cdots\cdots\binom{N_L}{n_L} < \binom{N}{n}$$

siendo $N_1, N_2, \ldots N_L$ los tamaños de los estratos, N el tamaño de la población y n el tamaño de la muestra.

Intuitivamente se aprecia que si los estratos se corresponden con clases de unidades similares, las muestras estratificadas serán más representativas que las muestras irrestrictas, aunque esta afirmación debe matizarse ya que, como veremos, la precisión de las estimaciones depende del reparto de la muestra a los estratos, del procedimiento probabilístico de selección y del estimador utilizado.

La discusión acerca de cuantos estratos se deben formar y cuales son los valores de características conocidas de la población que nos delimiten unos de otros, no tiene una solución matemática de fácil aplicación práctica y validez general; en cada caso hay que estudiar el problema concreto.

La forma óptima de delimitar los estratos es, obviamente, mediante valores de la propia variable que es objeto de estudio pero también es claro que la información necesaria no está disponible; por ello debe emplearse alguna otra variable lo más relacionada posible, de forma que permita clasificar las unidades en estratos homogéneos. Así por ejemplo el tamaño de las empresas de una rama de actividad (Número de empleados) con respecto a producción, venta, valor añadido, etc.

En cuanto al número óptimo de estratos que se deben formar, la decisión depende naturalmente de la distribución en la población de la característica objetivo y también, como en toda investigación, de la información marco disponible y del coste. Puede comprobarse que teóricamente al aumentar el número L de estratos, aumenta la precisión de las estimaciones en relación directa con L; esto es:

$$V(\hat{\theta}) = \frac{K}{L^2}$$

cuando la delimitación de los estratos se realiza respecto de la variable objetivo. En la realidad, esta delimitación se hace respecto de otra variable distinta y en consecuencia la anterior afirmación deja de ser cierta. Los estudios realizados, suponiendo que la variable utilizada para delimitar los estratos tiene una fuerte correlación lineal con la variable objetivo, muestran que los incrementos de precisión cuando L es mayor que 5 ó 6 son ya muy pequeños; por esta razón, en general, se recomienda tomar L entre 2 y 5 si no hay otras razones que obliguen a formar un mayor número de estratos.

Si, como es habitual, se desean estimaciones relativas a varias características $\mathfrak{X}$, $\mathfrak{Y}$, $\mathfrak{Z}$, etc. de la población, deberá estudiarse por separado el número y la delimitación de estratos; si la solución encontrada para una de ellas satisface los requerimientos de precisión para las restantes, hemos resuelto el problema; en otro caso tendremos que

adoptar una solución de compromiso, de acuerdo con los criterios de prioridad, coste, etc. Si a pesar de esto, alguna característica requiere una formación de estratos muy diferente, deberá pensarse en plantear y resolver por separado el problema de su estimación, reduciendo el objetivo a las restantes.

Acerca de estos temas puede verse un interesante estudio en Delenius (1957).

Como hemos dicho antes, la estratificación permite obtener estimaciones independientes acerca de subconjuntos de la población. Aunque no exista unanimidad en la nomenclatura, llamaremos *dominio de estudio* a un estrato, o conjunto de varios estratos, cuyas unidades constituyen un colectivo del que se requieren estimaciones por separado. En ocasiones se desea obtener estimaciones de algún subconjunto de la población que no constituye un dominio de estudio, en el sentido de que la muestra no se ha diseñado con ese fin; en este caso diremos que se trata de una *subpoblación*.

IV.2. ESTIMACION DE LA MEDIA $\bar{X}$, EL TOTAL X, LA PROPORCION P Y EL TOTAL DE CLASE C.

Utilizando el subíndice $h = 1, 2, \ldots, L$, para referirnos a cada uno de los L estratos, es inmediato traducir las notaciones empleadas en capítulos anteriores, especialmente el Cap. II.

Tanto en muestreo con reposición como sin reposición un estimador insesgado de X es:

$$\hat{\bar{X}} = \sum_{h=1}^{L} W_h \bar{x}_h$$

donde $W_h = N_h/N$ es el tamaño relativo del h° estrato. En efecto:

$$E(\hat{\bar{X}}) = E\left(\sum_{h=1}^{L} W_h \bar{x}_h\right) = \sum_{h=1}^{L} E(W_h \bar{x}_h) = \sum_{h=1}^{L} W_h E(\bar{x}_h) =$$

$$= \sum_{h=1}^{L} W_h \bar{X}_h = \frac{1}{N} \sum_{h=1}^{L} N_h \bar{X}_h = \frac{X}{N} = \bar{X}.$$

Teniendo en cuenta resultados ya conocidos, por ser muestras independientes las extraídas en los L estratos, la varianza de $\hat{\bar{X}}$ es:

$$V_{CR}(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2 \, V_{CR}(\bar{x}_h) = \sum_{h=1}^{L} W_h^2 \, \frac{\sigma_h^2}{n_h} \quad ; \text{ o bien}$$

$$V_{SR}(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_n} \cdot \frac{S^2}{n_h} \text{ según se trate}$$

de muestreo con o sin reposición:

Para el total X, se deduce inmediatamente el estimador insesgado.

$$\hat{X} = N \hat{\bar{X}} = \sum_{h=1}^{L} N_h \bar{x}_h$$

cuya varianza será $V(\hat{X}) = N^2 \, V(\hat{\bar{X}})$.

También teniendo en cuenta resultados conocidos, estimadores insesgados de la varianza son:

$$\hat{V}_{CR}(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2 \, \frac{s_h^2}{n_h} \quad ; \quad \text{o bien}$$

$$\hat{V}_{SR}(\hat{X}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \, \frac{s_h^2}{n_h}$$

siendo s_h^2 la cuasivarianza muestral del h° estrato. Análogamente,

$$\hat{V}(\hat{X}) = N^2 \, \hat{V}(\hat{X})$$

Para estimar la proporción, P, de unidades de una cierta clase en la población, siendo C el total de clase, tenemos:

$$X_{hi} = \begin{cases} 1 \text{ Si la unidad } i \text{ del estrato } h, \text{ pertenece a la clase.} \\ \\ 0 \text{ En caso contrario.} \end{cases}$$

entonces:

$$X = C\,; \quad \bar{X} = \frac{C}{N} = P\,; \quad \sigma^2 = P(1-P) = PQ\,; \quad S^2 = \frac{N}{N-1} PQ$$

y análogamente para cada estrato y para los datos muestrales. Por otra parte, son válidos los resultados obtenidos anteriormente identificando la media $\bar{X}$ con P y el total X con $C = NP$. Así, resulta:

En muestreo con probabilidades iguales, sea con o sin reposición, un estimador insesgado de P es:

$$\hat{P} = \sum_{h=1}^{L} W_h p_n\,; \quad p_h = \frac{c_h}{n_h} = \text{proporción muestral en el h° estrato}$$

cuya varianza viene dada por:

$$V(\hat{P})_{CR} = \sum_{h=1}^{L} W_h^2 \frac{P_h Q_h}{n_h}\,; \quad \text{(con reposición)}$$

$$V(\hat{P})_{SR} = \sum_{h=1}^{L} \frac{N_h - n_h}{N_h - 1} W_h^2 \frac{p_h Q_h}{n_h}\,; \text{(sin reposición)}$$

cuyos estimadores insesgados son:

$$\hat{V}(\hat{P})_{CR} = \sum_{h=1}^{L} W_h^2 \frac{P_h q_h}{n_h - 1}$$

$$\hat{V}(\hat{P})_{SR} = \sum_{h=1}^{L} \frac{N_h - n_h}{N_h} W_h^2 \frac{p_h q_h}{n_h - 1}$$

Para el total de clase, el estimador insesgado es $\hat{C} = N\hat{P}$; su varianza $V(\hat{C}) = N^2 V(\hat{P})$ y la estimación insesgada de ésta $\hat{V}(\hat{C}) = N^2 \hat{V}(\hat{P})$.

IV.3. TAMAÑO DE LA MUESTRA. AFIJACION OPTIMA

Se llama *afijación* de la muestra al reparto de n en L tamaños $n_1, n_2, \ldots, n_h, \ldots, n_L$ de modo que $n = \sum_{h=1}^{L} n_h$, siendo n_h el tamaño de muestra asignado al h° estrato. El problema del tamaño de la muestra está pues íntimamente ligado al de la afijación.

La afijación se llama *igual* cuando se asigna a todos los estratos igual tamaño de muestra; esto es:

$$n_h = \frac{n}{L} \; ; \qquad h = 1, 2, \ldots, L$$

En muestreo sin reposición con probabilidades iguales, la afijación igual es equivalente a exigir que probabilidad de que una unidad aparezca en la muestra sea inversamente proporcional al tamaño del estrato a que pertenece; en efecto, puesto que dicha probabilidad es la de ser elegida en la muestra de su estrato,

$$\Pr(A_{hi}) = \frac{n_h}{N_h} = \frac{n}{L\,N_h} = \frac{\text{constante}}{N_h}$$

y recíprocamente, si

$$\Pr(A_{hi}) = \frac{K}{N_h} \; ; \quad \text{por ser} \sum_{h=1}^{L} \sum_{1=i}^{N_h} \Pr(A_{hi}) = n, \text{ véase Cap. I,}$$

resulta

$$K = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \frac{1}{N_h} = K \sum_{h=1}^{L} N_h \cdot \frac{1}{N_h} = K \sum_{h=1}^{L} 1 = KL = n$$

de donde $K = \frac{n}{L}$, por tanto $\Pr(A_{hi}) = \frac{n}{L\,N_n}$

La afijación se llama *proporcional* cuando se asigna a cada estrato un tamaño de muestra proporcional al tamaño del estrato; esto es,

$$n_h = K N_h \; ; \qquad h = 1, 2, \ldots, L$$

y por ser $n = \sum_{h=1}^{L} n_h$; $N = \sum_{h=1}^{L} N_h$, la afijación proporcional equivale a exigir que la proporción de unidades del h° estrato en la muestra, sea igual a la proporción de unidades de dicho estrato en la población:

$$w_h = W_h \; ; \; \frac{n_h}{N} = \frac{N_h}{N} \; ; \quad h = 1, 2, \ldots, L$$

En muestreo sin reposición con probabilidades iguales, la afijación proporcional es equivalente a exigir que todas las unidades de la población tengan la misma probabilidad de aparecer en la muestra; en efecto, puesto que la probabilidad de una unidad es la de ser elegida en el estrato a que pertenece,

$$\Pr(A_{hi}) = \frac{n_h}{N_h} = \frac{n}{N} = \text{constante}$$

y recíprocamente, si

$$\Pr(A_{hi}) = K; \text{ por ser } \sum_{h=1}^{L} \sum_{i=1}^{N_h} \Pr(A_{hi}) = n, \text{ resulta}$$

$$\sum_{h=1}^{L} \sum_{i=1}^{N_h} K = K \sum_{h=1}^{L} N_h = KN = n; \text{ de donde } K = \frac{n}{N} = \text{cons-}$$

tante para toda A_{hi}.

Con afijación proporcional la varianza del estimador de la media es:

$$V(\hat{\bar{X}})_{EP} = \frac{1}{n} \sum_{h=1}^{L} W_h \, \sigma_h^2 \; ; \text{ (con reposición)}$$

$$V(\hat{\bar{X}})_{EP} = \frac{N-n}{Nn} \sum_{h=1}^{L} W_h S_h^2 \; ; \text{ (sin reposición)}$$

fórmulas que se obtienen inmediatamente, sustituyendo $n_h = W_h n$ en las generales para $V(\hat{\bar{X}})$ en muestreo estratificado. La varianza de los estimadores $\hat{X}$, $\hat{P}$ y $\hat{C}$ se obtiene de las anteriores sin dificultad.

Una propiedad interesante de la afijación proporcional es que, para los estimadores lineales estudiados de medias y totales, todas las observaciones muestrales vienen multiplicadas por un factor constante. En cualquier otra afijación, el estimador de la media es la suma de medias muestrales ponderadas por el tamaño relativo del estrato; por tanto las observaciones vienen multiplicadas por el factor W_h/n_h, que varía con cada estrato; esto es:

$$\hat{\bar{X}} = \sum_h W_h \bar{x}_h = \sum_h W_h \frac{x_h}{n_h} \; ; \text{ siendo } x_h \text{ el total muestral}$$

h°. Si la afijación es proporcional, por ser $n_h = W_h n$, resulta:

$$\hat{\bar{X}}_{EP} = \frac{1}{n} \sum_h x_h = \frac{1}{n} x = \bar{x}$$

esto es, $\hat{\bar{X}}_{EP}$ equivale a la media simple de las n observaciones muestrales. Para el estimador del total basta multiplicar x por N/n, que es también un "factor de elevación" constante. Este hecho tiene importancia práctica ya que simplifica el proceso de cálculo de las estimaciones.

TAMAÑO DE LA MUESTRA CON AFIJACION IGUAL Y CON AFIJACION PROPORCIONAL

AFIJACION IGUAL

Si se desea estimar $\bar{X}$ con una varianza dada, sea V_o, el tamaño necesario de la muestra se obtiene de la ecuación:

$$V(\hat{\bar{X}})_{EI(CR)} = \frac{L}{n} \sum_h W_h^2 \sigma_h^2 = V_o \; ; \text{ de donde resulta:}$$

$$n = \frac{L \sum_h W_h^2 \sigma_h^2}{V_o} \; ; \text{ (con reposición).}$$

En muestreo sin reposición,

$$V(\hat{\bar{X}})_{EI(SR)} = \frac{L}{n} \sum_h W_h^2 S_h^2 - \frac{1}{N} \sum_h W_h S_h^2 = V_o \text{ ; resulta:}$$

$$n = \frac{L \sum_h W_h^2 S_h^2}{V_o + \frac{1}{N} \sum_h W_h S_h^2} \quad ; \text{ (sin reposición)}$$

AFIJACION PROPORCIONAL. Operando de forma análoga:

$$V(\hat{\bar{X}})_{EP(CR)} = \frac{1}{n} \sum_h W_h \sigma_h^2 = V_o \text{ ; resulta}$$

$$n = \frac{\sum_h W_h \sigma_h^2}{V_o} \quad ; \text{ (con reposición)}$$

y por ser:

$$V(\hat{\bar{X}})_{EP(SR)} = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_h W_h S_h^2 \text{, el tamaño de la muestra}$$

para muestreo sin reposición resulta ser, tomando $\sigma_h^2 \simeq S_h^2$:

$$n = \frac{n'}{1 + n'/N}$$

de donde n′ es el que corresponde a muestreo con reposición para la misma precisión.

AFIJACION OPTIMA

La afijación se llama *óptima* cuando cumple algún requisito de mínimo, bajo unas condiciones dadas. Por ejemplo:

A) Varianza mínima para n dado.

B) Tamaño de muestra mínimo para Varianza dada.

C) Varianza mínima para un coste dado.

D) Coste mínimo para Varianza dada.

Consideremos el caso A). Utilizando el método de los multiplicadores de Lagrange formamos la función:

$$\phi = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

para $V(\hat{\bar{X}})$ en muestreo con reposición. Derivando respecto de n_1, $n_2, \ldots, n_h, \ldots, n_L$ y λ e igualando a cero las derivadas, se obtiene:

$$\begin{cases} \dfrac{-W_h^2 \sigma_h^2}{n_h^2} + \lambda = 0 \text{ ; para } h = 1, 2, \ldots, L \\ n = \sum_{h=1}^{L} n_h \end{cases}$$

de donde resulta:

$$\frac{W_1^2 \sigma_1^2}{n_1^2} = \frac{W_2^2 \sigma_2^2}{n_2^2} = \ldots\ldots = \frac{W_h^2 \sigma_h^2}{n_h^2} = \ldots\ldots = \frac{W_L^2 \sigma_L^2}{n_L^2}$$

extrayendo la raíz cuadrada e invirtiendo las fracciones, se conservan las igualdades y teniendo en cuenta la ecuación de condición, resulta:

$$\frac{n}{W_1 \sigma_1} = \frac{n_2}{W_2 \sigma_2} = \ldots\ldots\ldots = \frac{n_h}{W_h \sigma_h} = \ldots\ldots = \frac{n_L}{W_L \sigma_L} =$$

$$= \frac{n}{\sum_{h=1}^{L} W_h \sigma_h}$$

de donde se obtiene la afijación óptima, denominada de Neyman:

$$n_h = \frac{W_h\ \sigma_h}{\sum_{h=1}^{L} W_h\ \sigma_h}\ .\ n\ ; \qquad \text{(con reposición)}$$

y el valor mínimo de la varianza, sustituyendo n_h en $V(\hat{\bar{X}})$:

$$V(\hat{\bar{X}})_{E.OP} = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\ \sigma_h\right)^2 ; \text{(con reposición)}$$

Si el muestreo es sin reposición, el segundo sumando de la varianza:

$$V(\hat{\bar{X}}) = \sum_{h=1}^{L} W_h^2\ \frac{S_h^2}{n_h} - \frac{1}{N}\sum_{h=1}^{L} W_h\ S_h^2$$

es independiente de los n_h, por tanto las ecuaciones del sistema anterior son análogas, apareciendo S_h^2 en vez de σ_h^2 . Resulta:

$$n_h = \frac{W_h\ S_h}{\sum_{h=1}^{L} W_h\ S_h}\ n\ ; \qquad \text{(sin reposición)}$$

y la varianza mínima:

$$V(\hat{\bar{X}})_{E.OP} = \frac{1}{n}\left(\sum_{h=1}^{L} W_h\ S_h\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h\ S_h^2 ; \text{(sin reposición)}$$

Si tratamos de resolver el caso B) por el mismo procedimiento, se obtiene

$$n_h = \frac{W_h\ \sigma_h}{\sum_{h=1}^{L} W_h\ \sigma_h}\ .\ n\ ; \quad V_o = \sum_{h=1}^{L} W_h^2\ \frac{\sigma_h^2}{n_h}$$

de donde sustituyendo n_h en la ecuación de condición, resulta el valor de n mínimo:

$$n = \frac{\left(\sum_{h=1}^{L} W_h \sigma_h \right)^2}{V_o} \quad ; \quad \text{(con reposición)}$$

En muestreo sin reposición

$$n = \frac{\left(\sum_{h=1}^{L} W_h S_h \right)^2}{V_o + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2} \quad ; \quad \text{(sin reposición)}$$

De acuerdo con los resultados obtenidos la afijación óptima, tanto en el caso de varianza mínima para tamaño de muestra dado como en el de tamaño de muestra mínimo para varianza dada, consiste en asignar a cada estrato un tamaño de muestra proporcional al producto del tamaño del estrato por su desviación típica; esto es:

$$n_h = K W_h \sigma_h \ , \quad \text{o bien} \quad n_h = K N_h \sigma_h$$

y puede comprobarse fácilmente que en muestreo con probabilidades iguales sin reposición equivale a exigir que la probabilidad de que una unidad aparezca en la muestra, sea proporcional a la desviación típica del estrato a que pertenece; en efecto

$$\Pr(A_{hi}) = \frac{n_h}{N_h} = \frac{\sigma_h}{\sum N_h \sigma_h} \cdot n = K \sigma_h$$

y recíprocamente, si

$$\Pr(A_{hi}) = \lambda \sigma_h \ , \text{ por ser } \sum_h \sum_i \Pr(A_{hi}) = n \ , \text{ resulta}$$

$$\lambda = \frac{n}{\sum_h N_h \sigma_h} \ ;$$

de donde se obtiene:

$$\frac{n_h}{N_h} = \frac{\sigma_h}{\sum_h N_h \sigma_h} n \Longrightarrow n_h = \frac{N_h \sigma_h}{\sum_h N_h \sigma_h} \cdot n$$

Consideremos ahora los casos C) y D), en los que se hace intervenir el coste de la muestra en relación con la precisión de las estimaciones. Suponiendo que c_h ; $h = 1, 2, \ldots, L$ es el coste por unidad muestral en cada estrato, el coste total de la muestra viene dado por:

$$C = \sum_h c_h n_h$$

En el caso C) Varianza mínima para coste dado, siendo C_o el valor prefijado para el coste total, utilizando como antes el método de los multiplicadores de Lagrange:

$$\phi = \sum_h W_h^2 \frac{\sigma_h^2}{n_h} + \lambda \left(\sum_h c_h n_h - C_o \right)$$

para estimación de la media en muestreo con reposición. Derivando respecto a $n_1, n_2, \ldots, n_h, \ldots, n_L$ y λ e igualando a cero las derivadas, se obtiene el sistema de ecuaciones:

$$\begin{cases} \dfrac{-W_h^2 \sigma_h^2}{n_h^2} + \lambda c_h = 0 \; ; \text{ para } \quad h = 1, 2, \ldots, L \\ \\ \sum_h c_h n_h = C_o \end{cases}$$

de donde resulta (despejando $1/\sqrt{\lambda}$):

$$\frac{n_1}{W_1 \sigma_1 / \sqrt{c_1}} = \ldots\ldots = \frac{n_h}{W_h \sigma_h / \sqrt{c_h}} = \ldots\ldots = \frac{n_L}{W_L \sigma_L / \sqrt{c_L}} = \frac{n}{\sum_h W_h \sigma_h / \sqrt{c_h}}$$

por tanto la afijación óptima para costes variables de un estrato a otro, es:

$$n_h = \frac{W_h \, \sigma_h / \sqrt{c_h}}{\sum_h W_h \, \sigma_h / \sqrt{c_h}} \cdot n$$

y puesto que el coste está fijado en una cantidad C_o, sustituyendo n_h en la ecuación de condición resulta para el tamaño de la muestra:

$$n = \frac{C_o \sum_h W_h \, \sigma_h / \sqrt{c_h}}{\sum_h W_h \, \sigma_h / \sqrt{c_h}}$$

En muestreo sin reposición se obtiene, operando de forma análoga:

$$n_h = \frac{W_h \, S_h / \sqrt{c_h}}{\sum_h W_h \, S_h / \sqrt{c_h}} \, n \quad : \quad n = \frac{C_o \sum_h W_h \, S_h / \sqrt{c_h}}{\sum_h W_h \, S_h / \sqrt{c_n}}$$

Si se trata de resolver el caso D) Coste mínimo para varianza dada, siendo V_o el valor prefijado para la varianza en caso de estimación de la media, operando de la misma forma que en C) se llega a una afijación idéntica:

$$n_h = \frac{W_h \, \sigma_h / \sqrt{c_h}}{\sum_h W_h \, \sigma_h / \sqrt{c_h}} \, n$$

pero ahora por estar prefijada la varianza, obtenemos n después de sustituir los n_h en la ecuación de condición; así, resulta:

$$\frac{1}{n} \sum_h \frac{W_h^2 \, \sigma_h^2}{W_h \, \sigma_h / \sqrt{c_h}} \cdot \sum_h W_h \, \sigma_h / \sqrt{c_h} = V_o$$

de donde se obtiene:

$$n = \frac{\left(\sum_h W_h \sigma_h \sqrt{c_h} \right) \left(\sum_h W_h \sigma_h \sqrt{c_h} \right)}{V_o}$$

Si el muestreo es sin reposición se llega a la misma afijación, con S_h en vez de σ_h en las fórmulas, y el tamaño de la muestra resulta ser:

$$n = \frac{\left(\sum_h W_h S_h / \sqrt{c_h} \right) \left(\sum_h W_h S_h / \sqrt{c_h} \right)}{V_o + 1/N \sum_h W_h S_h^2}$$

Observemos que si los costes unitarios son iguales en todos los estratos, los resultados obtenidos en los casos C) y D) se reducen a los obtenidos en A) y B). Cuando los costes no son iguales, la afijación óptima asigna a cada estrato un tamaño de muestra proporcional al producto de su tamaño por su desviación típica e inversamente proporcional a la raíz cuadrada del coste unitario en dicho estrato.

IV.4. EFICIENCIA DEL MUESTREO ESTRATIFICADO.

Vamos a comparar la varianza del estimador de la media en muestreo estratificado, con afijación proporcional y afijación óptima, con el muestreo irrestricto por unidades (esto es, sin la previa estratificación de la población). El análisis se instrumenta de forma sencilla utilizando muestreo con reposición; para ello efectuemos en primer lugar la siguiente descomposición de la varianza, dentro de estratos y entre estratos:

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{\bar{X}})^2 =$$

$$= \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} \left[(X_{hi} - \bar{X}_h) + (\bar{X}_h - \bar{\bar{X}}) \right]^2 =$$

$$= \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2 + \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (\bar{X}_h - \bar{\bar{X}})^2 ;$$

por ser nula la suma de los productos cruzados.

$$= \frac{1}{N} \sum_{h=1}^{L} N_h \sigma_h^2 + \frac{1}{N} \sum_{h=1}^{L} N_h (\bar{X}_h - \bar{\bar{X}})^2 =$$

$$= \sum_{h=1}^{L} W_h \sigma_h^2 + \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{\bar{X}})^2 .$$

De acuerdo con resultados obtenidos anteriormente, podemos escribir las varianzas:

$$V(\bar{x})_{MI} = \frac{\sigma^2}{n} \quad ; \text{ muestreo irrestricto}$$

$$V(\hat{\bar{X}})_{EP} = \frac{1}{n} \sum_{h=1}^{L} W_h \sigma_h^2 \quad ; \text{ muestreo estratificado con afijación proporcional.}$$

$$V(X)_{E.OP} = \frac{1}{n} \left(\sum_{h=1}^{L} W_h \sigma_h \right)^2 \quad ; \text{ muestreo estratificado con afijación óptima.}$$

y teniendo en cuenta la descomposición de σ^2, podemos escribir:

$$V(\bar{x})_{MI} = V(\hat{\bar{X}})_{EP} + \frac{1}{n} \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{\bar{X}})^2$$

donde se observa que si las medias de los estratos no son todas iguales, la afijación proporcional, en estratificación, es más eficiente que el

muestreo irrestricto con el mismo tamaño de muestra. Esta ganancia en precisión es tanto mayor cuanto más variables sean las medias de los estratos.

Comparemos ahora la afijación óptima con la proporcional:

$$V(\hat{\bar{X}})_{EP} - V(\hat{\bar{X}})_{E.OP} = \frac{1}{n}\left\{ \sum_{h=1}^{L} W_h \sigma_h^2 - \left(\sum_{h=1}^{L} W_h \sigma_h \right)^2 \right\} =$$

$$= \frac{1}{n} \sum_{h=1}^{L} W_h \left(\sigma_h - \bar{\sigma}\right)^2$$

siendo $\bar{\sigma}$ la media ponderada de las desviaciones típicas de los estratos. Se tiene, entonces:

$$V(\hat{\bar{X}})_{EP} = V(\hat{\bar{X}})_{E.OP} + \frac{1}{n} \sum_{h=1}^{L} W_h \left(\sigma_h - \bar{\sigma} \right)^2$$

donde se observa que si las desviaciones típicas de los estratos no son todas iguales, la afijación óptima es más eficiente que la proporcional, para un mismo tamaño de muestra. Esta ganancia en precisión es tanto mayor cuanto más variables sean las desviaciones típicas de los estratos.

Como resumen, podemos escribir:

$$V(\bar{x})_{MI} = V(\hat{\bar{X}})_{E.OP} + \frac{1}{n} \sum_{h} W_h \left(\bar{X}_h - \bar{X}\right)^2 + \frac{1}{n} \sum_{h} W_h \left(\sigma_h - \bar{\sigma}\right)^2$$

fórmula que recoge los dos términos positivos que expresan el incremento de eficiencia de la afijación óptima respecto del muestreo irrestricto de unidades.

En muestreo sin reposición, admitiendo las aproximaciones: $N_h - 1 \simeq N_h$, $N - 1 \simeq N$; podemos descomponer S^2 de forma análoga a σ^2:

$$S^2 \simeq \sum_h W_h S_h^2 + \sum_h W_h (\bar{X}_h - \bar{X})^2$$

y siendo ahora:

$$V(\bar{x})_{MI} = \frac{S^2}{n} - \frac{S^2}{N}$$, y $\hat{V}(\bar{X})_{EP}$, $\bar{V}(\hat{\bar{X}})_{E.OP}$ ya obtenidas

en IV.3, resulta

$$V(\bar{x})_{MI} = V(\hat{\bar{X}})_{EP} + \frac{1}{n} (1-f) \sum_h W_h (\bar{X}_h - \bar{X})^2$$

cuya diferencia con el caso con reposición radica en el factor de finitud $(1-f)$ que afecta al segundo sumando. La comparación de las afijaciones óptima y proporcional resulta inmediata:

$$V(\hat{\bar{X}})_{EP} - V(\hat{\bar{X}})_{E.OP} = \frac{1}{n} \sum_h W_h S_h^2 - \frac{1}{n} \left(\sum_h X_h S_h \right)^2 =$$

$$= \frac{1}{n} \sum_h W_h (S_h - \bar{S})^2$$

siendo $\bar{S}$ la media ponderada de las cuasidesviaciones típicas de los estratos; por tanto la única diferencia con el caso con reposición radica en que aparecen S_h en vez de σ_h . Por último:

$$V(\bar{x})_{MI} = V(\hat{\bar{X}})_{E.OP} + \frac{1}{n} (1-f) \sum_h W_h (\bar{X}_h - \bar{X})^2 + \frac{1}{n} \sum_h W_h (S_h - \bar{S})^2$$

fórmula, análoga al caso de muestreo con reposición, que recoge los dos términos positivos que expresan el incremento de eficiencia de la afijación óptima respecto al muestreo irrestricto de unidades, uno debido a la variabilidad de las medias de los estratos y otro a la variabilidad de las cuasivarianzas de los estratos.

IV.5. ESTIMACION INSESGADA DE σ^2 Y S^2 A PARTIR DE UNA MUESTRA ESTRATIFICADA

Si se desea obtener una estimación de σ^2 o de S^2, cuando se dispo-

ne de los datos obtenidos mediante una muestra estratificada, la solución que consiste en calcular la cuasivarianza de los n datos muestrales:

$$s^2 = \frac{1}{n-1} \sum_{h,i} (X_{hi} - \bar{x})^2$$

no proporciona una estimación insesgada. Si la afijación es proporcional, la muestra se reparte entre los estratos de forma aproximada a una muestra aleatoria en la población y la fórmula anterior puede dar una buena estimación; en otro caso esta solución no es buena.

Veamos cómo puede obtenerse una estimación insesgada de σ^2 en muestreo con reposición. Sabemos que σ^2 puede descomponerse de la siguiente forma:

$$\sigma^2 = \Sigma\ W_h\ \sigma_h^2 + \Sigma\ W_h\ (\bar{X}_h - \bar{\bar{X}})^2$$

el primer sumando no presenta problema puesto que s_h^2 es estimador insesgado de σ_h^2. En cuanto al segundo sumando, sustituyamos $\bar{X}_h$ y $\bar{\bar{X}}$ por sus estimaciones y calculemos la esperanza matemática:

$$E \left\{ \Sigma\ W_h\ (\bar{x}_h - \hat{\bar{X}})^2 \right\} =$$

$$E \left\{ \Sigma\ W_h \left[(\bar{X}_h - \bar{\bar{X}}) + (\bar{x}_h - \bar{X}_h) - (\hat{\bar{X}} - \bar{\bar{X}}) \right]^2 \right\} =$$

$$E \left\{ \Sigma\ W_h\ (\bar{X}_h - \bar{\bar{X}})^2 \right\} + E \left\{ \Sigma\ W_h (\bar{x}_h - \bar{X}_h)^2 \right\} +$$

$$+ E \left\{ \Sigma\ W_h\ (\hat{\bar{X}} - \bar{\bar{X}})^2 \right\} - 2\,E \left\{ \Sigma\ W_h\ (\bar{x}_h - \bar{X}_h)\ (\hat{\bar{X}} - \bar{\bar{X}}) \right\}$$

dado que $\Sigma\ W_h\ (\bar{X}_h - \bar{\bar{X}})$ se anula.

Por otra parte $\Sigma\ W_h (\bar{x}_h - \bar{X}_h)\ (\hat{\bar{X}} - \bar{\bar{X}}) = (\hat{\bar{X}} - \bar{\bar{X}})^2$; $\Sigma\ W_h = 1$ los dos últimos sumandos equivalen a:

$-\ (\hat{\bar{X}} - \bar{X})^2$ cuya esperanza es:

$$E\left(-(\hat{\bar{X}} - \bar{X})^2\right) = -V(\hat{\bar{X}}) = \Sigma\ W_h^2\ \frac{\sigma_h^2}{n_h} \quad ;$$

El segundo sumando nos da:

$$E\left\{\Sigma\ W_h\ (\bar{x}_h - \bar{X}_h)^2\right\} = \Sigma\ W_h\ V(\bar{x}_h) = W_h\ \frac{\sigma_h^2}{n_h} \quad ,$$

asi resulta:

$$E\left\{\Sigma\ W_h\ (\bar{x}_h - \hat{\bar{X}})^2\right\} = \Sigma\ W_h\ (\bar{X}_h - \bar{X})^2 + \Sigma\ W_h\ \frac{\sigma_h^2}{n_h} - \Sigma\ W_h^2 \frac{\sigma_h^2}{n_h}$$

$$= \Sigma\ W_h\ (\bar{X}_h - \bar{X})^2 + \Sigma\ W_h\ (1 - W_h)\ \frac{\sigma_h^2}{n_h}$$

por tanto, el segundo sumando de esta última fórmula es el sesgo de $\Sigma\ W_h\ (\bar{x}_h - \hat{\bar{X}})^2$ como estimador de $\Sigma\ W_h\ (\bar{X}_h - \bar{X})^2$, y en consecuencia el estimador insesgado de σ^2 en muestreo con reposición, es:

$$\hat{\sigma}^2 = \Sigma\ W_h\ s_h^2 + \Sigma\ W_h\ (\bar{x}_h - \hat{\bar{X}})^2 - \Sigma\ W_h\ (1 - W_h)\ \frac{s_h^2}{n_h}$$

Si el muestreo es sin reposición, S^2 se descompone de la siguiente forma:

$$S^2 = \sum \frac{(N_h - 1)\ S_h^2}{(N - 1)} + \sum \frac{N_h}{(N - 1)}\ (\bar{X}_h - \bar{X})^2$$

y aplicando un razonamiento análogo se obtiene:

s_h^2 es estimador insesgado de S_h^2

$$\sum \frac{N_h}{(N-1)} (x_h - \hat{\bar{X}})^2 - \sum (1-f_h)(1-W_h) \frac{N_h}{(N-1)} \frac{s_h^2}{n_h}$$

es estimador insesgado del segundo sumando de S^2, por tanto:

$$\hat{S}^2 = \sum \frac{(N_h-1)}{(N-1)} s_h^2 + \sum \frac{N_h}{(N-1)} (x_h - \hat{\bar{X}})^2 - \sum (1-f)(1-W_h) \frac{N_h}{(N-1)} \frac{s_h^2}{n_h}$$

es un estimador insesgado de S^2 en muestreo sin reposición. Para σ^2 el estimador insesgado, es:

$$\hat{\sigma}^2 = \sum \frac{(N_h-1)}{N} s_h^2 + \sum W_h (\bar{x}_n - \hat{\bar{X}})^2 - \sum (1-f_h)(1-W_h) W_h \frac{s_h^2}{n_h}$$

donde simplificando, se obtiene:

$$\hat{\sigma}^2 = \sum W_h s_h^2 - \sum \frac{W_h s_h^2}{n_h} + \sum \frac{W_h^2 s_h^2}{n_h} - \sum \frac{W_h s_h^2}{N} +$$

$$+ \sum W_h \bar{x}_h^2 - (\Sigma W_h \bar{x}_h)^2 \qquad \text{(i)}$$

teniendo en cuenta que:

$$W_h (\bar{x}_h - \hat{\bar{X}})^2 = \Sigma W_h \bar{x}_h^2 - (\Sigma W_h \bar{x}_h)^2$$

y que:

$$(1-f)(1-W_h) \frac{W_h}{n_h} = \left(\frac{W_h}{n_h} - \frac{W_h^2}{n_h} - \frac{1}{N} + \frac{W_h}{N} \right)$$

La fórmula (i) admite algunas simplificaciones. Admitiendo que la población no es muy pequeña $N > 50$, el cuarto término de $\hat{\sigma}^2$ puede

omitirse ya que es un N–ésimo del primero; así resulta:

$$\hat{\sigma}^2 \simeq \Sigma\, W_h\, s_h^2 - \Sigma\, \frac{W_h s_h^2}{n_h} + \Sigma\, \frac{W_h^2 s_h^2}{n_h} + \Sigma\, W_h \bar{x}_h^2 - (\Sigma\, W_h \bar{x}_h)^2 \qquad \text{(ii)}$$

o bien:

$$\hat{\sigma}^2 \simeq \Sigma\, \frac{(n_h - 1)}{n_h}\, W_h\, s_h^2 + \Sigma\, \frac{W_h^2 s_h^2}{n_h} + \Sigma\, W_h \bar{x}_h^2 - (\Sigma\, W_h \bar{x}_h)^2$$

Si en todos los estratos el tamaño de la muestra es relativamente grande: $n_h > 50$ para todo h, puede omitirse en (ii) el segundo y tercer sumandos, resultando:

$$\hat{\sigma}^2 \simeq \Sigma\, W_h\, s_h^2 + \Sigma\, W_h\, \bar{x}_h^2 - (\Sigma\, W_h\, \bar{x}_h)^2 \qquad \text{(iii)}$$

Si la afijación es proporcional, por ser $n_h/n = W_h$ y teniendo en cuenta la descomposición de s^2 :

$$s^2 = \frac{\Sigma\, (X_{hi} - \bar{x})^2}{n-1} = \frac{1}{n-1}\left\{\Sigma\, (n_h - 1)\, s_h^2 + \Sigma\, n_h \bar{x}_h^2 - \frac{(\Sigma\, n_h \bar{x}_h)^2}{n}\right\}$$

resulta que admitiendo $n - 1 \simeq n$; $n_h - 1 \simeq n_h$; s^2 es prácticamente igual a $\hat{\sigma}^2$ en su forma (iii); en consecuencia s^2 es un buen estimador de σ^2 si la fijación es proporcional y si los términos en $1/n_h$ de (ii) son despreciables.

IV.6. ESTIMADORES DE RAZON EN MUESTREO ESTRATIFICADO

Existen dos formas de plantear un estimador de razón para el total X de la población. Una de ellas consiste en formar un estimador de razón del total de cada estrato y sumar estos totales estimados; con este procedimiento el estimador obtenido se denomina *estimador separado de razón.* Este planteamiento requiere el conocimiento del total de

la variable auxiliar, Y_h, en cada estrato por separado y supone que la razón de $\mathfrak{X}$ a $\mathfrak{Y}$ varía de un estrato a otro.

Otra forma de establecer un estimador de razón en muestreo estratificado consiste en obtener una sola razón con los totales $\hat{X}$ e $\hat{Y}$ estimados y multiplicarla por el total de la variable auxiliar, Y, que se supone conocido; se denomina *estimador combinado de razón*. Este planteamiento no requiere el conocimiento de los totales de la variable auxiliar en cada estrato, sino solamente el total de dicha variable en la población, pero supone implícitamente que la razón permanece constante de un estrato a otro.

EL ESTIMADOR SEPARADO

El estimador separado del total X, puede escribirse de cualquiera de las siguientes formas:

$$\hat{X}_{RS} = \sum_{h=1}^{L} \hat{R}_h Y_h = \sum_{h=1}^{L} \frac{x_h}{y_h} Y_h = \sum_{h=1}^{L} \frac{\bar{x}_h}{\bar{y}_h} Y_h$$

De acuerdo con los resultados obtenidos para la varianza del estimador de razón en el Capítulo II.9, tratando por separado cada estrato podemos escribir, para muestreo sin reposición:

$$V(\hat{X}_{RS}) = \sum_{h=1}^{L} V(\hat{R}_h) Y_h^2 =$$

$$= \sum_{h=1}^{L} \frac{N_h^2 (1-f_h)}{n_h} (S_{xh}^2 + R_h^2 S_{yh}^2 - 2\rho_h R_h S_{xh} S_{yh}) =$$

$$= \sum_{h=1}^{L} \frac{N_h^2 (1-f_h)}{n_h} (S_{xh}^2 + R_h^2 S_{yh}^2 - 2 R_h S_{xy,h})$$

Esta fórmula es válida cuando el tamaño de la muestra en cada estrato es suficientemente grande para admitir la aproximación de la fórmula de $V(\hat{R}_h)$.

El problema que debe tenerse en cuenta en relación con el uso de este estimador separado, es que si los sesgos de los estimadores de razón de cada estrato van en el mismo sentido (son del mismo signo), producirán en el estimador un sesgo acumulado del orden de L veces del sesgo medio de cada estrato. Sin embargo, la varianza será del mismo orden que la varianza media del estimador de razón de cada estrato; en consecuencia, la contribución del sesgo el error cuadrático medio puede ser importante.

En muestreo con reposición desaparece el factor $(1-f_h)$ de las fórmulas y las cuasivarianzas se sustituyen por varianzas. En cuanto a la estimación de la varianza, se aplica a cada estrato la fórmula obtenida en el capítulo II y se suman.

RELACION ENTRE EL ESTIMADOR LINEAL Y EL ESTIMADOR SEPARADO EN MUESTREO ESTRATIFICADO

Para estimar el total X, el estimador lineal estratificado hemos visto que puede escribirse como:

$$\hat{X}_L = N \sum_h W_h \bar{x}_h = \sum_h N_h \bar{x}_h = \sum_h \frac{N_h}{n_h} x_h$$

Si nos fijamos en la última forma de escribir $\hat{X}_L$, observamos que consiste en la suma de los totales muestrales de cada estrato multiplicados por su factor de elevación que es precisamente N_h/n_h; esto es, número de unidades en cada estrato dividido por el número de unidades muestrales en dicho estrato.

En ocasiones se forman estimadores para muestreo estratificado en los que el factor de elevación no está prefijado a priori, tomando una forma:

$$\sum_h F_h x_h$$

donde $F_h = \frac{M_h}{m_h}$ es una razón aproximadamente igual a $\frac{N_h}{n_h}$ cuando M_h, m_h representa el valor (total y muestral) de una caracterís-

tica del estrato que es proporcional al número de unidades. Por ejemplo: si M_h, m_h es el número de personas (total y muestral) en el estrato y N_h, n_h es el número de viviendas; entonces la hipótesis

$$\frac{M_h}{m_h} = \frac{N_h}{n_h}$$

que supone un promedio de personas por vivienda constante en el estrato y en la muestra (y que todas las M_h personas del estrato habitan viviendas incluídas en la lista que se muestrea) convierte el estimador $\hat{X}_L$ en un estimador de razón separado:

$$\hat{X}_S = \sum_h \frac{x_h}{m_h} M_n = \sum_h \frac{M_h}{m_h} x_h = \sum_h F_h x_h$$

que puede dar buenos resultados si la relación entre la característica observada X_{hi} y m_{hi} es una recta pasando por el origen, para cada estrato.

EL ESTIMADOR COMBINADO.

El estimador combinado puede escribirse en cualquiera de las formas:

$$\hat{X}_{RC} = \frac{\hat{X}_e}{\hat{Y}_e} Y = \frac{\hat{\bar{X}}_e}{\hat{\bar{Y}}_e} Y = \hat{R}_c Y$$

donde el suíndice "e" indica estimadores lineales en muestreo estratificado.

Utilizando el mismo procedimiento que en Cáp. II.8., podemos estudiar el sesgo de $\hat{X}_{RC}$ de la siguiente forma:

$$\text{Cov}(\hat{R}_c, \hat{Y}_e) = E\left(\frac{\hat{X}_e}{\hat{Y}_e} \hat{Y}_e\right) - E(\hat{R}_c) E(\hat{Y}_e) = X - Y E(\hat{R}_c)$$

y teniendo en cuenta que el sesgo puede escribirse como:

$$B(\hat{X}_{RC}) = E(\hat{X}_{RC}) - X = Y\,E(\hat{R}_c) - X$$

resulta el sesgo de $\hat{X}_{RC}$, en relación a su error de muestreo:

$$\frac{B(\hat{X}_{RC})}{\sigma(\hat{X}_{RC})} = \frac{-\,\mathrm{Cov}(\hat{R}_c,\ \hat{Y}_e)}{\sigma(\hat{X}_{RC})} = \frac{-\,\rho(\hat{R}_c,\ \hat{Y}_e)\,.\,\sigma(\hat{R}_c)\,\sigma(\hat{Y}_e)}{\sigma(\hat{X}_{RC})} =$$

$$= \frac{-\,\rho(\hat{R}_c,\ \hat{Y}_e)\,\sigma(\hat{Y}_e)}{Y} \quad ,$$

por ser $\sigma(\hat{X}_{RC}) = Y\,\sigma(\hat{R}_c)$

de donde resulta, por ser $|\rho(\hat{R}_c,\ \hat{Y}_e)| \leqslant 1$:

$$\frac{|B(\hat{X}_{RC})|}{(\hat{X}_{RC})} \leqslant \frac{\sigma(\hat{Y}_e)}{Y} = \text{coef. de variación de } \hat{Y}_e$$

por tanto el sesgo del estimador combinado de razón será despreciable, frente a su error de muestreo, si el coeficiente de variación de $\hat{Y}_e$ es pequeño, digamos menor del 10%.

La varianza de $\hat{X}_{RC}$, bajo la suposición de que el tamaño de la muestra es suficientemente grande, toma una forma análoga a la del estimador separado con la diferencia de que en

$V(\hat{X}_{RC})$ solamente interviene la razón $R = \dfrac{X}{Y} = \dfrac{\bar{X}}{\bar{Y}}$ y no cada una de las razones de los estratos. Veamos su cálculo:

$$\hat{X}_{RC} - X = \frac{\hat{\bar{X}}_e}{\hat{\bar{Y}}_e}\ .\ Y - X = \frac{N\bar{Y}}{\hat{\bar{Y}}_e}\,(\hat{\bar{X}}_e - R\hat{\bar{Y}}_e)$$

ahora suponemos que puede aproximarse por 1 la razón $\bar{Y}/\hat{\bar{Y}}_e$ en muestras grandes, supuesto $\hat{\bar{Y}}_e$ insesgado para $\bar{Y}$; tenemos:

$$\hat{X}_{RC} - X \simeq N\,(\hat{\bar{X}}_e - R\,\hat{\bar{Y}}_e)$$

consideremos ahora la variable

$$Z_{hi} = X_{hi} - R\,Y_{hi}$$

definida para cada una de las unidades de la población; $h = 1,2,\ldots L$; $i = 1, 2, \ldots, N_h$; entonces podemos denotar por $\bar{z}_e$ a la media ponderada de Z_{hi} en una muestra estratificada:

$$\bar{z}_e = \sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} Z_{hi} = \sum_{h=1}^{L} W_h \cdot \bar{z}_h$$

y por ser cero la media de $Z_{h,i}$:

$$\sum_{h,i} Z_{hi} = \sum_{h,i} X_{hi} - R \sum_{h,i} Y_{hi} = X - \frac{X}{Y} \cdot Y = 0$$

resulta:

$$V(\hat{X}_{RC}) = E(\hat{X}_{RC} - X)^2 \simeq N^2\,E(\bar{z}_e) = N^2\,V(\bar{z}_e);$$

$$V(\hat{X}_{RC}) = N^2\,V\left(\sum_h W_h \cdot \bar{z}_h\right) = \sum_h N^2\,W_h^2\,V(\bar{z}_h)$$

En muestreo con reposición, es:

$$V(\bar{z}_h) = \frac{1}{n_h}\left\{\sigma_{x,h}^2 + R^2\,\sigma_{y,h}^2 - 2R\sigma_{x,y;h}\right\}$$

de donde:

$$V(\hat{X}_{RC}) = \sum_{h=1}^{L} \frac{N_h^2}{n_h}\,(\sigma_{x,h}^2 + R^2\,\sigma_{y,h} - 2R\,\sigma_{x,y;h})$$

y en muestreo sin reposición

$$\hat{V}(\bar{X}_{RC}) = \sum_{h=1}^{L} \frac{N_h^2 (1-f_h)}{n_h} (S_{x,h}^2 + R^2 S_{y,h}^2 - 2RS_{x,y;h})$$

Para obtener estimaciones de la varianza, tanto del estimador separado como del combinado, sustituímos R_h y R así como las varianzas y covarianzas, por sus correspondientes estimaciones, en las fórmulas obtenidas de

$$\hat{V}(\bar{X}_{RS}) \quad y \quad \hat{V}(\bar{X}_{RC})$$

La comparación entre ambos estimadores debe hacerse en base a dos elementos fundamentales:

a) el tamaño de la muestra en cada estrato y b) la variabilidad de R_h de estrato a estrato. Si en cada estrato la muestra no es suficientemente grande como para admitir la validez de la varianza calculada, es recomendable el uso del estimador combinado; también será recomendable el estimador combinado si la razón R_h es constante en todos los estratos. Si por el contrario R_h varía de estrato a estrato es preferible el estimador separado, tanto más, cuanto más próxima a una recta pasando por el origen ($X_{hi} = R_{hi}\ Y_{hi}$) esté la relación entre X e Y en cada estrato. También el estimador separado tendrá ventajas cuando se trabaje con muestras grandes en todos los estratos. Digamos por último que no debe olvidarse el obtener evidencia empírica a priori, acerca de si el sesgo en los diferentes estratos es del mismo signo, lo cual sería un inconveniente para el estimador separado, o por el contrario los sesgos se equilibran entre unos y otros estratos.

IV.7 UNIDADES QUE ENTRAN CON CERTEZA EN LA MUESTRA

Por diversas razones puede requerirse al plantear una investigación por muestreo que algunas unidades de la población deben entrar con certeza en la muestra. Esta situación se presenta p.e. en poblaciones en las que la distribución de la característica en estudio es asimétrica, con un cierto número de unidades grandes cuya contribución al total de la población es muy notable; en otros casos, se desean datos por

separado de ciertas unidades y por tanto deben incluirse en la muestra.

Si las unidades en cuestión pueden identificarse, formamos un estrato con ellas, sea el L°, y el tamaño de muestra en dicho estrato será $n_L = N_L$, lo que significa que se observan todas sus unidades; el muestreo aleatorio se efectúa entonces en los (L−1) estratos restantes, siendo el tamaño de la muestra para este conjunto de estratos $n' = n - n_L = n - N_L$. Denotaremos por N', X', $\bar{X}'$ el número de unidades, el total y la media de este conjunto de (L−1) estratos, y en general utilizaremos los símbolos usuales afectados del signo $'$.

Por ser:

$$X = X' + X_L \qquad \text{y} \qquad \bar{X} = \frac{N'\,\bar{X}' + X_L}{N}$$

el total y la media de la población, los estimadores insesgados son:

Para $\bar{X}$:

$$\hat{\bar{X}} = \frac{N'\,\hat{\bar{X}}' + X_L}{N} \quad , \quad \text{donde} \quad \hat{\bar{X}}' = \sum_{1}^{L-1} W'_h\,\bar{x}_h$$

es estimador insesgado de la media de los (L−1) primeros estratos, siendo $W'_h = N_h / N'$.

Para X:

$$\hat{X} = N'\,\hat{\bar{X}}' + X_L = \sum_{1}^{L-1} N_h\,\bar{x}_h + X_L$$

Las varianzas se obtienen inmediatamente, teniendo en cuenta que X_L (total del L° estrato) se observa sin error de muestreo:

$$V(\hat{\bar{X}}) = \left(\frac{N'}{N}\right)^2 V(\hat{\bar{X}}') \quad ; \qquad V(\hat{X}) = N'^2\,V(\hat{\bar{X}}')$$

por tanto el problema se reduce a la aplicación de la teoría del muestreo estratificado al conjunto de los (L−1) primeros estratos. El cálculo de estimadores, varianzas y sus estimaciones, así como de las afijaciones queda resuelto fácilmente.

Más complicado es decidir el límite X^* a partir del cual las unidades con $X_i > X^*$ deben incluirse con certeza en la muestra. Dalenius (1957) ha estudiado este problema aproximando la distribución de frecuencias de X_i en la población, por una función de densidad; al mimso resultado llega Glasser (1962) que aplicó los métodos algebraicos usuales en poblaciones finitas. Consideran dos estratos, el primero formado por las unidades que se muestrean y el segundo por las que se incluyen con certeza en la muestra, y en esta situación el punto óptimo (varianza mínima para n dado) es:

$$X^* = X_1 + \sigma_1 \sqrt{\frac{N - n_2}{n - n_2}} = \bar{X}_1 + \sigma_1 \sqrt{\frac{N_1}{n_1}}$$

siendo $N_2 = n_2$, el número de unidades que se incluyen con certeza. La determinación de X^* a partir de esta fórmula es tediosa, debido a que al variar n_2 también lo hace $\bar{X}_1$, y σ_1; por ello se han buscado las siguientes cotas inferior y superior para X^*:

$$\bar{X} + \sigma\sqrt{\frac{N}{n} - 1} < X^* < \bar{X} + \sigma\sqrt{\frac{N}{n}}$$

que requieren el conocimiento de $\bar{X}$ y σ (poblacionales). En la práctica no conoceremos $\bar{X}$ y σ, pero podremos tener alguna aproximación y, en todo caso, estas fórmulas serán una guía para encontrar un valor de X^* que si no es el teóricamente óptimo, sea suficientemente bueno para el problema planteado.

Vamos a considerar aquí otro planteamiento para este problema que puede ser práctico en algunos casos. Se trata de determinar para un tamaño de muestra dado, si incluyendo un determinado número, N_L, de unidades con certeza en la muestra, la precisión es mayor o menor respecto al muestreo aleatorio en todos los estratos; usando para los (L−1) primeros estratos, en el primer caso, y para todos, en el segundo, el mismo tipo de afijación.

(a) N_L unidades fijas y muestreo en los (L−1) primeros estratos.

(b) Muestreo en todos los estratos.

Consideremos muestreo con reposición afijación óptima.

Siendo:

$N' = N - N_L$, se tiene:

$W'_h = N_h / N'$ y denotando por α la proporción de unidades en los (L−1) primeros estratos, resulta:

$$W'_h = \alpha \; W_h \quad ; \quad \alpha = N'/N$$

La varianza del estimador de $\bar{X}$, con afijación óptima para los (L−1) primeros estratos y $n_L = N_L$ unidades del L° estrato entrando con certeza en la muestra, caso (a), es:

$V(\hat{\bar{X}})_a = \frac{1}{n'} \left(\sum_{h=1}^{L-1} W_h \sigma_h \right)^2$; donde n' es el tamaño de la muestra, $n' = n - N_L$, para los (L−1) primeros estratos.

$$V(\hat{\bar{X}})_a = \frac{A}{n'} = \frac{A}{n - N_L}$$

La varianza con afijación óptima, repartiendo la muestra del tamaño n en los L estratos, es:

$$V(\hat{\bar{X}})_b = \frac{1}{n} \left(\sum_{h=1}^{L} W_h \sigma_h \right)^2 = \frac{B}{n} \text{ , por tanto}$$

$V(\hat{\bar{X}})_a \quad \frac{A}{n - N_L}$ es menor que $V(\hat{\bar{X}})_b$ si:

$\frac{A}{n - N_L} < \frac{B}{n}$, de donde resulta: $N_L < \left(1 - \frac{A}{B}\right) n$.

Por el contrario si $N_L > (1 - A/B)\, n$, tiene mayor varianza el estimador con las N_L unidades del último estrato entrando con certeza en la muestra. La ganancia o pérdida de precisión con el diseño (a) respecto

al (b) puede calcularse de la relación entre las varianzas:

$$\frac{V(\hat{\bar{X}})_a}{V(\hat{\bar{X}})_b} = \frac{A}{B}\,\frac{n}{n'} \qquad \text{(i)}$$

Debe tenerse en cuenta en este análisis que si es:

$$\frac{W_L\,\sigma_L}{\sum_1^L W_h\,\sigma_h}\,n \geqslant N_L\,,$$

la afijación óptima (b) ya exige un tamaño de muestra en el L° estrato igual o superior a su tamaño. En resumen, si para una estratificación dada y tamaño de muestra n, es:

$$N_L > \frac{W_L\,\sigma_L}{\sum_1^a W_h\,\sigma_h}\,n \qquad \text{y} \qquad N_L < \left(1 - \frac{A}{B}\right) n$$

es igual o más preciso el estimador con las N_L unidades del estrato L entrando con certeza en la muestra que tomando afijación óptima en todos los estratos. Si $N_L > (1 - A/B)\,n$, es menos preciso y en ambos casos la ganancia o pérdida de precisión viene dada por la fórmula (i).

Si el problema se plantea utilizando afijación proporcional, y efectuamos el análisis, como antes, para muestreo con reposición, se llega a un resultado formalmente análogo en el que siendo:

$$A' = \sum_{h=1}^{L-1} W_h\,\sigma_h^2 \quad ; \quad B' = \sum_{h=1}^{L} W_h\,\sigma_h^2\,. \qquad \text{(ii)}$$

resulta que si N_L es menor, o mayor, que $(1 - A'/B')\,n$, el procedimiento (a) es más, o menos, preciso que el (b). La relación entre las varianzas sigue siendo la expresada por la fórmula (i), con A' y B' dados por (ii). Es obvio que en este caso nunca será en (b) $n_L > N_L$ por tratarse de afijación proporcional.

IV.8. EJEMPLOS

Ejemplo IV.1

Supongamos que una población finita de N = 500 unidades, respecto de una característica $\mathfrak{X}$ presenta la distribución de frecuencias que se adjunta. Consideremos dos procedimientos para estimación de la media:

A) Estratificación:

$$\mathfrak{X} < 10 \quad ; \quad N_1 = 380$$

$$10 \leqslant \mathfrak{X} < 100 \quad ; \quad N_2 = 90$$

$$100 \leqslant \mathfrak{X} \quad ; \quad N_3 = 30$$

B) La misma estratificación, tomando en el tercer estrato ($100 \leqslant \mathfrak{X}$) todas las unidades con certeza en la muestra.

Valores de $\mathfrak{X}$	frecuencia
2	100
3	80
5	200
10	30
20	30
50	30
100	20
200	10
N =	500

Compruébese que con el procedimiento A, para $n = 100$, la afijación óptima exige tomar en el tercer estrato un tamaño de muestra ma-

yor que el número de unidades. Determínese la afijación óptima y el error de muestreo relativo con el procedimiento B, sin reposición;

Solución:

Cuasivarianzas de los estratos y tamaños realtivos:

$S_1^2 = 1{,}78498$; $S_2^2 = 292{,}1349$; $S_3^2 = 2298{,}8511$

$W_1 = 0{,}76$; $W_2 = 0{,}18$ $W_3 = 0{,}06$

Procedimiento A:

$W_1 S_1 = 1{,}0053$; $W_2 S_2 = 3{,}0766$; $W_3 S_3 = 2{,}8768$

$\sum W_h S_h = 6{,}9587$; de donde se obtiene

$n_1 = 15$; $n_2 = 44$; $n_3 = 41 > 30$

Procedimiento B:

Por ser $n_3 = 30$; la afijación óptima de las $100 - 30 = 70$ unidades restantes a los dos primeros estratos, teniendo en cuenta que ahora es $W_1' \doteq 0{,}8085$; $W_2' \doteq 0{,}1915$, resulta:

$$n_1 = 17 ; \qquad n_2 = 53 ; \qquad n_3 = 30$$

La media de la población puede escribirse en la forma:

$$\bar{X} = \frac{N' \bar{X}' + X_3}{N}$$

siendo $N' \doteq N - N_3$ y $\bar{X}'$ la media de los dos primeros estratos. Por tanto el estimador insesgado:

$$\hat{\bar{X}} = \frac{N' \sum_{h=1,2} W' \bar{x}_h + X_3}{N} = \frac{N'}{N} \hat{\bar{X}}' + \frac{X_3}{N}$$

y su varianza:

$$V(\hat{\bar{X}}) = \left(\frac{N'}{N}\right)^2 V(\hat{\bar{X}}') = \sum_{h=1,2} W_h^2 V(\bar{x}_h)$$

El cálculo para los datos del ejemplo, da un error de muestreo de 2,3 % .

Ejemplo IV.2

Se conocen los siguientes datos de una característica $\mathfrak{X}$, en una población clasificada en tres estratos:

$$\left.\begin{array}{l} \sigma_1 = 4 \\ W_1 = 0{,}6 \end{array}\right\} \quad \left.\begin{array}{l} \sigma_2 = 12 \\ W_2 = 0{,}3 \end{array}\right\} \quad \left.\begin{array}{l} \sigma_3 = 80 \\ W_3 = 0{,}1 \end{array}\right\} \quad N = 1000$$

En muestreo sin reposición, determínese el tamaño de muestra que con afijación proporcional da una varianza del estimador de la media, igual a 5. (Suponemos $\sigma_h \simeq S_h$).

Solución:

$n = 122$; $n_1 = 73$; $n_2 = 37$; $n_3 = 12$

Ejemplo IV.3

Resuélvase el ejemplo IV.2 para afijación óptima. (Suponemos $\sigma_h \simeq S_h$).

Solución:

$n = 35$: $n_1 = 6$; $n_2 = 9$; $n_3 = 20$

Ejemplo IV.4

Supongamos conocidos los siguientes datos de una población clasificada en tres estratos:

$$\left.\begin{array}{l} S_1^2 = 9 \\ N_1 = 1000 \end{array}\right\} \qquad \left.\begin{array}{l} S_2^2 = 225 \\ N_2 = 600 \end{array}\right\} \qquad \left.\begin{array}{l} S_3^2 = 1600 \\ N_3 = 200 \end{array}\right\}$$

Los costes unitarios son $c_1 = 1000$; $c_2 = 1200$; $c_3 = 2000$. Determínese el coste de una muestra estratificada, con afijación proporcional y con afijación óptima, que proporcione un error de muestreo del 5 % para estimar la media. (Se supone muestreo sin reposición y $\overline{X} \simeq 22$).

Solución:

Afijación proporcional

$$n = 191\,; \quad n_1 = 106\,; \quad n_2 = 64\,; \quad n_3 = 21\,; \quad C = 224.800$$

Afijación óptima:

$$n = 92\,; \quad n_1 = 14\,; \quad n_2 = 41\,; \quad n_3 = 37\,; \quad C = 137.200$$

Ejemplo IV.5

Una muestra aleatoria sin reposición, de tamaño $n = 35$, procedente de una población clasificada en tres estratos, uno de los cuales está formado por $N_3 = 4$ unidades que entran con certeza en la muestra, presenta los siguientes datos:

1er estrato:

$$\Sigma\, X_{1i} = 119\,; \qquad \Sigma\, X_{1i}^2 = 837\,; \qquad n_1 = 18$$

2º estrato:

$$\Sigma\, X_{2i} = 278\,; \qquad \Sigma\, X_{2i}^2 = 7.978\,; \qquad n_2 = 13$$

3er estrato:

$$\Sigma\, X_{3i} = 336\,; \qquad \Sigma\, X_{3i}^2 = 28.866\,; \qquad n_3 = N_3 = 4$$

Siendo $N_1 = 400$ y $N_2 = 200$, estímese la media de la población y su error de muestreo.

Solución:

$$\hat{\bar{X}} = 12{,}02 : \qquad \sigma(\hat{\bar{X}}) = 1{,}1853 ; \qquad \simeq 9{,}7\,\%$$

Ejemplo IV.6

Partiendo de las estimaciones de S_1^2 y S_2^2 proporcionadas por la muestra del ejemplo IV.5, compruebe que $n_1 = 6$; $n_2 = 25$ es aproximadamente afijación óptima y calcule en este caso el error de muestreo de $\hat{\bar{X}}$.

Solución:

Para $\hat{S}_1^2 = 2{,}9575$ y $\hat{S}_2^2 = 169{,}4230$, la afijación óptima es $n_1 = 6{,}48$; $n_2 = 24{,}52$. Si se toma $n_1 = 6$ y $n_2 = 25$, resulta:

$$\hat{\sigma}(\hat{\bar{X}}) = 0{,}9290; \simeq 7{,}7\,\%$$

Ejemplo IV.7

Una muestra aleatoria sin reposición, procedente de una población clasificada en tres estratos, presenta los siguientes datos:

1er estrato:

$$\Sigma\, X_{1i} = 260 ; \quad \Sigma\, X_{2i}^2 = 2.825 ; \quad n_1 = 25 ; \quad N_1 = 500$$

2° estrato:

$$\Sigma\, X_{2i} = 940 ; \quad \Sigma\, X_{2i}^2 = 23.240 ; \quad n_2 = 40 ; \quad N_2 = 400$$

3er estrato:

$$\Sigma\, X_{2i} = 650 ; \quad \Sigma\, X_{3i}^2 = 45{,}265 ; \quad n_3 = 10 ; \quad N_3 = 100$$

Estímese la eficiencia relativa de este procedimiento de muestreo, respecto a la selección aleatoria sin estratificación, con el mismo tamaño de muestra. Estímese también cual sería la afijación óptima.

Solución:

$s_1^2 = 5{,}0417$; $s_2^2 = 29{,}4872$; $s_3^2 = 335$

$w_1 = 0{,}5$; $w_2 = 0{,}4$; $w_3 = 0{,}1$

$\bar{x}_1 = 10{,}4$; $\bar{x}_2 = 23{,}5$; $\bar{x}_3 = 65$

La fórmula (ii) de IV.5, nos da la estimación, $\hat{\sigma}^2 = 296{,}8438$; $\hat{S}^2 = 297{,}1405$, por tanto el error de muestreo para la media, sin selección estratificada, se estima en:

$$\left(\frac{(1000 - 75)}{1000} \cdot \frac{297{,}1405}{75} \right)^{1/2} \simeq 1{,}9143$$

La selección estratificada nos da para la media, el error de muestreo estimado:

$$\left(\Sigma \quad (1 - f_h)\, w_h^2 \quad \frac{\hat{S}_h^2}{n_h} \right)^{1/2} \simeq 0{,}6759$$

que es el 35,26 % del que se obtendría en muestreo no estratificado.

La afijación óptima se estima en: $n_1 = 16$; $n_2 = 32$; $n_3 = 27$. El error de muestreo en este caso sería 0,55 que equivale al 28,73 % del que se tendría en muestreo no estratificado y al 81,49 % del que se tendría con la afijación planteada en el enunciado del ejemplo.

Ejemplo IV.8

Para la distribución $\bar{\mathfrak{X}}$ del ejemplo IV.1 y tamaño de muestra, sin reposición, $n = 100$, determínese por sus cotas el valor X^* óptimo, a partir del cual las unidades entran con certeza en la muestra.

Solución:

$$83{,}4 < X^{*} < 91{,}4$$

Ejemplo IV.9

Encontrar la condición necesaria y suficiente para que en muestreo estratificado sin reposición, la afijación igual sea mejor que la proporcional. Construya un ejemplo.

Solución:

$$\sum_{h=1}^{L} (1 - L\,W_h)\,W_h\,S_h^2 > 0\,; \qquad \text{ejemplo: } L = 2$$

$$\begin{cases} W_1 = 0{,}8 \\ S_1^2 = 20 \end{cases} \qquad \begin{cases} W_2 = 0{,}2 \\ S_2^2 = 100 \end{cases}$$

En muestreo sin reposición la condición es análoga, sustituyendo S_h^2 por σ_h^2.

Ejemplo IV.10

Determine el tamaño, n, de la muestra que con afijación proporcional a los estratos, proporcione la misma precisión que una muestra no estratificada de tamaño n_o, para estimar la proporción P de unidades de una cierta clase en la población. Suponga en ambos casos muestreo con reposición y aplique el resultado a los datos $W_1 = 0{,}5$, $P_1 = 0{,}5$; $W_2 = 0{,}3$, $P_2 = 0{,}6$; $W_3 = 0{,}2$, $P_3 = 0{,}4$ con $n_o = 600$.

Solución:

$$n = \frac{(\Sigma\ W_h\ P_h\ Q_h)}{(\Sigma\ W_h\ P_h)\ (1 - \Sigma\ W_h\ P_h)}\ n_o\ ; \quad n = 588$$

Ejemplo IV.11

Consideremos una población estratificada de dos formas:

I

Estrato	Valores de $\mathfrak{X}$	Frecuencia
	2	20
	3	30
$h = 1$	4	60
	7	100
	10	150
	12	200
$h = 2$	16	120
	20	80
	25	50
	30	20
	35	18
$h = 3$	50	10
	60	8
	100	4

II

Estrato	Valores de $\mathfrak{X}$	Frecuencia
	2	20
	3	30
$h = 1$	4	60
	7	100
	10	150
	12	200
	16	120
	20	80
$h = 2$	25	50
	30	20
	35	18
	50	10
$h = 3$	60	8
	100	4

Compruebe, suponiendo muestreo con reposición, que para un tamaño de muestra $n = 100$, es más preciso el procedimiento de estimación que incluye con certeza todas las unidades del tercer estrato, que el muestreo con afijación óptima en los tres estratos, en la forma II y no en la I.

Solución:

Para I, se tiene:

$W_1 = 0{,}3053$ $\quad W_2 = {,}6316$ $\quad W_3 = 0{,}0631$

$\sigma_1 = 1{,}6174$ $\quad \sigma_2 = 4{,}5623$ $\quad \sigma_3 = 18{,}4006$

$$A = \left(\sum_{1}^{2} W_h \sigma_h \right)^2 = 11{,}3929$$

$$B = \left(\sum_{1}^{3} W_h \sigma_h \right)^2 = 20{,}5791$$

Resulta $N_L = 60 > (1 - A/B) \cdot 100 = 44{,}6$, por tanto es mejor el muestreo en los tres estratos.

Para II, se tiene:

$W_1 = 0{,}4632$ $\quad W_2 = 0{,}5137$ $\quad W_3 = 0{,}0231$

$\sigma_1 = 2{,}6810$ $\quad \sigma_2 = 2{,}6120$ $\quad \sigma_3 = 18{,}1363$

$A = 6{,}6751$; $B = 9{,}0154$ de donde resulta:

$N_L = 22 < (1 - A/B)\ 100 = 26$, por tanto con la estratificación II, es mejor tomar con certeza las 22 unidades del tercer estrato.

Ejemplo IV.12

Determine con referencia al ejemplo IV.11, cúal de los procedimientos para estimación de la media, con reposición, es más preciso:

1. Estratificación I con afijación óptima a los tres estratos.
2. Estratificación II, incluyendo con certeza el tercer estrato y afijación óptima a los dos primeros.

Solución:

$V(\hat{\bar{X}})_1 = 0{,}2058$; $V(\hat{\bar{X}})_2 = 0{,}0856$, por tanto el error de muestreo es, en el segundo caso, un 65 % del primero, por tanto (2) es más preciso que (1).

Ejemplo IV.13

Supuesto que se conocen los siguientes datos de una característica ($\mathfrak{X}$, $\mathfrak{Y}$) en una población estratificada:

Estrato	W_h	σ_x^2	σ_y^2	ρ_{xy}	$\bar{X}_h$	$\bar{Y}_h$
1	0,65	20	250	0,72	10	50
2	0,35	85	100	0,65	40	60

se desea calcular la eficiencia relativa de los estimadores $\hat{X}$, $\hat{X}_{RS}$ y $\hat{X}_{RC}$. Supongamos muestreo con reposición y afijación proporcional al tamaño del estrato.

Solución:

$R_1 = 1/5$; $R_2 = 2/3$; $R = 0{,}3822$; $n_h = n W_h$

$$V(\hat{X}) = \frac{N^2}{n} \sum_1^2 W_h \sigma_{x,h}^2 = 42{,}75 \frac{N^2}{n}$$

$$V(\hat{X}_{RS}) = \frac{N^2}{n} \sum_1^2 W_h(\sigma_{xh}^2 + R^2 \sigma_{yh}^2 - 2\rho_h R_h \sigma_{xh} \sigma_{yh}) = 23{,}60 \frac{N^2}{n}$$

$$V(\hat{X}_{RC}) = \frac{N^2}{n} \sum_1^2 W_h(\sigma_{xh}^2 + R^2 \sigma_{yh}^2 - 2\rho_h R \sigma_{xh} \sigma_{yh}) = 30{,}31 \frac{N^2}{n}$$

Así resulta:

$V(\hat{X}_{RC}) = 70{,}9\%$ de $V(\hat{X})$

$V(\hat{X}_{RS}) = 55{,}2\%$ de $V(\hat{X})$

V. MUESTREO BIETAPICO

Vamos a estudiar en este capítulo el muestreo de poblaciones cuyas unidades, de una forma natural o artificial, se encuentran distribuidas en unidades mayores que venimos denominando conglomerados. El muestreo bietápico consiste en efectuar primero una selección aleatoria de n conglomerados de los N que suponemos constituyen el colectivo (esta es la primera etapa de muestreo) y después una selección de unidades, denominadas de segunda etapa, en cada una de las unidades primarias ya seleccionadas. El procedimiento de muestreo en primera etapa puede ser estratificado, o no, con o sin reposición y con probabilidades iguales o desiguales, pero en segunda etapa supondremos que se realiza sin reposición y con probabilidades iguales; además y esta es una condición fundamental, se efectúa de forma independiente, probabilísticamente hablando, en cada una de las unidades primarias de la muestra; esto obliga, p.e., a que si el muestreo en primera etapa es con reposición y aparece alguna unidad primaria más de una vez en la muestra, deberán hacerse tantas selecciones independientes en dicha unidad para la segunda etapa de muestreo; en la práctica suele multiplicarse por el número de veces que aparece en la muestra de unidad primaria, el resultado de un submuestreo de la misma, lo cual es una aproximación al modelo que describimos.

La utilización de este tipo de muestreo responde a un objetivo económico, bien sea para reducir costes y tiempo de preparación del diseño o por falta de la información básica necesaria (p.e. un directorio

de unidades de segunda etapa, convenientemente actualizado) para practicar un procedimiento de muestreo por unidades. Ejemplos típicos son:

Primer ejemplo
- Unidad primaria: Municipio
- Unidad de segunda etapa: Explotación agraria, Establecimiento comercial, etc.

Segundo ejemplo
- Unidad primaria: Sección censal
- Unidad de segunda etapa: Edificio, Vivienda, etc.

De una forma un tanto esquemática podemos explicar la reducción de coste de la siguiente forma: Si para una encuesta dirigida a los hogares se muestrean viviendas, la dispersión en el territorio obligará a efectuar una sola visita en muchos municipios distantes del centro de trabajo, con los consiguientes gastos de desplazamiento y tiempo para la recogida de los datos; en este caso el muestreo de municipios, o secciones censales, tomados como unidades primarias, concentra las visitas a las viviendas seleccionadas en segunda etapa.

La no existencia de un directorio actualizado puede presentarse p.e. cuando se desea una muestra de pequeños establecimientos de comercio al por menor; el uso de muestreo en dos etapas exige solamente actualizar el directorio (marco o lista de muestreo) en las unidades primarias seleccionadas para la muestra.

Respecto del muestreo unietápico presenta la ventaja de que al observar solamente una submuestra en cada unidad primaria, el muestreo bietápico, a igual tamaño de unidades simples permite extender mejor la muestra en la población; esto es, tomar un mayor número de unidades primarias a cambio de observar solamente una fracción de unidades secundarias. El hecho de que los conglomerados estén formados por unidades homogéneas (similares entre sí) contribuye a reforzar esta ventaja ya que una muestra de cada conglomerado será suficiente para una precisión aceptable. Por el contrario, cuantas más unidades primarias se tomen y menor fracción de submuestreo, tendremos una mayor dispersión territorial de la muestra que hará incrementar los costes. En la práctica el diseño de un procedimiento bietápico nos

conduce a tener que buscar una solución óptima de compromiso entre los requerimientos de precisión, para los distintos niveles o ámbitos territoriales, y los recursos disponibles.

V.1. CALCULO DE ESPERANZAS Y VARIANZAS EN MUESTREO BIETAPICO

Designemos por $W = \{U_1, U_2, \ldots, U_i, \ldots, U_n\}$ una muestra aleatoria de n unidades primarias y por Wr una muestra de unidades de segunda etapa constituída por n grupos de m_i unidades, cada uno seleccionado aleatoria e independientemente en cada U_i presente en W. En muestreo sin reposición tendríamos $\binom{N}{n}$ muestras W, no ordenadas, y para cada una de éstas son posibles

$$\binom{M_1}{m_1}\cdot\binom{M_2}{m_2}\cdot\ldots\cdot\binom{M_i}{m_i}\cdot\ldots\cdot\binom{M_n}{m_n} \qquad (1)$$

muestras Wr (siendo m_i el número de unidades secundarias seleccionadas en U_i cuando ésta aparece en W). El número total de muestras posibles Wr es igual a la suma de todos los productos del tipo (1) que se pueden formar con las $\binom{N}{n}$ combinaciones de n unidades primarias. En muestreo con reposición la situación es análoga, si bien tendríamos un número mayor de muestras de primera etapa, combinaciones con repetición, y por tanto de muestras Wr.

Para cualquier procedimiento de muestreo aleatorio están definidas las probabilidades:

$Pr(W)$ = Probabilidad de la muestra W de n unidades primarias = γ_W

$Pr(Wr / W)$ = Probabilidad de la muestra Wr cuando es W la muestra de primera etapa = $\gamma_{Wr/W}$

$Pr(Wr)$ = Probabilidad incondicional de la muestra Wr = γ_{Wr}

con las siguientes propiedades:

1) $\sum_{W} Pr(W) = \sum_{W} \gamma_W = 1$

2) $\sum_{r} Pr(Wr/W) = \sum_{r} \gamma_{Wr/W} = 1$, para cada W.

3) $Pr(Wr) = \gamma_{Wr} = \gamma_W \cdot \gamma_{Wr/W}$

4) En consecuencia se comprueba que $\sum_{W,r} Pr(Wr) = 1$

$$\sum_{W} \sum_{r} \gamma_W \, \gamma_{Wr/W} = \sum_{W} \gamma_W \sum_{r} \gamma_{Wr/W} = \sum_{W} \gamma_W \cdot 1 = 1$$

ESPERANZA Y VARIANZA CONDICIONADAS

Un estadístico se ha definido como una función de las observaciones muestrales; esto es, una aplicación que asocia a cada posible muestra, Wr, un valor numérico real:

$$T(Wr) : \{Wr\} \xrightarrow{T} \mathcal{R}$$

Emplearemos la notación $\underset{2}{E}(T) = T(W)$ para indicar la esperanza matemática de T(Wr) condicionada a una muestra W de primera etapa. Tenemos por tanto:

$$\bar{T}(W) = \underset{2}{E}(T/W) = \sum_{r} T(Wr) \cdot \gamma_{Wr/W}$$

donde la suma en r se extiende a todas las muestras posibles de segunda etapa, Wr, para una muestra dada, W, de primera etapa.

Análogamente la notación $\underset{2}{V}(T) = \bar{V}(W)$ expresa la varianza de T(Wr) condicionada a una muestra W de primera etapa. Tenemos ahora:

$$\bar{V}(W) = \underset{2}{V}(T) = E\left\{ (T - \underset{2}{E}(T))^2 / W \right\} = \sum_{r} (T(Wr) - \bar{T}(W))^2 \cdot \gamma_{Wr/W}$$

En estas condiciones vamos a probar dos resultados importantes para el cálculo de esperanzas y varianzas de los estimadores en muestreo bietápico:

I) $$E(T) = E_2(E(T))$$

II) $$V(T) = V_2(E(T)) + E_2(V(T))$$

Esperanza de T

$$E(T) = \sum_W \sum_r T(Wr) \cdot \gamma_{Wr} = \sum_W \sum_r T(Wr) \cdot \gamma_W \cdot \gamma_{Wr/W} =$$

$$= \sum_W \gamma_W \sum_r T(Wr) \cdot \gamma_{Wr/W} = \sum_W E_2(T) \cdot \gamma_W = E_2(E(T))$$

Este resultado nos dice que la esperanza matemática de un estadístico definido para cada muestra bietápica puede calcularse en dos pasos, primero la esperanza condicionada a una muestra de unidades primarias y después la esperanza de aquélla en el espacio de muestras de primera etapa.

Varianza de T

$$V(T) = \sum_W \sum_r \left\{ T(Wr) - E(T) \right\}^2 \cdot \gamma_{Wr} =$$

$$= \sum_W \sum_r \left\{ (T(Wr) - \bar{T}(W)) + (\bar{T}(W) - E(T)) \right\}^2 \cdot \gamma_{Wr}$$

$$= \sum_W \sum_r \left\{ (T(Wr) - \bar{T}(W))^2 + (\bar{T}(W) - E(T))^2 \right\} \cdot \gamma_{Wr}$$

puesto que la suma de productos cruzados se anula, teniendo en cuenta que se verifica:

$$\gamma_{Wr} = \gamma_W \cdot \gamma_{Wr/W} \qquad \text{y que} \qquad E(T(Wr)) = T(W)$$

seguimos con el cálculo de la varianza de T:

$$V(T) = \sum_W \gamma_W \left\{ \sum_r (T(Wr) - T(W))^2 \cdot \gamma_{Wr/W} \right\} +$$

$$+ \sum_W \gamma_W (T(W) - W(T))^2 \sum_r \gamma_{Wr/W}$$

y por ser igual a 1 la suma de las probabilidades condicionadas $\gamma_{Wr/W}$

$$V(T) = \underset{2}{E(V(T))} + \underset{2}{V(E(T))}$$

Este resultado nos dice que la varianza de un estadístico definido para cada muestra bietápica consta de dos sumandos: uno es la varianza de la esperanza condicionada a cada muestra de primera etapa y el otro es la esperanza de la varianza condicionada.

V.2. FORMACION DE ESTIMADORES LINEALES INSESGADOS. MUESTREO CON REPOSICION

Suponiendo que en primera etapa el muestreo es con reposición y P_i la probabilidad asignada a la unidad primaria U_i en una extracción, si $\hat{X}_i$ es un estimador insesgado del total X_i cuando aparece U_i en la muestra, entonces:

$$\hat{X} = \sum_{i=1}^{n} \frac{\hat{X}_i}{n P_i}$$

es un estimador insesgado del total X. En efecto:

$$E(\hat{X}) = E\left\{ \underset{2}{E}\left(\sum_{1}^{n} \frac{\hat{X}_i}{n P_i} \right) \right\} = E\left(\sum_{1}^{n} \frac{X_i}{n P_i} \right) =$$

$$= \sum_{1}^{N} \frac{X_i}{n P_i} \, n P_i = X$$

El submuestreo dentro de cada unidad primaria de la muestra suponemos siempre que se realiza sin reposición y probabilidades iguales, por tanto el estimador insesgado de X_i es:

$$\hat{X}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} X_{ij} = \frac{M_i}{m_i} x_i = M_i \bar{x}_i$$

donde m_i representa el número de unidades de segunda etapa seleccionadas en la unidad primaria U_i cuando ésta aparece en la muestra; x_i es el total muestral observado en las m_i unidades y $\bar{x}_i$ la media. El estimador $\hat{\bar{X}}$ puede entonces escribirse de la siguiente forma:

$$\hat{X} = \sum_{1}^{n} \frac{M_i}{n P_i} \cdot \frac{x_i}{m_i}$$

Si las probabilidades de selección de las unidades primarias son proporcionales al tamaño, medido por el número de unidades secundarias, tenemos $P_i = M_i/M$ para todo i ; por tanto el estimador del total admite la forma:

$$\hat{X} = \frac{M}{n} \sum_{1}^{n} \frac{x_i}{m_i} = \frac{M}{n} \sum_{1}^{n} \bar{x}_i$$

En el caso de muestreo de unidades primarias con probabilidades iguales, tenemos $P_i = 1/N$ para todo i ; por tanto el estimador del

total ahora es:

$$\hat{X} = \frac{N}{n} \sum_{1}^{n} \frac{M_i}{m_i} x_i = \frac{N}{n} \sum_{1}^{n} M_i \bar{x}_i$$

Estimadores de la media poblacional por unidad primaria, y por unidad secundaria, pueden formarse a partir de los estimadores del total de la siguiente forma:

$$\hat{\bar{X}} = \frac{1}{N} \hat{X} \quad ; \quad \hat{\bar{\bar{X}}} = \frac{1}{M} \hat{X}$$

V.3. FORMACION DE ESTIMADORES LINEALES INSESGADOS. MUESTREO SIN REPOSICION

Si el muestreo de unidades primarias se realiza sin reposición, siendo π_i la probabilidad de que la unidad U_i aparezca en la muestra, tendremos como estimador insesgado del total X.

$$\hat{X} = \sum_{1}^{n} \frac{\hat{X}_i}{\pi_i}$$

donde $\hat{X}_i$ tiene el mismo significado que antes, es decir un estimador insesgado del total de la unidad primaria i-ésima cuando ésta aparece en la muestra, basado en las m_i observaciones de submuestreo.

Si el muestreo es con probabilidades iguales, tenemos $\pi_i = n/N$ para todo i, por tanto el estimador del total ahora es:

$$\hat{X} = \frac{N}{n} \sum_{1}^{n} \hat{X}_i$$

Estimadores insesgados de la media por unidad primaria y de la media por unidad secundaria, obviamente se obtiene dividiendo $\hat{X}$ por N ó por M.

V.4. MUESTRAS AUTOPONDERADAS

Una muestra bietápica se dice que es *autoponderada,* o más correctamente que el procedimiento de muestreo origina muestras autoponderadas cuando se cumple:

$$E(\varepsilon_{ij}) = \text{constante para todo } i, j$$

siendo ε_{ij} la variable aleatoria: número de veces que la unidad secundaria u_{ij} aparece en la muestra.

Si el muestreo es sin reposición la condición es equivalente a exigir que todas las unidades (de segunda etapa) tengan la misma probabilidad de aparecer en la muestra.

La importancia de las muestras autoponderadas reside en que los estimadores insesgados del total poblacional admiten la forma

$$\hat{X} = F \, . \, x$$

donde F es un factor constante, denominado factor de elevación o de extrapolación, que multiplica al total muestral observado. Desde otro punto de vista esto significa que todas las observaciones X_{ij} efectuadas en las unidades secundarias de la muestra contribuyen por igual a la formación del estimador, ya que se multiplican por un mismo factor F. Otra ventaja reside en la sencillez del cálculo práctico puesto que si la muestra no es autoponderada el estimador lineal insesgado del total está constituído por una suma de observaciones afectadas de diferentes multiplicadores.

Vamos entonces a efectuar el cálculo de la esperanza matemática de ε_{ij} y a estudiar en qué condiciones el procedimiento origina muestras autoponderadas. Suponiendo como siempre, submuestreo con probabilidades iguales y sin reposición, realizado independientemente en cada unidad primaria de la muestra, para muestreo con reposición y probabilidades cualesquiera en primera etapa tenemos que en una extracción la probabilidad de que una determinada unidad u_{ij} aparezca es $\Pr(\varepsilon_{ij} = 1) = P_i \, . \, m_i / M_i$ por tanto en n extracciones la variable

ε_{ij} tiene distribución binomial de parámetros (n ; $P_i \cdot m_i / M_i$) ; en consecuencia:

$$E(\varepsilon_{ij}) = \frac{n\ P_i\ m_i}{M_i}$$

De este resultado se deduce que si en primera etapa se asignan probabilidades P_i proporcionales a los tamaños M_i, la muestra es autoponderada si el tamaño de muestra de segunda etapa es igual para todas las unidades primarias seleccionadas; en efecto, en estas condiciones se tiene:

$$E(\varepsilon_{ij}) = \frac{n\ M_i}{M} \cdot \frac{\bar{m}}{M_i} = \frac{n\ \bar{m}}{M} = \text{constante para todo } i\ ,\ j$$

siendo $\bar{m}$ el número fijo de unidades secundarias seleccionadas en cada unidad primaria de la muestra.

También se deduce que si en primera etapa se asignan probabilidades de selección iguales, $P_i = 1/N$, la muestra es autoponderada si en segunda etapa se toma una fracción de submuestreo igual para todas las unidades; esto es, si $f_2 = m_i/M_i$ es constante para todo i . En efecto, se tiene:

$$E(\varepsilon_{ij}) = \frac{n}{N} \cdot \frac{m_i}{M_i} = \frac{n}{N} \cdot f_2 = \text{Constante.}$$

Si aplicamos estos resultados a los estimadores $\hat{X}$ del total poblacional, se obtiene para muestreo con probabilidades proporcionales al tamaño:

$$\hat{X} = \sum_{i=1}^{n} \sum_{j=1}^{\bar{m}} \frac{M}{n\ \bar{m}} X_{ij} = \frac{M}{m}\ x = F\ .\ x$$

donde m es el número total de unidades secundarias en la muestra, x es el total muestral y F es el factor de expansión que coincide con el inverso de la fracción global de muestreo. Para muestreo con probabili-

dades iguales en primera etapa, tenemos:

$$\hat{X} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{N}{n} \cdot \frac{1}{f_2} X_{ij} = \frac{1}{f_1 \cdot f_2} x$$

donde $f_i = n/N$ es la fracción de muestreo en primera etapa. El factor de elevación es ahora el inverso del producto de las fracciones de muestreo en primera y segunda etapa.

Hasta aquí hemos supuesto muestreo con reposición en primera etapa, si suponemos muestreo sin reposición los resultados son análogos, teniendo en este caso $\pi_i = n\ M_i/M$ para probabilidades proporcionales al tamaño y $\pi_i = n/N$ para probabilidades iguales.

La relación que existe entre muestras autoponderadas y estimadores lineales insesgados de expansión simple, puede verse también de la siguiente forma. Siendo K_{ij} los coeficientes que se aplican a las observaciones muestrales X_{ij}, la forma general de un estimador lineal del total es:

$$\hat{X} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} K_{ij}\ X_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} K_{ij}\ X_{ij}\ \varepsilon_{ij}$$

en consecuencia la condición para que sea insesgado es:

$$E(\hat{X}) = \sum_{1}^{N} \sum_{1}^{M_i} K_{ij}\ X_{ij}\ E(\varepsilon_{ij}) = \sum_{1}^{N} \sum_{1}^{M_i} X_{ij}$$

de donde se deduce que debe ser:

$$K_{ij} = \frac{1}{E(\varepsilon_{ij})} \quad \text{, para todo } i, j$$

si además deseamos que $\hat{X}$ tome la forma:

$$\hat{X} = F \sum_{i=1}^{n} \sum_{j=1}^{m_i} X_{ij} = F \cdot x \qquad \text{(Expansión simple)}$$

debe ser k_{ij} constante para todo i, j, ó lo que es lo mismo, E (ε_{ij}) constante para todo i , j. Resulta obvio que el factor de elevación F es justamente el inverso de E (ε_{ij}).

V.5. VARIANZA DE LOS ESTIMADORES LINEALES EN MUESTREO BIETAPICO

Aplicaremos los resultados obtenidos en V.1. para efectuar el cálculo de la varianza de los estimadores estudiados. También usaremos como resultado conocido la varianza de estimadores en muestreo unietápico. Como siempre, suponemos que el submuestreo se realiza sin reposición, con probabilidades iguales, e independientemente en las unidades primarias de la muestra.

V.5.1. MUESTREO CON REPOSICION

Sabemos que $\hat{X} = \sum_{1}^{n} \frac{\hat{X}_i}{n P_i}$ es el estimador lineal insesgado del total. Para obtener una expresión explícita de su varianza aplicamos la fórmula general:

$$V(\hat{X}) = \underset{2}{V}(\underset{2}{E}(\hat{X})) + \underset{2}{E}(\underset{2}{V}(\hat{X}))$$

teniendo en cuenta que

$$\underset{2}{E}(\hat{X}) = \underset{2}{E}\left(\sum_{1}^{n} \frac{X_i}{n P_i}\right) = \sum_{1}^{n} \frac{X_i}{n P_i}$$

ya que se supone que $\hat{X}_i$ es insesgado de X_i para todo i. El primer sumando de la varianza es:

$$\underset{2}{V}(\underset{2}{E}(\hat{X})) = V\left(\sum_{1}^{n} \frac{X_i}{n P_i}\right) = \frac{1}{n}\left\{\sum_{1}^{N} \frac{X_i^2}{P_i} - X^2\right\}$$

de acuerdo con el resultado obtenido en III.8 para muestreo unietápico.

Calculemos ahora el segundo sumando. Comenzamos por la varianza condicionada en segunda etapa, teniendo en cuenta que el submuestreo se realiza de forma independiente en cada unidad primaria de la muestra:

$$V_2(\hat{X}) = V\left(\sum_1^n \frac{\hat{X}_i}{n\,P_i}\right) = \sum_1^n \frac{V(\hat{X}_i)}{n^2\,P_i^2}$$

y siendo $f_{2i} = m_i/M_i$ la fracción de submuestreo en la unidad primaria i-ésima cuando aparece en la muestra, sabemos que:

$$V(\hat{X}_i) = \frac{(1-f_{2i})\,M_i^2\,S_i^2}{m_i}$$

por tanto el segundo sumando de la varianza de $\hat{X}$ es:

$$E(V_2(\hat{X})) = E\left(\sum_1^n \frac{(1-f_{2i})\,M_i^2\,S_i^2}{n^2\,P_i^2\,m_i}\right) = \sum_1^N \frac{(1-f_{2i})\,M_i^2\,S_i^2}{n^2\,P_i^2\,m_i} \cdot n\,P_i$$

de donde resulta:

$$V(\hat{X}) = \frac{1}{n}\left\{\sum_1^N \frac{X_i^2}{P_i} - X^2\right\} + \frac{1}{n}\sum_1^N \frac{(1-f_{2i})\,M_i^2\,S_i^2}{P_i\,m_i}$$

Una observación interesante acerca de este resultado es que en muestreo con reposición, para cualquier sistema de probabilidades P_i, la varianza de $\hat{X}$ es inversamente proporcional al tamaño de la muestra de primera etapa. La fórmula obtenida admite la forma:

$$V(\hat{X}) = V_1 + V_2$$

donde V_1 es la varianza que tendría el estimador si no hubiera submues-

treo (que coincide con la obtenida en el capítulo III, punto III.8, para muestreo unietápico) y V_2 es la contribución a la varianza del estimador del total producida por el submuestreo.

Esta fórmula general de $V(\hat{X})$ para muestreo bietápico, con reposición y probabilidades cualesquiera en primera etapa, puede escribirse de otras formas equivalentes teniendo en cuenta que sus dos sumandos V_1 y V_2 admiten las siguientes expresiones:

$$V_1 = \frac{1}{n} \sum_1^N P_i \left(\frac{X_i}{P_i} - X \right)^2 = \frac{1}{n} \left\{ \sum_1^N \frac{X_i^2}{P_i} - X^2 \right\}$$

$$V_2 = \frac{1}{n} \sum_1^N \frac{(1 - f_{2i})}{f_{2i}} \frac{M_i S_i^2}{P_i} = \frac{1}{n} \sum_1^N \left(\frac{1}{m_i} - \frac{1}{M_i} \right) \frac{M_i^2 S_i^2}{P_i}$$

CASOS PARTICULARES

A) Probabilidades proporcionales al tamaño

Si las probabilidades de selección en primera etapa son proporcionales a los tamaños, tenemos $P_i = M_i/M$ para todo i, entonces resulta:

$$V(\hat{X}) = \frac{1}{n} \left\{ \sum_1^N \frac{M X_i^2}{M_i} - X^2 \right\} + \frac{M}{n} \sum_1^N \frac{(1 - f_{2i})}{f_{2i}} \cdot S_i^2$$

que también puede escribirse:

$$V(\hat{X}) = \frac{M}{n} \left\{ \sum_1^N \frac{X^2}{M_i} - \frac{X^2}{M} \right\} + \frac{M}{n} \sum_1^N \left(\frac{1}{m_i} - \frac{1}{M_i} \right) M_i S_i^2$$

B) Probabilidades iguales

Si el muestreo se realiza en primera etapa con probabilidades iguales, tenemos $P_i = 1/N$ para todo i, resultando entonces:

$$V(\hat{X}) = \frac{1}{n}\left\{\sum_1^N N X_i^2 - X^2\right\} + \frac{1}{n}\sum_1^N \frac{1-f_{2i}}{f_{2i}} M_i N S_i^2$$

que también puede escribirse:

$$V(\hat{X}) = \frac{N}{n}\left\{\sum_1^N X_i^2 - \frac{X^2}{N}\right\} + \frac{N}{n}\sum_1^N \frac{1-f_{2i}}{f_{2i}} M_i S_i^2$$

Si la muestra es autoponderada, en estas condiciones debe ser $f_{2i} = f_2$, constante para todo i; resultando:

$$V(\hat{X}) = \frac{N}{n}\left\{\sum_1^N X_i^2 - \frac{X^2}{N}\right\} + \frac{N(1-f_2)}{n f_2}\sum_1^N M_i S_i^2$$

V.5.2. MUESTREO SIN REPOSICION

La fórmula general de $V(\hat{X})$ depende en este caso del sistema de probabilidades de selección sin reposición utilizado en la primera etapa, que usualmente denotamos por π_i. Utilizando como antes la fórmula general de la varianza de un estadístico en muestreo bietápico, obtenida en el punto V.1, escribimos:

$$V(\hat{X}) = \underset{2}{V(E(\hat{X}))} + \underset{2}{E(V(\hat{X}))}$$

donde ahora es

$$\hat{X} = \sum_1^n \frac{\hat{X}_i}{\pi_i} \quad ; \quad \text{con } \hat{X}_i = \frac{M_i}{m_i} x_i$$

el estimador lineal insesgado del total poblacional $\hat{X}$.

Puesto que

$$\underset{2}{E}(\hat{X}) = E\left(\sum_{1}^{n} \frac{X_i}{\pi_i}\right) = \sum_{1}^{n} \frac{X_i}{\pi_i}$$

el primer sumando de $V(\hat{X})$ es la varianza del estimador del total en muestreo unietápico, sin reposición y probabilidades cualesquiera, obtenida en el punto III.9 del tercer capítulo, por tanto podemos escribir:

$$\underset{2}{V}(\underset{2}{E}(\hat{X})) = \sum_{1}^{N} \left(\frac{1-\pi_i}{\pi_i}\right) X_i^2 + \sum_{i \neq j}^{N} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j}\right) X_i X_j$$

Para calcular el segundo sumando de $V(X)$ escribimos:

$$\underset{2}{V}(\hat{X}) = \underset{2}{V}\left(\sum_{1}^{n} \frac{\hat{X}_i}{\pi_i}\right) = \sum_{1}^{n} \frac{\underset{2}{V}(\hat{X}_i)}{\pi_i^2} = \sum_{1}^{n} \frac{(1-f_{2i})\, M^2\, S_i^2}{\pi_i^2\, m_i}$$

y a continuación

$$\underset{2}{E}(\underset{2}{V}(\hat{X})) = \sum_{1}^{N} \frac{(1-f_{2i})\, M_i^2\, S_i^2}{\pi_i^2\, m_i} \cdot \pi_i = \sum_{1}^{N} \frac{(1-f_{2i})\, M_i^2\, S_i^2}{\pi_i\, m_i}$$

obteniendo finalmente:

$$\underset{2}{V}(\hat{X}) = \sum_{1}^{N} \left(\frac{1-\pi_i}{\pi_i}\right) X_i^2 + \sum_{i \neq j}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} X_i X_j + \sum_{1}^{N} \frac{(1-f_{2i})\, M_i^2\, S_i^2}{\pi_i\, m_i}$$

MUESTREO CON PROBABILIDADES IGUALES

Si las probabilidades de selección asignadas a las unidades primarias son iguales, sabemos que:

$$\pi_i = \frac{n}{N} \quad \text{para todo } i \; ; \qquad \pi_{ij} = \frac{n(n-1)}{N(N-1)} \quad \text{para todo } i \neq j$$

sustituyendo estos valores en los dos primeros sumandos de la fórmula obtenemos la varianza del estimador del total en muestreo unietápico:

$$\frac{N^2 (1-f_1)}{n} S_c^2 \quad ; \text{ siendo } S_c^2 = \frac{1}{N-1} \sum_1^N (X_i - \bar{X})^2$$

ahora, en el tercer sumando tenemos:

$$\frac{N}{m} \sum_1^N \frac{(1-f_{2i}) M_i^2 S_i^2}{m_i} = \frac{N}{n} \sum_1^N \frac{(1-f_{2i})}{f_{2i}} M_i S_i^2$$

quedando finalmente la siguiente fórmula:

$$V(\hat{X}) = \frac{N^2 (1-f_1)}{n} S_c^2 + \frac{N}{n} \sum_1^N \frac{(1-f_{2i})}{f_{2i}} M_i S_i^2$$

Si todos los conglomerados son del mismo tamaño, es $M_i = \bar{M}$:

$$V(\hat{X}) = \frac{N^2 (1-f_1)}{n} S_c^2 + \frac{N\bar{M}}{n} \sum_1^N \frac{(1-f_{2i})}{f_{2i}} S_i^2$$

Si además la muestra es autoponderada, $f_{2i} = m/\bar{M}$ es constante:

$$V(\hat{X}) = \frac{N^2 (1-f_1)}{n} S_c^2 + \frac{N M (1-f_2)}{n f_2} \sum_1^N S_i^2$$

V.6. METODO DE LOS CONGLOMERADOS ULTIMOS PARA LA ESTIMACION DE VARIANZAS

Si las unidades primarias se seleccionan con reposición y probabilidades cualesquiera, un estimador insesgado de $V(\hat{X})$ basado en los datos muestrales se obtiene mediante el siguiente procedimiento. Se denomi-

na *conglomerado último* al conjunto de unidades secundarias (o de k–ésima etapa en caso de muestreo polietápico) seleccionadas en una unidad primaria de la muestra. Por tratarse de muestreo con reposición en primera etapa, el estimador del total:

$$\hat{X} = \sum_{1}^{n} \frac{\hat{X}_i}{n\,P_i}$$

donde $\hat{X}_i$ es un estimador insesgado del total X_i de la unidad primaria i–ésima, basado en el conglomerado último observado en dicha unidad, puede escribirse en la forma de una media:

$$\hat{X} = \frac{1}{n} \sum_{1}^{n} \hat{X}_{(i)} \quad ; \quad \text{con} \quad \hat{X}_{(i)} = \frac{\hat{X}_i}{P_i}$$

donde $\hat{X}_{(1)}, \ldots, \hat{X}_{(i)}, \ldots, \hat{X}_{(n)}$ son estimadores insesgados de X, independientes entre sí, basados cada uno de ellos en el conglomerado último de cada una de las unidades primarias seleccionadas:

$$E\left(\frac{\hat{X}_i}{P_i}\right) = E\left(E_2\left(\frac{\hat{X}_i}{P_i}\right)\right) = E\left(\frac{X_i}{P_i}\right) = \sum_{1}^{N} \frac{X_i}{P_i} P_i = X$$

$$V(\hat{X}) = V\left(\frac{1}{n} \sum_{1}^{n} \hat{X}_{(i)}\right) = \frac{1}{n}\, \sigma^2\,(\hat{X}_{(i)})$$

por otro lado sabemos que la cuasivarianza de n observaciones independientes es un estimador insesgado de la varianza poblacional, entonces podemos afirmar que:

$$\frac{1}{n-1} \sum_{1}^{n} (\hat{X}_{(i)} - \hat{X})^2$$

es un estimador insesgado de $\sigma^2\ (\hat{X}_{(i)})$; por tanto:

$$V(\hat{X}) = \frac{1}{n\,(n-1)} \sum_{1}^{n} (\hat{X}_{(i)} - \hat{X})^2$$

es un estimador insesgado de $V(\hat{X})$.

La aplicación de este método a los distintos tipos de muestreo es inmediata. Si las P_i son proporcionales a M_i, tenemos:

$$\hat{X}_{(i)} = \frac{M\ x_i}{m_i}$$

si además la muestra es autoponderada ($m_i = \bar{m}$) ;

$$\hat{X}_{(i)} = \frac{M}{\bar{m}}\ x_i$$

Si las probabilidades P_i son iguales tendremos:

$$\hat{X}_{(i)} = \frac{N\,M_i}{m_i}\ x_i$$

en el caso general, y

$$\hat{X}_{(i)} = \frac{N}{f_2}\ x_i$$, si la muestra es autoponderada.

Este procedimiento se utiliza frecuentemente por su facilidad de

aplicación práctica ya que se reduce al cálculo de la cuasivarianza de los n valores $\hat{X}_{(i)}$, dividida por n. En el caso de muestras autoponderadas pueden obtenerse fórmulas simplificadas para el cálculo práctico. Si el muestreo se realiza sin reposición el método no proporciona una estimación insesgada, no obstante si la fracción de muestreo en primera etapa es pequeña puede aplicarse como una buena aproximación. También puede aplicarse a estimadores de razón, proporcionando una estimación consistente de la varianza.

V.7. PROCEDIMIENTO GENERAL PARA EL CALCULO DE VARIANZAS. Teorema I de Durbin.

En el punto V.5 de este capítulo hemos calculado la varianza de un estimador insesgado del total, tratando por separado los casos de muestreo con y sin reposición de unidades primarias. Ahora vamos a estudiar un método general válido para ambos casos de muestreo bietápico.

La notacion π_i que hemos reservado hasta hora para expresar la $E(\varepsilon_i)$ en muestreo sin reposición, será utilizada aquí también para muestreo con reposición, sabiendo que en este caso $E(\varepsilon_i) = \pi_i = n P_i$. En resumen:

$$\pi_i = \begin{cases} E(\varepsilon_i) = \Pr(U_i \text{ apareza en la muestra}), \text{ en M.S.R.} \\ E(\varepsilon_i) = n\, P_i, \text{ en M.C.R.} \end{cases}$$

Introducimos también la notación $Y_i = \dfrac{X_i}{\pi_i}$, y análogamente $\hat{Y}_i = \dfrac{\hat{X}_i}{\pi_i}$, resultando entonces:

$$X = \sum_1^N \pi_i\, Y_i = \sum_1^N X_i \text{ es el total poblacional}$$

$\hat{X} = \sum_{1}^{n} \frac{\hat{X}_i}{\pi_i} = \sum_{1}^{n} \hat{Y}_i$ es el estimador insesgado, donde $\hat{Y}_i$

puede interpretarse como la contribución al estimador obtenida del submuestreo de la unidad primaria i–ésima. En estas condiciones vamos a probar que:

$$V(\hat{X}) = V(\hat{X}_1) + \sum_{1}^{N} \pi_i \sigma_i^2$$

donde $\hat{X}_1 = \sum_{1}^{n} Y_i$, es el estimador insesgado del total X en muestreo unietápico (que resulta de sustituir en $\hat{X}$, $\hat{Y}_i$ por su valor esperado) y σ_i^2 es la varianza de $\hat{Y}_i$ en segunda etapa.

Esta fórmula general, debida a Durbin, nos dice que la varianza de un estimador de la forma:

$$\hat{X} = \sum_{1}^{n} \hat{Y}_i$$

en muestreo bietápico se compone de dos sumandos; el primero es la varianza en una etapa del estimador que resulta sustituyendo $\hat{Y}_i$ por su valor esperado en i, y el segundo sumando es la suma ponderada, por π_i, de las varianzas en segunda etapa de $\hat{Y}_i$.

Sabemos que la varianza de una variable aleatoria es igual al momento de orden dos menos el cuadrado de su valor esperado, por tanto podemos escribir:

$$V(\hat{X}) = E(\hat{X}^2) - E(\hat{X})^2 = E\left(\sum_{1}^{n} \hat{Y}_i^2 + \sum_{i \neq j}^{n} \hat{Y}_i \hat{Y}_j \right) - E(\hat{X})^2$$

y para calcular estas esperanzas en muestreo bietápico, consideramos

primero una muestra fija, w, de unidades primarias:

$$\underset{w}{E}(\hat{Y}_i^2) = \underset{w}{E}(\hat{Y}_i)^2 + \underset{w}{V}(\hat{Y}_i) = Y_i^2 + \underset{w}{V}(\hat{Y}_i) \quad ; \text{ por ser } \underset{w}{E}(\hat{Y}_i) = Y_i$$

$\underset{w}{E}(\hat{Y}_i \hat{Y}_j) = \underset{w}{E}(\hat{Y}_i) \underset{w}{E}(\hat{Y}_j) = Y_i Y_j$, por ser independientes los submuestreos efectuados en las distintas unidades de la muestra (Debe observarse que en caso de muestreo con reposición, si por ejemplo una unidad primaria aparece dos veces deberán tomarse dos submuestras independientes). Entonces podemos escribir:

$$\underset{w}{V}(\hat{X}) = \sum_1^n \left(Y_i^2 + \underset{w}{V}(\hat{Y}_i) \right) + \sum_{i \neq j}^n Y_i Y_j - E(\hat{X})^2 =$$

$$= \left(\sum_1^n Y_i \right)^2 - E(\hat{X})^2 + \sum_1^n \underset{w}{V}(\hat{Y}_i)$$

si ahora tomamos esperanza sobre las muestras posibles de primera etapa:

$$V(\hat{X}) = E \left(\sum_1^n Y_i \right)^2 - X^2 + \sum_1^N \pi_i \sigma_i^2 \quad ; \text{ y por ser } \sum_1^n Y_i = \hat{X}_i$$

se obtiene finalmente la fórmula propuesta:

$$V(\hat{X}) = V(\hat{X}_1) + \sum_1^N \pi_i \sigma_i^2$$

Puede comprobarse como ejercicio que las fórmulas para la varianza de $\hat{X}$ obtenidas en V.5.1 y V.5.2, se deducen fácilmente aplicando este

teorema. A modo de ejemplo, calculemos la varianza del estimador de un total de clase. Sea A_i el total de clase en el conglomerado U_i, y sea a_i el total de clase muestral; entonces tenemos:

$$\hat{A} = \sum_{1}^{n} \frac{M_i a_i}{m_i \pi_i} = \sum_{1}^{n} \hat{A}_i \quad ; \text{ siendo } \hat{A}_i = \frac{M_i a_i}{m_i \pi_i}$$

$$E_2(\hat{A}_i) = \frac{A_i}{\pi_i} \quad ; \quad E(\hat{A}) = E\left(\sum_{1}^{n} \frac{A_i}{\pi_i} \right) = \sum_{1}^{N} A_i = A$$

por tanto, de acuerdo con el teorema I de Durbin, resulta:

$$V(\hat{A}) = V\left(\sum_{1}^{n} \frac{A_i}{\pi_i} \right) + \sum_{1}^{N} \pi_i \sigma_i^2 (A_i)$$

si como caso particular consideramos muestreo con reposición y probabilidades iguales, para el primer sumando se tiene:

$$V\left(\sum_{1}^{n} \frac{A_i}{\pi_i} \right) = V\left(\sum_{1}^{n} \frac{N A_i}{n} \right) = \frac{N}{n} \sum_{1}^{N} \left(A_i - \frac{A}{N} \right)^2$$

y para el segundo sumando:

$$\sum_{1}^{N} \pi_i \sigma_i^2 (A_i) = \sum_{1}^{N} \frac{n}{N} \frac{M_i^2}{M_i} \frac{A_i}{M_i} \left(1 - \frac{A_i}{M_i} \right) \frac{N^2}{m_i n^2} =$$

$$= \sum_{1}^{N} \frac{N}{n} \frac{A}{m_i} (M_i - A_i) \text{ . La varianza de } \hat{A} \text{ resulta:}$$

$$V(\hat{A}) = \frac{N}{n} \sum_{1}^{N} \left(A_i - \frac{A}{N} \right)^2 + \frac{N}{n} \sum_{1}^{N} \frac{A_i (M_i - A_i)}{m_i}$$

Si ahora particularizamos para conglomerados de igual tamaño, $M_i = \overline{M}$, y muestra autoponderada, $f_{2i} = \text{cte.}$, denotando por $\alpha_i = \frac{A_i}{M_i} = \frac{A_i}{\overline{M}}$ la proporción de clase en U_i resulta:

$$V(\hat{A}) = \frac{N M^2}{n} \sum_{1}^{N} (\alpha_i - \overline{\alpha})^2 + \frac{N M}{n f_2} \sum_{1}^{N} \alpha_i (1 - \alpha_i)$$

habiendo prescindido en las dos etapas de los factores de finitud y de conversión de varianza en cuasivarianza; así pues en el segundo sumando siempre añadiríamos el factor $(1 - f_2)\ \overline{M}/(\overline{M} - 1)$, y en el primero $(1 - f_1)\ N/(N - 1)$ si el muestreo fuera sin reposición.

V.8. PROCEDIMIENTO GENERAL PARA LA ESTIMACION DE VARIANZAS EN MUESTREO SIN REPOSICION. Teorema II de Durbin.

Del teorema I de Durbin se deduce que la varianza del estimador insesgado del total:

$$\hat{X} = \sum_{1}^{n} \hat{Y}_i$$

en muestreo sin reposición y probabilidades de selección cualesquiera, viene dada por:

$$V(\hat{X}) = V\left(\sum_{1}^{n} \hat{Y}_i \right) + \sum_{1}^{N} \pi_i \sigma_i^2 =$$

$$= \sum_{1}^{N} Y_i^2 \pi_i (1 - \pi_i) + \sum_{i \neq j}^{N} Y_i Y_j (\pi_{ij} - \pi_i \pi_j) + \sum_{1}^{N} \pi_i \sigma_i^2$$

teniendo en cuenta que los dos primeros sumandos representan la varianza en primera etapa, obtenida en el capítulo III, y que aquí utilizamos las notaciones:

$$X_i = \pi_i \, Y_i \quad ; \quad \hat{X}_i = \pi_i \, \hat{Y}_i$$

y que se verifica:

$$\underset{2}{E}(\hat{Y}_i) = Y_i \quad ; \quad \underset{2}{V}(\hat{Y}_i) = \sigma_i^2$$

siendo $\hat{X}_1 = \overline{\sum_{1}^{n}} \, Y_i$ el estimador unietápico.

El teorema II de Durbin afirma que un estimador insesgado de la varianza de $\hat{X}$ se compone de dos sumandos, el primero es la fórmula del estimador insesgado de $V(\hat{X}_1)$, sustituyendo en ella Y_i por su estimador insesgado $\hat{Y}_i$; y el segundo es $\sum_{1}^{N} \pi_i^2 \, \hat{\sigma}_i^2$, donde σ^2 representa el estimador insesgado de σ_i^2 en el submuestreo de la unidad primaria i-ésima.

Veamos su demostración. El estimador insesgado de $V(\hat{X}_1)$ en muestreo unietápico viene dado por:

$$V(\hat{X}_1) = \overline{\sum_{1}^{n}} \, Y_i^2 \, (1 - \pi_i) + \overline{\sum_{i \neq j}^{n}} \, Y_i \, Y_j \, \frac{(\pi_{ij} - \pi_i \, \pi_j)}{\pi_{ij}}$$

según hemos visto en el capítulo III; sustituyendo ahora Y_i por $\hat{Y}_i$, de acuerdo con el enunciado del teorema, se tiene:

$$\hat{v}(\hat{X}_1) = \sum_{1}^{n} \hat{Y}_i^2 \, (1 - \pi_i) + \sum_{i \neq j}^{n} \hat{Y}_i \, \hat{Y}_j \, \frac{(\pi_{ij} - \pi_i \, \pi_j)}{\pi_{ij}}$$

calculemos pues su esperanza matemática en muestreo bietápico; el primer sumando nos da:

$$E\left(\sum_1^n \hat{Y}_i^2\ (1-\pi_i)\right) = E\left(\sum_1^n (1-\pi_i)\ \underset{2}{E}(\hat{Y}_i^2)\right) =$$

$$= E\left(\sum_1^n (1-\pi_i)\ (Y_i^2 + \sigma_i^2)\right) = \sum_1^N (Y_i^2 + \sigma_i^2)\ \pi_i (1-\pi_i)$$

y el segundo:

$$E\left(\sum_{i\neq j}^n \hat{Y}_i \hat{Y}_j \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}\right) = E\left(\sum_{i\neq j}^n \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}\ \underset{2}{E}(\hat{Y}_i)\ \underset{2}{E}(\hat{Y}_j)\right) =$$

$$= \sum_{i\neq j}^n (\pi_{ij} - \pi_i \pi_j)\ Y_i\ Y_j\ ,$$ teniendo en cuenta que el submuestreo se efectua en las distintas unidades primarias de la muestra de forma probabilísticamente independiente. Calculemos ahora:

$$E\ \sum_1^n \pi_i\ \hat{\sigma}_i^2 = E\left(\sum_1^n \pi_i\ \underset{2}{E}(\hat{\sigma}_i^2)\right) = E\left(\sum_1^n \pi_i\ \sigma_i^2\right) = \sum_1^N \pi_i^2\ \sigma_i^2$$

por último la suma de las tres esperanzas calculadas prueba el teorema:

$$E\left(\hat{v}(\hat{X}_1)\right) + E\left(\sum \pi_i\ \hat{\sigma}_i^2\right) =$$

$$= \sum_1^N (Y_i^2 + \sigma_i^2)\ \pi_i (1-\pi_i) + \sum_{i\neq j}^N (\pi_{ij} - \pi_i \pi_i)\ Y_i\ Y_j + \sum_1^N \pi_i^2\ \sigma_i^2 =$$

$$= \sum_1^N Y_i^2\ \pi_i (1-\pi_i) + \sum_{i\neq j}^N (\pi_{ij} - \pi_i \pi_j)\ Y_i\ Y_j + \sum_1^N \pi_i\ \sigma_i^2$$

V.9. TAMAÑO OPTIMO DE UNA MUESTRA BIETAPICA

Suponemos que se realiza muestreo con probabilidades iguales, sin reposición, en las dos etapas y que la muestra es autoponderada; admitimos además que todos los conglomerados son de igual tamaño. En la práctica no siempre se cumplen exactamente todas estas condiciones, sin embargo es una situación bastante común y por otra parte, simplifica los desarrollos de las fórmulas permitiendo llegar a un resultado suficientemente aproximado para conducir nuestra decisión acerca de los tamaños óptimos. Admitimos como función de coste:

$$C = c_1 \, n + c_2 \, n \, \bar{m}$$

siendo c_1 el coste unitario de una unidad primaria y c_2 el coste de cada unidad secundaria.

En estas condiciones la varianza del estimador insesgado del total

$$\hat{\bar{X}} = \frac{N}{n} \frac{\bar{M}}{\bar{m}} \sum_1^n x_i$$

obtenida en el punto V.5.2 es:

$$V(\hat{\bar{X}}) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) S_c^2 + \frac{N^2 (\bar{M} - \bar{m}) \bar{M}}{n \, \bar{m}} \sum_1^N S_i^2$$

y teniendo en cuenta las notaciones introducidas en el capítulo III, se puede escribir para el estimador insesgado de la media:

$$V(\hat{\bar{\bar{X}}}) = \frac{N}{(N-1)} \left(\frac{1}{n} - \frac{1}{N} \right) \sigma_b^2 + \frac{(\bar{M} - \bar{m})}{(\bar{M} - 1) \, n \, \bar{m}} \sigma_w^2$$

y con la simplificación

$$N = N - 1 \quad ; \quad \bar{M} = \bar{M} - 1 \, ,$$

la fórmula queda:

$$\hat{\hat{V}}(\bar{X}) = \frac{1}{n}\left(\sigma_b^2 - \frac{\sigma_w^2}{\bar{M}}\right) + \frac{\sigma_w^2}{n\,\bar{m}} - \frac{\sigma_b^2}{N}$$

Para determinar el mínimo de $\hat{\hat{V}}(\bar{X})$ con un coste total, C, dado, el último término de la varianza puede olvidarse puesto que es constante independientemente de los valores de n y de $\bar{m}$. Despejando de la función de coste n en función de $\bar{m}$, c_1 y c_2 y llevándolo a la varianza, el problema consiste en determinar el mínimo, en $\bar{m}$, de la función:

$$\phi = \frac{(c_1 + c_2\,\bar{m})}{C}\left(\sigma_b^2 - \frac{\sigma_w^2}{\bar{M}}\right) + \frac{\sigma_w^2\,(c_1 + c_2\,\bar{m})}{C\,\bar{m}}$$

cuya derivada respecto de $\bar{m}$ igualada a cero nos da:

$$c_2\,A - \frac{c_1\,\sigma_w^2}{\bar{m}^2} = 0 \quad ; \quad A = \sigma_b^2 - \frac{\sigma_w^2}{\bar{M}}$$

de donde resulta que el valor óptimo de $\bar{m}$, es:

$$\bar{m}^* = \sqrt{\frac{c_1}{c_2}}\sqrt{\frac{\sigma_w^2}{A}} = \sqrt{\frac{\sigma_w^2}{\sigma_b^2 - \sigma_w^2/\bar{M}}}$$

y el valor óptimo de n, que se obtiene llevando el óptimo de $\bar{m}$ a la función de coste, resulta:

$$n^* = \frac{C}{c_1 + c_2\,\bar{m}}$$

Observamos que el valor óptimo del tamaño de submuestreo es tanto mayor cuanto mayor sea la raiz cuadrada de la relación c_1/c_2 entre

entre los costes unitarios y cuanto mayor sea la cantidad:

$$\sqrt{\frac{\sigma_w^2}{\sigma_b^2 - \sigma_w^2 / \bar{M}}}$$

Si la variabilidad en la población de unidades secundarias es igual a la variabilidad dentro de los conglomerados; esto es, como si la formación de las unidades primarias fuese hecha "al azar", el denominador se anula (supuesto $\bar{M} = \bar{M} - 1$) y obtendríamos que $\bar{m}^*$ se hace infinito; en la práctica esto implicaría tomar $\bar{m} = \bar{M}$, o lo que es lo mismo no efectuar submuestreo. También se tomaría la misma decisión si el citado denominador es negativo. Este razonamiento resulta inmediato teniendo en cuenta que siendo δ el coeficiente de homogeneidad de conglomerados se cumple:

$$\sigma_b^2 - \sigma_w^2 / \bar{M} = \sigma^2 \delta \; ; \text{ con la aproximación dada}$$

y resulta para el óptimo:

$$\bar{m}^* = \sqrt{\frac{\sigma_w^2}{\sigma^2 \delta}}$$

y en resumen podemos afirmar que si es $\delta \leqslant 0$, la solución óptima consiste en efectuar muestreo unietápico y cuanto mayor que cero sea δ, menor tamaño de submuestra conviene tomar en cada unidad primaria. En el caso extremo de que sea $\delta = 1$, lo que implica que la varianza "dentro" es cero y la varianza "entre" es igual a la varianza total σ^2, se obtiene $\bar{m} = 0$ debido al uso de aproximaciones; en la práctica se tomaría $\bar{m} = 1$ ya que en esta situación, recordando lo estudiado en el cáp. III, las unidades secundarias dentro de un mismo conglomerado son todas iguales y por tanto una sola, nos da toda la información.

Se comprueba fácilmente que:

$$\frac{\sigma_w^2}{\sigma_b^2 - \sigma_w^2 / \bar{M}} = \frac{1 - \delta}{\delta}$$

utilizando como venimos haciendo la aproximación $\overline{M} = \overline{M} - 1$. Así resulta que el tamaño óptimo de $\bar{m}$, puede expresarse mediante la fórmula:

$$\bar{m}^* = \sqrt{\frac{c_1}{c_2}} \sqrt{\frac{1-\delta}{\delta}}$$

a la que puede llegarse de la misma forma que hemos hecho aquí, utilizando como hacen Hansen, Hurwitz y Madow, para la varianza del estimador la fórmula aproximada:

$$\hat{V}(\bar{X}) = \frac{\sigma^2}{n\,\bar{m}} \quad (1 - (\bar{m} - 1)\,\delta)$$

Por otra parte, también se comprueba fácilmente que se llega a los mismos resultados si se plantea el problema de minimizar el coste para un valor dado de la varianza.

V.10. EJEMPLOS

Ejemplo V.1

El siguiente cuadro muestra, para una población constituída por cinco unidades primarias, los valores X_{ij} correspondientes a cada una de las unidades secundarias

Unidades primarias	Unidades secundarias Valores de X_{ij} para j =									
	1	2	3	4	5	6	7	8	9	10
U_1	5	4	7	5	6	5				
U_2	8	6	8	9	6	7	10	8	6	
U_3	8	12	10	11	12	9	13	12		
U_4	12	9	11	10	10	8	12	10	13	12
U_5	14	12	10	11	12	13	13			

Supongamos que se extrae una muestra aleatoria en dos etapas, definida por los subíndices de las unidades de la siguiente forma:

i	j
2	1 , 4
3	2 , 5 , 7
5	3 , 4 , 6

A) Calcule la estimación del total poblacional, la media por unidad primaria y la media por unidad secundaria, en los siguientes casos de muestreo

1) En primera etapa probabilidades de selección iguales.

2) En primera etapa probabilidades proporcionales a los tamaños M_i.

B) En alguno de los casos anteriores ¿Es autoponderada la muestra?

Solución:

A1) $\hat{X} = 424$; $\hat{\bar{X}} = 84{,}8$; $\hat{\bar{\bar{X}}} = 10{,}6$

A2) $\hat{X} = 429$; $\hat{\bar{X}} = 85{,}8$; $\hat{\bar{\bar{X}}} = 10{,}7$

B) En ninguno de los dos casos es autoponderada.

Ejemplo V.2

Para la población del ejemplo V.1 calcule el error de muestreo relativo del estimador insesgado del total, en muestreo con probabilidades proporcionales al tamaño y tamaños $n = 3, m_i = 4$ para todo i.

Solución:

La fórmula que debemos aplicar es:

$$V(\hat{X}) = \frac{M}{n}\left\{\sum_{1}^{N}\frac{X_i^2}{M_i} - \frac{X^2}{M}\right\} + \frac{M}{n}\sum_{1}^{N}\frac{(1-f_{2i})}{f_{2i}} \cdot S_i^2$$

y haciendo los cálculos necesarios se obtiene:

$$V(\hat{\bar{X}}) = 2887{,}87 + 144{,}93 = 3032{,}80$$

el error de muestreo es 55,07 y el error de muestreo relativo, dado que el total poblacional es 379, es 14,5 % .

Ejemplo V.3

En una población de N = 10 conglomerados se efectúa muestreo en dos etapas, obteniéndose los siguientes datos:

Unidades primarias de la muestra $n = 3$	Tamaño M_i	Valores observados X_{ij} ; $m_i = \overline{m} = 5$
A_1	50	8, 6, 12, 14, 10
A_2	60	8, 10, 14, 14, 16
A_3	80	8, 10, 10, 16, 12

Sabiendo que la suma de los tamaños es M = 600, se pide formar un estimador insesgado del total X y calcular el valor particular correspondiente a los datos del problema, en los siguientes casos:

A) Muestreo con probabilidades iguales en las dos etapas.

B) Muestreo con probabilidades proporcionales al tamaño en primera etapa.

C) Estimar el error de muestreo relativo, en ambos casos, por el método de los conglomerados últimos.

D) ¿En qué caso la muestra es autoponderada?

Solución:

A) $\hat{X} = \frac{N}{m} \sum_{1}^{3} M_i x_i = 7133.$ B) $\hat{X} = \frac{M}{m} \sum_{1}^{3} x_i = 6720$

C) El error de muestreo relativo estimado por el método de los conglomerados últimos es en el caso A) 16,2 % y en el caso B) 15,2 % .

D) La muestra es autoponderada en el caso B.

Ejemplo V.4

Una población consta de N = 3 unidades primarias, conteniendo $M_1 = 100$, $M_2 = 200$, $M_3 = 300$ unidades secundarias. La proporción de unidades de una clase, dentro de cada unidad primaria, es

$$P_1 = 0{,}40 \; ; \; P_2 = 0{,}45 \; ; \; P_3 = 0{,}35$$

Calcule el E.C.M. del estimador del total de clase en los siguientes casos, siendo el tamaño de submuestreo $\overline{m} = 50$ y $n = 1$ el tamaño en primera etapa:

A) Probabilidades iguales en primera etapa. Estimador $\hat{A} = M\, p_i$, siendo p_i la proporción muestral.

B) Igual método de selección que en el caso A). Estimador $\hat{A} = N M_i p_i$

C) Probabilidades proporcionales al tamaño en primera etapa. Estimador $\hat{A} = M\, p_i$

Solución:

Efectuemos algunos cálculos previos. La proporción y el total de clase en la población, son:

$$P = \sum_1^3 \frac{M_i P_i}{M} = 0{,}392 \; ; \; A = \sum_1^3 M_i P_i = 235$$

las varianzas dentro de unidades primarias:

$$P_i Q_i = 0{,}24 \quad ; \quad 0{,}2475 \; ; \quad 0{,}2275 \; .$$

Caso (A)

$$E(\hat{A}) = \underset{2}{E}(M E(p_i) = E(MP_i) = M \sum_1^3 P_i \cdot \frac{1}{N} = \frac{M}{N} \sum_1^3 P_i = 240$$

por tanto el estimador tiene un sesgo $= 240 - 235 = 5$. Para la varianza:

$$V(\hat{A}) = \underset{1\ 2}{E(V(\hat{A}))} + \underset{1\ 2}{V(E(\hat{A}))}$$

el primer sumando es:

$$\frac{M^2}{N} \cdot \frac{M_i - \bar{m}}{M_i - 1} \cdot \frac{P_i Q_i}{\bar{m}} = 1188$$

y el segundo:

$$M^2 \left(\sum_{1}^{3} \frac{P_i^2}{N} - P^2 \right) = 2869. \text{ Por tanto resulta:}$$

$$V(\hat{A}) = 4057 \ ; \ ECM(\hat{A}) = 25 + 4057 = 4082 .$$

Caso (B)

$$E(\hat{A}) = E(N M_i p_i) = E(N \underset{2}{E}(M_i p_i) = E(N \hat{A}_i) = A. \text{ Es insesgado.}$$

El primer sumando de la varianza:

$$\underset{2}{E}(V(\hat{A})) = E(N^2 \underset{2}{V}(M_i p_i)) = 1548. \text{ El segundo sumando:}$$

$$\underset{2}{V}(E(\hat{A})) = V(N M_i P_i) = N^2 V(\hat{A}_i) = 6950. \text{ Por tanto ECM(A)=8498.}$$

Caso (C)

Operando de forma análoga se comprueba queses insesgado y su varianza es $V(\hat{A}) = 1307 + 716 = 2023$.

Ejemplo V.5

Para la función de coste $C = c_1 n + c_2 n \bar{m}$, se supone en un momento dado, los siguientes datos: $\sigma^2 = 1800$; $c_1 = 90$; $c_2 = 10$; $\delta = 0{,}2$ para $C = 3000$. Se pide calcular en cuanto aumenta, o disminuye, la varianza de $\hat{\bar{X}}$, utilizando tamaños óptimos de muestra, si investigaciones posteriores revelan que los verdaderos valores de δ y σ^2 son: 0,1 y 2000 respectivamente.

Solución:

Bajo la primera hipótesis se tiene:

$\bar{m} = 6$; $n = 20$; $V(\hat{\bar{X}}) = 30$ y con los verdaderos valores de δ y σ^2:

$\bar{m} = 9$; $n = 17$; $V(\hat{\bar{X}}) = 21{,}2$, por tanto la varianza disminuye en un 29,3%

Ejemplo V.6

En una ciudad hay M viviendas que se consideran agrupadas por secciones (unidades primarias) siendo N el número de éstas. Para estimar la proporción de viviendas alguno de cuyos miembros dispone de automóvil, se toma una muestra de la siguiente forma:

(A) n secciones con probabilidad proporcional al número de viviendas y submuestra de un número fijo $\bar{m}$ de viviendas.

(B) Lo mismo que en (A) pero con tamaño de submuestreo m_i variable.

(C) n secciones con probabilidades iguales y un número fijo $\bar{m}$ de viviendas en segunda etapa.

(D) Lo mismo que en (C) pero con tamaño de submuestreo m_i variable.

Formar en cada caso el estimador insesgado P y la fórmula para estimar su varianza por el método de los conglomerados últimos.

Solución:

Caso (A): $\hat{P} = \frac{1}{n\,\bar{m}} \sum_{1}^{n} a_i$; siendo a_i el total de clase muestral.

$$\hat{V}(\hat{P}) = \frac{1}{(n-1)\,\bar{m}^2\,n} \left\{ \sum_{1}^{n} a_i^2 - \frac{1}{n} \left(\sum_{1}^{n} a_i \right)^2 \right\}$$

Caso (B): $\hat{P} = \frac{1}{n} \sum_{1}^{n} p_i$; siendo p_i la proporción muestral. Para la estimación de la varianza queda:

$$\hat{V}(\hat{P}) = \frac{1}{n\,(n-1)} \left\{ \sum_{1}^{n} p_i^2 - \frac{1}{n} \left(\sum_{1}^{n} p_i \right)^2 \right\}$$

Caso (C): $\hat{P} = \frac{N}{M\,n\,m} \sum_{1}^{n} M_i\,a_i$. Para la estimación de la varianza:

$$\hat{V}(\hat{P}) = \frac{N^2}{n(n-1)\,\bar{m}^2\,M^2} \left\{ \sum_{1}^{n} M_i^2\,a_i^2 - \frac{1}{n} \left(\sum_{1}^{n} M_i\,a_i \right)^2 \right\}$$

Caso (D): $\hat{P} = \frac{N}{M\,n} \sum_{1}^{n} \frac{M_i\,a_i}{m_i}$. Para la estimación de la varianza, tenemos:

$$\hat{V}(\hat{P}) = \frac{N^2}{(n-1)\,M^2\,n} \left\{ \sum_{1}^{n} M_i^2\,p_i^2 - \frac{1}{n} \left(\sum_{1}^{n} M_i\,p_i \right)^2 \right\}$$

Ejemplo V.7

Para una característica cualitativa suponemos conocidos los datos $P = 0{,}2$ y $\sigma_b^2 = 0{,}02$, siendo $\overline{M} = 200$ el tamaño de las unidades primarias. Determine el coste total de una muestra con tamaños óptimos, con $c_1 = 525$ y $c_2 = 100$, para estimar la proporción P con un error de muestreo del 10% .

Solución:

De los datos se deduce $\delta = 0{,}1212$ y $\overline{m} = 6$, por tanto:

$$\text{de } \frac{V(P)}{P^2} = 0{,}16\,(1 + 0{,}1212 \,.\, 5)/6 \,.\, 0{,}04 \,.\, n = (0{,}10)^2$$

se obtiene $n = 107$, y por último $C = 120375$.

Ejemplo V.8

Supongamos que en una muestra aleatoria de tamaño n se observa una variable Y_i . A continuación se toma una submuestra de la primera de tamaño $n_1 < n$, en la que se observa la variable X_i. Determine la esperanza y varianza del estimador $\hat{\overline{D}} = \overline{y} - \overline{x}_1$ de la diferencia de medias, siendo $\overline{y}$ la media de la primera muestra y $\overline{x}_1$ la media de la submuestra. (Suponga muestreo sin reposición y probabilidades iguales).

Solución:

$$E(\underset{2}{E}(\hat{D})) = E(\underset{2}{E}(\overline{y} - \overline{x}_1)) = E(\overline{y} - \underset{2}{E}(\overline{x}_1)) = E(\overline{y} - \overline{x}) = \overline{X} - \overline{Y}$$

$V(\hat{D}) = \underset{2}{V}(E(\hat{D})) + E(\underset{2}{V}(\hat{D}))$. El primer sumando de la varianza nos da.

$$\frac{(1-f)\,(S_y^2 + S_x^2 - 2\,S_{xy})}{n}$$ y el segundo:

$\frac{S_x^2}{n_1}$. Por tanto la varianza es $V(\hat{D}) = \frac{(1-f)(S_y^2 + S_x^2 - 2S_{xy})}{n} + \frac{S_x^2}{n_1}$

Ejemplo V.9

En la situación del ejemplo V.8, suponiendo una función de coste $C = c_1 n + c_2 n_1$, determine los tamaños óptimos de $\pi = n_1 / n$ y de n, para conseguir varianza mínima con coste dado. Aplique la solución a los datos: $f = 0{,}05$; $c_1 = 50$; $c_2 = 100$ suponiendo que la relación S_y^2 / S_x^2 toma el valor 2 y el coeficiente de correlación vale 0,3.

Solución:

Llevando $n = C / (c_1 + c_2 \pi)$ obtenida de la función de coste, a la varianza e igualando a cero su derivada respecto de π, resulta:

$$\pi = \sqrt{\frac{c_1}{c_2}} \sqrt{\frac{1}{(1-f)\left(1 + \frac{S_y^2}{S_x^2} - 2\delta \frac{S_y}{S_x}\right)}}$$

y el valor óptimo de n se obtiene sustituyendo en la función de coste el valor óptimo encontrado para π.

Para los datos dados resulta: $\pi = 0{,}33$; $n = \frac{C}{83}$

Ejemplo V.10

Consideremos muestreo sin reposición y probabilidades iguales de una población de N unidades. Sea W una muestra de tamaño n, con media $\bar{x}(W)$ y cuasivarianza s^2 (W); ahora supongamos que se extrae aleatoriamente una submuestra de tamaño $m = n/K$, resultando $\bar{x}_r$ su media. Compruebe que $\hat{X}_r = N \bar{x}_r$ es un estimador insesgado del total poblacional $\hat{X}$, con varianza $(K - f) N^2 S^2 / n$.

Solución:

$E(\hat{X}_r) = E_W(E(\hat{X}_r)) = E_W(E(N\bar{x}_r)) = E(N\bar{x}) = X$. Por tanto es insesgado.

$V(\hat{X}_r) = V_W(E(\hat{X}_r)) + E_W(V(\hat{X}_r))$. El primer sumando de la varianza es:

$V_W(E(\hat{X}_r)) = V(N\bar{x}) = (1 - f)\dfrac{N^2 S^2}{n}$; y el segundo, teniendo en cuenta la fracción de submuestreo es $m/n = 1/K$, resulta:

$E_W(V(\hat{X}_r)) = E_W(V(N\bar{x}_r)) = E(N^2 (1 - \dfrac{1}{K}) \dfrac{s^2(W)}{n/K}$. Por tanto:

$V(\hat{X}_r) = \dfrac{N^2 S^2}{n} (1 - f + K - 1) = (K - f) \dfrac{N^2 S^2}{n}$. Observemos que en muestreo con reposición, $f = 0$ y $S^2 = \sigma^2$, por tanto en este caso la varianza de $\hat{X}_r$ sería exactamente K veces mayor que la de $\hat{X}$, resultado obvio por ser la submuestra de tamaño K veces menor.

VI. METODOS ESPECIALES PARA LA ESTIMACION DE VARIANZAS

El conocimiento de las varianzas es un elemento de especial interés, tanto para el usuario de estadísticas obtenidas mediante muestreo probabilístico, ya que le permite decidir acerca del grado de validez o confianza de los datos en relación con el uso que va a hacer de ellos, como al diseñador ya que además de servir a los intereses de los usuarios, le permite decidir entre varios procedimientos alternativos o mejorar un diseño ya establecido.

En diseños polietápicos complejos, cuando se estiman muchas características poblacionales, la aplicación de las fórmulas ordinarias de los estimadores de la varianza puede ocasionar tediosos problemas de cálculo. Por ello se han desarrollado algunas técnicas que permiten aplicar fórmulas más sencillas, aunque se pierda algo de precisión. Veamos en este capítulo algunos métodos más importantes.

VI.1. METODO DE LOS GRUPOS ALEATORIOS

Supongamos que la muestra consta de n unidades extraídas de una población de tamaño N. Si se subdivide aleatoriamente en K submuestras que se denominan *"Grupos"* de igual tamaño, m, de modo que $n = K\ m$; cada grupo además de ser una submuestra aleatoria de la muestra completa es una muestra de la población con las mismas propiedades probabilís-

ticas, si bien de menor tamaño. En estas condiciones cada grupo proporciona una estimación utilizando la misma fórmula que la diseñada para la muestra completa. Si es W la muestra, podemos imaginar que la formación de los grupos se hace numerando de 1 a n los elementos de W y después de elegir una permutación aleatoria de los números $1,2,\ldots,n$ asignamos al primer grupo los elementos correspondientes a los m primeros enteros de la permutación, al segundo los m siguientes y así sucesivamente.

Si $\hat{\theta}$ es un estimador insesgado de θ basado en la muestra completa W, y si $\hat{\theta}_r$ es el estimador basado en el r-ésimo grupo aleatorio, un estimador insesgado de la varianza de $\hat{\theta}$, en muestreo con reposición viene dado por:

$$\hat{v}(\hat{\theta}) = \frac{1}{K(K-1)} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$$

Para estudiar el fundamento de este método veamos primero el caso más simple: Muestreo con reposición y probabilidades iguales de una población de tamaño N. Para el total sabemos que:

$$\hat{X} = \frac{N}{n} x \quad ; \quad V(\hat{X}) = N^2 \frac{\sigma^2}{n}$$

entonces, siendo x_r el total de r-ésimo grupo:

$$\hat{X}_r = \frac{N}{m} x_r \quad : \quad V(\hat{X}_r) = N^2 \frac{\sigma^2}{m} = K\, V(\hat{X})$$

esta última fórmula se deduce inmediatamente pues el r° grupo es una muestra aleatoria de la población de tamaño $m = n/K$; una demostración formal puede verse en el ejemplo V.10. También se cumple:

$$\underset{W}{E}(\hat{X}_r) = \hat{X} \quad ; \quad \text{por tanto } E(\hat{X}_r) = E(\underset{W}{E}(\hat{X}_r)) = X$$

por ser el r° grupo una submuestra aleatoria de W.

Descomponemos $V(\hat{X}_r)$ de la siguiente forma:

$$V(\hat{X}_r) = E(\hat{X}_r - X)^2 = E(\hat{X}_r - \hat{X} + \hat{X} - X)^2 =$$

$$= E(X_r - \hat{X})^2 + E(\hat{X} - X)^2 + 2\,E(\hat{X}_r - X)\,(\hat{X} - X)$$

y puesto que el tercer sumando se anula, podemos escribir:

$E(\hat{X}_r - \hat{X})^2 = V(\hat{X}_r) - V(X) = (K-1)\,V(\hat{X})$ puesto que se cumple:

$$V(\hat{X}_r) = K\,V(\hat{X}) \qquad \text{y} \qquad E(\hat{X} - X)^2 = V(\hat{X})$$

Si ahora sumamos en $r := 1, 2, \ldots, K$ la igualdad obtenida:

$$\sum_{r=1}^{K} E(\hat{X}_r - \hat{X})^2 = E\left(\sum_{r=1}^{K} (\hat{X}_r - X)^2 \right) = K\,(K-1)\,V(\hat{X})$$

por tanto

$$\frac{E\left(\sum_{r=1}^{K} (\hat{X}_r - X)^2 \right)}{K\,(K-1)} = V(\hat{X}) \quad ; \quad \text{de donde resulta que}$$

$$\frac{\sum_{r=1}^{K} (\hat{X}_r - X)^2}{K\,(K-1)} \quad \text{es un estimador insesgado de } V(\hat{X}).$$

Veamos ahora el caso general. Si se cumplen las condiciones:

$$\text{(A)} \qquad V(\hat{\theta}_r) = K\,V(\hat{\theta}) \quad ; \qquad \text{(B)} \qquad \underset{W}{E}(\hat{\theta}_r) = \hat{\theta}$$

operando como antes, tenemos $V(\hat{\theta}_r) = E(\hat{\theta}_r - \hat{\theta})^2 + E(\hat{\theta} - \theta)^2 \Rightarrow$

$$K\,V(\hat{\theta}) = E(\hat{\theta}_r - \hat{\theta})^2 + V(\hat{\theta}) \Rightarrow E(\hat{\theta}_r - \hat{\theta})^2 = (K-1)\,V(\hat{\theta})$$

por tanto sumando en r las esperanzas se obtiene,

$$\frac{E\left\{\sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2\right\}}{K(K-1)} = V(\hat{\theta})$$

La condición (A) se cumple en muestreo con reposición y la (B) se cumplirá si $\hat{\theta}_r$ es una copia de $\hat{\theta}$ aplicada a una submuestra aleatoria de tamaño m. El método puede aplicarse a muestreo polietápico siempre que formemos los K grupos aleatorios con m unidades primarias de las n que constituyen la muestra completa. También se aplica al muestreo estratificado formando K grupos aleatorios en cada estrato. En muestreo de unidades primarias sin reposición y probabilidades iguales, debemos de multiplicar $\hat{v}(\hat{X})$ por el factor de finitud (1–f) para que el estimador permanezca insesgado. Para estimadores de razón el método es válido si son insesgados; en caso contrario debe tomarse como una buena aproximación si el sesgo es pequeño en relación con la varianza.

GRUPOS ALEATORIOS DE DISTINTO TAMAÑO

La condición de subdividir la muestra completa en K grupos aleatorios de igual tamaño no es necesaria para la aplicación del método. Si W se subdivide en grupos de tamaños $m_1, m_2, \ldots, m_r, \ldots, m_K$ cuya suma es n, sea $\lambda_r = m_r/n$; en estas condiciones la condición (A) se convierte en $V(\hat{\theta}_r) = (1/\lambda_r)\ V(\hat{\theta})$ permaneciendo inalterada la condición (B) y el estimador insesgado de $V(\hat{\theta})$ es:

$$\hat{v}(\hat{\theta}) = \frac{1}{K-1} \sum_{1}^{K} \lambda_r (\hat{\theta}_r - \hat{\theta})^2$$

tomando ahora $\hat{\theta} = \Sigma\ \lambda_r\ \hat{\theta}_r$.

Observemos por último que el método de los grupos aleatorios se ha descrito mediante una partición aleatoria de la muestra completa en K grupos de igual, o distinto tamaño; pues bien, en muestreo con reposición es equivalente a extraer K muestras de la población de igual, o distinto tamaño, formando la muestra completa como unión de aquéllas.

PRECISION DEL METODO

Estamos ahora interesados en estudiar la precisión del estimador de $V(\hat{\theta})$ dado por el método de los grupos aleatorios. Para ello calculamos la varianza relativa de $\hat{v}(\hat{\theta})$ usando dos resultados válidos en muestreo de poblaciones infinitas o muestreo con reposición:

(I). Si $z_1, z_2, \ldots, z_i, \ldots, z_n$ son observaciones independientes e igualmente distribuidas, la varianza relativa de:

$$s_1^2 = \frac{\sum_1^n (z_i - \bar{z})^2}{n} \quad \text{, viene dada por } \frac{1}{n}\left(\beta - \frac{n-3}{n-1}\right)$$

siendo β el coeficiente $\dfrac{\mu_4}{\sigma^4}$ de la variable z_i.

(II) Para la media de las n observaciones el coeficiente β es:

$$\beta(\bar{z}) = \frac{1}{n}(\beta + 3(n-1))$$

Cuando $\hat{\theta}$ es un estimador lineal también lo es cada uno de los $\hat{\theta}_r$, y pueden escribirse en forma de una media:

$$\hat{\theta} = \frac{1}{n}\sum_1^n t_i \quad ; \quad \hat{\theta}_r = \frac{1}{m}\sum_1^m t_i \,, \quad \text{por tanto } \hat{\theta} = \frac{1}{K}\sum_1^K \hat{\theta}_r$$

donde t_i es función de la observación efectuada en la unidad i-ésima de

la muestra; en el caso de un diseño simple para la media, t_i sería la observación X_i ; en caso del total sería N t_i y en diseños polietápicos aparecería otro factor debido al submuestreo.

Por otra parte puesto que K es una constante, la varianza relativa de $\hat{v}(\hat{\theta})$ es la misma que la de:

$\frac{1}{K} \sum_{1}^{K} (\hat{\theta}_r - \hat{\theta})^2$. Aplicando entonces la fórmula (I) a las estimaciones $\hat{\theta}_1$, $\hat{\theta}_2$, , . . . , $\hat{\theta}_r$, . . . , $\hat{\theta}_K$ podemos escribir:

$$VR(\hat{v}) = \frac{1}{K} \left(\beta (\hat{\theta}_r) - \frac{K-3}{K-1} \right) \text{ , y aplicando (II) tenemos:}$$

$$\beta (\hat{\theta}_r) = \frac{1}{m} (\beta + 3 (m-1)) \text{, que sustituido en la anterior, nos}$$

da para la varianza relativa de $\hat{v}$:

$$VR(\hat{v}) = \frac{1}{K} \left(\frac{\beta}{m} + \frac{3m}{m} - \frac{3}{m} - \frac{K-3}{K-1} \right) = \frac{\beta-3}{m} + \frac{3}{K} - \frac{K-3}{K-1},$$

$$VR(\hat{v}) = \frac{\beta-3}{n} + \frac{2}{K-1}$$

Se observa que la precisión de la varianza estimada por el método de los grupos aleatorios es tanto menor cuanto menor sea el número de grupos, puesto que al disminuir K aumenta el segundo sumando de la fórmula obtenida. Obviamente la precisión también depende del tamaño total de la muestra y del coeficiente de curtosis β de los valores poblacionales t_i.

VI.2. METODO DE LOS CONGLOMERADOS ULTIMOS

En el capítulo V hemos visto para estimadores lineales en muestreo bietápico este método, que ahora podemos expresar de una forma general teniendo en cuenta que puede considerarse como un caso particular del método de los grupos aleatorios. En efecto, si hacemos $K = n$, lo que equivale a tomar cada una de las n observaciones como un grupo aleatorio de tamaño 1, resulta:

$$\frac{1}{n\,(n-1)} \sum_{1}^{n} (\hat{\theta}_i - \hat{\theta})^2$$

estimador insesgado de $V(\hat{\theta})$, siendo $\hat{\theta} = \frac{1}{n} \sum_{1}^{n} \hat{\theta}_i$, y suponiendo $V(\hat{\theta}_i) = n\ V(\hat{\theta})$, equivalente a muestreo con reposición.

La precisión de este estimador de $V(\hat{\theta})$ viene dada por:

$$VR(\hat{v})_{n=K} = \frac{\beta - 3}{n} + \frac{2}{n-1}$$

donde β es el coeficiente de curtosis de t_i. Como referencia notemos que si $\beta = 3$ (Normal), para un tamaño $n = 40$ la varianza relativa de $\hat{v}$ es del 5%.

Observemos ahora que a igual tamaño de la muestra completa, la precisión del método de los conglomerados últimos es mayor que la del método de los grupos. En el primer caso se tiene $K = n$, y en el segundo $K = n/m$, por tanto la diferencia entre las varianzas relativas:

$$\frac{2}{K-1} - \frac{2}{n-1} = \frac{2\,(n-K)}{(K-1)\,(n-1)} > 0$$

por ser $n > K$. Así pues, se deduce que para estimar una varianza es mejor disponer de n estimaciones individuales $\hat{\theta}_i$ que de K estimaciones $\hat{\theta}_r$ basadas en m unidades cada una, siendo por supuesto $n = K\,m$.

VI.3 METODO DE SEMIMUESTRAS REITERADAS

Como hemos visto, la aplicación del método de los grupos aleatorios supone que disponemos de K muestras independientes de la población, que en conjunto constituyen una muestra completa, W, de tamaño n. Supongamos ahora que dada una muestra aleatoria, W, de tamaño n, que venimos denominando muestra completa, extraemos de ella una submuestra aleatoria de tamaño $n/2$, supuesto n par; tenemos entonces una "semimuestra". Si reponiendo la semimuestra repetimos K veces la selección, tendremos K semimuestras reiteradas de W. Observemos que ahora la unión de las K semimuestras no coincide con la muestra completa, no se trata pues de "grupos" en el sentido utilizado para los métodos anteriores.

Si $\hat{\theta}$ es un estimador insesgado de θ, basado en la muestra completa, y $\hat{\theta}_r$ es también un estimador de θ, basado en la r-ésima semimuestra reiterada a la que se aplica el mismo estimador

$$\frac{1}{K} \sum_{r=1}^{K} (\theta_r - \hat{\theta})^2$$

es un estimador insesgado de $V(\hat{\theta})$, supuesto que se cumplen las condiciones:

$$(A) \quad V(\hat{\theta}_r) = 2\,V(\hat{\theta}) \quad ; \quad (B) \quad \underset{W}{E}(\hat{\theta}_r) = \hat{\theta}$$

La condición (A) es obvia en muestreo con reposición puesto que por ser cada reiteración de tamaño $n/2$, la varianza de $\hat{\theta}_r$ será el doble que la de $\hat{\theta}$ si utilizamos el mismo estimador. La condición (B) también es inmediata por ser cada reiteración una muestra aleatoria de una W fija.

Descomponiendo $V(\hat{\theta}_r)$ de la siguiente forma:

$$V(\hat{\theta}_r) = E(\hat{\theta}_r - \hat{\theta} + \hat{\theta} - \theta)^2 = E(\hat{\theta}_r - \hat{\theta})^2 + E(\hat{\theta} - \theta)^2 +$$

$$+ 2E\,((\hat{\theta}_r - \hat{\theta})(\hat{\theta} - \hat{\theta}))$$

el tercer sumando se anula ya que $\underset{W}{E}(\hat{\theta}_r) = \bar{\hat{\theta}}$ implica $E(\hat{\theta}_r - \hat{\theta}) = 0$.

Teniendo en cuenta ahora la condición (A) podemos escribir:

$2\,V(\theta) = E(\hat{\theta}_r - \hat{\theta})^2 + V(\hat{\theta})$, de donde resulta $V(\hat{\theta}) = E(\hat{\theta}_r - \hat{\theta})^2$

y por último:

$$E\left(\frac{1}{K}\sum_{r=1}^{K}(\hat{\theta}_r - \hat{\theta})^2\right) = \frac{1}{K}\,K\,E(\hat{\theta}_r - \hat{\theta})^2 = V(\hat{\theta}).$$

PRECISION DEL METODO

Admitiendo que $\hat{\theta}$ es lineal, puede escribirse en forma de una media de n observaciones:

$$\hat{\theta} = \frac{1}{n}\sum_{1}^{n} t_i \text{ , por tanto también } \hat{\theta}_r = \frac{2}{n}\sum_{1}^{n/2} t_i$$

Por otra parte, para una muestra W fijada, $\hat{\theta}_1$, $\hat{\theta}_2$, . . . , $\hat{\theta}_r$, . . . , $\hat{\theta}_K$ son estimaciones independientes e insesgadas de $\hat{\theta}$; aplicando la fórmula (I) podemos escribir que la varianza relativa de:

$\hat{v}(\hat{\theta}) = \frac{1}{K}\sum_{1}^{K}(\hat{\theta}_r - \hat{\theta})^2$ viene dada por

$VR(\hat{v}\,|\,W) = \frac{1}{K}\left(\beta\,(\hat{\theta}_r) - \frac{K-3}{K-1}\right)$ donde $\beta\,(\hat{\theta}_r)$ es el coeficiente

$\frac{\mu_4}{\sigma^4}$ de $\hat{\theta}_r$, fijada W. Aplicando la fórmula (II) tenemos:

$$\beta\,(\hat{\theta}_r) = \frac{2}{n}\left(\beta_W + 3\left(\frac{n}{2} - 1\right)\right)$$

siendo β_W el coeficiente $\frac{\mu_4}{\sigma^4}$ de los valores t_i de la muestra W fijada.

Sustituyendo $\beta(\hat{\theta}_r)$ resulta:

$$VR(\hat{v}|W) = \frac{1}{K}\left(\frac{2\ \beta_W}{n} + \frac{3n}{n} - \frac{6}{n} - \frac{K-3}{K-1}\right), \text{y simplificando:}$$

$$VR(\hat{v}|W) = \frac{1}{K}\left(\frac{2\,(\beta_W - 3)}{n} + \frac{2K}{K-1}\right)$$

Debe observarse que esta fórmula nos da la varianza relativa del estimador, condicionada a la muestra W disponible. Si esta muestra completa es una buena representación de la población β_W será próximo al β; en caso contrario la variabilidad de β_W, al considerar las distintas posibles muestras W, nos hace aumentar la varianza del estimador de $V(\hat{\theta})$ dado por este métododo de semimuestras reiteradas.

Como indicación práctica, notemos que si $\beta_W = 3$ (Normal) con $K = 40$ reiteraciones se obtiene una varianza relativa del 5%.

FORMA DE INCREMENTAR LA PRECISION DEL METODO

Veamos ahora cómo pude incrementarse la precisión del método, particularmente cuando la muestra es pequeña. Antes de entrar en esta cuestión estudiemos otra forma de probar que el método de semimuestras proporciona una estimación insesgada de la varianza del estimador $\hat{\theta}$, con el fin de analizar con más detalle esta técnica. Puesto que cada reiteración puede contemplarse como una muestra de tamaño $n/2$ seleccionada en dos etapas, la varianza de $\hat{\theta}_r$ es:

$$V(\hat{\theta}_r) = E\left(\underset{W}{V}(\hat{\theta}_r)\right) + V\left(\underset{W}{E}(\hat{\theta}_r)\right)$$

y dado que se cumple: $V(\hat{\theta}_r) = 2\,V(\hat{\theta})$; $\underset{W}{E}(\hat{\theta}_r) = \hat{\theta}$, tenemos:

$$2\,V(\hat{\theta}) = E\left(\underset{W}{V}(\hat{\theta}_r)\right) + V(\hat{\theta}) \text{ , de donde resulta:}$$

$$E\left(\underset{W}{V}(\hat{\theta}_r)\right) = V(\hat{\theta})$$

En esta relación observamos que la varianza de $\hat{\theta}_r$, condicionada a la muestra completa, es un estimador insesgado de $V(\hat{\theta})$. Entonces, si podemos encontrar un estimador insesgado de $\underset{W}{V}(\hat{\theta}_r)$, al tomar esperanza matemática sobre el espacio de muestras W, resulta que también será insesgado pra $V(\hat{\theta})$. El estimador $\hat{v}$ cumple esta condición ya que:

$$\underset{W}{E}\left(\frac{1}{K}\sum_{1}^{K}(\hat{\theta}_r - \hat{\theta})^2\right) = \underset{W}{E}(\hat{\theta}_r - \hat{\theta})^2 = V(\hat{\theta}_r) \text{ , por tanto:}$$

$$E(\hat{v}) = E\left(\underset{W}{E}(\hat{v})\right) = E\left(\underset{W}{V}(\hat{\theta}_r)\right) = V(\hat{\theta})$$

haciendo uso de la relación obtenida anteriormente.

De acuerdo con estos resultados, cuando en la práctica tenemos una determinada muestra completa W. $\hat{v}$ estima a $\underset{W}{V}(\hat{\theta}_r)$, por tanto debemos intentar mejorar en lo posible esta estimación. Es claro que si tomamos todas las $n^{n/2}$ semimuestras posibles, $\hat{v}$ nos daría exactamente $\underset{W}{V}(\hat{\theta}_r)$, pero es evidente que incluso para valores moderados de n se trata de un número muy grande. Vamos a ver que si las reiteraciones se seleccionan sin reposición y de modo que no aparezcan dos complementarias podemos, para un K dado, mejorar la precisión de $\hat{v}$. En primer lugar es obvio que si las K semimuestras se toman sin reposición, se sigue cumpliendo que:

$$\underset{W}{E}(\hat{v}) = \underset{W}{V}(\hat{\theta}_r) = \frac{1}{H}\sum(\hat{\theta}_r - \hat{\theta})^2$$

donde $H = n^{n/2}$ es el número de todas las posibles reiteraciones. Por otro lado observemos que siendo $d_r = (\hat{\theta}_r - \hat{\theta})$ la desviación de una reiteración a su media, lo que se trata de estimar es $\frac{1}{H} \sum_{1}^{H} d_r^2$. Si damos el nombre de *reiteraciones complementarias* a dos semimuestras W_r, W_j cuya unión es W, se cumple:

$$\frac{1}{2} (\hat{\theta}_r + \hat{\theta}_j) = \hat{\theta}$$

ya que siendo $\hat{\theta}$ lineal:

$$\frac{1}{2} \left(\frac{2}{n} \sum_{K=1}^{n/2} t_{rk} + \frac{2}{n} \sum_{h=1}^{n/2} t_{jh} \right) = \frac{1}{n} \sum_{i=1}^{n} t_i = \hat{\theta}$$

en consecuencia, si dos semimuestras son complementarias nos dan la misma información acerca de d^2, puesto que:

$$d_r^2 = (\hat{\theta}_r - \hat{\theta})^2 = (2\hat{\theta} - \hat{\theta}_j - \hat{\theta})^2 = (\hat{\theta} - \hat{\theta}_j)^2 = d_j^2$$

De estos resultados se deduce que si del conjunto de todas las posible reiteraciones, fijada W, eliminamos las que tienen unidades repetidas y una de cada dos complementarias entre sí, resulta que con

$C = \frac{1}{2} \binom{n}{n/2}$ semimuestras obtendríamos exactamente $\underset{W}{V} (\hat{\theta}_r)$,

ya que:

$$V(\hat{\theta}_r) = \frac{1}{H} \sum d_r^2 = \frac{1}{C} \sum \delta_r^2$$

denotando por δ las diferencias $(\hat{\theta}_r - \hat{\theta})$ correspondientes a las C semimuestras sin elementos repetidos y no complementarias.

En resumen, cuando n es un número relativamente pequeño, tomando las K reiteraciones sin reposición y de modo que no aparezcan dos complementarias incrementamos la precisión de $\hat{v}$ como estimador de $V(\hat{\theta}_r)$. En el siguiente cuadro se muestra para unos cuantos valores

n	$H = n^{n/2}$	$C = \frac{1}{2}\binom{n}{n/2}$
4	16	3
6	216	10
8	4096	35
10	100000	126

de n, el número H de posibles reiteraciones y el número C aplicado a la técnica que acabamos de describir. Se deduce que si se seleccionan K semimuestras la varianza relativa de $\hat{v}$, calculada antes para muestreo aleatorio simple (con reposición), se multiplicaría por el factor:

$$\left(1 - \frac{K}{C}\right) = \frac{C - K}{C - 1} \leqslant 1, \text{ para todo } K > 1$$

así por ejemplo para $n = 8$, con $K = 40$ semimuestras $VR(\hat{v})$ se multiplicaría por $(70 - 40)/69 = 0{,}43$; esto es, se reduce a menos de la mitad. Obviamente para $K = C$, la varianza de $\hat{v}$, dada W, es cero y para K=1 no se distinguen los dos procedimientos.

VI.4. APLICACION AL MUESTREO ESTRATIFICADO

Suponemos que el estadístico cuya varianza se desea estimar puede escribirse como $\hat{\theta} = \Sigma\, \hat{\theta}_h$. Al considerar muestreo estratificado, teóricamente caben dos posibilidades de aplicar los métodos que estamos estudiando; bien aplicando "por separado" a cada estrato para obtener estimaciones de las $V(\hat{\theta}_h)$, de modo que por ser:

$$V(\hat{\theta}) = \sum V(\hat{\theta}_h)\text{, tendríamos } \hat{v}(\hat{\theta}) = \sum \hat{v}(\hat{\theta}_h)$$

o bien formando K estimaciones de θ, basadas en una muestra estratificada, cada una de ellas obtenida seleccionando un "grupo" o una semimuestra en cada estrato.

Desde el punto de vista de la precisión de las estimaciones de $V(\hat{\theta})$ es preferible la técnica por separado, pero el procedimiento combinado da lugar a una gran simplificación de los cálculos ya que solo necesita una suma de cuadrados de estadísticos que tienen la misma fórmula que la empleada en las estimaciones que proporciona la encuesta.

Por otra parte, como ya hemos visto, el método de los conglomerados últimos es preferible al de los grupos aleatorios, puesto que recoge información acerca de la varianza, de todas y cada una de las n unidades de la muestra. Otra observación interesante es que para muestreo estratificado la técnica combinada de conglomerados últimos solamente puede aplicarse en el caso, poco realista, de tener igual tamaño de muestra en todos los estratos.

El método de semimuestras reiteradas también puede aplicarse al muestreo estratificado con la técnica combinada o por separado; la primera de ellas que tiene la ventaja de la sencillez de la fórmula de cálculo práctico, impone al investigados la elección de las K semimuestras de forma tal que pierda la mínima información acerca de la varianza contenida en la muestra completa W dada.

METODO DE LOS CONGLOMERADOS ULTIMOS

Supuesto que $\hat{\theta}_h = \frac{1}{n_h} \sum_{1}^{n_h} \hat{\theta}_{hj}$ es el estimador de θ_h, y $\hat{\theta}_{hj}$ es la estimación de θ_h basada en la j-ésima unidad muestral del estrato h, la fórmula de la técnica por separado es:

$$\hat{v}(\hat{\theta}) = \sum_{h=1}^{L} \frac{1}{n_h (n_h - 1)} \sum_{j=1}^{n_h} (\hat{\theta}_{hj} - \hat{\theta}_h)^2$$

Para la técnica combinada, si tenemos m unidades muestrales en cada estrato, tomando una de cada estrato podemos formar el estimador global:

$$\hat{\theta}_j = \sum_{h=1}^{L} \hat{\theta}_{hj} \text{ , y siendo ahora } \hat{\theta} = \frac{1}{m} \sum_{j=1}^{m} \hat{\theta}_j$$

resulta:
$$\hat{v}(\hat{\theta}) = \frac{1}{m(m-1)} \sum_{1}^{m} (\hat{\theta}_j - \hat{\theta})^2$$

METODO DE SEMIMUESTRAS REITERADAS

La técnica por separado se aplica de la siguiente forma: Tomando K_h seminuestras en el estrato h, se forma el estimador $\hat{\theta}_{hr}$ de θ_h basado en la r-ésima reiteración, siendo θ_h basado en la muestra completa del estrato; entonces el estimador insesgado de la varianza de $\hat{\theta} = \Sigma \hat{\theta}_h$ viene dado por:

$$\hat{v}(\hat{\theta}) = \sum_{h=1}^{L} \frac{1}{K_h} \sum_{r=1}^{K_h} (\hat{\theta}_{hr} - \hat{\theta}_h)^2$$

Si en todos los estratos se toma igual número K de semimuestras:

$$\hat{v}(\hat{\theta}) = \frac{1}{K} \sum_{h=1}^{L} \sum_{r=1}^{K_h} (\hat{\theta}_{hr} - \hat{\theta}_h)^2$$

Este procedimiento, por separado, mejora su precisión si en cada estrato seleccionamos semimuestras sin reposición y no complementarias, siendo especialmente conveniente cuando no es muy grande el número de unidades muestrales por estrato. Tiene la desventaja de que se necesitan tantas sumas de cuadrados como estratos.

La técnica combinada se aplica de la siguiente forma: Seleccionamos una semimuestra en cada estrato, cuya unión es una semimuestra de la muestra completa, que nos da la estimación de θ:

$$\hat{\theta}_r = \sum_{h=1}^{L} \hat{\theta}_{hr}$$

si repetimos la selección K veces, $r = 1,2,\ldots,K$, el estimador insesgado de la varianza de $\hat{\theta}$ viene dado por:

$$\hat{v}(\hat{\theta}) = \frac{1}{K} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$$

VI.5. DISEÑOS CON DOS UNIDADES MUESTRALES POR ESTRATO

En ocasiones se utilizan diseños que constan de un número relativamente grande de estratos en cada uno de los que se toman dos unidades muestrales. Estudiemos la forma de estimar la varianza de un estimador lineal, utilizando la técnica de semimuestras reiteradas por "separado" y la "combinada".

TECNICA POR SEPARADO

Si denotamos por $\hat{\theta}_{h1}$ y $\hat{\theta}_{h2}$ las estimaciones de $\hat{\theta}_h$ basadas en cada una de las dos unidades disponibles en el estrato h, cuando el estimador es lineal $\hat{\theta}_h = (1/2)(\hat{\theta}_{h1} + \hat{\theta}_{h2})$, se cumple:

$$V(\hat{\theta}_{hi}) = \frac{1}{2}\left\{(\hat{\theta}_{h1} - \hat{\theta}_h)^2 + (\hat{\theta}_{h2} - \hat{\theta}_h)^2\right\}$$

y siendo $i = 1,2$ aleatorio: $(\hat{\theta}_{hi} - \hat{\theta}_h)^2$ es insesgado para $V(\hat{\theta}_{hi})$ en consecuencia el método de semimuestras aplicado por separado nos da:

$$\hat{v}(\hat{\theta}) = \sum_{h=1}^{L} (\hat{\theta}_{hi} - \hat{\theta}_h)^2$$

como estimador insesgado de $V(\hat{\theta})$, donde para cada L , i toma aleatoriamente el valor 1 ó 2. Pero observando que en todo estrato se cumple:

$$(\hat{\theta}_{h1} - \hat{\theta}_h)^2 = (\hat{\theta}_{h2} - \hat{\theta}_h)^2 = \frac{1}{4}(\hat{\theta}_{h1} - \hat{\theta}_{h2})^2 = \frac{1}{4} d_h^2$$

resulta la siguiente fórmula:

$$\hat{v}(\hat{\theta}) = \frac{1}{4} \sum_{h=1}^{L} (\hat{\theta}_{h1} - \hat{\theta}_{h2})^2 = \frac{1}{4} \sum_{h=1}^{L} d_h^2$$

que coincide con el método de Keyfitz.

TECNICA COMBINADA

Si elegimos en todos y cada uno de los estratos una de las dos unidades aleatoriamente, por ejemplo para L = 4, (1, 1, 2, 2) representa elegir la primera unidad en los dos primeros estratos y la segunda en los otros dos, podemos dar una estimación total de θ:

$$\hat{\theta}_r = \sum_{h=1}^{L} \hat{\theta}_{hi}$$

si repetimos K veces la selección aleatoria de una semimuestra:

$$\hat{v}(\hat{\theta}) = \frac{1}{K} \sum_{r=1}^{K} (\hat{\theta}_r - \hat{\theta})^2$$

es el estimador insesgado de $V(\hat{\theta})$.

Un problema que presenta esta técnica consiste en seleccionar dentro de las 2^L posibles semimuestras estratificadas, un conjunto de K de ellas que nos de la máxima precisión. Mc Carthy (1969) propone la solución que describimos en el siguiente párrafo.

SEMIMUESTRAS EQUILIBRADAS

Observemos la estimación dada por una reiteración $\hat{\theta}_r$ obtenida seleccionando un i aleatorio igual a 1 ó 2 en cada estrato:

$$(\hat{\theta}_r - \hat{\theta})^2 = \left(\sum_{h=1}^{L} \hat{\theta}_{hi} - \Sigma \hat{\theta}_h \right)^2 = \left\{ \sum_{h=1}^{L} (\hat{\theta}_{hi} - \hat{\theta}_h) \right\}^2 =$$

$$= \frac{1}{4} \left(\sum_{h=1}^{L} d_{hi} \right)^2 = \frac{1}{4} \sum_{h=1}^{L} d_{hi}^2 + \frac{1}{4} \sum_{h \neq j} d_{hi} \cdot d_{ji'}$$

según esto cada reiteración nos da la mejor estimación de la varianza, para una muestra completa dada, que es el primer sumando de la fórmula anterior, más un término debido a sumas de productos cruzados entre distintos estratos. Estos productos tendrán signos positivos o negativos que se anularán promediando sobre todas las 2^L posibles reiteraciones. Mc Carthy encuentra que es posible elegir un número $K \geqslant L$ de semimuestras "equilibradas" de forma que promediando sobre ellas los términos debidos a los productos cruzados se anulan, para ello denotando por $+1$ la elección de la primera unidad de un estrato y por -1 la elección de la segunda, la matriz formada escribiendo en filas las semimuestras (L números $+1$ ó -1) debe tener todas sus columnas ortogonales.

El sigiente problema reside en formar dicha matriz cuando L es grande; existe un método debido a Plackett y Burman para matrices de orden múltiplo de cuatro. De esta forma cuando L no es múltiplo de cuatro hay que tomar una, dos o tres reiteraciones más que número de estratos.

Veamos ahora la justificación de esta técnica: En cada estrato una de las diferencias d_{hi} es positiva y otra negativa, ambas iguales en valor absoluto, por tratarse de desviaciones respecto a su media común; entonces si elegimos en un estrato h una unidad tenemos d_h y si elegimos la otra $-d_h$. Observemos que ocurre en el caso L = 3, K = 4 utilizando una matriz de semimuestras ortogonal:

Matriz de columnas ortogonales			$\left(\sum_{L=1}^{3} d_h\right)^2$
+	+	+	$(d_1 + d_2 + d_3)^2$
+	−	−	$(d_1 - d_2 - d_3)^2$
−	−	+	$(-d_1 - d_2 + d_3)^2$
−	+	−	$(-d_1 + d_2 - d_3)^2$

Al efectuar la suma de las cuatro reiteraciones queda $\sum d_h^2$, anulándose los productos cruzados.

VI.6. EJEMPLOS

Ejemplo VI. 1

Supongamos que cinco investigadores toman muestras independientes de igual tamaño, constituidas por pequeñas parcelas de un campo de cultivo y obtienen estimaciones del rendimiento del campo, θ; sean: 97 ; 96 ; 100 ; 98 ; 94. Si tomamos como estimador de θ la media de las cinco, calcule el error de muestreo relativo.

Solución:

Admitiendo que los cinco investigadores utilizan el mismo procedimiento de medida, podemos aplicar el método de los grupos aleatorios; la muestra de cada uno puede considerarse como un grupo aleatorio, formando con su unión una muestra completa.

$\hat{\theta}_r$	$(\hat{\theta}_r - \hat{\theta})^2$
97	0
96	1
100	9
98	1
94	9
485	20

$$\hat{\theta} = \frac{485}{5} = 97$$

$$\hat{v}(\hat{\theta}) = \frac{20}{5.4} = 1$$

error de muestreo = 1,03%

Ejemplo VI.2

Suponga en el ejemplo VI.1 que las muestras son de distintos tamaños, respectivamente: 3 ; 1; 10; 10; 1.

Solución:

$\hat{\theta} = 98{,}44$; $v(\hat{\theta}) = 0{,}5816$; error de muestreo = 0,77%

Ejemplo VI.3

De una población N = 40, se ha seleccionado una muestra aleatoria de tamaño $n = 8$. En el siguiente cuadro aparecen los valores X_i observados, y con el signo * se indican cuatro semimuestras. Estime la varianza de $\hat{\theta} = (N/n) \sum X_i$ por el método de semimuestras reiteradas.

Semimuestras reiteradas	Valores de X_i							
	10	10	16	12	16	16	12	10
1	*		*	*				*
2			*		*	*		*
3		*	*	*		*		
4	*			*		*	*	

Solución:

$$\hat{\theta} = 510 \; ; \; \hat{v}(\hat{\theta}) = 1700 \; ; \text{ error de muestreo} = 8{,}1\%$$

Ejemplo VI.4

Se dispone de una muestra de tamaño $n = 80$. Aplicando el método de los grupos aleatorios ¿En que porcentaje disminuye la varianza relativa utilizando K = 20, en vez de K = 5? Particularice para los valores $\beta = 2 \; ; \; 3 \; ; \; 4$.

Solución:

$$\frac{15 \; . \; 160}{38 \, (2\beta + 74)} \; x \, 100$$

Ejemplo VI.5

En una muestra aleatoria de $n = 3$ unidades primarias se ha efectuado un submuestreo de m_i unidades secundarias, obteniéndose las siguientes observaciones p_i. Estime el error de muestreo del estadístico $\hat{\theta} = \sum_{1}^{n} m_i \, p_i$, por el método de los conglomerados últimos.

$$m_1 = 10 \; ; \; m_2 = 12 \; ; \; m_3 = 8 \; ;$$

$$p_1 = 0{,}20 \; ; \; p_2 = 0{,}25 \; ; \; p_3 = 0{,}15$$

Solución:

Escribiendo $\hat{\theta} = \frac{1}{n} \sum_{1}^{n} n \, m_i \, p_i$, es $\hat{\theta}_i = n \, m_i \, p_i$

se obtiene $\hat{v}(\hat{\theta}) = 2{,}44$; error de muestreo $= 1{,}56$

Ejemplo VI.6

Para el método de los grupos aleatorios de distinto tamaño compruebe que la fórmula simplificada es:

$$\hat{v}(\hat{\theta}) = \frac{1}{(K-1)} \left\{ \frac{1}{n} \sum_{1}^{K} m_r \, \hat{\theta}_r^2 - \hat{\theta}^2 \right\}$$

Ejemplo VI.7

En un diseño con $L = 3$ y dos unidades muestrales por estrato, se utiliza el estimador:

$$\hat{\theta} = \sum_{L=1}^{3} \hat{\theta}_h \qquad \text{siendo} \quad \hat{\theta}_h = F_h \, \frac{x_h}{n_h}$$

En el siguiente cuadro se dan los totales muestrales de cada unidad

x_{h1} y x_{h2} asi como los factores de elevación F_h.

Valores de x_{h1} y x_{h2} en el estrato h		
$h = 1$	$h = 2$	$h = 3$
8 , 10	11 , 13	12 , 16
$F_1 = 100$	$F_2 = 50$	$F_3 = 20$

(A) Determine la estimación de $\hat{\theta}$.

(B) Estime la varianza de $\hat{\theta}$ por el método de semimuestras separado"

(C) Estime la varianza de $\hat{\theta}$ por el método combinado, siendo K = 4 ; (1,2,1) ; (1,1,2) = (2,1,1) ; (2,2,1)

Solución:

(A) $\hat{\theta} = 1780$; (B) $\hat{v}(\hat{\theta}) = 14100$, error de muestreo 6,7% ;

(C) $\hat{v}(\hat{\theta}) = 8100$, error de muestreo 5 ,05%

Ejemplo VI.8

Utilice para la pregunta (C) el ejemplo VI.7 el conjunto de 4 semimuestras equilibradas (2,2,2) ; (2,1,1) ; (1,1,2) ; (1,2,1). Compruebe que obtiene la misma estimación que en el caso (B) del citado ejemplo.

Ejemplo VI.9

Tenemos una muestra de $n = 4$ unidades para el estimador $\hat{\theta} = \sum \hat{\theta}_i / n$. Utilizando un conjunto completo de semimuestras no complementarias, estime la varianza de $\hat{\theta}$. Los valores $\hat{\theta}_i$ son: 4, 10, 18, 28. Compruebe que llega al mismo resultado que si utiliza el estimador ordinario .

Solución: $\hat{v}(\hat{\theta}) = 81 / 3$

VII. ALGUNAS TECNICAS UTILES

Reunimos en este capítulo tres temas diferentes, en gran parte basados en fundamentos estudiados anteriormente, que aportan soluciones interesantes a algunos problemas que se presentan frecuentemente en la práctica.

VII.1. MUESTREO SISTEMATICO

Para una población $\mathcal{A} = \{A_1, A_2, \dots, A_N\}$ recibe el nombre de *muestreo sistemático,* el siguiente procedimiento de selección de una muestra de tamaño n. Se determina el número $K = N/n$, suponiendo que N es múltiplo de n, a continuación se elige un entero j con probabilidades iguales en el conjunto $1 \leqslant j \leqslant K$, resultando la muestra formada por las unidades:

$$\{A_j, A_{j+K}, A_{j+2K}, \dots, A_{j+(n-1)K}\}$$

Según esto para una numeración dada de las unidades de la población, el espacio muestral está constituido por K posibles muestras que podemos representar en el siguiente cuadro:

MUESTRAS POSIBLES					
W_1	W_2	–	W_j	–	W_K
A_1	A_2	–	A_j	–	A_K
A_{1+K}	A_{2+K}	–	A_{j+K}	–	A_{2K}
A_{1+2K}	A_{2+2K}	–	A_{j+2K}	–	A_{3K}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$A_{1+(n-1)K}$	$A_{2+(n-1)K}$	–	$A_{j+(n-1)K}$	–	A_{nK}

Por ser las K muestras equiprobables y formadas por n unidades cada una, la probabilidad de que una unidad aparezca en la muestra es igual para todas: $\pi_i = \Pr(A_i) = n/N$.

Si N no es exactamente múltiplo de n , caben dos posibilidades para determinar el período K:

(1) K = Parte entera de N/n
(2) K = (Parte entera de N/n) + 1

En el primer caso se formarían $H = N - Kn$ muestras de tamaño $n+1$, y $K - H$ muestras de tamaño n. En el segundo $H = H - K(n-1)$ de tamaño n y $K - H$ de tamaño $n - 1$. Cuando n no es pequeño, pongamos mayor o igual que 50, no cabe esperar que la variación en una unidad del tamaño muestral produzca una perturbación sustancial en la distribución de los estimadores. No obstante debe observarse que si las últimas son "unidades grandes" las muestras que tienen una unidad más, presentarían respecto de las restantes una cierta asimetría hacia la derecha.

Como ventajas de la selección sistemática citamos:

1. Suele ser fácil y rápida la selección de la muestra, tanto en el trabajo de oficina, utilizando ficheros u otros elementos similares donde las unidades puedan numerarse correlativamente, como en el trabajo de campo.

2. Si las unidades de la población están numeradas "al azar" el muestreo sistemático producirá muestras muy representativas, ya que ninguna sucesión grande quedará sin estar representada en la muestra.

3. Si las unidades de la población están numeradas de modo que dos de ellas son más parecidas cuanto más cerca están en la sucesión, el muestreo sistemático produce un efecto "similar" a la estratificación. Los falsos estratos serían los grupos de K unidades consecutivas, y la muestra contiene una unidad de cada uno. Es obvio que no se trata de una muestra estratificada pues ésta requiere selecciones independientes en cada estrato.

Una situación inconveniente se produce si la numeración de las unidades es tal que la sucesión de valores X_i presenta una tendencia periódica. En este caso las unidades de la muestra darían todas el mismo, o parecido valor, y en consecuencia pierden representatividad dotando a la distribución de los estimadores de una varianza grande.

VII.1.1. ANALISIS DEL MUESTREO SISTEMATICO. ESTIMACION DE LA MEDIA

Si observamos el cuadro de muestras posibles, resulta claro que la técnica del muestreo sistemático es equivalente a considerar una población de K grupos:

$$\Omega = \{ W_1, W_2, \ldots, W_j, \ldots, W_K \}$$

cada uno formado por n unidades de $\mathcal{A}$, de los cuales se selecciona uno con probabilidades iguales. El análisis puede hacerse tomando como referencia la teoría desarrollada en el capítulo III para conglomerados de

igual tamaño. Así pues, siendo:

$$\sigma^2 = \frac{1}{N} \sum_{1}^{N} (X_i - \bar{\bar{X}})^2 \qquad \text{Varianza poblacional}$$

$$\sigma_j^2 = \frac{1}{n} \sum_{1}^{n} (X_{jh} - \bar{X}_j)^2 \qquad \text{Varianza interna de } W_j$$

$$\sigma_w^2 = \frac{1}{K} \sum_{1}^{K} \sigma_j^2 \qquad \text{Varianza "dentro"}$$

$$\sigma_b^2 = \frac{1}{K} \sum_{1}^{K} (\bar{X}_j - \bar{\bar{X}})^2 \qquad \text{Varianza "entre"}$$

se cumple: $\sigma^2 = \sigma_b^2 + \sigma_w^2$

Análogamente definimos el *coeficiente de homogeneidad intramuestra* como:

$$\delta = \frac{\sigma_b^2 - \sigma_w^2/(n-1)}{\sigma^2}$$

La media de la muestra es un estimador insesgado de la media poblacional, ya que:

$$E(\bar{X}_j) = \sum_{1}^{K} \bar{X}_j \Pr(W_j) = \frac{1}{k} \sum_{1}^{K} \bar{X}_j = \bar{\bar{X}}$$

cuya varianza, obviamente, es: $V(\bar{X}_j) = \sigma_b^2 = \frac{1}{K} \sum_{1}^{K} (\bar{X}_j - \bar{\bar{X}})^2$

y teniendo en cuenta que se cumple: $\sigma_b^2 = \frac{\sigma^2}{n} (1 + (n-1)\delta)$

podemos escribir: $V(\bar{X}_j) = \frac{\sigma^2}{n} (1+(n-1)\delta) =$

$$= \frac{N-1}{N} \frac{S^2}{n} (1 + (n-1)\delta)$$

Para discutir la ganancia o pérdida de precisión del muestreo sistemático respecto del muestreo simple, comencemos por ver cuando coinciden las varianzas; de la ecuación:

$$\frac{N-n}{N} \frac{S^2}{n} = \frac{N-1}{N} \frac{S^2}{n} (1+(n-1)\delta)$$

se deduce inmediatamente que debe ser $\delta = -\frac{1}{N-1}$

En resumen, si el coeficiente de correlación intramuestra es menor que $-1/(N-1)$, prácticamente negativo para N grande, el muestreo sistemático proporciona mayor precisión que el muestreo simple: recíprocamente si δ es cero o positivo proporciona menos precisión. Este resultado confirma las ideas intuitivas expresadas al principio de que para muestreo sistemático conviene que las unidades estén numeradas u ordenadas de forma que las muestras resulten hetorogéneas dentro, con lo cual son más representativas, y en consecuencia homogéneas entre, con lo cual las estimaciones tienen menos varianza.

El análisis puede hacerse en función de las cuasivarianzas en vez de utilizar el coeficiente δ . Utilizando la notación:

$$S_w^2 = \frac{n}{n-1} \sigma_w^2$$

la igual precisión de ambos tipos de muestreo, nos lleva a la condición:

$$\frac{N-n}{Nn} S^2 = \sigma_b^2 = \sigma^2 - \sigma_w^2 = \frac{N-1}{N} S^2 - \frac{n-1}{n} S_w^2$$

de donde resulta $\left(\frac{N-1}{N} - \frac{N-n}{Nn} \right) S^2 = \frac{n-1}{n} S_w^2$, simplificando:

se obtiene $$S^2 = S_w^2$$

Esto significa que la variación total (S^2) es igual a la variación dentro de muestras (S_w^2) . Si S_w^2 es mayor que S^2 es más preciso el muestreo sistemático y menos preciso en caso contrario.

ESTIMACION DE LA VARIANZA

Observemos que con una muestra sistemática no es posible estimar la varianza. Ello es debido a que equivale a una muestra de tamaño uno entre un conjunto de K posibles; obviamente una sola observación no permite hacer inferencias acerca de la variabilidad.

Veamos que ocurre si tomamos como estimador de la varianza de la media muestral el estimador usual de muestreo simple:

$$E\left(\frac{N-n}{N} \cdot \frac{s_j^2}{n} \right) = \frac{N-n}{Nn} E(s_j^2) = \frac{N-n}{Nn} S_w^2$$

El valor esperado de la varianza de $\bar{X}_j$ estimada de la muestra, no coincide con la verdadera varianza, que como sabemos es σ_b^2 . Se observa también que el estimador sería insesgado bajo la hipótesis de que:

$$\frac{N-n}{Nn} S_w^2 = \sigma_b^2 = \frac{N-1}{N} S^2 - \frac{n-1}{n} S_w^2$$

o lo que es lo mismo: $S^2 = S_w^2$. Esta hipótesis, en rigor, solamente puede admitirse si, antes de seleccionar la muestra sistemática, se ha efectuado una aleatorización del orden de las unidades; en este caso el valor esperado, sobre las N ! ordenaciones posibles de la población, de S^2 coincide con el de S_w^2 .

Si se desea estimar, al mismo tiempo, la media de la población $\bar{\bar{X}}$ y la varianza del estimador (media muestral) debemos modificar el diseño de la siguiente forma. Se seleccionan h muestras sistemáticas de tamaño m, tal que $h.m = n$; entonces si es $\bar{X}_j$ la media de la j-êsima, la media de la muestra completa es:

$$\hat{\bar{\bar{X}}} = \frac{1}{h} \sum_{j=1}^{h} \bar{X}_j = \frac{1}{h} \sum_{j=1}^{h} \frac{1}{m} X_j = \frac{1}{n} x = \bar{x}$$

y podemos aplicar el método de los grupos aleatorios estudiado en el capítulo VI, que nos da como estimador insesgado de $V(\bar{x})$:

$$\hat{V}(\bar{x}) = \frac{1}{h(h-1)} \sum_{j=1}^{h} (\bar{X}_j - \bar{x})^2$$

VII.2. MUESTREO EN DOS OCASIONES

Cuando estamos interesados en estudiar la evolución de una determinada característica de la población a lo largo del tiempo, como p.e. la producción industrial, los salarios, la población activa, se toman periódicamente muestras del mismo colectivo. En esta situación es habitual que un objetivo sea el de estimar el cambio habido en la variable estudiada desde la ocasión anterior. Como veremos, un buen método consiste en diseñar una muestra que permanece fija de una ocasión a otra pero tiene el inconveniente de que las personas o entidades encuestadas son reacias a permanecer por un tiempo indefinido en dicha muestra. Para tratar de resolver este problema se utiliza un procedimiento que consiste en sustituir, en cada período de encuesta, una parte de la muestra. Conviene observar de pasada que esto no siempre puede practicarse, ya que cuando se trata de unidades muy grandes (grandes almacenes, siderúrgicas, astilleros, etc.) a veces una o unas pocas contribuyen al total estimado en una cantidad superior a todas las demás juntas; en este caso prescindiríamos del muestreo incluyéndolas en un estrato de probabilidad 1.

Otro problema que puede plantearse es el de la estimación óptima de la segunda ocasión, utilizando las informaciones disponibles, tanto de la ocasión presente como de la anterior.

En cualquier caso el valor X_i que toma la variable en la unidad A_i puede cambiar de una ocasión a la siguiente, por tanto denotaremos por Y_i el valor correspondiente a la ocasión $t+1$ en la misma unidad. El coeficiente de correlación lineal entre los valores de la variable en una y otra ocasión, que juega un papel importante en esta teoría, es:

$$\rho = \frac{\mathrm{cov}(X_i, Y_i)}{\sigma_x \ \sigma_y} = \frac{\sigma_{xy}}{\sigma_x \ \sigma_y}$$

Las unidades de la muestra en la ocasión $t+1$ pueden ser:

(A) Las mismas que en la ocasión anterior.
(B) Algunas son nuevas y otras permanecen.
(C) Seleccionadas independientemente de nuevo.

En el caso general denotamos por n_1 el tamaño de la muestra en primera ocasión, n_2 en la segunda y por n' el tamaño de muestra común a las dos ocasiones; también escribiremos $n' = \pi n_1$, siendo $0 \leqslant \pi \leqslant 1$ la proporción de unidades comunes.

VII.2.1. ESTIMACION DEL CAMBIO

Consideremos en primer lugar el problema de estimación del cambio entre las dos ocasiones, definido por la diferencia de medias $\overline{D} = \overline{Y} - \overline{X}$; para ello comenzamos por estudiar el estimador más simple:

$$\overline{d} = \overline{y} - \overline{x}$$

que obviamente es insesgado. Para el cálculo de su varianza procedemos de la siguiente forma:

$$V(\overline{d}) = V(\overline{y} - \overline{x}) = E\left\{(\overline{y} - \overline{Y}) - (\overline{x} - \overline{X})\right\}^2 =$$

$$E(\overline{y} - \overline{Y})^2 + E(\overline{x} - \overline{X})^2 - 2E\left\{(\overline{y} - \overline{Y})(\overline{x} - \overline{X})\right\} =$$

$$= \frac{\sigma_y^2}{n_2} + \frac{\sigma_x^2}{n_1} - \frac{2n'\sigma_{xy}}{n_1 \ n_2}$$

ya que existe covarianza solamente para las unidades comunes. Esta fórmula también puede escribirse:

$$V(\bar{d}) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - \frac{2\pi\sigma_{xy}}{n_2} = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - \frac{2\pi\rho\sigma_x\sigma_y}{n_2}$$

GANANCIA EN PRECISION CON EN SOLAPAMIENTO

Aplicando este resultado a las tres situaciones descritas al principio, resulta:

Solapamiento completo: $\pi = 1$; $V_{(A)} = V_{(C)} - \frac{2\rho\sigma_x\sigma_y}{n_2}$

Solapamiento parcial: $0 < \pi < 1$; $V_{(B)} = V_{(C)} - \frac{2\pi\rho\sigma_x\sigma_y}{n_2}$

Muestras independientes: $V_{(C)} = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2}$

de donde se deduce que si el coeficiente de correlación, ρ, entre los valores de una y otra ocasión es positivo:

$$V_{(A)} < V_{(B)} < V_{(C)}$$

y por tanto para estimar el cambio la precisión es tanto mayor cuanto mayor sea la proporción de unidades que permanece fija en la muestra, y obviamente cuanto mayor sea la correlación.

ESTIMADOR LINEAL DE MINIMA VARIANZA

En el siguiente esquema gráfico aparecen las cuatro posibles estimaciones de la media, en una y otra ocasión, en la parte común y no común

de las muestras.

	$\bar{y}'$	$\bar{y}''$	Segunda ocasión
Primera ocasión	$\bar{x}''$	$\bar{x}'$	

Estamos ahora interesados en determinar el estimador lineal insesgado de mínima varianza para el cambio $\bar{D} = \bar{Y} - \bar{X}$, que puede formarse utilizando como información las cuatro medias muestrales $(\bar{x}'; \bar{x}''; \bar{y}'; \bar{y}'')$. Para ello vamos a simplificar un poco la situación admitiendo que el tamaño de la muestra y la varianza poblacional no varian de una ocasión a otra; así pues: $n_1 = n_2 = n$; $\sigma_x^2 = \sigma_y^2 = \sigma^2$. La primera hipótesis no supone gran pérdida de generalidad ya que en la práctica es usual tomar muestras de igual tamaño en ocasiones sucesivas; en cuanto a la segunda, la varianza poblacional es bastante estable si no se producen cambios estructurales en la población entre las dos ocasiones.

Puesto que $(\bar{y}' - \bar{x}')$ y $(\bar{y}'' - \bar{x}'')$ son dos estimaciones insesgadas e independientes, del cambio $\bar{Y} - \bar{X}$, parece que una solución razonable consiste en formar una combinación lineal:

$$\bar{d} = \alpha\,(\bar{y}' - \bar{x}') + \beta\,(\bar{y}'' - \bar{x}'')$$

donde exigimos $\alpha + \beta = 1$ para que $\bar{d}$ sea insesgado. El problema se reduce a determinar el valor óptimo de α, para el cual $\bar{d}$ tiene mínima varianza.

Para simplificar las notaciones escribimos $\bar{d}' = \bar{y}' - \bar{x}'$; $\bar{d}'' = \bar{y}'' - \bar{x}''$, de esta forma es $\bar{d} = \alpha\,\bar{d}' + (1 - \alpha)\,\bar{d}''$. Ahora teniendo en cuenta las hipótesis y las fórmulas de la varianza de diferencia de medias obtenidas anteriormente, tenemos:

$$V(\bar{d}') = \frac{2\sigma^2}{\pi n}\,(1 - \rho) \quad ; \quad V(\bar{d}'') = \frac{2\sigma^2}{(1 - \pi)\,n}$$

Por otra parte el mínimo de $V(\bar{d}) = \alpha^2 V(\bar{d}') + (1-\alpha)^2 V(\bar{d}'') = \alpha^2 (V(\bar{d}') + V(\bar{d}'')) - 2\alpha V(\bar{d}'')$ se obtiene inmediatamente derivando respecto de α e igualando a cero, así resulta:

$$\alpha^* = \frac{V(\bar{d}'')}{V(\bar{d}') + V(\bar{d}'')}$$

de donde, sustituyendo las varianzas se obtiene el óptimo de α :

$$\alpha^* = \frac{\pi}{1-\rho(1-\pi)} \quad ; \quad 1-\alpha^* = \frac{(1-\pi)(1-\rho)}{1-\rho(1-\pi)}$$

por último, sustituyendo estos valores en $V(\bar{d})$ obtenemos la mínima varianza:

$$V(\bar{d})^* = \frac{2\sigma^2}{n} \quad \frac{(1-\rho)}{1-\rho(1-\pi)}$$

En esta fórmula observamos que la varianza es tanto menor, cuanto mayor sea el porcentaje de solapamiento y el coeficiente de correlación entre las dos ocasiones; esto mismo ocurría con el simple estimador de la diferencia de medias cuya varianza, teniendo en cuenta las hipótesis de igual tamaño de muestra y varianza poblacional, es:

$$V(\bar{d}) = \frac{2\sigma^2}{n} (1-\pi\rho)$$

Si calculamos la diferencia entre $V(\bar{d})$ y $V(d)^*$, se comprueba que efectivamente el estimador de mínima varianza produce una ganancia en precisión:

$$V(\bar{d}) - V(\bar{d})^* = \frac{2\sigma^2}{n} \left(\frac{\pi(1-\pi)\rho^2}{1-\rho(1-\pi)} \right)$$

También se comprueba inmediatamente que para $\pi = 1$, solapamiento completo, coinciden los dos estimadores y por tanto sus varianzas; para $\pi = 0$, muestras independientes, también coinciden y la varianza sería $2\sigma^2/n$.

VII.2.2. ESTIMACION DE LA SEGUNDA OCASION

El estudio anterior se dedicó al análisis de diferentes alternativas para estimar el cambio $\overline{Y} - \overline{X}$, entre dos ocasiones; ahora estamos interesados en estimar la media de la segunda ocasión. Bajo los mismos supuestos, un estimador insesgado es la simple media $\overline{y}$, cuya varianza será:

$$V(\overline{y}) = \frac{\sigma_y^2}{n_2} = \frac{\sigma^2}{n}$$

Veamos cómo puede obtenerse el estimador lineal insesgado de mínima varianza, utilizando como información las cuatro medias $\overline{x}''$; $\overline{x}'$; $\overline{y}'$; $\overline{y}''$. La forma general será:

$$\hat{\overline{Y}} = a\overline{x}'' + b\overline{x}' + c\overline{y}' + d\overline{y}''$$

donde la condición $E(\hat{\overline{Y}}) = \overline{Y}$ nos da las relaciones entre los coeficientes de la combinación lineal:

$$a + b = 0 \quad ; \quad c + d = 1$$

así pues el estimador insesgado tiene la forma:

$$\hat{\overline{Y}} = a(\overline{x}'' - \overline{x}') + c\overline{y}' + (1-c)\overline{y}''$$

cuya varianza, teniendo en cuenta que todas las variables son independientes excepto $\overline{x}'$, $\overline{y}'$, se obtiene inmediatamente:

$$V(\hat{\overline{Y}}) = a^2 V(\overline{x}'') + a^2 V(\overline{x}') + c^2 V(\overline{y}') - 2ac\,\mathrm{Cov.}(\overline{x}', \overline{y}') + (1-c)^2 V(y'') =$$

$$= a^2 \left(\frac{\sigma^2}{n - \pi n} + \frac{\sigma^2}{\pi n} \right) + c^2 \frac{\sigma^2}{\pi n} - \frac{2ac\,\rho\,\sigma^2}{\pi\, n} + (1-c)^2 \frac{\sigma^2}{n - \pi n}$$

simplificando:

$$V(\hat{\bar{Y}}) = \frac{\sigma^2}{n} \left\{ \frac{a^2 + c^2}{\pi(1-\pi)} - \frac{2c}{1-\pi} - \frac{2ac\,\rho}{\pi} + \frac{1}{1-\pi} \right\}$$

Para buscar el mínimo de esta varianza podemos prescindir del factor común constante σ^2/n y el término $1/(1-\pi)$ que no depende de a,c. Así pues el sistema de derivadas respecto de a y de c igualadas a cero nos da:

$$\left.\begin{aligned} \frac{2a}{\pi(1-\pi)} - \frac{2\rho c}{\pi} &= 0 \\ \frac{2c}{\pi(1-\pi)} - \frac{2}{1-\pi} - \frac{2\rho a}{\pi} &= 0 \end{aligned}\right\}$$

de cuya resolución se obtienen los valores de los coeficientes que nos dan la combinación lineal de mínima varianza:

$$a^* = \frac{\rho\,\pi\,(1-\pi)}{1 - \rho^2 (1-\pi)^2} \quad ; \quad c^* = \frac{\pi}{1 - \rho^2\,(1-\pi)^2}$$

Sustituyendo estos valores óptimos de a y c en $V(\hat{\bar{Y}})$, resulta la varianza mínima:

$$V(\hat{\bar{Y}}) = \frac{\sigma^2}{n} \frac{1 - \rho^2\,(1-\pi)}{1 - \rho^2\,(1-\pi)^2}$$

Enunciamos sin entrar en la demostración, que es análoga, que si se determina el estimador de mínima varianza lineal en (x'', x', y', y'')

para la media $\overline{X}$ de la primera ocasión, la diferencia entre ambos estimadores de mínima varianza $(\hat{\overline{Y}} - \hat{\overline{X}})$, coincide con el estimador del cambio de mínima varianza estudiado en VII.2.1. Si se desea la estimación de la media de las dos ocasiones (1/2) $(\overline{X} + \overline{Y})$, el mismo procedimiento nos conduce a que la semisuma de los estimadores de mínima varianza de una y otra ocasión, es también insesgado de mínima varianza.

VII.3. MUESTREO EN DOS FASES

Sabemos que en muchos casos es conveniente el uso de información previa que nos permite mejorar la exactitud de las estimaciones que obtendríamos con la media o total de expansión simple. Hemos visto como p.e. la estratificación produce muestras más representativas, a las que pueden aplicarse estimadores más precisos; lo mismo ocurre, bajo ciertas condiciones, con los estimadores de razón, de regresión lineal y con el uso de probabilidades desiguales de selección. En estos y otros casos similares la teoría ordinaria estudiada hasta aquí, supone que es conocida la información previa necesaria para la formación de los citados estimadores. En la práctica puede que esto no sea cierto; nos planteamos entonces la posibilidad de seleccionar una muestra, relativamente grande, en la que a bajo coste pueden observarse una o varias características generales de las unidades que nos proporcionan la información que necesitamos. En una segunda fase seleccionamos una submuestra de la primera, en la que observamos la característica objeto de estimación. Esta técnica se conoce con el nombre de *muestreo en dos fases,* o *muestro doble.*

VII.3.1. CONSIDERACION DEL COSTE

Es claro que la conveniencia de esta técnica de muestreo depende de los costes; si la observación de la característica que nos interesa no tiene coste, o es muy bajo, sencillamente tomaríamos una muestra del tamaño necesario para la precisión deseada y con ella haríamos las estimaciones. Supongamos entonces que disponemos de una presupuesto total C; que el coste por unidad de la primera muestra de tamaño n', es c' y que el coste por unidad de la segunda muestra, de tamaño $n < n'$, es c.

Frecuentemente c' es mucho más pequeño que c, bien sea porque la primera muestra se utiliza para obtener unos pocos datos generales de las unidades (en campo o en oficina, si se dispone de un fichero o registro) o bien porque la observación de la característica objetivo implica un proceso de observación más costoso. En estas condiciones si tomamos una sola muestra, tendremos:

$$C = c\, n_o$$

y si hacemos muestreo en dos fases:

$$C = c'\, n' + c\, n$$

igualando los dos costes totales, se obtiene:

$$n_o = n + \frac{c'}{c}\, n'$$

lo que nos dice que con la técnica de dos fases la observación efectiva se hace en una muestra de tamaño n, menor que n_o el cual corresponde a muestreo aleatorio simple en una sola fase con el mismo coste total. La cuestión que se plantea ahora es decidir si compensa la disminución del tamaño efectivo de la muestra, con el incremento de información adquirido en la primera fase; para ello debe calcularse la varianza correspondiente a muestreo doble y compararla con la del muestreo en una sola fase (σ^2 / n_o en caso de estimación de la media). Es obvio que cuanto menor sea la relación c' / c más favorece al muestreo doble, pero no es el único parámetro a tener en cuenta. En muestreo con reposición veremos que la varianza de los estimadores toma la forma:

$$V = \frac{K_1}{n} + \frac{K_2}{n'}$$

válida para muestreo sin reposición cuando las fracciones de muestreo son pequeñas, esta varianza puede minimizarse para un coste total dado, buscando los óptimos tamaños n' y n por el método de los multiplicadores de Lagrange.

VII.3.2. MUESTREO DOBLE PARA ESTRAFICACION

La primera muestra de tamaño n' se utiliza para estratificar sus unidades, atendiendo a una o varias características que observamos, así como para estimar la proporción de unidades de la población pertenecientes al h° estrato. Sea $n'_1 , \ldots , n'_h , \ldots , n'_L$ el número de unidades muestrales de cada estrato y $\hat{W}_h = n'_h / n'$ la proporción. La segunda fase consiste en tomar una submuestra aleatoria de tamaño $n_h \leqslant n'_h$ en cada estrato independientemente.

El estimador usual de la media en muestreo estratificado es:

$$\hat{\bar{X}} = \sum_h W_h \bar{x}_h \qquad ; \qquad W_h = \frac{N_h}{N}$$

en muestreo doble los W_h se estiman por los $\hat{W}_h$ obtenidos de la primera muestra, y con la segunda estimamos las medias, tomando $\bar{x}_h = x_h/n_h$; de esta forma resulta el estimador:

$$\hat{\bar{X}} = \sum_h \hat{W}_h \bar{x}_h \qquad ; \qquad \hat{W}_h = \frac{n'_h}{n'}$$

Utilizaremos la notación $\underset{W'}{E}(T)$ para expresar la esperanza matemática de un estadístico T, condicionada al conjunto de muestras de primera fase en las cuales $n'_1, \ldots , n'_h, \ldots , n'$ son fijos, o lo que es lo mismo, para un n' dado, $\hat{W}_1 , \ldots , \hat{W}_h , \ldots , \hat{W}_L$ son fijos. Análogamente $\underset{W'}{V}(T)$ expresará la varianza condicionada.

El estimador $\hat{\bar{X}}$ es insesgado:

$$E(\hat{\bar{X}}) = E\left\{ \underset{W'}{E}\left(\sum_h \hat{W}_h \bar{x}_h \right) \right\} = E\left(\sum_h \hat{W}_h \underset{W'}{E}(\bar{x}_h) \right) =$$

$$= E\left(\sum_h \hat{W}_h \bar{X}_h \right) = \sum_h E(\hat{W}_h) \bar{X}_h = \sum_h W_h \bar{X}_h = \bar{X}$$

puesto que para el conjunto de muestras de primera fase en las que n'_h es fijo, se cumple $E_{W'}(\bar{x}_h) = \bar{X}_h$; por otra parte en el conjunto de muestras variando n'_h se cumple $E(\hat{W}_h) = W_h$, siendo n' prefijado.

Para calcular la varianza del estimador utilizamos la descomposición:

$$V(\hat{\bar{X}}) = V\left(E_{W'}(\hat{\bar{X}})\right) + E\left(V_{W'}(\hat{\bar{X}})\right)$$

Veamos el primer sumando:

$$V\left(E_{W'}(\hat{\bar{X}})\right) = V\left(\sum_h \hat{W}_h \bar{X}_h\right) =$$

$$= \sum_h \bar{X}_h^2 \, V(\hat{W}_h) + \sum_{h \neq j} \bar{X}_h \bar{X}_j \, \text{Cov.}(\hat{W}_h, \hat{W}_j)$$

la varianza de $\hat{W}_h$ y la Cov. $(\hat{W}_h, \hat{W}_j)$ en muestreo sin reposición, haciendo uso de la distribución hipergeométrica para L clases, vienen dadas por:

$$V(\hat{W}_h) = \frac{N - n'}{N - 1} \frac{W_h (1 - W_h)}{n'} \quad ; \quad \text{Cov.}(\hat{W}_h, \hat{W}_j) = -\frac{N - n'}{N - 1} \frac{W_h W_j}{n'}$$

sustituyendo, resulta:

$$V\left(E_{W'}(\hat{\bar{X}})\right) = g' \left\{ \sum_h \bar{X}_h^2 \frac{W_h (1 - W_h)}{n'} - \sum_{h \neq j} \bar{X}_h \bar{X}_j \frac{W_h W_j}{n'} \right\}$$

donde g' es el factor de finitud $g' = (N - n') / (N - 1)$. La anterior expresión puede simplificarse teniendo en cuenta:

$$\sum_h \bar{X}_h^2 W_h - \sum_h \bar{X}_h^2 W_h^2 - \sum_{h \neq j} \bar{X}_h \bar{X}_j W_h W_j =$$

$$= \sum_h \bar{X}_h^2 W_h - \left(\sum_h W_h \bar{X}_h \right)^2 = \sum_h W_h (\bar{X}_h - \bar{X})^2$$

de donde resulta:

$$V\left(\underset{W'}{E} (\hat{\bar{X}}) \right) = \frac{g'}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2$$

Calculemos ahora el segundo sumando:

$$E\left(\underset{W'}{V} (\hat{\bar{X}}) \right) = E\left(\underset{W'}{V} \left(\sum_h \hat{W}_h \bar{x}_h \right) \right) = E\left(\sum_h (1-f_h) \frac{S_h^2}{n_h} \cdot \hat{W}_h^2 \right) =$$

$$= \sum_h (1-f_h) \frac{S_h^2}{n_h} E(\hat{W}_h^2) = \sum_h (1-f_h) \frac{S_h^2}{n_h} \left(V(\hat{W}_h) + W_h^2 \right) =$$

$$= \sum_h (1-f_h) \frac{S_h^2}{n_h} \left(\frac{g' W_h (1 - W_h)}{n'} + W_h^2 \right)$$

Por último agregando los dos sumandos la varianza del estimador de la media en muestreo doble para estratificación, es:

$$V(\hat{\bar{X}}) = \sum_h (1-f_h) \frac{S_h^2}{n_h} \left(W_h^2 + \frac{g' W_h (1-W_h)}{n'} \right) + \frac{g'}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2$$

donde f_h es la fracción de muestreo en la segunda fase, supuesto que el

muestreo se realiza con probabilidades iguales y sin reposición en las dos fases. Si el muestreo es con reposición en primera fase tendremos:

$$V(\hat{\bar{X}}) = \sum_h (1-f_h) \frac{S_h^2}{n_h} \left(W_h^2 + \frac{W_h (1-W_h)}{n'} \right) +$$

$$+ \frac{1}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2$$

fórmula aproximada para n' pequeño respecto de N en caso sin reposición. Si el muestreo es con reposición en las dos fases:

$$V(\hat{\bar{X}}) = \sum_h \frac{\sigma_h^2}{n_h} \left(W_h^2 + \frac{W_h (1-W_h)}{n'} \right) +$$

$$+ \frac{1}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2$$

fórmula aproximada para n_h pequeño respecto de N_h, en todo h, y n' pequeño respecto de N en caso sin reposición.

Para el total $X = N \bar{X}$, el estimador insesgado es $\hat{X} = N \hat{\bar{X}}$ y su varianza $V(\hat{X}) = N^2 V(\hat{\bar{X}})$.

Observemos por último que si la muestra de primera fase es de tamaño $n' = N$, esto es, se observan todas las unidades de la población para efectuar la estratificación, la fórmula general de la varianza del estimador en muestreo doble se convierte en:

$$V(\hat{\bar{X}}) = \sum_h (1-f_h) W_h^2 \frac{S_h^2}{n_h} \qquad ; \qquad g' = 0$$

que coincide con la del muestreo estratificado usual en una sola fase.

También se observa que n' aparece dividiendo, en consecuencia cuanto mayor es n' ($n' < N'$) la pérdida de precisión por el uso de muestreo doble disminuye. Obviamente el coste aumenta, razón por la cual conviene estudiar los tamaños y la afijación óptimos en función del coste.

ESTIMACION DE UNA PROPORCION

Si se desea estimar una proporción P en la población, siendo P_h la correspondiente al h° estrato, el estimador insesgado en muestreo doble es:

$$\hat{P} = \sum_h \hat{W}_h p_h \quad ; \quad p_h = \text{proporción muestral en segunda fase.}$$

y su varianza, aplicando el resultado anterior:

$$V(\hat{P}) = \sum_h (1-f_h) \frac{P_h Q_h}{n_h} \left(W_h^2 + \frac{g' W_h (1-W_h)}{n'} \right) +$$

$$+ \frac{g'}{n'} \sum_h W_h (P_h - P)^2$$

con la aproximación $S_h^2 = \frac{N_h}{N_h - 1} P_h Q_h \simeq P_h Q_h$.

En muestreo con reposición en las dos fases, o sin reposición y tamaños muestrales pequeños respecto de los correspondientes poblacionales ($f_h \simeq 1$; $g' \simeq 1$), se tiene:

$$V(\hat{P}) = \sum_h \frac{P_h Q_h}{n_h} \left(W_h^2 + \frac{W_h (1-W_h)}{n'} \right) + \frac{1}{n'} \sum_h W_h (P_h - P)^2$$

Para el total de clase, $A = NP$, el estimador es $\hat{A} = N \hat{P}$ y su varianza $V(\hat{A}) = N^2 V(\hat{P})$.

AFIJACION PROPORCIONAL

Si en la muestra de segunda fase asignamos a cada estrato un tamaño muestral n_h proporcional al tamaño del estrato, deberá ser:

$$n_h = W_h\, n$$

resultando para la varianza del estimador la fórmula:

$$V(\hat{\bar{X}}) = \frac{1}{n} \sum_h (1 - f_h)\, S_h^2 \left(W_h + \frac{g'\,(1 - W_h)}{n'} \right) +$$

$$+ \frac{g'}{n'} \sum_h W_h\, (\bar{X}_h - \bar{X})^2$$

aunque en la práctica para efectuar la afijación a los estratos utilizaremos $n_h = \hat{W}_h\, n$.

En muestreo con reposición:

$$V(\hat{\bar{X}}) = \frac{1}{n} \sum_h \sigma_h^2\, W_h + \frac{1}{n\,n'} \sum_h \sigma_h^2 (1 - W_h) +$$

$$+ \frac{1}{n'} \sum_h W_h\, (\bar{X}_h - \bar{X})^2$$

y teniendo en cuenta que si n no es pequeño el segundo sumando es del orden de un n-ésimo de los otros dos, podemos escribir la fórmula sufientemente aproximada:

$$V(\hat{\bar{X}}) = \frac{1}{n} \sum_h W_h\, \sigma_h^2 + \frac{1}{n'} \sum_h W_h\, (\bar{X}_h - \bar{X})^2$$

Para el total X, la proporción P y el total de clase, las fórmulas se deducen inmediatamente de las anteriores.

AFIJACION Y TAMAÑOS DE MUESTRA OPTIMOS

Determinemos ahora la afijación óptima en segunda fase; esto es, el reparto del tamaño n a los L estratos que para un coste total dado hace mínima la varianza del estimador; para ello utilizaremos la fórmula correspondiente a muestreo con reposición en las dos fases, válida en la práctica para muestreo sin reposición cuando la población es grande respecto de los tamaños muestrales. La función de Lagrange es:

$$\phi = \sum_h \frac{\sigma_h^2}{n_h}\left(W_h^2 + \frac{W_h\,(1-W_h)}{n'}\right) + \frac{1}{n'}\sum_h W_h\,(\bar{X}_h-\bar{X})^2 +$$

$$+ \lambda\,(c'n' + cn - C)$$

y derivando respecto de n_h se obtiene el sistema de ecuaciones para $h = 1, 2, \ldots, L$; para un n' dado:

$$\frac{\partial\,\Phi}{\partial\,n_h} = \frac{-\sigma_h^2}{n_h^2}\left(W_h^2 + \frac{W_h\,(1-W_h)}{n'}\right) + \lambda c = 0\,;$$

puesto que C, n', c' y c son fijos y $n = \sum_h n_h$.

Despejando n_h , resulta:

$$n_h = \frac{\sigma_h}{\sqrt{\lambda c}}\sqrt{W_h^2 + W_h\,(1-W_h)/n'}$$

lo que significa que n_h es proporcional a:

$$\sigma_h\ \sqrt{W_h^2 + W_h\,(1-W_h)/n'}$$

y en la práctica $W_h\,(1-W_h)/n'$ es pequeño con relación a W_h^2 , con lo que una buena aproximación a la afijación óptima viene dada por la

fórmula de Neyman. Así pues, tenemos:

$$n_h = \frac{W_h \sigma_h}{\sum_h W_h \sigma_h} . n \simeq \frac{W_h S_h}{\sum_h W_h S_h} . n$$

y llevando este resultado a la fórmula de la varianza del estimador, manteniendo que el término $W_h (1 - W_h)/n'$ es despreciable, resulta:

$$\hat{V}(\bar{X}) = \frac{1}{n}\left(\sum_h W_h \sigma_h \right)^2 + \frac{1}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2$$

Ahora para determinar los tamaños óptimos n' y n correspondientes a un coste total dado, escribimos la función de Lagrange:

$$\phi = \frac{1}{n} A + \frac{1}{n'} B + \lambda (c'n' + cn - C)$$

donde A y B representan las cantidades:

$$A = \left(\sum_h W_h \sigma_h \right)^2 \quad ; \quad B = \sum_h W_h (\bar{X}_h - \bar{X})^2$$

derivando respecto de n y n' obtenemos el sistema de ecuaciones:

$$\left.\begin{aligned} \frac{\partial \phi}{\partial n} &= -\frac{A}{n^2} + \lambda c = 0 \\ \frac{\partial \phi}{\partial n'} &= -\frac{B}{n'^2} + \lambda c' = 0 \end{aligned}\right\}$$

de donde despejando λ e igualando y teniendo en cuenta la ecuación de condición de coste, resulta el sistema que nos permite despejar los valores óptimos de n y n':

$$\left.\begin{aligned} \frac{n'}{\sqrt{A\,c'}} &= \frac{n'}{\sqrt{B\,c}} \\ C &= c'n' + c\,n \end{aligned}\right\}$$

y para obtener la fórmula de la varianza de $\hat{\bar{X}}$ con esta solución óptima, podemos operar de la siguiente forma:

$$\frac{c\,n}{c\sqrt{A\,c'}} = \frac{c'n'}{c'\sqrt{B\,c}} = \frac{c\,n + c'\,n'}{\sqrt{c\,c'}\left(\sqrt{A\,c} + \sqrt{B\,c'}\right)}$$

de donde:

$$n = \frac{C\sqrt{A}}{\sqrt{c}\left(\sqrt{A\,c} + \sqrt{B\,c'}\right)};\; n' = \frac{C\sqrt{B}}{\sqrt{c'}\left(\sqrt{A\,c} + \sqrt{B\,c'}\right)}$$

y llevando este resultado a la fórmula de la varianza, después de simplificar, resulta:

$$V_{\text{ópt.}}(\hat{\bar{X}}) = \frac{\left(\sqrt{A\,c} + \sqrt{B\,c'}\right)^2}{C}$$

Es interesante estudiar los tamaños de muestra óptimos, n' y n, para un coste total dado, cuando se aplica afijación proporcional ya que en la práctica ésta es más fácil de aplicar puesto que solamente necesita la información proporcionada por la primera muestra ($n_h = \hat{W}_h\, n$) frente a la afijación óptima que requiere información acerca de las varianzas internas de cada estrato. Veamos entonces este problema: La varianza en este caso puede escribirse, como hemos visto antes, en la forma:

$$V(\hat{\bar{X}}) = \frac{1}{n} A' + \frac{1}{n'} B$$

donde B es la constante utilizada antes y

$$A' = \sum_h W_h \sigma_h^2$$

en consecuencia, aplicando la misma técnica de los multiplicadores de Lagrange, resultan los tamaños óptimos del sistema de ecuaciones:

$$\left.\begin{aligned} \frac{n}{\sqrt{A' c'}} &= \frac{n'}{\sqrt{B c}} \\ C &= c' n' + c n \end{aligned}\right\}$$

y la varianza óptima:

$$V(\hat{\bar{X}}) = \frac{\left(\sqrt{A' c} + \sqrt{B c'}\right)^2}{C}$$

Se observa que esta varianza es mayor (o igual en el caso σ_h = etc.) que la obtenida para afijación óptima, puesto que la única diferencia reside en que aparece A' en vez de A en el numerador. Es inmediato comprobar que:

$$A' = \sum_h W_h (\sigma_h - \bar{\sigma})^2 + A \quad ; \quad A' \geqslant A$$

siendo $\bar{\sigma}$ la media ponderada de las σ_h.

ESTIMACION DE LA VARIANZA

Un estimador insesgado de la varianza de $\hat{\bar{X}}$ en muestreo doble para estratificación, con reposición, que como hemos visto es:

$$V(\hat{\bar{X}}) = \sum_h \frac{\sigma^2}{n_h} \left(W_h^2 + \frac{W_h(1 - W_h)}{n'}\right) + \frac{1}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2 \qquad (1)$$

viene dada por la fórmula:

$$v(\hat{\bar{X}}) = \frac{n'}{n'-1}\left\{\sum_h \frac{s_h^2}{n_h}\left(\hat{W}_h^2 - \frac{\hat{W}_h}{n'}\right) + \sum_h \frac{\hat{W}_h(\bar{x}_h - \hat{\bar{X}})^2}{n'}\right\} \quad (2)$$

que como veremos más adelante admite algunas simplificaciones bajo ciertos supuestos. Veamos primero la esperanza matemática de cada término sumado en h:

$$\sum_h E\left(\frac{s_h^2}{n_h}\hat{W}_h^2\right) = \sum_h E\left(\hat{W}_h^2 \underset{W'}{E}\left(\frac{s_h^2}{n_h}\right)\right) = \sum_h \frac{\sigma_h^2}{n_h} E(\hat{W}_h^2) =$$

$$= \sum_h \frac{\sigma_h^2}{n_h}\left(W_h^2 + \frac{W_h(1-W_h)}{n'}\right) \quad (3)$$

$$\sum_h E\left(\frac{s_h^2}{n_h}\frac{\hat{W}_h}{n'}\right) = \sum_h E\left(\frac{\hat{W}_h}{n'} \underset{W'}{E}\left(\frac{s_h^2}{n_h}\right)\right) = \frac{1}{n'}\sum_h \frac{\sigma_h^2}{n_h} E(\hat{W}_h) =$$

$$= \frac{1}{n'}\sum_h W_h \frac{\sigma_h^2}{n_h} \quad (4)$$

$$E\left(\sum_h \frac{\hat{W}_h(\bar{x}_h - \hat{\bar{X}})^2}{n'}\right) = \frac{1}{n'}\left\{E\left(\sum_h \hat{W}_h \bar{x}_h^2\right) - E(\hat{\bar{X}}^2)\right\} =$$

y tomando esperanza en dos fases,

$$= \frac{1}{n'}\sum_h W_h\left(\underset{W'}{V}(\bar{x}_h) + \bar{X}_h^2\right) - \frac{1}{n'}\left(V(\hat{\bar{X}}) + \bar{X}^2\right) =$$

$$= \frac{1}{n'} \sum_h W_h \frac{\sigma_h^2}{n_h} + \frac{1}{n'} \sum_h W_h (\bar{X}_h - \bar{X})^2 - \frac{1}{n'} V(\hat{\bar{X}}) \qquad (5)$$

La suma (3) − (4) + (5) nos da la fórmula (1) menos $\frac{1}{n'} V(\hat{\bar{X}})$:

$$V(\hat{\bar{X}}) - \frac{1}{n'} V(\hat{\bar{X}}) = \frac{n'-1}{n'} V(\hat{\bar{X}})$$

en consecuencia:

$$v(\hat{\bar{X}}) = \frac{n'}{n'-1} \{(3) - (4) + (5)\} \text{ es insesgado para } V(\hat{\bar{X}}).$$

El factor $n' / (n' - 1)$ prácticamente es próximo a la unidad si n' no es pequeño, también el término que aparece en segundo lugar en la fórmula (2) puede ser despreciable respecto de los otros dos ya que aparece el producto $n_h \cdot n'$ en el denominador, resultando la aproximación:

$$v(\hat{\bar{X}}) \simeq \sum_h \hat{W}_h^2 \frac{s_h^2}{n_h} + \frac{1}{n'} \sum \hat{W}_h (\bar{x}_h - \hat{\bar{X}})^2 \qquad (6)$$

y, por último, también en esta expresión el segundo sumando será pequeño respecto del primero para valores grandes de n', resultando como fórmula aproximada más sencilla:

$$v(\hat{\bar{X}}) \simeq \sum_h \hat{W}_h^2 \frac{s_h^2}{n_h} \qquad (7)$$

que es la correspondiente a muestreo estratificado en una sola fase, sustituyendo W_h por su estimación.

En caso de estimar la varianza de $\hat{P}$, sustituímos en la fórmula (2),

o en sus aproximaciones cuando sean válidas.

$$\frac{s_h^2}{n_h} = \frac{p_h q_h}{n_h - 1} \quad ; \quad (\bar{x}_h - \hat{\bar{X}})^2 = (p_h - \hat{P})^2$$

VIII.3.3. MUESTREO DOBLE PARA ESTIMADORES DE RAZON

El estimador usual de razón para la media $\bar{X}$, utiliza como información conocida previamente la media $\bar{Y}$ (o el total) de una característica $\mathfrak{Y}$, definida en todas las unidades de la población, elegida convenientemente de modo que su relación con $\mathfrak{X}$ sea lineal al menos aproximadamente. El muestreo doble utiliza la primera muestra de tamaño n' para obtener una buena estimación de $\bar{Y}$, ó de Y, y la segunda muestra de tamaño n para estimar $\bar{x}$ e $\bar{y}$, de esta forma el estimador de razón para la media en muestreo doble es:

$$\hat{\bar{X}}_R = \frac{\bar{x}}{\bar{y}} \cdot \bar{y}' \quad ; \quad \bar{y}' = \text{Media de la primera muestra.}$$

Con este procedimiento de muestreo doble cabe considerar dos posibilidades: (I) La segunda es una muestra aleatoria de la población, seleccionada independientemente de la primera; (II) La segunda muestra es una submuestra aleatoria de la primera; en ambos casos suponemos $n \leqslant n'$.

En cualquier caso la esperanza matemática del estimador, es:

$$E(\hat{\bar{X}}_R) = \bar{Y} E(\hat{R})$$

y será insesgado si lo es $\hat{R} = \bar{x}/\bar{y}$. Para calcular su ECM, que coincidirá con la varianza cuando se cumple $E(\hat{R}) = R = \bar{X}/\bar{Y}$, operamos de la siguiente forma:

$$(\hat{\bar{X}}_R - \bar{X}) = \left(\left(\frac{\bar{x}}{\bar{y}} \bar{y}' - \bar{X} \right) \right) = (\hat{R} \bar{y}' - R \bar{Y}) =$$

$$= (\hat{R}\bar{y}' - R\bar{Y} + \hat{R}\bar{Y} - \hat{R}\bar{Y}) = \bar{Y}(\hat{R} - R) + \hat{R}(\bar{y}' - \bar{Y}) =$$

$$= \frac{\bar{Y}}{\bar{y}} (\bar{x} - \hat{R}\bar{y}) + \hat{R}(\bar{y}' - Y)$$

Utilizando ahora las aproximaciones:

$$\hat{R} \simeq R \quad ; \quad \frac{\bar{Y}}{\bar{y}} \simeq 1$$

podemos escribir para el cálculo aproximado de la varianza del estimador:

$$V(\hat{\bar{X}}_R) = E\left\{(\bar{x} - R\bar{y}) + R(\bar{y}' - \bar{Y})\right\}^2 = V\left\{(\bar{x} - R\bar{y}) + R(\bar{y}' - \bar{Y})\right\} =$$

$$= V(\bar{x} - R\bar{y}) + V(R(\bar{y}' - \bar{Y})) + 2R\,\mathrm{Cov.}\left\{(\bar{x} - R\bar{y}), (\bar{y}' - \bar{Y})\right\} =$$

$$= V(\bar{x}) + R^2 V(\bar{y}) - 2R\,\mathrm{Cov.}(\bar{x}, \bar{y}) + R^2 V(\bar{y}') +$$

$$+ 2R\,\mathrm{Cov}(\bar{x}, \bar{y}') - 2R^2\,\mathrm{Cov}(\bar{y}, \bar{y}')$$

En el caso (I) de que las muestras de las dos fases son independientes, se anulan las covarianzas entre $(\bar{x}, \bar{y}')$ y entre $(\bar{y}, \bar{y}')$, resultando:

$$V(\hat{\bar{X}}_R) = V(\bar{x}) + R^2 V(\bar{y}) - 2R\,\mathrm{Cov.}(\bar{x}, \bar{y}) + R^2 V(\bar{y}');$$

$$V(\hat{\bar{X}}_R) = \frac{1}{n}\left\{\sigma_x^2 + R^2\sigma_y^2 - 2R\sigma_{xy}\right\} + \frac{1}{n'} R^2 \sigma_y^2 \qquad (1)$$

fórmula válida para muestreo con reposición; en caso sin reposición susti-

tuimos varianzas y covarianzas por cuasi-varianzas y cuasi-covarianzas, multiplicando el primer sumando por el factor de finitud en segunda fase y el segundo sumando por el de primera fase.

Para el caso (II) en que la segunda muestra de tamaño n es una submuestra aleatoria de la primera ($n \leqslant n'$), debemos de calcular previamente las covarianzas entre $(\bar{x}, \bar{y}')$ y entre $(\bar{y}, \bar{y}')$:

Teniendo en cuenta que fijada una muestra W' de primera fase, se cumple:

$$\underset{W'}{E}(\bar{x}) = \bar{x}' \quad ; \quad \underset{W'}{E}(\bar{y}) = \bar{y}'$$

por ser $\bar{x}$ e $\bar{y}$ nedias de una submuestra aleatoria $W \subset W'$, resulta:

$$\text{Cov.}(\bar{x}, \bar{y}') = E(\bar{x}\,\bar{y}') - E(\bar{x})\,E(\bar{y}') =$$

$$= E\left(\underset{W'}{E}(\bar{x}, \bar{y}')\right) - E\left(\underset{W'}{E}(\bar{x})\right) E\left(\underset{W'}{E}(\bar{y}')\right) =$$

$$= E(\bar{x}'\bar{y}') - E(\bar{x}')\,E(\bar{y}') = \text{Cov}(\bar{x}', \bar{y}') = \frac{\sigma_{xy}}{n'} .$$

Análogamente:

$$\text{Cov}(\bar{y}, \bar{y}') = \text{Cov}(\bar{y}', \bar{y}') = \frac{\sigma_y^2}{n'}$$

Sustituyendo estos dos resultados en la fórmula general, obtenemos:

$$V(\hat{\bar{X}}_R) = \frac{1}{n}\left\{\sigma_x^2 + R^2\,\sigma_y^2 - 2\,R\,\sigma_{xy}\right\} + \frac{1}{n'}\,R^2\,\sigma_y^2$$

$$- \frac{1}{n'} \cdot 2\,R^2\,\sigma_y^2 + \frac{1}{n'} \cdot 2\,R\,\sigma_{xy} .$$

de donde resulta:

$$V(\hat{\bar{X}}_R) = \frac{1}{n}\left\{ \sigma_x^2 + R^2 \sigma_y^2 - 2R\sigma_{xy} \right\} + \frac{1}{n'}\left\{ 2R\sigma_{xy} - R^2 \sigma_y^2 \right\} \quad (2)$$

fórmula válida para muestreo con reposición; en el caso sin reposición pondríamos donde corresponda las cuasivarianzas y cuasicovarianzas, así como los factores de finitud de segunda y primera fase.

Es interesante observar en la fórmula general de $V(\hat{\bar{X}}_R)$ para muestreo doble que si la primera muestra es de tamaño $n' = N$, esto es, se observan todas las unidades de la población para calcular $\bar{Y}$, resulta ser:

$$V(\bar{y}') = \mathrm{Cov}(\bar{x}, \bar{y}') = \mathrm{Cov}(y, \bar{y}') = 0$$

por tanto se reduce a la varianza del estimador usual de razón en una sola fase.

Para estimar el total en muestreo doble, tendremos:

$$\hat{X}_R = N\hat{\bar{X}}_R \quad ; \quad V(\hat{X}_R) = N^2 V(\hat{\bar{X}}_R)$$

TAMAÑOS OPTIMOS DE MUESTRA

El problema de la determinación de los tamaños óptimos n' y n, para un coste total dado por la función $C = c'n' + cn$, puede resolverse inmediatamente utilizando la misma técnica que en muestreo doble para estratificación. Observando las fórmulas (1) y (2) de la varianza vemos que son de la forma:

$$(I) \qquad V(\hat{\bar{X}}_R) = \frac{1}{n} a + \frac{1}{n'} b$$

con $$a = (\sigma_x^2 + R^2 \sigma_y^2 - 2R\sigma_{xy}) \; ; \quad b = R^2 \sigma_y^2$$

por lo tanto, los valores óptimos de n y n' se obtienen del sistema de ecuaciones:

$$\left.\begin{array}{l} \dfrac{n'}{n} = \dfrac{\sqrt{b\,c}}{\sqrt{a\,c'}} \\ \\ C = c'\,n' + c\,n \end{array}\right\}$$

y la varianza óptima:

$$V\left(\bar{\bar{X}}_R\right)_{opt.} = \frac{\left(\sqrt{a\,c} + \sqrt{b\,c'}\right)^2}{C}$$

(II) En el caso de que la segunda sea una submuestra de la primera:

$$V\left(\bar{\bar{X}}_R\right) = \frac{1}{n}\,a + \frac{1}{n'}\,b'$$

con $b' = 2\,R\,\sigma_{xy} - R^2\,\sigma_y^2$; por lo tanto:

$$\left.\begin{array}{l} \dfrac{n'}{n} = \dfrac{\sqrt{b'\,c}}{\sqrt{a\,c'}} \\ \\ C = c'\,n' + c\,n \end{array}\right\} \qquad V\left(\bar{\bar{X}}_R\right)_{opt.} = \frac{\left(\sqrt{a\,c} + \sqrt{b'\,c'}\right)^2}{C}$$

ESTIMACION DE LA VARIANZA

Dado que en la segunda muestra de tamaño n, obtenemos observaciones de la variable conjunta $(\mathfrak{X}, \mathfrak{Y})$ podemos calcular estimaciones de σ_x^2 y $\mathrm{Cov}(\mathfrak{X}, \mathfrak{Y})$:

$$s_x^2 = \frac{1}{n-1}\sum_1^n (X_i - \bar{x})^2 \quad ; \quad s_{xy} = \frac{1}{n-1}\sum_1^n (X_i - \bar{x})(Y_i - \bar{y})$$

y puesto que la primera muestra es de tamaño $n' > n$ nos permite una buena estimación de σ_y^2 :

$$s_y^2 = \frac{1}{n'-1} \sum_{1}^{n'} (Y_i - \bar{y}')^2$$

Para la razón R, pondremos la estimación $\hat{R}$.

VII.3.4 MUESTREO DOBLE PARA ESTIMADORES DE REGRESION LINEAL

El estimador usual para muestreo en una fase por regresión lineal viene dado por:

$$\hat{\bar{X}} = \bar{x} + K(\bar{Y} - \bar{y})$$

donde K es una constante prefijada e $\bar{Y}$ es la media poblacional de la variable auxiliar; $\bar{x}$, $\bar{y}$ se obtienen de las observaciones de una muestra (X_i, Y_i) de tamaño n. En muestreo doble, al suponer desconocida $\bar{Y}$, utilizamos la primera muestra de tamaño n' para estimar $\bar{Y}$, estimación dada por $\bar{y}'$; así con la muestra de tamaño n en segunda fase estimamos $\bar{x}, \bar{y}$ formando el estimador para la media poblacional:

$$\hat{\bar{X}}_{rl} = \bar{x} + K(\bar{y}' - \bar{y})$$

En esta situación caben dos posibilidades de actuación: (I) La segunda muestra es independiente de la primera. (II) La segunda es una submuestra aleatoria $n < n'$, de la primera.

El estimador $\hat{\bar{X}}_{rl}$ tiene como valor esperado:

$$E(\hat{\bar{X}}_{rl}) = E(\bar{x}) + K E(\bar{y}' - \bar{y})$$

y es insesgado en los dos casos. En el primero

$E(\bar{y}) = \bar{Y}$; $E(\bar{y}') = \bar{Y}$, por lo tanto:

$$E(\hat{\bar{X}}_{rl}) = E(\bar{x}) + K(\bar{Y} - \bar{Y}) = \bar{X}$$

y en el segundo $E_{W'}(\bar{y}) = \bar{y}'$, por tanto también se anula $E(\bar{y}' - \bar{y})$.

La fórmula general de su varianza se obtiene inmediatamente:

$$V(\hat{\bar{X}}_{rl}) = V(\bar{x} + K(\bar{y}' - \bar{y})) =$$

$$= V(\bar{x}) + K^2 V(\bar{y}' - \bar{y}) + 2K \operatorname{Cov.}(\bar{x}, (\bar{y}' - \bar{y}))$$

que desarrollada en todos sus términos, resulta:

$$V(\hat{\bar{X}}_{rl}) = V(\bar{x}) + K^2 V(\bar{y}') + K^2 V(\bar{y}) - 2K^2 \operatorname{Cov}(\bar{y}', \bar{y}) +$$

$$+ 2K \operatorname{Cov}(\bar{x}, (\bar{y}' - \bar{y})) \qquad (1)$$

En el caso I de muestras independientes, se cumple:

$$\operatorname{Cov}(\bar{y}', \bar{y}) = 0 \quad ; \quad \operatorname{Cov}(\bar{x}, (\bar{y}' - \bar{y})) = - \operatorname{Cov}(\bar{x}, \bar{y})$$

y resulta:

$$V(\hat{\bar{X}}_{rl}) = \frac{1}{n}(\sigma_x^2 + K^2 \sigma_y^2 - 2K\sigma_{xy}) + \frac{K^2 \sigma_y^2}{n'} . \qquad (2)$$

En el caso de que la segunda es una submuestra de la primera, se cumple:

$$\operatorname{Cov}(\bar{y}', \bar{y}) = \frac{\sigma_y^2}{n'} \; ; \; \operatorname{Cov}(\bar{x}, (\bar{y}' - \bar{y})) = \operatorname{Cov}(\bar{x}, \bar{y}') - \operatorname{Cov}(\bar{x}, \bar{y}),$$

y resulta, sustituyendo en la fórmula general y simplificando:

$$V(\hat{\bar{X}}_{rl}) = \frac{1}{n}(\sigma_x^2 + K^2 \sigma_y^2 - 2K\sigma_{xy}) + \frac{1}{n'}(2K\sigma_{xy} - K^2 \sigma_y^2) \qquad (3)$$

Observemos que la varianza del estimador de regresión lineal en muestreo doble puede expresarse de una forma cómoda de manejar; siendo:

$$A = \sigma_x^2 + K^2\,\sigma_y^2 - 2\,K\,\sigma_{xy} \quad ; \quad B = K^2\,\sigma_y^2 \quad ; \quad B' = 2\,K\,\sigma_{xy} - K^2\,\sigma_y^2$$

podemos escribir, para el caso de muestras independientes:

$$V\left(\hat{\bar{X}}_{rl}\right) = \frac{A}{n} + \frac{B}{n'} \tag{4}$$

y para el caso de que la segunda es una submuestra de la primera:

$$V\left(\hat{\bar{X}}_{rl}\right) = \frac{A}{n} + \frac{B'}{n'} \tag{5}$$

TAMAÑOS DE MUESTRA OPTIMOS

Si buscamos la varianza mínima para un coste total dado, podemos utilizar la función de Lagrange:

$$\phi = V\left(\hat{\bar{X}}_{rl}\right) + \lambda\,(C - c'\,n' - c\,n)$$

cuyo sistema de derivadas parciales, respecto de n y n' igualadas a cero, junto con la condición de coste, nos dará las soluciones buscadas. El procedimiento operativo es análogo al que hemos utilizado en muestreo doble para estratificación, por lo tanto en el caso I (muestras independientes) resulta:

$$n = \frac{C\sqrt{A}}{\sqrt{c}\,(\sqrt{A\,c} + \sqrt{B\,c'})} \quad ; \quad n' = \frac{C\sqrt{B}}{\sqrt{c'}\,(\sqrt{A\,c} + \sqrt{B\,c'})}$$

y la varianza, para estos tamaños óptimos:

$$V\left(\hat{\bar{X}}_{rl}\right) = \frac{(\sqrt{A\,c} + \sqrt{B\,c'})^2}{C}$$

Para el caso II (la segunda es una submuestra de la primera), resulta:

$$n = \frac{C\sqrt{A}}{\sqrt{c}\,(\sqrt{A\,c} + \sqrt{B'\,c'})} \; ; \; n' = \frac{C\sqrt{B'}}{\sqrt{c'}\,(\sqrt{A\,c} + \sqrt{B'\,c'})}$$

y la varianza para estos tamaños óptimos:

$$V(\hat{\bar{X}}_{rl}) = \frac{(\sqrt{A\,c} + \sqrt{B'\,c'})^2}{C}$$

ESTIMADOR OPTIMO

En el estudio realizado hasta aquí, hemos hecho caso omiso acerca de la constante K que aparece en el estimador:

$$\hat{\bar{X}}_{rl} = \bar{x} + K(\bar{y}' - \bar{y})$$

En muestreo de una sola fase, la varianza del estimador prescindiría del segundo término

$$\frac{1}{n'} B \text{ ó } \frac{1}{n'} B'$$

en cuyo caso ambas fórmulas coincidirían por ser iguales a la que corresponde al estimador usual de regresión lineal. En este caso se comprueba inmediatamente que el valor mínimo de:

$$V(\hat{\bar{X}}_{rl}) = \frac{1}{n}(\sigma_x^2 + K^2\,\sigma_y^2 - 2\,K\,\sigma_{xy})$$

se alcanza cuando se asigna a K el valor del coeficiente de regresión de $\mathfrak{X}$ sobre $\mathfrak{Y}$; derivando respecto de K e igualando a cero se obtiene:

$$K = b = \frac{\sigma_{xy}}{\sigma_y^2}$$

En el muestreo bifásico, esta solución óptima para K es aproximada

pero puede ser una buena aproximación en la práctica. Llevando este valor óptimo $K = b = \sigma_{xy} / \sigma_y^2$ a la fórmula de las varianzas, se obtiene en ambos casos I y II:

$$V(\hat{\bar{X}}_{rl}) = \frac{(1-\rho^2)\,\sigma_x^2}{n} + \frac{\rho^2\,\sigma_x^2}{n'}$$

de donde, con las notaciones:

$$A = (1-\rho^2)\,\sigma_x^2 \quad ; \quad B = \rho^2\,\sigma_x^2$$

pueden obtenerse los valores óptimos de n y n' así como la varianza óptima mínima, para un coste total dado, estudiados en el anterior apartado de este capítulo.

Mencionemos que b no será conocido a priori, por ello, en la práctica aplicaremos a K el valor: estimación de b mediante la segunda muestra:

$$K = \hat{b} = \frac{\sum_1^n (X_i - \bar{x})\,(Y_i - \bar{y})}{\sqrt{\sum_1^n (X_i - \bar{x})^2\,(Y_i - \bar{y})^2}}$$

ESTIMADOR POR DIFERENCIA

El estimador por diferencia en muestreo doble resulta del estimador de regresión, por lo que toda la teoría anterior es válida haciendo $K = 1$, resultando el estimador:

$$\hat{\bar{X}}_d = \bar{x} + (\bar{y}' - \bar{y})$$

Análogamente las fórmulas de las varianzas y tamaños óptimos de muestra se obtienen aplicando a K el valor $K = 1$.

VII.3.5. MUESTREO DOBLE PARA PROBABILIDADES PROPORCIONALES AL TAMAÑO

El estimador usual del total X, con probabilidades de selección de las unidades, proporcionales a una medida de su tamaño, sea M_i,

viene dado por:

$$\hat{X} = \sum_{1}^{n} \frac{X_i}{n P_i} \quad ; \quad P_i = \frac{M_i}{M}$$

Si no se conocen a priori los tamaños de las unidades de la población, podemos tomar una muestra aleatoria de tamaño n', con probabilidades iguales, para obtener información acerca de los tamaños $M_1, \ldots, M_i, \ldots, M_{n'}$, siendo $M' = \sum_{i=1}^{n'} M_i$. En estas condiciones, se toma una submuestra de tamaño $n < n'$, para formar el estimador bifásico del total, basándonos en que:

$$\frac{M_i}{\frac{N}{n'} M'} \quad \text{es un estimador de} \quad \frac{M_i}{M} = P_i$$

El estimador es por lo tanto:

$$\hat{X} = \sum_{1}^{n} \frac{N M' X_i}{n' n M_i} = \frac{N M'}{n n'} \sum_{1}^{n} \frac{X_i}{M_i}$$

Es insesgado:

$$E(\hat{X}) = E\left(\frac{N}{n'} \underset{W'}{E}\left(\sum_{1}^{n} \frac{M' X_i}{n M_i}\right)\right) = E\left(\frac{N}{n'} x'\right) = X$$

donde $\underset{W'}{E}$ indica esperanza para la primera muestra fija, con probabilidades proporcionales al tamaño; x' es el total de la muestra de primera fase, tamaño n', tomada con probabilidades iguales.

Supuesto que la primera muestra se extrae con probabilidades

iguales y sin reposición, y la segunda, submuestra de la primera, con probabilidades proporcionales al tamaño y con reposición, la varianza del estimador viene dada por:

$$V(\hat{X}) = \frac{N}{N-1}\,\frac{n'-1}{n\,n'}\,\sum P_i\left(\frac{X_i}{P_i} - X\right)^2 + \frac{N(N-n')}{n'}\,S^2$$

y si n' es grande $1/n\,n'$ puede despreciarse, resultando la fórmula aproximada:

$$\frac{1}{n}\,\sum_{1}^{N} P_i\left(\frac{X_i}{P_i} - X\right)^2 + \frac{N(N-n')}{n'}\,S^2$$

Estos resultados y otros desarrollos particulares pueden consultarse en Des Raj.

VII.4. EJEMPLOS

Ejemplo VII.1

De una población con N = 250 unidades, se han tomado tres muestras sistemáticas aleatorias de período K = 10. Los totales muestrales son 10, 12, 16. Estime la media poblacional y el error de muestreo del estimador.

Solución:

$\hat{\bar{X}} = 0{,}52$; $\hat{v}(\hat{\bar{X}}) = 0{,}0048$; error de muestreo 13,3%

Ejemplo VII.2

Compruebe que para el estimador lineal de mínima varianza, de la media de la segunda ocasión, el valor óptimo de la proporción de solapamiento π, viene dado por:

$$1 - \pi^* = \frac{1 - \sqrt{1 - \rho^2}}{\rho^2}$$

Ejemplo VII.3

Con los datos:

Primera ocasión	Segunda ocasión	
$\bar{x}'' = 150$	$\bar{y}'' = 160$	$n = 60$; $\sigma^2 = 20$
$\bar{x}' = 152$	$\bar{y}' = 158$	$\pi = 0{,}6$ $\rho = 0{,}7$

Determine las siguientes estimaciones y su error de muestreo:

(A) Del cambio $(\bar{y} - \bar{x})$
(B) Del cambio de mínima varianza
(C) De la media de la segunda ocasión $(\bar{y})$
(D) De la media de la segunda ocasión de mínima varianza

Solución:

(A) $\bar{d} = (\bar{y} - \bar{x}) = 7{,}6$; $\hat{v}(\bar{d}) = 0{,}386\ldots$; error de muestreo 8,2 %

(B) $\bar{d}^{*} = 6{,}668$; $\hat{v}(\bar{d}^{*}) = 0{,}27\ldots$; error de muestreo 7,9 %

(C) $\bar{y} = 158{,}8$; $\hat{v}(\bar{y}) = 0{,}3\ldots$; error de muestreo 0,36 %

(D) $\hat{\bar{Y}}^{*} = 159{,}062$; $\hat{v}(\hat{\bar{Y}}^{*}) = 0{,}2908$; error de muestreo 0,34%

Ejemplo VII.4

Supongamos que se conocen los siguientes datos de una población de unidades clasificadas en L = 3 estratos:

W_h	σ_h	$\bar{X}_h$
0,6	4	3
0,3	25	6
0,1	64	12

Para estimar la media, se pide decidir si es más conveniente el muestreo doble con $c' = 36$; $c = 225$ que el muestreo en una sola fase con el mismo coste total, en los casos:

(A) Afijación proporción en la segunda fase.

(B) Afijación óptima en la segunda fase.

Solución:

Teniendo en cuenta que para el mismo coste total el tamaño de muestra utilizable en muestreo aleatorio simple es $n_o = C / c$, la varianza del estimador de la media en este caso, es:

$$V(\bar{x}) = \frac{\sigma^2}{n_o} = \frac{c}{C}\,\sigma^2 = \frac{c}{C}\left\{\sum W_h\,(\bar{X}_h - \bar{X})^2 + \sum W_h\,\sigma_h^2\right\} =$$

$$= \frac{c}{C}\left\{B + \sum W_h\,\sigma_h^2\right\} = \frac{c}{C}\left\{B + A'\right\} \qquad (i)$$

por otra parte la varianza del estimador en muestreo doble con afijación proporcional y n' , n óptimos, es:

$$\hat{V}(\bar{X})_{prop.} = \frac{1}{C}\left\{\sqrt{A' c} + \sqrt{B c'}\right\}^2 \qquad \text{(ii)}$$

por tanto en el caso (A) del problema, el muestreo doble es mejor si se cumple (ii) < (i), o bien:

$$2\sqrt{A' B}\ \sqrt{\gamma} < B(1-\gamma)$$

siendo $\gamma = c'/c$.

En el caso (B) del problema, la varianza en muestreo doble, operando de dorma análoga es menor que (i) si se cumple:

$$A + 2\sqrt{A B}\ \sqrt{\gamma} < B(1-\gamma) + \Sigma\ W_h \sigma_h^2$$

Efectuando los cálculos necesarios con los datos dados:

$$A = 265{,}69\ ; \qquad B = 7{,}56\ ; \qquad A' = \Sigma\ W_h \sigma_h^2 = 606{,}7\ ; \qquad \gamma = 0{,}16$$

se obtiene que en el caso (A) no es 54,18 menor que 7,56 . 0,84 por tanto es menos conveniente el muestreo doble; en el caso (B) se cumple 301,54 menor que 613,05 por tanto es mejor el muestre doble.

Ejemplo VII.5

Determine el valor de la relación de costes c'/c, a partir del cual en el caso (A) del ejemplo VII.4, el muestreo doble es más conveniente.

Solución:

$$\frac{c'}{c} < 0{,}00309$$

Ejemplo VII.6

Supongamos que se conocen los siguientes datos de una variable ($\mathfrak{X}$, $\mathfrak{Y}$) definida en una población finita de tamaño grande:

$$\sigma_x^2 = 5 \quad ; \quad \sigma_y^2 = 25 \quad ; \quad \sigma_{xy} = 8 \quad ; \quad R = \bar{X}/\bar{Y} = 0{,}16$$

Para estimar la media $\bar{X}$, se pide decidir si es más conveniente el muestreo doble para estimador de razón con $c' = 10$, $c = 500$ que el muestreo aleatorio simple en una fase con el mismo coste total, en los casos:

(A) Muestra de segunda fase independiente de la primera.

(B) En segunda fase se toma una submuestra de la primera.

Solución:

En muestreo aleatorio simple, para el tamaño de muestra permitido por el coste, la varianza es:

$$V(\bar{x}) = \frac{\sigma_x^2}{n_o} = \frac{c}{C} \sigma_x^2$$

Por otra parte, en el caso (A) se cumple que el muestreo doble es más conveniente si:

$$a + b \frac{c'}{c} + 2 \sqrt{a\,b} \sqrt{c'/c} < \sigma_x^2$$

y en el caso (B):

$$a + b' \frac{c'}{c} + 2 \sqrt{a\,b'} \sqrt{c'/c} < \sigma_x^2$$

efectuando los cálculos necesarios resulta en ambos casos más conveniente el muestreo doble, ya que se cumple 3,491 y 3,806 menores que 5.

Ejemplo VII.7

En relación con el caso (A) del ejemplo VII.6, determine a partir de qué valor del coeficiente de correlación ρ entre $\mathfrak{X}$ e $\mathfrak{Y}$ no es conveniente el muestreo doble.

Solución:

$$\sigma_{xy} < 3{,}5409 \text{ , o bien } \rho < 0{,}3167$$

Ejemplo VII.8

En muestreo doble para estratificación se ha tomado en primera fase una muestra de tamaño $n' = 400$ y en la segunda, formados tres estratos, $n_1 = 20$; $n_2 = 10$; $n_3 = 10$. Los resultados se presentan en el siguiente cuadro:

$\hat{W}_h$	$\bar{x}_h$	s_h^2
0,55	2,8	15
0,32	8,2	200
0,13	26,0	1000

Se pide estimar el error de muestreo de $\hat{\bar{X}}$, utilizando las fórmulas (7) y (6) del texto.

Solución:

1,99 , o bien 26,4 %
2,03 , o bien 26,9 %

Ejemplo VII.9

Admitiendo los datos del ejemplo VII.8 como estimaciones de W_h, $\bar{X}_h$, σ^2 y siendo los costes unitarios $c' = 10$; $c = 500$; se pide

calcular los tamaños y la afijación que para un mismo coste total son óptimos.

Solución:

$$n' = 200\ ;\quad n = 44\ ;\quad n_1 = 9\ ;\quad n_2 = 18\ ;\quad n_3 = 17$$

Suponiendo que los valores W_h, $\bar{X}_h$, σ_h^2 son los verdaderos poblaciones, el error de muestreo en este caso es: 1,24, o bien 16,4 % .

Ejemplo VII.10

Supongamos conocidos para una variable ($\mathfrak{X}$, $\mathfrak{Y}$) los datos $\sigma_x = 2$ $\sigma_y = 4$; $\sigma_{xy} = 6$; $\bar{X} = 10$. Una primera muestra $n' = 100$, nos da $\bar{y}' = 40{,}6$; y una submuestra $n = 25$, $\bar{x} = 9{,}8$, $\bar{y} = 40{,}1$. Estime la media $\bar{X}$ por regresión lineal en muestreo doble óptimo; calcule el coste total siendo $c' = 10, c = 600$.

Solución:

$$\rho = 0{,}75\ ;\quad b = 6/16\ ;\quad \hat{\bar{X}}_{rl} = 9{,}8 + 6/16\ (40{,}6 - 40{,}1) = 9{,}988$$

C = 16000.

Ejemplo VII.11

Calcule el error de muestreo relativo correspondiente al ejemplo VII.10.

Solución:

$$V(\hat{\bar{X}}_{rl}) = \frac{(1 - 0{,}75^2)\ 4}{25} + \frac{0{,}75^2\ .\ 4}{100} = 0{,}0955.$$

El error de muestreo es 3,09 %

Ejemplo VII.12

Determine los tamaños de muestra óptimos n y n' para el coste total dado en el ejemplo VII.10. Suponga también estimador óptimo.

VIII. ERRORES AJENOS AL MUESTREO. MARCOS IMPERFECTOS

La teoría ordinaria de poblaciones finitas, como hemos visto en anteriores capítulos, supone que la lista de la que se selecciona la muestra aleatoria coincide con el colectivo de unidades que se desea investigar. En la práctica no siempre es esto cierto, ya que las listas, ficheros o registros, pueden presentar algunos defectos que dan origen a la aparición de sesgos y a la alteración de las varianzas de los estimadores usuales. Utilizamos aquí la denominación de Marco en sentido restringido para referirnos a la Lista de unidades de muestreo.

VIII.1. EL PROBLEMA DE LAS UNIDADES VACIAS O EXTRAÑAS

Por unidad vacía o extraña, entendemos una unidad que estando incluída en el marco, no contiene ninguna unidad estadística perteneciente al colectivo que se desea investigar, o que no es ella misma una unidad perteneciente al marco que se desea muestrear. En este segundo caso es más idónea la denominación de unidad extraña y unidad vacía en el primero, pero ambas situaciones son equivalentes en cuanto al problema metodológico que plantea su presencia en el marco de muestreo. Como ejemplos ilustrativos consideremos los siguientes:

1. Si para una encuesta de población se muestrea en una lista de viviendas, las deshabitadas son unidades vacías.
2. Si para estimar alguna característica de la producción de leche, se muestrea en una lista de explotaciones agrícolas, las que no se dediquen total o parcialmente a dicha actividad son unidades extrañas.

Estos dos ejemplos ponen de manifiesto que la presencia de unidades vacias o extrañas se produce principalmente en dos tipos de situación práctica, si bien conceptualmente equivalentes:

A. Por no estar actualizada, la lista incluye unidades que han dejado de pertenecer al colectivo que se desea muestrear.

B. La población que se desea muestrear es una subpoblación de la cubierta por el marco.

Ante esta situación pueden adoptarse varias reglas de conducta, no todas igualmente buenas, cuya puesta en práctica depende de los recursos disponibles.

(I) Depuración del marco:

Una primera solución consiste, obviamente, en eliminar del marco las unidades vacías o extrañas, al tiempo que se averigua cúantas incluía. Es claro que así queda resuelto el problema, puesto que a continuación puede seleccionarse la muestra en el marco deseado, conociendo el número de unidades (todas no vacías) que contiene. En muchos casos esto no será posible con los recursos dados, bien sea porque el marco disponible no contiene información acerca de cuales unidades son vacías (siendo necesario un trabajo de campo de carácter exhaustivo o la formación de un nuevo marco) o por limitaciones de tiempo, presupuesto o de cualquier otro tipo.

(II) Sustitución de las unidades vacías de la muestra:

Otra solución consiste en seleccionar la muestra en el marco disponible, no depurado, sustituyendo las unidades que resulten ser vacías por otras, aleatoriamente seleccionadas de entre las restantes del marco, hasta completar el tamaño de muestra prefijado con unidades todas no vacías. Se verá que para estimar un total, en ausencia de información acerca del número de unidades vacías en la lista de muestreo, esta solución da lugar a estimaciones sesgadas.

(III) Información acerca del número de unidades vacías:

Consideraremos en este problema la hipótesis de que el número de unidades vacías incluidas en el marco es conocido (pero no cúales son). Se verá que esta información es suficiente para resolver el problema, aunque el muestreo se efectúe sin la previa depuración del marco.

De acuerdo con este planteamiento vamos a considerar cinco procedimientos para estimación de un total; el primero consiste en seleccionar la muestra en el marco depurado, como caso ideal que servirá de comparación, y los otros cuatro resultan de combinar la sustitución o no de las unidades vacías de la muestra, con la hipótesis de que se conoce o no el número de unidades vacías, cuando se muestrea en el marco no depurado.

VIII.1.1. NOTACIONES Y RELACIONES BASICAS

Sin pérdida de generalidad podemos numerar las unidades no vacías de 1 a N', de modo que:

$\mathcal{A} = \{ A_1, \ldots, A_N \}$ es el marco disponible (no depurado).

$\mathcal{A}' = \{ A_1, \ldots, A_{N'} \}$ es el marco depurado.

La proporción de unidades vacías en el marco disponible la denotamos por W, y así podemos escribir:

$$W = \frac{N - N'}{N} = 1 - \frac{N'}{N} \quad ; \quad N' \leqslant N \qquad (1)$$

Por otra parte, si es X_i el valor de la característica que se desea investigar en la unidad A_i, atribuímos el valor $X_i = o$, cuando A_i es vacía o extraña ya que en cualquier caso su contribución al total X' del colectivo que se desea investigar, $\mathcal{A}'$, es nula. En consecuencia las sumas totales en ambos marcos son coincidentes, en particular:

$$\sum_1^N X_i = \sum_1^{N'} X_i = X = X'$$

$$\sum_1^N X_i^2 = \sum_1^{N'} X_i^2 \quad ; \quad \sum_{i \neq j}^{N} X_i X_j = \sum_{i \neq j}^{N'} X_i X_j$$

Las medias calculadas en uno u otro marco no son iguales, en particular:

$$\bar{X} = (1-W)\,\bar{X}' \qquad (2)$$

$$\sigma^2 = (1-W)\,\sigma'^2 + \frac{W}{1-W}\,\bar{X}^2 = (1-W)\,\sigma'^2 + W(1-W)\,\bar{X}'^2 \qquad (3)$$

La relación entre las medias se obtiene inmediatamente teniendo en cuenta (1); la relación entre las varianzas en el marco no depurado σ^2, y en el marco depurado σ'^2, se obtiene fácilmente operando de la siguiente forma:

$$N\sigma^2 = \sum_1^N X_i^2 - \frac{X^2}{N} = \sum_1^{N'} X_i^2 - \frac{X^2}{N'} + \frac{X^2}{N'} - \frac{X^2}{N} =$$

$$= N'\sigma'^2 + X^2\left(\frac{1}{N'} - \frac{1}{N}\right) = N'\sigma'^2 + X^2\,\frac{W}{N'}$$

y teniendo en cuenta (1), resulta la fórmula (3).

Excepto en poblaciones muy pequeñas puede suponerse $N \simeq N-1$; $N' \simeq N'-1$, resultando las varianzas práticamente iguales a las varianzas:

$$S^2 = (1-W)\,S'^2 + \frac{W}{1-W}\,\bar{X}^2 = (1-W)\,S'^2 + W(1-W)\,\bar{X}'^2 \qquad (4)$$

Otra relación que usuaremos más adelante es que siendo T_i definida de la siguiente forma:

$$T_i = \begin{cases} 1 & \text{si } A_i \text{ es no vacía.} \\ 0 & \text{si } A_i \text{ es vacía.} \end{cases}$$

Se cumple: $\mathrm{Cov}(X_i, T_i) = W\bar{X}$

En efecto,

$$\text{Cov}\ (X_i\ ,\ T_i) = \frac{1}{N}\sum_{1}^{N} X_i T_i - \left(\frac{1}{N}\sum_{1}^{N} X_i\right)\left(\frac{1}{N}\sum_{1}^{N} T_i\right) =$$

$$= \frac{X}{N} - \frac{X}{N}(1 - W) = W\,\overline{X}. \qquad (5)$$

VIII. 1. 2. ESTIMACION DEL TOTAL CUANDO SE MUESTREA EN EL MARCO DEPURADO

En muestreo sin reposición con probabilidades iguales, siendo n el tamaño prefijado de la muestra:

$$\hat{X} = \frac{N'}{n}\,x \quad ; \quad \text{es insesgado para } X.$$

$$V(\hat{X}) = \frac{N'(N' - n)}{n}\,S'^2 \quad ; \quad \text{es su varianza.}$$

VIII.1.3. ESTIMACION DEL TOTAL CUANDO SE MUESTREA EN EL MARCO NO DEPURADO.

Consideremos por separado las cuatro posibilidades mencionadas al final de VIII.1.

(I) Se desconoce N′ y no se sustituyen las unidades vacías que aparezcan en la muestra.

En estas condiciones, si se muestrea en $\mathcal{A}$ sin reposición y probabilidades iguales:

$$\hat{X}_1 = \frac{N}{n}\,x \qquad \text{es insesgado.}$$

$$V(\hat{X}_1) = \frac{N(N-n)}{n} S^2 > V(\hat{X})$$

Para probar que es insesgado tenemos en cuenta que la variable indicadora ε_i = uno ó cero según que $A_i \in \mathcal{A}$ aparezca en la muestra, verifica $E(\varepsilon_i) = n/N$ para todas las unidades A_i incluídas en el marco $\mathcal{A}$ no depurado:

$$E(\hat{X}_1) = \frac{N}{n} E\left(\sum_1^N X_i \varepsilon_i\right) = \frac{N}{n} \sum_1^N X_i E(\varepsilon_i) =$$

$$= \sum_1^N X_i = \sum_1^{N'} X_i = X'$$

El cálculo de su varianza también es inmediato:

$$V(\hat{X}_1) = V\left(\frac{N}{n} x\right) = N^2 V(\bar{x}) = \frac{N^2(N-n)}{N} \frac{S^2}{n} = \frac{N(N-n)}{n} S^2$$

Puede demostrarse exactamente (Mirás, 1979) que la varianza de $\hat{X}_1$ es mayor que la de $\hat{X}$, pero para mayor simplicidad veamos aquí una comparación admitiendo la aproximación $N \simeq N-1$; $N' \simeq N'-1$; en este supuesto la relación entre S^2 y S'^2 establecida en VIII.1.1. nos permite escribir:

$$\frac{V(\hat{X}_1)}{V(\hat{X})} = \frac{N(N-n)}{N'(N'-n)} \frac{S^2}{S'^2} = \frac{N(N-n)}{N'(N'-n)} \left\{1-W + W(1-W)\frac{\bar{X}'^2}{S'^2}\right\} =$$

$$= \frac{N(N-n)}{N'(N'-n)} \left\{\frac{N'}{N} + \frac{N'}{N} W \frac{\bar{X}'^2}{S'^2}\right\} = \frac{N-n}{N'-n}\left(1 + W \frac{\bar{X}'^2}{S'^2}\right)$$

de donde resulta cuando hay unidades vacías:

$$(N-n) > (N'-n)$$

y $W\,\bar{X}'^2 / S'^2$ es una cantidad positiva, por lo tanto:

$$\left\{V(\hat{X}_1) / V(\hat{X})\right\} > 1 \implies V(\hat{X}_1) > V(\hat{X}).$$

Es obvio que si no hay unidades vacías en $\mathcal{A}$, $(N-n) = (N'-n)$ y $W = 0$, por lo tanto los dos estimadores y sus varianzas coinciden. Si hay unidades vacías la varianza de $\hat{X}_1$ es tanto mayor que la de $\hat{X}$ cuanto mayor sea la proporción W de unidades vacías y menor sea la cuasivarianza relativa $S'^2 / \bar{X}'^2$ de la característica que se estudia en las unidades no vacías.

(II) Se desconoce N' y se sustituyen aleatoriamente las unidades vacías de la muestra hasta seleccionar n no vacías.

En estas coneiciones el estimador del total:

$$\hat{X}_2 = \frac{N}{n}\, x \quad ; \quad \text{tiene un sesgo positivo } B_2 = \frac{W}{1-W}\, X'.$$

y su varianza es:

$$V(\hat{X}_2) = \frac{N^2\,(N'-n)}{N'}\,\frac{S'^2}{n} > V(\hat{X})$$

Para calcular el sesgo y la varianza de $\hat{X}_2$ tenemos en cuenta que la sustitución aleatoria de las unidades vacías de la muestra hasta completar el tamaño prefijado, n, de unidades no vacías, es probabilísticamente equivalente a efectuar el muestreo en el marco depurado $\mathcal{A}'$ (véase ejemplo I.18); así resulta que se cumple:

$$E(\varepsilon_i) = \begin{cases} 0 & \text{si } A_i \text{ es vacía} \\ \dfrac{n}{N'} & \text{si } A_i \text{ es no vacía} \end{cases}$$

por lo tanto:

$$E(\hat{X}_2) = \frac{N}{n} \sum_{1}^{N} X_i \, E(\varepsilon_i) = \frac{N}{n} \frac{n}{N'} \sum_{1}^{N'} X_i = \frac{1}{1-W} X'$$

y su sesgo:

$$B_2 = \left(\frac{1}{1-W} - 1 \right) X' = \frac{W}{1-W} X' \; ; \quad B_2 > 0 \text{ si } W > 0.$$

El cálculo de su varianza es inmediato de acuerdo con las consideraciones probabilísticas mencionadas:

$$V(\hat{X}_2) = N^2 \, V(\bar{x}) = \frac{N^2 \, (N' - n)}{N'} \frac{S'^2}{n}$$

y teniendo en cuenta la $V(\hat{X})$, se obtiene:

$$\frac{V(\hat{X}_2)}{V(\hat{X})} = \frac{N^2}{N'^2} = \frac{1}{(1-W)^2} > 1$$

puesto que cuando existen unidades vacías, es $0 < (1 - W)^2 < 1$, por lo tanto $V(\hat{X}_2) > V(\hat{X})$. También es este caso, si no existen unidades vacías en el marco los estimadores $\hat{X}_2$ y $\hat{X}$ coinciden y tienen la misma varianza.

El Error Cuadrático Medio de $\hat{X}_2$, resulta ser:

$$ECM(\hat{X}_2) = \frac{1}{(1-W)^2} \left\{ V(\hat{X}) + W^2 \, X'^2 \right\}$$

(III) Se conoce N′ y no se sustituyen las unidades vacías que aparezcan en la muestra

Si denotamos por n' el número de unidades no vacías que aparecen en la muestra aleatoria de tamaño n, seleccionada en el marco no depurado, el estimador:

$$\hat{X}_3 = \frac{N'}{n'} x \quad , \quad \text{es insesgado.}$$

Su varianza viene dada por la fórmula aproximada:

$$V(\hat{X}_3) = \frac{N'(N-n)}{n} S'^2$$

y cumple la relación:

$$V(\hat{X}) < V(\hat{X}_3) < V(\hat{X}_1)$$

Si denotamos por $E_{n'}$ la esperanza matemática en el espacio de muestras en que n' es fijo, podemos escribir:

$$E(\hat{X}_3) = N' E\left(E_{n'} \left(\frac{x}{n'} \right) \right) = N' E(\bar{X}') = X'$$

Para calcular su varianza podemos actuar de forma análoga a los estimadores de razón ya que n' es variable aleatoria y tomará en general distintos valores en muestras distintas.

$$V(\hat{X}_3) = E\left(N' \frac{x}{n'} - X' \right)^2 = N'^2 E\left(\frac{x}{n'} - \bar{X}' \right)^2 = N'^2 E\left(\frac{x/n}{n'/n} - \bar{X}' \right)^2$$

ahora, denotando por $w = (n-n')/n$ la proporción de unidades vacías en

la muestra:

$$V(\hat{X}_3) = N'^2 \; E\left(\frac{\bar{x} - (1-w)\bar{X}'}{(1-w)}\right)^2$$

y utilizando las aproximación $w \simeq W$ en el denominador

$$V(\hat{X}_3) = \frac{N'^2}{(1-W)^2} \; E(\bar{x} - (1-w)\,\bar{X}')^2$$

$$= \frac{N'^2}{(1-W)^2} \left\{ V(\bar{x}) + \bar{X}'^2\, V(1-w) - 2\bar{X}'\, Cov(\bar{x}, (1-w)) \right\} =$$

$$= \frac{N'^2\,(N-n)}{(1-W)^2\, N\, n} \left\{ S^2 + \bar{X}'^2\, W(1-W) - 2\,\bar{X}'\, W\, \bar{X} \right\}$$

en cuyo último paso se ha utilizado la relación (5) y las aproximaciones $N \cong N - 1$; $N' \simeq N' - 1$; después de simplificar resulta:

$$V(\hat{X}_3) \simeq \frac{N(N-n)}{n} \; (S^2 - W(1-W)\,\bar{X}'^2)$$

y teniendo en cuenta la relación (4) se obtiene finalmente la fórmula propuesta:

$$V(\hat{X}_3) = \frac{N'(N-n)}{n} \; S'^2$$

La comprobación de que $V(\hat{X}_3) > V(\hat{X})$ resulta inmediata puesto que cuando hay unidades vacías es $N > N'$. Para comprobar que $V(\hat{X}_3)$ es menor que $V(\hat{X}_1)$ es suficiente tener en cuenta la fórmula (4) de la que se deduce inmediatamente $N'S'^2 < NS^2$ y por tanto $V(\hat{X}_3) < V(\hat{X}_1)$

(IV) **Se conoce N' y se sustituyen aleatoriamente las unidades vacías que aparezcan en la muestra hasta obtener n no vacías.**

Estas condiciones el estimador del total:

$\hat{X}_4 = N'/n = x$ es insesgado y su varianza es igual a la de $\hat{X}_1$. Para probar estas afirmaciones es suficiente con tener en cuenta que el factor de elevación N'/n que aparece en el estimador es una constante conocida y que la sustitución aleatoria de unidades vacías de la muestra hasta completar el tamaño n prefijo, equivale a efectuar el muestreo en el marco depurado; por lo tanto:

$$E(\hat{X}_4) = X' \; ; \; V(\hat{X}_4) = V(\hat{X}) = \frac{N'(N'-n)}{n} S'^2$$

VIII.1.4. RESUMEN Y CONCLUSIONES

En el cuadro adjunto se resumen los resultados obtenidos y a continuación repasamos las principales conclusiones.

1°) Si se conoce el número de unidades vacías (y por tanto el de no vacías) en el marco de muestreo $\mathcal{A}$, el problema queda resuelto aplicando el procedimiento IV ya que el estimador $\hat{X}_4$ es insesgado y su varianza es la misma que se obtendría efectuando el muestreo en el marco depurado $\mathcal{A}'$. Si en este caso no se sustituyen las unidades vacías, el estimador insesgado $\hat{X}_3$ tiene mayor varianza.

2°) Si no se conoce el número de unidades vacías es mejor no sustituir las que aparezcan en la muestra, ya que se puede formar un estimador insesgado $\hat{X}_1$. El estimador $\hat{X}_2$ formado cuando se sustituyen las unidades vacías tiene sesgo; si bien el $ECM(\hat{X}_2)$ podría ser del orden de la varianza de $\hat{X}_1$, debe tenerse en cuenta que ésta puede reducirse aumentando el tamaño de la muestra pero no el sesgo de $\hat{X}_2$ que es constante.

3°) En cuanto a la formación de un estimador para la media $\bar{X}' = X'/N'$, el conocimiento del número de unidades vacías en el marco es irrelevante ya que cualquiera de los dos estimadores:

CUADRO RESUMEN

PROCEDIMIENTO DE ESTIMACION				Número de unidades no vacías en la muestra	Varianza	Sesgo
Procedimiento de muestreo			Estimador			
Muestreo en el marco depurado			$\hat{X} = \frac{N'}{n}x$	n	$\frac{N'(N' - n)}{n}S'^2$	—
Muestreo en el marco no depurado	No se conoce N'	No se sustituyen las unidades vacías	$\hat{X}_1 = \frac{N}{n}x$	n'	$\frac{N(N - n)}{n}S^2$	—
		Se sustituyen las unidades vacías	$\hat{X}_2 = \frac{N}{n}x$	n	$\frac{N^2(N' - n)}{N'n}S'^2$	$\frac{W}{1 - W}X'$
	Se conoce N'	No se sustituyen las unidades vacías	$\hat{X}_3 = \frac{N'}{n'}x$	n'	$\frac{N'(N - n)}{n}S'^2$ (1)	—
		Se sustituyen las unidades vacías	$\hat{X}_4 = \frac{N'}{n}x$	n	$\frac{N'(N' - n)}{n}S'^2$	—

(1) Fórmula aproximada.

$$\bar{X}'_3 = \frac{x}{n'} \qquad \text{(Sin sustitución de unidades vacías)}$$

$$\bar{X}'_4 = \frac{x}{n} \qquad \text{(Con sustitución de unidades vacías)}$$

es insesgado. Tiene menor varianza $\bar{X}'_4$ por lo tanto es preferible a $\bar{X}'_3$.

VIII.2. EL PROBLEMA DE LAS UNIDADES REPETIDAS

La presencia de unidades repetidas en el marco de muestreo introduce sesgos en los estimadores usuales; ello se debe a que en tales condiciones se alteran las probabilidades de selección que en vez de ser iguales para todas las unidades, son proporcionales al número de veces que se repite en el marco la identificación de dichas unidades.

Por unidad repetida se entiende una unidad de muestreo cuya identificación figura más de una vez en la lista o marco disponible. Los siguientes ejemplos ilustran este concepto:

1. Si para una encuesta de empresas se muestrea en una lista de establecimientos, las que dispongan de más de un establecimiento son unidades repetidas.

2. Si para una encuesta el marco de muestreo se ha formado reuniendo listas procedentes de diversas fuentes, es posible que algunas unidades aparezcan repetidas.

Es obvio que el problema desaparece si se depura el marco disponible eliminando todas las identificaciones repetidas, no obstante veamos si es posible evitar los mencionados sesgos sin necesidad de efectuar previamente la depuración del marco.

VIII.2.1. PLANTEAMIENTO FORMAL DEL PROBLEMA

Suponemos que la población objeto de estudio consta de N′ unidades, que denotamos:

$$\mathcal{U} = \{U_1, U_2, \ldots, U_i, \ldots, U_{N'}\}$$

y que en la lista o marco disponible, la identificación de la unidad U_i

(dirección postal, ficha, número de registro, código, etc.) figura r_i veces, siendo r_i un entero igual o mayor que 1. Según esto el marco disponible es un conjunto de N ($N \geqslant N'$) elementos que denotamos:

$$\mathcal{A} = \{ A_1, A_2, \ldots, A_i, \ldots, A_N \} \quad ; \quad N = \sum_{1}^{N'} r_i$$

Si es X_i ; ($i = 1, 2, \ldots, N'$) el valor de la característica en estudio en la unidad U_i, el total y la media en la población objeto $\mathcal{U}$, son:

$$X' = \sum_{1}^{N'} X_i \quad ; \quad \bar{X}' = \frac{1}{N'} \sum_{1}^{N'} X_i$$

sin embargo el total y la media en el marco disponible $\mathcal{A}$, son:

$$X = \sum_{1}^{N} X_i = \sum_{1}^{N'} r_i X_i \quad ; \quad \bar{X} = \frac{1}{N} \sum_{1}^{N'} r_i X_i$$

Si se efectúa el muestreo en el marco $\mathcal{A}$ que contiene unidades repetidas, el estimador usual para el total:

$$\hat{X} = \frac{N}{n} x$$

tiene sesgo positivo (supuestos $X_i \geqslant 0$) como estimador del total X' que se desea; este sesgo es:

$$E(\hat{X}) - X' = \sum_{1}^{N'} (r_i - 1) X_i$$

El estimador usual de la media:

$$\hat{\bar{X}} = \frac{x}{n}$$

también tiene un sesgo:

$$E(\hat{\bar{X}}) - \bar{X}' = \frac{1}{NN'} \sum_{1}^{N'} (r_i N' - N) X_i$$

que puede ser positivo, negativo ó nulo. En el cuadro adjunto se presentan tres hipotéticos marcos con distinta estructura de unidades repetidas, para una pequeña población de cinco unidades distintas. Si se muestrea en $\mathcal{A}'$ el sesgo del estimador de la media es negativo, en el $\mathcal{A}''$ positivo

Valores X_i de la población	Marcos con unidades repetidas					
	$\mathcal{A}'$		$\mathcal{A}''$		$\mathcal{A}'''$	
	X_i	r_i	X_i	r_i	X_i	r_i
1	1	2	1	1	1	1
2	2	1	2	1	2	1
4	4	1	4	1	4	2
6	6	1	6	1	6	1
7	7	1	7	2	7	1
Total: 20	21		27		24	
Media: 4	3,5		4,5		4	

y en el $\mathcal{A}'''$ es nulo. El estimador del total tendrá siempre sesgo positivo.

VIII.2.2. MUESTREO EN EL MARCO NO DEPURADO PARA ESTIMACION DEL TOTAL

Suponemos muestreo sin reposición con probabilidades iguales. Veamos que pueden adoptarse varios procedimientos que permiten formar estimadores insesgados del total X', con la condición de que sea posible conocer el número r_i de veces que se repite en el marco cada una de las unidades seleccionadas para la muestra; esto evita la depuración completa del marco. Consideremos pues los tres siguientes procedimientos:

(I) Se efectúan n extracciones sin reposición en el marco no depurado, utilizándose para formar el total muestral las n observaciones, de modo que si aparecen unidades repetidas el correspondiente valor se repite tantas veces como haya aparecido en la muestra.

(II) Solamente se tienen en cuenta para formar el total muestral los

valores que corresponden a unidades no repetidas en la muestra; en este caso el tamaño efectivo $n' \leqslant n$ es una variable aleatoria que expresa el número de unidades distintas de la muestra.

(III) Se efectúan extracciones sucesivas comprobando si la unidad seleccionada ha aparecido anteriormente, en cuyo caso se elimina de la muestra; se continúa el proceso hasta obtener en la muestra n unidades distintas.

PROCEDIMIENTO I

Siendo ε_i la variable aleatoria que expresa el número de veces que la unidad U_i aparece en la muestra, la probabilidad conjunta responde a la hipergeométrica para N' clases:

$$P_r(\varepsilon_1 = t_1, \ldots, \varepsilon_i = t_i, \ldots, \varepsilon_N = t_N) = \frac{\binom{r_1}{t_1} \cdots \binom{r_i}{t_i} \cdots \binom{r_{N'}}{t_{N'}}}{\binom{N}{n}}$$

resultando:

$$E(\varepsilon_i) = \frac{r_i}{N} n \quad ; \quad V(\varepsilon_i) = \frac{(N-n)}{(N-1)} \frac{r_i\, n}{N} \left(1 - \frac{r_i}{N}\right)$$

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = -\frac{(N-n)\,n}{(N-1)} \frac{r_i}{N} \frac{r_j}{N} \; ; \quad i \neq j$$

Un estimador insesgado del total X' queda definido de la siguiente forma:

$$\hat{X}'_1 = \sum_1^n \frac{N}{n} \frac{X_i}{r_i} \text{ , o bién } \hat{X}'_1 = \sum_1^{n'} \frac{N X_i t_i}{n r_i}$$

siendo n' el número de unidades distintas en la muestra ($n' \leqslant n$) y t_i

el número de veces que en la muestra se repite la unidad, así pues: $\sum_{1}^{n'} t_i = n$. Se comprueba inmediatamente, haciendo uso de la distribución de probabilidades de ε_i, que el estimador es insesgado:

$$E\left(\sum_{1}^{n'} \frac{N X_i t_i}{n r_i}\right) = E\left(\sum_{1}^{N'} \frac{N X_i \varepsilon_i}{n r_i}\right) = \sum_{1}^{N'} \frac{N X_i E(\varepsilon_i)}{n r_i} =$$

$$= \sum_{1}^{N'} \frac{N X_i n r_i}{N n r_i} = \sum_{1}^{N'} X_i = X'$$

El cálculo de la varianza de $\hat{X}'_1$, también es sencillo:

$$V(X'_1) = V\left(\sum_{1}^{N'} \frac{N X_i \varepsilon_i}{n r_i}\right) =$$

$$= \frac{N^2}{n^2}\left\{\sum_{1}^{N'} \frac{X_i^2}{r_i^2} V(\varepsilon_i) + \sum_{j \neq i}^{N'} \frac{X_i}{r_i} \frac{X_j}{r_j} \text{Cov}(\varepsilon_i, \varepsilon_i)\right\}$$

y sustituyendo $V(\varepsilon_i)$; Cov. $(\varepsilon_i, \varepsilon_j)$ antes calculadas, se obtiene:

$$V(\hat{X}') = \frac{N-n}{(N-1)n}\left(\sum_{1}^{N'} \frac{X_i^2}{Z_i} - X'^2\right)$$

donde $Z_i = r_i / N$ es la proporción de veces que u_i está repetida en el marco no depurado. Mediante simples transformaciones se llega a la expresión equivalente:

$$V(\hat{X}'_1) = \frac{(N-n)}{(N-1)n} \sum_{1}^{N'} Z_i \left(\frac{X_i}{Z_i} - X'\right)^2$$

Puede probarse (Mirás, 1979) que un estimador insesgado de la varianza, viene dado por:

$$v(X') = \frac{(N-n)}{N(n-1)n} \sum_{1}^{n} \left(\frac{X_i}{Z_i} - \hat{X}'_1 \right)^2$$

PROCEDIMIENTO II

Si se utiliza el procedimiento II, la muestra queda constituída por las n' unidades distintas obtenidas en las n extracciones, es decir, cada unidad se "cuenta" una sola vez aunque haya aparecido repetida. En consecuencia, la probabilidad de que la unidad U_i aparezca en la muestra efectiva es:

$$\pi_i = \Pr(u_i) = 1 - \frac{\binom{N-r_i}{n}}{\binom{N}{n}}$$

y la probabilidad de que aparezca el par (u_i, u_j) en cualquier orden $(i \neq j)$:

$$\pi_{ij} = \Pr(u_i, u_j) = \frac{\sum_{h,K=1}^{r_i r_j} \binom{r_i}{h} \binom{r_j}{k} \binom{N-r_i-r_j}{n-2}}{\binom{N}{n}}$$

entonces podemos aplicar la teoría de H – H estudiada en cáp. III.9; así resulta:

$$\hat{X}'_2 = \sum_{1}^{n'} \frac{X_i}{\pi_i}$$

estimador insesgado del total X'. Su varianza así como estimaciones insesgadas vienen dadas por las fórmulas deducidas en III.9, sustituyendo π_i y π_{ij} por los valores aquí calculados.

PROCEDIMIENTO III

Este procedimiento es equivalente al esquema de urna en que cada unidad u_i ($i = 1, \ldots, N'$) viene representada por r_i bolas y que cuando por primera vez aparece la unidad u_i en la muestra se eliminan las r_i bolas que la representan. Como sabremos, este esquema para $n > 2$ da lugar a tediosos cálculos de las probabilidades de inclusión π_i y π_{ij} Fórmulas para un estimador insesgado del total, de su varianza y de su estimación, vendrían dadas por la teoría H – H, estudiada en III.9.

VIII.2.3. RESUMEN Y CONCLUSIONES

El método usualmente recomendado es el que denominamos procedimiento I; tiene la ventaja de que su aplicación es sencilla puesto que para obtener un estimador insesgado del total, solamente se requiere dividir el valor observado en cada unidad muestral X_i, por el número de veces que la citada unidad está repetida en la lista o marco de muestreo y acumular este valor ponderado X_i/r_i tantas veces como se haya repetido en la muestra la unidad. Por otra parte el cálculo y estimación de la varianza no requiere operaciones más complicadas de lo habitual. Una observación importante en cuanto a la precisión del estimador es la siguiente: la fórmula de su varianza:

$$V(\hat{X}'_1) = \frac{N - n}{(N-1)\,n} \left(\sum_1^{N'} \frac{X_i^2}{Z_i} - X'^2 \right)$$

es igual a la de muestreo con reposición y probabilidades proporcionales a r_i, en el marco $\mathfrak{U}$, multiplicada por el factor $(N - n)/(N - 1)n$; en consecuencia el estimador insesgado $\hat{X}'_1$ cuando existen unidades repetidas en el marco pudiera ser más o menos preciso que el estimador usual para el marco depurado (sin reposición y probabilidades iguales); esto dependerá de que los r_i (número de veces que se repite la unidad u_i en el marco no depurado) esté o no positivamente correlacionado con el valor X_i que tratamos de observar. A este respecto véase el ejemplo VIII.3.

El procedimiento II es viable en la práctica pero requiere operaciones de cálculo, de π_i para formar el estimador y también de π_{ij} para

estimar su varianza, que lo hacen menos deseable que el procedimiento I. No obstante tiene un interés metodológico, ya que entre otras propiedades nos hace ver la necesidad de repetir en el estimador I el valor X_i/r_i tantas veces como se haya repetido la unidad u_i en la muestra puesto que si nos quedamos solamente con las que son distintas, estaremos en la necesidad de aplicar el procedimiento II.

Acerca del procedimiento III debe decirse que en la práctica para $n > 2$, es poco útil debido a la complejidad del cálculo de π_i y π_{ij}. Pero además de estas dificultades, para el cálculo de las citadas probabilidades de selección no es suficiente el conocimiento de los r_i correspondientes a las unidades de la muestra, sino que se requieren los r_i de todas las N' unidades distintas de la población; este hecho invalida al procedimiento III como solución alternativa a la previa depuración completa del marco, si se desea un estimador insesgado.

VIII.2.4. ESTIMACION DEL NUMERO DE UNIDADES DISTINTAS EN EL MARCO.

Un estimador insesgado para N', utilizando el procedimiento I de muestreo, viene dado por:

$$\hat{N}' = \frac{N}{n} \sum_{1}^{n} \frac{1}{r_i} = \frac{N}{n} \sum_{i}^{n'} \frac{t_i}{r_i}$$

siendo n' el número de unidades distintas en la muestra, t_i el de veces que se repite dentro de la muestra la unidad A_i y r_i en el marco. La comprobación de que es insesgado, así como la fórmula de su varianza:

$$V(\hat{N}') = \frac{N-n}{(N-1)n} \left(\sum_{1}^{N'} \frac{N}{r_i} - N'^2 \right)$$

y de su estimador insesgado:

$$\hat{V}(\hat{N}') = \frac{N-n}{N(n-1)n} \sum_{1}^{n} \left(\frac{N}{r_i} - \hat{N}' \right)^2$$

se deducen de la teoría desarrollada en el estudio del procedimiento I para estimar un total, dando a X_i el valor 1.

VIII.2.5. ESTIMACION DE LA MEDIA EN UN MARCO CON UNIDADES REPETIDAS.

Puesto que la media objeto de estimación es: $\bar{X}' = X'/N'$

si el número N' de unidades distintas en el marco, es conocido, el problema está resuelto ya que se divide el estimador del total por N' y su varianza por N'^2. Si N' no es conocido podemos acudir a formar un estimador de razón:

$$\hat{\bar{X}}' = \frac{\hat{\bar{X}}'}{\hat{N}'} = \frac{\sum_1^n \frac{X_i}{r_i}}{\sum_1^n \frac{1}{r_i}}$$

utilizando el procedimiento de muestreo I.

Una expresión aproximada para el E.C.M. de $\hat{\bar{X}}'$, aplicando la teoría ordinaria del estimador de la razón viene dada por:

$$ECM(\hat{\bar{X}}') = \frac{N(N-n)}{N'^2 n}\left\{V\left(\frac{X_i}{r_i}\right) + \bar{X}'^2 V\left(\frac{1}{r_i}\right) - 2\bar{X}' \operatorname{Cov}\left(\frac{X_i}{r_i}, \frac{1}{r_i}\right)\right\}$$

que se obtiene inmediatamente, teniendo en cuenta que:

1°. R se identifica con $\bar{X}'$

2°. $E\left(\frac{1}{r_i}\right) = \frac{1}{N}\sum_1^N \frac{1}{r_i} = \frac{N'}{N}$

3°. Las varianzas que aparecen en la fórmula (suponiendo muestreo sin reposición y $N \simeq N-1$) son:

$$V\left(\frac{X_i}{r_i}\right) = \frac{1}{N}\sum_1^N \frac{X_i^2}{r_i^2} - \frac{X'^2}{N^2}$$

$$V\left(\frac{1}{r_i}\right) = \frac{1}{N}\sum_{1}^{N}\frac{1}{r_i^2} - \frac{N'^2}{N^2}$$

$$\mathrm{Cov}\left(\frac{X_i}{r_i}\ ,\ \frac{1}{r_i}\right) = \frac{1}{N}\sum_{1}^{N}\frac{X_i}{r_i^2} - \frac{N'\,X'}{N^2}$$

El estimador $\hat{\bar{X}}'$ es en general sesgado y de acuerdo con la teoría estudiada en cáp. II.8, su expresión aproximada viene dada por:

$$B(\hat{\bar{X}}') = \frac{(N-n)\,\bar{X}'}{(N-1)n}\left(\frac{N}{N'^2}\sum_{1}^{N}\frac{1}{r_i^2} - \frac{N}{N'X'}\sum_{1}^{N}\frac{X_i}{r_i^2}\right)$$

Las varianzas y covarianza de X_i/r_i, $1/r_i$ pueden estimarse de la muestra, conociendo los correspondientes r_i, por los métodos usuales, así pues con las notaciones:

$$Y_i = \frac{X_i}{r_i} \quad ; \quad \bar{y} = \frac{1}{n}\sum_{1}^{n} Y_i = \frac{1}{n}\sum_{1}^{n}\frac{X_i}{r_i}$$

$$R_i = \frac{1}{r_i} \quad ; \quad \bar{r} = \frac{1}{n}\sum_{1}^{n} R_i = \frac{1}{n}\sum_{1}^{n}\frac{1}{r_i}$$

resulta:

$$\hat{V}\left(\frac{X_i}{r_i}\right) = \frac{\sum_{1}^{n}(Y_i-\bar{y})^2}{n-1} \quad ; \quad \hat{V}\left(\frac{1}{r_i}\right) = \frac{\sum_{1}^{n}(R_i-\bar{r})^2}{n-1}$$

$$\hat{\mathrm{Cov}}\left(\frac{X_i}{r_i}\ ,\ \frac{1}{r_i}\right) = \frac{\sum_{1}^{n}(Y_i-\bar{y})(R_i-\bar{r})}{n-1}$$

VIII.3 EJEMPLOS

Ejemplo VIII.1

Supongamos que conocemos los datos de una población finita que tiene una proporción W = 0,05 de unidades vacías:

$$N = 2000 \quad ; \quad S'^2 = 2,2 \quad ; \quad X' = 3 \quad ; \quad n = 200$$

Exprese $\sqrt{ECM}$ de los estimadores I, II y III para el total X', en porcentaje del correspondiente al muestreo en el marco depurado.

Solución:

I) 112,93 % ; II) 196,13 % ; III) 102,90 %

El sesgo de $\hat{X}_2$ es 300 ; 5,26 % de X' .

Ejemplo VIII.2

Deduzca las fórmulas de la varianza de los estimadores I, II y III para el total cuando hay unidades vacías, en muestreo con reposición, y calcule su $\sqrt{ECM}$ para los datos:

$$N = 2000 \quad ; \quad \sigma'^2 = 2,2 \quad ; \quad X' = 3 \quad ; \quad n = 250 \quad ; \quad W = 0,10$$

Solución:

I) 211,28 ; II) 628,65 ; III) 177,99

El sesgo de $\hat{X}_2$ es 600 (11,11 % de X'). Las soluciones expresadas en porcentaje del error de muestreo en el marco depurado, son:

I) 125,13 % ; II) 372,31 % ; III) 105,41 %

Ejemplo VIII.3

Calcule la varianza del estimador del total según el procedimiento I, para los tres marcos del cuadro que aparece en el apartado VIII.2.1. del texto y compárela con la varianza del estimador usual en el marco depurado.

Solución:

$$\mathcal{A}' : V(\hat{X}'_1) = \frac{6-n}{n} \quad 46{,}6$$

$$\mathcal{A}'' : V(\hat{X}'_1) = \frac{6-n}{n} \quad 17{,}8$$

$$\mathcal{A}''' : V(\hat{X}') = \frac{6-n}{n} \quad 37{,}6$$

Marco depurado:

$$V(\hat{X}) = \frac{5-n}{n} \quad 32{,}5$$

Se observa que en $\mathcal{A}'$ y $\mathcal{A}'''$ la varianza de $\hat{X}'_1$ es mayor que la varianza de $\hat{X}$ (en marco depurado) pero en $\mathcal{A}''$ la varianza de $\hat{X}'_1$ es menor que en el marco depurado para tamaño de muestra $n \leqslant 3$; se deduce inmediatamente de la condición:

$$17{,}8\,(6-n) < 32{,}5\,(5-n)$$

Ejemplo VIII.4

En un marco $\mathcal{A}$ de tamaño N = 75 que tiene unidades repetidas, se ha extraído una muestra aleatoria de tamaño $n = 7$. Los resultados se presentan en el siguiente cuadro, donde t_i representa el número de veces

que se repiten las unidades dentro de la muestra:

Unidad	A_1	A_2	A_3	A_4	A_5	A_6
Valor X_i	3	2	1	2	4	1
r_i	2	1	1	5	1	1
t_i	1	1	1	2	1	1

Calcule la estimación insesgada del total por el procedimiento I y estime su varianza.

Solución:

$$\hat{X}'_1 = 110{,}36 \quad ; \quad V(\hat{X}'_1) = 1143{,}2 \quad ; \quad \sigma(\hat{X}'_1) \simeq 30{,}6\,\%$$

Ejemplo VIII.5

En las condiciones del ejemplo VIII.4, estime el número de unidades distintas que hay en el marco de muestreo y la varianza del estimador utilizado.

Solución:

$$\hat{N}' \simeq 53 \; ; \; (52{,}5) \quad \hat{V}(\hat{N}') \simeq 109{,}29 \quad ; \quad \sigma(\hat{N}') = 19{,}7\,\%$$

Ejemplo VIII.6

Con los datos del ejemplo VIII.4, estime la media $\bar{X}'$ y el error de muestreo (ECM) del estimador.

Solución

$$\hat{\bar{X}}' = 2{,}1 \quad ; \quad \hat{V}(\hat{\bar{X}}') \simeq 0{,}2699 \quad ; \quad \sigma(\hat{\bar{X}}') \simeq 24{,}7\,\%$$

Cálculos parciales son:

$$\hat{V}\left(\frac{X_i}{r_i}\right) = 1{,}5691 \; ; \; \hat{V}\left(\frac{1}{r_i}\right) = 0{,}15 \; ; \; \hat{Cov}\left(\frac{X_i}{r_i}, \frac{1}{r_i}\right) = 0{,}2833$$

IX ERRORES AJENOS AL MUESTREO. NO RESPUESTA. EFECTO DEL ENTREVISTADOR.

En el capítulo anterior nos hemos referido a las consecuencias que sobre los estimadores usuales producen ciertos defectos de la lista o marco de muestreo; también estudiamos la forma de resolver o de reducir el problema, estableciendo estimadores insesgados para el total. Ahora veremos los efectos producidos por dos acontecimientos que prácticamente se presentan en todas las encuestas, tanto exhaustivas como por muestreo, en las que se utilizan cuestionarios y agentes entrevistadores para la recogida de datos. Se trata de la denominada *No Respuesta* y del *Efecto del entrevistador.*

IX.1. EL PROBLEMA DE LA NO RESPUESTA

En los censos y encuestas por muestreo puede ocurrir que no se recojan datos de todas las unidades; en este caso falta información de una parte de la población, o de la muestra seleccionada (y por tanto del colectivo que representa), y se plantea el problema de establecer qué inferencias podemos hacer a partir de los datos obtenidos, y el de tomar acciones encaminadas a conseguir algún conocimiento de la parte de la población no representada en la encuesta.

Diversas circunstancias contribuyen a la existencia de este problema; citamos algunas más usuales:

1) No se encuentra en la dirección, o domicilio, la persona que ha de responder al cuestionario. En los censos este problema puede ser menor ya que normalmente se planifican varias visitas,

tienen carácter oficial, están respaldados por la legislación adecuada y generalmente la oficina de estadística cuenta con la colaboración de las autoridades y organismos públicos locales. En las encuestas (incluso las de carácter oficial que gozan del apoyo legislativo específico) el problema suele ser más acusado ya que, de una parte, el incremento del número de visitas hace aumentar el coste y de otra, el tiempo de recogida se retrasa pudiendo la respuesta llegar fuera del plazo establecido para obtener estimaciones en el momento oportuno. Este problema adquiere especial importancia si la fecha de la encuesta coincide con un período habitual de vacaciones o fiestas locales; por esta razón, entre otras, el censo general de población en España se realiza actualmente en el mes de Marzo, sustituyendo a la tradicional fecha de 31 de Diciembre. También contribuye a aumentar la falta de respuesta el hecho de que la encuesta esté deseñada de forma que el respondiente deba ser una persona específica, o especialmente cualificada para ello, ya que aumentará la probabilidad de "ausencia" en el momento de la entrevista, al no ser válida la respuesta dada por otro miembro del hogar o unidad a la que se dirige la encuesta.

2) En ocasiones la persona o entidad que se trata de entrevistar rehúsa su colaboración; es lo que los estadísticos denominamos "negativa". Además de las razones personales, subjetivas, por las que un entrevistado niega su respuesta, el problema es más frecuente cuando el cuestionario hace referencia a datos de conducta íntima (consumo de drogas, aborto, opiniones políticas en ciertos casos, etc.) o bién de carácter monetario (salarios, origen de sus rentas, etc.) especialmente si el entrevistado tiene la creencia de que su respuesta puede perjudicarle.

Cualquiera que sea la causa de la falta de datos de una parte de la muestra, o colectivo, se denomina en la terminología de los estadísticos *No respuesta o Falta de respuesta.*

La publicidad bien dirigida contribuye positivamente a la colaboración de los entrevistados, así como el adiestramiento, habilidad y capacidad de crear un clima de confianza por parte del agente entrevistador. En las encuestas "por correo", el problema tiene otros matices relacio-

nados con la facilidad de devolver el cuestionario (sobre con franqueo pagado, etc.), con la dificultad de cumplimentarlo sin ayuda del agente entrevistador, etc. Es muy conveniente, en cualquier caso, que el entrevistado considere que con su colaboración presta un servicio a la sociedad.

Por otra parte, la legislación estadística es también un elemento importante en ayuda de aminorar este problema. Prácticamente en todos los paises la Oficina Central de Estadística está autorizada a requerir información de los ciudadanos y personas jurídicas y éstos obligados a proporcionarla en las condiciones estipuladas por la ley; en contrapartida los datos individuales (confiados a la Oficina Central) susceptibles de ser identificados, están somtidos a secreto, incluso de los demás organismos del Estado como es el caso del Instituto Nacional de Estadística en España.

Conviene también mencionar que dentro de la falta de respuesta pueden entrar causas tales como inclemencias del tiempo, o dificultades de acceso de los entrevistadores a zonas de montaña. Las unidades ilocalizables producen falta de respuesta pero debe determinarse si en realidad se trata de un defecto del marco que incluye unidades no existentes; cuando se dispone de entrevistadores, la veracidad y esmero en su trabajo es elemento clave para dilucidar los casos dudosos.

IX.1.1. EFECTOS DE LA NO RESPUESTA

Un modelo ilustrativo consiste en suponer que las N unidades de la población, en caso de ser seleccionadas para una encuesta, se clasifican en dos "estratos", el primero formado por N_1 unidades que responderían y el otro formado por las N_2 unidades restantes que no responderían ($N = N_1 + N_2$). La proporción de no respondientes será $W = N_2/N$, y si denotamos por X_i, $\overline{X}_i$, σ_i el total, la media y la desviación típica del estrato de respondientes ($i = 1$) o de no respondientes ($i = 2$), se cumplen las relaciones:

$$\left.\begin{aligned} \overline{X} &= (1 - W)\,\overline{X}_1 + W\,\overline{X}_2 \\ \sigma^2 &= \sum_1^2 \frac{N_i}{N}\,\sigma_i^2 + \sum_1^2 \frac{N_i}{N}\,(\overline{X}_i - \overline{X})^2 \end{aligned}\right\} \text{Media y Varianza de la población}$$

En primer lugar la no respuesta ocasiona una disminución del tamaño de la muestra, si es n, tendremos n_1 respuestas y n_2 no respuestas ($n = n_1 + n_2$). Pero no es éste el principal problema ya que eligiendo n suficientemente grande, el número de respuestas n_1 podría ser suficiente; el verdadero problema reside en que no disponemos de información acerca de n_2 unidades muestrales que representarían a un colectivo de N_2 en la población.

La media muestral de las n_1 respuestas:

$$\bar{x}_1 = \frac{x_1}{n_1}$$

es un estimador insesgado de $\bar{X}_1$ (media del estrato poblacional de respondientes), véase II.11 por lo tanto su sesgo es:

$$B(\bar{x}_1) = \bar{X}_1 - \bar{X} = W(\bar{X}_1 - \bar{X}_2)$$

proporcional a la proporción de no respondientes y a la diferencia de medias de los dos estratos. Es obvio que si $W = 0$ (todos son respondientes) ó $\bar{X}_1 = \bar{X}_2$ (los dos estratos tienen la misma media) el estimador no tendría sesgo. En el caso de que $W > 0$, no siempre puede admitirse que $\bar{X}_1 = \bar{X}_2$, incluso como hipótesis de trabajo, ya que la condición de no respondiente puede estar muy relacionada con la característica $\mathfrak{X}$ que se investiga, dando lugar a dos medias $\bar{X}_1$ y $\bar{X}_2$ muy distintas. En consecuencia debemos suponer que el estimador es sesgado.

El efecto sobre la estimación del total también es un sesgo. Si se toma como estimador del total:

$\hat{X}_I = N\bar{x}_1$; por ser $E(\hat{X}_I) = N\bar{X}_1$, su sesgo es:

$$B(\hat{X}_I) = NW(\bar{X}_1 - \bar{X}_2) = \frac{W}{1-W}(X_1 - X_2).$$

Si se toma como estimador:

$$\hat{\bar{X}}_{II} = \frac{N}{n} x_1 \quad ; \text{ por ser } \quad E(\hat{\bar{X}}_{II}) = E\left(\frac{N}{n} n_1 \bar{x}_1\right) =$$

$$= N_1 \bar{X}_1 = X_1 \ , \quad \text{su sesgo es:} \quad B(\hat{\bar{X}}_{II}) = -X_2 \ .$$

Es obvio que también la falta de respuesta ejercerá un efecto sobre el estimador de σ^2, traducido en un sesgo. La cuasivarianza de los n_1 valores obtenidos, s_1^2, es un estimador insesgado de σ_1^2 ; por lo tanto el sesgo de s_1^2 , es:

$$B(s_1^2) = \sigma_1^2 - \sigma^2 = W(\sigma_1^2 - \sigma_2^2) - W(1-W)(\bar{X}_1 - \bar{X}_2)^2$$

que se obtiene después de algunas operaciones con la descomposición de σ^2 dada al comienzo de este apartado IX.1.1. Se observa que en ausencia de no respuesta (W = 0) el sesgo no existe; en otro caso la cuasivarianza de las unidades muestrales respondientes es un estimador sesgado.

IX.1.2. LA SUSTITUCION DE UNIDADES QUE FALTAN

Una primera idea para, de alguna forma, superar el problema del sesgo producido por la no respuesta, consiste en seguir seleccionando unidades aleatoriamente hasta completar el tamaño de muestra requerido. Es claro que entonces tenemos no n_1 unidades, sino n, pero todas pertenecientes al estrato de respondientes, con lo cual solamente conseguimos disminuir la varianza de los estimadores pero no el sesgo que permenece constante.

Si el diseño es estratificado y la sustitución de las no respuestas se hace con unidades que tienen características similares, podríamos admitir que no hay diferencia significativa entre $\bar{X}_1$ y $\bar{X}_2$ y por tanto que el sesgo es pequeño. No obstante, éste no es un razonamiento científicamente válido a menos que se disponga de información directa

del estrato de no respondientes que justifique la hipótesis $\bar{X}_1 \simeq \bar{X}_2$. El estimador ajustado del total sería:

$$\hat{X}_{III} = \frac{N}{n} \; x$$

insesgado bajo la fuerte hipótesis de que $\bar{X}_1 = \bar{X}_2$; en efecto:

$$E(\hat{X}_{III}) = N\bar{X}_1 = (N_1 + N_2)\bar{X}_1 = X_1 + N_2\bar{X}_1 = X_1 + X_2 = X.$$

Si $\bar{X}_1 \neq \bar{X}_2$ su sesgo coincide con el de $\hat{X}_I$, si bien tendrá menos varianza.

IX.1.3. ALGUNAS TECNICAS PARA LA NO RESPUESTA

Diversos autores han diseñado modelos aplicables a situaciones particulares en las que se produce no respuesta. Veamos algunos más conocidos.

METODO DE HANSEN Y HURWITZ (1946)

Este método, aunque es de aplicación bastante general, fué diseñado para encuestas que se envían por correo. De las n_2 unidades que no han respondido, se selecciona una muestra aleatoria de tamaño $n_{21} \leqslant n_2$ y se envían entrevistadores para conseguir las respuestas de las n_{21} unidades seleccionadas en la segunda ocasión de entre las no respondientes a la primera.

El total muestral podemos descomponerlo como:

$$x = x_1 + x_2$$

donde x_2 corresponde a las unidades sin respuesta en la primera ocasión. Si denotamos por $f_{21} = n_{21}/n_2$ la fracción de submuestreo de la segunda ocasión, un estimador insesgado de x_2, es:

$$\hat{x}_2 = \frac{1}{f_{21}} \; x_{21} = \frac{n_2}{n_{21}} \; x_{21}$$

siendo x_{21} el total de la segunda ocasión; en consecuencia un estimador insesgado del total poblacional, es:

$$\hat{X} = \frac{N}{n} (x_1 + \hat{x}_2) = \frac{1}{f} (x_1 + \hat{x}_2)$$

donde $f = n/N$ es la fracción de muestreo de la primera ocasión.

Para el cálculo de la varianza del estimador utilizamos la técnica:

$$V(\hat{X}) = V \underset{w}{E}(\hat{X}) + \underset{w}{E} V(\hat{X})$$

donde $\underset{w}{E}$ y $\underset{w}{V}$ representan esperanza y varianza condicionada al conjunto de muestras w en las que n_1 y n_2 son fijos. Para el primer sumando tenemos:

$$\underset{w}{E}(\hat{X}) = \underset{w}{E} \left\{ \frac{1}{f} (x_1 + \hat{x}_2) \right\} = \frac{1}{f} \left\{ (x_1 + \underset{w}{E}(\hat{x}_2) \right\} =$$

$$= \frac{1}{f} (x_1 + x_2) = \frac{1}{f} x = \frac{N}{n} x .$$

$$V\left(\underset{w}{E}(\hat{X}) \right) = V \left(\frac{N}{n} x \right) = \frac{N^2}{n} (1-f) S^2 ; \text{ (muestreo sin reposición)}$$

Para el segundo sumando, denotamos por s_2^2 la cuasivarianza de las n_2 unidades no respondientes en la primera ocasión, entonces tenemos:

$$\underset{w}{V}(\hat{X}) = \underset{w}{V} \left\{ \frac{1}{f} (x_1 + \hat{x}_2) \right\} = \underset{w}{V} \left(\frac{1}{f} \hat{x}_2 \right) = \frac{1}{f^2} \underset{w}{V}(x_2) =$$

$$= \frac{1}{f^2} \frac{n_2^2}{n_{21}} (1-f_{21}) s_2^2 = \frac{(1-f_{21})}{f^2 \cdot f_{21}} n_2 s_2^2$$

$$E\left(V_w(\hat{X})\right) = \frac{(1-f_{21})}{f^2 \cdot f_{21}} \quad E(n_2 s_2^2);$$

el cálculo de $E(n_2 s_2^2)$ puede hacerse:

$$E\left(n_2 \; E(s^2/n_2)\right) = E(n_2 S_2^2) = f N_2 S_2^2$$

por tanto:

$$V(\hat{X}) = N^2 (1-f) \frac{S^2}{n} + \frac{1}{f}\left(\frac{1}{f_{21}} - 1\right) N_2 S_2^2 \qquad (1)$$

donde S^2 es la cuasivarianza de las N unidades de la población y S_2^2 es la correspondiente al estrato de N_2 no respondientes.

De los dos sumandos de la fórmula (1), el primero representa la varianza del estimador del total si no hubiera falta de respuesta; el segundo expresa el incremento de varianza debido al submuestreo de las no respondientes. Si se trata de estimar la media $\bar{X}$, formamos el estimador:

$$\hat{\bar{X}} = \frac{1}{Nf}(x_1 + \hat{x}_2) = \frac{1}{n}(x_1 + \hat{x}_2)$$

y su varianza, será:

$$V(\hat{\bar{X}}) = (1-f)) \frac{S^2}{n} + \frac{N_2}{n\,N}\left(\frac{1}{f_{21}} - 1\right) S_2^2$$

Consideremos ahora la función de coste total:

$$C = c_o n + c_1 n_1 + c_2 n_{21}$$

donde el primer sumando representa el coste de selección de la muestra de tamaño n y los otros dos, los costes de las entrevistas en primera y

segunda fase. Puesto que n_2 es aleatorio, fijada una fracción de submuestreo f_{21} también $n_{21} = f_{21} \cdot n_2$ es aleatorio, entonces utilizamos el coste esperado que calculamos de la siguiente forma:

$$C = n\left(c_o + c_1 \frac{n_1}{n} + c_2 \frac{n_{21}}{n}\right) = n\left(c_o + c_1 \frac{n_1}{n} + c_2 f_{21} \frac{n_2}{n}\right)$$

$$E(C) = n\,(c_o + c_1 W' + c_2 f_{21} W)$$

siendo $W' = (1-W)$ la proporción teórica de respondientes.

Para optimizar el procedimiento podemos fijar la precisión (varianza del estimador) en un valor V y minimizar el coste esperado. Aplicando el método de multiplicadores de Lagrange para optimizar respecto de n, tamaño total de la muestra, y de la fracción f_{21} de submuestreo de no respondientes, escribimos:

$$\phi = n\,(c_o + c_1 W' + c_2 W f_{21}) +$$

$$+ \lambda \left\{ \frac{N(N-n)}{n} S^2 + \frac{N N_2}{n}\left(\frac{1}{f_{21}} - 1\right) S_2^2 - V \right\}$$

y derivando respecto de n, de f_{21} y junto con la ecuación de condición resulta el sistema de ecuacionees:

$$\left.\begin{aligned}
&(c_o + c_1 W' + c_2 W f_{21}) + \lambda\left\{\frac{-N^2}{n^2} S^2 - \frac{N N_2}{n^2}\left(\frac{1}{f_{21}} - 1\right) S_2^2\right\} = 0 \\
&c_2 W n + \lambda\left(\frac{-N N_2 S_2^2}{f_{21}^2}\right) = 0 \\
&\frac{N(N-n)}{n} S^2 + \frac{N N_2}{n}\left(\frac{1}{f_{21}} - 1\right) S_2^2 = V
\end{aligned}\right\}$$

despejando λ de la primera y de la segunda, igualamos y después de simplificar resulta:

$$W c_2 f_{21} \left\{ N S^2 f_{21} + (1 - f_{21}) N_2 S_2^2 \right\} = (c_0 + c_1 W') N_2 S_2 + N_2 S^2 W c_2 f_{21}$$

$$W c_2 f_{21} (N^2 S^2 - N_2 S_2^2) = (c_0 + W' c_1) N_2 S_2^2$$

de donde puede despejarse explícitamente f_{21}:

$$f_{21} = \sqrt{\frac{c_1}{c_2} \cdot \frac{S_2^2 (W' + c_0/c_1)}{S^2 - W S_2^2}}$$

donde se observa que cuanto mayor es c_2 respecto de c_1, menor es la fracción de seubmuestreo de no respondientes que debemos tomar. También puede escribirse f_{21} en la forma:

$$f_{21} = \sqrt{\frac{S_2^2 (W' c_1 + c_0)}{c_2 (S^2 - W S_2^2)}}$$

y la hipótesis de trabajo $S_2^2 = S^2$ (cuasivarianza del estrato de no respondientes igual a la cuasivarianza poblacional), nos da:

$$f_{21} \simeq \sqrt{\frac{W' c_1 + c_0}{W' c_2}}$$

Para despejar el valor óptimo de n, de la ecuación de condición obtenemos:

$$\frac{N^2 S^2}{n} - N S^2 + \frac{N N_2 \left(\frac{1}{f_{21}} - 1 \right) S_2^2}{n} = V \; ; \text{ de donde:}$$

$$n = \frac{N^2 S^2 + N N_2 \left(\frac{1}{f_{21}} - 1 \right) S_2^2}{V + N S^2}$$; que puede escribirse de

la siguiente forma:

$$n = \frac{N^2 S^2}{V + N S^2} + \frac{N N_2 \left(\frac{1}{f_{21}} - 1 \right) S_2^2}{V + N S^2}$$

donde el primer sumando representa el tamaño de la muestra necesario para estimar el total con un error de muestreo $E = \sqrt{V}$ si no hubiera falta de respuesta (véase II.6). El valor numérico del óptimo tamaño n, se obtiene sustituyendo el f_{21} calculado antes.

En caso de muestreo con reposición operando de forma análoga, se llega a los siguientes valores óptimos:

$$f_{21} = \frac{\sigma_2}{\sigma} \sqrt{\frac{c_o + c_1 W'}{c_2}}$$

donde σ es la desviación típica poblacional y σ_2 la del estrato de no respondientes; para n se obtiene:

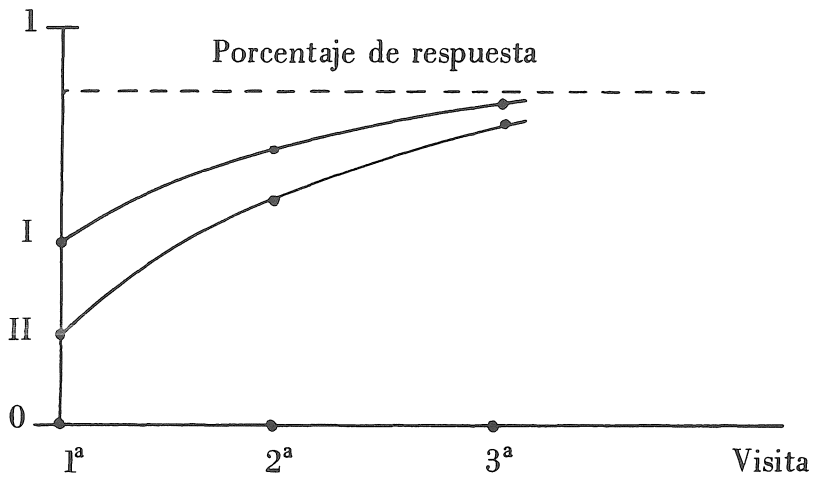
$$n = \frac{N^2 \sigma^2}{V} + \frac{N N_2 \sigma_2^2 / f_{21}}{V}$$

donde f_{21} debe tomarse de la fórmula anterior.

MODELO DE DEMING (1952)

Un procedimiento para aminorar la no respuesta, especialmente cuando es producida por ausencia de la persona específicamente seleccionada para la encuesta, consiste en realizar visitas sucesivas; para ello

se fija un número mínimo de "revisitas" que deben hacerse a cada unidad antes de abandonarla como "contacto imposible". Un elemento importante a considerar es si la entrevista se dirige a un adulto cualquiera del hogar seleccionado, o bien el cuestionario debe ser respondido por un adulto específico (cabeza de familia, por ejemplo). En el primer caso el número de respuestas en la primera visita es mucho mayor que en el segundo, pero a medida que aumenta el número de revisitas el porcentaje total de respuestas tiende a igualarse, si bien es mayor en el primer



caso. Stephan y Mc Carthy (1958) presentan datos de varias encuestas obteniendo unos valores promedios, que siguen la tendencia ilustrada en la figura adjunta. Para la aplicación y análisis del rendimiento de esta técnica es indispensable un buen estudio de los costes promedios por unidad en las sucesivas revisitas; Durbin (1954) ha efectuado algunos estudios sobre esta cuestión. Una medida que puede resultar en la práctica más útil, es el coste promedio por unidad de las respuestas obtenidas después de *i* visitas. Observemos por último, que el procedimiento de revisitas retrasa los resultados finales, aunque no siempre será necesariamente un retraso excesivo si de antemano se planifica la duración de la recogida de datos.

Deming desarrolló el siguiente modelo, muy útil para el estudio de esta técnica de revisitas. La población se divide en L clases de acuerdo

con la probabilidad de que el encuestado sea encontrado en casa; sean:

w_{ij} = Probabilidad de que un encuestado de la j clase sea encontrado en i visitas.

P_j = Proporción de la población en la j clase.

$\overline{X}_j$ = Media poblacional de la j clase.

σ_j^2 = Varianza poblacional de la j clase.

y supongamos $w_{ij} > 0$ para todas las clases, aunque el modelo puede adaptarse para incluir personas imposibles de encontrar. Después de i visitas la composición de la muestra es la siguiente: L clases, a cada una de las cuales pertenecen los elementos del estrato j de los cuales se ha obtenido respuesta durante alguno de los i primeros intentos, más una clase que incluye todos los elementos de la muestra inicial de los que no se ha obtenido respuesta en los mencionados intentos. Si es n_o el tamaño inicial de la muestra, el número n_{ij} de elementos observados de clase j después de i visitas, es una multinomial de parámetros:

$$n_o \; ; \left\{ w_{i1} \, P_1 \; ; \; w_{i2} \, P_2 \; ; \ldots ; w_{iL} P_L \; ; \; \left(1 - \sum_{j=1}^{L} w_{ij} \, P_j \right) \right\}$$

El número total de respuestas después de i visitas es una binomial de parámetros:

$$n_i : \left\{ n_o \quad ; \quad \sum_{j=1}^{L} w_{ij} \, P_j \right\}$$

por tanto el número esperado de respuestas en i visitas:

$$E(n_i) = n_o \sum_{i=1}^{L} w_{ij} \, P_j$$

Para un n_i fijo, el número de entrevistas obtenidas en cada una de las clases, viene dada por la multinomial condicionada que como sabemos

también es multinomial de parámetros:

$$\left\{ n_i \; ; \; \frac{w_{i1} \; P_1}{\Sigma \; w_{ij} \; P_j} \; , \; \frac{w_{i2} \; P_2}{\Sigma \; w_{ij} \; P_j} \; , \ldots , \; \frac{w_{iL} \; P_L}{\Sigma \; w_{ij} \; P_j} \right\}$$

por tanto:

$$E\,(n_{ij} \mid n_i) = \frac{n_i \; w_{ij} \; P_j}{\sum_j \; w_{ij} \; P_j}$$

y la esperanza matemática de la media muestral condicionada a n_i, es:

$$E(\bar{x}_i \mid n_i) = E\left(\frac{\Sigma \; n_{ij} \; \bar{x}_{ij}}{n_i}\right) = \frac{\sum_j n_i \; w_{ij} \; P_j \; \bar{X}_j}{n_i \sum_j \; w_{ij} \; P_i} =$$

$$= \frac{\Sigma \; w_{ij} \; P_j \; \bar{X}_j}{\Sigma \; w_{ij} \; P_j} = \bar{X}_i$$

por tanto vemos que no depende de n_i y en consecuencia podemos afirmar que la media de la muestra después de i visitas, independientemente de cual sea el número n_i de respuestas, es también $\bar{X}_i$. El sesgo de la media muestral después de i visitas es por tanto:

$$B\,(\bar{x}_i) = \bar{X}_i - \bar{X}$$

La varianza de $\bar{x}_i$ condicionada a un n_i, es:

$$V\,(\bar{x}_i \mid n_i) = \frac{\sum_j \; w_{ij} \; P_j \left(\sigma_j^2 + (\bar{X}_j - \bar{X}_i)^2 \right)}{n_i \sum_j \; w_{ij} \; P_j} \qquad (1)$$

y la varianza de $\bar{x}_i$, cualquiera que sea el número n_i de respuestas obtenidas después de i visitas, viene dada aproximadamente (despreciando términos de orden $1/n_i^2$) por la fórmula (1) sustituyendo en ella n_i por su valor esperado, resultando:

$$V(\bar{x}_i) \simeq \frac{\sum_j w_{ij} P_j \left(\sigma_j^2 + (\bar{X}_j - \bar{X}_i)^2 \right)}{n_o \left(\sum_j w_{ij} P_j \right)^2} \qquad (2)$$

Finalmente el Error Cuadrático Medio de la media $\bar{x}_i$ obtenida después de i vistas, es:

$$ECM(\bar{x}_i \mid i) = V(\bar{x}_i \mid i) + (\bar{X}_i - \bar{X})^2$$

Veamos ahora el coste de hacer i visitas. Sea c_r el coste por unidad en la r^a visita. Por otra parte el número esperado de entrevistas conseguidas al pasar de la visita (r -1) a la r, es:

$$\sum_j (w_{rj} - w_{r-1,j}) P_j$$

por tanto el coste esperado para i—visitas, siendo n_o el tamaño inicial de la muestra es:

$$n_o \left\{ c_1 \Sigma w_{1j} P_j + c_2 \Sigma (w_{2j} - w_{1j}) P_j + \cdots + c_i \Sigma (w_{ij} - w_{i-1,j}) P_j \right\}$$

METODO DE POLITZ Y SIMMONS (1949, 1950)

Este método fué propuesto por Hartley (1946) y desarrollado por Politz y Simmons (1949, 1950). No se realiza más que una visita pero se ponderan los resultados de acuerdo con la probabilidad de encontrar al encuestado. Dado que una persona que está en casa una proporción π del tiempo tiene una probabilidad relativa de ser entrevistado igual a π, su respuesta debería recibir una ponderación $1/\pi$. Suponiendo entonces que las visitas se realizan durante 6 tardes de cada semana, al entrevistado se le pregunta si estaba en casa cada una de las cinco tar-

des anteriores a la de la entrevista; si declara que estuvo $t = 0,1,2,3,4,5$ se toma como estimador de π:

$$\hat{\pi} = \frac{1 + t}{6}$$

El estimador ponderado de Politz y Simmons para la media es, entonces:

$$\bar{x}_{PS} = \frac{\sum_{t=0}^{5} 6\, n_t\, \bar{x}_t / (1+t)}{\sum_{t=0}^{5} 6\, n_t / (1+t)}$$

donde n_t es el número de entrevistados que declaran haber estado en casa t de las cinco noches anteriores y x_t su media. Este estimador es menos sesgado que la media simple de la muestra de respuestas obtenidas en la visita pero su varianza será mayor ya que las ponderaciones son estimadas. Desarrollos elaborados de este método y comparaciones con el método de revisitas pueden consultarse, además de los autores citados, en Deming (1953) y Durbin (1954).

IX.1.4. NO RESPUESTA PARCIAL

En el contexto del problema de la no respuesta existe otra situación algo distinta de la considerada hasta aquí; se trata de lo que podemos denominar *no respuesta parcial* que se presenta cuando falta la respuesta a una o varias de las variables ó preguntas incluídas en el cuestionario. También puede incluirse en este concepto una respuesta erróneamente codificada de modo que no corresponda al conjunto de posibles valores válidos de la variable; en este caso la situación es equivalente a no tener respuesta.

Para el tratamiento de este problema daremos algunas reglas de conducta útiles que nos conducen a técnicas de detección–imputación automática de errores e inconsistencias en los cuestionarios. Si bien se aplican usualmente en censos y encuestas, hasta hace relativamente poco tiempo no han aparecido trabajos planteando modelos teóricos que

permitan una instrumentación informática rápida y efectiva. Véase Fellegi y Oldt (1976), trabajo del que Gómez Alonso (1980) ha realizado una traducción al español; Rodríguez–Ponga y Villán (1985) han aportado recientemente alguna mejora.

En primer lugar la revisión de cuestionarios por encargados de grupo debe ser elemento básico para detectar la no respuesta parcial y las inconsistencias. Debe ponerse en contacto con el entrevistador para subsanar, si es posible, el defecto observado. Como segunda medida puede, en ocasiones, recurrirse a una llamada telefónica con la unidad informante. El último recurso consiste en un tratamiento informático (cuando se trata de muestras grandes) para el que son necesarios, al menos, los siguientes pasos:

1. Definición del campo de valores válidos para cada una de las variables.

2. Definición del campo de valores válidos para cada variable, condicionado por los valores de las restantes variables. Con esto se pretende evitar inconsistencias entre respuestas a distintas variables ya que es usual que el valor observado en una variable, aún estando dentro del conjunto de valores válidos, tenga que cumplir ciertos requisitos en relación con los valores de otras variables.

3. Establecer reglas determinísticas o aleatorias (imputando valores de cuestionarios "similares" al que está en revisión) para imputar los valores correspondientes a conjuntos de variables que presenten inconsistencias.

4. No debe olvidarse el establecimiento de la lista de variables en las que se concede posibilidad de "no consta".

5. Por último mencionemos la importancia de evaluar el efecto producido por este tratamiento de los datos faltantes o inconsistentes. Es corriente efectuar simplemente un recuento del número de imputaciones efectuadas pero pensamos que esto apenas es útil para el análisis; en cambio proponemos diseñar un recuento, en tabla de doble entrada, del número de unidades con valor X_{i1} en el cuestionario y con valor X_{i2} después del proceso de imputación automática. De esta forma si es n_{ij} la frecuencia de casos

(X_{i1}, X_{j2}) podemos estudiar los trasvases que producen el resultado neto de $n_{i.}$ unidades en el estado X_i en el cuestionario y $n_{.i}$ en el mismo estado después de la imputación. Obviamente n_{ii} es la frecuencia de unidades no afectadas por el tratamiento y $(n_{.i} - n_{i.})$ la diferencia neta. Si X es cuantitativa, debe obtenerse, además de la frecuencia n_{ij} en cada "casilla" de intervalos (I_i, I_j), la suma X_{ij} de los valores de la variable, de esta forma podemos evaluar el efecto producido que consistirá en una "deformación" de la distribución inicial (antes de la imputación automática) en la distribución final (después de la imputación).

IX.1.5. MODELOS DE RESPUESTA ALEATORIZADA

Las dificultades para contar con la colaboración veraz de los encuestados frente a preguntas de carácter íntimo, se traducen en falta de respuesta por negativa o en la obtención de respuestas deliberadamente falsas. Ejemplos de este tipo de preguntas son muy abundantes en las encuestas socio-económicas, como nivel de ingresos, trabajo sumergido, aborto, consumo de drogas, conducta sexual, etc. Warner (1965) publica un trabajo con el que se inicia una serie de investigaciones sobre el tema, basadas en esta primera idea denominada técnica o modelo de *Respuesta Aleatorizada.*

MODELO DE WARNER

La idea inicial de Warner consiste en lo siguiente: Si es $\mathcal{C}$ la clase de unidades de la población que poseen la característica en estudio y π su proporción poblacional que deseamos estimar, se presenta al encuestado una de dos preguntas:

(I) Pertenece a la clase $\mathcal{C}$. (Si = 1 ; No = 0)

(II) Pertenece a la clase $\bar{\mathcal{C}}$. (Si = 1 ; No = 0)

la primera con probabilidad P y la segunda con probabilidad (1−P). El entrevistador proporciona al encuestado una tarjeta con las dos preguntas y un mecanismo aleatorio (baraja, dado, caja con bolas de colores, etc.) y le pide que en privado, de forma que el entrevistador

no conozca la pregunta seleccionada, efectúe el sorteo y le comunique su respuesta. De esta forma queda garantizado el secreto ya que se desconoce a qué pregunta corresponde la respuesta.

Una observación trivial pero interesante para una mejor comprensión de la idea, es que esta técnica debería denominarse de "pregunta aleatorizada", ya que en efecto lo que se aleatoriza es la pregunta y no la respuesta.

En estas condiciones si $X_i =$ uno ó cero, es la respuesta obtenida, tenemos:

$$\begin{cases} \Pr(X_i = 1) = \lambda = P\pi + (1-P)(1-\pi) \\ \Pr(X_i = 0) = 1-\lambda = P(1-\pi) + (1-P)\pi \end{cases}$$

y siendo n_1 el número total de respuestas afirmativas, en una muestra aleatoria de tamaño n:

$$\hat{\lambda} = \frac{n_1}{n} \quad \text{es el estimador insesgado}$$

de λ, con varianza $\lambda\,(1-\lambda)/n$ en muestreo con reposición (multiplicaríamos por el factor de finitud en muestreo sin reposición).

Despejando de la ecuación:

$$\hat{\lambda} = P\,\hat{\pi} + (1-P)(1-\hat{\pi})$$

se obtiene el estimador de π:

$$\hat{\pi} = \frac{1}{2P-1}(P-1+\hat{\lambda}) = \frac{1}{2P-1}\left(P-1+\frac{n_1}{n}\right)$$

donde se observa que la probabilidad conocida P que el mecanismo aleatorio concede a la pregunta íntima, debe ser $P \neq 1/2$ para evitar la indefinición de $\hat{\pi}$. Debe advertirse que $\hat{\pi}$ puede tomar valores fuera del intervalo [0, 1] ; esto ocurre si $\hat{\lambda} < 1 - P$, ó si $\lambda > P$, Singh (1976)

propone el siguiente estimador:

$$\hat{\pi}_s = \begin{cases} 0 & \text{Si} \quad \dfrac{n_1}{n} < 1 - P \\ \hat{\pi} & \text{Si} \quad 1 - P \leqslant \dfrac{n_1}{n} \leqslant P \\ 1 & \text{Si} \quad \dfrac{n_1}{n} > P \end{cases}$$

La esperanza y varianza del estimador de Warner, son:

$$E(\hat{\pi}) = \pi \; ; \quad V(\hat{\pi}) = \frac{1}{(2P-1)^2} \quad V(\hat{\lambda}) = \frac{\lambda(1-\lambda)}{n(2P-1)^2}$$

También, sustituyendo λ en función de π:

$$V(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2}$$

donde se observa que la varianza de $\hat{\pi}$, tiene dos sumandos, el primero representa la varianza que tendriamos haciendo a todas las unidades de la muestra la pregunta íntima y el segundo es el incremento que experimenta como consecuencia de la aleatorización de la pregunta (tributo que se paga para recoger información más veraz).

MODELO CON PREGUNTAS NO RELACIONADAS

En el anterior modelo de Warner la segunda pregunta es complementaria de la primera. Horvitz, Shah y Simmons (1967) proponen un modelo basado en una idea de éste último, en que las preguntas no están relacionadas; la primera es la pregunta íntima y la segunda se denomina pregunta intrascendente (por ejemplo ser fumador, ser lector de un periódico, etc. supuesto que esta caracteríatica irrelevante no esté relacionada con la pregunta íntima). Si la proporción π_y de unidades pertenecientes a la clase intrascendente, es conocida, tenemos:

$$\left\{ \begin{array}{l} \Pr\ (X_i = 1) = \lambda = P\pi + (1-P)\pi_y \\ \Pr\ (X_i = 0) = 1-\lambda = P(1-\pi) + (1-P)(1-\pi_y) \end{array} \right.$$

Operando como en el modelo anterior, se obtiene el estimador insesgado:

$$\hat{\pi} = \frac{1}{P}\left((P-1)\pi_y + \hat{\lambda} \right) = \frac{1}{P}\left((P-1)\pi_y + \frac{n_1}{n} \right)$$

cuya varianza y su estimador insesgado, son:

$$V(\hat{\pi}) = \frac{\lambda(1-\lambda)}{n\,P^2} \quad ; \qquad \hat{V}(\hat{\pi}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)\,P^2}$$

Si π_y es desconocido, se toman dos muestras aleatorias de tamaños n_1, n_2; a la primera se le plantea la pregunta aleatorizada con un mecanismo de probabilidad P_1 y a la segunda con un mecanismo de probabilidad P_2; entonces:

$$\left. \begin{array}{l} \lambda_1 = P_1\,\pi + (1-p_1)\,\pi_y \\ \lambda_2 = P_2\,\pi + (1-P_2)\,\pi_y \end{array} \right\}$$

de donde podemos escribir las ecuaciones empíricas:

$$\left. \begin{array}{l} \hat{\lambda}_1 = P_1\,\hat{\pi} + (1-P_1)\,\hat{\pi}_y \\ \hat{\lambda}_2 = P_2\,\hat{\pi} + (1-P_2)\,\hat{\pi}_y \end{array} \right\}$$

y de este sistema, dados P_1, P_2 a priori, y $\hat{\lambda}_1, \hat{\lambda}_2$ por las observaciones muestrales podemos despejar $\hat{\pi}, \hat{\pi}_y$:

$$\left. \begin{array}{l} \hat{\pi} = \dfrac{\hat{\lambda}_1(1-P_2) - \hat{\lambda}_2(1-P_1)}{P_1 - P_2} \\ \hat{\pi}_y = \dfrac{P_2\,\hat{\lambda}_1 - P_1\,\hat{\lambda}_2}{P_2 - P_1} \end{array} \right\}$$

donde observamos que debe exigirse $P_1 \neq P_2$. Los estimadores son insesgados, pero estamos interesados en $\hat{\pi}$ solamente; su varianza es:

$$V(\hat{\pi}) = \left(\frac{1-P_2}{P_1-P_2}\right)^2 V(\hat{\lambda}_1) + \left(\frac{1-P_1}{P_1-P_2}\right)^2 V(\hat{\lambda}_2) =$$

$$= \left(\frac{1-P_2}{P_1-P_2}\right)^2 \frac{\lambda_1(1-\lambda_1)}{n_1} + \left(\frac{1-P_1}{P_1-P_2}\right)^2 \frac{\lambda_2(1-\lambda_2)}{n_2}$$

y su estimador insesgado:

$$\hat{V}(\hat{\pi}) = \left(\frac{1-P_2}{P_1-P_2}\right)^2 \frac{\hat{\lambda}_1(1-\hat{\lambda}_2)}{n_1-1} + \left(\frac{1-P_1}{P_1-P_2}\right)^2 \frac{\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2-1}$$

MODELO PARA VARIABLE CONTINUA

Si la pregunta íntima se refiere a una variable $\mathfrak{X}$ que tiene una función de distribución $F(x)$ y la pregunta intrascendente una distribución $H(x)$, la respuesta obtenida es una variable aleatoria $\mathfrak{Z}$ cuya distribución es:

$$\left.\begin{aligned} \phi_1(z) &= \Pr(Z \leqslant z)_1 = P_1 F(z) + (1-P_1) H(z) \\ \phi_2(z) &= \Pr(Z \leqslant z)_2 = P_2 F(z) + (1-P_2) H(z) \end{aligned}\right\}$$

Operando de forma análoga al caso anterior, las medias de $(Z)_1$ y de $(Z)_2$ verifican:

$$\left.\begin{aligned} \mu_1 &= P_1 \mu + (1-P_1) \mu_y \\ \mu_2 &= P_2 \mu + (1-P_2) \mu_y \end{aligned}\right\}$$

donde μ_y es la media correspondiente a la distribución H y μ la que deseamos estimar, correspondiente a la distribución F. se obtiene:

$$\hat{\mu} = \frac{(1 - P_2)\,\bar{z}_1 - (1 - P_1)\,\bar{z}_2}{P_1 - P_2}$$

siendo $\bar{z}_1$, $\bar{z}_2$ las medias de la primera y segunda muestra. La varianza viene dada por:

$$V(\hat{\mu}) = \left(\frac{1 - P_2}{P_1 - P_2}\right)^2 V(\bar{z}_1) + \left(\frac{1 - P_1}{P_1 - P_2}\right)^2 V(\bar{z}_2)$$

de la que se forma un estimador insesgado, sustituyendo:

$$V(\bar{z}_1) = \frac{s_1^2}{n_1} \quad ; \qquad V(\bar{z}_2) = \frac{s_2^2}{n_2}$$

Por último conviene decir que ha habido una gran profusión de estudios sobre este tipo de modelos, tratando K>2 preguntas, estudiando las probabilidades óptimas, etc. Un interesante trabajo de recopilación ha sido realizado en España por Ladoux, M. (1982), al que remitimos al lector.

IX.2. EFECTO DEL ENTREVISTADOR

Si bien es cierto que un censo o encuesta efectuado mediante entrevistas directas aporta una mayor calidad de datos que la "autoenumeración", y esto se debe a que disminuye la no respuesta y a que el entrevistador puede ayudar decisivamente a una correcta interpretación del cuestionario, especialmente cuando éste es complicado e incluye preguntas alternativas en función de la respuesta dada a las anteriores, también es cierto que distintos entrevistadores pueden ejercer efectos distintos sobre las respuestas. Esto puede ser debido a diversas causas derivadas de la conducta individual al efectuar la entrevista y del diferente grado de adiestramiento. Los cursos de formación tratan de conseguir uniformidad de criterio de los entrevistadores, pero a pesar todo, en la práctica, cabe esperar que permanezca en mayor o menor grado, un efecto sobre la respuesta producido por el entrevistador.

IX.2.1. MODELO PARA EL EFECTO ENTREVISTADOR

Supongamos una población finita de N unidades y un conjunto de L entrevistadores. Denotemos por Y_{ih} el dato recogido por el h° entrevistador en la unidad A_i ; denotemos por X_i el "verdadero valor" de la variable en la unidad A_i . Se denomina:

$$d_{ih} = Y_{ih} - X_i$$

Desviación de respuesta correspondiente al entrevistador h cuando observa la unidad A_i. En muestreo aleatorio simple, tenemos:

$$\underset{h}{E}\ \underset{i}{E}(Y_{ih}) = \underset{h}{E}(\bar{Y}_h) = \bar{Y}$$

donde $\bar{Y}_h = \underset{i}{E}(Y_{ih})$ es la media esperada de la población si es entrevistada por el h°; $\bar{Y}$ es la media esperada sobre todos los entrevistadores. La varianza de $\bar{Y}_{ih}$ puede descomponerse de la siguiente forma:

$$\sigma_y^2 = \underset{h}{E}\ \underset{i}{E}(Y_{ih} - \bar{Y})^2 = \underset{h}{E}\ \underset{i}{E}(Y_{ih} - \bar{Y}_h)^2 + \underset{h}{E}\ \underset{i}{E}(\bar{Y}_h - \bar{Y})^2$$

ya que se anula la esperanza:

$$\underset{h}{E}\ \underset{i}{E}(Y_{ih} - \bar{Y}_h)(\bar{Y}_h - \bar{Y}) = \underset{h}{E}(\bar{Y}_h - \bar{Y})\ \underset{i}{E}(Y_{ih} - \bar{Y}_h)$$

así, podemos escribir:

$$\sigma_y^2 = \underset{h}{E}\ \underset{i}{E}(Y_{ih} - \bar{Y}_h)^2 + \underset{h}{E}(\bar{Y}_h - \bar{Y})^2$$

donde el primer sumando:

$$\sigma_w^2 = \underset{h}{E}\ \underset{i}{E}(Y_{ih} - \bar{Y}_h)^2 = \underset{h}{E}(\sigma_h^2)$$

se denomina Varianza "dentro" de entrevistadores y representa la varianza promedio de las desviaciones de respuesta de cada entrevistador; el segundo sumando:

$$\sigma_b^2 = E\left(\bar{Y}_h - \bar{Y}\right)^2$$

se denomina Varianza "entre" entrevistadores y representa la varianza de los promedios $\bar{Y}_h$ obtenidas por cada entrevistador; es una medida del efecto del entrevistador ya que si todos obtienen el mismo valor esperado $\bar{Y}_h$ no hay efecto entrevistador y σ_b^2 es cero, por el contrario si al menos un entrevistador obtiene un promedio $\bar{Y}_h$ distinto a los demás $\sigma_b^2 > 0$. Así pues:

$$\sigma_y^2 = \sigma_w^2 + \sigma_b^2$$

Veamos ahora los distintos casos hipotéticos que recoge el modelo:

1. Todos los entrevistadores recogen datos exactos de todas las unidades.

   $Y_{ih} = X_i$ para todo i, h; por tanto:

   $$d_{ih} = 0 \quad ; \quad \bar{Y}_h = \bar{Y} = \bar{X} \quad \text{y} \quad \sigma_b^2 = 0.$$

   No existe efecto entrevistador y la varianza coincide con la varianza poblacional:

   $$\sigma_y^2 = \underset{h}{E}\,\underset{i}{E}\left(X_i - \bar{X}\right)^2 = \underset{i}{E}\left(X_i - \bar{X}\right)^2 = \sigma_x^2$$

   En este caso, por ser $\bar{Y} = \bar{X}$, tampoco hay sesgo.

2. Los errores o desviaciones de respuesta se compensan dentro de cada entrevistador.

   $\underset{i}{E}\left(d_{ih}\right) = \underset{i}{E}\left(Y_{ih} - X_i\right) = 0$, para todo h. En consecuencia los valores esperados de cada entrevistador coinciden con la verdadera media:

   $$E\left(Y_{ih}\right) = E\left(X_i + d_{ih}\right) = E\left(X_i\right) = \bar{X} \quad ; \quad \bar{Y}_h = \bar{Y} = \bar{X}$$

por tanto $\sigma^2 = 0$ y no existe efecto entrevistador. A diferencia del caso 1) la varianza del modelo aumenta debido a la variabilidad de la respuesta de los entrevistadores:

$$\sigma_y^2 = \sigma_w^2 = E\,(Y_{ih} - \bar{Y}_h)^2 = E(d_{ih} + X_i - \bar{Y}_h)^2 =$$

$$= E\,(d_{ih} + X_i - \bar{X})^2 =$$

$$= E\,(d_{ih}^2) + E\,(X_i - \bar{X})^2 + 2\,E\left\{d_{ih}(X_i - \bar{X})\right\}$$

de donde resulta que la varianza total del modelo es:

$$\sigma_y^2 = \sigma_w^2 = \sigma_x^2 + \sigma_d^2 + 2\,\sigma_{xd}$$

donde σ_x^2 es la varianza poblacional; σ_d^2 es la varianza de respuesta; σ_{xd} es la covarianza entre desviaciones de respuesta y de muestreo. En resumen, no existe efecto entrevistador ($\sigma_b^2 = 0$), no existe sesgo y la varianza del modelo aumenta debido a la variabilidad de las respuestas (centradas en media) de los entrevistadores y a la posible existencia de una covarianza entre desviaciones de respuesta y desviaciones de muestreo ($X_i - \bar{X}$).

3. Todos los entrevistadores recogen datos cuyos promedios tienen el mismo sesgo.
En este caso, $Y_h = Y$ para todo h, por tanto:

$$E\,(d_{ij}) = E\,(Y_{ih}) - \bar{X} = \bar{Y}_h - \bar{X} = \bar{Y} - \bar{X} = B\ (\text{sesgo}).$$

y en consecuencia no existe un efecto entrevistador ya que:

$$\sigma_b^2 = E\,(Y_h - Y)^2 = E(0) = 0$$

pero existe un sesgo $B = \bar{Y} - \bar{X}$, y la varianza del modelo aumenta, respecto a la varianza poblacional de la variable σ_x^2:

$$\sigma_y^2 = \sigma_w^2 = E\,(Y_{ih} - \bar{Y}_h)^2 = E\,(Y_{ih} - \bar{Y})^2 =$$

$$= E\,(d_{ih} + X_i - \bar{Y})^2 = E\,(d_{ih} + (X_i - \bar{X}) + B)^2 =$$

$$= \sigma_d^2 + \sigma_x^2 + 2\,\sigma_{dx} + B^2$$

Se diferencia del caso anterior solamente en la aparición del cuadrado del sesgo.

4. El valor promedio recogido por al menos un entrevistador, es distinto al obtenido por los demás.

   En este caso, por ser $\bar{Y}_h \neq \bar{Y}$, al menos para un h, resulta $\sigma_b^2 = E\,(\bar{Y}_h - \bar{Y})^2 > 0$ y aparece el efecto entrevistador. El sesgo puede existir o no según que se compensen o no, los promedios $\bar{Y}_h$; así pues si $E\,(\bar{Y}_h) = \bar{Y} = \bar{X}$ no hay sesgo, pero si $E\,(\bar{Y}_h) = \bar{Y} \neq \bar{X}$, aparecerá un sesgo $B = \bar{Y} - \bar{X}$.

El caso 1) representa la situación ideal donde no hay sesgo, no hay varianza de respuesta y no hay efecto de entrevistador; la varianza total coincide con la varianza poblacional de la variable $\sigma_y^2 = \sigma_x^2$. Los casos 2) y 3) representan situaciones en que hay varianza de respuesta, covarianza entre desviaciones de respuesta y de muestreo, puede haber o no (caso 2 ó caso 3) sesgo y no hay efecto de entrevistador; se trata de situaciones que pueden presentarse en encuestas con entrevistadores o sin entrevistadores ya que también las instrucciones y el cuestionario pueden inducir a los encuestados a producir sesgos y variabilidad en las respuestas. Así pues el caso 4) representa la situación donde hay efecto entrevistador, ya que no todos recogerían el mismo valor promedio.

IX.2.2. ESTIMACION Y CONTRASTE DEL EFECTO ENTREVISTADOR

Planteamos la experimentación de la siguiente forma:

a) Se toman l muestras aleatorias de tamaño m, de modo que la muestra total es de tamaño $n = l.m$.

b) Se asignan aleatoriamente uno de cada l entrevistadores a una submuestra.

Para detectar la presencia del efecto entrevistador procesamos los resultados de cada submuestra; si no existe tal efecto, la varianza estimada a partir de la variabilidad "entre" submuestras (entrevistadores) no ha de ser significativamente diferente de la varianza estimada a partir de la variabilidad "dentro" de submuestras. Se aplica entonces el modelo de análisis de la varianza para el contraste de la hipótesis $\sigma_b^2 = 0$ y si ésta se rechaza, el efecto entrevistador puede medirse mediante la estimación $\hat{\sigma}_b^2$, o bien mediante el porcentaje que supone en la varianza total. A continuación escribimos el cuadro de análisis de la varianza:

Fuente de variación	$g.l.$	Cuadrado medio
Entre	$l-1$	$A = \dfrac{m \sum_{h=1}^{l} (\bar{y}_h - \bar{y})^2}{l-1}$
Dentro	$l(m-1)$	$B = \dfrac{\sum_{h}^{l} \sum_{i}^{m} (y_{ih} - \bar{y}_h)^2}{l(m-1)}$

Para contrastar si A y B difieren significativamente se emplea la razón F = A/B que si la hipótesis nula, $\sigma_b^2 = 0$, es cierta se distribuye según una F de Snedecor $\left(l-1,\ l(m-1)\right)$. Si se acepta la hipótesis nula, la varianza de $\bar{y}$ se estima mediante el cuadrado medio total:

$$\hat{\sigma}^2_{\bar{y}} = \frac{1}{n(n-1)} \sum_{h=1}^{l} \sum_{i=1}^{m} (y_{ih} - \bar{y})^2$$

Si la hipótesis nula se rechaza, la varianza total de la media viene dada por:

$$\sigma^2_{\bar{y}} = \frac{\sigma_b^2}{l} + \frac{\sigma_w^2}{lm}$$

y el estimador insesgado viene dado por:

$$\hat{\sigma}^2_{\bar{y}} = \frac{1}{l}\left(\frac{A-B}{m}\right) + \frac{B}{lm}$$

y dado que B estima a σ^2_w, resulta que $\left(\frac{A-B}{m}\right)$ estima a σ^2_b . En consecuencia la medida del efecto entrevistador viene dada por:

$$\hat{\sigma}^2_b = \frac{A-B}{m} \quad ; \quad \hat{\sigma}^2_w = B$$

o más significativamente por el porcentaje que representa de la varianza total:

$$\frac{\frac{\hat{\sigma}^2_b}{l}}{\hat{\sigma}^2_{\bar{y}}} = \frac{A-B}{A} \times 100$$

X. MODELOS DEL ERROR TOTAL EN CENSOS Y ENCUESTAS

En el conjunto del proceso estadístico de un censo o una encuesta, intervienen diversos elementos (instrumentos y operaciones) que son fuentes potenciales de error. Además de los sesgos que puede producir un cuestionario mal diseñado, que en parte puede ser debido a una imprecisa definición de las unidades y conceptos, las operaciones de recogida, transcripción y grabación de los datos también pueden dar lugar a que se produzcan desviaciones de los valores observados respecto de los que podemos llamar para entendernos, verdaderos valores. Por otra parte no debemos olvidar los efectos debidos al trabajo de los encuestadores, a la situación objetiva o subjetiva de los entrevistados y a la interacción entre unos y otros. Tales errores se denominan "ajenos al muestreo" para distinguirlos de la variación que distintas muestras presentan, aún en ausencia de aquellos, que se denomina "error de muestreo" o desviación típica del estimador. De acuerdo con estas ideas es claro que un censo si bien no está sometido al error de muestreo sufre el efecto de los errores ajenos al muestreo; es más, la experiencia muestra que deben esperarse mayores errores que en una encuesta ya que debido al mayor número de operaciones y a la cantidad de personas que intervienen, es más difícil mantener bajo control la calidad de los trabajos.

Algunos investigadores, como Hansen, Hurwitz y Bershad, y otros, han mostrado interés por la construcción de modelos para describir el Error Cuadrático Medio de una estimación sometida a errores de muestreo y ajenos al muestreo.

X.1. FUNDAMENTO DE LOS MODELOS

Se utiliza el concepto de *superpoblación*, que consiste en lo siguiente: En la teoría ordinaria suponemos que para una población finita dada

$$\mathcal{A} = \{A_2, \ldots, A_i, \ldots, A_N\}$$

está definido unívocamente el conjunto de valores de una característica $\mathfrak{X}$; $\mathfrak{X}(A_i) = X_i$ y se supone que al observar cualquier unidad A_i obtenemos el valor X_i. Para representar el hecho de que, debido a los errores ajenos al muestreo en vez de X_i observamos en A_i un valor que puede coincidir o no con X_i, se supone que tenemos una población de variables aleatorias:

$$\mathfrak{y} = \{\mathfrak{y}_1, \mathfrak{y}_2, \ldots, \mathfrak{y}_i, \ldots, \mathfrak{y}_N\}$$

y que al realizar un censo o encuesta obtenemos una realización particular, t, que representamos por $Y_{1t}, Y_{2t}, \ldots, Y_{it}, \ldots, y_{Nt}$.

En estas condiciones se supone que el valor esperado de la observación efectuada en una unidad A_i es:

$$\underset{i}{E}(Y_{it}) = Y_i \qquad \text{(en general } Y_i \neq X_i\text{)}$$

y denotando por $\overline{Y}$ la media esperada en $\mathfrak{y}$:

$$E(Y_{it}) = E(E(Y_{it})) = E(Y_i) = \overline{Y}$$

La diferencia $d_{it} = Y_{it} - Y_i$ se denomina *desviación de respuesta* de la unidad A_i en la realizacion t. Para cada i, d_{it} es una variable con esperanza cero, por tanto:

$$V(d_{it}) = E(d_{it}^2) = \sigma_R^2$$

y se denomina *varianza unitaria de respuesta.*

La diferencia $\delta_i = Y_i - \bar{Y}$ se denomina *desviación de muestreo* de la unidad A_i. Su esperanza es cero, por tanto:

$$V(\delta_i) = E(\delta_i^2) = \sigma_M^2$$

y se denomina *varianza unitaria de muestreo.*

La diferencia $B_i = Y_i - X_i$ se denomina *sesgo de respuesta* de la unidad A_i . En general cabe suponer que los sesgos individuales no se compensan, de modo que:

$$B = E(Y_i) - E(X_i) = \bar{Y} - \bar{X}$$

es distinto de cero. Se denomina *sesgo total de respuesta.*

Cuando se realiza un censo, o una investigación muestral con $n > 1$ unidades, debemos suponer que en general las desviaciones de respuesta en dos unidades A_i, A_j $(i \neq j)$ están correlacionadas. Definimos el coeficiente de correlación entre desviaciones de respuesta como:

$$\rho_R = \frac{E(d_{it} \cdot d_{jt})}{\sigma_R^2} \quad ; \quad (i \neq j)$$

Análogamente debemos suponer que las desviaciones de respuesta y de muestreo en una misma unidad A_i también pueden estar correlacionadas. Definimos entonces el coeficiente de correlación:

$$\rho_{RM} = \frac{E(d_{it} \cdot \delta_i)}{\sigma_R \, \sigma_M}$$

X.2. EL ERROR TOTAL EN UN CENSO

Supongamos que el objetivo es $\bar{X}$; una realización particular nos

dará una media censal:

$$\bar{Y}_t = \frac{1}{N} \sum_{i=1}^{N} Y_{it} .$$

cuyo Error Cuadrático Medio pretende explicar el modelo.

$$ECM(\bar{Y}_t) = E\left(\frac{1}{N} \sum_{1}^{N} Y_{it} - \bar{X} \right)^2 =$$

$$= E\left(\frac{1}{N} \sum_{1}^{N} Y_{it} - \bar{Y} + \bar{Y} - \bar{X} \right)^2 =$$

$$= E\left(\frac{1}{N} \sum_{1}^{N} Y_{it} - \bar{Y} \right)^2 + E(\bar{Y} - \bar{X})^2 + 2E\left(\frac{1}{N} \sum_{1}^{N} (Y_{it} - \bar{Y})(\bar{Y} - \bar{X}) \right)$$

donde el tercer sumando se anula por ser $B = \bar{Y} - \bar{X}$ constante y $E(Y_{it}) = \bar{Y}$. Resulta que el ECM tiene dos componentes:

$$ECM(\bar{Y}_t) = E\left(\frac{1}{N} \sum_{1}^{N} Y_{it} - \frac{1}{N} \sum_{1}^{N} Y_i \right)^2 + B^2 =$$

$$= E\left(\frac{1}{N} \sum_{1}^{N} d_{it} \right)^2 + B^2$$

la primera es la Varianza censal de Respuesta y la segunda es el cuadrado del sesgo. Veamos ahora el análisis de la varianza de respuesta:

$$E\left(\frac{1}{N} \sum_{1}^{N} d_{it} \right)^2 = \frac{1}{N^2} E\left(\sum_{1}^{N} d_{it}^2 + \sum_{i \neq j}^{N} d_{it}\, d_{jt} \right) =$$

$$= \frac{1}{N^2} N \sigma_R^2 + \frac{1}{N^2} N(N-1) \rho_R \sigma_R^2 = \frac{1}{N} \sigma_R^2 + \frac{N-1}{N} \rho_R \sigma_R^2 =$$

$$= \frac{\sigma_R^2}{N} \quad (1 + (N-1)\ \rho_R)$$

En resumen, el ECM de la media censal tiene dos componentes: El cuadrado del sesgo y la varianza de respuesta, ésta a su vez se compone de dos sumandos, la varianza de respuesta simple y la varianza, o componente, correlacionada de respuesta.. Es interesante observar que si la población es grande, la varianza de respuesta simple podría tener poca importancia ya que aparece N en su denominador, pero no ocurre lo mismo con la componente correlacionada que suponiendo N = N−1 sería igual a $\rho_R \ \sigma_R^2$.

$$ECM\ (\bar{Y}_t) = \frac{\sigma_R^2}{N} + \frac{(N-1)}{N}\ \rho_R\ \sigma_R^2 + B^2$$

El sesgo de respuesta es el promedio (valor esperado) de los sesgos individuales y se asocia no a la variabilidad de las observaciones alrededor de su promedio, sino a características del procedimiento independientes de los errores accidentales. Cabe suponer en un modelo más simple que el sesgo es constante para todas las unidades, $B_i = B$.

X.3. EL ERROR TOTAL EN UNA ENCUESTA

Supongamos ahora que la investigación acerca de la media poblacional, se reduce a la media de una muestra aleatoria de tamaño n:

$$\bar{y}_t = \frac{1}{n} \sum_{i=1}^{n} Y_{it}$$

cuyo ECM vamos a calcular. Utilizando la notación:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} \underset{i}{E}\,(Y_{it}) = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

podemos escribir:

$$ECM(\bar{y}_t) = E(\bar{y}_t - \bar{X})^2 = E(\bar{y}_t - \bar{y} + \bar{y} - \bar{Y} + \bar{Y} - \bar{X})^2 =$$

$$= E(\bar{y}_t - \bar{y})^2 + E(\bar{y} - \bar{Y})^2 + 2E(\bar{y}_t - \bar{y})(\bar{y} - \bar{Y}) + E(\bar{Y} - \bar{X})^2$$

teniendo en cuenta que las restantes esperanzas de productos cruzados se anulan. Veamos por separado las componentes del $ECM(\bar{y}_t)$: La primera es la varianza de respuesta, la segunda es la varianza de muestreo, la tercera es debida a la correlación entre desviaciones de respuesta y desviaciones de muestreo en una misma unidad y la última es el cuadrado del sesgo.

Componente debida a la varianza de respuesta:

$$E(\bar{y}_t - \bar{y})^2 = E\left(\frac{1}{n} \sum_{i=1}^{n} (Y_{it} - Y_i)\right)^2 = E\left(\frac{1}{n} \sum_{i=1}^{n} d_{it}\right)^2 =$$

$$= \frac{1}{n^2}\left\{ n\,E(d_{it}^2) + n(n-1)\,\text{Cov.}(d_{it}, d_{jt}) \right\} =$$

$$= \frac{\sigma_R^2}{n} + \frac{n-1}{n}\,\sigma_R^2\,\rho_R = \frac{\sigma_R^2}{n}\left(1 + (n-1)\,\rho_R\right)$$

que consta de la componente debida a la varianza de respuesta simple y de la componente debida a la correlación entre desviaciones de respuesta de distintas unidades.

Componente debida a la varianza de muestreo:

$$E(\bar{y} - \bar{Y})^2 = \frac{\sigma_M^2}{n}$$

Componente debida a la correlación entre desviaciones de muestreo y de respuesta en una misma unidad:

$$2\,E\,(\bar{y}_t - \bar{y})\;(\bar{y} - Y) = \frac{2}{n}\;\rho_{RM}\;\sigma_R\;\sigma_M$$

En resumen, cuando se estima la media poblacional $\bar{X}$, mediante la media de una muestra aleatoria simple de tamaño n, el Error Cuadrático Medio admite la descomposición.

$$ECM(\bar{y}_t) = \frac{\sigma_R^2}{n}\left(1 + (n-1)\,\rho_R\right) + \frac{\sigma_M^2}{n} + \frac{2}{n}\;\rho_{RM}\,\sigma_R\,\sigma_M + B^2$$

Observamos que respecto de una investigación censal, aparecen dos nuevos términos: La varianza debida al muestreo y la componente debida a la correlación entre desviaciones de respuesta y de muestreo en una misma unidad; esta última se anula en repeticiones sobre una muestra fija. Por otra parte al aumentar el tamaño de la muestra, obviamente la componente debida a la varianza de muestreo disminuye hasta anularse en un censo; lo mismo ocurre con la componente de correlación entre desviación de respuesta y de muestreo. De las restantes, el sesgo es siempre constante y la componente de la varianza de respuesta correlacionada también permanece o incluso aumenta.

EFECTO DE LOS ENTREVISTADORES EN LA VARIANZA DE RESPUESTA

Si se admite que la correlación entre desviaciones de respuesta es debida fundamentalmente al efecto del entrevistador, la fórmula de la

$$\frac{(n-1)}{n}\;\rho_R\;\sigma_R^2$$

componente correlacionada se modificaría de la siguiente forma: Admi-

tiendo que solo están correlacionadas las desviaciones de respuesta entre unidades observadas por un mismo entrevistador, al efectuar el cálculo de la componente correlacionada, para K entrevistadores con $\bar{n}$ unidades cada uno ($N = K \cdot \bar{n}$), se tiene:

$$\frac{1}{n^2} E \sum_{i \neq j}^{n} d_{it} d_{jt} = \frac{1}{n^2} K \bar{n} (n-1) \rho_R \sigma_R^2 = \frac{(\bar{n}-1)}{n} \rho_R \sigma_R^2$$

por tanto al aumentar el tamaño de la muestra, bajo estas hipótesis, la componente correlacionada también podría disminuir si el número de unidades observadas por cada entrevistador es constante.

X.4. APLICACION DEL MODELO A UNA PROPORCION

Si la encuesta se dirige a estimar una proporción, las siguientes notaciones y fórmulas se deducen inmediatamente, teniendo en cuenta que ahora la variable observada Y_{it} toma el valor uno ó cero, según que en la realización t se clasifique la unidad en la clase de referencia o en su complementaria:

$$\underset{i}{E}(Y_{it}) = P_i \quad ; \quad \bar{Y} = \frac{1}{N} \sum_{1}^{N} P_i = P$$

$$\sigma_R^2 = V(d_{it}) = \frac{1}{N} \sum_{1}^{N} P_i Q_i \quad ; \quad \sigma_M^2 = \frac{1}{N} \sum_{1}^{N} (P_i - P)^2$$

La proporción muestral observada en la realización t, y la media muestral de valores esperados serán :

$$p_t = \frac{1}{n} \sum_{i=1}^{n} Y_{it} \quad ; \quad p = \frac{1}{n} \sum_{i=1}^{n} P_i$$

La varianza de respuesta y la varianza de muestreo:

$$\frac{1}{Nn}\left(\sum_{1}^{N} P_i Q_i\right)(1 + (n-1)\ \rho_R) \quad ; \quad \frac{1}{Nn}\sum_{1}^{N} (P_i - P)^2$$

El sesgo, suponiendo una "verdadera" P_o, será $B = P - P_o$

Una importante propiedad del modelo para una proporción es la siguiente: La varianza de respuesta simple más la varianza de muestreo es igual a PQ/n.

$$\frac{1}{n}\left\{\frac{1}{N}\sum_{1}^{N} P_i(1-P_i) + \frac{1}{N}\sum_{1}^{N} P_i^2 - P^2\right\} = \frac{1}{n}(P - P^2) = \frac{PQ}{n}$$

Esto implica que el estimador usual de la varianza de una proporción incluye la varianza de muestreo y la de respuesta simple:

$$E\left(\frac{pq}{n-1}\right) = \frac{PQ}{n} = \frac{\sigma_R^2}{n} + \frac{\sigma_M^2}{n}$$

X.5. EXPERIMENTACION Y ESTIMACION

En el contexto del modelo estudiado la única alternativa para obtener información acerca del ECM, consiste en repetir la observación en cada una de las unidades de una submuestra aleatoria que supondremos de tamaño m. Esta práctica que se conoce con el nombre de entrevista repetida permite disponer de dos observaciones. Sin entrar de momento en otras consideraciones, admitiremos la hipótesis de que la segunda es una observación independiente de la primera; es decir, sisndo Y_{it} e $Y_{it'}$ las dos realizaciones en una misma unidad, suponemos su independencia probabilística.

ESTIMACION DE LA VARIANZA DE RESPUESTA SIMPLE

Vamos a ver que siendo $G = \frac{1}{m} \sum_{i=1}^{m} (Y_{it} - Y_{it'})^2$ se cumple que $\frac{1}{2}$ G es un estimador insesgado de σ_R^2

$$E(G) = \frac{1}{m}\, m\, E(Y_{it} - Y_{it'})^2 = E(Y_{it} - Y_i + Y_i - Y_{it'})^2$$

$$= E(d_{it}^2 + d_{it'}^2 - 2E(d_{it}\, d_{it'}) = 2\,\sigma_R^2$$

puesto que $E(d_{it}) = E(d_{it'}) = \sigma_R^2$ y por otro lado, la independencia de d_{it} e $d_{it'}$ y por ser de media cero, implica la anulación del tercer término.

ESTIMACION DE LA VARIANZA TOTAL DE RESPUESTA

Veamos ahora que siendo $J = \left\{ \frac{1}{m} \sum_{i=1}^{m} (Y_{it} - Y_{it'}) \right\}^2$ se cumple que $\frac{1}{2}$ J es un estimador insesgado de $\sigma_R^2/m + (m-1)\,\rho_R\,\sigma_R^2/m$.

$$E(J) = \frac{1}{m^2}\, E \left\{ \sum_{i=1}^{m} (d_{it} - d_{it'})^2 + \sum_{j \neq i}^{m} (d_{it} - d_{it'})(d_{jt} - d_{jt'}) \right\}$$

el primer término ya lo calculamos antes, siendo igual $2\,\sigma_R^2$; en cuanto al segundo teniendo en cuenta que $E(d_{it}\, d_{jt}) = E(d_{it'}\, d_{jt'}) = \rho_R\,\sigma_R^2$, y que las esperanzas de los otros dos productos se anulan, resulta:

$$E(J) = \frac{2\,\sigma_R^2}{m} + \frac{2\,(m-1)}{m}\,\rho_R\,\sigma_R^2$$

Cuando se trata de una variable cualitativa binaria, podemos presen-

tar los resultados de las dos realizaciones (Original y Repetida) de la siguiente forma:

Entrevista repetida	Entrevista original		TOTAL
	$\mathcal{C}$	$\bar{\mathcal{C}}$	
$\mathcal{C}$	a	b	a + b
$\bar{\mathcal{C}}$	c	d	c + d
TOTAL	a + c	b + d	m

donde "a" es el número de unidades igualmente clasificadas en las dos entrevistas en la clase $\mathcal{C}$; "b" es el de unidades clasificadas en $\mathcal{C}$ en la repetida y en $\bar{\mathcal{C}}$ en la original. etc. Se comprueba fácilmente que

$$G = \frac{c + b}{m} \quad ; \text{se denomina Indice de cambio bruto.}$$

$$\sqrt{J} = \frac{(c - b)}{m} \quad ; \text{se denomina Indice de cambio neto.}$$

$$\frac{p_t q_t}{m-1} = \frac{(a + c)(b + d)}{(m-1)\, m^2} \quad ; \text{estimador de P Q / m}$$

Por último, de acuerdo con los resultados teóricos obtenidos, podemos establecer el siguiente cuadro de estimadores:

Estimador	Valor esperado
$T_1 = \frac{(c + b)}{2\, m^2}$	$\frac{\sigma_R^2}{m}$
$T_2 = \frac{(c - b)^2}{2\, m^2}$	$\frac{\sigma_R^2}{m} + \frac{\sigma_R^2}{m} (m-1)\, \rho_R$
$T_3 = \frac{(a + c)(b + d)}{(m-1)\, m^2}$	$\frac{\sigma_R^2}{m} + \frac{\sigma_M^2}{m}$

de donde se deduce:

T_1	estima la componente debida a la varianza de respuesta simple
$T_2 - T_1$	estima la componente debida a la correlación
$T_3 - T_1$	estima la componente debida al muestreo
$\frac{T_2 - T_1}{T_1(m-1)}$	estima el coeficiente de correlación entre desviaciones de respuesta

Algunas reflexiones finales son convenientes. Si bien el modelo del ECM presenta una explicación en términos de instrumentos teóricos usuales, como son sesgo, varianzas y covarianzas, presenta dificultades prácticas de estimación. En primer lugar raramente se podrán hacer más de dos observaciones en una misma unidad, con lo cual la estimación de la varianza de respuesta es relativamente débil. Por otra parte no siempre podemos estar seguros de que la entrevista repetida es una realización, t', independiente de la original, t, en cada unidad muestral; si no es así G/2 subestima a la varianza de respuesta simple, ya que aparece una covarianza entre d_{it}, $d_{it'}$ que podría ser positiva; en la práctica se recomienda que el segundo entrevistador no conozca los resultados de la primera entrevista. Por lo que se refiere al sesgo, para hacer algún tipo de inferencia es necesario disponer de una muestra de "contraste" en cuyas unidades conociésemos los verdaderos valores.

Recordemos por último que, como hemos visto en el caso de una proporción, el estimador usual de la varianza recoge la componente debida al muestreo y la debida a la varianza de respuesta simple, por tanto nuestras apreciaciones acerca de la precisión de una encuesta deberán corregirse "al alza" si la componente correlacionada es grande. La experiencia indica que en características que se recogen con bastante exactitud, como edad, rama de actividad, etc. σ_R^2 puede oscilar entre un 3 y un 10 por ciento de PQ, sin embargo en otras como Número de horas efectivamente trabajadas en una semana, puede superar el 50 por ciento. Pero como hemos visto es la componente correlacionada a la que debe prestarse atención ya que puede superar ampliamente a la de respuesta simple y, además, no está incluída en el estimador usual de la varianza. En los ejemplos X.3 y X.4 se muestran dos casos interesantes, uno con muy grande componente correlacionada y otro en la que es pequeña.

X.6. EJEMPLOS

Ejemplo X.1

Compruebe que en el caso de una variable binaria, el índice de cambio bruto, G, es igual a $1/n \sum (Y_{it} - Y_{it'})^2$.

Solución:

Se deduce inmediatamente teniendo en cuenta que $Y_{it}^2 = Y_{it}$ y que $Y_{it} \cdot Y_{it'} = 1$ si ambas valen uno y cero en caso contrario.

Ejemplo X.2

Compruebe que J es igual al cuadrado del índice de cambio neto.

Ejemplo X.3

En el siguiente cuadro se presentan los resultados de la entrevista original y repetida a 2337 Activos Ocupados, correspondiente al 4° trimestre de 1983 de la Encuesta de Población Activa del INE. La clase $\mathbb{C}$ considerada incluye los que han trabajado menos de 40 horas en la semana de referencia. Estime la varianza de respuesta y sus componentes, la varianza debida al muestreo y el coeficiente de correlación entre desviaciones de respuesta.

E.O. / E.R.	< 40 h	⩾ 40 h	TOTAL
< 40 h	261	352	613
⩾ 40 h	238	1486	1724
TOTAL	499	1838	2337

Solución:

Varianza de respuesta	$118{,}976 \times 10^{-5}$
Varianza de respuesta simple	$5{,}401 \times 10^{-5}$
Varianza de respuesta correlacionada	$113{,}575 \times 10^{-5}$
Varianza debida al muestreo	$1{,}787 \times 10^{-5}$
Ceoficiente de correlación	0,009

Ejemplo X.4

El siguiente cuadro también corresponde a la misma encuesta y se refiere a la Rama de Actividad de 2674 activos, siendo Agricultura la clase de referencia. Calcule las mismas estimaciones que en el ejemplo X.3. Intente alguna explicación para resultados tan dispares en uno y otro caso.

E.O. / E.R.	Agricultura	Otra	TOTAL
Agricultura	470	51	521
Otra	40	2113	2153
TOTAL	519	2164	2674

Solución:

Varianza de respuesta	$0{,}84612 \times 10^{-5}$
Varianza de respuesta simple	$0{,}63634 \times 10^{-5}$
Varianza de respuesta correlacionada	$0{,}20978 \times 10^{-5}$
Varianza debida al muestreo	$5{,}13804 \times 10^{-5}$
Coeficiente de correlación	0,000123

BIBLIOGRAFIA

Libros y monografías

Azorín Poch, F.—Curso de Muestreo y aplicaciones. Ed. Aguilar. Madrid.

Cochran, W.G.—Sampling Techniques. Ed. John Wiley.

Cramer, H.—Métodos Matemáticos de Estadística. Ed. Aguilar. Madrid.

Dalenius, T.—Sampling in Sweden: Contributions to the Methods and Theories of Sample Survey Practice. Ed. Almqvist and Wiksell.

Deming, W.E.—Sample Design in Bisiness Research. Ed. John Wiley.

Desabie, J.—Théorie et Practique des Sondages. Ed. Dunod.

Des Raj.—Sampling Theory. Ed. McGraw-Hill.

Hansen, Hurwitz y Madow.—Sample Survey Methods and Theory. Ed. John Wiley.

Kish, L.—Survey Sampling. Ed. John Wiley.

Konijn, H.S.—Statistical Theory of Sample Survey. Ed. North Holland.

Ladoux Aguilar, M.—Modelos de Respuesta Aleatorizada. INE. Madrid.

Ruiz-Maya, L.—Métodos Estadísticos de Investigación (Introducción al Análisis de la Varianza). INE. Madrid.

Sánchez-Crespo.—Curso intensivo de Muestreo en Poblaciones Finitas. INE. Madrid.

Stephan y Mc Carthy.—Sampling Opinions. Ed. John Wiley and Sons.

Sukhatme, P.V.—Teoría de Encuestas por Muestreo con Aplicaciones. Fondo de Cultura Económica.

Zarkovich, S.S.—Calidad de los Datos Estadísticos. Traducción de A. Meredíz, editado por la FAO.

Zarkovich, S.S.—Los Métodos de Muestreo y los Censos. Ed. FAO.

Artículos

Bailar, B.A. (1968).—Recent Research in Reinterview Procedures. JASA, 63.

Brewer, K.R.W. (1963).—A Model of Systematic Sampling with Unequal Probabilities. Austral. Jour. Statist. 5.

Brever, K.R.W. (1967).—JASA, 62.

David, F.N. y Neyman, J. (1938).—Extension of the Markoff Theorem of Least Squares. Statistical Research Memoirs, 2.

Deming, W.E. (1953).—On a Probabilitys Mechanism to Attain an Economic Balance Between the Resultant Error of Non-response and the Bias of Non-response. JASA, 48.

Durbin, J. (1954).—Non-response and Call-backs in Surveys. Bull. ISI, 34.2.

Durbin, J. y Stuart, A. (1954).—Call-backs and Clustering in Sample Surveys: An Experimental Study. JRSS, A, 117.

Durbin, J. (1953).—Some Results in Sampling Theory when the Units are Selected with Unequal Probabilities. JRSS, B. 15.

Durbin, J. (1967).—Design of Multi-stage Surveys for the Estimation of Sampling Errors. Apl. Statist. 16.

Fellegi, I.P. (1963).—JSA, 58.

Fellegi, I.P. y Oldt. D. (1976).— JASA, 71.

Glasser, G.J. (1962).—On the Complete Coverage of Large Units in Statistical Study. Rev. of the ISI, 30.

Gómez Alonso, J.M. (1980).—Traducción del artículo de Fellegi y Oldt. Estadística Española nº 88, INE. Madrid.

Hansen, M.H. y Hurwitz, W.N. (1943).—On the Theory of Sampling from Finite Populations. Ann. Math. Stat. 14.

Hansen, M.H. y Hurwitz, W.N. (1946).—The Problem of Nonresponse in Sample Surveys. JASA, 41.

Hansen, Hurwitz y Bershad (1961).—Measurement Errors in Censuses and Surveys. Bull. of the ISI. 38.

Hartley y Rao (1962).—Sampling with Unequal Probabilities and Without Replacement. Ann. Math. Stat. 33.

Horvitz, D.G. y Thompson, D.J. (1952).—A Generalization of Sampling Without Replacement from a Finite Universe. JASA, 47.

Horvitz, Shah y Simmons (1967).—The Unrelated Question Randomized Response Model. Procc. of Am. Stat. Assoc. (Social Stat. Section).

Jessen, R.J. (1942).—Statistical Investigation of a Sample Survey For Obtaining Farm Facts. Iowa Agricultural Experiment Station Research Bulletin, 304.

Keifitz, N. (1957).—Estimates of Sampling Variance where Two Units are Selected from each Stratum. JASA. 52.

Lahiri (1951).—A Method of Sample Selection Providing Unbiased Ratio Estimates. Bull. ISI, 33, Pt. II.

Mc Carthy, P.J. (1969).—Pseudo-Replication: Half Samples. Rev. ISI. 37.

Midzuno, H. (1951).—On the Sampling System with Probabilitys Proporcionate to Sum of Sizes. Ann. Ins. Stat. Math. 2.

Mirás, J. (1979).—Defectos en el Marco de Muestreo. Estadística Española, nº 82-83. INE. Madrid.

Narain, R.D. (1951).—On Sampling Without Replacement With Varying Probabilities. Jour. Ind. Soc. Agric. Stat. 3.

Neyman, J. (1934).—On the two diferrent espects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. JRSS, 97.

Politz, A.N. y Simmons, W.R. (1949-50).—An attempt to get the "No at homes" into the Sample without Callbacks. JASA, 44 y 45.

Rodriguez-Ponga y Villán (1985).—Propuesta de una Regla para simplificiar el proceso de obtención del Conjunto Completo en la metodología de Fellegi y Olt. Estadística Española, nº 106. INE. Madrid.

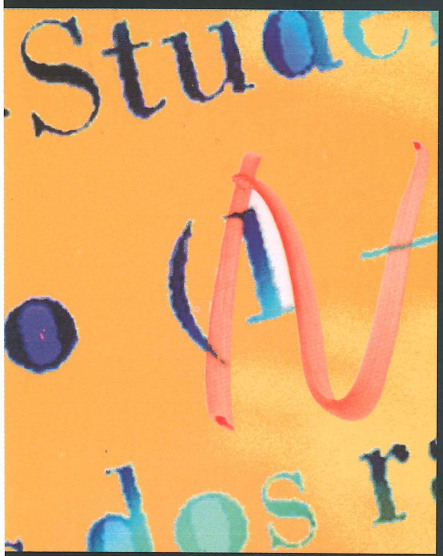
Singh, J. (1976).— A Note on Randomized Response Tecniques. Procc. Am. Stat. Assoc. (Social Stat. Sect.).

Smith, H.F. (1938).—An Empirical Law Describing Heterogeneity in the Yields of Agricultural Crops. Jour. Agr. Sci. 28.

Van Beek y Vermetten (1966).—Appl. Statist. 15.

Warner, S.L. (1965).—Randomized Response: A Survey technique for Eliminating Evasive Answer Bias. JASA, 60.

Yates y Grundy (1953).—Selection Without Replacement from within strata with probability proportional to size. JRSS, B 15.



Colección Libros de autor.

"La toma de decisiones en el ámbito de las actividades económicas y sociales, tanto del sector privado como del sector público, así como en la investigación científica de las mismas, se viene haciendo a partir de informaciones de índole cuantitativa acerca de campos de estudio cada vez más variados y con mayor intensidad por lo que se refiere a las variables en cuestión y a las relaciones entre ellas.

A menudo nos encontramos con necesidades de datos acerca de colectivos finitos como son por ejemplo los habitantes, las familias, las unidades productivas, pero cuyo número de unidades es muy elevado.

Las técnicas de muestreo para poblaciones finitas se han desarrollado para satisfacer estas demandas, permitiendo obtener información de forma más rápida y barata, así como, mediante encuestas sucesivas, suplir la falta de información que se tendría en los períodos intercensales."

Estas palabras del autor resumen a la perfección el interés que las técnicas de muestreo para poblaciones finitas tienen, no sólo para los responsables de tomar las decisiones políticas o económicas, sino también para todos los investigadores de las mismas. Por ello, el Instituto Nacional de Estadística ha considerado oportuno reimprimir una obra que, a buen seguro, constituirá un instrumento básico de consulta y estudio para profesionales y estudiantes de la teoría y técnica del muestreo.

INSTITUTO NACIONAL DE ESTADÍSTICA
INE
INSTITUTO NACIONAL DE ESTADÍSTICA