José Luis Sánchez-Crespo
Javier de Parada

Ejercicios y problemas resueltos de muestreo en poblaciones finitas

INE

1856 2006
150 ANIVERSARIO
ESTADÍSTICA OFICIAL ESPAÑOLA

Colección Libros de autor

José Luis Sánchez-Crespo
Javier de Parada

Ejercicios y problemas resueltos de muestreo en poblaciones finitas

INE

Colección de la. Libros de autor

INSTITUTO NACIONAL DE ESTADISTICA

150 1856 2006 ANIVERSARIO ESTADÍSTICA OFICIAL ESPAÑOLA

Madrid, 2007

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

Ficha editorial

Título: Ejercicios y problemas resueltos de muestreo en poblaciones finitas

Nº INE: 181

NIPO: 605-07-031-X

Depósito Legal: NA-954-2007

ISBN: 978-84-260-1967-7

Tarifa: 3

Edita: INE
Paseo de la Castellana, 183 - 28046 Madrid

Impreso en España/Printed in Spain
Gráficas Lizarra
Ctra. de Tafalla, km 1 31132 Villatuerta (Navarra)

NOTA DE LOS AUTORES

Este libro trata de satisfacer una creciente demanda de estudiantes, opositores e investigadores que se inician en las técnicas del muestreo en poblacones finitas. Se encuentran con que los libros de muestreo abordan la teoría llegando a fórmulas más o menos complejas, presentan el enunciado de algunos ejercicios o problemas, pero rara vez contienen la resolución completa de los mismos, quedándoles la duda sobre la aplicación correcta de los conocimientos adquiridos. La experiencia docente de los autores ha demostrado que muchos estudiantes con un perfecto conocimiento teórico de las técnicas del muestreo son incapaces de resolver un sencillo supuesto práctico, naufragando en la aplicación de los conceptos teóricos.

Lo expuesto anteriormente nos ha movido a publicar este libro en el que se resuelven completamente, con desmenuzamiento que puede parecer trivial en ciertos casos, una serie de ejercicios y problemas de muestreo que abarcan los temas clásicos por los que discurre cualquier libro teórico. En su mayoría, los ejercicios y problemas resueltos corresponden a enunciados propuestos por los autores en la Facultad de Ciencias Económicas y Empresariales de la Universidad Autónoma de Madrid, la Escuela de Estadística y las oposiciones a los Cuerpos de Estadísticos Facultativos y Estadísticos Técnicos Diplomados del Instituto Nacional de Estadística.

Para mayor facilidad del lector hemos seguido la terminología del libro *Métodos y Aplicaciones del Muestreo* (F. Azorín, J. L. Sánchez-Crespo, Ed. Alianza, 1986), si bien se ha procurado explicar suficientemente cualquier símbolo introducido para que sea asequible a cualquier lector.

Los problemas resueltos siguen el orden lógico con el que se enseñan las técnicas de muestreo, así en el capítulo I se presentan las distribuciones en el muestreo de los principales estimadores, en el II se aplican las técnicas relacionadas con el muestreo aleatorio simple, el capítulo III trata del muestreo estratificado, el IV recoge problemas relacionados con estimadores mejorados, estimadores de razón y de regresión, en el capítulo V se presentan técnicas relativas al muestreo con probabilidades desiguales, el capítulo VI se refiere al muestreo de conglomerados, sin y con submuestreo, dedicándose finalmente el último capítulo a repasar algunas técnicas especiales: muestreo sistemático,

muestreo bifásico, muestreo en ocasiones sucesivas, y errores ajenos al muestreo.

Agradecemos los ánimos recibidos para abordar esta ingrata tarea de escribir un libro de problemas de muestreo en poblaciones finitas, y si ello contribuye a un mejor aprendizaje de las técnicas de muestreo, consideraremos cumplido nuestro objetivo.

INDICE

CAPITULO I

Espacio muestral y distribuciones en el muestreo

1.1. En la población (u_1, u_2, u_3) se obtienen todas las muestras posibles de tamaño $n = 2$, con probabilidades iguales y bajo los siguientes supuestos:

a) Muestreo sin reposición con $(u_i, u_j) = (u_j, u_i)$.

b) Muestreo sin reposición con $(u_i, u_j) \neq (u_j, u_i)$.

c) Muestreo con reposición con $(u_i, u_j) = (u_j, u_i)$.

d) Muestreo con reposición con $(u_i, u_j) \neq (u_j, u_i)$.

es decir, en los casos *a*) y *c*) las muestras que constan de las mismas unidades en distinto orden se consideran idénticas, mientras que en los casos *b*) y *d*) tales muestras se consideran distintas.

Se pide: El espacio muestral y los valores de la función de probabilidad en cada supuesto.

Solución:

Denominando espacio muestral $S(\mathbf{x})$, al conjunto de muestras posibles $\mathbf{x}$ seleccionadas con un procedimiento de muestreo dado, tendremos:

Caso *a*):

$S(\mathbf{x})$	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)\mathrm{P}(u_i/u_j)$
u_1u_2	$\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3}$
u_1u_3	$\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3}$
u_2u_3	$\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{3}$

Caso *b*):

$S(\mathbf{x})$	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i)$
u_1u_2	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
u_1u_3	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
u_2u_1	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
u_2u_3	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
u_3u_1	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$
u_3u_2	$\frac{1}{3} \times \frac{1}{2} = \frac{1}{6}$

Caso *c*):

$S(\mathbf{x})$	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)\mathrm{P}(u_i/u_j)$
u_1u_1	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_1u_2	$\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{2}{9}$
u_1u_3	$\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{2}{9}$
u_2u_2	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_2u_3	$\frac{1}{3} \times \frac{1}{3} + \frac{1}{3} \times \frac{1}{3} = \frac{2}{9}$
u_3u_3	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

Caso *d*):

$S(\mathbf{x})$	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i)$
u_1u_1	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_1u_2	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_1u_3	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_2u_1	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_2u_2	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_2u_3	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_3u_1	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_3u_2	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$
u_3u_3	$\frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$

Se observa que en el caso *c*) las muestras no son equiprobables. En caso de muestras equiprobables, la probabilidad de cada muestra es 1/número de muestras posibles. En el caso *a*), el número de muestras posibles son las combinaciones de N elementos tomados de *n* en *n*, es decir $\binom{3}{2} = 3$. En el caso b), el número de muestras posibles son las variaciones de *N* elementos tomados de *n* en *n*, es decir $\binom{3}{2}2! = 6$. Finalmente, en el caso *d*), el número de muestras posibles sería el de variaciones con repetición de *N* elementos, tomados de *n* en *n*, es decir $3^2 = 9$.

En el caso *c*) el número de muestras posibles sería el de combinaciones con repetición de *N* elementos, tomados de *n* en *n*, es decir $\binom{3+2-1}{2} = 6$, aunque debido a la no equiprobabilidad de las muestras, no sería correcto en este caso asignar la probabilidad 1/6 a cada muestra.

1.2. Dada la población (u_1, u_2, u_3) se obtienen todas las muestras posibles de tamaño $n = 2$, con las siguientes especificaciones:

I) Muestreo con reposición y probabilidades iguales.

II) $\{u_i, u_j\} = \{u_j, u_i\}$.

Se pide: El espacio muestral y los valores de la probabilidad de cada muestra utilizando la distribución polinomial.

Solución:

En general, en el muestreo con reposición en una población de tamaño N, si designamos por P_i la probabilidad de ser seleccionada la unidad u_i en cada extracción, y e_i el número de veces que resulta seleccionada u_i en n extracciones, la distribución polinomial nos dice que:

$$P(e_1, e_2, ..., e_N) = \frac{n!}{e_1!\, e_2! \cdots e_N!} P_1^{e_1} P_2^{e_2} \cdots P_N^{e_N}$$

En el caso particular $N = 3$, $n = 2$, $P_1 = P_2 = P_3 = 1/3$, resulta:

$$P(u_1, u_1) = P(e_1 = 2; e_2 = 0; e_3 = 0) = \frac{2!}{2!\,0!\,0!}\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^0\left(\frac{1}{3}\right)^0 = \frac{1}{9}$$

análogamente para $(u_2; u_2)$ y $(u_3; u_3)$.

$$P(u_1, u_2) = P(e_1 = 1; e_2 = 1; e_3 = 0) = \frac{2!}{1!\,1!\,0!}\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^1\left(\frac{1}{3}\right)^0 = \frac{2}{9}$$

y análogamente para $(u_1; u_3)$ y $(u_2; u_3)$.

Por tanto:

$S(\mathbf{x})$	Valores de la distribución polinomial
u_1u_1	$P(2,0,0) = \frac{1}{9}$
u_1u_2	$P(1,1,0) = \frac{2}{9}$
u_1u_3	$P(1,0,1) = \frac{2}{9}$
u_2u_2	$P(0,2,0) = \frac{1}{9}$
u_2u_3	$P(0,1,1) = \frac{2}{9}$
u_3u_3	$P(0,0,2) = \frac{1}{9}$

Obsérvese que coincide con el caso *c*) del ejercicio anterior.

1.3. Una población está formada por 5 unidades de muestreo a las que se asignan probabilidades iguales. Se obtiene una muestra de 2 unidades siguiendo cada uno de los siguientes procesos de selección:

a) Muestreo sin reposición considerando idénticas las muestras que contengan las mismas unidades.

b) Muestreo con reposición considerando idénticas las muestras que contengan las mismas unidades.

c) Muestreo sin reposición considerando distintas las muestras que contengan las mismas unidades.

d) Muestreo con reposición considerando distintas las muestras que contengan las mismas unidades.

Se pide:

1.º) Calcular el número de muestras posibles en cada uno de los casos anteriores.

2.º) ¿Cuál será, en media, el número de unidades distintas contenidas en la muestra, en el caso *b*)?

Solución:

1.º)

Caso *a*): Combinaciones $\binom{N}{n} = \binom{5}{2} = 10.$

Caso *c*): Variaciones $\binom{N}{n} n! = \binom{5}{2} 2! = 20.$

Caso *b*): Combinaciones con repetición $\binom{N+n-1}{n} = \binom{6}{2} = 15.$

Caso *d*) Variaciones con repetición $N^n = 5^2 = 25.$

2.º) Si definimos por v la variable aleatoria «número de unidades distintas en la muestra», a partir de la tabla siguiente correspondiente al caso *b*):

ESPACIO MUESTRAL	MATRIZ DE PROBABILIDAD
$\{u_1u_1\}\{u_1u_2\}\{u_1u_3\}\{u_1u_4\}\{u_1u_5\}$	1/25 2/25 2/25 2/25 2/25
$\{u_2u_2\}\{u_2u_3\}\{u_2u_4\}\{u_2u_5\}$	1/25 2/25 2/25 2/25
$\{u_3u_3\}\{u_3u_4\}\{u_3u_5\}$	1/25 2/25 2/25
$\{u_4u_4\}\{u_4u_5\}$	1/25 2/25
$\{u_5u_5\}$	1/25

podemos formar la distribución de probabilidad de la variable v y proceder al cálculo de su esperanza:

v	$P(v)$
1	$\frac{5}{25}$
2	$\frac{20}{25}$

$$E(v) = 1 \times \frac{1}{5} + 2 \times \frac{4}{5} = \frac{9}{5}$$

1.4. Dada la población $U_i\{1, 2, 3\}$ con probabilidades $P_i\left\{\frac{1}{6}, \frac{2}{6}, \frac{3}{6}\right\}$ y las especificaciones: I) $n = 2$; II) Muestreo sin reposición; III) $(u_i, u_j) = (u_j, u_i)$.

Se pide:

a) Espacio muestral $S(\mathbf{x})$ y la función de probabilidad $P(\mathbf{x})$.

b) Valores de $\pi_i = \text{Prob.}\,(u_i \in \text{muestra})$ a partir de las $\pi_{ij} = P(u_i, u_j)$.

c) Comprobar que se verifica $\sum_{i=1}^{N} \pi_i = n = 2$.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1975.)

Solución

a)

$S(\mathbf{x})$	Función de probabilidad $\pi_{ij} = P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j)$
u_1u_2	$\pi_{12} = \frac{1}{6} \times \frac{2}{5} + \frac{2}{6} \times \frac{1}{4} = \frac{9}{60}$
u_1u_3	$\pi_{13} = \frac{1}{6} \times \frac{3}{5} + \frac{3}{6} \times \frac{1}{3} = \frac{16}{60}$
u_2u_3	$\pi_{23} = \frac{2}{6} \times \frac{3}{4} + \frac{3}{6} \times \frac{2}{3} = \frac{35}{60}$

b) Las probabilidades de inclusión de cada unidad en la muestra son:

$$\pi_1 = \mathrm{P}(u_1) = \pi_{12} + \pi_{13} = \frac{25}{60}$$

$$\pi_2 = \mathrm{P}(u_2) = \pi_{12} + \pi_{23} = \frac{44}{60}$$

$$\pi_3 = \mathrm{P}(u_3) = \pi_{13} + \pi_{23} = \frac{51}{60}$$

c) Se comprueba fácilmente que:

$$\sum_{i}^{N} \pi_i = \frac{25}{60} + \frac{44}{60} + \frac{51}{60} = \frac{120}{60} = 2$$

1.5. Dada la población $X_i = 1, 2, 3$ se selecciona mediante muestreo aleatorio simple una muestra de tamaño $n = 2$, y se da la regla circular de sustituir una unidad por la siguiente unidad no seleccionada cuando no se obtenga información en el primer intento.

Se pide:

a) El espacio muestral, manteniendo el tamaño de la muestra, cuando u_1 no constesta en el primer intento.

b) El sesgo de estimación en el estimador del total.

Solución:

a)

Espacio muestral original		Espacio muestral distorsionado	
$S(\mathbf{x})$	$P(\mathbf{x})$	$S'(\mathbf{x})$	$P'(\mathbf{x})$
u_1u_2	$\frac{1}{3}$	u_3u_2	$\frac{1}{3}$
u_1u_3	$\frac{1}{3}$	u_2u_3	$\frac{1}{3}$
u_2u_3	$\frac{1}{3}$	u_2u_3	$\frac{1}{3}$

b) El total poblacional X viene dado por:

$$X = 1 + 2 + 3 = 6$$

y el estimador del total en un muestreo aleatorio simple, se obtiene mediante:

$$\hat{X} = \frac{N}{n} \sum_{i}^{n} x_i$$

A causa de la regla de sustitución la distribución de este estimador en el muestreo será:

Muestra original	Muestra distorsionada	$\hat{X}_{s'}$	$P(\hat{X}_{s'})$
u_1u_2	u_3u_2	$\frac{3}{2}(3+2) = 7{,}5$	$\frac{1}{3}$
u_1u_3	u_2u_3	7,5	$\frac{1}{3}$
u_2u_3	u_2u_3	7,5	$\frac{1}{3}$

siendo $E\hat{X} = \sum_{s'} \hat{X}_{s'}[P(\hat{X}_{s'})] = 7{,}5$.

Por definición:

$$\text{sesgo } B = E\hat{X}_{s'} - X = 7{,}5 - 6 = 1{,}5$$

puede observarse que la sustitución de una unidad originalmente seleccionada ocasiona un sesgo en la estimación.

1.6. Dada la población

X_i	1	3	4
P_i	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Se pide:

1.°) Formar el espacio muestral y matriz de probabilidades con $n = 2$, selección con reposición y probabilidades P_i, considerándose idénticas las muestras (u_i, u_j) y (u_j, u_i).

2.°) Calcular $V(\hat{X}_{HH})$, siendo $\hat{X}_{HH}$ el estimador insesgado de Hansen y Hurwitz.

3.°) A partir de la distribución en el muestreo de $\hat{X}_{HH}$ y $\hat{V}(\hat{X}_{HH})$, calcular $V(\hat{X}_{HH})$, $E(\hat{X}_{HH})$ y $E[\hat{V}(\hat{X}_{HH})]$.

Solución:

1.°) En un muestreo con reposición cada extracción es un suceso aleatorio independiente del anterior, por tanto, las probabilidades de selección de cada muestra pueden obtenerse mediante:

$$P(u_i, u_i) = P(u_i)P(u_i) = [P(u_i)]^2$$

$$P(u_i, u_j) = P(u_i)P(u_j) + P(u_j)P(u_i) = 2P(u_i)P(u_j)$$

resultando la siguiente matriz de probabilidades:

ESPACIO MUESTRAL

$$S(\mathbf{x}) = \left\{ \begin{matrix} (u_1u_1)\,(u_1u_2)\,(u_1u_3) \\ (u_2u_2)\,(u_2u_3) \\ (u_3u_3) \end{matrix} \right\}$$

MATRIZ DE PROBABILIDADES

$$P(\mathbf{x}) = \left\{ \begin{matrix} 1/36 & 4/36 & 6/36 \\ & 4/36 & 12/36 \\ & & 9/36 \end{matrix} \right\}$$

2.º) El estimador insesgado del total (también llamado estimador de Hansen y Hurwitz: $\hat{X}_{HH}$) viene dado por:

$$\hat{X}_{HH} = \sum_{i}^{n} \frac{X_i}{nP_i}$$

siendo su varianza

$$V(\hat{X}_{HH}) = \frac{1}{n} \sum_{i}^{N} \left(\frac{X_i}{P_i} - X \right)^2 P_i \quad ; \quad X = \sum_{i}^{N} X_i$$

obteniéndose para los datos del enunciado:

$$V(\hat{X}_{HH}) = \frac{1}{2} \left[\frac{1}{6} (6 - 8)^2 + \frac{2}{6} (9 - 8)^2 + \frac{3}{6} (8 - 8)^2 \right] = 0{,}5$$

3.º) Un estimador insesgado de $V(\hat{X}_{HH})$ es:

$$\hat{V}(\hat{X}_{HH}) = \frac{\sum_{i}^{n} \left(\frac{X_i}{P_i} - \hat{X}_{HH} \right)^2}{n(n-1)}$$

resultando la siguiente distribución en el muestreo:

$S(\mathbf{x})$	$P(\mathbf{x})$	$\hat{X}_{HH}$	$\hat{V}(\hat{X}_{HH})$
u_1u_1	$\frac{1}{36}$	$\frac{1}{2}(6+6)=6$	$\frac{1}{2}[(6-6)^2+(6-6)^2]=0$
u_1u_2	$\frac{4}{36}$	$\frac{1}{2}(6+9)=7{,}5$	$\frac{1}{2}[(6-7{,}5)^2+(9-7{,}5)^2]=2{,}25$
u_1u_3	$\frac{6}{36}$	$\frac{1}{2}(6+8)=7$	$\frac{1}{2}[(6-7)^2+(8-7)^2]=1$
u_2u_2	$\frac{4}{36}$	$\frac{1}{2}(9+9)=9$	$\frac{1}{2}[(9-9)^2+(9-9)^2]=0$
u_2u_3	$\frac{12}{36}$	$\frac{1}{2}(9+8)=8{,}5$	$\frac{1}{2}[(9-8{,}5)^2+(8-8{,}5)^2]=0{,}25$
u_3u_3	$\frac{9}{36}$	$\frac{1}{2}(8+8)=8$	$\frac{1}{2}[(8-8)^2+(8-8)^2]=0$

Aplicando a la anterior distribución la definición de esperanza matemática:

$$E(\hat{X}_{HH}) = 6 \times \frac{1}{36} + 7{,}5 \times \frac{4}{36} + 7 \times \frac{6}{36} + 9 \times \frac{4}{36} +$$

$$+ 8{,}5 \times \frac{12}{36} + 8 \times \frac{9}{36} = 8 = X$$

estimador insesgado.

$$E[\hat{V}(\hat{X}_{HH})] = 2{,}25 \times \frac{4}{36} + 1 \times \frac{6}{36} + 0{,}25 \times \frac{12}{36} =$$

$$= \frac{18}{36} = 0{,}5 = V(\hat{X}_{HH})$$

estimador insesgado.

Asimismo, por definición de varianza:

$$V(\hat{X}_{HH}) = (6-8)^2\frac{1}{36} + (7{,}5-8)^2\frac{4}{36} + (7-8)^2\frac{6}{36} + (9-8)^2\frac{4}{36} +$$

$$+ (8{,}5-8)^2\frac{12}{36} + (8-8)^2\frac{9}{36} = 0{,}5$$

1.7. Dada la población

X_i	1	3	4
P_i	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Se pide:

1.º) Formar el espacio muestral y matriz de probabilidades con $n = 2$, selección sin reposición y probabilidades P_i, considerándose idénticas las muestras (u_i, u_j) y (u_j, u_i).

2.º) Calcular $V(\hat{X}_{HT})$, siendo $\hat{X}_{HT}$ el estimador insesgado de Horvitz y Thompson.

3.º) A partir de la distribución en el muestreo de $\hat{X}_{HT}$ y $\hat{V}(\hat{X}_{HT})$, calcular $V(\hat{X}_{HT})$, $E(\hat{X}_{HT})$ y $E[\hat{V}(\hat{X}_{HT})]$.

4.º) Comprobar que $\sum_{i}^{N} \pi_i = n$, siendo $\pi_i = P(u_i \in \text{muestra})$.

Solución:

1.º) Las probabilidades de selección de cada muestra se obtienen mediante:

$$\pi_{ij} = P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j) =$$

$$= P(u_i)\frac{P(u_j)}{1 - P(u_i)} + P(u_j)\frac{P(u_i)}{1 - P(u_j)}$$

resultando la siguiente matriz de probabilidades:

Espacio muestral	Probabilidades (π_{ij})
u_1u_2	$\frac{9}{60}$
u_1u_3	$\frac{16}{60}$
u_2u_3	$\frac{35}{60}$

2.º) El estimador insesgado del total (también llamado estimador de Horvitz y Thompson: $\hat{X}_{HT}$) viene dado por:

$$\hat{X}_{HT} = \sum_{i}^{n} \frac{X_i}{\pi_i} \quad , \quad \text{siendo } \pi_i = P(u_i \in \text{muestra})$$

En nuestro caso:

$$\pi_1 = \pi_{12} + \pi_{13} = \frac{25}{60}$$

$$\pi_2 = \pi_{12} + \pi_{23} = \frac{44}{60}$$

$$\pi_3 = \pi_{13} + \pi_{23} = \frac{51}{60}$$

siendo la varianza:

$$V(\hat{X}_{HT}) = \sum_{i}^{N} \frac{X_i^2}{\pi_i}(1 - \pi_i) + 2 \sum_{i<j}^{N} \frac{X_i X_j}{\pi_i \pi_j}(\pi_{ij} - \pi_i \pi_j)$$

sustituyendo los correspondientes valores, resulta:

$$V(\hat{X}_{HT}) = \frac{60}{25}\left(1 - \frac{25}{60}\right) + \frac{540}{44}\left(1 - \frac{44}{60}\right) + \frac{960}{51}\left(1 - \frac{51}{60}\right) +$$

$$+ 2\left[3\left(\frac{540}{25 \times 44} - 1\right) + 4\left(\frac{960}{25 \times 51} - 1\right) + 12\left(\frac{2.100}{44 \times 51} - 1\right)\right] = 0{,}925$$

3.°) Un estimador insesgado de $V(\hat{X}_{HT})$ debido a Yates Grundy es:

$$\hat{V}_{YG}(\hat{X}_{HT}) = \sum_{i<j}^{n} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$$

resultando la siguiente distribución en el muestreo:

$S(\mathbf{x})$	$P(\mathbf{x})$	$\hat{X}_{HT}$	$\hat{V}(\hat{X}_{HT})$
$u_1 u_2$	$\frac{9}{60}$	$\frac{60}{25} + \frac{180}{44} = 6{,}5$	$\frac{\left(\frac{25 \times 44}{60^2} - \frac{9}{60}\right)\left(\frac{60}{25} - \frac{180}{44}\right)^2}{\frac{9}{60}} = 2{,}965$
$u_1 u_3$	$\frac{16}{60}$	$\frac{60}{25} + \frac{240}{51} = 7{,}1$	$\frac{\left(\frac{25 \times 51}{60^2} - \frac{16}{60}\right)\left(\frac{60}{25} - \frac{240}{51}\right)^2}{\frac{16}{60}} = 1{,}745$
$u_2 u_3$	$\frac{35}{60}$	$\frac{180}{44} + \frac{240}{51} = 8{,}8$	$\frac{\left(\frac{44 \times 51}{60^2} - \frac{35}{60}\right)\left(\frac{180}{44} - \frac{240}{51}\right)^2}{\frac{35}{60}} = 0{,}026$

Aplicando las definiciones de esperanza y varianza, resulta:

$$E(\hat{X}_{HT}) = 6{,}5 \times \frac{9}{60} + 7{,}1 \times \frac{16}{60} + 8{,}8 \times \frac{35}{60} = 8 = X$$

estimador insesgado.

$$V(\hat{X}_{HT}) = (6{,}5-8)^2\frac{9}{60} + (7{,}1-8)^2\frac{16}{60} + (8{,}8-8)^2\frac{35}{60} = 0{,}925$$

$$E[\hat{V}(\hat{X}_{HT})] = 2{,}965 \times \frac{9}{60} + 1{,}745 \times \frac{16}{60} + \frac{0{,}026 \times 35}{60} =$$

$$= 0{,}925 = V(\hat{X}_{HT})$$

estimador insesgado.

4.°) De los valores obtenidos en el apartado 2.°), se comprueba fácilmente que:

$$\sum_i^N \pi_i = \frac{25}{60} + \frac{44}{60} + \frac{51}{60} = 2 = n$$

1.8. Dada la población

X_i	1	3	4
P_i	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Se pide:

1.°) Formar el espacio muestral y matriz de probabilidades con $n = 2$, utilizando el esquema de selección con probabilidades gradualmente variables (Sánchez-Crespo, 1977).

2.°) Calcular $V(\hat{X}_{SC})$, siendo $\hat{X}_{SC}$ el estimador insesgado de Sánchez-Crespo.

3.°) A partir de la distribución en el muestreo de $\hat{X}_{SC}$ y $\hat{V}(\hat{X}_{SC})$, calcular $V(\hat{X}_{SC})$, $E(\hat{X}_{SC})$ y $E[\hat{V}(\hat{X}_{SC})]$.

4.°) Comprobar que

$$V(\hat{X}_{SC}) = \frac{M-n}{M-1} V(\hat{X}_{HH})$$

donde $M = \sum_i^N M_i$, siendo M_i el número de bolas que representa a la unidad u_i en el esquema de selección mencionado.

Solución:

1.º) Según el esquema de selección de probabilidades gradualmente variables se supone que existen 6 bolas en una urna, de las que 1 bola representa la unidad u_1, 2 bolas la unidad u_2 y 3 bolas la unidad u_3 (de acuerdo con las probabilidades originales 1/6, 2/6 y 3/6, respectivamente). En cada extracción la bola extraída no se restituye a la urna, por lo que las probabilidades varían de una extracción a otra. Es un método intermedio entre el muestreo con y sin reposición, pues una unidad representada por dos o más bolas en la urna puede resultar seleccionada más de una vez.

Según este esquema, el espacio muestral y las probabilidades de cada muestra será:

$S(\mathbf{x})$	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j)$
u_1u_2	$\frac{1}{6} \times \frac{2}{5} + \frac{2}{6} \times \frac{1}{5} = \frac{2}{15}$
u_1u_3	$\frac{1}{6} \times \frac{3}{5} + \frac{3}{6} \times \frac{1}{5} = \frac{3}{15}$
u_2u_2	$\frac{2}{6} \times \frac{1}{5} = \frac{1}{15}$
u_2u_3	$\frac{2}{6} \times \frac{3}{5} + \frac{3}{6} \times \frac{2}{5} = \frac{6}{15}$
u_3u_3	$\frac{3}{6} \times \frac{2}{5} = \frac{3}{15}$

2.º) El estimador insesgado del total (véase Azorín-Sánchez-Crespo, 1986, pág. 206) viene dado por:

$$\hat{X}_{SC} = \sum_i^n \frac{X_i}{nP_i}$$

con varianza:

$$V(\hat{X}_{SC}) = \frac{M-n}{M-1} \times \frac{1}{n} \sum_i^N P_i \left(\frac{X_i}{P_i} - X \right)^2$$

donde M representa el número total inicial de bolas en la urna; en este caso $M=6$.

Sustituyendo valores:

$$V(\hat{X}_{SC}) = \frac{6-2}{5\times 2}\left[\frac{1}{6}(6-8)^2 + \frac{2}{6}(9-8)^2 + \frac{3}{6}(8-8)^2\right] = 0{,}4$$

3.º) Un estimador insesgado de $V(\hat{X}_{SC})$ es:

$$\hat{V}(\hat{X}_{SC}) = \frac{M-n}{M}\,\frac{\sum_i^n \left(\frac{X_i}{P_i} - \hat{X}_{SC}\right)^2}{n(n-1)}$$

resultando la siguiente distribución en el muestreo:

$S(\mathbf{x})$	$P(\mathbf{x})$	$\hat{X}_{SC}$	$\hat{V}(\hat{X}_{SC})$
u_1u_2	$\frac{2}{15}$	$\frac{1}{2}(6+9)=7{,}5$	$\frac{1}{3}[(6-7{,}5)^2+(9-7{,}5)^2]=1{,}5$
u_1u_3	$\frac{3}{15}$	$\frac{1}{2}(6+8)=7$	$\frac{1}{3}[(6-7)^2+(8-7)^2]=\frac{2}{3}$
u_2u_2	$\frac{1}{15}$	$\frac{1}{2}(9+9)=9$	$\frac{1}{3}[(9-9)^2+(9-9)^2]=0$
u_2u_3	$\frac{6}{15}$	$\frac{1}{2}(9+8)=8{,}5$	$\frac{1}{3}[(9-8{,}5)^2+(8-8{,}5)^2]=\frac{0{,}5}{3}$
u_3u_3	$\frac{3}{15}$	$\frac{1}{2}(8+8)=8$	$\frac{1}{3}[(8-8)^2+(8-8)^2]=0$

Aplicando las definiciones de esperanza y varianza, resulta:

$$E(\hat{X}_{SC}) = 7{,}5\times\frac{2}{15} + 7\times\frac{3}{15} + 9\times\frac{1}{15} + 8{,}5\times\frac{6}{15} + 8\times\frac{3}{15} =$$

$$= \frac{120}{15} = 8 = \bar{X}$$

estimador insesgado.

$$V(\hat{X}_{SC}) = (7{,}5-8)^2\frac{2}{15} + (7-8)^2\frac{3}{15} + (9-8)^2\frac{1}{15} + (8{,}5-8)^2\frac{6}{15} +$$

$$+ (8-8)^2\frac{3}{15} = \frac{6}{15} = 0{,}4$$

$$E[\hat{V}(\hat{X}_{SC})] = 1{,}5 \times \frac{2}{15} + \frac{2}{3} \times \frac{3}{15} + 0 + \frac{0{,}5}{3} \times \frac{6}{15} + 0 =$$

$$= \frac{18}{45} = 0{,}4 = V(\hat{X}_{SC})$$

estimador insesgado.

4.º) Teniendo en cuenta que $V(\hat{X}_{HH}) = 0{,}5$ (véase ejercicio n.º 6), se comprueba fácilmente que:

$$V(\hat{X}_{SC}) = \frac{M - n}{M - 1} V(\hat{X}_{HH}) = \frac{6 - 2}{5} \times 0{,}5 = 0{,}4$$

1.9. Dada la población

X_i	1	3	4
P_i	3	5	7

con $M = \sum_i^N M_i = 15$, y probabilidades iniciales proporcionales a los tamaños,

$$P_i = \left\{\frac{3}{15} \; ; \; \frac{5}{15} \; ; \; \frac{7}{15}\right\},$$

se seleccionan todas las muestras posibles de tamaño $n = 2$.

Considerando idénticas las muestras $\{u_i, u_j\}$ y $\{u_j, u_i\}$, se pide:

1.º) Formar el espacio muestral y la matriz de probabilidades utilizando el esquema mixto de selección debido a Sánchez-Crespo y Gabeiras, en el que

$$\hat{X}_{SCG} = \sum_i^n X_i / nP_i.$$

En este método, siguiendo el esquema de urna, después de cada selección se retiran b bolas, siendo

$$b = \frac{\text{mín. } (M_i)}{n - 1} \quad , \quad (i = 1, 2, ..., N),$$

donde M_i es el número de bolas que representan a la unidad u_i.

2.º) Utilizando las distribuciones en el muestreo de los estimadores $\hat{X}_{SCG}$ y $\hat{X}_{HH}$, comprobar las relaciones siguientes:

$$E(\hat{X}_{SCG}) = E(\hat{X}_{HH}) = X \quad , \quad V(\hat{X}_{SCG}) = \frac{M - nb}{M - b} V(\hat{X}_{HH})$$

donde $\hat{X}_{HH}$ es el estimador de Hansen y Hurwitz (selección con probabilidades proporcionales a los tamaños y con reemplazamiento).

3.º) Considerando el estimador debido a Horwitz y Thompson (selección con probabilidades desiguales y sin reemplazamiento) y el método de selección de Brewer en el que:

$$\hat{X}_{HT} = \sum_{i}^{n} \frac{X_i}{\pi_i} \quad , \quad \pi_i = nP_i \quad , \quad \pi_{ij} = \frac{2P_iP_j}{1 + \sum_{i}^{N} \frac{P_i}{1-P_i}} \left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j} \right] \quad , \quad j > i$$

calcular: $E(\hat{X}_{HT})$ y $V(\hat{X}_{HT})$ utilizando la distribución en el muestreo de $\hat{X}_{HT}$.

Solución:

1.º) Siendo

$$b = \frac{\text{mín. } (M_i)}{n - 1} = 3,$$

tendremos el siguiente espacio muestral:

Muestras posibles	Función de probabilidad $P(u_i, u_j) = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j)$
u_1u_2	$\left(\frac{3}{15}\right)\left(\frac{5}{12}\right) + \left(\frac{5}{15}\right)\left(\frac{3}{12}\right) = \frac{15}{90}$
u_1u_3	$\left(\frac{3}{15}\right)\left(\frac{7}{12}\right) + \left(\frac{7}{15}\right)\left(\frac{3}{12}\right) = \frac{21}{90}$
u_2u_2	$\left(\frac{5}{15}\right)\left(\frac{2}{12}\right) = \frac{5}{90}$
u_2u_3	$\left(\frac{5}{15}\right)\left(\frac{7}{12}\right) + \left(\frac{7}{15}\right)\left(\frac{5}{12}\right) = \frac{35}{90}$
u_3u_3	$\left(\frac{7}{15}\right)\left(\frac{4}{12}\right) = \frac{14}{90}$

2.º) Los estimadores $\hat{X}_{SCG}$ y $\hat{X}_{HH}$ tienen idéntica expresión:

$$\hat{X}_{SCG} = \hat{X}_{HH} = \sum_{i}^{n} \frac{X_i}{nP_i}$$

aunque distinta distribución en el muestreo:

Distribución en el muestreo de $\hat{X}_{SCG}$

Muestras posibles	$\hat{X}_{SCG} = \frac{1}{n}\sum_{i}^{n}\frac{X_i}{P_i}$	$P(\hat{X}_{SCG})$	$\hat{X}_{SCG} \cdot P(\hat{X}_{SCG})$	$(\hat{X}_{SCG} - X)^2 \cdot P(\hat{X}_{SCG})$
u_1u_2	7	$\frac{15}{90}$	$\frac{14}{12}$	$\frac{15}{90}$
u_1u_3	$\frac{95}{14}$	$\frac{21}{90}$	$\frac{19}{12}$	$\frac{864}{2.520}$
u_2u_2	9	$\frac{5}{90}$	$\frac{6}{12}$	$\frac{5}{90}$
u_2u_3	$\frac{123}{14}$	$\frac{35}{90}$	$\frac{41}{12}$	$\frac{605}{2.520}$
u_3u_3	$\frac{60}{7}$	$\frac{14}{90}$	$\frac{16}{12}$	$\frac{128}{2.520}$
			$E(\hat{X}_{SCG}) = \frac{96}{12} = 8 = X$	$V(\hat{X}_{SCG}) = \frac{2.160}{2.520} = \frac{6}{7} = 0{,}8571$

Distribución en el muestreo de $\hat{X}_{HH}$

Muestras posibles	$\hat{X}_{HH}=\frac{1}{n}\sum_{i}^{n}\frac{X_i}{P_i}$	$P(\hat{X}_{HH})$	$\hat{X}_{HH}\cdot P(\hat{X}_{HH})$	$(\hat{X}_{HH}-X)^2\cdot P(\hat{X}_{HH})$
u_1u_1	5	$\frac{9}{225}$	$\frac{45}{225}$	$\frac{81}{225}$
u_1u_2	7	$\frac{30}{225}$	$\frac{210}{225}$	$\frac{30}{225}$
u_1u_3	$\frac{95}{14}$	$\frac{42}{225}$	$\frac{285}{225}$	$\frac{867}{3.150}$
u_2u_2	9	$\frac{25}{225}$	$\frac{225}{225}$	$\frac{25}{225}$
u_2u_3	$\frac{123}{14}$	$\frac{70}{225}$	$\frac{615}{225}$	$\frac{605}{3.150}$
u_3u_3	$\frac{60}{7}$	$\frac{49}{225}$	$\frac{420}{225}$	$\frac{16}{225}$
			$E(\hat{X}_{HH})=\frac{1.800}{225}$	$V(\hat{X}_{HH})=\frac{3.600}{3.150}$
			$=8=X$	$=\frac{8}{7}=1,1428$

Por tanto:

$$\frac{V(\hat{X}_{SCG})}{V(\hat{X}_{HH})}=\frac{6/7}{8/7}=\frac{3}{4}$$

y como

$$\frac{M-nb}{M-b}=\frac{15-2\times 3}{15-3}=\frac{3}{4}$$

se verifica

$$V(\hat{X}_{SCG})=\frac{M-nb}{M-b}V(\hat{X}_{HH})$$

3.°) Los valores de π_{ij} son:

$$\pi_{12} = \frac{\frac{2}{15}}{1+\frac{25}{3}}\left[\frac{1}{1-\frac{6}{15}}+\frac{1}{1-\frac{10}{15}}\right]=\frac{1}{15}$$

$$\pi_{13} = \frac{\frac{42}{225}}{1+\frac{25}{3}}\left[\frac{1}{1-\frac{6}{15}}+\frac{1}{1-\frac{14}{15}}\right]=\frac{5}{15}$$

$$\pi_{23} = \frac{\frac{70}{225}}{1+\frac{25}{3}}\left[\frac{1}{1-\frac{10}{15}}+\frac{1}{1-\frac{14}{15}}\right]=\frac{9}{15}$$

Distribución en el muestreo de $\hat{X}_{HT}$ con selección Brewer

Muestras posibles	$\hat{X}_{HT}=\frac{1}{n}\sum_{i}^{n}\frac{X_i}{\pi_i}$	$P(\hat{X}_{HT})$	$\hat{X}_{HT}\cdot P(\hat{X}_{HT})$	$(\hat{X}_{HT}-X)^2P(\hat{X}_{HT})$
u_1u_2	7	$\frac{1}{15}$	$\frac{7}{15}$	$\frac{1}{15}$
u_1u_3	$\frac{95}{14}$	$\frac{5}{15}$	$\frac{475}{210}$	$\frac{1.445}{2.940}$
u_2u_3	$\frac{123}{14}$	$\frac{9}{15}$	$\frac{1.107}{210}$	$\frac{1.089}{2.940}$
			$E(\hat{X}_{HT})=\frac{1.680}{210}$	$V(\hat{X}_{HT})=\frac{2.370}{2.940}$
			$=8=X$	$=\frac{13}{14}=0{,}9286$

Puede observarse que siempre será $V(\hat{X}_{SCG}) < V(\hat{X}_{HH})$, y en este supuesto práctico:

$$V(\hat{X}_{SCG}) < V(\hat{X}_{HT}) < V(\hat{X}_{HH})$$

1.10. Con los datos del ejercicio anterior, obtener las distribuciones en el muestreo de $\hat{V}(\hat{X}_{SCG})$ y $\hat{V}(\hat{X}_{HH})$, siendo:

$$\hat{V}(\hat{X}_{SCG}) = \frac{M - nb}{M} \times \frac{1}{n(n-1)} \left[\sum_{i}^{n} \left(\frac{X_i}{P_i} \right)^2 - n\hat{X}_{SCG}^2 \right]$$

$$\hat{V}(\hat{X}_{HH}) = \frac{1}{n(n-1)} \left[\sum_{i}^{n} \left(\frac{X_i}{P_i} \right)^2 - n\hat{X}_{HH}^2 \right]$$

Comprobar que son estimadores insesgados de $V(\hat{X}_{SCG})$ y $V(\hat{X}_{HH})$, respectivamente.

Solución:

Distribución en el muestreo de $\hat{V}(\hat{X}_{SCG})$

Muestras posibles	$\hat{V}(\hat{X}_{SCG})$	$P(X_{SCG})$	$\hat{V}(\hat{X}_{SCG}) \cdot P(\hat{X}_{SCG})$
u_1u_2	$\frac{3}{10}[5^2+9^2-2\times 7^2]=\frac{24}{10}$	$\frac{15}{90}$	$\frac{36}{90}$
u_1u_3	$\frac{3}{10}\left[5^2+\left(\frac{60}{7}\right)^2-2\left(\frac{95}{14}\right)^2\right]=\frac{375}{196}$	$\frac{21}{90}$	$\frac{7.875}{17.640}$
u_2u_2	0	$\frac{5}{90}$	0
u_2u_3	$\frac{3}{10}\left[9^2+\left(\frac{60}{7}\right)^2-2\left(\frac{123}{14}\right)^2\right]=\frac{54}{1.960}$	$\frac{35}{90}$	$\frac{189}{17.640}$
u_3u_3	0	$\frac{14}{90}$	0

$$E\hat{V}(X_{SCG}) = \frac{15.120}{17.640} =$$

$$= \frac{6}{7} = V(\hat{X}_{SCG})$$

Distribución en el muestreo de $\hat{V}(\hat{X}_{HH})$

Muestras posibles	$\hat{V}(\hat{X}_{HH})$	$P(X_{HH})$	$\hat{V}(\hat{X}_{HH}) \cdot P(\hat{X}_{HH})$
u_1u_1	0	$\frac{9}{225}$	0
u_1u_2	$\frac{1}{2}[5^2+9^2-2\times 7^2]=4$	$\frac{30}{225}$	$\frac{8}{15}$
u_1u_3	$\frac{1}{2}\left[5^2+\left(\frac{60}{7}\right)^2-2\left(\frac{95}{14}\right)^2\right]=\frac{3.750}{1.176}$	$\frac{42}{225}$	$\frac{125}{210}$
u_2u_2	0	$\frac{25}{225}$	0
u_2u_3	$\frac{1}{2}\left[9^2+\left(\frac{60}{7}\right)^2-2\left(\frac{123}{14}\right)^2\right]=\frac{54}{1.176}$	$\frac{70}{225}$	$\frac{3}{210}$
u_3u_3	0	$\frac{49}{225}$	0
			$E\hat{V}(X_{HH})=\frac{240}{210}=$
			$=\frac{8}{7}=V(\hat{X}_{HH})$

1.11. Dada la población:

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
x_i	1	3	5	2	4	6	2	7

se desea obtener una muestra sistemática de tamaño 3, para lo cual se determina un período $K=\frac{N}{n}$, redondeado al entero más próximo, y un arranque aleatorio entre 1 y K. Se pide:

1.°) Determinar el espacio muestral y la función de probabilidad.

2.°) La probabilidad π_i que tiene la unidad u_i de pertenecer a la muestra.

3.º) Calcular $V(\hat{X})$, siendo $\hat{X}$ el estimador insesgado del total.

4.º) A partir de la distribución en el muestreo de $\hat{X}$, calcular $E(\hat{X})$ y $V(\hat{X})$.

Solución:

1.º) Siendo $K = \frac{8}{3} \simeq 3$, tendremos el siguiente espacio muestral:

Muestras posibles (S)	Función de probabilidad $P(S) = \frac{1}{K}$
$u_1u_4u_7$	$\frac{1}{3}$
$u_2u_5u_8$	$\frac{1}{3}$
u_3u_6	$\frac{1}{3}$

2.º) Puesto que cada unidad sólo puede pertenecer a una muestra, siendo todas las muestras equiprobables, la probabilidad de cualquier unidad será:

$$P(u_i) = \frac{1}{K} = \frac{1}{3}$$

3.º) El estimador insesgado del total es: $\hat{X} = Kx$, siendo x el total muestral. Su varianza es:

$$V(\hat{X}) = K \sum_s (x_s - \bar{x})^2$$

donde x_s es el total muestral proporcionado por la muestra s, y

$$\bar{x} = \frac{\sum^k x_s}{K}$$

De los datos del enunciado obtenemos:

$$x_{(1)} = 1+2+2 = 5 \ ; \ x_{(2)} = 3+4+7 = 14 \ ; \ x_{(3)} = 5+6 = 11 \ , \ \bar{x} = \frac{30}{3} = 10$$

Por tanto:

$$V(\hat{X}) = 3(5^2 + 4^2 + 1^2) = 126$$

4.º) La distribución en el muestreo de $\hat{X}$ se presenta a continuación:

(S)	$\hat{X}_s$	Prob $(\hat{X}_s) = P(s)$
$u_1u_4u_7$	$3(1 + 2 + 2) = 15$	$\frac{1}{3}$
$u_2u_5u_8$	$3(3 + 4 + 7) = 42$	$\frac{1}{3}$
u_3u_6	$3(5 + 6) = 33$	$\frac{1}{3}$

$$E(\hat{X}) = \sum_s (\hat{X}_s)P(\hat{X}_s) = \frac{1}{3}(15 + 42 + 33) = 30 = X, \text{ estimador insesgado}$$

$$V(\hat{X}) = \sum_s (\hat{X}_s^2)P(\hat{X}_s) - X^2 = \frac{1}{3}(15^2 + 42^2 + 33^2) - 30^2 = 126$$

1.12. En una población con $N = 3$ unidades, $u_i (i = 1, 2, 3)$, la variable t_i asociada a cada unidad toma los valores $(1; 3; 5)$.

Se considera un proceso de muestreo sin reposición, probabilidades iniciales $P_i(1/5; 2/5; 2/5)$, y tamaño muestral $n = 2$. Se pide:

a) Distribuciones en el muestreo de las variables ξ y η, sus medias y varianzas. $(\xi = t_i + t_j;\ \eta = \text{mín. } (t_i; t_j))$.

b) Comprobar que con este procedimiento de selección $nP_i \neq \pi_i$, siendo $\pi_i = P(u_i \in \text{muestra})$.

c) Distribución en el muestreo de la media $\alpha = \frac{t_i + t_j}{2}$, y comprobar que no es un estimador insesgado de la media poblacional.

Solución:

a) Espacio muestral: $(u_1; u_2)$, $(u_1; u_3)$, $(u_2; u_3)$

$$P_{12} = P(u_1; u_2) = P(u_1) \cdot P(u_2/u_1) + P(u_2) \cdot P(u_1/u_2) =$$

$$= \frac{1}{5} \cdot \frac{\frac{2}{5}}{1 - \frac{1}{5}} + \frac{2}{5} \cdot \frac{\frac{1}{5}}{1 - \frac{2}{5}} = \frac{7}{30}$$

y, análogamente, se obtienen

$$P_{13} = \frac{7}{30} \quad \text{y} \quad P_{23} = \frac{16}{30}$$

con $\sum_s P_{ij} = 1$ indicando con s las muestras posibles.

Las distribuciones en el muestreo de ξ y η, son:

Muestras posibles	Probabilidades	$\xi = t_i + t_j$	$\eta = \text{mín.}\ (t_i; t_j)$
(u_1u_2)	$\frac{7}{30}$	4	1
(u_1u_3)	$\frac{7}{30}$	6	1
(u_2u_3)	$\frac{16}{30}$	8	3
	1		

$$E[\xi] = \sum_s \xi P(s) = \frac{33}{5} = 6{,}6$$

$$V[\xi] = \sum_s (\xi - E(\xi))^2 P_s$$

$$V(\xi) = \left(4 - \frac{33}{5}\right)^2 \cdot \frac{7}{30} + \left(\frac{6 - 33}{5}\right)^2 \cdot \frac{7}{30} + \left(8 - \frac{33}{5}\right) \cdot \frac{16}{30} = 2{,}70$$

y análogamente:

$$E[\eta] = \frac{31}{15} = 2{,}0667 \quad y \quad V(\eta) = 0{,}9956$$

b)

$$\pi_1 = P_{12} + P_{13} = \frac{14}{30} \qquad nP_1 = \frac{2}{5} = \frac{12}{30}$$

$$\pi_2 = P_{12} + P_{23} = \frac{23}{30} \qquad nP_2 = \frac{4}{5} = \frac{24}{30}$$

$$\pi_3 = P_{13} + P_{23} = \frac{23}{30} \qquad nP_3 = \frac{4}{5} = \frac{24}{30}$$

vemos que $\pi_i \neq nP_i$ aunque $\sum \pi_i = n = 2 = \sum nP_i$.

c) Siendo $\alpha = \frac{1}{2}\xi$, $E(\alpha) = \frac{1}{2}E(\xi) = 3{,}3 \neq 3$.

CAPITULO II
Muestreo aleatorio simple

2.1. En una población de 10.000 viviendas se obtuvo una muestra aleatoria simple de $n = 11$ viviendas con el siguiente número de personas: 2, 3, 6, 1, 4, 3, 8, 2, 2, 1, 1.

Se pide:

1.º) Estimar el número de personas por vivienda y su error de muestreo.

2.º) Demostrar que la expresión:

$$\hat{V}(\bar{x}) = \frac{N-n}{Nn}\left[\frac{\sum_i^n x_i^2 - n\bar{x}^2}{n-1}\right] \qquad (1)$$

es un estimador insesgado de

$$V(\bar{x}) = \frac{N-n}{Nn}\left[\frac{\sum_i^N x_i^2 - N\bar{X}^2}{N-1}\right]$$

Solución:

1.º) En un muestreo aleatorio simple, designándose como tal el muestreo sin reposición y probabilidades iguales, un estimador insesgado de la media poblacional (en este caso, número medio de personas por vivienda) es la media muestral, por tanto:

$$\hat{\bar{X}} = \bar{x} = \frac{\sum_i^n x_i}{n} = \frac{33}{11} = 3$$

estima $\bar{X} = \dfrac{\sum_i^N x_i}{N}$ = número de personas por vivienda, siendo $N = 10.000$ el total de unidades de muestreo (viviendas) en la población.

La varianza de este estimador viene dada por:

$$V(\bar{x}) = \frac{N-n}{Nn}\left[\frac{\sum_{i}^{N} x_i^2 - N\bar{X}^2}{N-1}\right]$$

Se demuestra en el siguiente apartado que un estimador insesgado de esta varianza es:

$$\hat{V}(\bar{x}) = \frac{N-n}{Nn}\left[\frac{\sum_{i}^{n} x_i^2 - n\bar{x}^2}{n-1}\right]$$

que con los valores obtenidos en la muestra, resulta:

$$\hat{V}(\bar{x}) = \frac{10.000-11}{110.000}\left[\frac{149-11\times 9}{10}\right] = 0{,}454$$

El error de muestreo, definido como la desviación típica del estimador, se estima mediante:

$$\hat{\sigma}_{\bar{x}} = \sqrt{\hat{V}(\bar{x})} = \sqrt{0{,}454} = 0{,}67$$

A partir de este valor pueden construirse diferentes intervalos de confianza del estimador.

2.°) Si designamos por e_i una variable aleatoria que toma valores 1 ó 0, según la unidad u_i pertenezca o no pertenezca a la muestra, la esperanza matemática de esta variable es $Ee_i = n/N$, ya que la probabilidad de que u_i pertenezca a la muestra es, precisamente, n/N.

Tomando esperanzas en (1):

$$E[\hat{V}(\bar{x})] = \frac{N-n}{Nn}\frac{1}{n-1}\left[E\sum_{i}^{n} x_i^2 - nE\bar{x}^2\right] \qquad (2)$$

y

$$E\sum_{i}^{n} x_i^2 = E\sum_{i}^{N} x_i^2 e_i = \sum_{i}^{N} x_i^2 Ee_i = \frac{n}{N}\sum_{i}^{N} x_i^2$$

Por otra parte, si $\bar{x}$ y $\hat{V}(\bar{x})$ son estimadores insesgados de $\bar{X}$ y $V(\bar{x})$, respectivamente, resulta:

$$V(\bar{x}) = E\bar{x}^2 - \bar{X}^2$$

de donde

$$E\bar{x}^2 = V(\bar{x}) + \bar{X}^2 = E\hat{V}(\bar{x}) + \bar{X}^2$$

Sustituyendo en (2) los valores obtenidos de las esperanzas:

$$E[\hat{V}(\bar{x})] = \frac{N-n}{N}\,\frac{1}{n-1}\left[\frac{1}{N}\sum_{i}^{N} x_i^2 - E\hat{V}(\bar{x}) - \bar{X}^2\right]$$

y, finalmente, despejando $E\hat{V}(\bar{x})$, obtenemos:

$$E\hat{V}(\bar{x}) = \frac{N-n}{n(N-1)}\left[\frac{1}{N}\sum_{i}^{N} x_i^2 - \bar{X}^2\right] = \frac{N-n}{Nn}\left[\frac{\sum_{i}^{N} x_i^2 - N\bar{X}^2}{N-1}\right]$$

Nótese que el último término entre corchetes $\left[\dfrac{\sum_{i}^{N} x_i^2 - N\bar{X}^2}{N-1}\right]$ es la denominada cuasi-varianza poblacional, también definida por

$$S^2 = \frac{\sum_{i}^{N} (x_i - \bar{X})^2}{N-1}$$

Por tanto, un estimador insesgado de esta cuasi-varianza es la correspondiente cuasi-varianza muestral:

$$s^2 = \frac{\sum_{i}^{n} (x_i - \bar{x})^2}{n-1} = \left(\frac{\sum_{i}^{n} x_i^2 - n\bar{x}^2}{n-1}\right)$$

2.2. En una facultad $N = 200$ alumnos han aprobado por curso, mediante tres exámenes parciales una asignatura. Entre ellos, se efectúa una entrevista a una muestra de $n = 25$ alumnos, elegidos sin reposición y probabilidades iguales.

De las contestaciones, se obtienen entre otros, los siguientes resultados:

a) 18 alumnos están satisfechos con la enseñanza recibida en la citada asignatura.

b) El número de horas que han necesitado para preparar el último examen parcial viene dado por la siguiente distribución de frecuencias:

horas	x_i	16	20	30
alumnos	n_i	6	12	7

Se pide, estimar:

1.°) La proporción del número de alumnos satisfechos y su error de muestreo.

2.°) La media de horas por alumno dedicadas a preparar el último examen, su error de muestreo y el intervalo de confianza del 95 %.

Solución:

1.°) Siendo la proporción muestral un estimador insesgado de la proporción poblacional, resulta:

$$\hat{P} = \frac{18}{25} = 0{,}72$$

es decir, el 72 % de los alumnos aprobados por curso están satisfechos con la enseñanza recibida.

Un estimador insesgado de la varianza de $\hat{P}$, es:

$$\hat{V}(\hat{P}) = (1 - f)\frac{\hat{P}(1 - \hat{P})}{n - 1}$$

donde f es la fracción de muestreo aplicada. Por tanto

$$\hat{V}(\hat{P}) = \left(1 - \frac{25}{200}\right)\frac{0{,}72 \times 0{,}28}{24} = 0{,}00735$$

y

$$\hat{\sigma}_{\hat{P}} = \sqrt{\hat{V}(\hat{P})} = \sqrt{0{,}00735} = 0{,}0857$$

estima el error de muestreo.

2.º) Siendo, igualmente, la media muestral un estimador insesgado de la media poblacional, resulta:

$$\bar{x} = \frac{\sum_i n_i x_i}{\sum_i n_i} = \frac{540}{25} = 21{,}6 \text{ horas}$$

Un estimador insesgado de la varianza de $\bar{x}$ viene dado por:

$$\hat{V}(\bar{x}) = (1 - f)\frac{s^2}{n}$$

donde s^2 es la cuasivarianza muestral que toma el valor:

$$s^2 = \frac{\sum_i n_i (x_i - \bar{x})^2}{\sum_i n_i - 1} = \frac{786}{24} = 32{,}75$$

resultando:

$$\hat{V}(\bar{x}) = \left(1 - \frac{25}{200}\right)\frac{32{,}75}{25} = 1{,}14625$$

De donde $\hat{\sigma}_{\bar{x}} = \sqrt{1{,}14625} = 1{,}0706$, estima el error de muestreo. Suponiendo la normalidad de la distribución de la media muestral:

$$\text{Prob}\{\bar{x} - 2\sigma_{\bar{x}} \leqslant \bar{X} \leqslant \bar{x} + 2\sigma_{\bar{x}}\} = 0{,}95$$

siendo el intervalo de confianza:

$$\bar{x} \pm 2\hat{\sigma}_{\bar{x}} = (19{,}46;\ 23{,}74)$$

2.3. Para estimar el número total de unidades de la clase A en el dominio de estudio 1, se recomienda utilizar el estimador $\hat{A}_1 = N_1 \hat{P}_1$ si se conoce N_1, frente al estimador $\hat{A}'_1$ si no se conoce N_1 siendo $\hat{A}'_1 = N \times a_1/n$. Ignorando el factor de corrección en poblaciones finitas, demostrar que para muestras grandes la razón de varianza $V(\hat{A}_1)/V(\hat{A}'_1)$ es aproximadamente igual a $Q_1/(Q_1 + \pi P_1)$ donde π es la proporción de la población que no pertenece al dominio 1, y P_1 la proporción de las unidades en el dominio 1 que pertenecen a la clase A. Establecer bajo qué condiciones el conocimiento de N_1 produce mayor reducción en varianza. (Cochran W. C., 1967. Sampling Techniques.)

[Símbolos utilizados:

N = Total de unidades en la población.

N_1 = Total de unidades en el dominio de estudio 1.

n_1 = Unidades que pertenecen al dominio 1 en una muestra aleatoria de n unidades.

a_1 = Unidades que pertenecen a la clase A del dominio 1, en una muestra aleatoria de n unidades.

$$P_1 = \frac{A_1}{N_1} \quad ; \quad \hat{P}_1 = \frac{a_1}{n_1} \quad ; \quad Q_1 = 1 - P_1 \quad ; \quad \pi = 1 - \frac{N_1}{N}.\]$$

Solución:

Prescindiendo del factor de corrección de poblaciones finitas, las varianzas de los respectivos estimadores $\hat{A}_1$ y $\hat{A}'_1$ del total de clase en el dominio 1, vienen dadas por:

$$V(\hat{A}_1) = N_1^2 \frac{P_1 Q_1}{n_1} \quad , \quad V(\hat{A}'_1) = N^2 \frac{PQ}{n}$$

donde

$$P = A_1 N \quad \text{y} \quad Q = 1 - P$$

siendo el cociente de varianzas:

$$\frac{V(\hat{A}_1)}{V(\hat{A}'_1)} = \frac{\left(\frac{N_1}{N}\right)^2 P_1 Q_1}{\left(\frac{n_1}{n}\right) PQ} \tag{1}$$

Ahora bien,

$$P = \frac{A_1}{N} = \frac{N_1 P_1}{N} = (1 - \pi) P_1 \quad \text{y} \quad Q = 1 - (1 - \pi) P_1 = Q_1 + \pi P_1$$

Por otra parte, en muestras grandes

$$\frac{n}{n_1} \simeq \frac{N}{N_1} = \frac{1}{1 - \pi}$$

Sustituyendo estos valores en (1), resulta:

$$\frac{V(\hat{A}_1)}{V(\hat{A}'_1)} \simeq \frac{(1-\pi)P_1Q_1}{(1-\pi)P_1[Q_1+\pi P_1]} = \frac{Q_1}{Q_1+\pi P_1}$$

Este cociente será tanto menor (y, por tanto, la reducción en varianza que produce el conocimiento de N_1 tanto mayor) cuanto mayor sea P_1.

2.4. En un área existen $N = 10.000$ viviendas. Los datos de un censo anterior hacen suponer que, aproximadamente, los 2/3 corresponde a régimen de alquiler.

Se pide: El tamaño de muestra necesario para estimar la proporción actual de viviendas en alquiler, con un error de muestreo igual a 0,04, en caso de:

a) Muestreo con reposición y probabilidades iguales.

b) Muestreo aleatorio simple (muestreo sin reposición y probabilidades iguales).

Solución:

a) En caso de muestreo con reposición, la varianza de una proporción viene dada por

$$V(p) = \frac{P(1-P)}{n}$$

siendo P la proporción poblacional y n el tamaño de muestra.

La precisión requerida nos conduce a la ecuación:

$$\text{error } \sigma_p = \sqrt{V(p)} = 0{,}04$$

es decir

$$\frac{P(1-P)}{n} = 0{,}0016$$

Si conjeturamos $P = \frac{2}{3}$, resulta

$$n = \frac{\frac{2}{3} \times \frac{1}{3}}{0{,}0016} = 139$$

Puede observarse que, en este caso, el tamaño de muestra no depende del tamaño de población N.

b) En caso de muestreo sin reposición, interviene el factor de corrección de poblaciones finitas en la varianza de p, siendo

$$V(p) = \frac{N - n}{N - 1} \frac{P(1 - P)}{n}$$

de donde

$$n = \frac{NP(1 - P)}{P(1 - P) + (N - 1)V(p)}$$

Bajo los supuestos anteriores:

$$P = \frac{2}{3} \quad \text{y} \quad \text{error } \sigma_p = \sqrt{V(p)} = 0{,}04$$

resulta

$$n = \frac{10.000 \times \frac{2}{3} \times \frac{1}{3}}{\frac{2}{3} \times \frac{1}{3} + 9.999 \times 0{,}0016} = 137$$

puede observarse una ligera reducción en el tamaño de muestra, consecuencia de una mayor precisión del muestreo sin reposición.

2.5. El número de viviendas de un municipio es igual a 5.200. Se pide:

1.º) Calcular el tamaño de muestra necesario para estimar el número de viviendas desocupadas con un error de muestreo igual a 10, sabiendo que una encuesta piloto ha mostrado que la proporción de viviendas desocupadas era 0,12. ¿Cuál sería el tamaño de muestra si el error de muestreo fuese igual a 30? Se utiliza muestreo aleatorio simple.

2.º) ¿Cuáles serían los tamaños de muestra bajo los supuestos anteriores si se utiliza muestreo con reposición?

Solución:

1.º) El error de muestreo, $\sigma_{\hat{X}}$, del estimador de un total de clase y el tamaño de muestra vienen relacionados a través de la expresión:

$$\sigma_{\hat{X}}^2 = N^2 \frac{N-n}{N-1} \frac{P(1-P)}{n} \quad , \quad \text{siendo} \begin{cases} N = \text{tamaño de la población} \\ P = \text{proporción poblacional} \\ n = \text{tamaño de muestra} \end{cases}$$

despejando n, resulta:

$$n = \frac{N^3 P(1-P)}{\sigma_{\hat{X}}^2 (N-1) + N^2 P(1-P)} \qquad (1)$$

si hacemos $\sigma_{\hat{X}} = 10$ y sustituimos P por el valor conjeturado 0,12, obtenemos:

$$n = \frac{5.200^3 (0{,}12 \times 0{,}88)}{10^2 (5.200-1) + 5.200^2 (0{,}12 \times 0{,}88)} = 4.399 \text{ viviendas}$$

el elevado tamaño de muestra es debido a la alta precisión con que se requieren los resultados, ya que un error de muestreo igual a 10 equivale a estimar una proporción de 0,12 con un error de muestreo de 0,002.

Si la precisión es menor y hacemos $\sigma_{\hat{X}} = 30$, aplicando la fórmula (1) se obtiene:

$$n = \frac{5.200^3 (0{,}12 \times 0{,}88)}{30^2 (5.200-1) + 5.200^2 (0{,}12 \times 0{,}88)} = 1.971 \text{ viviendas}$$

Si observamos los tamaños de muestra requeridos por las distintas precisiones, notaremos que está lejos de cumplirse la teoría del muestreo en poblaciones infinitas que determina que el error de muestreo disminuye proporcionalmente a la raíz cuadrada del incremento de muestra o, dicho de otra manera, un error de muestreo 3 veces mayor exigirá una muestra 9 veces menor. El no cumplimiento de este principio es debido a la influencia del factor de corrección de poblaciones finitas $\frac{N-n}{N-1}$, tanto más relevante cuanto mayor sea la muestra en relación con la población.

2.º) En caso de muestreo con reposición:

$$\sigma_{\hat{X}}^2 = N^2 \frac{P(1-P)}{n}, \quad \text{de donde} \quad n = \frac{N^2 P(1-P)}{\sigma_{\hat{X}}^2}$$

si hacemos $\sigma_{\bar{X}} = 10$, y sustituimos P por el valor conjeturado 0,12, obtenemos:

$$n = \frac{5.200^2 \times 0{,}12 \times 0{,}88}{10^2} = 28.554 \text{ viviendas}$$

Se observará que resulta un tamaño de muestra superior al tamaño de población, lo que hace que este muestreo, semejante al muestreo de poblaciones infinitas, no sea práctico cuando se requiere una alta precisión y, por tanto, un elevado tamaño de muestra.

Para $\sigma_{\bar{X}} = 30$, resulta

$$n = \frac{5.200^2 \times 0{,}12 \times 0{,}88}{30^2} = 3.173 \text{ viviendas}$$

Resulta palpable la menor precisión de este muestreo en relación con el anterior, ya que a igualdad de error de muestreo exige un tamaño de muestra considerablemente mayor. Sin embargo, el tamaño de muestra es mucho más sensible a variaciones en la precisión requerida.

2.6. En una zona de 1.000 viviendas determinar el tamaño de la muestra necesario para que, con un grado de confianza del 95 %, la estimación de la proporción de viviendas sin agua corriente no difiera en más de 0,10 del valor verdadero. Se utiliza muestreo aleatorio simple.

Solución:

Bajo la hipótesis de normalidad en la distribución del estimador, si designamos por $\hat{P}$ el estimador de la proporción de viviendas sin agua corriente, ha de ser:

$$2\sigma_{\hat{P}} = 0{,}10$$

donde

$$\sigma_{\hat{P}} = \sqrt{\frac{N-n}{N-1}\frac{P(1-P)}{n}} \tag{1}$$

y, por tanto,

$$n = \frac{NP(1-P)}{\sigma_{\hat{P}}^2(N-1) + P(1-P)} \tag{2}$$

Si no disponemos de ninguna conjetura acerca del valor de P, adoptaremos la actitud conservadora de ponernos en el caso más desfavorable que es considerar el valor de $P = 1/2$, para el cual el valor del error de muestreo dado en (1) se hace máximo.

Según lo anterior, sustituyendo en (2) los correspondientes valores de N, P y σ_p requerido, resulta:

$$n = \frac{1.000 \times \frac{1}{2} \times \frac{1}{2}}{0{,}05^2(1.000 - 1) + \frac{1}{2} \times \frac{1}{2}} = 91 \text{ viviendas}$$

2.7. Dada una población de $N = 1.000$ establecimientos que se dedican a la producción de un determinado artículo, se desea conocer el tamaño n de la muestra necesario para estimar la producción total de modo que la estimación quede dentro del 10 % del valor del parámetro con una confianza de 0,95. Se utiliza muestreo aleatorio simple y se conoce mediante una encuesta piloto que el coeficiente de variación de la población es 0,6.

Solución:

Según las condiciones establecidas en el problema designando por X la producción total, ha de ocurrir que:

$$\text{Prob}\,(|\hat{X} - X| \leqslant 0{,}10X) = 0{,}95 \qquad (1)$$

Por otra parte, supuesta la normalidad del estimador $\hat{X}$, se tiene que:

$$\text{Prob}\,(|\hat{X} - X| \leqslant 2\sigma_{\hat{X}}) = 0{,}95 \qquad (2)$$

donde $\sigma_{\hat{X}}$ es el error de muestreo o desviación típica del estimador.

De las relaciones (1) y (2), resulta la ecuación básica:

$$0{,}10X = 2\sigma_{\hat{X}} \qquad (3)$$

siendo

$$\sigma^2_{\hat{X}} = N^2 \frac{N - n}{N - 1} \frac{\sigma^2}{n}$$

donde σ^2 es la varianza de la población.

El conocimiento del coeficiente de variación poblacional conduce a la relación:

$$CV = \frac{\sigma}{\bar{X}} = 0{,}6$$

siendo

$$\bar{X} = \frac{X}{N}$$

la media poblacional. Según lo anterior, la relación (3) puede transformarse en:

$$(0{,}10N\bar{X})^2 = 4N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n}$$

es decir

$$0{,}10^2 = 4 \frac{N-n}{N-1} \frac{(CV)^2}{n}$$

y, finalmente, despejando n y sustituyendo los valores conocidos, se obtiene:

$$n = \frac{4N(CV)^2}{0{,}10^2(N-1) + 4(CV)^2} = \frac{4 \times 1.000 \times 0{,}6^2}{0{,}10^2(1.000-1) + 4 \times 0{,}6^2} = 126$$

establecimientos.

2.8. En un muestreo aleatorio simple, determinar el tamaño de muestra necesario para estimar la producción total de 500 fábricas dedicadas a un determinado producto, con un coeficiente de variación $C(\hat{X})$ para el estimador igual al 5 %, sabiendo que en un estudio anterior se encontró un coeficiente de variación en la población igual a 115 %.

Solución:

Si $\hat{X}$ es un estimador insesgado de la producción total X, por definición:

$$C(\hat{X}) = \frac{\sigma_{\hat{X}}}{E(\hat{X})} = \frac{\sigma_{\hat{X}}}{X}$$

Por otra parte, en un muestreo aleatorio simple:

$$\sigma_{\bar{X}} = \sqrt{N^2 \frac{N-n}{N-1} \frac{\sigma^2}{n}}$$

dividiendo por X en ambos miembros, se obtiene:

$$\frac{\sigma_{\bar{X}}}{X} = \frac{N\sigma}{X} \sqrt{\frac{N-n}{(N-1)n}}$$

sustituyendo en esta ecuación los valores dados en el enunciado del problema:

$$\frac{\sigma_{\bar{X}}}{X} = 0{,}05 \quad \text{y} \quad \frac{N\sigma}{X} = \frac{\sigma}{\bar{X}} = 1{,}15$$

el tamaño de muestra ha de satisfacer la ecuación:

$$0{,}05 = 1{,}15 \sqrt{\frac{500-n}{499n}}$$

resultando:

$$n = \frac{500 \times 1{,}15^2}{499 \times 0{,}05^2 + 1{,}15^2} = 257 \text{ fábricas}$$

2.9. Se quiere estimar la producción total de cada uno de dos tipos de artículos *a* y *b* fabricados en 1.000 establecimientos (que producen ambos tipos) con un error de muestreo en cada estimador del total inferior a 300.000 unidades, mediante un muestreo aleatorio sin reposición. Las cuasivarianzas poblacionales S_a^2 y S_b^2 de las producciones vienen determinadas por $S_a = 5.000$ y $S_b = 4.000$, según las estimaciones realizadas en una ocasión anterior, y se cree que no han variado para la presente.

El coste de trabajo de campo supone:

1) Gastos de desplazamiento a cada fábrica visitada ..... 1.000 pts.

2) Gastos de determinación de la producción del artículo *a* en cada fábrica en que se investigue la misma ..... 500 pts.

3) Idem para el artículo *b* ..... 400 pts.

Se sabe que la producción de los artículos *a* y *b*, dentro de cada fábrica y en conjunto, son totalmente independientes.

¿Cuántas fábricas deberán incluirse en la muestra para conocer en ellas:

1.° la producción del artículo *a* exclusivamente,

2.° la producción del artículo *b* exclusivamente,

3.° la producción de ambos artículos,

para que el coste sea el menor posible? ¿Cuál es este coste?
(Propuesto en las Oposiciones al Cuerpo de Estadísticos Técnicos Diplomados.)

Solución:

1.°) Si designamos por $\hat{a}$ el estimador de la producción total del artículo *a*, y n_a el número de fábricas de la muestra donde se ha de investigar la producción de dicho artículo, el error de muestreo y el tamaño de muestra vienen relacionados a través de la expresión:

$$\sigma^2_{\hat{a}} = N^2\left(1 - \frac{n_a}{N}\right)\frac{S_a^2}{n_a}$$

donde conocemos:

$$\sigma_{\hat{a}} = 300.000 \quad ; \quad N = 1.000 \quad ; \quad S_a = 5.000$$

por tanto:

$$n_a = \frac{N^2 S_a^2}{\sigma^2_{\hat{a}} + NS_a^2} = \frac{(1.000 \times 5.000)^2}{300.000^2 + 1.000 \times 5.000^2} = 217 \text{ establecimientos}$$

2.°) Análogamente para el artículo *b*:

$$n_b = \frac{N^2 S_b^2}{\sigma^2_{\hat{b}} + NS_b^2} = \frac{(1.000 \times 4.000)^2}{300.000^2 + 1.000 \times 4.000^2} = 151 \text{ establecimientos}$$

3.°) De acuerdo con las cifras anteriores se seleccionarían aleatoriamente 151 fábricas donde se investigaría la producción de ambos artículos y 66 fábricas adicionales donde únicamente se investigaría la producción del artículo *a*.

El coste de esta investigación seria:

Para cada fábrica donde se investigan ambos artículos:

$$1.000 + 500 + 400 = 1.900 \text{ pts.}$$

Para cada fábrica donde sólo se investiga el artículo *a*:

$$1.000 + 500 = 1.500 \text{ pts.}$$

Coste total:

$$151 \times 1.900 + 66 \times 1.500 = 385.900 \text{ pts.}$$

2.10. Se supone que en una población de 100.000 habitantes existen dos enfermedades: una en la zona periférica, que se cree afecta al 1 por 100 de sus habitantes, y otra que afecta al resto, con una incidencia que se cree en torno al 20 %. Sabiendo que en la zona periférica residen 30.000 personas.

Se pide: Determinar el tamaño de muestra en cada zona para que, utilizando un muestreo aleatorio simple, los errores de muestreo relativos al estimar las proporciones de enfermos no sean superiores al 3 y 2 %, respectivamente.

Solución:

En un muestreo aleatorio simple el error de muestreo de una proporción P y el tamaño de muestra vienen relacionados a través de la expresión:

$$\sigma_{\hat{p}} = \sqrt{\frac{N-n}{N-1}\frac{P(1-P)}{n}}$$

y dividiendo por P obtenemos la expresión en función del error relativo:

$$\frac{\sigma_{\hat{p}}}{P} = \sqrt{\frac{N-n}{N-1}\frac{(1-P)}{nP}} \qquad (1)$$

Designando mediante el subíndice 1 los datos relativos a la zona periférica, la relación (1) se transforma en:

$$\frac{\sigma_{\hat{p}_1}}{P_1} = \sqrt{\frac{N_1-n_1}{N_1-1}\frac{(1-P_1)}{n_1P_1}}$$

si P_1 lo sustituimos por el valor conjeturado 0,01 y aplicamos los valores dados en el enunciado:

$$\frac{\sigma_{\hat{p}_1}}{P_1} = 0{,}03 \quad ; \quad N_1 = 30.000$$

resulta la siguiente ecuación que determinará el tamaño de muestra requerido:

$$0{,}03 = \sqrt{\frac{30.000 - n_1}{(30.000 - 1)n_1} \cdot \frac{1 - 0{,}01}{0{,}01}}$$

de donde

$$n_1 = \frac{30.000 \times 99}{0{,}03^2 \times 29.999 + 99} = 23.572 \text{ habitantes}$$

el elevado tamaño de muestra, casi un 80 % de la población, es debido a la rareza de la característica a estimar y al alto grado de precisión requerido. Análogamente, designando mediante el subíndice 2 la zona restante, obtenemos la relación:

$$\frac{\sigma_{\hat{P}_2}}{P_2} = \sqrt{\frac{N_2 - n_2}{N_2 - 1} \frac{(1 - P_2)}{n_2 P_2}}$$

y sustituyendo los correspondientes valores dados en el enunciado:

$$\frac{\sigma_{\hat{P}_2}}{P_2} = 0{,}02 \quad ; \quad N_2 = 70.000 \quad ; \quad P_2 = 0{,}20$$

resulta la siguiente ecuación que determinará el tamaño de muestra:

$$0{,}02 = \sqrt{\frac{70.000 - n_2}{(70.000 - 1)n_2} \cdot \frac{(1 - 0{,}2)}{0{,}2}}$$

de donde

$$n_2 = \frac{70.000 \times 4}{0{,}02^2 \times 69.999 + 4} = 8.750 \text{ habitantes}$$

2.11. Para estimar el volumen de ventas del millón de comercios de un país se ha seleccionado una muestra aleatoria, con probabilidades iguales y sin reemplazamiento, con una fracción de muestreo del medio por ciento extendida a todas las provincias del país. La precisión lograda ha sido satisfactoria a nivel nacional. Sin embargo, para una determinada provincia, con un tamaño poblacional de 20.000 comercios y una varianza estimada en el volumen de ventas del 80 % de la varianza nacional, se desea obtener la estimación provincial con idéntica precisión relativa. ¿Es suficiente la fracción de muestreo del medio por

ciento seleccionada para esa provincia? En caso contrario, ¿qué fracción de muestreo se precisaría?

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1988.)

Solución:

La igualdad de las precisiones relativas en la estimación del volumen de ventas nacional y provincial requiere la igualdad de las varianzas del estimador del volumen medio de ventas, suponiendo que no existen diferencias significativas en las medias nacional y provincial. Es decir, si designamos por X la característica volumen de ventas, N y n los tamaños respectivos de población y muestra, y utilizamos los subíndices 1 y 2 para expresar los niveles nacional y provincial, respectivamente, la igualdad de las precisiones relativas exige que:

$$\hat{V}(\hat{\bar{X}}_1) = \hat{V}(\hat{\bar{X}}_2)$$

donde

$$\hat{V}(\hat{\bar{X}}_1) = (1 - f_1)\frac{\hat{S}_1^2}{n_1} = \frac{1 - f_1}{f_1}\frac{\hat{S}_1^2}{N_1} \qquad (1)$$

$$\hat{V}(\hat{\bar{X}}_2) = (1 - f_2)\frac{\hat{S}_2^2}{n_2} = \frac{1 - f_2}{f_2}\frac{\hat{S}_2^2}{N_2} \qquad (2)$$

Como $\hat{S}_2^2 = 0{,}8\hat{S}_1^2$, las fracciones de muestreo tendrían que ser iguales sólo si fuese

$$N_2 = 0{,}8N_1$$

Puesto que

$$N_2 = 0{,}02N_1 \quad \text{ha de ser} \quad f_2 > f_1$$

Dividiendo (1) y (2), los valores de f_1 y f_2 que igualan ambas ecuaciones han de cumplir la condición:

$$\frac{1 - f_1}{f_1} = 40 \cdot \frac{1 - f_2}{f_2}$$

de donde

$$f_2 = \frac{40f_1}{1 + 39f_1}$$

sustituyendo f_1 por su valor 0,005, se obtiene la fracción de muestreo requerida a nivel provincial:

$$f_2 = \frac{40(0{,}005)}{1 + 39(0{,}005)} = 0{,}167 \simeq 17\,\%$$

es decir, la fracción de muestreo a nivel provincial ha de ser 34 veces mayor que la nacional.

CAPITULO III
Muestreo estratificado

3.1. Dos estadísticos A y B investigaron el estado de 200 cuestionarios. El estadístico A seleccionó una muestra aleatoria con reemplazamiento de 20 cuestionarios y contó el número de errores por cuestionario con los siguientes resultados:

Número de errores por cuestionario:	0	1	2	3	4	5	6	7	8	9	10
Número de cuestionarios:	8	4	2	2	1	1	0	0	0	1	1

El estadístico B examinó los 200 cuestionarios, registrando únicamente aquéllos que no tenían ningún error, encontrando 60 cuestionarios sin error.

Estimar el número total de errores utilizando un estimador de expansión y:

a) Sólo los resultados del estadístico A.

b) Los resultados de ambos estadísticos.

c) ¿Son insesgados los estimadores? ¿Qué estimador tiene más precisión?

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos.)

Solución:

a) Si designamos por n_i el número de cuestionarios en la muestra con i errores, y N el número de cuestionarios en la población, el estimador insesgado del total en un muestreo aleatorio con reposición, utilizando los resultados del estadístico A, viene dado por:

$$\hat{X}_A = N\bar{x} = N\frac{\sum_{i=0}^{10} in_i}{\sum_i n_i} = 200 \times \frac{42}{20} = 420 \text{ errores}$$

b) Utilizando los datos del estadístico B podemos considerar la población dividida en dos estratos:

Estrato 1: Cuestionarios sin error, $N_1 = 60$.

Estrato 2: Cuestionarios con algún error, $N_2 = 140$.

Según los datos del estadístico A, del estrato 2 se han investigado 12 cuestionarios, con media:

$$\bar{x}_2 = \frac{\sum i_2 n_{i2}}{\sum n_{i2}} = \frac{42}{12} = 3{,}5 \text{ errores por cuestionario}$$

El estimador insesgado del total en el muestreo estratificado viene dado por:

$$\hat{X}_{AB} = \sum_h N_h \bar{x}_h = 60 \times 0 + 140 \times 3{,}5 = 490 \text{ errores}$$

c) Por propia definición los estimadores $\hat{X}_A$ y $\hat{X}_{AB}$ son insesgados.
Para estudiar la precisión comparemos sus varianzas:

$$V(\hat{X}_A) = N^2 \frac{\sigma^2}{n}$$

$$V(\hat{X}_{AB}) = \sum_h N_h^2 \frac{\sigma_h^2}{n_h} = N_2^2 \frac{\sigma_2^2}{n_2} \qquad (1)$$

ya que $\sigma_1^2 = 0$ (varianza del número de errores en el estrato con cero errores).
Por otra parte, aplicando la ecuación fundamental del análisis de la varianza en el muestreo estratificado:

$$N\sigma^2 = \sum_h N_h \sigma_h^2 + \sum_h N_h (\bar{X}_h - \bar{X})^2$$

Puesto que:

$$\sigma_1^2 = 0 \quad \text{y} \quad \bar{X}_1 = 0$$

se obtiene:

$$\sigma^2 = w_2 \sigma_2^2 + w_2(1 - w_2)\bar{X}_2^2 \qquad (2)$$

donde

$$w_2 = \frac{N_2}{N}$$

Si en (1) hacemos $N_2 = Nw_2$, y sustituimos n_2 *por su esperanza matemática* nw_2, *tenemos*:

$$V(\hat{X}_{AB}) = N^2 \frac{w_2\sigma_2^2}{n} < V(\hat{X}_A)$$

ya que según (2):

$$w_2\sigma_2^2 = \sigma^2 - w_2(1 - w_2)\bar{X}_2^2$$

luego el estimador $\hat{X}_{AB}$ es más preciso que el estimador $\hat{X}_A$.

[*Nota:* En el estudio de las precisiones se ha considerado n_2 como constante, prescindiendo del efecto de la estratificación *a posteriori* que tendría la varianza de $\hat{X}_{AB}$ si se considerase n_2 variable aleatoria.]

3.2. Una población se divide en dos estratos de igual tamaño, de los que se obtienen muestras aleatorias simples. En el supuesto de afijación proporcional y una fracción de muestreo global igual al 5 %. ¿Qué tamaño n de muestra es necesario tomar para obtener un error de muestreo para la media igual a 0,5? Un estudio piloto ha mostrado los siguientes valores para las cuasivarianzas de los estratos: $S_1^2 = 25$; $S_2^2 = 15$.

Solución:

Si en la expresión general de la varianza de la media en un muestreo estratificado:

$$V(\bar{x}) = \sum_h (1 - f_h) \frac{S_h^2}{n_h} W_h^2$$

donde

$$W_h = \frac{N_h}{N} = \frac{1}{2}$$

se aplica el criterio de afijación proporcional:

$$n_h = nW_h \quad \text{y} \quad 1 - f_h = 1 - f$$

se obtiene:

$$n = (1 - f)\frac{\sum S_h^2 W_h}{V(\bar{x})} \qquad (1)$$

La precisión requerida exige: $\sigma_{\bar{x}} = 0{,}5$; es decir, $V(\bar{x}) = 0{,}25$.
Por tanto, sustituyendo valores en (1):

$$n = (1 - 0{,}05)\frac{\frac{1}{2}(25 + 15)}{0{,}25} = 76$$

es decir, 38 unidades en cada estrato.

3.3. Para una muestra $n = 1.000$ y una función de coste

$$c = 2\sqrt{n_1} + 3\sqrt{n_2}$$

determinar n_1 y n_2 tal que hagan mínima $V(\bar{x}_{st})$ bajo las condiciones siguientes:

Estratos	W_h	S_h
1	0,5	3
2	0,5	5

($\bar{x}_{st}$ designa el estimador insesgado de la media en un muestreo estratificado).

Solución:

Siendo la varianza de la media en el muestreo estratificado:

$$V(\bar{x}_{st}) = \sum_h W_h^2(1 - f_h)\frac{S_h^2}{n_h}$$

se trata de minimizar dicha expresión para los valores de n_1 y n_2 con la condición impuesta por la función de coste. Para ello, formamos la función de Lagrange:

$$\Phi = \sum_h W_h^2(1 - f_h)\frac{S_h^2}{n_h} + \lambda(c - 2\sqrt{n_1} - 3\sqrt{n_2})$$

derivando respecto a n_1 y n_2:

$$\frac{\partial \Phi}{\partial n_1} = -\frac{W_1^2 S_1^2}{n_1^2} - \frac{2\lambda}{2\sqrt{n_1}} = 0; \text{ de donde } \lambda = -\frac{W_1^2 S_1^2 \sqrt{n_1}}{n_1^2} = -\frac{9}{4\sqrt{n_1^3}}$$

$$\frac{\partial \Phi}{\partial n_2} = -\frac{W_2^2 S_2^2}{n_2^2} - \frac{3\lambda}{2\sqrt{n_2}} = 0; \text{ de donde } \lambda = -\frac{2W_2^2 S_2^2 \sqrt{n_2}}{3n_2^2} = -\frac{25}{6\sqrt{n_2^3}}$$

y eliminando λ:

$$\sqrt{\left(\frac{n_1}{n_2}\right)^3} = \frac{9 \times 6}{25 \times 4} = 0{,}54 \quad ; \quad \frac{n_1}{n_2} = \sqrt[3]{0{,}54^2} = 0{,}663$$

Finalmente, resolviendo el sistema:

$$\left.\begin{aligned} n_1 + n_2 &= 1.000 \\ \frac{n_1}{n_2} &= 0{,}663 \end{aligned}\right\}$$

se obtiene, $n_1 = 399$; $n_2 = 601$.

3.4. En un país formado por L provincias se llevó a cabo una encuesta por muestreo consultando a n_h ocupantes de fincas en la provincia h-ésima ($h = 1, 2, ..., L$). Se sabe que el número de consultados era el que daba menos error de muestreo al estimar una cierta característica X que se desea volver a estimar en una próxima ocasión.

En esta nueva ocasión el costo por entrevista en la provincia h-ésima es de C_h pesetas. (En el caso anterior no se tuvieron en cuenta los costos.) Se supone que las varianzas poblacionales de X no han variado, así como tampoco el número de fincas.

Se dispone para realizar la nueva encuesta de un presupuesto total de C pesetas, y se desea saber cuáles deben ser los nuevos tamaños de la muestra n'_h en cada provincia, en relación a los antiguos tamaños n_h, para tener el menor error de muestreo posible.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Técnicos Diplomados.)

Solución:

La afijación óptima, teniendo en cuenta una función lineal de costes del tipo $C = \sum_h n'_h C_h$, determina los tamaños de muestra en cada estrato en función del

tamaño poblacional de cada estrato, su cuasivarianza y el coste unitario, mediante la expresión:

$$n'_h = n' \frac{\frac{N'_h S'_h}{\sqrt{C_h}}}{\sum_h \frac{N'_h S'_h}{\sqrt{C_h}}} \tag{1}$$

En la ocasión anterior, la afijación óptima sin tener en cuenta costes, condujo a:

$$n_h = n \frac{N_h S_h}{\sum_h N_h S_h} \tag{2}$$

Puesto que $N'_h = N_h$ y $S'_h = S_h$, dividiendo (1) y (2), se obtiene:

$$\frac{n'_h}{n_h} = \frac{1}{\sqrt{C_h}} \cdot \frac{\frac{n'}{\sum_h \frac{N_h S_h}{\sqrt{C_h}}}}{\frac{n}{\sum_h N_h S_h}} = \frac{1}{\sqrt{C_h}} \cdot \frac{K_1}{K_2} \tag{3}$$

siendo

$$K_1 = \frac{n'}{\sum \frac{N_h S_h}{\sqrt{C_h}}} \quad \text{y} \quad K_2 = \frac{n}{\sum N_h S_h}$$

Puesto que K_1 y K_2 no son valores conocidos explícitamente en el problema, se ha de buscar alguna expresión de los mismos en función de los valores conocidos C, C_h y n_h.

De (3)

$$n'_h = \frac{n_h}{\sqrt{C_h}} \cdot \frac{K_1}{K_2}$$

si multiplicamos ambos términos por C_h y sumamos en h, obtenemos:

$$C = \sum_h n'_h C_h = \frac{K_1}{K_2} \sum_h n_h \sqrt{C_h}$$

es decir

$$\frac{K_1}{K_2} = \frac{C}{\sum_h n_h \sqrt{C_h}}$$

sustituyendo este valor en (3), se obtiene finalmente:

$$n'_h = n_h \frac{C}{\sqrt{C_h} \sum_h n_h \sqrt{C_h}}$$

3.5. De una población con $N = 6$ unidades, se obtiene una muestra de $n = 2$ unidades sin reposición y con probabilidades iguales.

Posteriormente se agregan las unidades en dos estratos:

Estratos	X_{hi}
I	0, 1, 3
II	5, 6, 9

Se pide:

a) Varianza del estimador de la media $V(\bar{x})$, sin estratificar.

b) Varianza de la media con estratificación $V(\bar{x}_{st})$, y fracción de muestreo $f_h = 1/3$, en cada estrato.

c) ¿Cuáles serán los valores de K_1 y K_2 $\left(\text{siendo } K_h = \frac{n_h}{n}\right)$ que minimizan el coste total, sabiendo que el coste por unidad en el primer estrato $c_1 = 4$ es doble que en el segundo?

d) Determinar $n = n_1 + n_2$ para $V(\bar{x}_{st}) = 0{,}6$.

Solución:

a) En un muestreo aleatorio simple la varianza de la media muestral $\bar{x}$, se obtiene mediante la expresión:

$$V(\bar{x}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

siendo S^2 la cuasivarianza poblacional que toma el valor:

$$S^2 = \frac{\sum_i X_i^2 - N\bar{X}^2}{N-1} = \frac{152 - 6 \times \left(\frac{24}{6}\right)^2}{6-1} = \frac{56}{5}$$

Por tanto,

$$V(\bar{x}) = \left(1 - \frac{2}{6}\right)\frac{56}{10} = 3{,}73$$

b) En el muestreo estratificado,

$$V(\bar{x}_{st}) = \sum_h \left(\frac{N_h}{N}\right)^2 (1 - f_h)\frac{S_h^2}{n_h}$$

designando por S_h^2 la cuasivarianza del estrato h que, en este caso, toma los valores:

$$S_1^2 = \frac{\sum_i X_{1i}^2 - N_1\bar{X}_1^2}{N_1 - 1} = \frac{10 - 3\left(\frac{4}{3}\right)^2}{3-1} = \frac{7}{3}$$

$$S_2^2 = \frac{\sum_i X_{2i}^2 - N_2\bar{X}_2^2}{N_2 - 1} = \frac{142 - 3\left(\frac{20}{3}\right)^2}{3-1} = \frac{13}{3}$$

Por tanto,

$$V(\bar{x}_{st}) = \frac{1}{4} \times \frac{2}{3}\left(\frac{7}{3} + \frac{13}{3}\right) = \frac{10}{9} = 1{,}1$$

La fracción de muestreo $f_h = \frac{1}{3}$ equivale a seleccionar 1 unidad en cada estrato. Comparada esta varianza con la del muestreo aleatorio simple puede observarse una considerable ganancia en precisión debida a la estratificación.

c) Se trata de calcular los valores de n_1 y n_2 que minimizan la función de Lagrange:

$$\varphi = c_1 n_1 + c_2 n_2 + \lambda\left[\sum_h W_h^2(1 - f_h)\frac{S_h^2}{n_h} - V(\bar{x}_{st})\right]$$

donde

$$W_h = \frac{N_h}{N}$$

expresa el tamaño relativo del estrato h. Derivando respecto a n_1 y n_2:

$$\left.\begin{aligned} \frac{\partial \varphi}{\partial n_1} &= c_1 - \lambda \frac{W_1^2 S_1^2}{n_1^2} = 0 \\ \frac{\partial \varphi}{\partial n_2} &= c_2 - \lambda \frac{W_2^2 S_2^2}{n_2^2} = 0 \end{aligned}\right\} \quad \begin{aligned} n_1 &= \sqrt{\lambda}\frac{W_1 S_1}{\sqrt{c_1}} \\ n_2 &= \sqrt{\lambda}\frac{W_2 S_2}{\sqrt{c_2}} \end{aligned}$$

$$n = n_1 + n_2 = \sqrt{\lambda} \sum_h \frac{W_h S_h}{\sqrt{c_h}}$$

Por tanto:

$$K_1 = \frac{n_1}{n} = \frac{\dfrac{W_1 S_1}{\sqrt{c_1}}}{\displaystyle\sum_h \frac{W_h S_h}{\sqrt{c_h}}} = \frac{\dfrac{1}{2} \times \sqrt{\dfrac{7}{3}} \times \dfrac{1}{2}}{\dfrac{1}{2}\left[\dfrac{\sqrt{7/3}}{2} + \dfrac{\sqrt{13/3}}{\sqrt{2}}\right]} = 0{,}34 \quad ; \quad K_2 = 1 - K_1 = 0{,}66$$

d) El valor de n se obtiene de la expresión:

$$V(\bar{x}_{st}) = \sum_h W_h^2 (1 - f_h) \frac{S_h^2}{n_h}$$

sustituyendo

$$n_h = nK_h$$

se obtiene:

$$n = \frac{\displaystyle\sum_h \frac{W_h^2}{K_h} S_h^2}{V(\bar{x}_{st}) + \dfrac{1}{N} \displaystyle\sum_h W_h S_h^2} = \frac{\dfrac{1}{4}\left[\dfrac{7/3}{0{,}34} + \dfrac{13/3}{0{,}66}\right]}{0{,}6 + \dfrac{1}{6} \times \dfrac{1}{2}\left(\dfrac{7}{3} + \dfrac{13}{3}\right)} = 2{,}9$$

Es decir, debería tomarse una muestra de 3 unidades para lograr una varianza inferior a 0,6, de las cuales 2 unidades serían seleccionadas en el estrato 2, y 1 unidad en el estrato 1.

3.6. En una población $L = 2$; $N_1 = N_2 = 5$; $n_1 = n_2 = 2$, se obtiene la siguiente muestra:

Estrato 1 (X_{1i})	Estrato 2 (X_{2i})
10	20
2	4

Se pide:

a) Estimar la ganancia o pérdida en precisión, para el estimador de la media, con relación al muestreo aleatorio simple.

b) Estimar la ganancia en precisión si la muestra hubiese sido:

X_{1i}	X_{2i}
10	2
20	4

Solución:

a) Para estimar la ganancia o pérdida en precisión nos basaremos en la siguiente relación debida a J. N. K. Rao que permite estimar, a partir de muestras estratificadas, la varianza que se obtendría en un muestreo aleatorio simple con el mismo tamaño de muestra:

$$\hat{V}(\bar{x}) = \frac{N-n}{N-1} \times \frac{1}{n}\left[\frac{1}{N}\sum_{h}^{L}\frac{N_h}{n_h}\sum_{i}^{n_h} X_{hi}^2 - \hat{\bar{X}}_{st}^2 + \hat{V}(\hat{\bar{X}}_{st})\right] \qquad (1)$$

La anterior expresión exige calcular los estimadores $\hat{\bar{X}}_{st}$ y $\hat{V}(\hat{\bar{X}}_{st})$ en el muestreo estratificado, mediante:

$$\hat{\bar{X}}_{st} = \sum_{h}\frac{N_h}{N}\bar{x}_h = \frac{5}{10}(6 + 12) = 9$$

$$\hat{V}(\hat{\bar{X}}_{st}) = \sum_{h}\left(\frac{N_h}{N}\right)^2(1 - f_h)\frac{\hat{S}_h^2}{n_h} = \frac{1}{4}\left(1 - \frac{2}{5}\right)\frac{1}{2}(32 + 128) = 12$$

ya que:

$$\hat{S}_1^2 = \frac{\sum_i X_{1i}^2 - n_1\bar{x}_1^2}{n_1 - 1} = (100 + 4) - 2 \times 6^2 = 32$$

$$\hat{S}_2^2 = \frac{\sum_i X_{2i}^2 - n_2\bar{x}_2^2}{n_2 - 1} = (400 + 16) - 2 \times 144 = 128$$

Sustituyendo los anteriores valores en (1), obtenemos:

$$\hat{V}(\bar{x}) = \frac{10-4}{10-1} \times \frac{1}{4}\left[\frac{1}{10} \times \frac{5}{2}(100 + 4 + 400 + 16) - 9^2 + 12\right] = \frac{61}{6} = 10{,}17$$

Por tanto, la diferencia relativa de varianzas entre ambos tipos de muestreo es:

$$\frac{\hat{V}(\bar{x}_{st}) - \hat{V}(\bar{x})}{\hat{V}(\bar{x})} = \frac{12 - 10{,}17}{10{,}17} = +0{,}18$$

es decir, hay una pérdida en precisión de un 18 % debido a la estratificación.

b) Procediendo análogamente al caso anterior, con los nuevos valores de la muestra obtenemos:

$$\hat{\bar{X}}_{st} = \frac{5}{10}(15 + 3) = 9$$

$$\hat{V}(\hat{\bar{X}}_{st}) = \frac{1}{4}\left(1 - \frac{2}{5}\right)\frac{1}{2}(50 + 2) = 3{,}9$$

ya que:

$$\hat{S}_1^2 = (100 + 400) - 2 \times 15^2 = 50$$

$$\hat{S}_2^2 = (4 + 16) - 2 \times 3^2 = 2$$

Sustituyendo los anteriores valores en (1), se obtiene:

$$\hat{V}(\bar{x}) = \frac{6}{9} \times \frac{1}{4}\left[\frac{1}{10} \times \frac{5}{2}(100 + 400 + 4 + 16) - 9^2 + 3{,}9\right] = \frac{52{,}9}{6} = 8{,}82$$

Siendo la diferencia relativa de varianzas:

$$\frac{\hat{V}(\bar{x}_{st}) - \hat{V}(\bar{x})}{\hat{V}(\bar{x})} = \frac{3{,}9 - 8{,}82}{8{,}82} = -0{,}56$$

es decir, hay una ganancia en precisión del 56 % al utilizar la estratificación.

Los diferentes resultados obtenidos en los casos *a*) y *b*) se debe a la diferente construcción de los estratos. El caso *b*) cumple el principio básico de la estratificación de lograr la mayor homogeneidad posible dentro de los estratos.

3.7. Determine el tamaño n de la muestra que con afijación óptima produzca la misma precisión que una muestra simple (no estratificada) de tamaño n', para estimar la proporción P de una cierta clase en la población. Suponga en ambos casos muestreo con reposición y aplique el resultado a los siguientes datos con $n' = 1.000$. Datos:

	Estratos		
	I	II	III
W_h	0,2	0,3	0,5
P_h	0,5	0,6	0,4

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos.)

Solución:

En una muestra simple con reposición de tamaño n', la varianza del estimador de la proporción viene dada por:

$$V(\hat{P}) = \frac{P(1-P)}{n'} \tag{1}$$

Y en un muestreo estratificado con reposición, en general:

$$V(\hat{P}_{st}) = \sum_h W_h^2 \frac{P_h(1-P_h)}{n_h}$$

y siendo afijación óptima:

$$n_h = n \frac{W_h\sqrt{P_h(1-P_h)}}{\sum_h W_n\sqrt{P_h(1-P_h)}}$$

de donde:

$$V(\hat{P}_{st}) = \sum_h W_h^2 \frac{P_h(1-P_h)}{n\dfrac{W_h\sqrt{P_h(1-P_h)}}{\sum_h W_h\sqrt{P_h(1-P_h)}}} = \frac{\left[\sum_h W_h\sqrt{P_h(1-P_h)}\right]^2}{n} \tag{2}$$

Solución:

Problema análogo al anterior sin más que variar la naturaleza de la afijación. En este caso, siendo afijación proporcional: $n_h = nW_h$, y, por tanto:

$$V(\hat{P}_{st}) = \sum_h W_h^2 \frac{P_h(1 - P_h)}{n_h} = \frac{\sum_h W_h P_h(1 - P_h)}{n} \qquad (1)$$

Por otra parte, en una muestra simple sin estratificar:

$$V(\hat{P}) = \frac{P(1 - P)}{n'} \qquad (2)$$

Igualando (1) y (2), a fin de obtener la misma precisión, resulta:

$$n = \frac{n' \sum_h W_h P_h(1 - P_h)}{P(1 - P)}$$

donde

$$P = \sum_h W_h P_h$$

De los datos del enunciado:

Estratos	W_h	P_h	$1 - P_h$	$W_n P_h$	$W_h P_h(1 - P_h)$
I	0,5	0,52	0,48	0,26	0,1248
II	0,3	0,40	0,60	0,12	0,0720
III	0,2	0,60	0,40	0,12	0,480
				$\sum = 0,5$	$\sum = 0,2448$

Por tanto:

$$n = \frac{600 \times 0,2448}{0,5 \times 0,5} = 588$$

Igualando las expresiones (1) y (2) para obtener la misma precisión, resulta:

$$n = \frac{n'\left[\sum_h W_h\sqrt{P_h(1-P_h)}\right]^2}{P(1-P)}$$

De los datos del enunciado, y siendo $P = \sum_h W_h P_h$, obtenemos:

Estratos	W_h	P_h	$1-P_h$	W_hP_h	$\sqrt{(P_h(1-P_h)}$	$W_h\sqrt{P_h(1-P_h)}$
I	0,2	0,5	0,5	0,10	0,5	0,1
II	0,3	0,6	0,4	0,18	0,49	0,147
III	0,5	0,4	0,6	0,20	0,49	0,245
				$\sum = 0{,}48$		$\sum = 0{,}492$

Por tanto:

$$n = \frac{1.000 \times 0{,}492^2}{0{,}48 \times 0{,}52} = 970$$

ligeramente inferior al tamaño 1.000 requerido en una muestra simple para obtener la misma precisión. En este caso, la estratificación no aporta sustancial ganancia en precisión.

3.8. En una población con tres estratos, los pesos relativos de los estratos W_h, son:

$$W_h : (0{,}5;\ 0{,}3;\ 0{,}2)$$

y en una encuesta piloto se han encontrado los valores:

$$P_h : (0{,}52;\ 0{,}40;\ 0{,}60)$$

N_h es suficientemente grande con relación a n_h para prescindir de f_h.

Se pide: Determinar el tamaño de una muestra estratificada, con afijación proporcional, que nos dé la misma precisión para estimar P que una muestra de tamaño $n' = 600$ sin estratificar.

3.9. En una población de $N = 24$ unidades dividida en dos estratos del mismo tamaño se obtiene una muestra estratificada que proporciona los siguientes valores:

$$\text{Estrato 1 : } (3; 2; 4) \quad , \quad \text{Estrato 2 : } (6; 4; 5)$$

Se pide:

a) Error de muestreo $\hat{\sigma}_{\bar{x}_{st}}$, siendo $\bar{x}_{st} = \sum_k W_h \bar{x}_h$.

b) Error de muestreo $\hat{\sigma}_{\bar{x}}$, que se hubiese obtenido sin estratificar la población con un esquema sin reposición y probabilidades iguales, utilizando los resultados obtenidos con la muestra estratificada. ¿Cuál es la ganancia en precisión expresada en porcentaje?

c) Suponiendo ahora que, en un muestreo aleatorio simple, los datos muestrales del estrato 1 corresponden a la variable en estudio y los del estrato 2 a una variable auxiliar que se supone relacionada con la anterior, estimar el valor aproximado de la componente sistemática del error debido al muestreo al utilizar un estimador de la razón.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos.)

Solución:

a) La varianza del estimador de la media, en un muestreo estratificado, se estima insesgadamente mediante la expresión:

$$\hat{V}(\bar{x}_{st}) = \sum_h W_h^2 (1 - f_h) \frac{\hat{S}_h^2}{n_h}$$

siendo

$$W_1 = W_2 = \frac{1}{2} \quad ; \quad f_1 = f_2 = \frac{3}{12} \quad ; \quad n_1 = n_2 = 3$$

$$\hat{S}_1^2 = \frac{\sum_i x_{1i}^2 - n_1 \bar{x}_1^2}{n_1 - 1} = \frac{29 - 27}{2} = 1 \quad \text{y} \quad \hat{S}_2^2 = \frac{\sum_i x_{2i}^2 - n_2 \bar{x}_2^2}{n_2 - 1} = \frac{77 - 75}{2} = 1$$

se obtiene:

$$\hat{V}(\bar{x}_{st}) = \frac{1}{4} \times \frac{3}{4} \times \frac{1}{3} (1 + 1) = \frac{1}{8} = 0{,}125$$

y el error de muestreo:

$$\hat{\sigma}_{\bar{x}_{st}} = \sqrt{0{,}125} = 0{,}35$$

b) Para estimar la ganancia en precisión nos basaremos en la expresión debida a J. N. K. Rao que relaciona la varianza en muestras aleatorias simples con resultados obtenidos en muestras estratificadas:

$$\hat{V}(\bar{x}) = \frac{N-n}{N-1} \times \frac{1}{n}\left[\frac{1}{N}\sum_{h}\frac{N_h}{n_h}\sum_{i}^{n_h} X_{hi}^2 - \bar{x}_{st}^2 + V(\bar{x}_{st})\right]$$

con el valor de

$$\bar{x}_{st} = \sum_{h} W_h \bar{x}_h = \frac{1}{2}(3+5) = 4$$

se obtiene:

$$\hat{V}(\bar{x}) = \frac{24-6}{24-1} \times \frac{1}{6}\left[\frac{1}{24} \times \frac{12}{3} \times 106 - 4^2 + 0{,}125\right] = \frac{43}{184} = 0{,}23$$

La estimación de la ganancia en precisión, en términos porcentuales, será:

$$\frac{\hat{V}(\bar{x}) - V(\bar{x}_{st})}{\hat{V}(\bar{x})} \times 100 = \frac{0{,}23 - 0{,}125}{0{,}23} \times 100 = 46\,\%$$

es decir, se obtiene una ganancia en precisión del 46 % debido a la estratificación.

c) Si utilizamos un estimador de razón, en un muestreo aleatorio simple, con los siguientes valores muestrales de la variable X objeto de estudio, y la variable auxiliar Y:

X_i	Y_i	X_iY_i	Y_i^2
3	6	18	36
2	4	8	16
4	5	20	25
		$\sum = 46$	$\sum = 77$

el sesgo del estimador de razón se estima, aproximadamente, mediante la expresión:

$$\hat{B} \simeq (1-f)\hat{R}\left[\frac{1}{\bar{y}^2}\frac{\sum y_i^2 - n\bar{y}^2}{n(n-1)} - \frac{1}{\bar{x}\bar{y}}\frac{\sum x_i y_i - n\bar{x}\bar{y}}{n(n-1)}\right]$$

y siendo en este caso:

$$f = \frac{n}{N} = \frac{3}{24} \quad ; \quad \hat{R} = \frac{\bar{x}}{\bar{y}} = \frac{3}{5}$$

se obtiene:

$$\hat{B} \simeq \left(1 - \frac{1}{8}\right)\frac{3}{5}\left[\frac{1}{25} \times \frac{77 - 75}{6} - \frac{1}{15} \times \frac{46 - 45}{6}\right] = \frac{7}{6.000} = 0{,}00117$$

sesgo prácticamente despreciable.

3.10. En una población dividida en dos estratos del mismo peso se ha obtenido una muestra con reposición y probabilidades iguales de 2 unidades por estrato, $n_h = 2$ $(h = 1; 2)$ que proporcionó las observaciones siguientes: $X_{1i}(5; 3)$, $X_{2i}(12; 16)$.

Se pide: Calcular $\hat{V}(\hat{\bar{X}}_{st})$, directamente y por el método de las pseudo-reiteraciones con semimuestras utilizando las 2^2 semimuestras posibles. (Comprobar que se obtiene el mismo resultado).

Solución:

Los estimadores insesgados de la media y de su varianza, obtenidos directamente, arrojan los siguientes resultados:

$$\hat{\bar{X}}_{st} = \sum_h W_h \bar{x}_h = \frac{1}{2}(4 + 14) = 9$$

$$\hat{V}(\hat{\bar{X}}_{st}) = \sum_h W_h^2 \frac{\hat{S}_h^2}{n_h} \qquad (1)$$

siendo $\hat{S}_h^2$ la cuasivarianza muestral en el estrato h. En el caso de 2 unidades por estrato, la expresión (1) se transforma en:

$$\hat{V}(\hat{\bar{X}}_{st}) = \frac{1}{4}\sum_h W_h^2 d_h^2$$

donde

$$d_h = X_{h1} - X_{h2}$$

Por tanto,

$$\hat{V}(\hat{\bar{X}}_{st}) = \frac{1}{4}\left(\frac{1}{2}\right)^2 [(5-3)^2 + (12-16)^2] = \frac{5}{4} = 1{,}25$$

Seleccionando aleatoriamente una unidad en cada estrato tendríamos una semimuestra o pseudoreiteración i, a partir de la cual puede obtenerse una estimación $\hat{\bar{X}}_i$ del parámetro poblacional a investigar. Designando por $\hat{\bar{X}}_{st}$ el estimador obtenido con la muestra total, un estimador de $V(\hat{\bar{X}}_{st})$, por el método de las pseudoreiteraciones con semimuestras, sería:

$$\hat{V}(\hat{\bar{X}}_{st}) = \frac{\sum_{i}^{r} (\hat{\bar{X}}_i - \hat{\bar{X}}_{st})^2}{r} \qquad (2)$$

donde r es un número arbitrario de pseudoreiteraciones o semimuestras aleatorias utilizado para la estimación. Cuanto mayor sea r tanto más preciso será el método. Si r es el número total de semimuestras posibles ($r = 2^L$, siendo L el número de estratos), la expresión (2) coincide con la estimación directa de la varianza dada en (1).

En nuestro caso, las 2^2 semimuestras posibles proporcionan los siguientes valores:

Pseudoreiteraciones ó semimuestras	$\hat{\bar{X}}_i = \sum W_h \bar{x}_h$	$\hat{\bar{X}}_i - \hat{\bar{X}}_{st}$	$(\hat{\bar{X}}_i - \hat{\bar{X}}_{st})^2$
X_{11} ; X_{21}	$\frac{1}{2}(5+12) = 8{,}5$	$-0{,}5$	0,25
X_{11} ; X_{22}	$\frac{1}{2}(5+16) = 10{,}5$	$-1{,}5$	2,25
X_{12} ; X_{21}	$\frac{1}{2}(3+12) = 7{,}5$	1,5	2,25
X_{12} ; X_{22}	$\frac{1}{2}(3+16) = 9{,}5$	0,5	0,25
			$\sum = 5$

Por tanto,

$$\hat{V}(\hat{\bar{X}}_{st}) = \frac{1}{4}\sum_{i}^{4} (\hat{\bar{X}}_i - \hat{\bar{X}}_{st})^2 = \frac{5}{4} = 1{,}25$$

que coincide con el resultado obtenido mediante la expresión (1).

3.11. Se desea conocer el salario medio mensual, por establecimiento, para una población de $N = 2.000$ establecimientos industriales de un cierto tipo. Para ello se obtiene una muestra de $n = 200$ establecimientos, sin reemplazamiento y probabilidades iguales, a los que se solicita por correo este dato.

Después de varios recordatorios se consigue una respuesta del 75 % con los siguientes valores para la media y cuasivarianza muestrales:

$$\bar{x}_1 = 20.000 \text{ pts.} \qquad \hat{S}_1^2 = 1.400.000$$

donde el subíndice 1 expresa el estrato de establecimientos respondientes.

Seguidamente se envían agentes entrevistadores a una submuestra, seleccionada mediante muestreo aleatorio simple, del 40 % de establecimientos seleccionados de una lista formada por los establecimientos que no contestaron.

Los resultados obtenidos en esta submuestra fueron:

$$\bar{x}_2 = 16.000 \textit{ pts.} \qquad \hat{S}_2^2 = 1.600.000$$

Se pide:

a) Salario medio estimado por establecimiento.

b) Componentes de la varianza estimada para la media, siendo aproximadamente:

$$\hat{S}^2 \doteq \sum_h \hat{W}_h (\bar{x}_h - \hat{\bar{X}})^2 + \sum_h \hat{W}_h \hat{S}_h^2$$

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos.)

Solución:

a) Si consideramos la población dividida en dos estratos: I) los que responden a la encuesta por correo, II) los no respondentes a la misma, nos encontramos ante un muestreo estratificado en el que la información del segundo estrato se consigue mediante un nuevo esfuerzo o entrevista personal a una submuestra de no respondentes. Esta técnica para el tratamiento de la no respuesta, debida a Hansen y Hurwitz, es una aplicación del muestreo doble para estratificación.

El estimador insesgado de la media viene dado por:

$$\hat{\bar{X}} = \sum_h \hat{W}_h \bar{x}_h \qquad (1)$$

donde $\hat{W}_h$ son los estimadores de los pesos relativos de los estratos. En nuestro caso:

$$\hat{W}_1 = 0{,}75 \quad , \quad \hat{W}_2 = 0{,}25$$

Por tanto:

$$\hat{\bar{X}} = 0{,}75 \times 20.000 + 0{,}25 \times 16.000 = 19.000 \ pts.$$

b) La varianza del estimador dado en (1) tiene dos componentes:

$$V(\hat{\bar{X}}) = \frac{N-n}{N}\frac{S^2}{n} + \frac{N^2}{N}\left(\frac{1}{f_{21}} - 1\right)\frac{S_2^2}{n} \qquad (2)$$

siendo f_{21} la fracción de muestreo en 2.ª fase entre los no respondentes en 1.ª fase.

La 1.ª componente de la varianza es la que se obtendría en un muestreo aleatorio simple sin falta de respuesta. Por tanto, la 2.ª componente es el incremento en la varianza debido a la falta de respuesta.

A partir de los datos del enunciado podemos obtener el estimador:

$$\hat{V}(\hat{\bar{X}}) = \frac{N-n}{N}\frac{\hat{S}^2}{n} + \hat{W}_2\left(\frac{1}{f_{21}} - 1\right)\frac{\hat{S}_2^2}{n}$$

con los siguientes resultados:

Estrato	$\hat{W}_h$	$\bar{x}_h$	$(\bar{x}_h - \hat{\bar{X}})$	$\hat{W}_h(\bar{x}_h - \hat{\bar{X}})^2$	$\hat{W}_h\hat{S}_h^2$	
I	0,75	20.000	1.000	750.000	1,.050.000	
II	0,25	16.000	−3.000	2.250.000	400.000	
				$\sum = 3.000.000$	$\sum = 1.450.000$	$\hat{S}^2 \simeq 4.450.000$

Por tanto:

1.ª componente: $\dfrac{N-n}{N}\dfrac{\hat{S}^2}{n} = \dfrac{2.000-200}{2.000} \times \dfrac{4.450.000}{200} = 20.025$

2.ª componente: $\hat{W}_2\left(\dfrac{1}{f_{21}} - 1\right)\dfrac{\hat{S}_2^2}{n} = 0{,}25\left(\dfrac{1}{0{,}4} - 1\right)\dfrac{1.600.000}{200} = 3.000$

Varianza total estimada: $\hat{V}(\hat{\bar{X}}) = 23.025$.

CAPITULO IV

Estimadores mejorados (estimadores de razón y de regresión)

4.1. Los valores de dos variables X e Y en las $N = 4$ unidades de una población, son:

u_i	X_i	Y_i
u_1	1	1
u_2	2	3
u_3	3	4
u_4	4	6

Se obtienen todas las muestras posibles de tamaño $n = 2$, sin reposición y con probabilidades iguales. Se pide:

1.º) El valor exacto del sesgo y de la varianza del estimador de la razón poblacional.

2.º) Comparar estos resultados con los obtenidos mediante las fórmulas aproximadas del sesgo y de la varianza.

Solución:

1.º) La razón poblacional

$$R = \frac{X}{Y} = \frac{\sum_i^N X_i}{\sum_i^N Y_i} = \frac{10}{14} = \frac{5}{7}$$

se estima, en un muestreo con probabilidades iguales, por la correspondiente razón de totales muestrales:

$$\hat{R} = \frac{\sum_i^n X_i}{\sum_i^n Y_i} = \frac{x}{y}$$

Los valores exactos del sesgo y varianza de este estimador pueden obtenerse a partir de la siguiente distribución en el muestreo del estimador

Muestras posibles	Probabilidades	$x = \sum^{n} x_i$	$y = \sum^{n} y_i$	$\hat{R} = \frac{x}{y}$
u_1u_2	$\frac{1}{6}$	3	4	$\frac{3}{4}$
u_1u_3	$\frac{1}{6}$	4	5	$\frac{4}{5}$
u_1u_4	$\frac{1}{6}$	5	7	$\frac{5}{7}$
u_2u_3	$\frac{1}{6}$	5	7	$\frac{5}{7}$
u_2u_4	$\frac{1}{6}$	6	9	$\frac{6}{9}$
u_3u_4	$\frac{1}{6}$	7	10	$\frac{7}{10}$

siendo

$$E(\hat{R}) = \frac{1}{6}\left[\frac{3}{4} + \frac{4}{5} + \frac{5}{7} + \frac{5}{7} + \frac{6}{9} + \frac{7}{10}\right] = \frac{365}{504}$$

obtenemos como valor exacto del sesgo:

$$\text{sesgo de } \hat{R} = E(\hat{R}) - R = \frac{365}{504} - \frac{5}{7} = \frac{5}{504} = 0{,}0099$$

Por otra parte, siendo el estimador sesgado, utilizaremos el error medio cuadrático como medida de la variabilidad del estimador (mejor que la varianza). Obtenemos:

$$E(\hat{R}-R)^2 = \frac{1}{6}\left[\left(\frac{3}{4}-\frac{5}{7}\right)^2 + \left(\frac{4}{5}-\frac{5}{7}\right)^2 + \left(\frac{6}{9}-\frac{5}{7}\right)^2 + \left(\frac{7}{10}-\frac{5}{7}\right)^2\right] = 0{,}00185$$

Puede observarse que, en este caso, el sesgo es importante en relación con la varianza.

2.º) Una expresión aproximada del sesgo viene dada por:

$$E(\hat{R} - R) \simeq \frac{1 - f}{n\bar{Y}^2}(RS_y^2 - S_{xy})$$

siendo

$$f \text{ (fracción de muestreo)} = \frac{1}{2} \quad ; \quad n = 2 \quad ; \quad \bar{Y} = \frac{7}{2}$$

$$S_y^2 = \frac{\sum Y_i^2 - N\bar{Y}^2}{N - 1} = \frac{62 - 49}{3} = \frac{13}{3} \quad ; \quad S_{xy} = \frac{\sum X_i Y_i - N\bar{X}\bar{Y}}{N - 1} = \frac{43 - 35}{3} = \frac{8}{3}$$

resulta

$$E(\hat{R} - R) \simeq \frac{\frac{1}{2}}{2\left(\frac{7}{2}\right)^2}\left(\frac{5}{7} \times \frac{13}{3} - \frac{8}{3}\right) = \frac{3}{343} = 0{,}009$$

que subestima ligeramente el verdadero valor del sesgo.

Respecto a la varianza, una expresión aproximada viene dada por:

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

siendo

$$S_x^2 = \frac{\sum X_i^2 - N\bar{X}^2}{N - 1} = \frac{30 - 25}{3} = \frac{5}{3}$$

y los demás valores ya conocidos, se obtiene:

$$V(\hat{R}) = \frac{\frac{1}{2}}{2\left(\frac{7}{2}\right)^2}\left[\frac{5}{3} + \frac{5^2}{7^2} \times \frac{13}{3} - 2 \times \frac{5}{7} \times \frac{8}{3}\right] = 0{,}00139$$

que también subestima el verdadero valor.

4.2. De una población de $N = 40$ hogares se obtiene una muestra aleatoria simple, de tamaño $n = 4$ hogares, que proporciona los siguientes valores anuales expresados en miles de pesetas:

Gastos en alimentación (x_i)	Gasto total (y_i)
125	250
135	300
70	200
158	350

Se pide: estimar el porcentaje de gasto en alimentación y su error de muestreo.

Solución:

El porcentaje estimado se obtiene mediante el cociente entre los totales muestrales de las variables «gastos en alimentación» y «gasto total». Se trata, pues, de un estimador de razón del tipo:

$$\hat{R} = \frac{\sum_{i}^{n} x_i}{\sum_{i}^{n} y_i} = \frac{488}{1.100} = 0{,}4436 \simeq 44{,}4\,\%$$

La varianza se estima mediante:

$$\hat{V}(\hat{R}) = \frac{1-f}{\bar{y}^2(n-1)n}\left[\sum_{i}^{n} x_i^2 + \hat{R}^2 \sum_{i}^{n} y_i^2 - 2\hat{R} \sum_{i}^{n} x_i y_i\right] =$$

$$= \frac{\frac{9}{10}}{\left(\frac{1.100}{4}\right)^2 12}\left[63.714 + \left(\frac{488}{1.100}\right)^2 315.000 - 2 \times \frac{488}{1.100} \times 141.050\right] =$$

$$= 0{,}000556$$

siendo el error de muestreo:

$$\hat{\sigma}_{\hat{R}} = \sqrt{0{,}000556} = 0{,}0236 \simeq 2{,}4\,\%$$

4.3. Una muestra de tamaño $n = 4$ proporciona los valores siguientes:

x_i	y_i
1	1
2	3
3	4
4	5

Prescindiendo del factor de corrección para poblaciones finitas, en un muestreo aleatorio simple, se pide estimar el sesgo del estimador de razón en valor absoluto, y su relación con el error de muestreo.

Solución:

La expresión del estimador del sesgo, prescindiendo del factor de corrección de poblaciones finitas, es:

$$\hat{B} \simeq \frac{1}{n\bar{y}^2}(\hat{R}\hat{S}_y^2 - \hat{S}_{xy})$$

siendo

$$\hat{R} = \frac{\sum_n^n x_i}{\sum_n^n y_i} = \frac{10}{13} \quad , \quad \bar{y} = \frac{\sum_n^n y_i}{n} = \frac{13}{4}$$

$$\hat{S}_y^2 = \frac{\sum_n^n y_i^2 - n\bar{y}^2}{n-1} = \frac{51 - 42{,}25}{3} = \frac{8{,}75}{3}$$

$$\hat{S}_{xy} = \frac{\sum_n^n x_i y_i - n\bar{x}\bar{y}}{n-1} = \frac{39 - 4\left(\frac{10}{4}\right)\left(\frac{13}{4}\right)}{3} = \frac{6{,}5}{3}$$

se obtiene:

$$\hat{B} = \frac{1}{4\left(\frac{13}{4}\right)^2}\left(\frac{10}{13} \times \frac{8{,}75}{3} - \frac{6{,}5}{3}\right) = 0{,}0018$$

Por otra parte:

$$\hat{V}(\hat{R}) \simeq \frac{1}{n(n-1)\bar{y}^2}\left[\sum_i^n x_i^2 + \hat{R}^2 \sum_i^n y_i^2 - 2\hat{R}\sum_i^n x_i y_i\right] =$$

$$= \frac{1}{12\left(\frac{13}{4}\right)^2}\left[30 + \left(\frac{10}{13}\right)^2 51 - 2\left(\frac{10}{13}\right)39\right] = 0{,}0014$$

obteniéndose como error de muestreo:

$$\hat{\sigma}_{\hat{R}} = \sqrt{0{,}0014} = 0{,}037$$

Por tanto,

$$\frac{\hat{B}}{\hat{\sigma}_{\hat{R}}} = \frac{0{,}0018}{0{,}037} = 0{,}048$$

es decir, el sesgo apenas representa el 5 % en relación con el error de muestreo, siendo prácticamente despreciable.

4.4. Una población se ha estratificado en dos estratos con $N_1 = N_2 = 3$ unidades. Los pares de valores, para cada unidad, de la variable objeto de estudio X y una variable correlacionada Y, son los siguientes:

Estrato 1		Estrato 2	
X_{1i}	Y_{1i}	X_{2i}	Y_{2i}
2	1	5	4
4	2	7	5
5	3	12	6

Se pide:

1.° Obtener para cada estrato los coeficientes de regresión β_h.

2.° Calcular las varianzas mínimas de los estimadores de regresión simple y combinado de la media poblacional, para $n_h = 2$ y muestreo aleatorio simple en cada estrato.

3.° Comprobar que con las elecciones óptimas el estimador simple o separado tiene una varianza menor que el estimador combinado.

Solución:

1.º) Los coeficientes de regresión en cada estrato vienen dados por la expresión:

$$\beta_h = \frac{\sigma_{xyh}}{\sigma^2_{yh}} = \frac{\sum_i^{N_h} X_{ih}Y_{ih} - N_h\bar{X}_h\bar{Y}_h}{\sum_i^{N_h} Y^2_{ih} - N_h\bar{Y}^2_h}$$

Con los valores del enunciado, obtenemos:

Estrato 1			Estrato 2		
X^2_{1i}	Y^2_{1i}	$X_{1i}Y_{1i}$	X^2_{2i}	Y^2_{2i}	$X_{2i}Y_{2i}$
4	1	2	25	16	20
16	4	8	49	25	35
25	9	15	144	36	72
45	14	25	218	77	127

Por tanto:

$$\beta_1 = \frac{25 - 3 \times \frac{11}{3} \times \frac{6}{3}}{14 - 3\left(\frac{6}{3}\right)^2} = \frac{3}{2} \quad ; \quad \beta_2 = \frac{127 - 3 \times \frac{24}{3} \times \frac{15}{3}}{77 - 3\left(\frac{15}{3}\right)^2} = \frac{7}{2}$$

2.º) El estimador de regresión separado viene dado por la expresión:

$$\bar{x}_{rgs} = \sum_h W_h[\bar{x}_h + b_h(\bar{Y}_h - \bar{y}_h)]$$

donde $W_h = \frac{N_h}{N}$, y b_h es una constante predeterminada.

Los valores de b_h que hacen mínima la varianza del estimador son, precisamente, los coeficientes de regresión β_h, obteniéndose la siguiente expresión para la varianza mínima:

$$V_{\min}(\bar{x}_{rgs}) = \sum_h \frac{W^2_h(1 - f_h)}{n_h}(S^2_{xh} + \beta^2_h S^2_{yh} - 2\beta_h S_{xyh}) \qquad (1)$$

A partir de los valores del enunciado, calculamos:

$$S_{1x}^2 = \frac{\sum_i^N X_{1i}^2 - N_1 \bar{X}_1^2}{N_1 - 1} = \frac{45 - 3\left(\frac{11}{3}\right)^2}{2} = \frac{7}{3}$$

$$S_{1y}^2 = \frac{\sum_i^{N_1} Y_{1i}^2 - N_1 \bar{Y}_1^2}{N_1 - 1} = \frac{14 - 3(2)^2}{2} = 1$$

$$S_{1xy} = \frac{\sum_i^{N_1} X_{1i} Y_{1i} - N_1 \bar{X}_1 \bar{Y}_1}{N_1 - 1} = \frac{25 - 3 \times \frac{11}{3} \times 2}{2} = \frac{3}{2}$$

Análogamente, para el estrato 2:

$$S_{2x}^2 = \frac{\sum_i^{N_2} X_{2i}^2 - N_2 \bar{X}_2^2}{N_2 - 1} = \frac{218 - 3 \times 8^2}{2} = 13$$

$$S_{2y}^2 = \frac{\sum_i^{N_2} Y_{2i}^2 - N_2 \bar{Y}_2^2}{N_2 - 1} = \frac{77 - 3 \times 5^2}{2} = 1$$

$$S_{2xy} = \frac{\sum_i^{N_2} X_{2i} Y_{2i} - N_2 \bar{X}_2 \bar{Y}_2}{N_2 - 1} = \frac{127 - 3 \times 8 \times 5}{2} = \frac{7}{2}$$

$$W_1 = W_2 = \frac{1}{2} \quad ; \quad f_1 = f_2 = \frac{2}{3} \quad ; \quad n_1 = n_2 = 2$$

Por tanto:

$$V_{\min}(\bar{x}_{rgs}) = \frac{1}{24}\left[\left(\frac{7}{3} + 13\right) + \left(\frac{3}{2}\right)^2 \times 1 + \left(\frac{7}{2}\right)^2 \times 1 - 2 \times \frac{3}{2} \times \frac{3}{2} - 2 \times \frac{7}{2} \times \frac{7}{2}\right] =$$

$$= \frac{5}{144} = 0{,}0347$$

Por otra parte, el estimador de regresión combinado viene dado por la expresión:

$$\bar{x}_{rgc} = \sum_h W_h \bar{x}_h + b_c(\bar{Y} - \sum_h W_h \bar{y}_h)$$

el valor de b_c que minimiza esta varianza es:

$$\bar{\beta}_c = \frac{\sum_h \omega_h \beta_h}{\sum_h \omega_h}$$

es decir, una media ponderada de los coeficientes de regresión de los estratos, con factores de ponderación:

$$\omega_h = \frac{W_h^2(1 - f_h)}{n_h} S_{yh}^2$$

Siendo

$$\omega_1 = \frac{1}{24} S_{1y}^2 = \frac{1}{24} \quad ; \quad \omega_2 = \frac{1}{24} S_{2y}^2 = \frac{1}{24}$$

resulta

$$\bar{\beta}_c = \frac{\beta_1 + \beta_2}{2} = \frac{5}{2}$$

Finalmente, la varianza mínima del estimador de regresión combinado, se obtiene mediante la expresión:

$$V_{\min}(\bar{x}_{rgc}) = \sum_h \frac{W_h^2(1 - f_h)}{n_h} (S_{x_h}^2 + \bar{\beta}_c^2 S_{yh}^2 - 2\bar{\beta}_c S_{xyh}) \qquad (2)$$

resultando con los datos del enunciado:

$$V_{\min}(\bar{x}_{rgc}) = \frac{1}{24}\left[\left(\frac{7}{3} + 13\right) + \left(\frac{5}{2}\right)^2 (1 + 1) - 2 \times \frac{5}{2}\left(\frac{3}{2} + \frac{7}{2}\right)\right] = \frac{17}{144} = 0{,}118$$

3.°) Comparando las expresiones (1) y (2), se obtiene:

$$V_{\min}(\bar{x}_{rgc}) - V_{\min}(\bar{x}_{rgs}) = \sum_h \omega_h (\bar{\beta}_c - \beta_h)^2 \qquad (3)$$

expresión no negativa. Es decir, con las elecciones óptimas de b_n y b_c, el estimador separado tiene una varianza menor que el combinado, a menos que β_h sea idéntico para todos los estratos.

De (3) se obtiene:

$$V_{\min}(\bar{x}_{rgc}) - V_{\min}(\bar{x}_{rgs}) = \frac{1}{24}\left[\left(\frac{5}{2} - \frac{3}{2}\right)^2 + \left(\frac{5}{2} - \frac{7}{2}\right)^2\right] = \frac{1}{12}$$

que es, precisamente, la diferencia entre los valores

$$\frac{17}{144} \quad \text{y} \quad \frac{5}{144}$$

de las varianzas mínimas.

4.5. Utilizando una variable Y inversamente relacionada con la variable X objeto de estudio, se define el estimador de producto para el total X mediante:

$$\hat{X}_p = \hat{X} \cdot \frac{\hat{Y}}{Y}$$

donde $\hat{X}$ e $\hat{Y}$ son los estimadores simples de expansión de las variables X e Y, respectivamente.

Sabiendo que la varianza aproximada del estimador de producto viene dada por la expresión:

$$V(\hat{X}_p) \simeq V(\hat{X}) + 2R\,\text{Cov}\,(\hat{X}, \hat{Y}) + R^2 V(\hat{Y})$$

donde

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

Se pide: Determinar la condición que ha de cumplir el coeficiente de correlación $\rho(\hat{X}, \hat{Y})$ para que el estimador de producto $\hat{X}_p$ sea más preciso que el estimador de expansión simple $\hat{X}$.

Solución:

El estimador de producto será más preciso que el estimador simple, si:

$$V(\hat{X}_p) - V(\hat{X}) < 0 \qquad (1)$$

Para que se cumpla (1) ha de ser:

$$R[2\,\mathrm{Cov}\,(\hat{X},\hat{Y}) + RV(\hat{Y})] = \frac{X}{Y}\left[2\rho\sqrt{V(\hat{X})}\,\sqrt{V(\hat{Y})} + \frac{X}{Y}V(\hat{Y})\right] < 0$$

lo que implica:

$$2\rho\sqrt{V(\hat{X})}\,\sqrt{V(\hat{Y})} < -\frac{X}{Y}V(\hat{Y})$$

o bien

$$2\rho < -\frac{X}{Y}\frac{\sqrt{V(\hat{Y})}}{\sqrt{V(\hat{X})}}$$

y en función de los coeficientes de variación de $\hat{X}$ e $\hat{Y}$:

$$2\rho < -\frac{CV(\hat{Y})}{CV(\hat{X})}$$

resultando la condición:

$$\rho < -\frac{1}{2}\frac{CV(\hat{Y})}{CV(\hat{X})}$$

4.6. En una determinada localidad de 500 viviendas se desea hacer un estudio sobre el hábito de fumar entre las personas mayores de 16 años. Para ello se estratifica la población en dos estratos, en el estrato I (estrato de viviendas de clase alta) se encuentran clasificadas 200 viviendas, mientras que en el estrato II (estrato de viviendas de clase baja) existen 300 viviendas. De cada uno de los estratos se selecciona una muestra aleatoria de 5 viviendas que arroja los siguientes resultados:

Estrato I

Viviendas en la muestra	1	2	3	4	5
Número de personas mayores de 16 años	4	3	2	1	2
Número de fumadores mayores de 16 años	1	1	0	1	1

Estrato II

Viviendas en la muestra	1	2	3	4	5
Número de personas mayores de 16 años	5	6	4	4	3
Número de fumadores mayores de 16 años	3	3	1	2	2

Se pide:

a) Estimar la proporción total de fumadores, entre las personas mayores de 16 años, en la localidad.

b) Calcular el error de muestreo de la estimación anterior.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1988.)

Solución:

Nos encontramos en un muestreo estratificado cuya unidad primaria de muestreo es la vivienda o conglomerado de personas. Puesto que no se realiza submuestreo de personas, se aplica en cada estrato un muestreo aleatorio simple de conglomerados sin submuestreo.

Para cada vivienda de la muestra se obtiene la pareja de datos:

x_{hi} = número de fumadores >16 años, en la vivienda 1.ª del estrato *h.*

y_{hi} = número de personas >16 años, en la vivienda 1.ª del estrato *h.*

Un estimador consistente de R es:

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

donde $\hat{X}$ e $\hat{Y}$ son estimadores insesgados de X e Y, respectivamente.

Puesto que los estimadores $\hat{X}$ e $\hat{Y}$ los obtenemos a través de un muestreo estratificado, $\hat{R}$ será un *estimador de razón combinado* en un muestreo estratificado. Por tanto:

$$\hat{X} = \sum_h N_h \bar{x}_h \qquad \hat{Y} = \sum_h N_h \bar{y}_h \qquad \hat{R} = \frac{\hat{X}}{\hat{Y}} = \frac{\sum_h N_h \bar{x}_h}{\sum_h N_h \bar{y}_h}$$

y la varianza se estima mediante la expresión:

$$\hat{V}(\hat{R}) \simeq \frac{1}{\hat{Y}^2}\left[\hat{\sigma}_{\hat{X}}^2 + \hat{R}^2\hat{\sigma}_{\hat{Y}}^2 - 2\hat{R}\hat{\sigma}_{\hat{X}\hat{Y}}\right]$$

siendo:

$$\hat{\sigma}_{\hat{X}}^2 = \sum_h N_h(N_h - n_h)\frac{\sum_i^{n_h}(x_{hi} - \bar{x}_h)^2}{n_h - 1} \quad ; \quad \bar{x}_h = \frac{\sum_i x_{hi}}{n_h}$$

$$\hat{\sigma}_{\hat{Y}}^2 = \sum_h N_h(N_h - n_h)\frac{\sum_i^{n_h}(y_{hi} - \bar{y}_h)^2}{n_h - 1} \quad ; \quad \bar{y}_h = \frac{\sum_i y_{hi}}{n_h}$$

$$\hat{\sigma}_{\hat{X}\hat{Y}} = \sum_h N_h(N_h - n_h)\frac{\sum_i^{n_h}(x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{n_h - 1}$$

donde N_h es el número total de conglomerados (viviendas) en el estrato h, n_h el tamaño de la muestra, y $\bar{x}_h$ e $\bar{y}_h$ las medias muestrales de fumadores y personas, respectivamente, en el estrato h.

Siendo

$$N_1 = 200 \quad ; \quad N_2 = 300 \quad ; \quad n_1 = n_2 = 5$$

$$\bar{x}_1 = \frac{4}{5} \quad ; \quad \bar{x}_2 = \frac{11}{5} \quad ; \quad \bar{y}_1 = \frac{12}{5} \quad ; \quad \bar{y}_2 = \frac{22}{5}$$

se obtiene:

$$\hat{R} = \frac{200 \times \frac{4}{5} + 300 \times \frac{11}{5}}{200 \times \frac{12}{5} + 300 \times \frac{22}{5}} = \frac{820}{1.800} = 0{,}455 \simeq 46\,\%$$

porcentaje de fumadores en la población.

Para estimar la varianza, a partir de los datos del enunciado, obtenemos el siguiente cuadro:

Estrato	$N_h(N_h-n_h)$	$\sum_i (x_{hi}-\bar{x}_h)^2 = \sum x_{hi}^2 - n_h\bar{x}_h^2$	$\sum_i y_{hi}^2 - n_h\bar{y}_h^2$	$\sum_i x_{hi}y_{hi} - n_h\bar{x}_h\bar{y}_h$
I	$200 \times 195 = 39.000$	$4 - \frac{16}{5} = \frac{4}{5}$	$34 - \frac{144}{5} = \frac{26}{5}$	$10 - \frac{48}{5} = \frac{2}{5}$
II	$300 \times 295 = 88.500$	$27 - \frac{121}{5} = \frac{14}{5}$	$102 - \frac{484}{5} = \frac{26}{5}$	$51 - \frac{242}{5} = \frac{13}{5}$

Por tanto:

$$\hat{\sigma}_{\hat{X}}^2 = \frac{1}{4}\left[39.000 \times \frac{4}{5} + 88.500 \times \frac{14}{5}\right] = 69.750$$

$$\hat{\sigma}_{\hat{Y}}^2 = \frac{1}{4}\left[39.000 \times \frac{26}{5} + 88.500 \times \frac{26}{5}\right] = 165.750$$

$$\hat{\sigma}_{\hat{X}\hat{Y}} = \frac{1}{4}\left[39.000 \times \frac{2}{5} + 88.500 \times \frac{13}{5}\right] = 61.425$$

$$\hat{V}(\hat{R}) = \frac{1}{1.800^2}\left[69.750 + \left(\frac{820}{1.800}\right)^2 165.750 - 2\left(\frac{820}{1.800}\right)61.425\right] = 0{,}015$$

Siendo el error de muestreo de $\hat{R}$:

$$\hat{\sigma}_{\hat{R}} = \sqrt{0{,}015} = 0{,}12$$

CAPITULO V
Muestreo con probabilidades desiguales

5.1. Para estudios relacionados con el maíz en una determinada comarca, que tiene una extensión total de 25.000 Ha con 100 explotaciones agrarias, se selecciona una muestra con reemplazamiento de 10 explotaciones agrarias con probabilidad proporcional a la superficie total de la explotación. La proporción de superficie dedicada al cultivo de maíz en cada unidad de la muestra es:

0,05; 0,25; 0,10; 0,30; 0,15; 0,25; 0,35; 0,25; 0,10; 0,20

Se pide:

a) Un estimador insesgado de la superficie total de la comarca dedicada al cultivo del maíz.

b) El coeficiente de variación del estimador anterior.

c) El intervalo de confianza del 95 % para la superficie total de la comarca dedicada al cultivo del maíz.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1988)

Solución:

Siendo:

$M_i \rightarrow$ superficie explotación i-ésima.

$X_i \rightarrow$ superficie dedicada al cultivo del maíz en la explotación i-ésima.

a) Un estimador insesgado del total es:

$$\hat{X} = \sum_i^n \frac{X_i}{nP_i} = \frac{1}{n} \sum_i \frac{X_i}{M_i/M}$$

donde n es el tamaño de la muestra,

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

es la probabilidad de selección de la unidad i-ésima, y

$$M = \sum_i M_i = 25.000 \text{ Ha}$$

Puesto que en el enunciado se dan los valores X_i/M_i, obtenemos:

$$\hat{X} = \frac{M}{n} \sum_i^n \frac{X_i}{M_i} = \frac{25.000}{10} \times 2 = 5.000 \text{ Ha}$$

b) El coeficiente de variación es:

$$CV(\hat{X}) = \frac{\sigma_{\hat{X}}}{\hat{X}}$$

donde

$$\hat{\sigma}^2_{\hat{X}} = \frac{\sum_i^n \left(\frac{X_i}{P_i} - \hat{X}\right)^2}{n(n-1)} = \frac{\sum_i \left(M\frac{X_i}{M_i} - \hat{X}\right)^2}{n(n-1)} = \frac{M^2 \sum_i \left(\frac{X_i}{M_i}\right)^2 - n\hat{X}^2}{n(n-1)} =$$

$$= \frac{25.000^2(0,485) - 10(5.000^2)}{10 \times 9} = 590.278 \quad ; \quad \hat{\sigma}_{\hat{X}} = \sqrt{590.278} = 768$$

Por tanto:

$$C\hat{V}(\hat{X}) = \frac{768}{5.000} = 0,15$$

obteniéndose un error de muestreo relativo del 15 %.

c) Suponiendo la normalidad del estimador, un intervalo de confianza del 95 %, será:

$$\hat{X} \pm 2\hat{\sigma}_{\hat{X}}$$

es decir (3.464; 6.536).

5.2. En un muestreo con reposición y probabilidades proporcionales a los tamaños (P_i), se pide:

1.º) Demostrar que la expresión

$$\hat{V}(\hat{X}) = \frac{1}{n(n-1)} \left[\sum_i^n \left(\frac{X_i}{P_i}\right)^2 - n\hat{X}^2 \right] \qquad (1)$$

es un estimador insesgado de

$$V(\hat{X}_{HH}) = \frac{1}{n}\left[\sum_i^N \frac{X_i^2}{P_i} - X^2\right]$$

siendo

$$X = \sum_i^N X_i$$

2.°) Demostrar que la expresión

$$V(\hat{X}_{HH}) = \frac{1}{n}\sum_i^N \sum_{j>i}^N \left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2 P_i P_j \qquad (2)$$

donde el sumatorio se refiere a todos los pares de unidades distintas, es una forma alternativa de

$$V(\hat{X}_{HH}) = \frac{1}{n}\left[\sum_i^N \frac{X_i^2}{P_i} - X^2\right]$$

3.°) Calcular la expresión (2) con los datos del ejercicio número 1.6 y comprobar el resultado.

Solución:

1.°) Si designamos por e_i la variable aleatoria «número de veces que la unidad u_i es seleccionada en una muestra de tamaño n», e_i sigue una distribución binomial de parámetros (n, P_i), por tanto:

$$Ee_i = nP_i$$

Tomando esperanzas en (1):

$$E\hat{V}(\hat{X}) = \frac{1}{n(n-1)}\left[E\sum_i^n \left(\frac{X_i}{P_i}\right)^2 - nE\hat{X}^2\right] \qquad (3)$$

y

$$E\sum_i^n \left(\frac{X_i}{P_i}\right)^2 = E\sum_i^N \left(\frac{X_i}{P_i}\right)^2 e_i = \sum_i^N \left(\frac{X_i}{P_i}\right)^2 nP_i = n\sum_i^N \frac{X_i^2}{P_i}$$

Por otra parte, si $\hat{X}$ y $\hat{V}(\hat{X})$ son estimadores insesgados de X y $V(\hat{X}_{HH})$, respectivamente, resulta:

$$V(\hat{X}_{HH}) = E\hat{X}^2 - X^2$$

de donde

$$E\hat{X}^2 = V(\hat{X}_{HH}) + X^2 = E\hat{V}(\hat{X}) + X^2$$

sustituyendo en (3) los valores obtenidos de las esperanzas:

$$E\hat{V}(\hat{X}) = \frac{1}{n-1}\left[\sum_i^N \frac{X_i^2}{P_i} - E\hat{V}(\hat{X}) - X^2\right]$$

de donde despejando $E\hat{V}(\hat{X})$, obtenemos:

$$E\hat{V}(\hat{X}) = \frac{1}{n}\left[\sum_i^N \frac{X_i^2}{P_i} - X^2\right]$$

2.º) En efecto:

$$\sum_i^N \frac{X_i^2}{P_i} - X^2 = \sum_i^N \frac{X_i^2}{P_i} - \left(\sum_i^N X_i\right)^2 = \sum_i^N \frac{X_i^2}{P_i} - \sum_i^N X_i^2 - 2\sum_i^N \sum_{j>i}^N X_iX_j =$$

$$= \sum_i^N \frac{X_i^2}{P_i}(1-P_i) - 2\sum_i \sum_{j>i} X_iX_j = \sum_i^N \frac{X_i^2}{P_i}\left(\sum_{j\neq i} P_j\right) - 2\sum_i \sum_{j>i} X_iX_j =$$

$$= \sum_i^N \sum_{j>i} \left(\frac{X_i^2}{P_i}P_j + \frac{X_j^2}{P_j}\right) - 2\sum_i \sum_{j>i} X_iX_j = \sum_i^N \sum_{j>i} \left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right)^2 P_iP_j$$

3.º) De los datos del ejercicio número 1.6, obtenemos:

$$\frac{X_1}{P_1} = 6 \;\; ; \;\; \frac{X_2}{P_2} = 9 \;\; ; \;\; \frac{X_3}{P_3} = 8 \;\; ; \;\; P_1 = \frac{1}{6} \;\; ; \;\; P_2 = \frac{2}{6} \;\; ; \;\; P_3 = \frac{3}{6}$$

resultando:

$$V(\hat{X}_{HH}) = \frac{1}{2}\left[(6-9)^2\frac{1}{6}\times\frac{2}{6} + (6-8)^2\frac{1}{6}\times\frac{3}{6} + (8-9)^2\frac{2}{6}\times\frac{3}{6}\right] = \frac{36}{72} = 0{,}5$$

5.3. En un muestreo sin reposición y probabilidades desiguales, se pide:

1.°) Demostrar que la expresión

$$V(\hat{X}_{HT}) = \sum_{i}^{N} \sum_{j>i}^{N} (\pi_i \pi_j - \pi_{ij}) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j} \right)^2 \tag{1}$$

es una forma alternativa de

$$V(\hat{X}_{\mathrm{HT}}) = \sum_{i}^{N} \left(\frac{1 - \pi_i}{\pi_i} \right) X_i^2 + 2 \sum_{i}^{N} \sum_{j>i}^{N} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} X_i X_j \tag{2}$$

donde

π_i = probabilidad de u_i de pertenecer a la muestra y

π_{ij} = probabilidad de u_i y u_j de pertenecer conjuntamente a la muestra.

2.°) Calcular la expresión (1) con los datos del ejercicio número 1.7 y comprobar el resultado.

Solución:

1.°) Teniendo en cuenta las siguientes relaciones importantes en este tipo de muestreo:

a) $\sum_{i}^{N} \pi_i = n$

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

c) $\sum_{j(\neq i)}^{N} \pi_i \pi_j = \pi_i (n - \pi_i)$

se deduce

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

de donde

$$\frac{1 - \pi_i}{\pi_i} = \frac{\sum_{j(\neq i)}^{N} (\pi_i \pi_j - \pi_{ij})}{\pi_i^2}$$

que sustituida en (2), nos da:

$$V(\hat{X}_{HT}) = \sum_{i}^{N} \sum_{j(\neq i)}^{N} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_i^2} X_i^2 - 2 \sum_{i}^{N} \sum_{j(>i)}^{N} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_i\pi_j} X_iX_j =$$

$$= \sum_{(i\neq j)}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{X_i^2}{\pi_i^2}\right) - 2 \sum_{i}^{N} \sum_{j(>i)}^{N} (\pi_i\pi_j - \pi_{ij}) \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} =$$

$$= \sum_{i}^{N} \sum_{j(>i)}^{N} (\pi_i\pi_j - \pi_{ij}) \left[\left(\frac{X_i}{\pi_i}\right)^2 + \left(\frac{X_j}{\pi_j}\right)^2 - 2\frac{X_i}{\pi_i}\frac{X_j}{\pi_j}\right] =$$

$$= \sum_{i}^{N} \sum_{j(>i)}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j}\right)^2$$

2.°) De los datos del ejercicio número 1.7, se tiene:

$$\pi_1 = \frac{25}{60} \quad ; \quad \pi_2 = \frac{44}{60} \quad ; \quad \pi_3 = \frac{51}{60}$$

$$\pi_{12} = \frac{9}{60} \quad ; \quad \pi_{13} = \frac{16}{60} \quad ; \quad \pi_{23} = \frac{35}{60}$$

$$\frac{X_1}{\pi_1} = \frac{60}{25} = 2{,}4 \quad ; \quad \frac{X_2}{\pi_2} = \frac{180}{44} = 4{,}091 \quad ; \quad \frac{X_3}{\pi_3} = \frac{240}{51} = 4{,}706$$

que reemplazando en (1) da:

$$V(\hat{X}_{HT}) = \left(\frac{25 \times 44}{60^2} - \frac{9}{60}\right)\left(\frac{60}{25} - \frac{180}{44}\right)^2 + \left(\frac{25 \times 51}{60^2} - \frac{16}{60}\right)\left(\frac{60}{25} - \frac{240}{51}\right)^2 +$$

$$+ \left(\frac{44 \times 51}{60^2} - \frac{35}{60}\right)\left(\frac{180}{44} - \frac{240}{51}\right)^2 = 0{,}445 + 0{,}465 + 0{,}015 = 0{,}925$$

5.4. En un muestreo de unidades elementales se utiliza el esquema mixto de selección de Sánchez-Crespo y Gabeiras (véase ejercicio 1.9), con probabilidad iniciales proporcionales a los tamaños

$$P_i = \frac{M_i}{M} \quad , \quad (i = 1, 2, ..., N)$$

Se pide:

a) Demostrar que la probabilidad de obtener la unidad u_i en la 2.ª extracción es igual a la de obtenerla en la primera: P_i.

b) Demostrar que la varianza del estimador insesgado

$$\hat{X}_{SCG} = \sum_{i}^{n} \frac{X_i}{nP_i}$$

es

$$V(\hat{X}_{SCG}) = \frac{M - nb}{M - b} \times \frac{1}{n} \sum_{i}^{N} \left(\frac{X_i}{P_i} - X \right)^2 P_i$$

c) Calcular la expresión anterior con los datos del ejercicio 1.9 y comprobar el resultado.

Solución:

a) La probabilidad que tiene u_i de ser elegida en la 2.ª extracción es:

$$P(u_i \; ; \; 2.^{a}) = P(u_i \; ; \; 2.^{a}/u_i \; ; \; 1.^{a}) + P(u_i \; ; \; 2.^{a}/u_{j \neq i} \; ; \; 1.^{a}) =$$

$$= \frac{M_i}{M} \times \frac{M_i - b}{M - b} + \sum_{j \neq i} \frac{M_j}{M} \times \frac{M_i}{M - b} = \frac{1}{M(M-b)} \left[M_i(M_i - b) + M_i \sum_{j \neq i} M_j \right] =$$

$$= \frac{1}{M(M-b)} [M_i(M_i - b) + M_i(M - M_i)] = \frac{M_i(M-b)}{M(M-b)} = \frac{M_i}{M} = P_i$$

b)

$$V(\hat{X}_{SCG}) = V\left(\frac{1}{n} \sum_{i}^{n} \frac{X_i}{P_i} \right) =$$

$$= \frac{1}{n^2} \left[\sum_{i}^{N} \frac{X_i^2}{P_i^2} V(e_i) + \sum_{i \neq j}^{N} \frac{X_i X_j}{P_i P_j} \text{Cov}\,(e_i, e_j) \right] \qquad (1)$$

donde e_i es la variable aleatoria «número de veces que la unidad u_i puede resultar seleccionada en una muestra de tamaño n», y sigue una distribución factorial (véase «Un esquema mixto de muestreo con probabilidades desigua-

les», Sánchez-Crespo y Gabeiras, *Rev. Estadística Española*, núm. 15, 1987), con

$$V(e_i) = \frac{M - nb}{M - b} nP_i(1 - P_i)$$

$$\text{Cov}\,(e_i, e_j) = -\frac{M - nb}{M - b} nP_iP_j$$

sustituyendo estos valores en (1), se obtiene:

$$V(\hat{X}_{SCG}) = \frac{M - nb}{M - b} \times \frac{1}{n}\left[\sum_i^N \frac{X_i^2}{P_i^2} P_i(1 - P_i) - \sum_{i \neq j}^N X_iX_j\right] =$$

$$= \frac{M - nb}{M - b} \times \frac{1}{n}\left[\sum_i^N \frac{X_i^2}{P_i} - \left(\sum_i^N X_i\right)^2\right] = \frac{M - nb}{M - b} \times \frac{1}{n} \sum_i^N \left(\frac{X_i}{P_i} - X\right)^2 P_i$$

ya que

$$\sum_i^N \left(\frac{X_i}{P_i} - X\right)^2 P_i = \sum_i^N \frac{X_i^2}{P_i^2} P_i + X^2 \sum_i^N P_i - 2X \sum_i^N X_i = \sum_i^N \frac{X_i^2}{P_i} - X^2$$

c) Siendo:

$$M = 15 \;\; ; \;\; b = 3 \;\; ; \;\; n = 2 \;\; ; \;\; P_1 = \frac{3}{15} \;\; ; \;\; P_2 = \frac{5}{15} \;\; ; \;\; P_3 = \frac{7}{14}$$

$$X_1 = 1 \;\; ; \;\; X_2 = 3 \;\; ; \;\; X_3 = 4$$

se tiene:

$$\frac{X_1}{P_1} = 5 \;\; ; \;\; \frac{X_2}{P_2} = 9 \;\; ; \;\; \frac{X_3}{P_3} = \frac{60}{7} \;\; ; \;\; X = 8$$

Por tanto:

$$V(\hat{X}_{SCG}) = \frac{3}{8}\left[3^2 \times \frac{3}{15} + 1^2 \times \frac{5}{15} + \left(\frac{4}{7}\right)^2 \times \frac{7}{15}\right] = \frac{6}{7}$$

que es el resultado que se obtuvo en el ejercicio 1.9 a partir de la distribución en el muestreo de $\hat{X}_{SCG}$.

5.5. Midzuno propuso el método de selección siguiente: la primera unidad se obtiene con probabilidad proporcional al tamaño y las $n - 1$ unidades restantes con igual probabilidad.

Se pide:

a) La probabilidad π_i de inclusión de la unidad u_i en una muestra de n unidades.

b) La probabilidad π_{ij} de que las unidades u_i y u_j figuren simultáneamente en la muestra.

c) La probabilidad que una determinada muestra tiene de ser elegida y demostrar que es proporcional a la suma de los tamaños de las unidades de la muestra.

Solución:

Sea

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

la probabilidad asignada a la unidad u_i, siendo M_i una medida de su tamaño y

$$M = \sum_{i}^{N} M_i$$

a) La probabilidad de inclusión π_i es igual a la probabilidad de obtener u_i en la primera selección más la probabilidad de que no se obtenga en la primera selección y sí en cualquiera de las $n - 1$ restantes.

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

b) La probabilidad π_{ij} será igual a la probabilidad de que u_i se obtenga en la primera selección y u_j en cualquiera de las restantes, más la de que u_j se obtenga en la primera y u_i en cualquiera de las restantes, más la de que ni u_i ni u_j se obtengan en las dos primeras selecciones y sí se obtengan en las $n - 2$ restantes.

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

$$= \frac{(P_i + P_j) \times (n-1)}{(N-1) \times (N-2)} \times [N-n] + \frac{(n-1) \times (n-2)}{(N-1) \times (N-2)} =$$

$$= \frac{n-1}{N-1} \times \left[\frac{N-n}{N-2} (P_i + P_j) + \frac{n-2}{N-2} \right]$$

c) La probabilidad de obtener u_i en la primera selección es P_i y la de las $n-1$ unidades restantes de la muestra $1 \Big/ \binom{N-1}{n-1}$, y como u_i puede ser cualquiera de las n unidades de la muestra s, tendremos:

$$P(s) = \sum_i^n P_i \times \frac{1}{\binom{N-1}{n-1}} = \frac{1}{M} \times \frac{1}{\binom{N-1}{n-1}} \times \sum_i^n M_i = K \times \sum_i^n M_i$$

luego es proporcional a la suma de los tamaños de las unidades muestrales.

5.6. En una población de $N = 6$ unidades, la respuesta correcta a una determinada pregunta es «SI» en tres unidades y «NO» en las otras tres.

Debido a errores de medida, la probabilidad de obtener respuesta «SI» en las unidades que sin tales errores la darían, es igual a 0,9. Análogamente la probabilidad de dar respuesta «NO» en las unidades que sin errores de medida la darían es también igual a 0.9.

A partir de las respuestas posibles en muestras de tamaño 2, sin reposición y probabilidades iguales se calculan las probabilidades P_1, P_2 y P_3 de que una muestra proporcione 0, 1, 2 respuesta «SI».

Se considera ahora un muestreo con reposición y probabilidades de selección iguales a P_1, P_2 y P_3 con muestras de tamaño $n = 2$ en la población siguiente:

x_i	1	2	3
$P(x_j)$	P_1	P_2	P_3

Se pide:

a) Valores de P_1, P_2 y P_3.

b) Error de muestreo del estimador

$$\hat{X}_{HH} = \sum_i^n \frac{x_i}{nP_i}$$

para la muestra (1; 2).

[Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos (1982).]

Solución:

a) Clases de resultados posibles:

$$A(z = 0,\ t = 2) \quad ; \quad B(z = 2,\ t = 0) \quad ; \quad C(z = 1,\ t = 1)$$

representando con z la respuesta afirmativa y con t la negativa.

Debido a errores de medida, las probabilidades condicionadas a estas tres situaciones serán:

$$P(z = 0/A) = 0{,}9 \times 0{,}9 = 0{,}81$$

$$P(z = 1/A) = 0{,}1 \times 0{,}9 + 0{,}1 \times 0{,}9 = 0{,}18$$

$$P(z = 2/A) = 0{,}1 \times 0{,}1 = 0{,}01$$

$$P(z = 0/B) = 0{,}1 \times 0{,}1 = 0{,}01$$

$$P(z = 1/B) = 0{,}1 \times 0{,}9 + 0{,}9 \times 0{,}1 = 0{,}18$$

$$P(z = 2/B) = 0{,}9 \times 0{,}9 = 0{,}81$$

$$P(z = 0/C) = 0{,}1 \times 0{,}9 = 0{,}09$$

$$P(z = 1/C) = 0{,}9 \times 0{,}9 + 0{,}1 \times 0{,}1 = 0{,}82$$

$$P(z = 2/C) = 0{,}9 \times 0{,}1 = 0{,}09$$

Las probabilidades que se obtienen (para el caso de no cometer errores) con la distribución hipergeométrica, son:

$$P(z=1, t=1) = \frac{\binom{3}{1} \times \binom{3}{1}}{\binom{6}{2}} = \frac{3}{5} \quad ; \quad P(z=2, t=0) = \frac{1}{5} \quad ; \quad P(z=0, t=2) = \frac{1}{5}$$

y, por lo tanto, las probabilidades finales serían:

$$P(z=0) = 0{,}81 \times \frac{1}{5} + 0{,}01 \times \frac{1}{5} + 0{,}09 \times \frac{3}{5} = 0{,}218 = P_1$$

$$P(z=1) = 0{,}18 \times \frac{1}{5} + 0{,}18 \times \frac{1}{5} + 0{,}82 \times \frac{3}{5} = 0{,}564 = P_2$$

$$P(z=2) = 0{,}01 \times \frac{1}{5} + 0{,}81 \times \frac{1}{5} + 0{,}09 \times \frac{3}{5} = 0{,}218 = P_3$$

$$\hat{X} = \frac{1}{2}\left[\frac{1}{0{,}218} + \frac{2}{0{,}564}\right] = 4{,}07$$

b) $$\hat{V}(\hat{X}) = \frac{1}{n}\sum_{i}^{n} \frac{\sum_{i}^{n}\left(\frac{X_i}{P_i} - \hat{X}\right)^2}{n-1}$$

$$\hat{V}(\hat{X}) = \frac{1}{2(2-1)}\left[\left[\frac{1}{0{,}218} - 4{,}07\right]^2 + \left[\frac{2}{0{,}564} - 4{,}07\right]^2\right] = 0{,}271$$

$$\hat{\sigma}_{\hat{X}} = 0{,}52$$

5.7. Brewer (1963) propuso el siguiente método de selección: La primera unidad se extrae con probabilidad proporcional a

$$k_i = P_i \frac{(1 - P_i)}{(1 - 2P_i)}$$

siendo

$$P_i < \frac{1}{2}$$

La segunda extracción se realiza sin reposición y con probabilidades proporcionales a P_j. El tamaño de la muestra es $n = 2$.

Se pide:

a) Demostrar que la probabilidad de inclusión de la unidad u_i en la muestra es igual a nP_i.

b) Calcular la probabilidad conjunta que tienen las unidades u_i y u_j de pertenecer simultáneamente a la muestra.

c) Indicar si π_i es proporcional al correspondiente tamaño de la unidad u_i.

Solución:

a) $\pi_i = P(u_i \in 1.^{a}) + P(u_i \in 2.^{a}/u_{j\neq i} \in 1.^{a}) =$

$$= \frac{\frac{P_i(1-P_i)}{1-P_i}}{\sum_i^N \frac{P_i(1-P_i)}{1-2P_i}} + \frac{\sum_{j\neq i}^N \frac{P_i(1-P_j)}{1-2P_j} \times \frac{P_i}{1-P_j}}{\sum_i^N \frac{P_i(1-P_i)}{1-2P_i}} =$$

$$= \frac{P_i\left(\frac{1-2P_i+P_i}{1-2P_i} + \sum_{j\neq i}^N \frac{P_j}{1-2P_j}\right)}{\frac{1}{2}\sum_i^N \frac{P_i(1+(1-2P_i))}{1-2P_i}} =$$

$$= \frac{P_i\left(1+\sum_j^N \frac{P_j}{1-2P_j}\right)}{\frac{1}{2}\sum_i^N \frac{P_i(1+(1-2P_i))}{1-2P_i}} = \frac{P_i\left(1+\sum_j^N \frac{P_j}{1-2P_j}\right)}{\frac{1}{2}\left(1+\sum_i^N \frac{P_i}{1-2P_i}\right)} = 2P_i$$

b) $K = \sum_i^N k_i = \sum_i^N \frac{P_i(1-P_i)}{1-2P_i} = \frac{1}{2}\sum_i^N \frac{P_i(2-2P_i)}{1-2P_i} =$

$$= \frac{1}{2}\left[\sum_i^N P_i + \sum_i^N \frac{P_i}{1-2P_i}\right] = \frac{1+\sum_i^N \frac{P_i}{1-2P_i}}{2}$$

$$\pi_{ij} = \frac{P_i(1-P_i)}{(1-2P_i)\times K} \times \frac{P_j}{1-P_j} + \frac{P_j(1-P_j)}{(1-2P_j)\times K} \times \frac{P_i}{1-P_j} = \frac{P_iP_j}{K} \times \left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right]$$

y, por consiguiente:

$$\pi_{ij} = \frac{2P_iP_j}{1+\sum_i^N \frac{P_i}{1-2P_i}} \times \left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right]$$

c) Como

$$\pi_i = nP_i = n \times \frac{M_i}{M} = \frac{n}{M} \times M_i$$

es proporcional a M_i.

5.8. En una población con $N = 3$ unidades, los valores de la variable en estudio son: $X_i = (1; 3; 4)$. *Se selecciona una muestra de* $n = 2$ unidades utilizando el procedimiento de selección de Brewer, siendo:

$$P_i = \left(\frac{3}{15}; \frac{5}{15}; \frac{7}{15}\right)$$

Calcular la varianza del estimador del total.

Solución:

La expresión general de la varianza es:

$$V(\hat{X}_{HT}) = \sum_i^N \frac{1 - \pi_i}{\pi_i} X_i^2 + 2 \sum_{i<j} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \times X_i X_j$$

sustituyendo los valores

$$\pi_i = 2P_i \quad \text{y} \quad \pi_{ij} = \frac{2P_iP_j}{1 + \sum_i^N \frac{P_i}{1-2P_i}} \times \left[\frac{1}{1-2P_i} + \frac{1}{1-2P_j}\right]$$

en la expresión general de la varianza, para el estimador del total, en el muestreo sin reposición debido a Horvitz y Thompson, tenemos:

$$\sum_i^3 \frac{1 - \pi_i}{\pi_i} \times X_i^2 = 1{,}5 + 4{,}5 + 1{,}1439 = 7{,}143$$

$$2 \sum_{i<j} \left(\frac{\pi_{ij}}{\pi_i \pi j} - 1\right) \times X_i X_j = -4{,}5 - 0{,}857 - 0{,}857 = -6{,}214$$

$$V(\hat{X}_{HT}/B) = 7{,}143 = 6{,}214 = 0{,}929$$

por ser:

$$\pi_1 = \frac{6}{15} \quad \pi_2 = \frac{10}{15} \quad \pi_3 = \frac{14}{15} \quad \sum_i^N \pi_i = 2$$

$$\pi_{12} = \frac{1}{15} \quad \pi_{13} = \frac{5}{15} \quad \pi_{23} = \frac{9}{15} \quad \sum_{i<j}^N \pi_{ij} = 1$$

NOTA: Una expresión alternativa para la varianza del total con el procedimiento de Horvitz y Thompson es:

$$V(\hat{X}_{HT}) = \sum_{i<j}^{N} (\pi_i\pi_j - \pi_{ij}) \times \left(\frac{X_i}{\pi_i} - \frac{X_j}{\pi_j}\right)^2$$

En efecto:

$$(\pi_1\pi_2 - \pi_{12}) \times \left(\frac{X_1}{\pi_1} - \frac{X_2}{\pi_2}\right)^2 = 0{,}800$$

$$(\pi_1\pi_3 - \pi_{13}) \times \left(\frac{X_1}{\pi_1} - \frac{X_3}{\pi_3}\right)^2 = 0{,}128 \qquad V(\hat{X}_{HT}) = 0{,}929$$

$$(\pi_2\pi_3 - \pi_{23}) \times \left(\frac{X_2}{\pi_2} - \frac{X_3}{\pi_3}\right)^2 = 0{,}001$$

5.9. Una población está formada por tres conglomerados, siendo los valores de la variable en que estamos interesados $X_i = (40; 42; 43)$ y los tamaños de los mencionados conglomerados $M_i = (120; 125; 130)$. En un muestreo con reposición y probabilidades proporcionales a los tamaños, se pide:

a) Calcular la varianza del total $\hat{X}_{HH} = \frac{1}{n}\sum_{i}^{n}\frac{X_i}{P_i}$ debido a Hansen y Hurwitz utilizando las expresiones alternativas:

$$(1) \quad V(\hat{X}_{HH}) = \frac{1}{n}\left[\sum_{i}^{N}\frac{X_i^2}{P_i} - X^2\right]$$

y

$$(2) \quad V(\hat{X}_{HH}) = \frac{1}{n}\sum_{i<j}^{N}\left(\frac{X_i}{P_i} - \frac{X_j}{P_j}\right) \times P_iP_j$$

b) Calcular la varianza de dicho estimador utilizando el método de Sánchez-Crespo y Gabeiras y expresar en porcentaje la ganancia en varianza

$$(b = 120;\ M = 375) \quad , \quad V(\hat{X}_{SCG}) = \left[\frac{(M - nb)}{M - b}\right] \times (V(X_{HH}))$$

En los dos apartados el tamaño de la muestra es $n = 2$.

Solución:

a)

(1)

$\frac{X_i^2}{P_i}$
5.000,0000
5.292,5292
5.333,1410
$\sum_i^N \frac{X_i^2}{P_i} = 15.626{,}6702$

$X^2 = 15.625$

$\frac{1}{2}\sum_1^N \frac{X_i^2}{P_i} - \frac{15.625}{2} = 0{,}3351$

$V(\hat{X}_{HH}) = 0{,}3351$

(2)

$A = \left(\frac{X_i}{P_i} - \frac{X_i}{P_j}\right)^2$	P_iP_j	$\frac{A \times P_iP_j}{2}$
$(125 - 126{,}0126)^2$	0,1066	0,0547
$(125 - 124{,}0265)^2$	0,1109	0,0525
$(126{,}0126 - 124{,}0265)^1$	0,1156	0,2279
		$V(\hat{X}_{HH}) = 0{,}3351$

b)

$$\frac{M - nb}{M - b} = \frac{375 - 240}{375 - 120} = \frac{135}{255} = 0{,}5294$$

$$V(\hat{X}_{SCG}) = 0{,}5294 \times 0{,}3351 = 0{,}1774$$

La ganancia en varianza sería:

$$\frac{0{,}3351 - 0{,}1774}{0{,}3351} \times 100 = 47\,\%$$

5.10. En un muestreo con probabilidades desiguales y sin reemplazamiento, con $n = 2$, Durbin (1967) propuso el siguiente método de selección: la primera unidad es seleccionada con probabilidad dada P_i. Si la unidad U_i se extrajo primeramente, la probabilidad de que la unidad u_j se extraiga en 2.° lugar se hace proporcional a K_j, siendo

$$K_j = P_j\left[\frac{1}{1 - 2P_i} + \frac{1}{1 - 2P_j}\right]$$

Se pide:

1.º) La probabilidad π_i de inclusión de la unidad u_i en la muestra.

2.º) La probabilidad de que las unidades u_i y u_j figuren simultáneamente en la muestra.

Comprobar que los valores de π_i y π_{ij} son idénticos a los obtenidos con el método de selección de Brewer (véase ejercicio 5.7).

Solución:

$$1.^{o}) \quad \pi_i = P_i + \sum_{j(j\neq i)} P_j \frac{P_i\left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right]}{\sum_{i(i\neq j)} P_i\left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right]} =$$

$$= P_i \left\{ 1 + \sum_{j(j\neq i)} P_j \frac{\left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right]}{\sum_{i(i\neq j)} P_i\left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right]} \right\}$$

El denominador de la anterior expresión, resulta:

$$\sum_{i(i\neq j)} P_i\left[\frac{1}{1-2P_j} + \frac{1}{1-2P_i}\right] = \frac{1-P_j}{1-2P_j} + \sum_{i(i\neq j)} \frac{P_i}{1-2P_i} =$$

$$= \frac{1-P_j}{1-2P_j} - \frac{P_j}{1-2P_j} + \frac{P_j}{1-2P_j} + \sum_{i(i\neq j)} \frac{P_i}{1-2P_i} = 1 + \sum_i \frac{P_i}{1-2P_i}$$

que es una constante.

Una vez reducido a constante el denominador, el numerador tiene la misma estructura que el denominador, luego:

$$\pi_i = P_i \left\{ 1 + \frac{1 + \sum_j \frac{P_j}{1-2P_j}}{1 + \sum_i \frac{P_i}{1-2P_i}} \right\} = 2P_i$$

2.º) $\pi_{ij} = P(u_i)P(u_j/u_i) + P(u_j)P(u_i/u_j) =$

$$= P_i \frac{P_j\left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]}{\sum_{j(\neq i)} P_j\left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]} + P_j \frac{P_i\left[\frac{1}{1-2P_j}+\frac{1}{1-2P_i}\right]}{\sum_{i(\neq j)} P_i\left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]} =$$

$$= \frac{2P_iP_j}{1+\sum_i \frac{P_i}{1-2P_i}}\left[\frac{1}{1-2P_i}+\frac{1}{1-2P_j}\right]$$

obteniéndose para π_i y π_{ij} expresiones idénticas a las de Brewer.

CAPITULO VI
Muestreo de conglomerados

6.1. De una población formada por N conglomerados se selecciona una muestra de tamaño n con el procedimiento siguiente: la 1.ª extracción se realiza con probabilidades desiguales P_i, siendo

$$\sum_{i}^{N} P_i = 1$$

los $n - 1$ conglomerados restantes de la muestra se eligen con probabilidades iguales. Todas las extracciones se hacen sin reposición. Se pide:

a) La probabilidad π_i de que el conglomerado u_i aparezca en la muestra.

b) Comprobar que $\sum_{i}^{N} \pi_i = n$.

c) Calcular una estimación insesgada del total poblacional X, siendo $N = 50$, $n = 4$, X_i el total del conglomerado i-ésimo, y conociendo los siguientes datos de los conglomerados de la muestra:

P_i	0,026	0,017	0,022	0,013
X_i	100	80	120	60

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1984.)

Solución:

a) La probabilidad de que la unidad u_i pertenezca a la muestra es igual a la de que aparezca en la 1.ª extracción, más la probabilidad de que no aparezca en la primera multiplicada por la de que aparezca en una de las $n - 1$ extracciones restantes a partir de las $N - 1$ unidades que quedan en la población. Es decir:

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

b)

$$\sum_{i}^{N} \pi_i = N\frac{n-1}{N-1} + \frac{N-n}{N-1}\sum_{i}^{N} P_i = \frac{Nn-n}{N-1} = n\frac{N-1}{N-1} = n$$

c)

$$\hat{X} = \sum_{i}^{n} \frac{X_i}{\pi_i} = \frac{100}{\frac{3}{49} + 0{,}026 \times \frac{46}{49}} + \frac{80}{\frac{3}{49} + 0{,}017 \times \frac{46}{49}} +$$

$$+ \frac{120}{\frac{3}{49} + 0{,}022 \times \frac{46}{49}} + \frac{60}{\frac{3}{49} + 0{,}013 \times \frac{46}{49}} = 4.487$$

6.2. Supongamos que en un estudio de muestreo por conglomerados se obtuvo una muestra de n conglomerados de $\bar{M}$ unidades cada uno, con reposición. Sean b y w estimadores insesgados de las varianzas «entre» y «dentro» de conglomerados, respectivamente. Expresando el tamaño de la muestra en unidades elementales, obtener una estimación de la eficiencia relativa del muestreo por conglomerados comparado con el de unidades elementales, estimando en ambos casos las varianzas de muestreo insesgadamente.

(M. N. Murthy, 1967, *Sampling Theory and Methods*, p. 315.)

Solución:

Si designamos por $\bar{\bar{X}}$ y $\bar{\bar{X}}_i$ la media por unidad elemental en la población y en el conglomerado i-ésimo, respectivamente, la identidad fundamental del análisis de la varianza en el muestreo por conglomerados, tiene la siguiente expresión:

$$\sum_{i}^{N}\sum_{j}^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2 = \sum_{i}^{N}\sum_{j}^{\bar{M}} (X_{ij} - \bar{\bar{X}}_i)^2 + \sum_{i}^{N}\sum_{j}^{\bar{M}} (\bar{\bar{X}}_i - \bar{\bar{X}})^2$$

o bien, en función de las varianzas:

$$N\bar{M}\sigma^2 = N\bar{M}\sigma_w^2 + N\sigma_b^2 \qquad (1)$$

siendo

$$\sigma^2 = \frac{\sum_i \sum_j (X_{ij} - \bar{\bar{X}})^2}{N\bar{M}} \qquad \text{varianza poblacional}$$

$$\sigma_w^2 = \frac{\sum_i \sum_j (X_{ij} - \bar{X}_i)^2}{N\bar{M}} \qquad \text{varianza «dentro» de conglomerados}$$

$$\sigma_b^2 = \frac{\sum_i^N \bar{M}(\bar{X}_i - \bar{\bar{X}})^2}{N} \qquad \text{varianza «entre» conglomerados}$$

La varianza de $\hat{X}$ en un muestreo de conglomerados con reposición viene dada por:

$$V(\hat{X})_{\text{cong.}} = N^2 \frac{\bar{M}\sigma_b^2}{n}$$

mientras que en un muestreo aleatorio de unidades elementales con reposición e igual tamaño de muestra:

$$V(\hat{X})_{\text{al.}} = (N\bar{M})^2 \frac{\sigma^2}{n\bar{M}}$$

Por tanto, la eficiencia relativa será:

$$\frac{V(\hat{X})_{\text{congl.}}}{V(\hat{X})_{\text{al.}}} = \frac{\sigma_b^2}{\sigma^2}$$

Si b es un estimador insesgado de σ_b^2, y w es un estimador insesgado de σ_w^2, la estimación de la eficiencia relativa, resulta:

$$\text{Eficiencia relativa estimada} = \frac{b}{\frac{b}{\bar{M}} + w}$$

ya que, según (1):

$$\sigma^2 = \frac{\sigma_b^2}{\bar{M}} + \sigma_w^2$$

6.3. Consideremos una población de $N = 100$ conglomerados del mismo tamaño $\bar{M} = 4$ unidades elementales, en la que la proporción de personas con un cierto atributo es $P = 0{,}5$. En una muestra de $n = 5$ conglomerados se obtuvieron los siguientes resultados:

Conglomerado (i)	Unidades elementales con el atributo (A_i)
1	2
2	3
3	1
4	2
5	1

Se pide: Estimar la eficiencia relativa del muestreo por conglomerados respecto a la del muestreo aleatorio simple.

Solución:

Si definimos la eficiencia relativa mediante la relación de varianzas, tendremos:

En caso de muestreo aleatoria: $V_a(\hat{P}) = (1 - f)\dfrac{S^2}{n\bar{M}}$

En caso de muestreo por conglomerados: $V_c(\hat{P}) = (1 - f)\dfrac{S_P^2}{n}$

donde S_P^2 es la cuasivarianza entre proporciones en los conglomerados, es decir:

$$S_P^2 = \frac{\sum_i^N (P_i - \bar{P})^2}{N - 1} \quad \text{donde} \quad \bar{P} = \frac{\sum_i^N P_i}{N}$$

siendo la eficiencia relativa:

$$\frac{V_c(\hat{P})}{V_a(\hat{P})} = \frac{\bar{M}S_P^2}{S^2}$$

De los datos del enunciado se obtiene

$$S^2 = \frac{N\bar{M}P(1 - P)}{N\bar{M} - 1} = \frac{400 \times 0{,}5 \times 0{,}5}{399} = 0{,}2506$$

S_P^2 es estimada a través de la muestra mediante:

$$\hat{S}_P^2 = \frac{\sum_i^n (P_i - \hat{\bar{P}})^2}{n-1} = \frac{1}{\bar{M}_2} \frac{\sum_i^n (A_i - \bar{A})^2}{n-1} = \frac{1}{\bar{M}^2(n-1)} \left(\sum_i^n A_i^2 - n\bar{A}^2 \right) =$$

$$= \frac{1}{64} \left(19 - 5 \times \frac{81}{25} \right) = 0{,}04375$$

Luego, la eficiencia relativa es:

$$\frac{4 \times 0{,}04375}{0{,}2506} = 0{,}7$$

es decir, al utilizar muestreo por conglomerados se tiene, en este caso, una ganancia en precisión del 30 % respecto al muestreo aleatorio simple.

6.4. A partir de las expresiones:

$$V(\bar{\bar{x}}) \doteqdot (1-f) \times \frac{S^2}{n\bar{M}} (1 + (\bar{M} - 1) \times \delta) \qquad V(\bar{\bar{x}}) = (1-f) \times \frac{S_b^2}{n\bar{M}}$$

y del análisis de la varianza:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios
«Entre submuestras»	$N-1$	$\sum_i^N \sum_j^{\bar{M}} (\bar{X}_i - \bar{\bar{X}})^2$	S_b^2
«Dentro de submuestras»	$N(\bar{M}-1)$	$\sum_i^N \sum_j^{\bar{M}} (X_{ij} - \bar{X}_i)^2$	S_w^2
Total	$N\bar{M}-1$	$\sum_i^N \sum_j^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2$	S^2

demostrar que:

$$\delta = \doteqdot \frac{S_b^2 - S^2}{S^2(\bar{M}-1)} \qquad S_w^2 \doteqdot S^2(1-\delta)$$

Solución:

De la expresión

$$S_b^2 \doteqdot S^2 \times (1 + (\bar{M} - 1) \times \delta)$$

se deduce

$$\delta \doteqdot \frac{(S_b^2 - S^2)}{S^2 \times (\bar{M} - 1)}$$

y de la tabla del análisis de la varianza, se obtiene:

$$(N\bar{M} - 1) \times S^2 = (N - 1) \times S_b^2 + N \times (\bar{M} - 1) \times S_w^2$$

que para la aproximación

$$N - 1 \doteqdot N \quad \bar{M} - 1 \doteqdot M \quad N\bar{M} - 1 \doteqdot N\bar{M}$$

nos da:

$$N\bar{M}S^2 \doteqdot NS_b^2 + N\bar{M}S_w^2$$

o bien

$$\bar{M}S^2 \doteqdot S_b^2 + \bar{M}S_w^2$$

de donde

$$S_w^2 \doteqdot \frac{\bar{M}S^2 - S_b^2}{M}$$

y como

$$S_b^2 = S^2 \times (1 + (\bar{M} - 1)\delta)$$

tendremos:

$$S_w^2 \doteqdot \frac{(\bar{M} - 1)(S^2 - S^2\delta)}{\bar{M}} \doteqdot S^2 - S^2\delta$$

$$S_w^2 \doteqdot S^2 \times (1 - \delta)$$

6.5. Una población de 5.000 elementos es estratificada en 5 estratos, conteniendo cada uno 20 conglomerados de 50 elementos. Un análisis de la varianza en la población objeto de estudio ofrece los siguientes datos:

Causas de variación	Cuadrados medios
Entre estratos	300
Entre conglomerados dentro de estratos	30
Entre elementos dentro de los conglomerados	15

Despreciando el factor de corrección de poblaciones finitas, ¿qué relación existiría entre las precisiones de una muestra por conglomerados monoetápica y una muestra aleatoria simple del mismo tamaño, en el caso:

a) De no hacerse estratificación (con selección aleatoria simple).

b) De hacerse estratificación (con afijación proporcional)?

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1984.)

Solución:

Designando los estratos, conglomerados y elementos mediante los subíndices h, i, j, respectivamente, los datos del enunciado permiten construir el siguiente cuadro completo de análisis de la varianza:

	Causas de variación	Cuadrados medios	Grados de libertad (g.l.)	Suma de cuadrados
Datos del enunciado	Entre estratos	$A=300$	4	$1.200=\sum_h\sum_i\sum_j(\bar{X}_h-\bar{X})^2=A'$
	Entre conglomerados dentro de estratos	$B=30$	95	$2.850=\sum_h\sum_i\sum_j(\bar{X}_{hi}-\bar{X}_h)^2=B'$
	Entre elementos dentro de conglomerados	$C=15$	4.900	$73.550=\sum_h\sum_i\sum_j(X_{hij}-\bar{X}_{hi})^2=C'$
Datos calculados	Total	$D=15{,}51$	4.999	$77.550=\sum_h^5\sum_i^{20}\sum_j^{50}(X_{hij}-\bar{X})^2=D'$
	Entre conglomerados	$E=40{,}91$	99	$4.050=\sum_h\sum_i\sum_j(\bar{X}_{hi}-\bar{X})^2=E'$
	Dentro de estratos	$F=15{,}28$	4.995	$76.350=\sum_h\sum_i\sum_j(X_{hij}-\bar{X}_h)^2=F'$

Datos calculados:

Total $\qquad D' = A' + B' + C' \quad ; \quad D = \dfrac{D'}{g.l.}$

Entre conglomerados $\qquad E' = D' - C' = A' + B' \quad ; \quad E = \dfrac{E'}{\text{g.l.}}$

Dentro de estratos $\qquad F' = D' - A' \quad ; \quad F = \dfrac{F'}{g.l.}$

a) Si no hay estratificación:

$$\frac{\text{Varianza conglomerados}}{\text{Varianza aleatorio}} = \frac{S_b^2}{S^2} = \frac{E}{D} = \frac{40{,}91}{15{,}51} = 2{,}64$$

Es decir, la varianza del muestreo por conglomerados es 2,64 veces mayor que la del aleatorio. El efecto del diseño es, pues, 2,64.

b) Si se estratifica (con afijación proporcional):

$$\frac{\text{Varianza conglomerados estratificado}}{\text{Varianza aleatorio estratificado}} = \frac{S_{b(h)}^2}{S_{w(h)}^2} = \frac{B}{F} = \frac{30}{15{,}28} = 1{,}96$$

El efecto del diseño disminuye con la estratificación.

6.6. En un muestreo de conglomerados de tamaño $\bar{M}$ sin submuestreo, a partir del siguiente cuadro de análisis de la varianza para características cuantitativas:

Fuente variación	g.l.	Suma de cuadrados	Cuadrados medios	Estimadores insesgados
Entre	$N-1$	$A=\sum_i^N \bar{M}(\bar{X}_i-\bar{\bar{X}})^2$	$S_b^2=\dfrac{A}{N-1}$	$\hat{S}_b^2=\dfrac{\sum_i^n\left(\bar{M}\bar{X}_i-\dfrac{\sum_i^n \bar{M}\bar{X}_i}{n}\right)^2}{\bar{M}(n-1)}$
Dentro	$N(\bar{M}-1)$	$B=\sum_i^N\sum_j^{\bar{M}}(X_{ij}-\bar{X}_i)^2$	$S_w^2=\dfrac{B}{N(\bar{M}-1)}$	$\hat{S}_w^2=\dfrac{\sum_i^n\sum_j^{\bar{M}}(X_{ij}-\bar{X}_i)^2}{n(\bar{M}-1)}$
Total	$N\bar{M}-1$	$C=\sum_i^N\sum_j^{\bar{M}}(X_{ij}-\bar{\bar{X}})^2$	$S^2=\dfrac{C}{N\bar{M}-1}$	$\hat{S}^2=\dfrac{[N(\bar{M}-1)\hat{S}_w^2+(N-1)\hat{S}_b^2]}{N\bar{M}-1}$

Se pide: Obtener un cuadro análogo para el estudio de proporciones o características cualitativas.

Solución:

En el caso de características cualitativas, las X_{ij} sólo tomarán valores 1 ó 0 según la unidad u_{ij} posea o no la característica. De ello resultan las siguientes relaciones:

$$\bar{\bar{X}} = \frac{\sum_{i}^{N} \sum_{j}^{\bar{M}} X_{ij}}{N\bar{M}} = P$$

proporción poblacional

$$\bar{\bar{X}}_i = \frac{\sum_{j}^{\bar{M}} X_{ij}}{\bar{M}} = P_i$$

proporción en el conglomerado i-ésimo

$$\sum_{i}^{N} \bar{M}(\bar{\bar{X}}_i - \bar{\bar{X}})^2 = \sum_{i}^{N} \bar{M}(P_i - P)^2$$

$$\sum_{i}^{N} \sum_{j}^{\bar{M}} (X_{ij} - \bar{\bar{X}}_i)^2 = \sum_{i}^{N} \left[\sum_{j}^{\bar{M}} X_{ij}^2 + \bar{M}\bar{\bar{X}}_i^2 - 2\bar{\bar{X}}_i \sum_{j}^{\bar{M}} X_{ij} \right] =$$

$$= \sum_{i}^{N} [\bar{M}P_i + \bar{M}P_i^2 - 2P_i\bar{M}P_i] = \sum_{i}^{N} \bar{M}P_i(1 - P_i)$$

$$\sum_{i}^{N} \sum_{j}^{\bar{M}} (X_{ij} - \bar{\bar{X}})^2 = \sum_{i} \sum_{j} (X_{ij}^2) + N\bar{M}\bar{\bar{X}}^2 - 2\bar{\bar{X}} \sum_{i} \sum_{j} X_{ij} =$$

$$= N\bar{M}P + N\bar{M}P^2 - 2PN\bar{M}P = N\bar{M}P(1 - P)$$

Obteniéndose el siguiente cuadro de análisis de la varianza para características cualitativas:

Fuente de variación	g.l.	Suma de cuadrados	Cuadrados medios	Estimadores insesgados
Entre	$N-1$	$A=\sum_i^N \bar{M}(P_i-P)^2$	$S_b^2=\dfrac{A}{N-1}$	$\hat{S}_b^2=\sum_i^n \bar{M}\dfrac{\left(P_i-\frac{1}{n}\sum_i^n P_i\right)^2}{n-1}$
Dentro	$N(\bar{M}-1)$	$B=\sum_i^N \bar{M}P_i(1-P_i)$	$S_w^2=\dfrac{B}{N(\bar{M}-1)}$	$\hat{S}_w^2=\dfrac{\sum_i^n \bar{M}P_i(1-P_i)}{n(\bar{M}-1)}$
Total	$N\bar{M}-1$	$C=N\bar{M}P(1-P)$	$S^2=\dfrac{C}{N\bar{M}-1}$	$\hat{S}^2=\dfrac{[N(\bar{M}-1)\hat{S}_w^2+(N-1)\hat{S}_b^2]}{N\bar{M}-1}$

6.7. Una población está formada por $N = 300$ conglomerados de $\bar{M} = 50$ elementos. Se obtiene una muestra de $n = 5$ conglomerados, sin reposición y probabilidades iguales. La proporción de unidades elementales que pertenecen a una cierta clase en cada uno de los conglomerados muestrales es:

$$P_i = (0{,}14 \;;\; 0{,}20 \;;\; 0{,}18 \;;\; 0{,}12 \;;\; 0{,}16)$$

Se pide:

a) Calcular los estimadores insesgados de las varianzas «entre» (S_b^2) y «dentro» (S_w^2) de conglomerados, definidas mediante:

$$S_b^2 = \frac{1}{N-1}\sum_i^N \bar{M}(P_i - P)^2 \quad , \quad S_w^2 = \frac{1}{N(\bar{M}-1)}\sum_i^N \bar{M}P_i(1-P_i)$$

b) Estimar el total de clase y su error de muestreo absoluto y relativo.

c) Estimar el coeficiente de homogeneidad.

Solución:

a) Estimaciones insesgadas de S_b^2 y S_w^2 se obtienen, respectivamente, mediante las expresiones:

$$\hat{S}_b^2 = \frac{\bar{M}}{n-1}\left[\sum_i^n P_i^2 - n\bar{P}^2\right] \quad ; \quad \hat{S}_w^2 = \frac{\bar{M}}{n(\bar{M}-1)}\left[\sum_i^n P_i - \sum_i^n P_i^2\right]$$

donde

$$\bar{P} = \frac{1}{n} \sum_{i}^{n} P_i = 0{,}16$$

Sustituyendo los valores de $\bar{P}, n, \bar{M}, N$ y P_i, obtenemos:

$$\hat{S}_b^2 = \frac{50}{4}(0{,}132 - 0{,}128) = 0{,}05 \quad ; \quad \hat{S}_w^2 = \frac{50}{5 \times 49}(0{,}8 - 0{,}132) = 0{,}1363$$

b) Estimador del total de clase:

$$\hat{A} = N\bar{M}\bar{P} = 15.000 \times 0{,}16 = 2.400$$

siendo el estimador insesgado de la varianza:

$$\hat{V}(\hat{A}) = (N\bar{M})^2 \frac{\sum_{i}^{n} (P_i - \bar{P})^2}{n(n-1)} = \frac{(N\bar{M})^2}{n(n-1)} \left[\sum_{i}^{n} P_i^2 - n\bar{P}^2 \right] =$$

$$= \frac{15.000^2}{5 \times 4} \times 0{,}004 = 45.000$$

Los errores de muestreo absoluto y relativos son, respectivamente:

$$\sqrt{\hat{V}(\hat{A})} = 212 \quad , \quad \frac{\sqrt{\hat{V}(\hat{A})}}{\hat{A}} \times 100 = \frac{212}{2.400} \times 100 = 8{,}8\,\%$$

c) El coeficiente de homogeneidad puede calcularse en función de S_b^2 y S_w^2 mediante la expresión:

$$\delta = 1 - \frac{N\bar{M}S_w^2}{N(\bar{M}-1)S_w^2 + (N-1)S_b^2} = 1 - \frac{2.045}{2.019} = -0{,}0129$$

6.8. En una muestra aleatoria de $n = 10$ viviendas, el número de personas (M_i) y sus contestaciones afirmativas (A_i) a una determinada pregunta son:

Número de vivienda	M_i	A_i
1	4	2
2	2	1
3	6	4
4	1	1
5	5	2
6	3	1
7	3	2
8	8	5
9	1	0
10	4	3

Se pide:

a) Estimar la proporción P de respuestas afirmativas y su error de muestreo.

b) Estimar el valor del coeficiente de correlación intraconglemerados δ, suponiendo $\bar{M} = 4$.

Solución:

a) Como no se dice nada sobre el tamaño de la población, supondremos que N es grande en relación a n de manera que pueda prescindirse del factor de corrección sobre poblaciones finitas.

El valor poblacional

$$P = \frac{\sum_i^N A_i}{\sum_i^N M_i}$$

se estima mediante el estimador de razón

$$\hat{P} = \frac{\sum_i^n A_i}{\sum_i^n M_i} = \frac{21}{37} = 0{,}57$$

Su varianza se estima mediante la expresión aproximada:

$$V(\hat{P}) \simeq \frac{1}{n\left(\frac{\sum_{i}^{n} M_i}{n}\right)^2} \frac{\sum_{i}^{n} (A_i - \hat{P}M_i)^2}{n-1}$$

siendo:

$$\sum_{i}^{n} (A_i - \hat{P}M_i)^2 = \sum_{i}^{n} A_i^2 - 2\hat{P} \sum_{i}^{n} A_i M_i + \hat{P}^2 \sum_{i}^{n} M_i^2$$

resulta:

$$V(\hat{P}) \simeq \frac{1}{10(3{,}7)^2 \times 9}\left[65 - 2 \times \frac{21}{37} \times 106 + \left(\frac{21}{37}\right)^2 \times 181\right] = 0{,}00242$$

Obteniéndose la estimación del error de muestreo mediante:

$$\hat{\sigma}_{\hat{P}} = \sqrt{0{,}00242} = 0{,}049$$

b) El coeficiente de correlación intraconglomerados puede definirse mediante la expresión:

$$\delta = 1 - \frac{N\bar{M}S_w^2}{N(\bar{M}-1)S_w^2 + (N-1)S_b^2} \simeq \frac{\bar{M}S_w^2}{(\bar{M}-1)S_w^2 + S_b^2}$$

donde:

$$S_b^2 = \frac{\sum_{i}^{N} \bar{M}(P_i - P)^2}{N-1}$$

varianza «entre» conglomerados

$$S_w^2 = \frac{\sum_{i}^{N} \bar{M}P_i(1 - P_i)}{N(\bar{M}-1)}$$

varianza «dentro» de conglomerados.

Los estimadores de S_w^2 y S_b^2 se obtienen, a partir de la muestra, mediante las siguientes expresiones:

$$\hat{S}_w^2 = \frac{\sum_i^n \bar{M} P_i(1 - P_i)}{n(\bar{M} - 1)} \quad ; \quad \hat{S}_b^2 = \frac{\sum_i^n \bar{M}(P_i - \bar{P})^2}{n - 1} \quad ; \quad \bar{P} = \frac{\sum_i^n P_i}{n}$$

con los datos del enunciado construimos el siguiente cuadro de valores:

n	P_i	$1-P_i$	$P_i(1-P_i)$	P_i^2
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
3	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{4}{9}$
4	1	0	0	1
5	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{6}{25}$	$\frac{4}{25}$
6	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{9}$	$\frac{1}{9}$
7	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{4}{9}$
8	$\frac{5}{8}$	$\frac{3}{8}$	$\frac{15}{64}$	$\frac{25}{64}$
9	0	1	0	0
1	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{9}{16}$
	$\sum P_i = \frac{653}{120} = 5{,}44$		$\sum P_i(1-P_i) = 1{,}8285$	$\sum P_i^2 = 3{,}613$

$$\bar{P} = 0{,}544 \quad ; \quad \sum_i^n (P_i - \bar{P})^2 = \sum_i^n P_i^2 - n\bar{P}^2 = 3{,}613 - 10(0{,}544)^2 = 0{,}652$$

Por tanto:

$$\hat{S}_w^2 = \frac{4(1{,}8285)}{10(4-1)} = 0{,}2438 \quad ; \quad \hat{S}_b^2 = \frac{4(0{,}652)}{10-1} = 0{,}29$$

Siendo

$$\delta = 1 - \frac{4(0{,}2438)}{3(0{,}2438) + 0{,}29} = 0{,}045$$

6.9. De una población con $L = 4$ estratos se obtuvo una muestra de $n = 15$ unidades primarias. A partir de éstas se han calculado los valores de $\hat{X}_{hi}$ que figuran en la tabla siguiente:

i \ h	1	2	3
1	500	800	600
2	300	1.000	500
3	400	700	700
4	—	900	400
5	—	—	800

siendo $\hat{X}_{hi}$ un estimador insesgado del total X_h basado en la información recogida en la unidad primaria $i$ésima del estrato h-ésimo.

Calcular las estimaciones del total X y las varianzas de $\hat{X}_h$ y $\hat{X}$.

Solución:

De los datos de la tabla se deducen los siguientes valores:

h	n_h	$\hat{X}_h$	$\sum_i^{n_h} \hat{X}_{hi}^2$	$n_h \hat{X}_h^2$	$\hat{V}(\hat{X}_h)$	
1	3	400	50×100^2	3×400^2	3.333,3	$\hat{X} = \sum_h \hat{X}_h = 5.850$
2	4	850	294×100^2	4×850^2	4.166,7	
3	5	600	190×100^2	5×600^2	5.000,0	$V(\hat{X}) = \sum_h V(\hat{X}_h) =$
4	3	4.000	50×1.000^2	5×4.000^2	333.333,3	$= 345.833,3$
	15	5.850			345.833,3	

siendo:

$$\hat{X}_h = \frac{\sum_i^{n_h} \hat{X}_{hi}}{n_h} \quad ; \quad \hat{V}(\hat{X}_h) = \frac{\sum_i^{n_h} \hat{X}_{hi}^2 - n_h \hat{X}_h^2}{n_h(n_h - 1)}$$

6.10. En una población dividida en conglomerados de igual tamaño $\bar{M} = 25$, se obtiene una muestra de $n = 10$ conglomerados. De experiencias anteriores parece razonable asumir la relación $S_b^2 = 8S^2$. Si se obtiene una varianza para el estimador de la media, igual a 0,01, se pide:

a) Valor del coeficiente de homogeneidad δ.

b) Valor de S_b^2 y S^2.

c) Factor por el que habría que multiplicar el tamaño de una muestra aleatoria simple para obtener la misma precisión que en un muestreo de conglomerados.

(Prescíndase del factor $1 - f$.)

Solución:

a) Por ser

$$V(\bar{x}) = \frac{S_b^2}{n\bar{M}} = \frac{S^2}{n\bar{M}}[1 + (\bar{M} - 1)\delta]$$

resulta:

$$8 = 1 + 24\delta \quad ; \quad \delta = 0{,}2916$$

b)

$$0{,}01 = \frac{S_b^2}{250} \quad ; \quad S_b^2 = 2{,}5 \quad ; \quad S^2 = \frac{1}{8} S_b^2 = 0{,}3125$$

c)

$$F = [1 + (\bar{M} - 1)\delta] = 8.$$

6.11. Se decide la realización de una encuesta utilizando n conglomerados de tamaño $\bar{M}$, siendo la función de coste

$$C = 100\sqrt{n} + 200n\bar{M}$$

Por experiencias anteriores se cree que el valor de $\delta = 0{,}8$ varía muy poco para valores de $\bar{M}$ próximos a 4, así como que un valor conjeturado de $\tilde{P} = 0{,}5$ podría considerarse aceptable para diseñar la muestra.

Se pide: Calcular los valores de n y $\bar{M}$, enteros y por defecto, que resultan óptimos dentro de las alternativas $\bar{M} = 3$ y $\bar{M} = 4$ para un coste fijo $C = 30.500$ pts, utilizando la expresión aproximada:

$$V(\hat{P}) \doteqdot \tilde{P}(1 - \tilde{P}) \times \frac{(1 + \delta(\bar{M} - 1))}{n\bar{M}}$$

Solución:

Despejando n en la función de coste, se obtiene:

$$n = \frac{(-100 + \sqrt{100^2 + 800 \times 30.500\bar{M}})}{400\bar{M}}$$

que para $\bar{M} = 3$ y $\bar{M} = 4$ nos da, respectivamente, $n = 49$ y $n = 37$.

De aquí:

$\bar{M}$	n	$n\bar{M}$	$\frac{\tilde{P}(1) - \tilde{P})}{n\bar{M}}$	$1 + \delta(\bar{M} - 1)$	$V(\hat{P})$
3	49	147	0,001701	2,6	0,0044
4	37	148	0,001736	3,4	0,0057

luego los valores óptimos de $\bar{M}$ y n son:

$$\bar{M}_{op} = 3 \qquad n_{op} = 49$$

6.12. En una población compuesta por 10 conglomerados de 100 elementos, se toma una muestra monoetápica de n conglomerados. Por experiencias anteriores se sabe que el modelo de F. Smith:

$$\log S_b^2 = \log S^2 + t \log \bar{M} \quad \text{con} \quad t > 0$$

se ajusta bien en la proximidad de $\bar{M} = 100$ y se conoce el valor de $S_b^2 = 1.173$.

Se pide:

a) Calcular el valor de t y S_w^2 en el supuesto que $\frac{S_b^2}{S^2} = 13{,}8$.

b) Formar la tabla poblacional del análisis de la varianza.

c) Expresar $V(\bar{x})$ en función de S^2, n y $\bar{M}$, utilizando el modelo.

Solución:

a) Del modelo de Smith se deduce:

$$t = \frac{\log\left(\frac{S_b^2}{S^2}\right)}{\log \bar{M}} = \frac{\log 13{,}8}{\log 100} = \frac{1{,}14}{2} = 0{,}57$$

Por otra parte, de la identidad fundamental del análisis de la varianza:

$$(N\bar{M} - 1)S^2 = N(\bar{M} - 1)S_w^2 + (N - 1)S_b^2$$

se obtiene

$$S_w^2 = \frac{(N\bar{M} - 1)S^2 - (N - 1)S_b^2}{N(\bar{M} - 1)} = \frac{999 \times \frac{1.173}{13{,}8} - 9 \times 1.173}{990} = 75{,}11$$

b)

Fuente de Variación	Grados de libertad	Cuadrados medios	Sumas de de cuadrados
Conglomerados	9	$S_b^2 = 1.173$	10.557
Elementos	990	$S_w^2 = 75{,}11$	74.358
Total	999	$S^2 = 85$	84.915

c) Siendo

$$V(\bar{x}) = \left(1 - \frac{n}{N}\right)\frac{S_b^2}{n\bar{M}}$$

sustituyendo S_b^2 por el valor dado por el modelo

$$S_b^2 = S^2\bar{M}^{0{,}57}$$

resulta

$$V(\bar{x}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n\bar{M}^{0,43}}$$

6.13. Consideremos una población de 10.000 elementos, con una varianza conjeturada $S^2 = 85$, que sigue el modelo de Jessen, el cual supone la existencia de una relación matemática entre la varianza dentro de los conglomerados y el tamaño de éstos que se expresa mediante la fórmula $S_w^2 = A\bar{M}^t$, $t > 0$, donde A y t son constantes que no dependen de $\bar{M}$. Se pide:

a) Obtener los valores de t y A sabiendo que con conglomerados de tamaño $\bar{M} = 5$ se ha estimado un valor de $S_w^2 = 64$.

b) Elección del tamaño óptimo del conglomerado (en el sentido de minimizar la varianza del estimador) bajo el siguiente supuesto: Se dispone de un presupuesto de 5.000 unidades monetarias para estimar la media de la población mediante un muestreo de conglomerados en una etapa, pudiendo construirse conglomerados de los siguientes tamaños: $\bar{M} = 1$ (muestreo de unidades elementales), $\bar{M} = 2$, $\bar{M} = 3$, $\bar{M} = 4$, $\bar{M} = 5$ y $\bar{M} = 10$. La función de coste es del tipo $C = 400\sqrt{n} + 10n\bar{M}$, donde n es el número de conglomerados en la muestra y $\bar{M}$ el tamaño de los mismos. (Prescíndase del factor de corrección de poblaciones finitas.)

Solución:

a) Según el modelo se verifica:

$$\log S_w^2 = \log A + t \log \bar{M} \tag{1}$$

Por otra parte, si se considera la población como un único conglomerado de $N\bar{M}$ elementos, el modelo de Jessen conduce a:

$$S^2 = A(N\bar{M})^t$$

donde

$$\log S^2 = \log A + t \log N\bar{M} \tag{2}$$

De (1) y (2) se obtienen:

$$t = \frac{\log S^2 - \log S_w^2}{\log N} = \frac{\log 85 - \log 64}{\log 1.000} = 0{,}03$$

$$\log A = \log S^2 - t \log 10.000 = 1{,}9292 - 0{,}12 = 1{,}8094 \quad ; \quad A = 64{,}48$$

b) De la función de coste:

$$5.000 = 400\sqrt{n} + 10n\bar{M}$$

resulta:

$$\sqrt{n} = \frac{-400 + \sqrt{160.000 + 200.000\bar{M}}}{20\bar{M}} \tag{3}$$

Para los valores de $\bar{M}$ dados en el enunciado se obtienen los siguientes pares de valores que satisfacen la ecuación (3):

$\bar{M}$	n
1	100
2	76
3	62
4	52
5	46
10	28

Por otra parte, podemos definir la varianza de la media, mediante:

$$V(\hat{\bar{X}}) = \frac{S_b^2}{n\bar{M}} = \frac{(N\bar{M} - 1)S^2 - N(\bar{M} - 1)S_w^2}{n\bar{M}(N - 1)}$$

con la simplificación $N\bar{M} - 1 \simeq N\bar{M}$ y $N - 1 \simeq N$

$$V(\hat{\bar{X}}) \simeq \frac{\bar{M}S^2 - (\bar{M} - 1)S_w^2}{n\bar{M}} = \frac{MS^2 - (\bar{M} - 1)A\bar{M}^t}{n\bar{M}} = \frac{S^2 - (\bar{M} - 1)A\bar{M}^{t-1}}{n} =$$

$$= \frac{1}{n}[85 - 64{,}48(\bar{M} - 1)\bar{M}^{-0{,}97}] \tag{4}$$

Reemplazando en (4) los pares de valores $\bar{M}$, n encontrados anteriormente, se obtienen los siguientes valores de la varianza:

$\bar{M}$	n	$V(\hat{\bar{X}})$
1	100	0,85
2	76	0,68
3	62	0,65
4	52	0,67
5	46	0,67
10	28	0,81

El diseño óptimo se obtiene formando conglomerados de tamaño $\bar{M} = 3$, y seleccionando 62 conglomerados en la muestra.

6.14. De una población formada por 1.000 conglomerados de 50 elementos cada uno, se extrae una muestra de 30 conglomerados en 1.ª etapa y 5 elementos de cada conglomerado en 2.ª etapa, utilizando muestreo con probabilidades iguales y con reemplazamiento en ambas etapas. El análisis de la varianza de la muestra presenta los siguientes resultados:

Fuente de variación	Grados de libertad	Cuadrados medios
Entre conglomerados	29	600
Entre elementos dentro de conglomerados	120	400

Se pide:

a) Estimar la varianza total de la media muestral.

b) Estimar los componentes de la varianza debidos a cada etapa de muestreo.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1988.)

Solución:

Si designamos por x_{ij} el valor del elemento j-ésimo en el conglomerado i-ésimo, n el número de conglomerados en la muestra, y $\bar{n}$ el número de elemen-

tos por conglomerado en la muestra, la ecuación fundamental del análisis de la varianza en la muestra nos determina que:

$$\sum_{i}^{n}\sum_{j}^{\bar{m}}(x_{ij}-\bar{\bar{x}})^2=\sum_{i}^{n}\sum_{j}^{\bar{m}}(x_{ij}-\bar{x}_i)^2+\sum_{i}^{n}\sum_{j}^{\bar{m}}(\bar{x}_i-\bar{\bar{x}})^2$$

donde

$$\bar{x}_i=\frac{\sum_{j}^{\bar{m}}x_{ij}}{\bar{m}} \quad ; \quad \bar{\bar{x}}=\frac{\sum_{i}^{n}\sum_{j}^{\bar{m}}x_{ij}}{n\bar{m}}$$

De los datos del problema obtenemos las diferentes sumas de cuadrados:

$$\sum_{i}\sum_{j}(x_{ij}-\bar{x}_i)^2=120\times 400=48.000$$

$$\sum_{i}\sum_{j}(\bar{x}_i-\bar{\bar{x}})^2=29\times 600=17.400$$

$$\sum_{i}\sum_{j}(x_{ij}-\bar{x}_i)^2=48.000\times 17.400=65.400$$

Aplicando el método de conglomerados últimos para la estimación insesgada de la varianza total, obtenemos:

$$\hat{V}(\bar{\bar{x}})=\frac{\sum_{i}^{n}(\bar{x}_i-\bar{\bar{x}})^2}{n(n-1)}=\frac{\frac{17.400}{5}}{30\times 29}=4$$

Por otra parte, un estimador insesgado de la componente de la varianza debida al submuestreo, o componente «dentro», se obtiene mediante:

$$\hat{w}=\frac{\hat{\sigma}_w^2}{n\bar{m}}$$

donde $\hat{\sigma}_w^2$ es la media de las cuasivarianzas muestrales en cada conglomerado, es decir:

$$\hat{\sigma}_w^2=\frac{\sum_{i}\sum_{j}(x_{ij}-\bar{x}_i)^2}{n(\bar{m}-1)}=\frac{48.000}{30\times 4}=400$$

Por tanto:

$$\hat{w}=\frac{400}{150}=2{,}67$$

Finalmente, por diferencia obtenemos la componente debida al muestreo en 1.ª etapa, o componente «entre»: $\hat{B} = 4 - 2,\ 67 = 1,\ 33$.

6.15. En una población de 100 conglomerados de 40 elementos cada uno, se obtiene una muestra de $n = 6$ conglomerados. Dentro de los conglomerados muestrales se obtiene otra muestra con fracción de muestreo igual $f_2 = 0{,}1$, que proporciona los siguientes valores de una característica X:

$$X_{1j} = \{4\ ;\ 3\ ;\ 2\ ;\ 3\}$$
$$X_{2j} = \{3\ ;\ 4\ ;\ 5\ ;\ 4\}$$
$$X_{3j} = \{2\ ;\ 3\ ;\ 3\ ;\ 4\}$$
$$X_{4j} = \{5\ ;\ 4\ ;\ 5\ ;\ 5\}$$
$$X_{5j} = \{6\ ;\ 3\ ;\ 2\ ;\ 1\}$$
$$X_{6j} = \{2\ ;\ 4\ ;\ 3\ ;\ 3\}$$

Se pide: Calcular el estimador de la media y su error de muestreo (se supone muestreo sin reposición y selección con probabilidades iguales en cada etapa).

Solución:

El estimador insesgado de la media viene dado por:

$$\bar{\bar{x}} = \frac{\sum_{i}^{n} \bar{x}_i}{n}$$

donde

$$\bar{x}_i = \frac{\sum_{j}^{\bar{m}} x_{ij}}{\bar{m}}$$

siendo $\bar{m}$ el tamaño de muestra en cada conglomerado.

De los datos del enunciado, se obtiene:

$$\bar{x}_1 = 3\ \ ;\ \ \bar{x}_2 = 4\ \ ;\ \ \bar{x}_3 = 3\ \ ;\ \ \bar{x}_4 = 5\ \ ;\ \ \bar{x}_5 = 3\ \ ;\ \ \bar{x}_6 = 3\ \ ;\ \ \bar{m} = 4$$

Por tanto:

$$\bar{\bar{x}} = \frac{21}{6} = 3{,}5$$

La estimación de la varianza de este estimador se obtiene, mediante:

$$\hat{V}(\bar{\bar{x}}) = (1 - f_1)\frac{\sum^{n}(\bar{x}_i - \bar{\bar{x}})^2}{n(n-1)} + f_1(1 - f_2)\frac{\sum_i^n \sum_j^{\bar{m}}(x_{ij} - \bar{x}_i)^2}{n^2\bar{m}(\bar{m}-1)}$$

Con los datos del enunciado procederemos a realizar los siguientes cálculos:

Conglomerado	$\bar{x}_i$	$(x_i - \bar{\bar{x}})^2$	$\sum_j (x_{ij} - \bar{x}_i)^2$
1	3	0,25	2
2	4	0,25	2
3	3	0,25	2
4	5	2,25	2
5	3	0,25	14
6	3	0,25	2
		$\sum_i = 3,5$	$\sum_i = 24$

Por tanto:

$$\hat{V}(\bar{\bar{x}}) = \left(1 - \frac{6}{100}\right)\frac{3,5}{30} + \frac{6}{100}(1 - 0,1)\frac{24}{36 \times 12} = 0,1097 + 0,003 = 0,1127$$

Estimándose el error de muestreo

$$\hat{\sigma}_{\bar{\bar{x}}} = \sqrt{0,1127} = 0,3357$$

6.16. En una población se obtiene una muestra de $n = 6$ conglomerados de $\bar{M} = 30$ unidades elementales, con probabilidades iguales. Dentro de los conglomerados de la muestra se realiza submuestreo sin reposición con fracción de muestreo $f_2 = \frac{1}{6}$, obteniéndose los siguientes valores de elementos que poseen una determinada característica A:

$$A_i\{4, 3, 5, 2, 1, 5\}$$

Se pide:

a) Estimador de la proporción y su error de muestreo, en caso de muestreo con reposición de unidades primarias.

b) Estimador de la proporción y su error de muestreo, en caso de muestreo sin reposición de unidades primarias con fracción de muestreo $f_1 = \frac{1}{2}$.

Solución:

a) Un estimador insesgado de la proporción es:

$$\hat{P} = \frac{1}{n\bar{m}} \sum_{i}^{n} A_i$$

siendo $\bar{m}$ el número de elementos de la muestra por conglomerado.

Con los valores:

$$n = 6 \quad , \quad \bar{m} = 5 \quad , \quad \sum_{i}^{n} A_i = 20$$

obtenemos:

$$\hat{P} = \frac{2}{3}$$

En caso de muestreo con reposición de unidades primarias, un estimador insesgado de la varianza total puede obtenerse por el método de conglomerados últimos, resultando:

$$\hat{V}(\hat{P}) = \frac{1}{n(n-1)\bar{m}^2} \sum_{i}^{n} (A_i - \bar{A})^2 = \frac{1}{n(n-1)\bar{m}^2} \left\{ \sum_{i}^{n} A_i^2 - n\bar{A}^2 \right\}$$

con los valores:

$$\sum_{i}^{n} A_i^2 = 80 \quad , \quad \bar{A} = \frac{\sum_{i}^{n} A_i}{n} = \frac{10}{3}$$

Se obtiene

$$\hat{V}(\hat{P}) = \frac{1}{30 \times 25} \left(80 - \frac{200}{3} \right) = 0{,}018$$

y como error de muestreo:

$$\hat{\sigma}_{\hat{P}} = \sqrt{0{,}018} = 0{,}13$$

b) El estimador insesgado es idéntico al caso *a*):

$$\hat{P} = \frac{2}{3}$$

La estimación insesgada de la varianza la obtendremos a través de la estimación de dos componentes:

$$\hat{V}(\hat{P}) = (1 - f_1)\frac{\sum_i^n (\hat{P}_i - \hat{\bar{P}})^2}{n(n-1)} + f_1(1 - f_2)\frac{\sum_i^n \hat{P}_i(1 - \hat{P}_i)}{n^2(\bar{m} - 1)} = B + W$$

siendo $f_1 = \frac{1}{2}$; $f_2 = \frac{1}{6}$;

$$\sum_i^n (\hat{P}_i - \hat{\bar{P}})^2 = \frac{1}{\bar{m}^2}\left(\sum_i^n A_i^2 - n\bar{A}^2\right) = \frac{1}{25}\left(80 - \frac{200}{3}\right) = \frac{8}{15}$$

$$\sum_i^n \hat{P}_i(1 - \hat{P}_i) = \frac{1}{\bar{m}}\sum_i^n A_i - \frac{1}{\bar{m}^2}\sum_i^n A_i^2 = \frac{4}{5}$$

resulta:

$$B = \frac{1}{2} \times \frac{1}{30} \times \frac{8}{15} = 0{,}0089 \quad ; \quad W = \frac{1}{2} \times \frac{5}{6} \times \frac{1}{144} \times \frac{4}{5} = 0{,}0023$$

Por tanto:

$$\hat{V}(\hat{P}) = 0{,}0089 + 0{,}0023 = 0{,}0112$$

y el error de muestreo:

$$\hat{\sigma}_{\hat{P}} = \sqrt{0{,}0112} = 0{,}106$$

ligeramente inferior al caso de muestreo con reposición.

6.17. En una población de $N = 10$ conglomerados de tamaños desiguales (M_i) se realiza muestreo en dos etapas. En la primera se toman tres unidades primarias y en la segunda cinco unidades en cada unidad primaria. Se pide formar el estimador lineal insesgado del total X en el caso de muestreo sin reposición con probabilidades iguales en las dos etapas. Probar que si se aplica el teorema de Durbin para la estimación de la varianza del estimador, se obtiene:

$$\hat{V}(\hat{X}) = \frac{14}{45}\sum_{i=1}^{3} M_i^2 x_i^2 - \frac{7}{45}\sum_{i\neq j} M_i M_j x_i x_j + \frac{2}{3}\sum_{i=1}^{3} M_i(M_i - 5)s_i^2$$

siendo x_i el total muestral y s_i^2 la cuasivarianza dentro de la unidad primaria i[a] de la muestra.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1986.)

Solución:

En un muestreo bietápico con probabilidades iguales el estimador insesgado del total adopta la expresión general:

$$\hat{X} = \frac{N}{n}\sum_{i=1}^{3} \frac{M_i x_i}{m_i} \qquad (1)$$

si

$$N = 10 \quad , \quad n = 3 \quad \text{y} \quad m_i = 5$$

resulta

$$\hat{X} = \frac{2}{3}\sum_{i=1}^{3} M_i x_i$$

La regla de Durbin dice que el estimador de la varianza es la suma de dos partes: la 1.[a] es igual al estimador de la varianza calculado como si el muestreo se hubiese realizado en una etapa (copia del estimador del muestreo en una etapa sin más que sustituir X_i por su estimación $\hat{X}_i$). La 2.[a] es igual al estimador de la varianza calculado como si las unidades seleccionadas en 1.[a] etapa fueran estratos, multiplicando la contribución a la varianza de cada unidad primaria seleccionada por la probabilidad que tiene dicha unidad de pertenecer a la muestra.

Aplicando la regla anterior al estimador general (1), resulta:

$$\hat{V}(\hat{X})_{1.^a\,\text{parte}} = N^2\left(1-\frac{n}{N}\right)\frac{\sum_{i=1}^{n}(\hat{X}_i-\hat{\bar{X}})^2}{n(n-1)} = 100\left(1-\frac{3}{10}\right)\frac{1}{6}\sum_{i=1}^{3}(\hat{X}_i-\hat{\bar{X}})^2 =$$

$$= \frac{35}{3}\sum_{i=1}^{3}(\hat{X}_i-\hat{\bar{X}})^2$$

Ahora bien, siendo:

$$\hat{X}_i = \frac{M_i}{m_i}x_i = \frac{1}{5}M_i x_i \quad \text{y} \quad \hat{\bar{X}} = \frac{\sum_{i=1}^{3}\hat{X}_i}{3} = \frac{1}{15}\sum_{i}^{3}M_i x_i$$

resulta:

$$\sum_{i}^{3}(\hat{X}_i-\hat{\bar{X}})^2 = \sum_{i}^{3}\hat{X}_i^2 - 3\hat{\bar{X}}^2 = \sum_{i}^{3}\left(\frac{M_i x_i}{5}\right)^2 - 3\left(\frac{\sum_{i}^{3}M_i x_i}{15}\right)^2 =$$

$$= \frac{1}{25}\sum_{i}^{3}M_i^2x_i^2 - \frac{1}{75}\left(\sum_{i}^{3}M_i^2x_i^2 + \sum_{i\neq j}M_iM_jx_ix_j\right) =$$

$$= \frac{2}{75}\sum_{i}M_i^2x_i^2 - \frac{1}{75}\sum_{i\neq j}M_iM_jx_ix_j$$

Por tanto:

$$\hat{V}(\hat{X})_{1.^a\,\text{parte}} = \frac{35}{3}\left(\frac{2}{75}\sum_{i}M_i^2x_i^2 - \frac{1}{75}\sum_{i\neq j}M_iM_jx_ix_j\right) =$$

$$= \frac{14}{45}\sum_{i=1}^{3}M_i^2x_i^2 - \frac{7}{45}\sum_{i\neq j}M_iM_jx_ix_j$$

Por otra parte, el estimador (1) es equivalente a:

$$\hat{\bar{X}} = \frac{N}{n}\sum_{i}^{3}\hat{X}_i$$

por lo que aplicando la regla de Durbin, resulta:

$$\hat{V}(\hat{\bar{X}})_{2.^a\,\text{parte}} = \frac{N^2}{n^2}\sum_{i=1}^{3}\hat{V}(\hat{X}_i)\pi_i$$

donde

$$\pi_i \text{ (probabilidad de } u_i \text{ de pertenecer a la muestra)} = \frac{n}{N} = \frac{3}{10}$$

y

$$\hat{V}(\hat{X}_i) = M_i(M_i - m_i)\frac{s_i^2}{m_i} = M_i(M_i - 5)\frac{s_i^2}{5}$$

Por tanto:

$$\hat{V}(\hat{X})_{2.^a \text{parte}} = \frac{100}{9}\sum_i^3 M_i(M_i - 5)\frac{s_i^2}{5} \times \frac{3}{10} = \frac{2}{3}\sum_{i=1}^3 M_i(M_i - 5)s_i^2$$

6.18. En una población de $N = 1.000$ conglomerados de tamaño $\bar{M} = 50$, se obtiene una muestra de $n = 5$ unidades de primera etapa, de las que se obtiene, a su vez, una submuestra de $\bar{m} = 6$ unidades elementales que proporcionan los datos siguientes:

$$\bar{x}_i = \{3 \ ; \ 5 \ ; \ 2 \ ; \ 4 \ ; \ 6\}$$

El muestreo se realiza con reposición y probabilidades iguales en ambas etapas.

Se pide:

a) Estimar la varianza de la media utilizando el método de conglomerados últimos.

b) Comparar este resultado con el que se hubiera obtenido con la fórmula del muestreo bietápico sin reposición, suponiendo que $\hat{S}_w^2 = 1{,}5$.

Solución:

Un estimador insesgado de la media es:

$$\bar{\bar{x}} = \frac{1}{n}\sum_i \bar{x}_i = \frac{20}{5} = 4$$

siendo, en caso de muestreo con reemplazamiento, un estimador insesgado de la varianza total el estimador de últimos conglomerados:

$$\hat{V}(\bar{\bar{x}}) = \frac{\hat{\sigma}^2_{\bar{x}i}}{n} = \frac{\sum^{n} (\bar{x}_i - \bar{\bar{x}})^2}{n(n-1)} = \frac{10}{20} = 0{,}5$$

En caso de muestreo sin reposición, el estimador insesgado de la varianza es:

$$\hat{V}(\bar{\bar{x}}) = (1 - f_1)\frac{\hat{S}^2_b}{n\bar{m}} + f_1(1 - f_2)\frac{\hat{S}^2_w}{n\bar{m}}$$

donde

$$\hat{S}^2_b = \frac{\bar{m}\sum^{n}_{i} (\bar{x}_i - \bar{\bar{x}})^2}{n-1} = 15 \quad \text{y} \quad \hat{S}^2_w = \frac{\sum^{n}_{i}\sum^{\bar{m}}_{j} (x_{ij} - \bar{x}_i)^2}{n(\bar{m}-1)} = 1{,}5$$

Por tanto:

$$\hat{V}(\bar{\bar{x}}) = \left(1 - \frac{5}{1.000}\right)0{,}5 + \frac{5}{1.000}(1 - 0{,}1)\frac{1{,}5}{30} = 0{,}4975 + 0{,}000225 \simeq 0{,}5$$

Obteniéndose, prácticamente, la misma varianza que con el método de conglomerados últimos.

6.19. Si $\bar{m}/\bar{M}$ y n/N son pequeños y la función de coste es lineal, demostrar que $\bar{m} = 2$ da un valor más pequeño de $V(\bar{\bar{x}})$ que $\bar{m} = 1$ si $c_1/c_2 > 2S_1^2/S_2^2$.

$$V(\bar{\bar{x}}) = \frac{S_1^2}{n} + \frac{S_2^2}{n\bar{m}} \qquad C = c_1 n + c_2 n\bar{m}$$

(W. G. Cochran, 1977, p. 291.)

Solución:

Los valores de n y $\bar{m}$ que minimizan $V(\bar{\bar{x}})$ se obtendrían derivando la función de Lagrange

$$F = \frac{S_1^2}{n} + \frac{S_2^2}{n\bar{m}} + \lambda \times (c_1 n + c_2 n\bar{m} - C)$$

y resolviendo el sistema de ecuaciones

$$\frac{\partial F}{\partial n} = 0 \qquad \frac{\partial F}{\partial \bar{m}} = 0$$

Puede demostrarse que

$$\bar{m}_{op}^2 = \frac{c_1 S_2^2}{c_2 S_1^2}$$

(ver, por ejemplo, Azorín y Sánchez-Crespo, 1986).

Si sustituimos en esta expresión los valores $\bar{m} = 1$ y $\bar{m} = 2$, tendremos:

$$\frac{c_1 S_2^2}{c_2 S_1^2} = 4 \quad \text{y} \quad \frac{c_1 S_2^2}{c_2 S_1^2} = 1$$

luego si

$$\frac{c_1 S_2^2}{c_2 S_1^2} > 2$$

$\bar{m}$ óptimo ha de ser mayor que $\sqrt{2}$ lo que se cumple para $\bar{m} = 2$ y no para $\bar{m} = 1$.

6.20. Una población está formada por M unidades elementales agrupadas en 50 conglomerados de tamaños desiguales M_i, ($i = 1, 2, ..., N$). El valor de $M = \sum_i M_i$ es conocido e igual a 1.000. Con objeto de estimar la proporción de unidades elementales que pertenecen a una cierta clase se decide utilizar un muestreo de conglomerados con submuestreo. En ambas etapas se emplea un procedimiento de selección con probabilidades iguales sin reposición.

En la 1.ª etapa se obtienen 5 conglomerados muestrales con los siguientes valores de M_i: 6; 10; 8; 20; 60. En la segunda etapa realizada con fracción de muestreo $f_{2i} = 4/M_i$, se obtienen en los cinco conglomerados de la muestra los valores 1; 3; 2; 2; 3 para el número de elementos que pertenecen a la clase. Se pide:

a) Una estimación insesgada de la mencionada proporción y su error de muestreo.

b) La estimación y el correspondiente error de muestreo utilizando el estimador de la razón al tamaño.

c) Comentar las ventajas e inconvenientes del estimador *b*) respecto al *a*).

Solución:

a) La probabilidad que tiene la unidad elemental *j* del conglomerado *i*, de pertenecer a la muestra, es:

$$P(u_{ij}) = \frac{n}{N} \times \frac{4}{M_i} = \frac{4}{10M_i}$$

Por tanto, una estimación insesgada del total de clase es:

$$\hat{A} = \sum_i^n \frac{10M_i}{4} \times x_i = 10 \sum_i \hat{P}_i M_i$$

siendo

$$\hat{P}_i = \frac{x_i}{m_i} = \frac{x_i}{4}$$

y como estimación insesgada de la proporción, tendremos:

$$\hat{P} = \frac{\hat{A}}{M} = \frac{10}{1.000} \sum_i \hat{P}_i M_i =$$

$$= \frac{10}{1.000}\left(6 \times \frac{1}{4} + 10 \times \frac{3}{4} + 8 \times \frac{2}{4} + 20 \times \frac{2}{4} + 60 \times \frac{3}{4}\right) = 0{,}68$$

La varianza de $\hat{P}$ se obtiene mediante la expresión

$$V(\hat{P}) = \frac{1-f_1}{n\bar{M}^2} \frac{\sum_i^N (M_i P_i - \bar{M}P)^2}{N-1} + \frac{1}{nN\bar{M}^2} \frac{\sum_i^N M_i^2 (M_i - m_i) P_i (1-P_i)}{m_i (M_i - 1)} \qquad (1)$$

Una estimación insesgada de (1) obtenida a partir de la muestra, es:

$$\hat{V}(\hat{P}) = \frac{1-f_1}{n\bar{M}^2} \sum_i^n \left(M_i \hat{P}_i - \frac{\sum_i^n M_i \hat{P}_i}{n}\right)^2 \frac{1}{n-1} +$$

$$+ \frac{1}{nN\bar{M}^2} \frac{\sum_i^n M_i (M_i - m_i)}{m_i - 1} \hat{P}_i (1 - \hat{P}_i)$$

Los cálculos necesarios se presentan en el siguiente cuadro:

M_i	$\hat{P}_i$	$M_i\hat{P}_i$	$(M_i\hat{P}_i)^2$	$A=M_i\hat{P}_i(1-\hat{P}_i)$	$B=\frac{(M_i-m_i)}{(m_i-1)}$	$A\times B$
6	$\frac{1}{4}$	$\frac{6}{4}$	$\frac{36}{16}$	$\frac{18}{16}$	$\frac{2}{3}$	$\frac{36}{48}$
10	$\frac{3}{4}$	$\frac{30}{4}$	$\frac{900}{16}$	$\frac{30}{16}$	$\frac{6}{3}$	$\frac{180}{48}$
8	$\frac{2}{4}$	$\frac{16}{4}$	$\frac{256}{16}$	$\frac{32}{16}$	$\frac{4}{3}$	$\frac{128}{48}$
20	$\frac{2}{4}$	$\frac{40}{4}$	$\frac{1.600}{16}$	$\frac{80}{16}$	$\frac{16}{3}$	$\frac{1.280}{48}$
60	$\frac{3}{4}$	$\frac{180}{4}$	$\frac{32.400}{16}$	$\frac{180}{16}$	$\frac{56}{3}$	$\frac{10.080}{48}$
104		68	$\frac{35.192}{16}$			$\frac{11.704}{48}$

$$\hat{V}(\hat{P}) = \frac{0{,}9}{5 \times 20^2}\left[\frac{35.192}{16} - 5 \times \left(\frac{68}{5}\right)^2\right]\frac{1}{4} + \frac{1}{5 \times 50 \times 20^2} \times \frac{11.704}{48} =$$

$$= 0{,}1434 + 0{,}0024 = 0{,}1458$$

y el error de muestreo:

$$\hat{\sigma}_{\hat{P}} = \sqrt{0{,}1458} = 0{,}38$$

b) El estimador de razón se obtiene mediante la expresión:

$$\hat{P}_R = \frac{\hat{A}}{\hat{M}} = \frac{\frac{N}{n}\sum_i^n M_i\hat{P}_i}{\frac{N}{n}\sum_i^n M_i} = \frac{\sum_i^n M_i\hat{P}_i}{\sum_i^n M_i} = \frac{68}{104} = 0{,}65$$

Por ser estimador de razón no existe una expresión exacta de la varianza; la varianza aproximada es:

$$V(\hat{P}_R) \simeq \frac{1-f_1}{n\bar{M}^2}\frac{\sum_i^N M_i^2(P_i-P)^2}{N-1} + \frac{1}{nN\bar{M}^2}\sum_i^N \frac{M_i^2(M_i-m_i)}{M_i(M_i-1)}P_i(1-P_i)$$

que se estima mediante:

$$\hat{V}(\hat{P}_R) = \frac{1-f_1}{n\bar{M}^2}\frac{\sum_i^n M_i^2(\hat{P}_i-\hat{P}_R)^2}{n-1} + \frac{1}{nN\bar{M}^2}\sum_i^n \frac{M_i(M_i-m_i)}{m_i-1}\hat{P}_i(1-\hat{P}_i)$$

Puesto que la 2.ª componente es idéntica al caso *a*), realicemos los cálculos para obtener la 1.ª componente:

M_i	$\hat{P}_i$	$(\hat{P}_i - \hat{P}_R)$	$[M_i(\hat{P}_i - \hat{P}_R)]^2$
6	$\frac{1}{4}$	0,4	5,76
10	$\frac{3}{4}$	0,1	1
8	$\frac{2}{4}$	0,15	1,44
20	$\frac{2}{4}$	0,15	9
60	$\frac{3}{4}$	0,1	36
			53,2

Por tanto:

$$\hat{V}(\hat{P}_R) = \frac{0,9}{5 \times 20^2} \times \frac{53,2}{4} + 0,0024 = 0,006 + 0,0024 = 0,0084$$

siendo el error de muestreo:

$$\hat{\sigma}_{\hat{P}_R} = \sqrt{0,0084} = 0,09$$

c) El principal inconveniente del estimador *b*) es el sesgo, inherente a todo estimador de razón, aún cuando puede despreciarse si n es grande.

A pesar de lo anterior, el estimador *b*) presenta notables ventajas. En general, será más eficiente, en cuanto permite corregir los desequilibrios de la muestra de unidades primarias. Así, por ejemplo, si en la selección de la muestra hemos tenido la mala suerte de seleccionar muchos conglomerados pequeños, la estimación del total de clase será una infraestimación del valor verdadero y, por tanto, la estimación insesgada $\hat{P}$ proveniente de esa muestra resultará notablemente infraestimada. El estimador de razón corrige esa infraestimación.

El estimador insesgado será, en general, mucho más ineficiente que el estimador de razón al ser muy sensible a valores de las M_i, como puede observarse en el supuesto práctico.

Una segunda ventaja del estimador de razón $\hat{P}_R$ es que puede aplicarse aunque no se conozca el total M de unidades elementales en la población, lo que no es posible con el estimador insesgado.

CAPITULO VII
Otras técnicas de muestreo

7.1. Se dispone de una lista de 14 viviendas. De ellas, las cuatro primeras han tenido protección oficial.

El número de personas por vivienda es: 2, 4, 3, 5, 2, 1, 6, 4, 3, 5, 2, 3, 4, 1.

Con muestras de tamaño igual a 15 personas se pide calcular la varianza de la proporción estimada, de personas que habitan en viviendas protegidas, utilizando:

a) Muestreo aleatorio simple de personas.

b) Muestreo sistemático, formando las 3 muestras posibles de 15 personas.

Solución:

a) Para obtener la varianza del estimador $\hat{P}$ de la proporción P, en el muestreo aleatorio simple (sin reposición y probabilidades iguales) puede utilizarse la expresión:

$$V(\hat{P}) = \left[\frac{(N-n)}{(N-1)}\right] \times \frac{P(1-P)}{n}$$

donde en nuestro caso

$$N = 45 \quad , \quad n = 15 \quad , \quad P = \frac{14}{45} = 0{,}31 \quad , \quad 1 - P = 0{,}69$$

y, por lo tanto:

$$V(\hat{P}) = 0{,}0097$$

b) Las tres muestras posibles de quince personas son:

s_1	s_2	s_3
1	2	3
4	5	6
7	8	9
10	11	12
13	14	15
16	17	18
19	20	21
22	23	24
25	26	27
28	29	30
31	32	33
34	35	36
37	39	39
40	41	42
43	44	45

y los valores de las proporciones muestrales $\hat{P}_j\{j = 1, 2, 3\}$:

$$\hat{P}_1 = \frac{5}{15}$$

$$\hat{P}_2 = \frac{5}{15}$$

$$\hat{P}_3 = \frac{4}{15}$$

Las muestras sistemáticas están formadas por las personas a las que corresponden los respectivos números de orden.

La varianza de la proporción $\hat{P}_j$ (correspondiente a la muestra sistemática j) viene dada por:

$$V(\hat{P}_j) = \frac{1}{k} \sum_{j}^{k} (\hat{P}_j - P)^2 = 0{,}001$$

siendo $k = 3$ el número de muestras posibles.

7.2. Dada la población

u_i	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
X_i	1	3	5	2	4	6	2	7

se obtienen las dos muestras sistemáticas posibles de 4 unidades.

Se pide:

a) Calcular $V(\bar{x}_j)$.

b) Estimar $V(\bar{x}_j)$ con cada muestra utilizando $\hat{V}(\bar{x})$.

c) Estimar $V(\bar{x}_{st})$ considerando las dos primeras unidades de $S_1(x)$ como procedentes del estrato $(u_1\, u_2\, u_3\, u_4)$ y las dos restantes como obtenidas de un segundo estrato $(u_5\, u_6\, u_7\, u_8)$.

Solución:

a) La varianza de la media viene dada por:

$$V(\bar{x}_j) = \frac{1}{2}[(\bar{x}_1 - \bar{x})^2 + (\bar{x}_2 - \bar{x})^2] = \frac{1}{2}[(3 - 3{,}75)^2 + (4{,}5 - 3{,}75)^2] = 0{,}5625$$

$$\text{por ser } \bar{X} = \frac{1}{8}\sum_{i}^{N} X_i = 3{,}75$$

la media poblacional y

$$\bar{x}_1 = \frac{(1 + 5 + 4 + 2)}{4} = 3 \quad , \quad \bar{x}_2 = \frac{3 + 2 + 6 + 7}{4} = 4{,}5$$

sus estimaciones basadas en cada muestra sistemática posible.

b) Utilizando como estimador la varianza correspondiente al muestreo aleatorio simple:

$$\hat{V}_1(\bar{x}) = \frac{(1 - f)\hat{S}_1^2}{n} = \frac{0{,}5 \times 3{,}33}{4} = 0{,}42$$

$$\hat{V}_2(\bar{x}) = \frac{0{,}5 \times 5{,}67}{4} = 0{,}71$$

serían las estimaciones de $V(\bar{x}_j)$ con cada muestra sistemática.

c) $W_1 = W_2 = \frac{1}{2}$ serían las ponderaciones de los estratos. Las varianzas estimadas para las muestras (1; 5) y (4; 2) serían:

$$\hat{S}_1^2 = 8 \quad \text{y} \quad \hat{S}_2^2 = 2$$

Sustituyendo estos valores en

$$\hat{V}(\bar{x}_{st}) = \sum_h W_h^2 \times \hat{S}_h^2 \times \frac{1 - f_h}{n_h}$$

se obtiene

$$\hat{V}(\hat{\bar{x}}_{st}) = \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{2}\right) \times \left(\frac{1}{2}\right) \times 10 = 0{,}625$$

7.3. La población: $u_i (i = 1; 2; ...; 8)$ con valores $X_i (1; 3; 5; 2; 4; 6; 2; 7)$ se considera dividida en n zonas (estratos) de k unidades consecutivas. Si $n = 4$ es el número de zonas, será $k = 2$ el número de unidades por zona ya que $N/n = k$. El muestreo sistemático puede considerarse como un muestreo estratificado con una unidad por estrato o zona, con la particularidad de que la selección en una zona no es independiente de la realizada en otra.

Se pide:

a) Calcular el cuadrado medio correspondiente a la variación «dentro» de zonas.

b) El coeficiente de correlación entre las desviaciones a las medias de las zonas, para todos los pares de unidades pertenecientes a la misma muestra sistemática.

c) La varianza $V(\bar{x}_j)$ en función de los valores obtenidos en los apartados anteriorexs.

Solución:

a) El cuadrado medio «dentro» de zonas viene dado por

$$S_{wst}^2 = \frac{\sum_i^n \sum_j^k (X_{ij} - \bar{X}_i)^2}{n(k-1)} = \frac{2(1^2 + 1{,}5^2 + 1^2 + 2{,}5^2)}{4} = 5{,}25$$

b) El coeficiente de correlación pedido es:

$$\rho_{wst} = \frac{E_{ij}(X_{ij} - \bar{X}_i)(X_{1j} - \bar{X}_1)}{E_{1j}(X_{ij} - \bar{X}_i)} = \frac{(-0{,}125)}{2{,}625} = -0{,}05$$

c) La varianza es:

$$V(\bar{x}_j) = (N - n) \times S^2_{wst} \times \frac{[1 - (n - 1) \times \rho_{wst}]}{N \times n} = 0{,}5625$$

varianza idéntica a la que se obtuvo directamente en el ejercicio anterior.

7.4. En una población de N unidades se pueden formar dos estratos de aproximadamente el mismo peso ($W_h = N_h/N$). El presupuesto para realizar una encuesta es de un millón de pesetas, siendo el coste por unidad de unas 500 pts., y el coste unitario para obtener información de la variable auxiliar que permitiría realizar la estratificación es de 50 ptas.

Si las proporciones, de interés para la encuesta, en los estratos fuesen $P_1 = 0{,}4$ y $P_2 = 0{,}6$: ¿Cuáles serían los valores óptimos de n y n'? ¿Cuál sería la varianza obtenida con una muestra de $n = 3.000$ unidades y afijación proporcional?

Se propone utilizar la función de coste:

$$1.000^2 = 500n + 50n'$$

y la fórmula aproximada

$$V(\hat{P}'_{st}) = \frac{\sum_h^L W_h P_h(1 - P_h)}{n' v_h} + \frac{\sum_h^L W_h(P_h - P)^2}{n'}$$

siendo

$$v_h = \frac{n}{n'} = v_1 = v_2$$

n' y n son los tamaños muestrales en la primera y segunda fase, respectivamente.

Solución:

La varianza de $\hat{P}'_{st}$ sería:

$$V(\hat{P}'_{st}) = \frac{0{,}24}{n} + \frac{0{,}01}{n'}$$

y al introducir la condición que imponen los costes:

$$\phi = \frac{0{,}24}{n} + \frac{0{,}01}{n'} + \lambda(500n + 50n' - 1.000^2)$$

Derivando respecto de n y n', tenemos:

$$\frac{\partial \phi}{\partial n} = -\frac{0{,}24}{n^2} + \lambda \times 500 = 0$$

$$\frac{\partial \phi}{\partial n'} = -\frac{0{,}01}{n'^2} + \lambda \times 50 = 0$$

$$\frac{\partial \phi}{\partial \lambda} = 500n + 50n' - 1.000^2 = 0$$

de donde se deduce:

$$\frac{n'^2}{n^2} = 0{,}4167 \qquad \frac{n'}{n} = 0{,}6455$$

$$n' = 1.213$$

$$n = 1.879$$

$$V(\hat{P}'_{st}) = 0{,}0001359$$

En el caso de $n = 3.000$ y afijación proporcional tendríamos:

$$V(\hat{P}_P) = \left(\frac{1}{2}\right)^2 \times \frac{0{,}4 \times 0{,}6 \times 2}{1.500} = 0{,}00008$$

7.5. En una población de $N = 1.000$ personas, la experiencia de una encuesta piloto proporciona los siguientes datos:

$$S^2 = S_2^2 = 0{,}001 \quad ; \quad P_1 = 0{,}8$$

$$C_0 = 50 \quad ; \quad C_1 = 300 \quad ; \quad C_2 = 1.000$$

Siendo C_o, C_1 y C_2, respectivamente, los costes unitarios de envío de cuestionarios por correo, proceso de cuestionarios obtenidos por correo, y entrevistas en intentos posteriores. P_1 es la proporción de los que contestan por correo y S_2^2 la cuasivarianza del estrato no-respuesta.

Se pide: El número óptimo n, de cuestionarios a enviar por correo, y la fracción de muestreo $f_{21} = n_{21}/n_2$, en el estrato de los que no contestaron, para obtener una varianza $V(\hat{X}) = 10$.

Solución:

Las fórmulas siguientes pueden verse, por ejemplo, en Azorín y Sánchez-Crespo (1986), páginas 241-242.

$$f_{21} = \sqrt{\frac{P_1 S_2^2}{S^2 - P_2 S_2^2} \times \frac{C_1 + \dfrac{C_0}{P_1}}{C_2}} \qquad n' = \frac{N^2 S^2}{V_0(\hat{X}) + N^2 S^2}$$

$$n = n' \times \left[1 + \left(\frac{1}{f_{21}} - 1\right) P_2 - \frac{S_2^2}{S^2}\right]$$

siendo n' el tamaño de la muestra necesario para obtener una varianza $V_0(\hat{X})$ en el supuesto de no existir falta de respuesta. Es decir:

$$V_0(\hat{X}) = N^2 \times \left(1 - \frac{n'}{N}\right) \times \frac{S^2}{n'}$$

sustituyendo la información obtenida en la encuesta piloto, se obtiene:

$$f_{21} = \sqrt{\frac{0{,}8 \times 0{,}001}{0{,}^{\circ}001 \times 0{,}8} \times \frac{300 + \dfrac{5}{0{,}8}}{1.000}} = \sqrt{0{,}3625} = 0{,}6021$$

$$n' = \frac{1.000^2 \times 0{,}001}{10 + 1.000 \times 0{,}001} \doteqdot 91$$

$$n = 91 \times \left[1 + \left(\frac{1}{0{,}6021} - 1\right) \times 0{,}2\right] = 103$$

7.6. Para estimar una media se realizó una encuesta por correo seleccionando, con reposición, una muestra de n unidades elementales con probabilidades iguales.

De las $n_2 = n - n_1$ unidades que no contestaron se eligió, siguiendo el mismo procedimiento, una muestra con fracción de muestreo $f_{21} = n_{21}/n_2$ de las que se obtuvo la información mediante entrevista.

Un estudio piloto previo proporcionó la información siguiente:

a) El cociente de la varianza para las unidades que no respondieron a la varianza total σ^2 fue 4,65.

b) El coste unitario de envío por correo fue 25 pts., el de proceso para cada una de las n_1 unidades que respondieron fue 100 pts., y el de cada unidad entrevistada 540 pts.

c) La proporción de falta de respuesta al envío por correo fue $P_2 = 0{,}35$.

Calcular los valores óptimos de n y f_{21} que minimizarían el coste total de la encuesta, de tal forma que el error de muestreo del estimador de la media fuese igual al que se obtendría con una muestra de 100 unidades elegida por el mismo procedimiento en el supuesto de no existir falta de respuesta.

Solución:

La varianza de la media en el muestreo con reposición y probabilidades iguales es:

$$V(\bar{x}) = \frac{\sigma^2}{n} + \frac{1}{nf_{21}} \times P_2\sigma_2^2$$

y al ser $\sigma_2^2 = 4{,}65\sigma^2$, tenemos:

$$V(\bar{x}) = \frac{\sigma^2}{n} \times \left[1 + \frac{4{,}65P_2}{f_{21}}\right]$$

que igualado a

$$V_0(\bar{x}) = \frac{\sigma^2}{100}$$

nos da:

$$n = 100 \times \left[1 + \frac{4{,}65P_2}{f_{21}}\right] \qquad (1)$$

Una estimación insesgada de la varianza de $\hat{P}_{st}$ puede obtenerse mediante la siguiente expresión debida a J. N. K. Rao:

$$\hat{V}(\hat{P}_{st}) = \frac{N-1}{N} \sum \left(\frac{n'_h - 1}{n' - 1} - \frac{n_h - 1}{N - 1} \right) \frac{\hat{W}_h \hat{P}_h (1 - \hat{P}_h)}{n_h - 1} +$$

para optimizar tendremos:

$$\phi = n \times (C_0 + C_1 P_1 + C_2 P_2 f_{21}) + \lambda \left[\frac{\sigma^2}{n} \left[1 + \frac{4{,}65 P_2}{f_{21}} \right] - V_0(\bar{x}) \right]$$

de donde

$$\frac{\partial \phi}{\partial n} = C_0 + C_1 P_1 + C_2 P_2 f_{21} - \frac{\lambda \sigma^2}{n^2} \left[1 + \frac{4{,}65 P_2}{f_{21}} \right]$$

de donde

$$\frac{\lambda \sigma^2}{n^2} = f_{21} \times \frac{C_0 + C_1 P_1 + C_2 P_2 f_{21}}{f_{21} + 4{,}65 P_2} \qquad (2)$$

$$\frac{\partial \phi}{\partial f_{21}} = n C_2 P_2 - \frac{\lambda \sigma^2}{n} \times \frac{4{,}65}{f_{21}^2} P_2 = 0$$

de donde

$$\frac{\lambda \sigma^2}{n^2} = \frac{C_2 f_{21}^2}{4{,}65} \qquad (3)$$

Eliminando $\lambda \sigma^2 / n$ entre (2) y (3) y siendo $C_0 = 25$; $C_1 = 100$; $C_2 = 540$ se obtiene

$$418{,}5 = 540 f_{21}^2 \qquad f_{21} = 0{,}88$$

y sustituyendo en (1)

$$n = 100 \times [1 + 1{,}849] \div 285$$

7.7. Para estimar la proporción de viviendas en régimen de propiedad o en régimen de alquiler se decide utilizar un muestreo bifásico. En la primera fase las viviendas se estratifican de acuerdo con su época de construcción, anterior o posterior al primero de enero de 1935, utilizando una muestra aleatoria simple de $n' = 1.000$ viviendas de las que 380 resultaron construidas con anterioridad a la mencionada fecha y 620 lo fueron con posterioridad.

En una segunda fase se obtuvo una submuestra aleatoria simple con fracción

Sea $\hat{D} = \hat{P}_2 - \hat{P}_1$ un estimador de la diferencia de proporciones con contestación afirmativa, entre la segunda y la primera ocasión.

Se pide estimar:

1.° Error de muestreo de $\hat{D}$.

2.° Error de muestreo de $\hat{D}$ si las muestras hubiesen sido independientes.

3.° Coeficiente de correlación ϱ_{12}.

Solución:

La varianza estimada de $\hat{D}$ (cuadrado del error de muestreo) viene dada por la expresión:

$$\hat{V}(\hat{D}) = \hat{V}(\hat{P}_1) + \hat{V}(\hat{P}_2) - 2\,\hat{\text{Cov}}\,(\hat{P}_1; \hat{P}_2)$$

donde

$$\hat{V}(\hat{P}_1) = (1 - f) \times \frac{\hat{P}_1(1 - \hat{P}_1)}{(n - 1)}$$

y análogamente para $\hat{V}(\hat{P}_2)$. La covarianza estimada es:

$$\hat{\text{Cov}}\,(\hat{P}_1; \hat{P}_2) = (1 - f) \times \frac{\left(\sum_i^n X_{1i} \times X_{2i} - n\hat{P}_1\hat{P}_2\right)}{n(n - 1)}$$

Sustituyendo los valores:

$$1 - f = 1 - \frac{10}{100} = 0{,}9 \quad \hat{P}_1 = \frac{90}{100} \quad \hat{P}_2 = \frac{85}{100} \quad n = 100$$

$$\sum_i^n X_{1i} \times X_{2i} = 80$$

se obtiene:

$$\hat{V}(\hat{P}_1) = 0{,}00082 \quad \hat{V}(\hat{P}_2) = 0{,}00116 \quad \hat{\text{Cov}}\,(\hat{P}_1; \hat{P}_2) = 0{,}00032$$

y, por consiguiente:

1.° $\hat{V}(\hat{D}) = 0{,}00134 \quad \hat{\sigma}_D = 0{,}0366.$

$$\hat{P}_{st} = 0{,}38 \times \frac{20}{38} + 0{,}62 \times \frac{40}{62} = 0{,}6$$

2.° Si las muestras hubiesen sido independientes, la covarianza sería cero, y la varianza estimada de $\hat{D}$:

$$\hat{V}(\hat{D}) = \hat{V}(\hat{P}_1) + \hat{V}(\hat{P}_2) = 0{,}00198$$

y el error de muestreo

$$\hat{\sigma}_{\hat{D}} = 0{,}0445$$

3.° El coeficiente de correlación es:

$$\rho_{12} = \frac{\text{Cov}\,(P_1; P_2)}{\sigma_{P_1}\sigma_{P_2}} = 0{,}3$$

7.9. En una población de $N = 200$ personas se obtiene una muestra aleatoria simple (sin reposición y probabilidades iguales) de $n = 100$. No existe falta de respuesta y al repetir la encuesta con idéntica muestra se obtienen los resultados siguientes, para una determinada pregunta A:

Ocasión 1

Ocasión 2	Sí	No	Total
Sí	60	10	70
No	5	25	30
Total	65	35	100

Se pide:

a) Error de muestreo de la diferencia $\hat{d} = \hat{P}_2 - \hat{P}_1$ de proporciones con contestaciones afirmativas a la pregunta A en las correspondientes ocasiones.

b) Si la población se divide en dos estratos, de igual tamaño, con varianzas S_1^2 9 $S_2^2 = 4$ y costes unitarios $c_1 = 16$, $c_2 = 25$, determinar el tamaño n de muestra que proporcione una varianza $V(\bar{x}_{st}) = 0{,}035$, utilizando valores de n_h/n que minimicen el coste total $c = c_1 \times n_1 + c_2 \times n_2$.

(Propuesto en las Oposiciones al Cuerpo de Estadísticos Facultativos, 1978.)

Solución:

a) La varianza de $\hat{P}_2 - \hat{P}_1$, estimada, es:

$$\hat{V}(\hat{d}) = \frac{1-f}{n-1}\left[\hat{P}_1(1-\hat{P}_1) + \hat{P}_2(1-\hat{P}_2) - 2\frac{\sum_i^n X_{i1}X_{i2} - n\hat{P}_1\hat{P}_2}{n}\right]$$

y sustituyendo los valores:

$$n = 100 \quad N = 200 \quad f = \frac{1}{2} \quad \hat{P}_1 = 0{,}65 \quad \hat{P}_2 = 0{,}70 \quad \hat{d} = 0{,}05$$

se obtiene:

$$\hat{V}(\hat{d}) = \frac{\left(\frac{0{,}2275}{99} + \frac{0{,}21}{99}\right)}{2} - \frac{(60 - 45{,}5)}{9.900} = 0{,}00221 - 0{,}00146 = 0{,}00075$$

de donde

$$\hat{\sigma}_{\hat{d}} = 0{,}0274$$

b) Para minimizar el coste utilizaremos la función:

$$\phi = 16n_1 + 25n_2 + \frac{\lambda}{4}\left[\left(1 - \frac{n_1}{100}\right) \times \frac{S_1^2}{n_1} + \left(1 - \frac{n_2}{100}\right) \times \frac{S_2^2}{n_2} - 0{,}035\right]$$

y derivando respecto de n_1 y n_2, tenemos:

$$\frac{\partial \phi}{\partial n_1} = 16 - \frac{\lambda}{4} \times \frac{S_1^2}{n_1^2} = 0$$

$$\frac{\partial \phi}{\partial n_2} = 25 - \frac{\lambda}{4} \times \frac{S_2^2}{n_2^2} = 0$$

de donde

$$n_1 = 1{,}875 \times n_2 \tag{1}$$

y como

$$\frac{1}{4}\left[\left(1 - \frac{n_1}{100}\right) \times \frac{9}{n_1} + \left(1 - \frac{n_2}{100}\right) \times \frac{4}{n_2}\right] = 0{,}035 \tag{2}$$

de (1) y (2) obtenemos:

$$n_1 = 77 \quad n_2 = 41 \quad n = 118$$

7.10. A fin de evaluar la calidad en las respuestas para investigar una determinada característica cualitativa se utiliza un modelo de entrevista repetida. Los entrevistadores E.O. (entrevista original) y E.R. (entrevista repetida) obtienen los siguientes resultados al visitar una muestra de $m = 100$ personas:

E.R. \ E.O.	Con la característica	Sin la característica	Total
Con la característica	$a = 12$	$b = 18$	30
Sin la característica	$c = 14$	$d = 56$	70
Total	26	74	100

Se supone que ambos tipos de entrevistadores han recibido un adiestramiento similar y no existe correlación entre los errores de respuesta de una misma unidad.

Se pide: Estimar la varianza simple de respuesta y el índice de inconsistencia de Hanson y Pritzker.

Solución:

Designando por X_{i1} y X_{i2} la respuesta de la unidad i-ésima al entrevistador E.O. y E.R., respectivamente, la varianza de respuesta simple se estima mediante:

$$\hat{\sigma}_d^2 = \frac{1}{2m}\sum_i^m (x_{i1} - X_{i2})^2 = \frac{1}{2m}\left[\sum_i^m X_{i1}^2 + \sum_i^m S_{i2}^2 - 2\sum_i^m X_{i1}X_{i2}\right] =$$

$$= \frac{b+c}{2m} = \frac{32}{200} = 0{,}16$$

El índice de inconsistencia de Hanson y Pritzker mide la relación entre la varianza de respuesta y la de muestreo, y se define mediante:

$$I = \frac{\sigma_d^2}{V(\hat{P})}$$

Un estimador es

$$\hat{I} = \frac{\hat{\sigma}_d^2}{\hat{V}(\hat{P})}$$

siendo

$$\hat{\sigma}_d^2 = \frac{b + c}{2m}$$

y

$$\hat{V}(\hat{P}) = \hat{P}(1 - \hat{P})$$

la media de $\hat{P}_1(1 - \hat{P}_1)$ y $\hat{P}_2(1 - \hat{P}_2)$, es decir:

$$\hat{P}(1 - \hat{P}) = \frac{1}{2}\left[\frac{a + b}{m} \times \frac{c + d}{m} + \frac{a + c}{m} \times \frac{b + d}{m}\right] = 0{,}2012$$

Por tanto:

$$\hat{I} = \frac{0{,}16}{0{,}2012} \simeq 0{,}8$$

lo que indica cierto equilibrio entre la varianza de respuesta y la de muestreo.

7.11. Una muestra de $n = 100$ personas se divide aleatoriamente en dos submuestras de igual tamaño $m = 50$ que se asignan, respectivamente, a los entrevistadores E_1 y E_2. El supervisor es común para ambos y los resultados obtenidos para un carácter cualitativo son:

	E_1	E_2
$\sum_{j}^{50} X_{ijt}$	12	18
$\hat{P}_{it}$	$\frac{12}{50}$	$\frac{18}{50}$

Se pide: Obtener una estimación insesgada de la varianza total:

$$V(\hat{P}_{it}) = \frac{\sigma_0^2 + \sigma_d^2(1 + (m - 1)\rho_w)}{n}$$

y comprobar que la diferencia entre dicha estimación y la varianza de muestreo es una estimación de la componente correlacionada de la varianza de respuesta.

Solución:

El cuadrado medio

$$\hat{S}_b^2 = \frac{m}{K-1}\left[\sum_i^K \hat{P}_{it}^2 - K\hat{P}_t^2\right]$$

«entre submuestras» es:

$$\hat{S}_b^2 = \frac{50}{2-1} \times \left[\frac{144}{50^2} + \frac{324}{50^2} - 2 \times 0{,}3^2\right] = 50 \times [0{,}1872 - 0{,}13] = 0{,}36$$

y un estimador insesgado de la varianza total es:

$$\frac{\hat{S}_b^2}{n} = \frac{0{,}36}{100} = 0{,}0036$$

El cuadrado medio

$$\hat{S}_w^2 = \frac{m}{K \times (m-1)} \times \sum_i^K \hat{P}_{it}(1 - \hat{P}_{it})$$

«dentro de submuestras» viene dado por:

$$\hat{S}_w^2 = \frac{50}{2 \times 49}[0{,}24 \times 0{,}76 + 0{,}36 \times 0{,}64] = 0{,}2106$$

y como

$$\frac{\hat{S}_b^2 - \hat{S}_w^2}{K \times (m-1)}$$

es un estimador insesgado de la componente correlacionada, tendremos:

$$\hat{\rho}_w\hat{\sigma}_d^2 = \frac{0{,}36 - 0{,}2106}{2 \times 49} = 0{,}0015$$

valor al que llegaríamos también restando a la varianza total estimada la varianza de muestreo

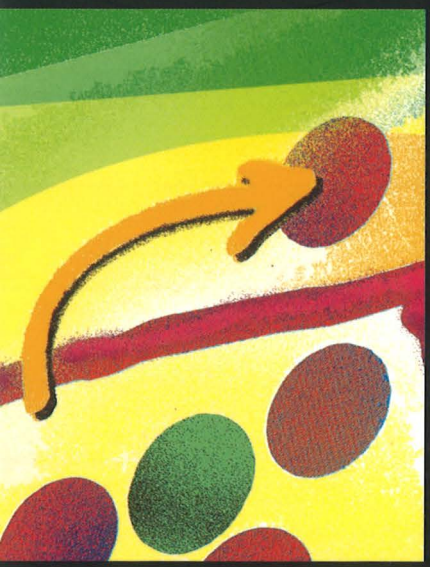
$$\sigma_0^2 = \frac{\hat{P}(1-\hat{P})}{n} = 0{,}3 \times 0{,}7 = 0{,}0021$$

es decir

$$0{,}0036 - 0{,}0021 = 0{,}0015$$

Frente al enfoque habitual en este tipo de obras —que aborda los problemas teóricamente y llega a un conjunto de fórmulas, más o menos complejas, sin desarrollar su resolución completa—, este libro trata de satisfacer la creciente demanda de estudiantes, opositores e investigadores que buscan una resolución pormenorizada de los problemas clásicos de muestreo en poblaciones finitas.

La experiencia docente de sus autores les ha permitido confirmar el hecho de que *muchos estudiantes con un perfecto conocimiento teórico de las técnicas de muestreo son incapaces de resolver un sencillo supuesto práctico, naufragando en la aplicación de los conceptos teóricos.* Para hacer frente a esta dificultad, el presente libro ofrece un conjunto de problemas resueltos —*con desmenuzamiento que puede parecer trivial en ciertos casos*— que sigue el orden lógico con el que se enseñan las técnicas de muestreo y que abarca desde los conceptos básicos de espacio muestral y distribuciones en el muestreo hasta algunas técnicas especiales (muestreo sistemático, bifásico, en ocasiones sucesivas y errores ajenos al muestreo).



Colección Libros de autor.

150 1856 2006 ANIVERSARIO ESTADÍSTICA OFICIAL ESPAÑOLA

INSTITUTO NACIONAL DE ESTADÍSTICA INE

ISBN-978-84-260-1967-7

9 788426 019677