Problemas resueltos de

Teoría de Muestras en Poblaciones Finitas

Prof. Alfonso García Pérez

Departamento de Estadística

UNED

Edita: Universidad Nacional de Educación a Distancia

Prólogo

El presente texto es un libro de apoyo dirigido a los alumnos de la UNED que cursan la asignatura *Diseño de Experimentos y Teoría de Muestras* de cuarto curso de Ciencias Matemáticas.

Se compone, básicamente, de problemas resueltos de exámenes que han ido apareciendo en los últimos años, por lo que además de servir de ayuda a un trabajo práctico, constituye un texto de orientación de las Primeras Pruebas Presenciales.

Esperamos que se consigan los objetivos para los que fué concebido.

Alfonso García Pérez

Madrid, Octubre de 1994

Contenido

		Página
1	Muestreo aleatorio	1
2	Muestreo estratificado	33
3	Estimadores de la razón	75
4	Estimadores de regresión lineal	85
5	Muestreo sistemático	91
6	Muestreo por conglomerados	95
7	Muestreo con probabilidades desiguales	103
8	Muestreo polifásico	109

CAPÍTULO 1

Muestreo aleatorio

Supondremos una población finita formada por N individuos o *unidades elementales*, $\mathcal{A} = \{U_1, ..., U_N\}$.

En dicha población se quiere investigar alguna característica de dichos individuos X_i , $i = 1, ..., N$, tales como su peso, su talla, etc., asociados a la cual están definidos los parámetros poblacionales

media poblacional	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
total poblacional	$X = \sum_{i=1}^{N} X_i = N\,\overline{X}$
varianza poblacional	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \overline{X})^2$
cuasivarianza poblacional	$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2 = \frac{N\,\sigma^2}{N-1}$

Si la característica en estudio es de tipo cualitativo, se asocia a ella una variable dicotómica, denominada ahora A_i, la cual toma el valor 1 si se presenta la característica en estudio, ó 0 si no se presenta tal característica. La media y el total poblacionales se convierten ahora en

proporción poblacional	$P = \frac{1}{N}\sum_{i=1}^{N} A_i$
total de clase poblacional	$A = \sum_{i=1}^{N} A_i = N\,P$

siendo ahora la varianza poblacional $\sigma^2 = P\,Q$.

El objetivo habitual será el de obtener estimaciones de dichos parámetros poblacionales, observando la característica en estudio sólamente en una parte de la población, denominada *muestra*, $X_1, ..., X_n$.

En todo el libro consideraremos únicamente *muestreo aleatorio*, el cual podrá ser *sin reemplazamiento* de las unidades elementales, o *con reemplazamiento* de las mismas.

Si además, como ocurre en los primeros capítulos del libro, el muestreo es con *probabilidades iguales*, en el sentido de que en cada extracción —la t-ésima—, todos los individuos tienen la misma probabilidad de ser seleccionados: $1/(N - t + 1)$, $t = 1, ...n$, si el muestreo es sin reposición y $1/N$, $t = 1, ...n$, si el muestreo es con reposición, los dos tipos de muestreo antes mencionados recibirán el nombre de *muestreo aleatorio simple* en el caso de un muestreo aleatorio sin reemplazamiento y probabilidades iguales y de *muestreo aleatorio con reemplazamiento* en el otro.

Tanto en uno como en otro tipo de muestreo, los estimadores insesgados de los parámetros poblacionales antes mencionados son

media muestral	$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$
total muestral	$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} X_i = N\,\overline{x}$
proporción muestral	$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} A_i$
total de clase muestral	$\hat{A} = \frac{N}{n} \sum_{i=1}^{n} A_i = N\,\hat{p}$

de varianzas respectivas, en un muestreo aleatorio simple

$$V(\overline{x}) = \frac{N-n}{N}\,\frac{S^2}{n}$$

$$V(\hat{X}) = N(N-n)\,\frac{S^2}{n}$$

$$V(\hat{p}) = \frac{N-n}{N-1}\,\frac{PQ}{n}$$

$$V(\hat{A}) = N^2\,\frac{N-n}{N-1}\,\frac{PQ}{n}$$

y en un muestreo aleatorio con reemplazamiento

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

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

$$V(\hat{p}) = \frac{PQ}{n}$$

$$V(\hat{A}) = N^2 \frac{PQ}{n}$$

denominándose *error de muestreo* de un estimador a la raiz cuadrada de la varianza de dicho estimador y *error relativo de muestreo* o *coeficiente de variación* de un estimador T al cociente $\sqrt{V(T)}/E[T]$.

Como las varianzas de los estimadores dependen de parámetros poblacionales y éstos, habitualmente, serán desconocidos, habrá que estimar las varianzas anteriores, utilizando la *cuasivarianza muestral*

$$\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2$$

el cual, para variables cualitativas, es igual a $\hat{S}^2 = n\,\hat{p}\,\hat{q}/(n-1)$, siendo $\hat{q} = 1 - \hat{p}$.

Como en un muestreo aleatorio simple es $E[\hat{S}^2] = S^2$ y en un muestreo aleatorio con reemplazamiento $E[\hat{S}^2] = \sigma^2$, los estimadores insesgados de la varianza de los estimadores son, en un muestreo aleatorio simple

$$\hat{V}(\overline{x}) = \frac{N-n}{N}\,\frac{\hat{S}^2}{n}$$

$$\hat{V}(\hat{X}) = N(N-n)\,\frac{\hat{S}^2}{n}$$

$$\hat{V}(\hat{p}) = \frac{N-n}{N}\,\frac{\hat{p}\,\hat{q}}{n-1}$$

$$\hat{V}(\hat{A}) = N(N-n)\,\frac{\hat{p}\,\hat{q}}{n-1}$$

y en un muestreo aleatorio con reemplazamiento

$$\hat{V}(\overline{x}) = \frac{\hat{S}^2}{n}$$

$$\hat{V}(\hat{X}) = N^2\,\frac{\hat{S}^2}{n}$$

$$\hat{V}(\hat{p}) = \frac{\hat{p}\,\hat{q}}{n-1}$$

$$\hat{V}(\hat{A}) = N^2\,\frac{\hat{p}\,\hat{q}}{n-1}$$

Problema 1.1.- Dada una población finita de tamaño N, determinar:

(a) La probabilidad de obtener una muestra sin reemplazamiento de tamaño n.

(b) El número de muestras posibles de tamaño n.

(c) La probabilidad de que un elemento determinado pertenezca a la muestra de tamaño n.

(a) Suponemos un muestreo con *probabilidades iguales* en el sentido de que, en cada extracción, t, todos los individuos tienen la misma probabilidad de ser seleccionados: $1/(N-t+1)$, $t = 1,...,n$, si el muestreo es sin reposición —*muestreo aleatorio simple*— ó $1/N$, $t = 1,...,n$, si el muestreo es con reposición —*muestreo aleatorio con reemplazamiento*—.

El enunciado del problema nos habla de un muestreo aleatorio simple, por lo que la probabilidad de obtener una muestra específica $\{u_1,...,u_n\}$, será, en ese orden

$$\frac{1}{N}\cdot\frac{1}{N-1}\cdot...\cdot\frac{1}{N-n+1}$$

y como la misma muestra se obtendrá independientemente del lugar en el que se extrajo cada elemento, la probabilidad pedida será el resultado de multiplicar la probabilidad anterior por el número de reordenaciones posibles de los n elementos, $n!$. Es decir,

$$n!\cdot\frac{1}{N}\cdot\frac{1}{N-1}\cdot...\cdot\frac{1}{N-n+1}=\frac{n!(N-n)!}{N!}=\frac{1}{\binom{N}{n}}.$$

(b) Por otro lado, el número de muestras posibles serán las combinaciones de los N elementos tomados de n en n, es decir $\binom{N}{n}$, con lo que del apartado (a) se deduce que todas las muestras posibles tienen la misma probabilidad de ser seleccionadas.

(c) Por último, como la probabilidad que tiene un individuo determinado de incorporarse a la muestra, exactamente en la extracción i-ésima es

$$\frac{N-1}{N}\cdot\frac{N-2}{N-1}\cdot...\cdot\frac{N-(i-1)}{N-(i-2)}\cdot\frac{1}{N-(i-1)}=\frac{1}{N}$$

para todo $i = 1,...,n$, la probabilidad que tiene ese individuo de pertenecer a la muestra será

$$\frac{1}{N}+\frac{1}{N}+...+\frac{1}{N}=\frac{n}{N}.$$

Problema 1.2.- Supongamos una población compuesta por mil fincas donde la varianza del número de vacunos por finca es 250. Estimemos el número medio de vacunos por finca mediante una muestra aleatoria simple. Suponiendo hipótesis de normalidad, determinar el tamaño mínimo de muestra necesario para que, con probabilidad 0'95, el error que se cometa al estimar dicho número medio con la media muestral, no sea mayor de 1.

La condición requerida en el enunciado se puede expresar de la forma

$$P\{|\overline{x}-\overline{X}|\leq 1\}=0'95.$$

Bajo la hipótesis de que el número de vacunos por finca sigue una distribución normal, la media muestral sigue también una distribución normal de media $\overline{X}$ y de varianza

$$V(\overline{x})=\frac{N-n}{N-1}\frac{\sigma^2}{n}.$$

Tipificando en la expresión anterior se tendrá, por un lado que

$$P\left\{|Z|\leq\frac{1}{\sqrt{\frac{1000}{999}\frac{250}{n}-\frac{250}{999}}}\right\}=0'95$$

con $Z\rightsquigarrow N(0,1)$, pero por otro que

$$P\{|Z|\leq 1'96\}=0'95$$

con lo que deberá ser,

$$\frac{1}{\sqrt{\frac{1000}{999}\frac{250}{n}-\frac{250}{999}}}=1'96$$

de donde se obtiene, $n=490'15$. Por tanto, un tamaño de muestra de 491 unidades será el requerido.

Problema 1.3.- En una Facultad de 200 alumnos se hizo una encuesta sobre una determinada asignatura, a 25 estudiantes elegidos al azar y sin reemplazamiento. De ellos, 18 estuvieron satisfechos con la enseñanza recibida, expresando el número de horas que necesitaron para preparar el examen mediante la siguiente distribución de frecuencias, en donde X es el tiempo empleado:

X_i	15	20	30
n_i	6	12	7

Se pide, estimar:

(a) La proporción del número de alumnos satisfechos, y su error de muestreo.

(b) La media de horas por alumno, su error de muestreo y un intervalo de confianza para un coeficiente de confianza de 0'95, suponiendo normalidad.

(a) Un estimador de la proporción de alumnos satisfechos será la proporción muestral

$$\hat{p} = \frac{18}{25} = 0'72$$

siendo su varianza estimada,

$$\hat{V}(\hat{p}) = \frac{N-n}{N} \cdot \frac{\hat{p}\hat{q}}{n-1} = \frac{200-25}{200} \cdot \frac{0'72 \cdot 0'28}{24} = 0'00735$$

y, por tanto, su error de muestreo,

$$\hat{\sigma}_{\hat{p}} = \sqrt{\hat{V}(\hat{p})} = 0'086.$$

(b) La media muestral de horas por alumno será,

$$\overline{x} = \frac{\sum_{i=1}^{3} x_i n_i}{25} = 21'6$$

y su varianza estimada,

$$\hat{V}(\overline{x}) = \frac{N-n}{N} \cdot \frac{\hat{S}^2}{n} = \frac{200-25}{200} \cdot \frac{32'75}{25} = 1'14625$$

siendo, por tanto, su error de muestreo

$$\hat{\sigma}_{\overline{x}} = 1'07.$$

Por último, el intervalo de confianza pedido será,

$$\left[\overline{x} \mp t_{n-1;\alpha/2} \frac{\hat{S}}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \right]$$

es decir,

$$\left[21{,}6 \mp 2{,}064 \sqrt{1'14625} \right] = [19'39 \ , \ 23'81].$$

Problema 1.4.- En una muestra aleatoria simple constituida por 25 presos condenados por el mismo tipo de delito, se observó que la duración media de la pena impuesta era de 132'3 meses con una desviación típica de 34'7 meses. Calcule un intervalo de confianza para la media de la población, con un coeficiente de confianza de 0'95.

La característica en estudio, X, *duración de la pena impuesta en meses*, no sabemos si sigue una distribución normal, y además, el tamaño de la muestra no es suficientemente grande como para poder utilizar el teorema central del límite, el cual permitiría suponer una distribución normal para la media muestral $\overline{x}$. La varianza poblacional tampoco es conocida.

Por estas razones, debemos de utilizar el intervalo de confianza basado en la desigualdad de Chebychev, el cual viene dado por

$$\left[\overline{x} - \frac{1}{\sqrt{\alpha}} \sqrt{\left(\frac{\hat{S}^2}{n}\right)\left(\frac{N-n}{N}\right)} \ ; \ \overline{x} + \frac{1}{\sqrt{\alpha}} \sqrt{\left(\frac{\hat{S}^2}{n}\right)\left(\frac{N-n}{N}\right)} \right]$$

siendo $\overline{x} = 132'3$, $\hat{S}^2 = 25(34'7)^2/24 = 1254'26$ y $\alpha = 0'05$.

Como no se hace referencia al tamaño de la población, la supondremos infinita, lo que hace que el factor de corrección para poblaciones finitas, $1 - f = 1 - (n/N)$, sea igual a 1 , siendo, por tanto, el intervalo pedido,

$$\left[132'3 - \sqrt{\frac{1254'26}{0'05 \cdot 25}} \ , \ 132{,}3 + \sqrt{\frac{1254'26}{0'05 \cdot 25}} \right] = [100'62 \ , \ 163'98].$$

Problema 1.5.- Si la función de pérdida debida a un error en la estimación de la media poblacional $\overline{X}$ por la media muestral $\overline{x}$ es proporcional a $|\overline{x} - \overline{X}|$ y si el coste total del estudio es $C = a + cn$, demuestre que para un muestreo aleatorio simple, el valor óptimo de n es

$$\left(\frac{k\sigma}{c\sqrt{2\pi}}\right)^{2/3}$$

siendo k la constante de proporcionalidad y $\overline{x} \rightsquigarrow N(\overline{X}, \sigma/\sqrt{n})$.

La pérdida total que comete el estadístico al realizar el muestreo será

$$L = k|\overline{x} - \overline{X}| + a + cn$$

y su esperanza (pérdida esperada) igual a

$$E[L] = k\int_{\mathbb{R}} |\overline{x} - \overline{X}| f(\overline{x})\, d\overline{x} \; + a + cn.$$

Como es

$$\int_{\mathbb{R}} |\overline{x} - \overline{X}| f(\overline{x})\, d\overline{x} = -\int_{-\infty}^{\overline{X}} (\overline{x} - \overline{X}) f(\overline{x})\, d\overline{x} + \int_{\overline{X}}^{\infty} (\overline{x} - \overline{X}) f(\overline{x})\, d\overline{x} =$$

$$= 2\int_{\overline{X}}^{\infty} (\overline{x} - \overline{X}) f(\overline{x})\, d\overline{x} = 2\frac{\sigma}{\sqrt{2\pi}\sqrt{n}}$$

será,

$$E[L] = 2k\frac{\sigma}{\sqrt{2\pi}\sqrt{n}} + a + cn$$

función de n que alcanza su máximo en

$$n = \left(\frac{k\sigma}{c\sqrt{2\pi}}\right)^{2/3}$$

Problema 1.6.- En un estudio sobre el posible empleo de muestreo para ahorrar trabajo en la realización de inventarios en un almacén, se llevó la cuenta del valor de los artículos en cada uno de los 36 estantes del mencionado almacén. Los valores, en pesetas, obtenidos fueron los siguientes:

$$\sum_{i=1}^{36} X_i = 2.138 \ , \ \sum_{i=1}^{36} X_i^2 = 131.682$$

siendo X_i el valor de los artículos del estante i-ésimo.

El estimador del valor total obtenido mediante la muestra se considera correcto si difiere en menos de 200 pesetas del verdadero, con probabilidad 0'95.

Un asesor sugiere que un muestreo sin reemplazamiento de 12 estantes cumpliría los requisitos. Admitiendo la hipótesis de normalidad, ¿podemos estar de acuerdo con esta afirmación?.

Se trata de determinar el tamaño de la muestra en la estimación del total para alcanzar en precisión dada.

La condición expuesta en el enunciado es

$$P\{|\hat{X} - X| < 200\} = 0'95$$

o, suponiendo normalidad,

$$P\left\{|Z| < \frac{200}{\sqrt{V(\hat{X})}}\right\} = 0'95$$

con $Z \rightsquigarrow N(0,1)$.

De una tablas de la $N(0,1)$ se obtiene que

$$P\{|Z| < 1'96\} = 0'95$$

con lo que deberá ser

$$\frac{200}{\sqrt{V(\hat{X})}} = 1'96$$

es decir,

$$\left(\frac{200}{1'96}\right)^2 = V(\hat{X}) = N(N-n)\frac{S^2}{n} = \frac{N^2 S^2}{n} - N S^2$$

o bien, despejando,

$$n = \frac{1'96^2 N^2 S^2}{200^2 + 1'96^2 N S^2} = \frac{1'96^2 \cdot 36^2 \cdot 134'5303}{200^2 + 1'96^2 \cdot 36 \cdot 134'5303} = 11'4288.$$

Por tanto, la respuesta es sí.

Problema 1.7.- Los siguientes coeficientes de variación, o error relativo de muestreo, poblacionales, fueron determinados observando una encuesta piloto de granjas en el estado de Iowa, siendo la unidad de área 1 milla cuadrada.

Características	Coefic. de variación poblacional (en %)
Acres de granja	38
Acres de trigo	39
Acres de avena	44
Nº de familias trabajadoras	100
Nº de trabajadores asalariados	110
Nº de parados	317

Se quiere determinar el tamaño muestral de una encuesta que permita estimar las características de área (total de acres de granja, trigo y avena) y el total de trabajadores (excluidos los parados) de forma que los coeficientes de variación no superen el 2'5 % y el 5 % respectivamente. Con un muestreo aleatorio con reemplazamiento, ¿cuál debe ser el tamaño de la muestra?.

Con el tamaño de muestra calculado, ¿qué coeficiente de variación deberá esperarse para la estimación del número de parados?.

Nota: *Se denomina coeficiente de variación del total poblacional X al valor $N\sigma/X$.*

La condición exigida para las características de área (total de acres) es

$$C.V.(\hat{X}) \leq 0'025$$

es decir,

$$\frac{D(\hat{X})}{E(\hat{X})} = \frac{N\sigma}{\sqrt{n}X} \leq 0'025. \qquad [1]$$

Por otro lado, el coeficiente de variación de un parámetro poblacional es el cociente entre su desviación típica y él mismo. Es decir, para el total poblacional $X = \sum_{i=1}^{N} X_i$

$$C.V.(X) = \frac{D(X)}{X} = \frac{N \cdot \sigma}{X}$$

como decía la nota del enunciado. La condición [1] es, por tanto,

$$C.V.(X)/\sqrt{n} \leq 0'025$$

o bien,

$$\sqrt{n} \geq \frac{C.V.(X)}{0'025}$$

Como de los coeficientes de variación (para las características de área) que aparecen en la tabla, el mayor es 0'44, tomando un n

$$n \geq \frac{(0'44)^2}{(0'025)^2} = 309'76$$

cumpliremos la primera restricción solicitada.

Para el total de trabajadores deberá ser también

$$C.V.(\hat{X}) \leq 0'05$$

o bien,

$$\sqrt{n} \geq \frac{C.V.(X)}{0'05}.$$

Eligiendo de nuevo el mayor coeficiente de variación poblacional, de entre los referentes al total de trabajadores (excluidos los parados), se obtiene una condición para n

$$n \geq \frac{(1'1)^2}{(0'05)^2} = 484.$$

Tomando, por tanto, un tamaño muestral $n = 484$ cumpliremos los dos requerimientos.

Para la estimación del número de parados cabe esperar un coeficiente de variación

$$C.V.(\hat{X}) = \frac{C.V.(X)}{\sqrt{n}} = \frac{3'17}{\sqrt{484}} = 0'144.$$

Problema 1.8.- Sea una población finita de tamaño $N = 1000$ y de cuasivarianza poblacional $S^2 = 500$, de la que extraemos una muestra aleatoria simple, con la que queremos estimar la media poblacional $\overline{X}$ con un error absoluto menor o igual que 10 y con un coeficiente de confianza del 90%. Hallar el tamaño de muestra mínimo necesario para cumplir las condiciones requeridas, primero sin suponer la población normal y luego suponiéndola.

La condición expresada por el enunciado es

$$P\{|\overline{x} - \overline{X}| \leq 10\} = 0'9$$

es decir,

$$P\{|\overline{x} - \overline{X}| > 10\} = 0'1.$$

De la desigualdad de Chebychev,

$$P\{|\overline{x} - \overline{X}| > k\} \leq \frac{V(\overline{x})}{k^2} = \frac{(N-n)S^2}{k^2\, N\, n}$$

se deduce que el n mínimo será aquel que verifique la igualdad

$$\frac{(N-n)S^2}{k^2\, N\, n} = 0'1$$

es decir,

$$\frac{(1000-n)500}{100 \cdot 1000 \cdot n} = 0'1$$

de donde se obtiene el valor $n = 47'619$. Por tanto, un tamaño muestral de $n = 48$ cumple las condiciones del enunciado.

Suponiendo normalidad, será por un lado

$$P\left\{|Z| \leq \frac{10\sqrt{N\, n}}{S\sqrt{N-n}}\right\} = 0'9$$

con $Z \rightsquigarrow N(0,1)$, y por otro

$$P\{|Z| \leq 1'645\} = 0'9$$

con lo que debe ser

$$\frac{10\sqrt{N\,n}}{S\sqrt{N-n}} = \frac{10\sqrt{1000\,n}}{\sqrt{(1000-n)500}} = 1'645$$

obteniéndose el valor $n = 13'3495$. Por tanto, un tamaño muestral de $n = 14$ unidades cumpliría las condiciones del enunciado.

Problema 1.9.- Se extrajo una muestra aleatoria simple de 290 familias de un área urbana que contenía 14.828 familias. A cada una de ellas se le preguntó si tenía la vivienda en propiedad o no, y si hacía uso exclusivo de los servicios o no. Los resultados fueron los recogidos en la siguiente tabla.

	Uso exclusivo Servicios		Totales
	Si	*No*	
Propietario	141	6	147
Alquiler	109	34	143
Totales	250	40	290

Se pide:

(a) Para familias en alquiler, estimar el porcentaje de las que hacen uso exclusivo de los servicios y su error de muestreo.

(b) Estimar el número total de familias en alquiler que no hacen uso exclusivo de los servicios y su error de muestreo.

(c) Si suponemos ahora que el número total de familias en alquiler en el área urbana citada es 7.526, hacer una nueva estimación del número de familias que no hacen uso exclusivo de los servicios y de su error de muestreo. Comparar los resultados obtenidos con los del apartado anterior.

(a) La distribución —condicionada— para familias en alquiler será

Si	109
No	34
	143

de la que se obtiene el estimador proporción muestral $\hat{p}_2 = 109/143 = 0'762$.

Su error de muestreo estimado será

$$\hat{D}(\hat{p}_2) = \sqrt{\frac{N_2 - n_2}{N_2}\,\frac{\hat{p}_2\,\hat{q}_2}{n_2 - 1}}.$$

Como el número de familias en alquiler de la población, N_2, es desconocido, lo estimaremos por

$$\hat{N}_2 = \frac{143}{290} \cdot 14828 \approx 7312$$

quedando el error de muestreo igual a

$$\hat{D}(\hat{p}_2) = \sqrt{\frac{7312 - 143}{7312} \cdot \frac{0'762 \cdot 0'238}{142}} = 0'035386.$$

(b) El estimador del total de clase será

$$\hat{A} = N\hat{p} = 14828 \cdot \frac{34}{290} = 1738'46$$

y su error de muestreo

$$\hat{D}(\hat{A}) = \sqrt{N(N-n)\frac{\hat{p}\,\hat{q}}{n-1}} = \sqrt{14828 \cdot (14828 - 290)\frac{34 \cdot 256}{290^2 \cdot 289}} = 277'847.$$

(c) Ahora la estimación del total de clase (efectuada utilizando la distribución condicionada anterior) será

$$\hat{A}_2 = 7526 \cdot \frac{34}{143} = 1789'4$$

y su error de muestreo

$$\hat{D}(\hat{A}_2) = \sqrt{N_2(N_2 - n_2)\frac{\hat{p}_2\,\hat{q}_2}{n_2 - 1}} = \sqrt{7526 \cdot (7526 - 143)\frac{0'238 \cdot 0'762}{142}} = 266'39.$$

Como se ve, esta segunda estimación tiene un error de muestreo menor, por lo que es más fiable; claro está, supuesto conocido el número total de familias en alquiler del área urbana en cuestión.

Problema 1.10.- Se extrajo una muestra aleatoria con reemplazamiento de tamaño $n = 3$ de una población finita de N elementos. Se pide:

(a) Calcular la probabilidad π_1 de que la muestra esté formada por tres elementos iguales, π_2 de que esté formada por dos elementos iguales y otro distinto y π_3 de que esté formada por tres elementos distintos.

(b) Si $\overline{y}'$ es la media aritmética de los elementos distintos de la muestra seleccionada, estudiar si es insesgado y calcular su varianza esperada.

(c) Comprobar que dicha varianza esperada es menor que la varianza de la media muestral $\overline{y}$.

(a) Las probabilidades pedidas serán

$$\pi_1 = \frac{N}{N^3} = \frac{1}{N^2}$$

$$\pi_2 = \frac{3!\,N(N-1)/2}{N^3} = \frac{3(N-1)}{N^2}$$

$$\pi_3 = \frac{N(N-1)(N-2)}{N^3} = \frac{(N-1)(N-2)}{N^2}$$

(b) $\overline{y}'$ será una variable aleatoria con distribución

$$\overline{y}' = \begin{cases} \dfrac{x_1+x_2+x_3}{3} & \text{con probabilidad } \pi_3 \\ \dfrac{x_1+x_2}{2} & \text{con probabilidad } \pi_2 \\ x_1 & \text{con probabilidad } \pi_1 \end{cases}$$

Si D es el número de observaciones muestrales distintas, la esperanza de $\overline{y}'$ será

$$E[\overline{y}'] = E[\overline{y}'/D=1]\cdot\pi_1 + E[\overline{y}'/D=2]\cdot\pi_2 + E[\overline{y}'/D=3]\cdot\pi_3.$$

Obsérvese que, condicionado a $D=2$ y a $D=3$, las variables X_1, X_2 y X_1, X_2, X_3 no son independientes aunque el muestreo sea con reemplazamiento, ya que D condiciona a que sean distintas. Es decir, que las esperanzas y varianzas —condicionadas— han de ser calculadas mediante las correspondientes fórmulas de un muestreo aleatorio simple.

Así, por ser, en un muestreo aleatorio simple, la media muestral un estimador insesgado de la media poblacional, será

$$E[\overline{y}'] = \overline{Y}\cdot\pi_1 + \overline{Y}\cdot\pi_2 + \overline{Y}\cdot\pi_3 = \overline{Y}.$$

De la misma manera, $V(\overline{y}')$ será una variable aleatoria con distribución

$$V(\overline{y}') = \begin{cases} \sigma^2 & \text{con probabilidad } \pi_1 \\ \dfrac{(N-2)}{N} \cdot \dfrac{S^2}{2} & \text{con probabilidad } \pi_2 \\ \dfrac{(N-3)}{N} \cdot \dfrac{S^2}{3} & \text{con probabilidad } \pi_3 \end{cases}$$

de esperanza

$$E[V(\overline{y}')] = \sigma^2 \cdot \frac{1}{N^2} + \frac{(N-2)}{N} \cdot \frac{S^2}{2} \cdot \frac{3(N-1)}{N^2} +$$

$$+\frac{(N-3)}{N} \cdot \frac{S^2}{3} \cdot \frac{(N-1)(N-2)}{N^2} = \frac{(2N-1)(N-1)S^2}{6N^2}.$$

(c) Por último, demostrar que $E[V(\overline{y}')] < V(\overline{y})$ equivale a demostrar que

$$\frac{(2N-1)(N-1)S^2}{6N^2} < \frac{\sigma^2}{3} = \frac{(N-1)S^2}{3N}$$

desigualdad que fácilmente se puede verificar simplificando.

Observese que $\overline{y}'$ es un estimador insesgado, al igual que $\overline{y}$, ¡de menor varianza!, aunque en promedio, ya que $V(\overline{y}')$ es una variable aleatoria, no siéndolo la varianza de $\overline{y}$.

Problema 1.11.- En un área existen $N = 10.000$ viviendas. Los datos de un censo anterior hacen suponer que aproximadamente dos terceras partes corresponden a régimen de alquiler. Determinar:

1. El tamaño de muestra necesario para estimar la proporción de viviendas en alquiler, con un error de muestreo igual a $0'04$ si se realiza un muestreo

   (a) Sin reemplazamiento. (b) Con reemplazamiento.

2. El tamaño de muestra necesario para estimar la proporción de viviendas en alquiler, con un error absoluto máximo admisible $e_0 = 0'08$ y un coeficiente de confianza $\alpha = 0'95$, supuesto normalidad y si se realiza un muestreo

   (a) Sin reemplazamiento. (b) Con reemplazamiento.

Como la varianza de la proporción muestrasl es $V(\hat{p}) = \dfrac{N-n}{N-1}\ \dfrac{PQ}{n}$ para un muestreo sin reemplazamiento, y $V(\hat{p}) = \dfrac{PQ}{n}$ para un muestreo con reemplazamiento, en el primer caso deberá ser

$$\sqrt{\frac{10000-n}{9999}\cdot\frac{2}{9n}} = 0'04$$

y en el segundo

$$\sqrt{\frac{2}{9n}} = 0'04$$

obteniéndose, respectivamente, unos tamaños muestrales de $136'99 \approx 137$ y $138'88 \approx 139$.

Como la condición exigida en el segundo apartado es

$$P\{|\hat{p}-P| \leq 0'08\} = 0'95$$

en caso de no haber reemplazamiento deberá ser

$$P\left\{|Z| \leq \frac{0'08\sqrt{(N-1)n}}{\sqrt{(N-n)PQ}}\right\} = 0'95$$

y en caso de haberlo

$$P\left\{|Z| \leq \frac{0'08\sqrt{n}}{\sqrt{PQ}}\right\} = 0'95$$

con $Z \rightsquigarrow N(0,1)$.

Como es $P\{|Z| \leq 1'96\} = 0'95$, deberá ser,

$$0'08 = 1'96\sqrt{\frac{NPQ}{(N-1)n} - \frac{PQ}{N-1}} = 1'96\sqrt{\frac{10000\cdot 2}{9\cdot 9999\cdot n} - \frac{2}{9\cdot 9999}}$$

en el primer caso, y

$$0'08\sqrt{n} = 1'96\sqrt{PQ} = 1'96\sqrt{\frac{2}{9}}$$

en el segundo, ecuaciones de donde se obtienen respectivamente unos valores de 131'646 y 133'388, es decir, unos tamaños muestrales mínimos de 132 y 134 unidades respectivamente.

Problema 1.12.- Dos estadísticos A y B investigaron el estado de 200 cuestionarios. El estadístico A seleccionó una muestra aleatoria simple de 20 cuestionarios, contando el número de errores por cuestionario y obteniendo los siguientes datos

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 aquellos que no tenían ningún error, encontrando 60 cuestionarios sin ningún error. Estimar el número total de errores,

(a) Sólo con los resultados del estadístico A.
(b) Con los resultados de A y B.
(c) ¿Son insesgados los estimadores?.
(d) ¿Qué estimador tiene más precisión?.

(a) De la distribución de frecuencias dada se obtiene una media muestral $\overline{x} = 42/20 = 2'1$, con lo que un estimador del total será $\hat{X} = 200 \cdot 2'1 = 420$.

(b) La información suministrada por el estadístico B, permite centrar la investigación en la subpoblación de cuestionarios con algún error, de tamaño —según la información proporcionada por B—, $N_2 = 140$.

La muestra obtenida por A, podemos considerarla ahora como una muestra aleatoria de tamaño $n_2 = 12$ (8 cuestionarios de la muestra anterior no presentaron ningún error), para la que se obtuvieron los siguientes datos

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

que proporcionan una media muestral $\overline{x}_2 = 42/12 = 3'5$ y un estimador del total $\hat{X}_2 = 140 \cdot 3'5 = 490$.

(c) Ambos estimadores son insesgados para el total poblacional X, tanto el habitual total muestral $\hat{X}$, como el utilizado en el apartado (b), ya que si por X_i' representamos los cuestionarios con algún error, será $X = \sum_{i=1}^{N} X_i = \sum_{i=1}^{N_2} X_i'$, con lo que $\hat{X}_2$, al ser un estimador insesgado en la subpoblación de tamaño N_2 será $E[\hat{X}_2] = \sum_{i=1}^{N_2} X_i' = \sum_{i=1}^{N} X_i = X$.

(d)
Las varianzas de ambos estimadores son

$$\hat{V}(\hat{X}) = N(N-n)\frac{\hat{S}^2}{n} = 200(200-20)\frac{8'621053}{20} = 15.517'895$$

y

$$\hat{V}(\hat{X}_2) = N_2(N_2-n_2)\frac{\hat{S}_2^2}{n_2} = 140(140-12)\frac{9'545455}{12} = 14.254'545$$

Luego el segundo estimador tiene un menor error de muestreo.

Problema 1.13.- Se sabe que el intervalo de confianza de la media poblacional, para un coeficiente de confianza de 0'95, tiene una longitud k. Se conoce el tamaño de la población, N, así como el de la muestra n. Determinar la varianza poblacional en función de k, N y n.

El intervalo de confianza para la media poblacional en el caso de población no necesariamente normal es, utilizando la desigualdad de Chebychev,

$$\left[\overline{x} - \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} \ , \ \overline{x} + \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} \right]$$

de longitud

$$\overline{x} + \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} - \overline{x} + \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}} = \frac{2}{\sqrt{\alpha}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}.$$

Por tanto, deberá ser

$$k = \frac{2}{\sqrt{0'05}}\sqrt{\frac{\sigma^2}{n}\frac{N-n}{N-1}}$$

de donde se obtiene que

$$\sigma^2 = \frac{k^2 n(N-1)}{80(N-n)}.$$

Problema 1.14.- Determinar el intervalo de confianza para el total poblacional, suponiendo un tamaño de muestra grande y la varianza desconocida, en el caso de un muestreo aleatorio simple: (a) con reemplazamiento y (b) sin reemplazamiento.

Aunque la variable aleatoria en estudio no siga una distribución normal, si el tamaño muestral es suficientemente grande, la distribución de la media muestral $\overline{x}$, se puede aproximar por una normal de media la media poblacional $\overline{X}$ y de varianza $V(\overline{x})$

$$\frac{\overline{x} - \overline{X}}{\sqrt{V(\overline{x})}} \rightsquigarrow N(0,1).$$

Por tanto, multiplicando el numerador y el denominador por N, tendremos que será

$$\frac{\hat{X} - X}{N\sqrt{V(\overline{x})}} \rightsquigarrow N(0,1)$$

con lo que, fijado un nivel de confianza $1-\alpha$, se podrá determinar el valor de la abscisa de una normal $N(0,1)$ que deje a la derecha un área de probabilidad $\alpha/2$, $z_{\alpha/2}$, de forma que sea

$$P\left\{-z_{\alpha/2} < \frac{\hat{X} - X}{N\sqrt{V(\overline{x})}} < z_{\alpha/2}\right\} = 1-\alpha$$

de donde se obtiene

$$\left[\hat{X} - z_{\alpha/2} N\sqrt{V(\overline{x})}\ ,\ \hat{X} + z_{\alpha/2} N\sqrt{V(\overline{x})}\right]$$

como intervalo de confianza. Si el muestreo es con reemplazamiento, el intervalo de confianza buscado será

$$\left[\hat{X} - z_{\alpha/2} N\frac{\sigma}{\sqrt{n}}\ ,\ \hat{X} + z_{\alpha/2} N\frac{\sigma}{\sqrt{n}}\right]$$

y si es sin reemplazamiento

$$\left[\hat{X} - z_{\alpha/2} S \sqrt{\frac{N(N-n)}{n}} \ , \ \hat{X} + z_{\alpha/2} S \sqrt{\frac{N(N-n)}{n}} \right]$$

Problema 1.15.- Con objeto de evitar el deterioro medio ambiental de una determinada zona geográfica, un grupo ecologista recogió firmas en toda la región. Las firmas a favor de la petición fueron recogidas en 676 hojas, cada una de las cuales tenía espacio suficiente para 42 firmas, aunque muchas de ellas no llegaron a completarse.

A la hora de contar el número total de firmas recogidas a favor de la petición, se decidió seleccionar al azar y sin reemplazamiento 50 hojas en lugar de examinar las 676, obteniéndose la siguiente distribución de frecuencias absolutas del número X_i de firmas por hoja

X_i	2	4	5	6	8	10	11	13	15	16	18	19	23	27	30	31	41	42
n_i	1	2	2	3	1	1	1	2	1	1	1	1	1	2	1	1	5	23

Estimar el número total de firmas obtenidas a favor de la petición y determinar un intervalo de confianza para esta cantidad con un coeficiente de confianza del 80 %.

De la distribución de frecuencias anterior se obtiene una media muestral $\overline{x} = 29'4$ y una cuasivarianza muestral de $\hat{S}^2 = 236'53$.

La estimación del total poblacional es $\hat{X} = N \cdot \overline{x} = 676 \cdot 29'4 = 19.874'4$ y, al ser un muestreo sin reemplazamiento, suponiendo normalidad, el intervalo de confianza pedido será

$$\left[\hat{X} - z_{\alpha/2} \hat{S} \sqrt{\frac{N(N-n)}{n}} \ , \ \hat{X} + z_{\alpha/2} \hat{S} \sqrt{\frac{N(N-n)}{n}} \right]$$

$$= \left[19.874'4 \mp 1'28 \cdot 15'38 \cdot \sqrt{\frac{676(676-50)}{50}} \right] = [18.063'3 \ , \ 21.685'5].$$

Problema 1.16.- Interesados en conocer la opinión de los colegios de un país acerca de unos nuevos planes de educación, se tomó una muestra aleatoria simple de 200 colegios de los 2.000 con que cuenta el país, contestando 120 de ellos a favor, 57 en contra y 23 en blanco.

Determinar un intervalo de confianza para el número de colegios de la población a favor del nuevo proyecto educativo, con un coeficiente de confianza del 95 %.

Como el tamaño de la muestra es suficientemente grande, $n = 200$, podemos utilizar la aproximación normal en la determinación del intervalo de confianza del total de clase, el cual será $N \cdot I$, siendo I el correspondiente a la proporción poblacional.

Éste es,

$$I = \left[\hat{p} \mp \left(z_{\alpha/2} \sqrt{\frac{N-n}{N} \cdot \frac{\hat{p}\,\hat{q}}{n-1}} + \frac{1}{2n} \right) \right]$$

en donde se ha utilizado el factor de corrección $1/(2n)$ al aproximarse una distribución discreta —la de $\hat{p}$— por una continua —la normal.

El intervalo de confianza para el total de clase poblacional será, por tanto,

$$N \cdot I = \left[\hat{A} \mp \left(z_{\alpha/2} \sqrt{N(N-n) \cdot \frac{\hat{p}\,\hat{q}}{n-1}} + \frac{N}{2n} \right) \right]$$

que para los datos del problema resulta igual a

$$\left[1.200 \mp \left(1'96 \sqrt{2.000 \cdot 1.800 \cdot \frac{0'6 \cdot 0'4}{199}} + \frac{2.000}{2 \cdot 200} \right) \right] = [1.065'85\ ,\ 1.334'15].$$

Problema 1.17.- Con objeto de estimar la diferencia de estatura a lo largo del día en una población de tamaño $N = 100$, se seleccionaron al azar 10 individuos sin reemplazamiento y probabilidades iguales, a los que se les midió la estatura (en cm.) por la mañana al levantarse, X_i, y por la noche antes de acostarse, Y_i, obteniéndose los siguientes datos,

X_i	169'7	168'5	165'9	177'8	179'6	168'9	169'2	167'9	181'8	163'3
Y_i	168'2	165'5	164'4	175'7	176'6	166'1	167'1	166'3	179'7	161'5

Admitiendo una distribución normal para la estatura, determinar un intervalo de confianza para la diferencia media de estaturas, con un coeficiente de confianza del 0'95.

La variable de interés es la diferencia de estaturas $D = X - Y$, y al no ser independientes X e Y, tendremos un caso de datos apareados.

Al suponerse que la estatura —tanto por la mañana como por la noche— siguen distribuciones normales, $N(\overline{X}, \sigma_x)$ y $N(\overline{Y}, \sigma_y)$, la variable diferencia $D = X - Y$ seguirá una distribución normal de parámetros,

$$N\left(\overline{X} - \overline{Y}\ ,\ \sqrt{\sigma_x^2 + \sigma_y^2 - 2\,\rho_{xy}\,\sigma_x\,\sigma_y}\ \right)$$

o brevemente, $N(\overline{D}, \sigma_d)$.

El objetivo es determinar un intervalo de confianza para $\overline{D}$, basado en una muestra aleatoria simple de D.

Como la media muestral —de las diferencias— sigue una distribución normal de parámetros,

$$\overline{d} \rightsquigarrow N\left(\overline{D}\,,\,\sqrt{V(\overline{d})}\right) = N\left(\overline{D}\,,\,\sqrt{\frac{N-n}{N}}\cdot\frac{S_d}{\sqrt{n}}\right)$$

si la cuasivarianza poblacional de las diferencias, S_d^2 fuera conocida, el intervalo de confianza buscado sería

$$\left[\overline{d} \mp z_{\alpha/2}\,\frac{S_d}{\sqrt{n}}\,\sqrt{\frac{N-n}{N}}\right]$$

pero como, en nuestro caso, S_d es desconocida habrá que estimarla y utilizar una distribución t de Student, siendo el intervalo de confianza buscado,

$$\left[\overline{d} \mp t_{n-1;\alpha/2}\,\frac{\hat{S}_d}{\sqrt{n}}\,\sqrt{\frac{N-n}{N}}\right]$$

con $\hat{S}_d$ la cuasidesviación típica muestral (de las diferencias).

Y ésto porque el tamaño muestral es pequeño. Si fuera suficientemente grande, en el intervalo anterior podríamos sustituir los valores $t_{n-1;\alpha/2}$ por $z_{\alpha/2}$, o seguir buscándolos en una tabla de la t de Student si ésta es lo suficientemente precisa.

De los datos del problema se obtienen los valores $\overline{d} = 2'15$ y $\hat{S}_d^2 = 0'349$, con los que se obtiene el intervalo de confianza de coeficiente de confianza $0'95$,

$$\left[2'15 \mp 2'262 \cdot \sqrt{\frac{0'349}{10} \cdot \frac{90}{100}}\right] = .'75\ ,\ 2'55].$$

Problema 1.18.- En un distrito de 4.000 viviendas se desea estimar el porcentaje de familias propietarias de su vivienda, con un error de muestreo menor del 2%, así como el porcentaje de familias con dos coches, con un error de muestreo no mayor del 1%. Se cree que el verdadero porcentaje de propietarios se encuentra entre el 45 y el 65 % y que el porcentaje de familias con dos vehículos está entre el 5 y el 10%. ¿Cuál debe ser el tamaño muestral para que se satisfagan ambas condiciones?.

Llamando P_1 y P_2 a los porcentajes poblacionales, respectivamente, de familias propietarias de su vivienda y de familas con dos coches, las condiciones exigidas en el enunciado son

$$\sqrt{V(\hat{p}_1)} < 0'02$$

$$\sqrt{V(\hat{p}_2)} \leq 0'01$$

siendo $\hat{p}_1$ y $\hat{p}_2$ las correspondientes proporciones muestrales. Es decir,

$$\frac{N-n}{N-1}\frac{P_1 Q_1}{n} < 0'02^2$$

$$\frac{N-n}{N-1}\frac{P_2 Q_2}{n} \leq 0'01^2$$

Como la función $f(P) = P(1-P)$ alcanza su máximo en $P = 0'5$, la información adicional sobre P_1 y P_2, hace que el caso más desfavorable (el de mayor varianza) sea en la primera estimación $P_1 = 0'5$, al ser $0'45 < P_1 < 0'65$ y estar, por tanto, el máximo entre los valores posibles de la variable. En dicho máximo la función toma el valor $P_1 Q_1 = 1/4$.

Para la segunda condición, al suponerse que es $0'05 < P_2 < 0'1$, el máximo de la función $P_2(1-P_2)$, y por tanto de la varianza, se alcanzará en el extremo, $P_2 = 0'1$, en donde la función toma el valor $P_2\, Q_2 = 0'09$.

Las ecuaciones anteriores quedan, por tanto, de la forma

$$\frac{4.000-n}{3.999}\,\frac{1}{4\,n} < 0'0004$$

$$\frac{4.000-n}{3.999}\,\frac{0'09}{n} \leq 0'0001.$$

De la primera ecuación se obtiene la condición $n > 540'657$ y de la segunda $n \geq 734'844$. Un tamaño muestral de $n = 735$ unidades cumplirá ambas condiciones, aún en las peores situaciones: $P_1 = 0'5$ y $P_2 = 0'1$.

Problema 1.19.- Se desea realizar un estudio sobre enfermedades comunes en grandes poblaciones. Para una enfermedad que afecta, al menos, al 1% de los individuos de la población, se desea estimar el número total de casos, con un coeficiente de variación no mayor del 20%.

(a) Utilizando un muestreo aleatorio simple, ¿qué tamaño muestral debe elegirse?.

(b) ¿Cuál sería el tamaño muestral a elegir si la estimación del número total de casos se realizase separadamente para hombres y para mujeres, y se desease la misma precisión anterior supuesto que la enfermedad afecta de igual manera a ambos colectivos?.

(a) Ignorando el factor de corrección por ser grandes poblaciones, la varianza del total de clase es

$$V(\hat{A}) = N^2\,\frac{P\,Q}{n}$$

con lo que la condición exigida por el enunciado será

$$C.V.(\hat{A}) = \frac{\sqrt{V(\hat{A})}}{A} = \frac{\sqrt{N^2\,P\,Q}}{N\,P\,\sqrt{n}} \leq 0'2$$

es decir,

$$\frac{1-P}{P\,n} \leq 0'04$$

o bien,

$$n \geq 25\,\frac{1-P}{P}.$$

Como la función $f(P) = (1-P)/P$ es decreciente y debe ser $P \geq 0'01$, la peor de la situaciones será aquella en la que es $P = 0'01$, para la que se obtiene un tamaño muestral

$$n \geq 25 \frac{0'99}{0'01} = 2.475$$

Tomando un tamaño de muestra igual a 2.475 cumpliremos las condiciones del enunciado.

(b) Si la estimación del total de clase se realizase separadamente para hombres y para mujeres manteniendo la misma precisión, las condiciones requeridas serían

$$\frac{1-P_1}{P_1\, n_1} \leq 0'04$$

$$\frac{1-P_2}{P_2\, n_2} \leq 0'04$$

de donde, como antes, se obtendrían las soluciones $n_1 = 2.475$ hombres y $n_2 = 2.475$ mujeres y, por tanto, un tamaño muestral de $n = 4.950$ individuos.

Problema 1.20.- Se eligió una muestra aleatoria simple de 30 hogares de una ciudad que contenía 14.848 hogares. El número de personas por casa en la muestra fue:

$$5,6,3,3,2,3,3,3,4,4,3,2,7,4,3,5,4,4,3,3,4,3,3,1,2,4,3,4,2,4$$

Estimar el número total de personas de la ciudad y calcular la probabilidad de que este estimador esté dentro del $\pm 10\%$ del verdadero valor.

El estimador del total poblacional es el total muestral

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{n} X_i = N\, \overline{x} = 51.473'066.$$

Por otro lado, la probabilidad que se pide es

$$P\{|\hat{X} - X| < 0'1\, X\}.$$

Suponiendo que el número de personas por casa siga una distribución normal, será

$$\hat{X} \rightsquigarrow \text{Normal}\left(X\ ,\ N\sqrt{\frac{N-n}{N}}\ \frac{\hat{S}}{\sqrt{n}}\right)$$

con lo que la probabilidad pedida será igual a

$$P\left\{|Z| < \frac{0'1\,X}{N\sqrt{\frac{N-n}{N}}\ \frac{\hat{S}}{\sqrt{n}}}\right\}$$

con $Z \rightsquigarrow N(0,1)$. Como el total muestral X es desconocido, lo reemplazaremos en la expresión anterior por su estimador $\hat{X}$, el cual como vimos, para los datos del ejemplo toma el valor 51.473'066.

Por otro lado, de esos mismos datos se obtiene que es $\hat{S}^2 = 1'498851$, con lo que en definitiva, la probabilidad pedida será igual a

$$P\left\{|Z| < \frac{0'1 \cdot 51.473'066}{14.848 \cdot \sqrt{\frac{14.848-30}{14.848}\ .\ \frac{1'498851}{30}}}\right\} = P\{|Z| < 1'5525\} = 0'8788.$$

Problema 1.21.- Se quiere estimar el número medio de gusanos, $\overline{X}$, en un campo de dimensiones suficientemente grandes como para poder prescindir del factor de corrección para poblaciones finitas.

Para ello se utilizó un instrumento muestral de 9 × 9 pulgadas en superficie ×5 pulgadas en profundidad, observándose la variable X_i, *número de gusanos en la muestra i-ésima*, $i = 1,...,n$, de media $\overline{X}$ y cuasivarianza $S^2 = 1'2\,\overline{X}$.

Sabiendo que en un sondeo anterior, el número medio de gusanos en un volumen de tierra de 1 acre de extensión y 5 pulgadas de profundidad fue de 200.000, determinar el menor tamaño de muestra necesario para que la media muestral $\overline{x}$ estime $\overline{X}$ con un error menor del 30%, con probabilidad 0'95.

Nota: *Datos adicionales: 1 acre= 43.560 pies cuadrados. 1 pie = 12 pulgadas.*

La condición requerida por el enunciado es

$$P\{|\overline{x} - \overline{X}| < 0'3\,\overline{X}\} = 0'95.$$

Es decir,

$$P\left\{|Z| < \frac{0'3\,\overline{X}}{\sqrt{V(\overline{x})}}\right\} = 0'95$$

de donde se obtiene que debe de ser

$$0'09\,\overline{X}^2 = 1'96^2 \cdot V(\overline{x}) = 1'96^2 \cdot \frac{S^2}{n} = 1'96^2 \cdot \frac{1'2\,\overline{X}}{n}$$

es decir,

$$n = \frac{4'6099}{0'09\,\overline{X}}.$$

Como la profundidad es siempre de 0'5 pulgadas, podemos omitirla en la estimación de $\overline{X}$. El enunciado nos dice que hay aproximadamente 200.000 gusanos en un acre, es decir, en 43.560 pies cuadrados, o equivalentemente en 6.272.640 pulgadas al cuadrado. Como la unidad muestral tiene 81 pulgadas al cuadrado en superficie, cabe esperar que se recoja una media de $81 \cdot (200.000/6.272.640) = 2'5826$ gusanos en cada unidad muestral.

Por tanto, el tamaño muestral requerido será

$$n = \frac{4'6099}{0'09 \cdot 2'5826} = 19'83.$$

Un tamaño de 20 unidades muestrales bastará para obtener la precisión deseada.

Problema 1.22.- En una comunidad americana, de aproximadamente 50.000 habitantes, se desea determinar el menor tamaño de muestra necesario para estimar, con probabilidad 0'95, la proporción de habitantes de dicha ciudad sin cobertura de la Sanidad Pública con un error menor del 15% de la verdadera proporción, suponiendo que ésta está entre el 10% y el 20%.

La condición expresada por el enunciado es

$$P\left\{|\hat{p} - P| < 0'15P\right\} = 0'95.$$

Es decir,

$$P\left\{|Z| < \frac{0'15\,P}{\sqrt{V(\hat{p})}}\right\} = 0'95.$$

De donde, por las tablas de una normal $N(0,1)$, debe ser

$$0'15^2\,P^2 = \frac{N-n}{N-1}\,\frac{P(1-P)}{n}\,1'96^2$$

es decir,

$$0'15^2\,P^2 = \frac{50.000\,P(1-P)\,1'96^2}{49.999\,n} - \frac{P(1-P)\,1'96^2}{49.999}$$

de donde se obtiene que debe ser

$$n = \frac{192.080\,(1-P)}{3'8416 + 1.121'1359\,P}.$$

Ésta es una función decreciente de P, por lo que, para el rango de posibles valores de la variable, la peor situación se tiene cuando P sea 0'1, ya que es el valor para el que se obtiene un mayor tamaño muestral. Por tanto, para estar seguros de que se cumple la condición exigida por el enunciado, deberemos tomar un tamaño muestral igual a

$$n = \frac{192.080\,(1-0'1)}{3'8416 + 1.121'1359\,0'1} = 1.490'85.$$

Por tanto, un tamaño de muestra de $n = 1.491$ unidades será suficiente.

Problema 1.23.- En una población de 1000 habitantes se desea determinar el menor tamaño de muestra necesario para estimar, con probabilidad 0'95, la proporción de votantes a un determinado partido político, con un error menor del 10% de la verdadera proporción, suponiendo que ésta está entre el 40% y el 65%.

La condición expresada por el enunciado es

$$P\{|\hat{p} - P| < 0'1P\} = 0'95.$$

Es decir,

$$P\left\{|Z| < \frac{0'1\,P}{\sqrt{V(\hat{p})}}\right\} = 0'95.$$

De donde, por las tablas de una normal $N(0,1)$, debe ser

$$0'1^2\, P^2 = \frac{N-n}{N-1}\,\frac{P(1-P)}{n}\,1'96^2$$

de donde se obtiene que debe ser

$$n = \frac{3'845\,(1-P)}{0'0038 + 0'0062\,P}.$$

Ésta es una función decreciente de P, por lo que, para el rango de posibles valores de la variable, la peor situación se tiene cuando P sea $0'4$, ya que es el valor para el que se obtiene un mayor tamaño muestral. Por tanto, para estar seguros de que se cumple la condición exigida por el enunciado, deberemos tomar un tamaño muestral igual a

$$n = \frac{3'845\,(1-0'4)}{0'0038 + 0'0062\,0'4} = 367'36.$$

Por tanto, un tamaño de muestra de $n = 368$ unidades será suficiente.

CAPÍTULO 2

Muestreo estratificado

El muestreo estratificado se utiliza cuando se tiene una población de tamaño N dividida en L grupos de tamaño N_h, $h = 1, ...L$, denominados *estratos*, bastante homogéneos respecto, a la característica en estudio, de forma que, con objeto de estimar algún parámetro poblacional, se seleccionan, mediante un muestreo aleatorio simple, n_h individuos del estrato h, constituyendo, de esta manera, la muestra estratificada de tamaño $n = \sum_{h=1}^{L} n_h$. El cociente $W_h = N_h/N$ se denomina *peso* o *tamaño relativo del estrato h*.

Las principales características poblacionales en los estratos son

media del estrato h	$\overline{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} X_{hi}$
total del estrato h	$X_h = \sum_{i=1}^{N_h} X_{hi} = N_h \, \overline{X}_h$
varianza del estrato h	$\sigma_h^2 = \frac{1}{N_h} \sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2$
cuasivarianza del estrato h	$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (X_{hi} - \overline{X}_h)^2 = \frac{N_h \, \sigma_h^2}{N_h - 1}$

proporción del estrato h	$P_h = \frac{1}{N_h} \sum_{i=1}^{N_h} A_{hi}$
total de clase del estrato h	$A_h = \sum_{i=1}^{N_h} A_{hi} = N_h \, P_h$

Y en la población estratificada,

media poblacional	$\overline{X} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi} = \frac{1}{N} \sum_{h=1}^{L} X_h = \sum_{h=1}^{L} W_h \overline{X}_h$
total poblacional	$X = \sum_{h=1}^{L} \sum_{i=1}^{N_h} X_{hi} = \sum_{h=1}^{L} X_h = N \, \overline{X}$
varianza poblacional	$\sigma^2 = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \overline{X})^2$
cuasivarianza poblacional	$S^2 = \frac{1}{N-1} \sum_{h=1}^{L} \sum_{i=1}^{N_h} (X_{hi} - \overline{X})^2 = \frac{N \, \sigma^2}{N-1}$
proporción poblacional	$P = \sum_{h=1}^{L} W_h \, P_h$
total de clase poblacional	$A = N \, P$

Como dentro del estrato h se realiza un muestreo aleatorio simple de tamaño n_h, la media muestral observada en dicho estrato

$$\overline{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi}$$

será un estimador insesgado de $\overline{X}_h$, el total muestral del estrato h, $\hat{X}_h$, lo será de X_h, etc.; con lo que los estimadores insesgados de los parámetros poblacionales serán

media muestral estratificada	$\overline{x}_{st} = \sum_{h=1}^{L} W_h \overline{x}_h$
total muestral estratificado	$\hat{X}_{st} = N \overline{x}_{st}$
proporción muestral estratificada	$\hat{P}_{st} = \sum_{h=1}^{L} W_h \hat{p}_h$
total de clase muestral estratificado	$\hat{A}_{st} = N \hat{P}_{st}$

de varianzas respectivas

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{S_h^2}{n_h}$$

$$V(\hat{X}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h) \frac{S_h^2}{n_h}$$

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

$$V(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h - 1} \frac{P_h Q_h}{n_h}$$

las cuales se estiman, respectivamente, por los estimadores

$$\hat{V}(\overline{x}_{st}) = \sum_{h=1}^{L} \frac{N_h(N_h - n_h)}{N^2} \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^{L} N_h(N_h - n_h) \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{P}_{st}) = \sum_{h=1}^{L} \frac{N_h}{N^2} \frac{N_h - n_h}{n_h - 1} \hat{p}_h \hat{q}_h$$

$$\hat{V}(\hat{A}_{st}) = \sum_{h=1}^{L} N_h \frac{N_h - n_h}{n_h - 1} \hat{p}_h \hat{q}_h$$

Si los estratos son de gran tamaño, se puede omitir el factor de corrección para poblaciones finitas, haciendo en todas las fórmulas anteriores $1 - \frac{n_h}{N_h} \approx 1$, y quedando

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}$$

$$V(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{S_h^2}{n_h}$$

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{P_h Q_h}{n_h}$$

$$V(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{P_h Q_h}{n_h}$$

$$\hat{V}(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{\hat{S}_h^2}{n_h}$$

$$\hat{V}(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\hat{p}_h \hat{q}_h}{n_h - 1}$$

$$\hat{V}(\hat{A}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{\hat{p}_h \hat{q}_h}{n_h - 1}$$

Un problema clásico en el muestreo estratificado es el de la *afijación*; es decir, el del reparto de los n individuos muestrales a seleccionar entre los L estratos.

Una posibilidad es la denominada *Afijación igual*, la cual consiste en tomar

$$n_h = \frac{n}{L} \qquad h = 1, ..., L.$$

Otra posibilidad es la *Afijación proporcional*, consistente en asignar a cada estrato un tamaño muestral proporcional a su tamaño; es decir,

$$n_h = \frac{N_h}{N} \cdot n.$$

Cuando se supone este tipo de afijación, los estimadores habituales tienen, respectivamente, como varianzas y varianzas aproximadas para grandes estratos, las siguientes

$$V(\overline{x}_{ap}) = \frac{N-n}{N \cdot n} \sum_{h=1}^{L} W_h \, S_h^2 \approx \frac{1}{n} \sum_{h=1}^{L} W_h \, S_h^2$$

$$V(\hat{X}_{ap}) = \frac{N(N-n)}{n} \sum_{h=1}^{L} W_h \, S_h^2 \approx \frac{N^2}{n} \sum_{h=1}^{L} W_h \, S_h^2$$

estimadas por

$$\hat{V}(\overline{x}_{ap}) = \frac{N-n}{N \cdot n} \sum_{h=1}^{L} W_h \, \hat{S}_h^2 \approx \frac{1}{n} \sum_{h=1}^{L} W_h \, \hat{S}_h^2$$

$$\hat{V}(\hat{X}_{ap}) = \frac{N(N-n)}{n} \sum_{h=1}^{L} W_h \, \hat{S}_h^2 \approx \frac{N^2}{n} \sum_{h=1}^{L} W_h \, \hat{S}_h^2$$

Otro tipo de afijación es la *Afijación óptima*, mediante la cual se determinan los n_h que minimizan la varianza del estimador en estudio, sujeta a la restricción de ser $\sum_{h=1}^{L} n_h = n$; en concreto, si se trata de estimar la media, se trata de minimizar la función

$$\phi = \sum_{h=1}^{L} \frac{N_h \, (N_h - n_h)}{N^2} \, \frac{S_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} n_h - n \right)$$

con la que se obtienen las afijaciones

$$n_h = \frac{N_h \, S_h}{\sum_h N_h \, S_h} \cdot n$$

y una expresión para la varianza y su aproximación para grandes estratos de la media

$$V(\overline{x}_{ao}) = \frac{(\sum_h W_h \, S_h)^2}{n} - \frac{\sum_h W_h \, S_h^2}{N} \approx \frac{(\sum_h W_h \, S_h)^2}{n}.$$

Si existe un coste de observación, c_h, por unidad muestral en el estrato h, y el coste total C es fijo, se utiliza la *Afijación óptima para costes variables*, en donde, para la estimación de la media, se minimiza la función

$$\phi = \sum_{h=1}^{L} \frac{N_h\,(N_h - n_h)}{N^2}\,\frac{S_h^2}{n_h} + \lambda\left(\sum_{h=1}^{L} c_h\,n_h - C\right)$$

proporcionando como afijaciones

$$n_h = \frac{N_h\,S_h/\sqrt{c_h}}{\sum_h (N_h\,S_h/\sqrt{c_h})}\cdot n$$

Problema 2.1.- Para un estudio socio-económico, se agrupó en estratos a todas las localidades de una región, incluyendo las deshabitadas, utilizando como criterios la altitud sobre el nivel del mar y la densidad de población. En cada estrato se seleccionó mediante un muestreo aleatorio con reemplazamiento, a 10 unidades. Los datos sobre el número de familias residentes en las localidades observadas fueron los siguientes:

Estrato	nº de	nº de familias residentes									
	locali.	1	2	3	4	5	6	7	8	9	10
I	1411	43	84	98	0	10	44	0	124	13	0
II	4705	50	147	62	87	84	158	170	104	56	160
III	2558	228	262	110	232	139	178	334	0	63	220
IV	14997	17	34	25	34	36	0	25	7	15	31

Se pide:

(a) Obtener una estimación del número total de familias y de su error de muestreo.

(b) Estimar la ganancia debida al uso de estratificación.

(c) Comparar la eficiencia de la afijación considerada con la óptima, manteniendo constante el tamaño total de la muestra.

A partir de la tabla del enunciado se obtienen los valores

$$N_1 = 1411 \quad \overline{x}_1 = 41'6 \quad \hat{S}_1^2 = 2087'156 \quad n_1 = 10$$

$$N_2 = 4705 \quad \overline{x}_2 = 107'8 \quad \hat{S}_2^2 = 2198'4 \quad n_2 = 10$$

$$N_3 = 2558 \quad \overline{x}_3 = 176'6 \quad \hat{S}_3^2 = 9956'267 \quad n_3 = 10$$

$$N_4 = 14997 \quad \overline{x}_4 = 22'4 \quad \hat{S}_4^2 = 151'6 \quad n_4 = 10$$

$$N = 23671 \quad \hat{S}^2 = 7089'528$$

correspondientes, respectivamente, a los tamaños, medias muestrales, cuasivarianzas muestrales y tamaños muestrales de los cuatro estratos, así como el tamaño de la población y cuasivarianza muestral correspondiente a una muestra aleatoria simple de tamaño $n = 40$.

(a) El estimador del total en un muestreo estratificado es

$$\hat{X}_{st} = N \cdot \overline{x}_{st}$$

siendo

$$\overline{x}_{st} = \sum_{h=1}^{L} \frac{N_h}{N} \overline{x}_h = \sum_{h=1}^{L} W_h \overline{x}_h \qquad \text{y} \qquad \overline{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} X_{hi}.$$

La estimación de su varianza es

$$\hat{V}(\hat{X}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{\hat{S}_h^2}{n_h} \qquad \text{con} \qquad \hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (X_{hi} - \overline{x}_h)^2.$$

Así pues, la media muestral estratificada será

$$\overline{x}_{st} = \frac{1411}{23671} \cdot 41'6 + \frac{4705}{23671} \cdot 107'8 + \frac{2558}{23671} \cdot 176'6 + \frac{14997}{23671} \cdot 22'4 = 57'182721$$

y la estimación del número total de familias,

$$\hat{X}_{st} = 23671 \cdot 57'182721 = 1.353.572'1.$$

Su varianza estimada será,

$$\hat{V}(\hat{X}_{st}) = \frac{1411^2 \cdot 2087'156}{10} + \frac{4705^2 \cdot 2198'4}{10} + \frac{2558^2 \cdot 9956'267}{10} +$$

$$+ \frac{14997^2 \cdot 151'6}{10} = 15.206.523.489'7264$$

y por tanto, su error de muestreo estimado,

$$\sqrt{\hat{V}(\hat{X}_{st})} = \sqrt{15.206.523.489'7264} = 108.613'47869.$$

(b) Como el muestreo es con reemplazamiento dentro de los estratos, se puede ignorar el factor de correción, con lo que será

$$\hat{V}(\hat{X}_{as}) = N^2 \cdot \hat{S}^2/n, \quad \text{con} \quad \hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{x})^2.$$

En nuestro problema es $\hat{V}(\hat{X}_{st}) = 15.206.523.489'7264$ y $\hat{V}(\hat{X}_{as}) = 99.309.441.985'6062$, con lo que la ganancia relativa debido al uso de la estratificación será,

$$\frac{\hat{V}(\hat{X}_{as})}{\hat{V}(\hat{X}_{st})} = 6'53071.$$

(c) La afijación óptima, en un muestreo con reemplazamiento dentro de los estratos, dará como varianza estimada para el estimador del total un valor de

$$\hat{V}(\hat{X}_{ao}) = \frac{1}{n}(\sum_{h=1}^{L} N_h \hat{S}_h)^2 = \frac{724.957'89^2}{40} = 13.139.098.556'8313$$

y términos relativos,

$$\frac{\hat{V}(\hat{X}_{st})}{\hat{V}(\hat{X}_{ao})} = 1'1573491.$$

Problema 2.2.- Se clasificaron las factorías de una ciudad de acuerdo con su número de trabajadores. Con posterioridad, se extrajo una muestra estratificada de 3.000 factorías cuyos resultados, junto con los datos poblacionales, fueron los siguientes

Estrato h	nº de trabajadores	nº de factorías	S_h	Produc. por factoría
I	1-49	18260	80	100
II	50-99	4315	200	250
III	100-249	2233	600	500
IV	250-999	1057	1900	1760
V	1000 o más	567	2500	2250

Con objeto de estimar la producción total, compare las eficiencias de las afijaciones, Proporcional, Proporcional a la producción y Óptima.

Supondremos, como es habitual, dentro de cada estrato un muestreo aleatorio simple.

(a) Bajo afijación proporcional, la varianza del estimador del total es,

$$V(\hat{X}_{ap}) = \frac{N-n}{n} \sum_{h=1}^{L} N_h S_h^2$$

con S_h^2 la cuasivarianza poblacional del estrato h.

Es decir, en nuestro ejemplo,

$$V(\hat{X}_{ap}) = \frac{26.432 - 3.000}{3.000}(18.260 \cdot 80^2 + 4.315 \cdot 200^2 + 2.233 \cdot 600^2 +$$

$$+1.057 \cdot 1.900^2 + 567 \cdot 2.500^2) = \frac{23.432}{3.000} \cdot 8.452.864.000 = 66.022.503.082'667$$

(b) En este caso, la afijación tendrá que determinarse por la expresión

$$\frac{n_h}{n} = \frac{\text{Produc. estrato } h}{\text{Produc. total}} = \frac{P^h}{P^t}$$

en donde la última igualdad se entiende como notación. Será pues,

$$n_h = \frac{P^h}{P^t} \cdot n$$

siendo, por tanto, la varianza del estimador del total,

$$V(\hat{X}_{app}) = \sum_{h=1}^{L} N_h(N_h - n_h)\frac{S_h^2}{n_h} = \sum_{h=1}^{L} \frac{N_h^2 \cdot S_h^2}{n_h} - \sum_{h=1}^{L} N_h \cdot S_h^2 =$$

$$= \frac{P^t}{n} \sum_{h=1}^{L} \frac{N_h^2 \cdot S_h^2}{P^h} - \sum_{h=1}^{L} N_h \cdot S_h^2 = 41.918.161.360'441.$$

(c) Respecto a la afijación óptima, la varianza del estimador del total será,

$$V(\hat{X}_{ao}) = \frac{(\sum_{h=1}^{L} N_h \cdot S_h)^2}{n} - \sum_{h=1}^{L} N_h \cdot S_h^2 = 8.300.333.453'33.$$

Las eficiencias, medidas en términos relativos, serán

- De la afijación proporcional respecto de la afijación proporcional a la producción,

$$\frac{V(\hat{X}_{ap})}{V(\hat{X}_{app})} = \frac{66.022.503.082'667}{41.918.161.360'441} = 1'57503.$$

- De la afijación proporcional respecto de la afijación óptima,

$$\frac{V(\hat{X}_{ap})}{V(\hat{X}_{ao})} = \frac{66.022.503.082'667}{8.300.333.453'33} = 7'9542.$$

- De la afijación proporcional a la producción total respecto de la afijación óptima,

$$\frac{V(\hat{X}_{app})}{V(\hat{X}_{ao})} = \frac{41.918.161.360'441}{8.300.333.453'33} = 5'05018.$$

Problema 2.3.- Se hizo un estudio dirigido a estimar el número total de literatos de una ciudad habitada por tres comunidades. Un trabajo piloto dió los siguientes resultados:

Comunidad	nº total de personas	Porcentaje de literatos
1	60.000	40
2	10.000	80
3	30.000	60

Se pide:

(a) Tratando las comunidades como estratos, asigne una muestra de 2000 personas a los estratos de forma óptima, para estimar la proporción de literatos en la ciudad.

(b) Estime la eficiencia de la estratificación comparada con un muestreo aleatorio simple no estratificado.

(a) La fórmula de la afijación óptima a utilizar es

$$n_h = \frac{N_h\sqrt{P_h \cdot Q_h}}{\sum_h N_h\sqrt{P_h \cdot Q_h}} \cdot n$$

que nos da unos tamaños muestrales para cada estrato de

$$n_1 = \frac{60.000\sqrt{0'4 \cdot 0'6}}{60.000\sqrt{0'4 \cdot 0'6} + 10.000\sqrt{0'8 \cdot 0'2} + 30.000\sqrt{0'6 \cdot 0'4}} \cdot 2000 = 1222'4 \approx 1.223$$

$$n_2 = \frac{10.000\sqrt{0'8 \cdot 0'2}}{60.000\sqrt{0'4 \cdot 0'6} + 10.000\sqrt{0'8 \cdot 0'2} + 30.000\sqrt{0'6 \cdot 0'4}} \cdot 2000 = 166'35 \approx 166$$

$$n_3 = \frac{30.000\sqrt{0'6 \cdot 0'4}}{60.000\sqrt{0'4 \cdot 0'6} + 10.000\sqrt{0'8 \cdot 0'2} + 30.000\sqrt{0'6 \cdot 0'4}} \cdot 2000 = 611'2 \approx 611$$

(b) La varianza del estimador de la proporción para el muestreo estratificado es

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_h \cdot Q_h}{n_h} = 0'0001133.$$

Respecto a un muestreo aleatorio simple, ésta sería

$$V(\hat{p}_{as}) = \frac{N - n}{N - 1} \cdot \frac{P \cdot Q}{n} = \frac{100.000 - 2.000}{99.999} \cdot \frac{0'5 \cdot 0'5}{2.000} = 0'0001225$$

al ser la proporción poblacional

$$P = \sum_{h=1}^{L} W_h P_h = 0'6 \cdot 0'4 + 0'1 \cdot 0'8 + 0'3 \cdot 0'6 = 0'5.$$

Así pues, la eficiencia, medida en términos relativos, será

$$\frac{V(\hat{p}_{as})}{V(\hat{P}_{st})} = \frac{0'0001225}{0'0001133} = 1'081.$$

Problema 2.4.- Sea una población de tamaño N, con estratos de tamaño N_h.

(a) Si la función de coste es de la forma $C = c_0 + \sum_h t_h\sqrt{n_h}$ en donde c_0 y t_h son números conocidos, demostrar que, en orden a minimizar $V(\overline{x}_{st})$ para un coste total fijo, n_h debe ser proporcional a $(W_h^2 \cdot S_h^2/t_h)^{2/3}$, en donde $W_h = N_h/N$.

(b) Encontrar, utilizando el resultado obtenido en (a), los valores de n_h para una muestra de tamaño 1.000 siendo

Estrato	I	II	III
W_h	0'4	0'3	0'3
S_h	4	5	6
t_h	1	2	4

(c) Si $C = 500$ y $c_0 = 300$, determinar el tamaño muestral y las afijaciones en este caso, utilizando (a) y (b).

(d) Supuesto que n=2.969, determinar la afijación proporcional y comparar los resultados obtenidos con los del apartado anterior, utilizando $V(\overline{x}_{st})$.

(a) La función a minimizar, es decir la $V(\overline{x}_{st})$ más la restricción igualada a cero, será

$$\Phi = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2 \frac{S_h^2}{n_h} + \lambda \left(c_0 + \sum_{h=1}^{L} t_h \sqrt{n_h} - C\right)$$

Como es

$$\frac{\partial \Phi}{\partial n_h} = -\frac{W_h^2 S_h^2}{n_h^2} + \lambda\, t_h \frac{1}{2\sqrt{n_h}} = 0 \qquad \forall h = 1, ..., L$$

será

$$\lambda = \frac{W_h^2\, S_h^2\, 2\sqrt{n_h}}{n_h^2\, t_h} \qquad \forall h = 1, ..., L$$

es decir,

$$\lambda^{2/3} = 2^{2/3} \frac{(W_h^2\, S_h^2/t_h)^{2/3}}{n_h} \qquad \forall h = 1, ..., L$$

con lo que será

$$2^{2/3} \frac{(W_h^2\, S_h^2/t_h)^{2/3}}{n_h} = 2^{2/3} \frac{\sum_{h=1}^{L} (W_h^2\, S_h^2/t_h)^{2/3}}{\sum_{h=1}^{L} n_h}$$

de donde se obtiene en definitiva,

$$n_h = \frac{(W_h^2\, S_h^2/t_h)^{2/3}}{\sum_{h=1}^{L} (W_h^2\, S_h^2/t_h)^{2/3}} \cdot n$$

es decir, proporcional a

$$\left(W_h^2\, S_h^2/t_h\right)^{2/3}$$

(b) Como es

W_h	S_h	t_h	$W_h^2\, S_h^2/t_h$
0'4	4	1	2'56
0'3	5	2	1'125
0'3	6	4	0'81

será

$$n_1 = \frac{\left(W_1^2\, S_1^2/t_1\right)^{2/3}}{\sum_{h=1}^{L}\left(W_h^2\, S_h^2/t_h\right)^{2/3}} \cdot n = \frac{1'87137}{3'822} \cdot 1.000 = 489'63 \approx 490.$$

$$n_2 = \frac{\left(W_2^2\, S_2^2/t_2\right)^{2/3}}{\sum_{h=1}^{L}\left(W_h^2\, S_h^2/t_h\right)^{2/3}} \cdot n = \frac{1'08169}{3'822} \cdot 1.000 = 283'02 \approx 283.$$

$$n_3 = \frac{\left(W_3^2\, S_3^2/t_3\right)^{2/3}}{\sum_{h=1}^{L}\left(W_h^2\, S_h^2/t_h\right)^{2/3}} \cdot n = \frac{0'86894}{3'822} \cdot 1.000 = 227'35 \approx 227.$$

(c) Ahora es

$$n_1 = 0'48963 \cdot n$$

$$n_2 = 0'28302 \cdot n$$

$$n_3 = 0'22735 \cdot n$$

y como es

$$C = c_0 + \sum_{h=1}^{3} t_h \sqrt{n_h}$$

será

$$500 = 300 + \sqrt{0'48963}\,\sqrt{n} + 2\sqrt{0'28302}\,\sqrt{n} + 4\sqrt{0'22735}\,\sqrt{n}$$

de donde se obtiene $n = 2.968'22$, por lo que tomaremos $n = 2.969$.

Las afijaciones serán, por tanto,

$$n_1 = 0'48963 \cdot 2.969 = 1.453'7 \approx 1.454$$

$$n_2 = 0'28302 \cdot 2.969 = 840'3 \approx 840$$

$$n_3 = 0'22735 \cdot 2.969 = 675$$

(d) De la afijación proporcional se obtienen los valores,

$$n_1 = W_1 \cdot n = 0'4 \cdot 2.969 = 1.187'6 \approx 1.187$$

$$n_2 = W_2 \cdot n = 0'3 \cdot 2.969 = 890'7 \approx 891$$

$$n_3 = W_3 \cdot n = 0'3 \cdot 2.969 = 890'7 \approx 891$$

Calcularemos la varianza $V(\overline{x}_{st})$ directamente, utilizando la expresión

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2 \frac{S_h^2}{n_h} \approx \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h}$$

que para las afijaciones determinadas en el apartado anterior resulta igual a $0'0092392$ y que en el caso de afijación proporcional toma el valor $0'0083183$.

Problema 2.5.- En una población de 6 unidades se obtuvo una muestra aleatoria simple de dos unidades. Posteriormente se agregaron las unidades en dos estratos tomando la variable en estudio los valores

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

obteniéndose una muestra estratificada con factor de corrección para poblaciones finitas $f_h = n_h/N_h = 1/3$. Se pide:

(a) V($\overline{x}$).

(b) V($\overline{x}_{st}$).

(c) ¿Cuáles serán los valores de las afijaciones que minimicen el coste total, sabiendo que el coste por unidad en el primer estrato es cuatro y dos en el segundo?.

(d) Determinar el tamaño muestral $n = n_1 + n_2$ para que sea V($\overline{x}_{st}$) = 0'2, utilizando las afijaciones calculadas en el apartado anterior.

(a) Sabemos que es

$$V(\overline{x}) = \frac{N-n}{N} \cdot \frac{S^2}{n}$$

Como es $S^2 = 11'2$, será

$$V(\overline{x}) = \frac{6-2}{6} \cdot \frac{11'2}{2} = 3'73.$$

(b) Sabemos que es

$$V(\overline{x}_{st}) = \sum_{h=1}^{2} W_h^2 \, \frac{N_h - n_h}{N_h} \, \frac{S_h^2}{n_h}$$

y como es $S_1^2 = 2'33$ y $S_2^2 = 4'33$, será $V(\overline{x}_{st}) = 1'11$.

(c) La expresión de la *afijación para costes variables* es

$$\omega_h = \frac{N_h \, S_h/\sqrt{c_h}}{\sum N_h \, S_h/\sqrt{c_h}}$$

que proporciona unos valores de $\omega_1 = 0'34$ y $\omega_2 = 0'66$.

(d) De la expresión

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 \, (1 - f_h) \, \frac{S_h^2}{n_h}$$

se obtiene que es

$$n = \frac{\sum_h W_h^2 S_h^2/\omega_h}{V(\bar{x}_{st}) + \sum_h W_h S_h^2/N}$$

con lo que sustituyendo los valores queda, $n = 4'44 \simeq 4$.

Problema 2.6.- Se quiere estimar el total de elementos pertenecientes a una cierta clase en una población finita constituida por dos estratos de tamaños $N_1 = 2.000$ y $N_2 = 6.000$. Para ello se decide tomar una muestra aleatoria con reemplazamiento, independientemente en cada uno de los dos estratos.

Se sabe que las proporciones en cada estrato son $P_1 = 0'3$ y $P_2 = 0'7$. Se pide:

1. Calcular la varianza del estimador del total de clase en los siguientes casos:
   (a) Afijación óptima para costes variables, siendo $c_1 = 100$ pts y $c_2 = 200$ pts los costes por unidad muestral en cada estrato y siendo $C = 120.000$ el presupuesto disponible.
   (b) Afijación óptima para un tamaño muestral total igual al que resulte en el caso anterior, suponiendo iguales los costes unitarios en ambos estratos. Determinar su coste.
   (c) Muestreo no estratificado, con un tamaño de muestra igual a los casos anteriores.
2. ¿Por qué la afijación proporcional coincide con la que resulta en el caso (b)?.
3. Determinar el tamaño de la muestra para que con afijación proporcional el error de muestreo relativo del estimador del total de clase sea del dos por ciento.

(1) El estimador del total de clase en el muestreo estratificado es

$$\hat{A}_{st} = N \cdot \hat{P}_{st} = N \sum_{h=1}^{2} W_h \hat{p}_h$$

en donde $\hat{p}_h$ es la proporción muestral del estrato h. La varianza de dicho estimador es en nuestro caso

$$V(\hat{A}_{st}) = N^2\, V(\hat{P}_{st}) = N^2 \sum_{h=1}^{2} W_h^2\, \frac{P_h\, Q_h}{n_h}.$$

(a) La afijación óptima consiste en determinar el tamaño de la muestra a seleccionar en el estrato h, es decir n_h, de forma que la $V(\hat{A}_{st})$ sea mínima, con la restricción de ser $\sum n_h\, c_h = C$.

Para ello consideremos la función

$$\phi = V(\hat{A}_{st}) + \lambda\Big(\sum n_h\, c_h - C\Big) =$$

$$= \sum_{h=1}^{2} N_h^2\, \frac{P_h\, Q_h}{n_h} + \lambda(100 \cdot n_1 + 200 \cdot n_2 - 120.000)$$

y resolvamos el sistema

$$\left.\begin{array}{l} \dfrac{\partial \phi}{\partial n_h} = 0 \\ \\ h = 1,2 \end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} \dfrac{-N_1^2\, P_1\, Q_1}{n_1^2} + 100\,\lambda = 0 \\ \\ \dfrac{-N_2^2\, P_2\, Q_2}{n_2^2} + 200\,\lambda = 0 \end{array}\right. \Longleftrightarrow$$

$$\frac{N_1^2\, P_1\, Q_1}{100\, n_1^2} = \frac{N_2^2\, P_2\, Q_2}{200\, n_2^2} \Longrightarrow \frac{n_2^2}{n_1^2} = \frac{N_2^2\, P_2\, Q_2\, 100}{N_1^2\, P_1\, Q_1\, 200}$$

$$\Longrightarrow n_2 = \frac{3\, n_1}{\sqrt{2}}.$$

Por otro lado, debe ser

$$100\, n_1 + \frac{600}{\sqrt{2}}\, n_1 = 120.000$$

es decir, $n_1 = 228'89 \simeq 229$. Por tanto será $n_2 = 3\, n_1/\sqrt{2} \simeq 486$ y $n = 715$.

En consecuencia, será

$$V(\hat{A}_{st}) = \sum_{h=1}^{2} \frac{N_h^2\, P_h\, Q_h}{n_h} = 19.223'68.$$

(b) Hemos visto en el apartado anterior que debía ser

$$\frac{n_2}{n_1} = \frac{N_2}{N_1} \cdot \sqrt{\frac{P_2 \cdot Q_2}{P_1 \cdot Q_1}} \cdot \sqrt{\frac{c_1}{c_2}}$$

es decir,

$$\frac{n_2}{n_1} = \frac{N_2}{N_1} \cdot \sqrt{\frac{c_1}{c_2}}$$

y como los costes unitarios se suponen iguales, (i.e., $c_1 = c_2$), será

$$\frac{n_2}{n_1} = \frac{N_2}{N_1} = 3$$

con lo que deberá ser

$$n_2 = 3 \cdot n_1$$

y por tanto

$$n_2 = 3(n - n_2) = 3(715 - n_2)$$

es decir, $n_2 = 536'25$ y en consecuencia $n_1 = 178'75$. Tomando $n_1 = 179$ y $n_2 = 536$, queda $V(\hat{A}_{st}) = 18.797'21$.

(c) En este caso es

$$V(\hat{A}_{st}) = N^2 \frac{PQ}{n}$$

siendo

$$P = \sum_{h=1}^{2} W_h P_h = \frac{2.000 \cdot 0'3 + 6.000 \cdot 0'7}{8.000} = 0'6$$

resultando, por tanto, $V(\hat{A}_{st}) = 21.482'52$.

(2) La respuesta es porque en el caso (b) era

$$\frac{n_2}{n_1} = \frac{W_2}{W_1}$$

es decir, afijación proporcional.

(3) Como en el caso de afijación proporcional es $n_1 = W_1 \cdot n$ y $n_2 = W_2 \cdot n$, será

$$V(\hat{A}_{st}) = N^2 \sum_{h=1}^{2} W_h^2 \frac{P_h Q_h}{n_h} = N^2 \sum_{h=1}^{2} W_h^2 \frac{P_h Q_h}{W_h n} =$$

$$= N \sum_{h=1}^{2} \frac{N_h P_h Q_h}{n} = \frac{0'21 \cdot 64 \cdot 10^6}{n}.$$

Por otro lado es $A = N P = 8.000 \cdot 0'6 = 4.800$, con lo que el error de muestreo relativo del estimador del total de clase al cuadrado es

$$\frac{V(\hat{A}_{st})}{A^2} = \frac{0'21 \cdot 64 \cdot 10^6}{n \cdot (4.800)^2} = 0'02^2$$

de donde,

$$n = \frac{21 \cdot 64 \cdot 10^4}{48^2 \cdot 4} = 1.458'33 \simeq 1.459$$

Problema 2.7.- Una población de N unidades se dividió en tres estratos siendo los pesos relativos $W_1 = 0'5, W_2 = 0'3, W_3 = 0'2$ y las proporciones poblacionales $P_1 = 0'52, P_2 = 0'4, P_3 = 0'6$.

Suponiendo que los tres estratos tienen un tamaño suficiente para poder prescindir del factor de corrección, determinar el tamaño de muestra estratificada con afijación proporcional que de la misma precisión que una muestra aleatoria simple sin estratificar de tamaño 600, en la estimación de la proporción poblacional.

De los datos del enunciado se deduce que la proporción poblacional es

$$P = \sum_{h=1}^{3} W_h P_h = 0'5 \cdot 0'52 + 0'3 \cdot 0'4 + 0'2 \cdot 0'6 = 0'5.$$

Por otro lado, la proporción muestral

$$\hat{P}_{st} = \sum_{h=1}^{3} W_h \hat{p}_h$$

tiene una varianza

$$V(\hat{P}_{st}) = \sum_{h=1}^{3} W_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_h Q_h}{n_h}$$

que al poder prescindir del factor de corrección se reduce a

$$V(\hat{P}_{st}) = \sum_{h=1}^{3} W_h^2 \cdot \frac{P_h (1 - P_h)}{n_h}.$$

Por último, bajo afijación proporcional ($n_h = W_h \cdot n$), tomará un valor,

$$V(\hat{P}_{st}) = \sum_{h=1}^{3} W_h^2 \cdot \frac{P_h (1 - P_h)}{W_h n} = \frac{1}{n} \sum_{h=1}^{3} W_h \cdot P_h \cdot (1 - P_h) = \frac{0'2448}{n}.$$

Por otro lado, el estimador de la proporción en un muestreo aleatorio simple de tamaño 600 tiene una varianza

$$V(\hat{p}) = \frac{PQ}{600} = \frac{0'25}{600}.$$

De la condición impuesta en el enunciado

$$\frac{0'2448}{n} = \frac{0'25}{600}$$

se obtiene un tamaño muestral igual a $n = 587'52 \approx 588$.

Problema 2.8.- Comparar los valores obtenidos para $V(\hat{P}_{st})$ en los casos de afijación proporcional y afijación óptima para un tamaño muestral fijo, en las siguientes dos poblaciones.

Población 1		**Población 2**	
Estrato	P_h	*Estrato*	P_h
I	0'1	I	0'01
II	0'5	II	0'05
III	0'9	III	0'1

Supónganse los estratos de igual tamaño y que el factor de corrección para poblaciones finitas puede ser ignorado.

Supuesto que el factor de corrección para poblaciones finitas puede ser ignorado, la varianza de $\hat{P}_{st}$

$$V(\hat{P}_{st}) = \sum W_h^2 \frac{P_h Q_h}{n_h}$$

bajo afijación proporcional y bajo afijación óptima es, respectivamente,

$$V(\hat{P}_{ap}) = \frac{1}{n} \sum W_h P_h Q_h = \frac{1}{3n} \sum P_h Q_h$$

$$V(\hat{P}_{ao}) = \frac{1}{n} \left(\sum W_h \sqrt{P_h Q_h} \right)^2 = \frac{1}{9n} \left(\sum \sqrt{P_h Q_h} \right)^2$$

en donde la última igualdad se tiene al ser los tres estratos de igual tamaño, $W_h = 1/3$.

En la Población 1, la reducción en la varianza obtenida con la afijación óptima es

$$\frac{V(\hat{P}_{ap})}{V(\hat{P}_{ao})} = \frac{0'143/n}{0'134/n} = 1'067$$

Y en la Población 2,

$$\frac{V(\hat{P}_{ap})}{V(\hat{P}_{ao})} = \frac{0'0491/n}{0'0424/n} = 1'158$$

Problema 2.9.- Se desea realizar un muestreo estratificado en el cual se espera que el coste del trabajo de campo sea de la forma $\sum c_h n_h$. Por un censo anterior se conocen los siguientes valores para los dos estratos

Estratos	W_h	S_h	c_h
I	0'4	10	400 *Pts.*
II	0'6	20	900 *Pts.*

(a) Encontrar los valores de n_1/n y n_2/n que minimizan el coste total del trabajo de campo para un valor dado de la varianza del estimador de la media del muestreo estratificado.

(b) Encontrar el tamaño muestral requerido, con la afijación óptima anterior, para hacer la varianza igual a 1. Ignórese el factor de corrección.

(c) Calcular el coste total resultante de los trabajos de campo.

(a) La función a minimizar es

$$\Phi = \sum c_h n_h + \lambda \left(\sum W_h^2 \frac{S_h^2}{n_h} - \frac{1}{N} \sum W_h S_h^2 - V_0 \right)$$

de la cual se obtiene la afijación

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})}$$

que para los datos proporcionados por la tabla del enunciado, resultan los valores

$$\frac{n_1}{n} = \frac{1}{3} \qquad \frac{n_2}{n} = \frac{2}{3}$$

(b) La varianza del estimador de la media, prescindiendo del factor de corrección para poblaciones finitas es

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}.$$

De la condición exigida

$$1 = \frac{W_1^2 S_1^2}{n_1} + \frac{W_2^2 S_2^2}{n_2} = \frac{0'4^2 \cdot 10^2}{n/3} + \frac{0'6 \cdot 20^2}{2n/3}$$

se obtiene el valor $n = 264$, y las afijaciones $n_1 = 88$, $n_2 = 176$.

(c) El coste total de los trabajos de campo será

$$C = c_1 \cdot n_1 + c_2 \cdot n_2 = 400 \cdot 88 + 900 \cdot 176 = 193.600.$$

Problema 2.10.- Los siguientes datos muestran la estratificación de las granjas de un pais por su tamaño (en acres), incluyendo el número de granjas de cada estrato, N_h, el promedio de acres dedicados al cultivo de maiz por granja, $\overline{Y}_h$, para cada estrato, así como la cuasidesviación típica en cada uno de ellos, S_h

Tamaño	N_h	$\overline{Y}_h$	S_h
$0-40$	394	5'4	8'3
$41-80$	461	16'3	13'3
$81-120$	391	24'3	15'1
$121-160$	334	34'5	19'8
$161-200$	169	42'1	24'5
$201-240$	113	50'1	26
más de 240	148	63'8	35'2

Para un tamaño muestral de 100 granjas, calcular los tamaños muestrales para cada estrato bajo

(a) Afijación proporcional.

(b) Afijación óptima.

(c) Comparar las precisiones de estos dos métodos con un muestreo aleatorio simple.

A partir de la tabla del enunciado se obtiene la siguiente,

Tamaño	N_h	$\overline{Y}_h$	S_h	S_h^2	$N_h \cdot S_h$
$0-40$	394	5'4	8'3	68'89	327'02
$41-80$	461	16'3	13'3	176'89	6131'3
$81-120$	391	24'3	15'1	228'01	5904'1
$121-160$	334	34'5	19'8	392'04	6613'2
$161-200$	169	42'1	24'5	600'25	4140'5
$201-240$	113	50'1	26	676	2938
más de 240	148	63'8	35'2	1239'04	5209'6

la cual es utilizada en los cálculos del problema.

(a) La afijación proporcional viene dada por

$$n_h = \frac{N_h}{N} \cdot n = \frac{N_h}{2.010} \cdot 100 = \begin{cases} n_1 = 19'60 \approx 20 \\ n_2 = 22'94 \approx 23 \\ n_3 = 19'45 \approx 19 \\ n_4 = 16'62 \approx 17 \\ n_5 = 8'41 \approx 8 \\ n_6 = 5'62 \approx 6 \\ n_7 = 7'36 \approx 7 \end{cases}$$

(b) La afijación óptima viene dada por

$$n_h = \frac{N_h \cdot S_h}{\sum N_h \cdot S_h} \cdot n = \frac{N_h \cdot S_h}{34.206'9} \cdot 100 = \begin{cases} n_1 = 9'56 & \approx & 10 \\ n_2 = 17'92 & \approx & 18 \\ n_3 = 17'26 & \approx & 17 \\ n_4 = 19'33 & \approx & 19 \\ n_5 = 12'10 & \approx & 12 \\ n_6 = 8'59 & \approx & 9 \\ n_7 = 15'23 & \approx & 15 \end{cases}$$

(c) La varianza del estimador de la media en el muestreo estratificado

$$V(\overline{x}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h \left(N_h - n_h\right) \frac{S_h^2}{n_h}$$

toma un valor, en el caso de afijación proporcional de

$$V(\overline{x}_{ap}) = \frac{13.345.197}{2.010^2} = 3'303$$

y en el caso de afijación óptima de

$$V(\overline{x}_{ao}) = \frac{11.017.293}{2.010^2} = 2'727.$$

Para calcular la varianza del estimador de la media en un muestreo aleatorio simple,

$$V(\overline{x}) = \frac{N-n}{N} \cdot \frac{S^2}{n}$$

debemos calcular la cuasivarianza de la población, S^2, que es igual a

$$S^2 = \frac{1}{N-1} \sum_{h=1}^{L} \sum_{j=1}^{N_h} \left(Y_{hj} - \overline{Y}\right)^2 = \frac{1}{N-1} \sum_{h=1}^{L} \sum_{j=1}^{N_h} \left(Y_{hj} - \overline{Y}_h\right)^2 +$$

$$+ \frac{1}{N-1} \sum_{h=1}^{L} N_h \left(\overline{Y}_h - \overline{Y}\right)^2 = \frac{686.609'27}{2009} + \frac{556.546'91}{2009} = 618'79$$

al ser la media poblacional igual a

$$\overline{Y} = \frac{1}{N}\sum_{h=1}^{L} N_h \cdot \overline{Y}_h = \frac{52.884'8}{2.010} = 26'31.$$

La varianza de la media muestral en un muestreo aleatorio simple será, por tanto,

$$V(\overline{x}) = \frac{N-n}{N} \cdot \frac{S^2}{n} = \frac{2.010-100}{2.010} \cdot \frac{618'79}{100} = 5'88.$$

Realizando las comparaciones solicitadas por cociente, se obtienen unos valores de

$$\frac{V(\overline{x})}{V(\overline{x}_{ap})} = \frac{5'88}{3'303} = 1'78$$

y

$$\frac{V(\overline{x})}{V(\overline{x}_{ao})} = \frac{5'88}{2'727} = 2'16.$$

Problema 2.11.- Se quiere estimar la media de una población estratificada en tres estratos, de tamaños $N_1 = 3000$, $N_2 = 2000$ y $N_3 = 5000$ y cuasivarianzas $S_1^2 = 100$, $S_2^2 = 400$ y $S_3^2 = 900$, por medio de una muestra de tamaño 100. Determinar el error de muestreo que se comete si se utiliza: (a) Afijación igual, (b) Afijación proporcional, y (c) Afijación óptima.

La varianza del estimador de la media en el muestreo estratificado es

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2 \, \frac{N_h - n_h}{N_h} \, \frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 \, S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$

(a) Bajo afijación igual, $\omega_h = n_h/n = 1/L$, ésta es

$$V(\overline{x}_{ai}) = \sum_{h=1}^{L} W_h^2 \, S_h^2 \left(\frac{L}{n} - \frac{1}{N_h}\right)$$

que con los datos del problema resulta igual a

$$V(\overline{x}_{ai}) = (0'3)^2 \cdot 100 \cdot \left(\frac{3}{100} - \frac{1}{3000}\right) +$$

$$+(0'2)^2 \cdot 400 \cdot \left(\frac{3}{100} - \frac{1}{2000}\right) + (0'5)^2 \cdot 900 \cdot \left(\frac{3}{100} - \frac{1}{5000}\right) = 7'444$$

con lo que el error de muestreo bajo afijación igual será $\sqrt{V(\overline{x}_{ai})} = 2'73$.

(b) Bajo ajijación proporcional, $\omega_h = W_h$, la varianza del estimador de la media queda,

$$V(\overline{x}_{ap}) = \frac{N-n}{N\,n} \sum_{h=1}^{L} W_h\, S_h^2$$

que para los datos del problema resulta igual a

$$V(\overline{x}_{ap}) = \frac{9.900}{1.000.000}(0'3 \cdot 100 + 0'2 \cdot 400 + 0'5 \cdot 900) = 5'544$$

el error de muestreo será, por tanto $\sqrt{V(\overline{x}_{ap})} = 2'35$.

(c) Bajo afijación óptima, $\omega_h = N_h S_h / (\sum N_h S_h)$, la varianza buscada será

$$V(\overline{x}_{ao}) = \frac{(\sum W_h\, S_h)^2}{n} - \frac{\sum W_h\, S_h^2}{N} =$$

$$= \frac{(0'3 \cdot 10 + 0'2 \cdot 20 + 0'5 \cdot 30)^2}{100} - \frac{0'3 \cdot 100 + 0'2 \cdot 400 + 0'5 \cdot 900}{10.000} = 4'784$$

y el error de muestreo, $\sqrt{V(\overline{x}_{ao})} = 2'19$.

Problema 2.12.- Determinar el tamaño n de una muestra estratificada que con afijación proporcional permita estimar la proporción poblacional P con un coeficiente de variación del 5 %. (Utilizar fórmulas simplificadas despreciando la fracción de muestreo).

La varianza del estimador de la proporción para el muestreo estratificado es

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \cdot \frac{N_h - n_h}{N_h - 1} \cdot \frac{P_h \cdot Q_h}{n_h}.$$

Despreciando la fracción de muestreo y utilizando la afijación proporcional, $\omega_h = W_h$, ésta queda,

$$V(\hat{P}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{P_h \cdot Q_h}{n_h} = \sum_{h=1}^{L} W_h \frac{P_h \cdot Q_h}{n}.$$

Por otro lado, el coeficiente de variación de $\hat{P}_{st}$ debe ser

$$C.V.(\hat{P}_{st}) = \frac{\sqrt{V(\hat{P}_{st})}}{E[\hat{P}_{st}]} = 0'05.$$

Es decir,

$$\frac{\sqrt{\sum W_h P_h Q_h}}{\sqrt{n} \sum W_h P_h} = 0'05$$

de donde despejando se obtiene el valor,

$$n = \frac{400 \sum W_h P_h Q_h}{(\sum W_h P_h)^2}$$

Problema 2.13.- En una población finita se consideran dos estratos de tamaños $N_1 = 400$ y $N_2 = 600$ siendo las correspondientes cuasidesviaciones típicas $S_1 = 12$ y $S_2 = 20$. Se pide:

(a) Determinar la varianza del estimador centrado de la media, siendo el tamaño muestral $n = 200$ y la afijación proporcional.

(b) Determinar, para el mismo tamaño muestral anterior, la afijación óptima y la varianza que tendría el estimador en ese caso.

(c) Si los costes por unidad muestral son $c_1 = 225$ pts. y $c_2 = 169$ pts., y suponiendo que se dispone de un presupuesto total de 30.000 pts. y se desea afijación óptima, ¿cuál será el tamaño muestral?. Determinar la varianza del estimador en esta situación.

La varianza del estimador de la media en el muestreo estratificado es

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{N_h - n_h}{N_h} \frac{S_h^2}{n_h} = \sum_{h=1}^{L} W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$$

que bajo afijación proporcional adopta la forma

$$V(\overline{x}_{ap}) = \frac{N-n}{N\,n}\sum_{h=1}^{L} W_h\,S_h^2$$

resultando igual a

$$V(\overline{x}_{ap}) = \frac{1000-200}{1000\cdot 200}(0'4\cdot 12^2 + 0'6\cdot 20^2) = 1'1904.$$

(b) Las afijaciones óptimas,

$$n_h = \frac{N_h S_h}{\sum N_h S_h}\cdot n$$

serán

$$n_1 = \frac{400\cdot 12}{400\cdot 12 + 600\cdot 20}\cdot 200 = 57'14 \approx 57$$

$$n_2 = \frac{600\cdot 20}{400\cdot 12 + 600\cdot 20}\cdot 200 = 142'86 \approx 143.$$

La varianza para estas afijaciones la determinaremos mediante la expresión

$$V(\overline{x}_{ao}) = \frac{(\sum W_h\,S_h)^2}{n} - \frac{\sum W_h\,S_h^2}{N}$$

en lugar de utilizar las afijaciones acabadas de calcular, con objeto de obtener mayor precisión. Será

$$V(\overline{x}_{ao}) = \frac{(0'4\cdot 12 + 0'6\cdot 20)^2}{200} - \frac{0'4\cdot 12^2 + 0'6\cdot 20^2}{1000} = 1'1136.$$

(c) Suponiendo que el coste total es de la forma

$$C = c_1\,n_1 + c_2\,n_2$$

y que el tipo de afijación es el óptimo para costes variables,

$$n_h = \frac{N_h\,S_h/\sqrt{c_h}}{\sum N_h\,S_h/\sqrt{c_h}}\cdot n$$

se obtiene la ecuación

$$C = \left(\frac{\sqrt{c_1}\, N_1\, S_1}{\sum N_h\, S_h/\sqrt{c_h}} + \frac{\sqrt{c_2}\, N_2\, S_2}{\sum N_h\, S_h/\sqrt{c_h}} \right) \cdot n$$

es decir,

$$30.000 = \left(\frac{15 \cdot 400 \cdot 12}{1.243'08} + \frac{13 \cdot 600 \cdot 20}{1.243'08} \right) \cdot n$$

de donde se obtiene un tamaño muestral de $n = 163'56$. Tomaremos $n = 163$ con objeto de no sobrepasar el presupuesto.

Para este tamaño muestral, las afijaciones deberán ser

$$n_1 = \frac{400 \cdot 12/\sqrt{225}}{(400 \cdot 12/\sqrt{225}) + (600 \cdot 20/\sqrt{169})} \cdot 163 = 41'96 \approx 42$$

$$n_2 = \frac{600 \cdot 20/\sqrt{169}}{(400 \cdot 12/\sqrt{225}) + (600 \cdot 20/\sqrt{169})} \cdot 163 = 121'03 \approx 121.$$

Comprobemos que no superamos el presupuesto:

$$225 \cdot 42 + 169 \cdot 121 = 29.899 < 30.000.$$

Por último, con estas afijaciones, la varianza del estimador de la media será,

$$V(\bar{x}_{st}) = \sum_{h=1}^{L} W_h^2\, S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) =$$

$$= 0'4^2 \cdot 12^2 \left(\frac{1}{42} - \frac{1}{400} \right) + 0'6^2 \cdot 20^2 \left(\frac{1}{121} - \frac{1}{600} \right) = 1'441.$$

Problema 2.14.- De una población estratificada en 4 estratos se extrajo una muestra estratificada de tamaño $n = 15$. A partir de ella se han calculado los valores de $\hat{X}_{hi}$ que figuran en la siguiente tabla

h \ i	1	2	3	4	5
I	500	300	400	– – –	– – –
II	800	1000	700	900	– – –
III	600	500	700	400	800
IV	4000	3000	5000	– – –	– – –

siendo $\hat{X}_{hi}$ un estimador insesgado del total del estrato h, X_h, basado en la información recogida en la unidad i-ésima seleccionada en el estrato h-ésimo. Se pide:

(a) Determinar un estimador insesgado $\hat{X}_h$ del total del estrato h-ésimo X_h, basado en los estimadores $\hat{X}_{h1}, \hat{X}_{h2}, \ldots$. Calcular sus valores para la muestra obtenida.

(b) Determinar un estimador insesgado $\hat{X}$ del total poblacional X, basado en $\hat{X}_1$, ..., $\hat{X}_4$. Calcular su valor para la muestra obtenida.

(c) Despreciando los factores de corrección para poblaciones finitas, determinar estimadores de las varianzas de $\hat{X}_h$ y de $\hat{X}$, así como sus valores en la muestra obtenida.

(a) Como los $\hat{X}_{hi}$ son estimadores insesgados de X_h, un estimador insesgado de X_h será

$$\hat{X}_h = \frac{1}{k}\left\{\hat{X}_{h1} + \ldots + \hat{X}_{hk}\right\}$$

por la linealidad de la esperanza, supuesto que se seleccionaron k unidades en el estrato h-ésimo.

Para los datos del enunciado se tendrán las estimaciones $\hat{X}_1 = 400$, $\hat{X}_2 = 850$, $\hat{X}_3 = 600$ y $\hat{X}_4 = 4.000$.

(b) De nuevo por la linealidad de la esperanza, un estimador insesgado del total poblacional será

$$\hat{X} = \hat{X}_1 + \hat{X}_2 + \hat{X}_3 + \hat{X}_4$$

por ser los $\hat{X}_k$ estimadores insesgados de los totales de los estratos y ser el total poblacional suma de éstos.

Para la muestra obtenida toma un valor

$$\hat{X} = 400 + 850 + 600 + 4.000 = 5.850$$

(c) Como $\hat{X}_h$ es una "media muestral", un estimador de su varianza será (prescindiendo del factor de corrección) la "cuasivarianza muestral" dividida por el número de observaciones

$$\hat{V}(\hat{X}_h) = \frac{1}{(k-1)\,k}\sum_{i=1}^{k}\left(\hat{X}_{hi} - \hat{X}_h\right)^2$$

que para los cuatro estimadores del enunciado toma los valores

$$\hat{V}(\hat{X}_1) = \frac{1}{2 \cdot 3}\left[(500-400)^2+(300-400)^2+(400-400)^2\right] =$$

$$= \frac{10.000}{3} = 3.333'33$$

$$\hat{V}(\hat{X}_2) = \frac{1}{3 \cdot 4}\left[(800-850)^2+(1000-850)^2+\right.$$

$$\left.+(700-850)^2+(900-850)^2\right] = \frac{12.500}{3} = 4.166'67$$

$$\hat{V}(\hat{X}_3) = \frac{1}{4 \cdot 5}\left[(600-600)^2+(500-600)^2+\right.$$

$$\left.+(700-600)^2+(400-600)^2+(800-600)^2\right] = 5.000$$

$$\hat{V}(\hat{X}_4) = \frac{1}{2 \cdot 3}\left[(4000-4000)^2+(3000-4000)^2+\right.$$

$$\left.+(5000-4000)^2\right] = \frac{1.000.000}{3} = 333.333'33.$$

Por último, como es .

$$\hat{X} = \hat{X}_1 + \hat{X}_2 + \hat{X}_3 + \hat{X}_4$$

será

$$\hat{V}(\hat{X}) = \hat{V}(\hat{X}_1) + \hat{V}(\hat{X}_2) + \hat{V}(\hat{X}_3) + \hat{V}(\hat{X}_4)$$

al ser independientes los cuatro estimadores del segundo miembro, por ser un muestreo estratificado. Por tanto, será $\hat{V}(\hat{X}) = 345.833'33$.

Problema 2.15.- En un muestreo estratificado se ha obtenido para cada estrato una muestra aleatoria simple de n_h elementos, con $\sum_{h=1}^{L} n_h = n$.

Para estimar la media poblacional se quiere utilizar un estimador del tipo

$$\widehat{\overline{X}} = \sum_{h=1}^{L} c_h \overline{x}_h$$

con $\sum_{h=1}^{L} c_h = 1$ y siendo $\overline{x}_h$ la media muestral en el estrato h.
Se pide:

(a) Determinar el sesgo del estimador $\widehat{\overline{X}}$. ¿Qué condición deberán cumplir los c_h para que éste sea insesgado?.

(b) Calcular los c_h que minimizan la varianza del estimador.

(a) La esperanza de $\widehat{\overline{X}}$ es

$$E\left[\widehat{\overline{X}}\right] = \sum_{h=1}^{L} c_h E[\overline{x}_h] = \sum_{h=1}^{L} c_h \overline{X}_h$$

con lo que su sesgo será

$$E\left[\widehat{\overline{X}}\right] - \overline{X} = \sum_{h=1}^{L} c_h \overline{X}_h - \sum_{h=1}^{L} W_h \overline{X}_h = \sum_{h=1}^{L} (c_h - W_h) \overline{X}_h$$

que se hace cero si $c_h = W_h \ \forall h$. Es decir, si $\widehat{\overline{X}} = \overline{x}_{st}$.

(b) La varianza de $\widehat{\overline{X}}$ será

$$V(\widehat{\overline{X}}) = \sum_{h=1}^{L} c_h^2 V(\overline{x}_h) = \sum_{h=1}^{L} c_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

La función a minimizar será, por tanto,

$$\Phi = \sum_{h=1}^{L} c_h^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{S_h^2}{n_h} + \lambda \left(\sum_{h=1}^{L} c_h - 1\right)$$

Como es

$$\frac{\partial \Phi}{\partial c_h} = 2\, c_h \left(\frac{N_h - n_h}{N_h}\right) \frac{S_h^2}{n_h} + \lambda = 0 \qquad \forall h = 1, ..., L$$

será

$$\lambda = -2 \frac{c_h}{n_h / [(1 - f_h)\, S_h^2]} \qquad \forall h = 1, ..., L$$

con lo que será

$$-2\,\frac{c_h}{n_h/[(1-f_h)\,S_h^2]} = -2\,\frac{\sum c_h}{\sum n_h/[(1-f_h)\,S_h^2]}$$

es decir,

$$c_h = \frac{n_h/[(1-f_h)\,S_h^2]}{\sum n_h/[(1-f_h)\,S_h^2]}$$

Problema 2.16.- Una población estratificada en dos estratos presenta los siguiente valores poblacionales

Estrato	W_h	S_h
I	0'8	2
II	0'2	4

Ignorando el factor de corrección para poblaciones finitas, determinar los tamaños muestrales n_1 y n_2 en cada uno de los tres casos siguientes:

(a) Que minimicen el tamaño muestral $n = n_1 + n_2$, siendo el error de muestreo del estimador de la media 0'1.

(b) Que el error de muestreo del estimador de la media de cada estrato sea 0'1.

(c) Que minimicen el tamaño muestral $n = n_1 + n_2$, siendo el error de muestreo de la diferencia entre los estimadores de las medias de los estratos 0'1.

(a) La varianza del estimador de la media en un muestreo estratificado es

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2\,\frac{S_h^2}{n_h}$$

Ignorando el factor de corrección para poblaciones finitas, queda

$$V(\overline{x}_{st}) = \sum_{h=1}^{L} W_h^2\,\frac{S_h^2}{n_h} = \frac{2'56}{n_1} + \frac{0'64}{n_2}$$

La función a minimizar en este primer apartado es, por tanto,

$$\Phi = n_1 + n_2 + \lambda\left(\frac{2'56}{n_1} + \frac{0'64}{n_2} - 0'01\right)$$

Derivando parcialmente respecto a n_1 y n_2 se obtiene el sistema

$$\left.\begin{array}{l}\dfrac{\partial\Phi}{\partial n_h} = 0 \\ h = 1,2\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l}1 - \lambda\,\dfrac{2'56}{n_1^2} = 0 \\ \\ 1 - \lambda\,\dfrac{0'64}{n_2^2} = 0\end{array}\right\} \Longrightarrow \lambda = \frac{n_1^2}{2'56} = \frac{n_2^2}{0'64}$$

es decir, $n_1 = 2\,n_2$.

Como además debe ser

$$\frac{2'56}{n_1} + \frac{0'64}{n_2} = 0'01$$

se obtienen los valores $n_1 = 384$, $n_2 = 192$ y por tanto $n = 576$.

(b) Como en un muestreo estratificado la muestra de cada estrato es aleatoria simple, ignorando el factor de corrección será $V(\overline{x}_h) = S_h^2/n_h$ con lo que las condiciones exigidas en el enunciado serán

$$\left.\begin{array}{l}V(\overline{x}_1) = \dfrac{4}{n_1} = 0'01 \\ \\ V(\overline{x}_2) = \dfrac{16}{n_2} = 0'01\end{array}\right\} \Longrightarrow \left\{\begin{array}{l}n_1 = \dfrac{4}{0'01} = 400 \\ \\ n_2 = \dfrac{16}{0'01} = 1.600\end{array}\right.$$

siendo, por tanto, el tamaño muestral $n = 2.000$.

(c) Como en un muestreo estratificado las muestras son seleccionadas independientemente en cada uno de los estratos, será

$$V(\overline{x}_1 - \overline{x}_2) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

con lo que la función a minimizar en esta ocasión es

$$\Phi = n_1 + n_2 + \lambda\left(\frac{4}{n_1} + \frac{16}{n_2} - 0'01\right)$$

Derivando parcialmente respecto a n_1 y n_2 se obtiene el sistema

$$\left.\begin{array}{l}\dfrac{\partial \Phi}{\partial n_h} = 0 \\ \\ h = 1,2\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l}1 - \lambda \dfrac{4}{n_1^2} = 0 \\ \\ 1 - \lambda \dfrac{16}{n_2^2} = 0\end{array}\right\} \Longrightarrow \lambda = \frac{n_1^2}{4} = \frac{n_2^2}{16}$$

es decir, $n_1 = 0'5\, n_2$.

Como además debe ser

$$\frac{4}{n_1} + \frac{16}{n_2} = 0'01$$

se obtienen los valores $n_1 = 1.200$, $n_2 = 2.400$ y por tanto $n = 3.600$.

Problema 2.17.- Se desea estimar el número medio de activos financieros por familia de una población. Para ello se estratificó dicha población en dos estratos, el correspondiente a rentas altas, de tamaño $N_1 = 4.000$ familias, y el correspondiente a rentas bajas, de $N_2 = 20.000$ familias. Se piensa que una familia de renta alta posee, aproximadamente, nueve veces más activos financieros que una de la clase baja, y que la cuasidesviación típica de cada estrato es proporcional a la raiz cuadrada de la media del estrato, $S_h = c\sqrt{\overline{X}_h}$, $h = 1,2$.

Con objeto de obtener la máxima precisión en la estimación, ¿cómo distribuiría entre los dos estratos una muestra de $n = 1.000$ familias?. ¿Y si se quisiera estimar la diferencia entre el número medio de activos financieros de los dos estratos?.

La función a minimzar será

$$\Phi = V(\overline{x}_{st}) + \lambda(n_1 + n_2 - 1.000) = \sum_{h=1}^{L} \left(1 - \frac{n_h}{N_h}\right) W_h^2 \frac{S_h^2}{n_h} + \lambda(n_1 + n_2 - 1.000)$$

Como, según el enunciado es $\overline{X}_1 = 9\,\overline{X}_2$ y además $S_h^2 = k\,\overline{X}_h$, $(k = c^2)$, será

$$\Phi = \left(1 - \frac{n_1}{4.000}\right) \frac{k\,\overline{X}_2}{4\,n_1} + \left(1 - \frac{n_2}{20.000}\right) \frac{25\,k\,\overline{X}_2}{36\,n_2} + \lambda(n_1 + n_2 - 1.000)$$

Derivando parcialmente respecto a n_1 y n_2 se obtiene el sistema

$$\left.\begin{array}{l}\dfrac{\partial \Phi}{\partial n_h} = 0 \\ h = 1,2\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} -\dfrac{k\,\overline{X}_2}{4\,n_1^2} + \lambda = 0 \\ \\ -\dfrac{25\,k\,\overline{X}_2}{36\,n_2^2} + \lambda = 0 \end{array}\right\} \Longrightarrow \lambda = \frac{k\,\overline{X}_2}{4\,n_1^2} = \frac{25\,k\,\overline{X}_2}{36\,n_2^2}$$

es decir, $n_1 = 0'6\,n_2$. Como además debe ser

$$n_1 + n_2 = 1.000$$

se obtienen los valores $n_1 = 375$ y $n_2 = 625$.

Si se quisiera estimar la diferencia entre las medias de los dos estratos, al realizarse independientemente en ambos estratos un muestreo aleatorio simple, la varianza de la diferencia sería

$$V(\overline{x}_1 - \overline{x}_2) = \sum_{h=1}^{2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

siendo, por tanto, la función a minimizar

$$\Phi = \left(1 - \frac{n_1}{4.000}\right) \frac{9\,k\,\overline{X}_2}{n_1} + \left(1 - \frac{n_2}{20.000}\right) \frac{k\,\overline{X}_2}{n_2} + \lambda(n_1 + n_2 - 1.000)$$

Derivando de nuevo parcialmente respecto a n_1 y n_2 se obtiene el sistema

$$\left.\begin{array}{l}\dfrac{\partial \Phi}{\partial n_h} = 0 \\ h = 1,2\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l} -\dfrac{9\,k\,\overline{X}_2}{n_1^2} + \lambda = 0 \\ \\ -\dfrac{k\,\overline{X}_2}{n_2^2} + \lambda = 0 \end{array}\right\} \Longrightarrow \lambda = \frac{9\,k\,\overline{X}_2}{n_1^2} = \frac{k\,\overline{X}_2}{n_2^2}$$

de donde se obtiene que debe ser $n_1 = 3\,n_2$. Como además es

$$n_1 + n_2 = 1.000$$

deberá ser $n_1 = 750$ y $n_2 = 250$.

Problema 2.18.- Se quiere estimar la media de una población estratificada en dos estratos. Para ello se puede elegir entre una afijación igual o una afijación óptima; es decir, entre asignar el mismo tamaño muestral a los dos estratos, $n/2$, o aquellos n_1 y n_2 que resulten de minimizar la varianza del estimador de la media en el muestreo estratificado, sujeto a la restricción $n_1 + n_2 = n$.

Si las varianzas obtenidas con ambas afijaciones son, respectivamente, $V(\overline{x}_{ai})$ y $V(\overline{x}_{ao})$, se pide, ignorando el factor de corrección,

(a) Demostrar que

$$\frac{V(\overline{x}_{ai}) - V(\overline{x}_{ao})}{V(\overline{x}_{ao})} = \left(\frac{r-1}{r+1}\right)^2$$

con $r = n_1/n_2$, siendo n_1 y n_2 las afijaciones óptimas.

(b) Si $W_1 = 0'8$, $W_2 = 0'2$, $S_1 = 2$ y $S_2 = 4$, ¿cuánto crece la anterior fracción $(V(\overline{x}_{ai}) - V(\overline{x}_{ao}))/V(\overline{x}_{ao})$ al pasar de una afijación igual a una óptima?.

(a) Con afijación igual es

$$V(\overline{x}_{ai}) = \frac{2\,W_1^2\,S_1^2}{n} + \frac{2\,W_2^2\,S_2^2}{n}$$

y con afijación óptima,

$$V(\overline{x}_{ao}) = \frac{W_1^2\,S_1^2}{n_1} + \frac{W_2^2\,S_2^2}{n_2}$$

siendo

$$n_h = \frac{W_h\,S_h}{\sum W_h\,S_h} \cdot n$$

La igualdad a demostrar es, por tanto,

$$\frac{\frac{2\,W_1^2\,S_1^2}{n} + \frac{2\,W_2^2\,S_2^2}{n}}{\frac{W_1^2\,S_1^2}{n_1} + \frac{W_2^2\,S_2^2}{n_2}} - 1 = \left(\frac{\frac{n_1}{n_2} - 1}{\frac{n_1}{n_2} + 1}\right)^2$$

o bien,

$$\frac{2}{n}\,\frac{W_1^2\,S_1^2 + W_2^2\,S_2^2}{\frac{W_1^2\,S_1^2}{n_1} + \frac{W_2^2\,S_2^2}{n_2}} = \frac{n_1^2 + n_2^2}{n_1 + n_2}$$

es decir,

$$n_1^2\, n_2\, W_2^2\, S_2^2 + n_1\, n_2^2\, W_1^2\, S_1^2 = n_1^3\, W_2^2\, S_2^2 + n_2^3\, W_1^2\, S_1^2.$$

Igualdad que se obtiene, sustituyendo n_h por su valor

$$n_h = \frac{W_h\, S_h}{\sum W_h\, S_h} \cdot n \qquad\qquad h = 1, 2.$$

(b) La fracción

$$\frac{V(\overline{x}_{ai}) - V(\overline{x}_{ao})}{V(\overline{x}_{ao})}$$

es claramente cero con afijación igual, ya que para ese caso debemos sustituir $n_1 = n_2$ en

$$\left(\frac{r-1}{r+1}\right)^2$$

es decir, hacer $r = 1$.

Con afijación óptima y los valores dados en el enunciado, es

$$n_1 = \frac{W_1\, S_1}{W_1\, S_1 + W_2\, S_2} \cdot n = \frac{0'8 \cdot 2}{0'8 \cdot 2 + 0'2 \cdot 4} \cdot n = \frac{2}{3} \cdot n$$

y

$$n_2 = \frac{1}{3} \cdot n$$

con lo que será $r = 2$ y

$$\left(\frac{r-1}{r+1}\right)^2 = \frac{1}{9}.$$

La fracción se incrementa, por tanto, 1/9.

Problema 2.19.- En una empresa de gran tamaño, lo que permite ignorar el factor de corrección para poblaciones finitas, el 62 % de los empleados son obreros de sexo masculino, el 31 % obreros de sexo femenino y el 7 % directivos. Se desea utilizar una muestra de 400 empleados de la empresa, para estimar la proporción de empleados que usa determinadas instalaciones deportivas de la empresa. Unos estudios previos sugirieron que las proporciones de empleados,

de los grupos antes mencionados, que usan las instalaciones en cuestión son respectivamente, el 45 %, el 25 % y el 7'5%. Se pide:

(a) ¿Cómo distribuiría las 400 unidades muestrales a seleccionar entre los tres grupos, de forma que se minimice la varianza del estimador?.

(b) Si las verdaderas proporciones de usuarios fueron el 48 %, el 21 % y el 4 % respectivamente, ¿cuál sería el valor del error de muestreo del estimador, utilizando las afijaciones determinadas en el apartado anterior?.

(c) Si utilizáramos un muestreo aleatorio simple de tamaño $n = 400$, ¿cuál sería el error de muestreo del estimador?.

(a) Del enunciado se deduce que la población está estratificada, con tamaños relativos, $W_1 = 0'62$, $W_2 = 0'31$ y $W_3 = 0'07$, siendo las proporciones de la característica en estudio para cada estrato, $P_1 = 0'45$, $P_2 = 0'25$ y $P_3 = 0'075$.

La afijación óptima es

$$n_h = \frac{W_h \sqrt{P_h Q_h}}{\sum W_h \sqrt{P_h Q_h}} \cdot n$$

que proporciona unos valores de

$$n_1 = \frac{0'62\sqrt{0'45 \cdot 0'55}}{0'4611} \cdot 400 = 267'564 \simeq 268$$

$$n_2 = \frac{0'31\sqrt{0'25 \cdot 0'75}}{0'4611} \cdot 400 = 116'442 \simeq 116$$

$$n_3 = \frac{0'07\sqrt{0'075 \cdot 0'925}}{0'4611} \cdot 400 = 15'99 \simeq 16$$

(b) La varianza, simplificada, del estimador de la proporción en un muestreo estratificado es

$$V(\hat{P}_{st}) = \sum_{h=1}^{3} W_h^2 \frac{P_h Q_h}{n_h}$$

que para los datos del problema toma el valor,

$$\begin{aligned} V(\hat{P}_{st}) &= 0'62^2 \frac{0'48 \cdot 0'52}{268} + 0'31^2 \frac{0'21 \cdot 0'79}{116} + 0'07^2 \frac{0'04 \cdot 0'96}{16} \\ &= 0'0004957. \end{aligned}$$

De donde el error de muestreo será la raiz cuadrada, es decir, 0'022264.

(c) La proporción poblacional será

$$P = W_1 P_1 + W_2 P_2 + W_3 P_3 = 0'2976 + 0'0651 + 0'0028 = 0'3655.$$

La varianza, simplificada, del estimador de la proporción en un muestreo aleatorio simple es

$$V(\tilde{p}_{as}) = \frac{P\,Q}{n} = \frac{0'3655 \cdot 0'6345}{400} = 0'0005797$$

con lo que el error de muestreo será 0'0240785; es decir, mayor que en el caso de un muestreo estratificado.

CAPÍTULO 3

Estimadores de la razón

Cuando se dispone de información adicional sobre una variable Y_i observada en la misma población que la objeto de estudio X_i, se utilizan los denominados *estimadores de la razón.*

Para ello, si los totales poblacionales de estas variables son, respectivamente, Y y X, se define la *razón poblacional* R como

$$R = \frac{X}{Y} = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i} = \frac{\overline{X}}{\overline{Y}}$$

la cual se estima por la denominada *razón muestral*

$$\hat{R} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{\hat{X}}{\hat{Y}} = \frac{\overline{x}}{\overline{y}}.$$

Este estimador, no siempre insesgado, tiene como error cuadrático medio

$$E[(\hat{R} - R)^2] = \frac{N-n}{n\,N\,\overline{Y}^2}\left(S_x^2 - 2\,R\,S_{xy} + R^2\,S_y^2\right)$$

valor que coincidirá con la varianza cuando $\hat{R}$ sea insesgado para R. En la expresión anterior es

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2 \qquad S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

y

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y}).$$

El *estimador de razón para la media* se define ahora como

$$\hat{\overline{X}}_R = \hat{R}\,\overline{Y}$$

de error cuadrático medio

$$E[(\hat{\overline{X}}_R - \overline{X})^2] = \frac{N-n}{n\,N}\left(S_x^2 - 2\,R\,S_{xy} + R^2\,S_y^2\right)$$

el cual, habitualmente habrá que estimar por el estimador

$$\hat{E}[(\hat{\overline{X}}_R - \overline{X})^2] = \frac{N-n}{n\,N}\left(\hat{S}_x^2 - 2\,\hat{R}\,\hat{S}_{xy} + \hat{R}^2\,\hat{S}_y^2\right)$$

siendo

$$\hat{S}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2 \qquad\qquad \hat{S}_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{y})^2$$

y

$$\hat{S}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})(Y_i - \overline{y}).$$

El *estimador de razón para el total* se define como

$$\hat{X}_R = N\,\hat{\overline{X}}_R = \hat{R}\,Y$$

de error cuadrático medio

$$E[(\hat{X}_R - X)^2] = N\,\frac{N-n}{n}\left(S_x^2 - 2\,R\,S_{xy} + R^2\,S_y^2\right)$$

el cual, habitualmente habrá que estimar por el estimador

$$\hat{E}[(\hat{\overline{X}}_R - \overline{X})^2] = N\,\frac{N-n}{n}\left(\hat{S}_x^2 - 2\,\hat{R}\,\hat{S}_{xy} + \hat{R}^2\,\hat{S}_y^2\right)$$

Problema 3.1.- Se desea estudiar el comportamiento de la característica bidimensional (X,Y), en una población formada por 500 unidades. Para ello se extrajo una muestra aleatoria simple de tamaño 80, que proporcionó los siguientes datos:

$$\sum Y_i = 190 \quad \sum X_i = 420 \quad \sum Y_i X_i = 1045 \quad \sum Y_i^2 = 512 \quad \sum X_i^2 = 2284$$

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

El sesgo de $\hat{R}$ es

$$B = E[\hat{R} - R] = \frac{\overline{X}}{\overline{Y}} \left[\frac{V(\overline{y})}{\overline{Y}^2} - \frac{Cov(\overline{y}, \overline{x})}{\overline{Y}\ \overline{X}} \right] = \frac{1-f}{n\overline{Y}^2} \left[\frac{\overline{X}}{\overline{Y}} S_y^2 - S_{xy} \right].$$

Al depender este sesgo de parámetros desconocidos, habrá que sustituirlos por estimaciones suyas, quedando como expresión para la estimación del sesgo,

$$\hat{B} = \frac{1-f}{n\overline{y}^2} \left[\frac{\overline{x}}{\overline{y}} \hat{S}_y^2 - \hat{S}_{xy} \right].$$

En nuestro problema quedará,

$$\hat{B} = \frac{0'84}{80 \cdot 2'375^2} \left[\frac{420}{190} \cdot 0'76899 - 0'60127 \right] = 0'002045.$$

El error de muestreo es la desviación típica del estimador. Como éste no es, en general, insesgado, utilizaremos la raiz cuadrada del error cuadrático medio,

$$\begin{aligned} E[(\hat{R} - R)^2] &= \frac{N-n}{n(N-1)\overline{Y}^2}(\sigma_x^2 - 2R\sigma_{xy} + R^2\sigma_y^2) \\ &= \frac{1-f}{n\overline{Y}^2} \left[S_x^2 - 2\frac{\overline{X}}{\overline{Y}} S_{xy} + \left(\frac{\overline{X}}{\overline{Y}} \right)^2 S_y^2 \right]. \end{aligned}$$

De nuevo debemos sustituir los parámetros desconocidos por sus estimadores, quedando,

$$\hat{E}[(\hat{R}-R)^2] = \frac{1-f}{n\overline{y}^2}\left[\hat{S}_x^2 - 2\frac{\overline{x}}{\overline{y}}\,\hat{S}_{xy} + \left(\frac{\overline{x}}{\overline{y}}\right)^2 \hat{S}_y^2\right]$$

$$= \frac{0'84}{80\cdot 2'375^2}\left(1 - 2\cdot\frac{5'25}{2'375}\cdot 0'60127 + \left(\frac{5'25}{2'375}\right)^2\cdot 0'76899\right)$$

$$= 0'003908$$

siendo la raiz cuadrada del error cuadrático medio igual a 0'06251.

Problema 3.2.- En una ciudad que tiene 15.000 viviendas, se ha tomado una muestra aleatoria simple de tamaño 600. En cada vivienda se han observado dos variables: Y = nº de personas y X = nº de habitaciones, obteniéndose los siguientes datos:

$$\sum Y_i = 2946 \quad \sum X_i = 2150 \quad \sum X_iY_i = 12800 \quad \sum Y_i^2 = 18694 \quad \sum X_i^2 = 10997$$

Se pide:

(a) Estime el número medio de personas por vivienda, dando un intervalo de confianza.

(b) Sabiendo que el número total de personas en la ciudad es de 74.000, estime el total de habitaciones mediante un estimador de razón.

(c) Obtenga un intervalo de confianza para el cociente nº total de habitaciones / nº total de personas.

(a) Como la muestra es suficientemente grande, podemos suponer una distribución normal para la media muestral $\overline{y}$, obteniéndose un intervalo de confianza igual a

$$\left[\overline{y} - z_{\alpha/2}\frac{\hat{S}_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\ ,\ \overline{y} + z_{\alpha/2}\frac{\hat{S}_y}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right].$$

Tomando un nivel de significación $\alpha = 0'05$ se obtiene el intervalo

$$\left[4'91 \mp 1'96\cdot\frac{2'657}{\sqrt{600}}\sqrt{1-\frac{600}{15000}}\right] = [4'702\ ,\ 5'118].$$

(b) Como la razón poblacional es

$$R = \frac{X}{Y} = \frac{\sum_{i=1}^{N} X_i}{\sum_{i=1}^{N} Y_i}$$

será $X = RY$, con lo que un buen estimador para el total X será, $\hat{X}_R = \hat{R}\,Y$ siendo

$$\hat{R} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} Y_i} = \frac{2.150}{2.946} = 0'7298$$

la razón muestral.

Así pues, la estimación pedida será

$$\hat{X}_R = 0'7298 \cdot 74.000 = 54.005'2 \approx 54.005.$$

(c) El intervalo de confianza pedido será

$$\left[\hat{R} \mp z_{\alpha/2} \sqrt{\frac{N-n}{\overline{y}^2\, N\, n} \left(\hat{S}_x^2 + (\hat{R})^2\, \hat{S}_y^2 - 2\,\hat{R}\,\hat{S}_{xy} \right)} \right] =$$

$$\left[0'7298 \mp 1'96 \sqrt{\frac{15.000 - 600}{4'91^2 \cdot 15.000 \cdot 600} \left(5'497 + 0'7298^2 \cdot 7'06 - 2 \cdot 0'7298 \cdot 3'7618 \right)} \right]$$

$$= [0'6988\ ,\ 0'7608]$$

Obsérvese que en la situación particular de este problema, al venir dado en (b) el valor del total poblacional, y en consecuencia conocer el valor de $\overline{Y} = 74000/15000 = 4'93$, hubiera sido más adecuado utilizar este valor en lugar de $\overline{y} = 4'91$, aunque, dada la diferencia entre ambos, los resultados apenas si hubieran variado.

Problema 3.3.- Sabiendo que la razón muestral tiene como error cuadrático medio

$$E[(\hat{R} - R)^2)] = \frac{N-n}{n(N-1)\overline{Y}^2} [\sigma_x^2 - 2\,R\,\sigma_{xy} + R^2 \sigma_y^2]$$

siendo σ_{xy} la covarianza entre X e Y, determinar el error cuadrático medio del estimador de razón para la media, $\hat{\overline{X}}_R$.

Como es,

$$\hat{\overline{X}}_R = \overline{Y}\,\hat{R} \quad \text{y} \quad \overline{X} = \overline{Y}\,R$$

será

$$E[(\hat{\overline{X}}_R - \overline{X})^2] = E[(\overline{Y}\,\hat{R} - \overline{Y}\,R)^2] = \overline{Y}^2\,E[(\hat{R} - R)^2]$$

$$= \frac{N-n}{n(N-1)}[\sigma_x^2 - 2\,R\,\sigma_{xy} + R^2\sigma_y^2].$$

Problema 3.4.- De una población de 468 colegios pequeños se extrajo una muestra aleatoria simple de tamaño 100, conteniendo 54 públicos y 46 privados. Los datos muestrales del número de estudiantes (variable X) y del número de profesores (variable Y) fueron los siguientes:

	n	$\sum X_i$	$\sum Y_i$	$\sum X_i^2$	$\sum X_iY_i$	$\sum Y_i^2$
Públicos	54	31.281	2.024	29.881.219	1.729.349	111.090
Privados	46	13.707	1.075	6.366.775	431.041	33.119

Se pide:

(a) Para cada tipo de colegio en la población, estimar la razón *número de estudiantes/número de profesores.*

(b) Calcular el error de muestreo de las estimaciones efectuadas en el apartado anterior, tomando en ambos casos como fracción de muestreo el valor $f = 100/468$.

(a) El estimador de la razón para los *colegios públicos* será

$$\hat{R} = \frac{\sum X_i}{\sum Y_i} = \frac{31.281}{2.024} = 15'455$$

y para los *colegios privados*

$$\hat{R} = \frac{\sum X_i}{\sum Y_i} = \frac{13.707}{1.075} = 12'7507.$$

(b) El error de muestreo (estimado) de la razón muestral es la raiz cuadrada de

$$\hat{E}[(\hat{R}-R)^2] = \frac{1}{\overline{y}^2}\frac{1}{n}(1-f)\frac{1}{n-1}\left[\sum^n X_i^2 - 2\hat{R}\sum^n X_i Y_i + \sum^n Y_i^2(\hat{R})^2\right]$$

que en el caso de los *colégios públicos* será

$$\begin{aligned}\hat{E}[(\hat{R}-R)^2] &= \frac{1}{(2.024/54)^2}\cdot\frac{1}{54}\left(1-\frac{100}{468}\right)\frac{1}{53} \\ &\cdot [29.881.219 - 2\cdot 15'455\cdot 1.729.349 + 111.090\cdot(15'455)^2] \\ &= 0'5792091\end{aligned}$$

siendo, por tanto, el error de muestreo igual a 0'761058.

De la misma manera, para los *colegios privados* se obtendría un error de muestreo igual a 0'727.

Problema 3.5.- De una población de $N = 40$ hogares se obtuvo una muestra aleatoria simple de tamaño $n = 4$, la cual proporcionó 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
488	1100

Estimar el porcentaje de gasto en alimentación y su error de muestreo.

El parámetro a estimar es el porcentaje de gasto, es decir X/Y, la razón poblacional. Su estimador natural es la razón muestral

$$\hat{R} = \frac{\sum X_i}{\sum Y_i} = \frac{488}{1.100} = 0'4436.$$

Su error de muestreo (estimado) será la raiz cuadrada de

$$\hat{E}[(\hat{R}-R)^2] = \frac{1}{\bar{y}^2}\frac{1}{n}(1-f)\frac{1}{n-1}\left[\sum^n X_i^2 - 2\hat{R}\sum^n X_i Y_i + \sum^n Y_i^2(\hat{R})^2\right]$$

$$= \frac{1}{275^2}\frac{0'9}{12}\left[63.714 - 2\cdot 0'4436\cdot 141.050 + 315.000\cdot 0'1968\right]$$

$$\doteq 0'0006241.$$

El error de muestreo será, por tanto, 0'025; es decir, del 2'5%.

Problema 3.6.- Una muestra piloto de 21 hogares dió los siguientes resultados sobre el número de miembros, y, el número de niños, x_1, el número de coches, x_2, y el número de televisores, x_3

y	x_1	x_2	x_3	y	x_1	x_2	x_3	y	x_1	x_2	x_3
5	3	1	3	2	0	0	1	6	3	2	0
2	0	1	1	3	1	1	1	4	2	1	1
4	1	2	0	2	0	2	0	4	2	1	1
4	2	1	1	6	4	2	1	3	1	0	1
6	4	1	1	3	1	0	0	2	0	2	1
3	1	1	2	4	2	1	1	4	2	1	1
5	3	1	1	5	3	1	1	3	1	1	1

Suponiendo que el número total de habitantes de la población en estudio fuera conocido, Y, ¿qué sería preferible, el estimador de la razón o el estimador basado en un muestreo aleatorio simple, para estimar, (a) el total de niños en la población, (b) el total de coches poblacional, (c) el total de televisores en la población mencionada?.

El estimador de razón para el total poblacional es

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

siendo Y el número total de habitantes de la población en estudio y

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

la razón muestral.

El error cuadrático medio del estimador $\hat{X}_R$ será

$$E[(\hat{X}_R - X)^2] = E[(\hat{R}Y - RY)^2] = Y^2 E[(\hat{R} - R)^2]$$
$$= N^2 \frac{1-f}{n} \left[S_x^2 - 2\frac{\overline{X}}{\overline{Y}} S_{xy} + \left(\frac{\overline{X}}{\overline{Y}}\right)^2 S_y^2 \right]$$

el cual, al contener parámetros desconocidos, habrá que estimar por

$$N^2 \frac{1-f}{n} \left[\hat{S}_x^2 - 2\frac{\overline{x}}{\overline{y}} \hat{S}_{xy} + \left(\frac{\overline{x}}{\overline{y}}\right)^2 \hat{S}_y^2 \right]$$

expresión que, denotando por $\hat{S}_c^2$

$$\hat{S}_c^2 = \hat{S}_x^2 - 2\frac{\overline{x}}{\overline{y}} \hat{S}_{xy} + \left(\frac{\overline{x}}{\overline{y}}\right)^2 \hat{S}_y^2$$

quedará igual a

$$N^2 \frac{1-f}{n} \hat{S}_c^2.$$

Por otro lado, como la varianza estimada del estimador del total en un muestreo aleatorio simple es

$$\hat{V}(\hat{X}) = N^2 \frac{1-f}{n} \hat{S}_x^2$$

la comparación entre el estimador de la razón y el basado en un muestreo aleatorio simple queda reducida a comparar $\hat{S}_c^2$ y $\hat{S}_x^2$ en cada uno de los tres apartados.

(a) De los datos del enunciado se obtiene que es

$$\hat{S}_x^2 = 1'614286$$

y

$$\hat{S}_c^2 = \hat{S}_x^2 - 2\frac{\overline{x}}{\overline{y}}\hat{S}_{xy} + \left(\frac{\overline{x}}{\overline{y}}\right)^2 \hat{S}_y^2$$

$$= 1'614286 - 2 \cdot \frac{1'714286}{3'809524} \cdot 1'642856 + \left(\frac{1'714286}{3'809524}\right)^2 \cdot 1'761905$$

$$= 0'4925013.$$

Por tanto, tiene más precisión el estimador de la razón.

(b) En este caso es

$$\hat{S}_x^2 = 0'3904762$$

y

$$\hat{S}_c^2 = 0'3904762 - 2 \cdot \frac{1'095238}{3'809524} \cdot 0'2190479 + \left(\frac{1'095238}{3'809524}\right)^2 \cdot 1'761905$$

$$= 0'4101561$$

siendo ahora más preciso el basado en un muestreo aleatorio simple.

(c) Por último, para el total de televisores, será

$$\hat{S}_x^2 = 0'4476191$$

y

$$\hat{S}_c^2 = 0'4476191 - 2 \cdot \frac{0'952381}{3'809524} \cdot 0'0904759 + \left(\frac{0'952381}{3'809524}\right)^2 \cdot 1'761905$$

$$= 0'5125002$$

siendo de nuevo más preciso el estimador del total basado en un muestreo aleatorio simple.

CAPÍTULO 4

Estimadores de regresión lineal

Problema 4.1.- Se quiere estimar la media poblacional $\overline{Y}$ de una variable positiva Y, cuya varianza poblacional es $\sigma_y^2 = 10$, mediante una muestra aleatoria con reemplazamiento. Se dispone de la siguiente información adicional sobre otra variable X relacionada con la Y : $\overline{X} = 80$ y $\sigma_x^2 = 2$, siendo el coeficiente de correlación poblacional $\rho_{xy} = 0'7$. Se pide:

(a) ¿Para que valores de $\overline{Y}$ el estimador de razón para la media tiene menor error cuadrático medio que la media muestral $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$?.

(b) Si llamamos sesgo relativo de un estimador al cociente entre su sesgo y su esperanza, estudiar el sesgo relativo del estimador de razón para la media, expresándolo en función de ésta.

(c) Una muestra de tamaño $n = 10$ dió los siguientes resultados: $\overline{x} = 82$, $\overline{y} = 180'4$. Calcular las estimaciones $\overline{y}$, $\hat{\overline{Y}}_R$ e $\hat{\overline{Y}}_{rl}$, siendo $\hat{\overline{Y}}_R$ el estimador de razón e $\hat{\overline{Y}}_{rl}$ el estimador de regresión para la media; comparar los tres estimadores mediante sus errores cuadráticos medios.

(a) El error cuadrático medio de $\overline{y}$ es $V(\overline{y}) = \sigma_y^2/n$, y el error cuadrático medio de $\hat{\overline{Y}}_R$ es, al ser un muestreo con reemplazamiento,

$$\frac{N-n}{n(N-1)}\left[\sigma_y^2 - 2\,R\,\sigma_{xy} + R^2\,\sigma_x^2\right] = \frac{1}{n}\left[\sigma_y^2 - 2\,R\,\rho_{xy}\,\sigma_y\,\sigma_x + R^2\,\sigma_x^2\right].$$

Será

$$E.C.M.(\hat{\overline{Y}}_R) < E.C.M.(\overline{y}) \Longleftrightarrow \sigma_y^2 - 2\,R\,\rho_{xy}\,\sigma_y\,\sigma_x + R^2\,\sigma_x^2 < \sigma_y^2 \Longleftrightarrow$$

$$-2\,\rho_{xy}\,\sigma_y\,\sigma_x + R\,\sigma_x^2 < 0 \Longleftrightarrow R = \frac{\overline{Y}}{\overline{X}} < \frac{2\rho_{xy}\sigma_y}{\sigma_x} = \frac{2 \cdot 0'7\sqrt{10}}{\sqrt{2}} \Longleftrightarrow$$

$$\Longleftrightarrow \overline{Y} < 250'4396.$$

(b) Como es $\hat{\overline{Y}}_R = \overline{X}\,\hat{R}$, será $E[\hat{\overline{Y}}_R] = \overline{X}\,E[\hat{R}]$, con lo que

$$\text{sesgo}(\hat{\overline{Y}}_R) = \overline{X}\,E[\hat{R}] - \overline{Y} = \overline{X}\left(E[\hat{R}] - R\right) = \overline{X}\,\text{sesgo}(\hat{R})$$

de donde

$$\text{sesgo relat.}(\hat{\overline{Y}}_R) = \frac{\text{sesgo}(\hat{\overline{Y}}_R)}{\overline{Y}} = \frac{\overline{X}\text{sesgo}(\hat{R})}{\overline{Y}} = \frac{\text{sesgo}(\hat{R})}{R}.$$

Sabemos que

$$\text{sesgo}(\hat{R}) \approx \frac{N-n}{n(N-1)\overline{X}^2}\left[\frac{\overline{Y}}{\overline{X}}\sigma_x^2 - \rho_{xy}\cdot\sigma_x\cdot\sigma_y\right] =$$

al ser un muestreo con reemplazamiento,

$$= \frac{1}{n\overline{X}^2}\left[\frac{\overline{Y}}{\overline{X}}\sigma_x^2 - \rho_{xy}\cdot\sigma_x\cdot\sigma_y\right] = \frac{1}{n\overline{X}^2}\left[\frac{\overline{Y}}{80}\cdot 2 - 0'7\cdot\sqrt{2}\cdot\sqrt{10}\right]$$

con lo que

$$\text{sesgo relat.}(\hat{\overline{Y}}_R) = \frac{\overline{X}}{n\,\overline{Y}\,\overline{X}^2}\left[\frac{\overline{Y}}{40} - 3'13\right] = \frac{\overline{Y} - 125'22}{n\,\overline{Y}\,3.200}$$

(c) Será

$$\overline{y} = 180'4$$

$$\hat{\overline{Y}}_R = \overline{X}\hat{R} = 80\,\frac{\overline{y}}{\overline{x}} = 176$$

Por otro lado, como la recta de regresión poblacional es

$$y - \overline{Y} = \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - \overline{X})$$

el estimador de regresión para $\overline{Y}$ será

$$\hat{\overline{Y}}_{rl} = \overline{y} + \rho_{xy}\frac{\sigma_y}{\sigma_x}(\overline{X} - \overline{x})$$

que en nuestro caso toma un valor de

$$\hat{\overline{Y}}_{rl} = 180'4 + 0'7 \cdot \sqrt{\frac{10}{5}}(80 - 82) = 177'26951$$

Respecto a sus errores cuadráticos medios, éstos serán,

$$V(\overline{y}) = \frac{\sigma_y^2}{n} = 1$$

$$E.C.M.(\hat{\overline{Y}}_R) = \frac{1}{n}\left[\sigma_y^2 - 2\,\hat{R}\,\rho_{xy}\,\sigma_y\,\sigma_x + \hat{R}^2\,\sigma_x^2\right] =$$

$$= \frac{1}{10}\left[10 - 2 \cdot \frac{180'4}{82} \cdot 0'7 \cdot \sqrt{10} \cdot \sqrt{2} + \left(\frac{180'4}{82}\right)^2 \cdot 2\right] = 0'59058.$$

Por último, será

$$E.C.M.(\hat{\overline{Y}}_{rl}) = \frac{1}{n}\sigma_y^2(1 - \rho_{xy}^2) = \frac{1}{10} \cdot 10(1 - 0'7^2) = 0'51.$$

Problema 4.2.- En una población de tamaño $N = 256$ se están investigando dos variables fuertemente correladas X e Y.

Para tal fin se seleccionó una muestra aleatoria simple de tamaño $n = 100$ de la mencionada población que proporcionó los siguientes valores,

Cuasivarianza muestral de Y :	$\hat{S}_y^2 = 6409$
Cuasivarianza muestral de X :	$\hat{S}_x^2 = 3898$
Cuasicovarianza muestral :	$\hat{S}_{xy} = 4434$
Razón muestral :	$\hat{R} = 1'27$

Si el propósito es estimar la media de Y, comparar los estimadores de regresión lineal, de la razón y el de un muestreo aleatorio simple, mediante sus errores cuadrático medios.

El error cuadrático medio estimado de la media muestral en el muestreo aleatorio simple es su varianza estimada

$$\hat{V}(\overline{y}) = \frac{N-n}{N}\,\frac{\hat{S}_y^2}{n} = \frac{256-100}{256}\,\frac{6.409}{100} = 39'05$$

El del estimador de la razón será

$$E.C.M.(\hat{\overline{Y}}_R) = \frac{N-n}{n\,N}\left[\hat{S}_y^2 - 2\,\hat{R}\,\hat{S}_{xy} + \hat{R}^2\,\hat{S}_x^2\right] =$$

$$= \frac{256-100}{100\cdot 256}\left[6.409 - 2\cdot 1'27\cdot 4.434 + 1'27^2\cdot 3.898\right] = 8'74$$

es decir, sensiblemente menor, ya que aprovecha la información aportada por la variable X, fuertemente correlada con Y al ser $\hat{\rho}_{xy} = 0'887$.

Por último, el error cuadrático medio estimado del estimador de regresión lineal es

$$E.C.M.(\hat{\overline{Y}}_{rl}) = \frac{N-n}{n\,N}\,\hat{S}_y^2\left(1-\hat{\rho}_{xy}^2\right) =$$

$$= \frac{N-n}{n\,N}\,\hat{S}_y^2\left(1-\left[\frac{\hat{S}_{xy}}{\hat{S}_x\,\hat{S}_y}\right]^2\right) =$$

$$= \frac{256-100}{100\cdot 256}\cdot 6.409\left(1-\left[\frac{4.434}{\sqrt{3.898\cdot 6.409}}\right]^2\right) = 8'32$$

el cual mejora, ligeramente, al estimador de la razón.

Problema 4.3.- Un experimentado granjero es capaz de realizar estimaciones visuales del peso, X_i, de los melocotones de un árbol.

Con objeto de utilizar conjuntamente la experiencia del granjero y la estadística, éste estimó visualmente un peso total de $X = 11.600$ libras en un huerto de $N = 200$ árboles, seleccionándose además una muestra aleatoria simple de $n = 10$ árboles, para la que se obtuvieron los siguientes resultados sobre el peso real observado en la muestra, Y_i, y el peso estimado visualmente, X_i,

	Árbol número									
	1	2	3	4	5	6	7	8	9	10
Y_i	61	42	50	58	67	45	39	57	71	53
X_i	59	47	52	60	67	48	44	58	76	58

Se desea emplear el siguiente estimador del peso total real Y,

$$\hat{Y} = N\left[\overline{X} + (\overline{y} - \overline{x})\right]$$

siendo $\overline{X} = X/N$, y siendo $\overline{x}$ e $\overline{y}$ las medias muestrales de las observaciones X_i e Y_i respectivamente. Se pide:

(a) Determinar el error de muestreo del estimador $\hat{Y}$.

(b) Determinar, para los datos obtenidos, el valor de $\hat{Y}$ y de su error de muestreo.

(a) Si llamamos T al estimador $T = \overline{X} + (\overline{y} - \overline{x})$, es decir, a la "media muestral"

$$T = \frac{1}{n}\sum_{i=1}^{n} T_i = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X} + (Y_i - X_i)\right)$$

su varianza será la correspondiente a un muestreo aleatorio simple de T_i

$$\begin{aligned} V(T) &= \frac{N-n}{N}\frac{S_T}{n} \\ &= \frac{N-n}{N}\frac{1}{n}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{X} + Y_i - X_i - (\overline{X} + \overline{y} - \overline{x})\right)^2\right] \\ &= \frac{N-n}{N}\frac{1}{n}\left[\hat{S}_y^2 - 2\,\hat{S}_{xy} + \hat{S}_x^2\right] \end{aligned}$$

con lo que la varianza de $\hat{Y} = N\,T$ será

$$V(\hat{Y}) = \frac{N(N-n)}{n}\left[\hat{S}_y^2 - 2\,\hat{S}_{xy} + \hat{S}_x^2\right].$$

Su error de muestreo será la raíz cuadrada del valor anterior.

(b) Para los datos del problema queda,

$$\hat{Y} = N\left[\overline{X} + (\overline{y} - \overline{x})\right] = 200\,[58 + (54'3 - 56'9)] = 11.080$$

y

$$V(\hat{Y}) = \frac{200 \cdot 190}{10}\left[110'9 - 2\,\frac{897'3}{9} + 94'54\right] = 22.952.$$

Su error de muestreo será, por tanto, 151'5.

CAPÍTULO 5

Muestreo sistemático

Problema 5.1.- Una población de $N = 360$ casas (numeradas del 1 al 360) en Baltimore es ordenada alfabeticamente por la inicial del apellido del cabeza de la familia que vive en cada una de las casas. Los números de las casas en las cuales el cabeza de familia es una persona de color son: 28, 31-33, 36-41, 44, 45, 47, 55, 56, 58, 68, 69, 82, 83, 85, 86, 89-94, 98, 99, 101, 107-110, 114, 154, 156, 178, 223, 224, 296, 298-300, 302-304, 306-323, 325-331, 333, 335-339, 341, 342.

Comparar la precisión de un muestreo sistemático de período $k = 8$ con uno aleatorio simple del mismo tamaño muestral, en la estimación de la proporción de casas en las que el cabeza de familia es de color.

La varianza de la media muestral en un muestreo sistemático es

$$V_{sys} = \frac{N-1}{N} S^2 - \frac{k(n-1)}{N} \left[\frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)^2 \right]$$

Al estar estimando proporciones, ésta queda de la forma

$$V_{sys} = PQ - \frac{k(n-1)}{N} \left[\frac{1}{k} \sum_{i=1}^{k} \frac{n}{n-1} P_i Q_i \right] = PQ - \frac{n}{N} \sum_{i=1}^{k} P_i Q_i$$

en donde P_i es la "proporción poblacional" de cada una de las k "poblaciones" —posibles muestras.

De los datos del enunciado se desprende que es $n = N/k = 360/8 = 45$, $P = 81/360 = 0'225$ y además, $P_1 = 7/45$, $P_2 = 13/45$, $P_3 = P_4 = P_7 = P_8 = 10/45$, $P_5 = 12/45$ y $P_6 = 9/45$, con lo que será

$$V_{sys} = 0'225 \cdot 0'775 - \frac{45}{360}\left[\frac{7 \cdot 38 + 13 \cdot 32 + 10 \cdot 35 \cdot 4 + 12 \cdot 33 + 9 \cdot 36}{45^2}\right] =$$

$$= 0'0014121$$

Por otra parte, la varianza de la proporción muestral en un muestreo aleatorio simple es

$$V_{as} = \frac{N-n}{N-1} \cdot \frac{PQ}{n} = \frac{360-45}{359} \cdot \frac{0'225 \cdot 0'775}{45} = 0'0034$$

con lo que comparando ambas varianzas por cociente se obtiene

$$\frac{V_{as}}{V_{sys}} = \frac{0'0034}{0'0014121} = 2'41$$

Problema 5.2.- En una población formada por los individuos de 13 casas de una calle, se listaron las personas como sigue: *H=hombre adulto, M=mujer adulta, h=niño, m=niña,* obteniéndose los siguientes resultados

				Casa								
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
H	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*	*H*
M	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*	*M*
m	*m*	*h*		*h*	*m*	*m*	*h*	*h*	*h*	*m*	*m*	
h	*h*	*m*		*h*	*h*	*m*	*m*		*m*	*h*		
m	*m*			*m*		*h*						

Comparar las varianzas dadas por un muestreo sistemático de período $k = 5$ y uno aleatorio simple del mismo tamaño, para estimar, (a) la proporción de individuos del sexo masculino, (b) la proporción de personas no adultas, (c) la proporción de personas que viven en casas altas (las casas 1, 2, 3, 12 y 13 fueron descritas como altas).

Para la elección de la muestra sistemática los 50 individuos de la población se numeraron por columnas, empezando por la casa 1, y de arriba abajo:

$$H,\ M,\ m,\ h,\ m,\ H,\ M, \cdots,\ h,\ H,\ M,\ m,\ H,\ M.$$

Como la población está formada por $N = 50$ individuos y el muestreo sistemático tiene período $k = 5$, el tamaño de la muestra será $n = N/k = 10$.

Las cinco posibles muestras que se podrían extraer mediante un muestreo sistemático son

Muestra 1: *H, H, H, M, m, H, H, M, m, H*
Muestra 2: *M, M, M, H, H, M, M, h, H, M*
Muestra 3: *m, m, h, M, M, m, h, H, M, m*
Muestra 4: *h, h, m, h, m, m, m, M, m, H*
Muestra 5: *m, m, H, h, h, h, H, h, h, M*

(a) De los datos del enunciado se desprende que la proporción poblacional P de individuos del sexo masculino es $P = 24/50$, con lo que la varianza de la proporción muestral en un muestreo aleatorio simple será

$$V_{as} = \frac{N-n}{N-1} \cdot \frac{PQ}{n} = \frac{50-10}{49} \cdot \frac{24/50 \cdot 26/50}{10} = 0'0203755$$

siendo $Q = 1 - P$. Por otro lado, las proporciones de individuos del sexo masculino en cada una de las cinco muestras que se pueden extraer con un muestreo sistemático son

$$P_1 = 0'6 \ , \ P_2 = 0'4 \ , \ P_3 = 0'3 \ , \ P_4 = 0'4 \ , \ P_5 = 0'7$$

con lo que la varianza de la proporción muestral en un muestreo sistemático será (con $Q_i = 1 - P_i$)

$$\begin{aligned} V_{sys} &= PQ - \frac{n}{N} \sum_{i=1}^{k} P_i Q_i \\ &= \frac{24}{50} \cdot \frac{26}{50} - \frac{10}{50} \cdot (0'24 + 0'24 + 0'21 + 0'24 + 0'21) \\ &= 0'0216 \end{aligned}$$

siendo más preciso, por tanto, un muestreo aleatorio simple.

(b) La proporción de personas no adultas de la población es $P = 24/50$ —la misma del apartado anterior—, por lo que de nuevo será

$$V_{as} = \frac{50-10}{49} \cdot \frac{24/50 \cdot 26/50}{10} = 0'0203755.$$

Por otro lado, de las cinco muestras que se pueden extraer con un muestreo sistemático se obtienen las siguientes proporciones de personas no adultas

$$P_1 = 0'2 \ , \ P_2 = 0'1 \ , \ P_3 = 0'6 \ , \ P_4 = 0'8 \ , \ P_5 = 0'7$$

con lo que la varianza de la proporción muestral en un muestreo sistemático será

$$V_{sys} = \frac{24}{50} \cdot \frac{26}{50} - \frac{10}{50} \cdot (0'16 + 0'09 + 0'24 + 0'16 + 0'21) = 0'0776.$$

De nuevo el muestreo aleatorio simple es más preciso.

(c) Por último, como la proporción poblacional de personas que viven en casas altas es $P = 19/50$, la varianza de la proporción muestral en un muestreo aleatorio simple será

$$V_{as} = \frac{50-10}{49} \cdot \frac{19/50 \cdot 31/50}{10} = 0'0192326.$$

Como las proporciones de personas que viven en casas altas, en cada una de las cinco muestras posibles de un muestreo sistemático, son

$$P_1 = 0'4 \ , \ P_2 = 0'4 \ , \ P_3 = 0'4 \ , \ P_4 = 0'4 \ , \ P_5 = 0'3$$

la varianza de la proporción muestral en un muestreo sistemático será

$$V_{sys} = \frac{19}{50} \cdot \frac{31}{50} - \frac{10}{50} \cdot 1'17 = 0'0016$$

con lo que, en este caso, se obtendrá mayor precisión con un muestreo sistemático.

CAPÍTULO 6

Muestreo por conglomerados

Problema 6.1.- Suponiendo que en una comunidad de 300.000 personas la proporción de desempleados es 0'08 y que la población se encuentra repartida en 600 conglomerados de 500 personas cada uno, determine el tamaño de una muestra de conglomerados que garantice la estimación del total de desempleados con un error relativo de muestreo del 5 por ciento, sabiendo que en una encuesta previa la varianza dentro de los conglomerados era de

$$\sigma_w^2 = \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2 = \frac{1}{N}\sum_{i=1}^{N} P_i\, Q_i = 0'072$$

Llamaremos N al número de conglomerados en los que se ha dividido la población, n al tamaño de la muestra, es decir al número de conglomerados elegidos al azar —mediante un muestreo aleatorio simple en todo el capítulo—, M al tamaño de los conglomerados —que en todo el capítulo supondremos de igual dimensión—, $P_i = A_i/M$ la proporción de individuos del conglomerado (*cluster*) i-ésimo que presentan la característica en estudio, $P = \sum_{i=1}^{N} P_i/N$ la proporción poblacional, y $A = N\ M\ P$ el total de clase poblacional.

Con esta notación, la descomposición de la varianza

$$N\ M\ P\ Q = M\sum_{i=1}^{N}(P_i - P)^2 + M\sum_{i=1}^{N} P_i\, Q_i$$

es decir

$$P\ Q = \frac{1}{N}\sum_{i=1}^{N}(P_i - P)^2 + \frac{1}{N}\sum_{i=1}^{N} P_i\, Q_i$$

la expresaremos de la forma

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

De ahí y de los datos del enunciado se deduce que debe ser

$$\sigma_b^2 = \sigma^2 - \sigma_w^2 = 0'08 \cdot 0'92 - 0'072 = 0'0016$$

Por otro lado, como el estimador de la proporción (recuerdese que en los n conglomerados seleccionados se observan todos los individuos. De ahí que sea una media muestral de las proporciones poblacionales de los n conglomerados elegidos —en nuestro caso— mediante un muestreo aleatorio simple)

$$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} P_i$$

tiene varianza

$$V(\hat{P}) = \frac{N-n}{N\,(N-1)\,n} \sum_{i=1}^{N} (P_i - P)^2 = \frac{N-n}{(N-1)\,n}\,\sigma_b^2$$

el estimador insesgado del total de clase, $\hat{A} = N \cdot M \cdot \hat{P}$, tendrá varianza

$$V(\hat{A}) = (N \cdot M)^2\, V(\hat{P}) = (N \cdot M)^2\, \frac{N-n}{(N-1)\,n}\,\sigma_b^2$$

La condición exigida por el enunciado es que el coeficiente de variación o error relativo de muestreo sea del 5%, lo que lleva a la ecuación

$$\left(C.V.(\hat{A})\right)^2 = \frac{V(\hat{A})}{A^2} = (0'05)^2$$

es decir,

$$(N \cdot M)^2\, \frac{N-n}{(N-1)\,n}\,\sigma_b^2 = A^2 \cdot (0'05)^2$$

o bien

$$(300.000)^2 \cdot \frac{600-n}{599 \cdot n} \cdot 0'0016 = (300.000 \cdot 0'08)^2 \cdot (0'05)^2$$

ecuación de la que se obtiene el valor $n = 85'837$. Tomando un tamaño muestral de $n = 86$ conglomerados cumpliremos los requisitos del enunciado.

Problema 6.2.- Sea una población finita $\mathcal{A}$ = {a,b,c,d,e,f} en la que está definida una variable que toma respectivamente los valores 4,6,2,4,8,6. Supongamos que la población se divide en tres grupos, $G_1 = \{a,b\}, G_2 = \{c,d\}$ y $G_3 = \{e,f\}$.

1. Se efectua un muestreo que consiste en elegir uno de estos grupos al azar, observando los dos individuos del grupo seleccionado.

   (a) ¿Es la media muestral $\overline{y}$ un estimador centrado de la media poblacional $\overline{Y}$?.

   (b) Calcular la varianza de $\overline{y}$ y comprobar que en este caso es mayor que si se efectuase un muestreo aleatorio simple de $\mathcal{A}$ del mismo tamaño. ¿Qué le sucede al coeficiente de correlación intraconglomerados?.

2. Si el procedimiento de muestreo consiste en elegir un elemento al azar de cada grupo,

   (a) ¿Es la media muestral $\overline{y}$ un estimador centrado de la media poblacional $\overline{Y}$?.

   (b) Calcular la varianza de $\overline{y}$ y comprobar que en este caso es menor que si se hubiese efectuado un muestreo aleatorio simple del mismo tamaño.

3. Se define el estimador $\hat{\overline{Y}} = \frac{1}{3}y_1 + \frac{1}{3}y_2 + \frac{1}{3}y_3$, en donde y_i representa el valor de la variable del elemento seleccionado en el grupo $i = 1,2,3$, por el procedimiento de muestreo expuesto en el apartado (2). Se pide,

   (a) ¿Es el estimador $\hat{\overline{Y}}$ centrado para $\overline{Y}$?.

   (b) Determinar la varianza de $\hat{\overline{Y}}$.

   (c) Determinar el tamaño de muestra necesario para que con el procedimiento de muestreo expuesto en el apartado (2), la varianza de $\hat{\overline{Y}}$ sea igual a 1/6.

   (d) Con el tamaño muestral calculado en el apartado anterior, determinar el número de elementos a elegir en cada grupo para que la varianza de $\hat{\overline{Y}}$ sea mínima.

(e) Si el muestrear con un elemento de G_1 nos cuesta 100 pts, el muestrear uno de G_2, 200 pts y el muestrear uno de G_3, 300 pts, y si suponemos que contamos con un presupuesto total de 725 pts, ¿cuál debe de ser el número de elementos a elegir en cada grupo, con el mismo tamaño muestral calculado en (c), de modo que la $V(\hat{\overline{Y}})$ sea mínima?.

(1) La media poblacional es

$$\overline{Y} = 4 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 5$$

(a) Como las muestras posibles (y los valores de la media muestral en cada una de ellas) son

$$\left.\begin{array}{l} \omega_1 = (a,b) \\ \omega_2 = (c,d) \\ \omega_3 = (e,f) \end{array}\right\} \xrightarrow{\overline{y}} \left\{\begin{array}{lll} 5 & \text{con probabilidad} & 1/3 \\ 3 & \text{con probabilidad} & 1/3 \\ 7 & \text{con probabilidad} & 1/3 \end{array}\right.$$

será

$$E[\overline{y}] = 5 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} + 7 \cdot \frac{1}{3} = 5$$

es decir, centrado. De hecho lo que se nos propone es un muestreo por conglomerados, de ahí que la respuesta no sea sorprendente.

(b) Como es

$$E\left[\overline{y}^2\right] = 25 \cdot \frac{1}{3} + 9 \cdot \frac{1}{3} + 49 \cdot \frac{1}{3} = 83/3$$

será

$$V(\overline{y}) = E\left[\overline{y}^2\right] - (E[\overline{y}])^2 = 83/3 - 25 = 8/3.$$

Por otro lado, en un muestreo aleatorio simple de $\mathcal{A}$ de tamaño dos, la varianza de $\overline{y}_{as}$ es

$$V(\overline{y}_{as}) = \frac{N-n}{N} \cdot \frac{S^2}{n}$$

Como la varianza poblacional es

$$\sigma^2 = \frac{1}{6}\sum_{i=1}^{6} Y_i^2 - \left(\frac{1}{6}\sum_{i=1}^{6} Y_i\right)^2 = 172/6 - 25 = 11/3$$

será

$$S^2 = \frac{N\sigma^2}{N-1} = 22/5$$

y por tanto,

$$V(\overline{y}_{as}) = \frac{6-2}{6} \cdot \frac{22/5}{2} = 22/15 < 8/3.$$

Por último, el coeficiente de correlación intraconglomerados es

$$\delta = \frac{1}{(M-1)(NM-1)S^2}\left\{\sum_{i=1}^{N}\sum_{j\neq k}^{M}(Y_{ij}-\overline{Y})(Y_{ik}-\overline{Y})\right\} =$$

$$= \frac{1}{1\cdot 5\cdot 4'4}\{(4-5)(6-5)+(2-5)(4-5)+(8-5)(6-5)\} = 5/22$$

al ser $M = 2$, $N = 3$, $\overline{Y} = 5$ y $S^2 = 4'4$.

El hecho de que sea $\delta > 0$ indica que es peor agrupar, como observamos más arriba al comparar las varianzas.

(2) En este caso las muestras posibles son

$$\left.\begin{array}{llll} \omega_1 = (a,c,e) & ; & \omega_5 = (b,c,e) \\ \omega_2 = (a,c,f) & ; & \omega_6 = (b,c,f) \\ \omega_3 = (a,d,e) & ; & \omega_7 = (b,d,e) \\ \omega_4 = (a,d,f) & ; & \omega_8 = (b,d,f) \end{array}\right\} \xrightarrow{\overline{y}} \left\{\begin{array}{lll} 14/3 & ; & 16/3 \\ 12/3 & ; & 14/3 \\ 16/3 & ; & 18/3 \\ 14/3 & ; & 16/3 \end{array}\right.$$

(a) Como todas las muestras tienen la misma probabilidad de ser seleccionadas, será

$$E[\overline{y}] = \frac{1}{8}\sum_{i=1}^{8}\overline{y}(\omega_i) = 5 = \overline{Y}$$

es decir, centrado.

(b) Analogamente será $E[\overline{y}^2] = 1.824/72$, con lo que

$$V(\overline{y}) = 1/3 < 22/15 = V(\overline{y}_{as})$$

(3) Los grupos pueden ser considerados como estratos y el estimador propuesto

$$\hat{\overline{Y}} = \frac{1}{3}y_1 + \frac{1}{3}y_2 + \frac{1}{3}y_3$$

el correspondiente al estimador de la media en dicho muestreo, ya que puede expresarse de la forma

$$\hat{\overline{Y}} = \sum_{h=1}^{3} W_h \overline{y}_h$$

con $W_h = N_h/N = 1/3$. Será, por tanto, centrado. De hecho, el muestreo que se propone aquí es el mismo del apartado (2). Allí se hicieron los cálculos directamente y aquí se utilizan resultados del muestreo estratificado.

(b) Respecto a su varianza, al corresponder a un caso de afijación proporcional, será

$$V(\hat{\overline{Y}}) = \frac{N-n}{N \cdot n} \sum_{h=1}^{3} W_h S_h^2$$

Como es

$$\sigma_1^2 = \frac{1}{2}(16+36) - 25 = 1$$

$$\sigma_2^2 = \frac{1}{2}(4+16) - 9 = 1$$

$$\sigma_3^2 = \frac{1}{2}(64+36) - 49 = 1$$

será

$$S_h^2 = 2\,\sigma_h^2 = 2\,, \quad \forall h = 1,2,3$$

con lo que

$$V(\hat{\overline{Y}}) = \frac{6-3}{6 \cdot 3} \cdot \frac{1}{3} \cdot 2 \cdot 3 = \frac{1}{3}$$

(c) Para este tipo de muestreo estratificado el tamaño de muestra requerido será,

$$n = \frac{\sum_{h=1}^{3} W_h S_h^2}{V + 1/N \sum_{h=1}^{3} W_h S_h^2} = \frac{1/3 \cdot 2 \cdot 3}{1/6 + 1/6 \cdot 1/3 \cdot 2 \cdot 3} = 4.$$

(d) La *afijación* requerida será

$$n_h = \frac{N_h S_h}{\sum N_h S_h} \cdot n = \frac{2\sqrt{2}}{3 \cdot 2\sqrt{2}} \cdot 4 = 4/3$$

$\forall h = 1, 2, 3.$

(e) La *afijación para costes variables* proporciona unos valores

$$n_h = \frac{N_h S_h / \sqrt{c_h}}{\sum_1^3 (N_h S_h / \sqrt{c_h})} \cdot n$$

y sustituyendo se obtiene, $n_1 = 1'75$, $n_2 = 1'25$, y $n_3 = 1$.

CAPÍTULO 7

Muestreo con probabilidades desiguales

Los resultados que se utilizan en los problemas de este capítulo pueden encontrarse, por ejemplo, en el capítulo 9 del libro de *Azorín y Sánchez-Crespo*, Métodos y Aplicaciones del muestreo, 1986, Alianza editorial.

Problema 7.1.- Demostrar las siguientes relaciones para las probabilidades de selección en un muestreo sin reemplazamiento y probabilidades desiguales de tamaño n, de una población de tamaño N

$$\sum_{i=1}^{N} f_i = n$$

$$\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} f_{ij} = n(n-1)$$

siendo f_i la probabilidad de que el elemento poblacional U_i pertenezca a la muestra y f_{ij} la de que pertenezca el par (U_i, U_j).

Si consideramos la variable, $i = 1, ..., n$

$$e_i = \begin{cases} 1 & \text{si } U_i \text{ pertenece a la muestra} \\ 0 & \text{si } U_i \text{ no pertenece a la muestra} \end{cases}$$

será $e_i \rightsquigarrow B(1, f_i)$ y por tanto $E[e_i] = f_i$. En consecuencia,

$$\sum_{i=1}^{N} f_i = \sum_{i=1}^{N} E[e_i] = E\left[\sum_{i=1}^{N} e_i\right] = E[n] = n.$$

Analogamente, definiendo la variable

$$e_{ij} = \begin{cases} 1 & \text{si } (U_i, U_j) \text{ pertenece a la muestra} \\ 0 & \text{si } (U_i, U_j) \text{ no pertenece a la muestra} \end{cases}$$

será $e_{ij} \rightsquigarrow B(1, f_{ij})$ y en consecuencia $E[e_{ij}] = f_{ij}$. Por tanto,

$$\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} f_{ij} = E\left[\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} e_{ij}\right] = E[n(n-1)] = n(n-1).$$

Problema 7.2.- Sea una población $\mathcal{P} = \{U_1, U_2, U_3\}$ con probabilidades de selección asociadas $P_i = \{1/6\,,\, 2/6\,,\, 3/6\}$ respectivamente. Se obtienen todas las muestras posibles S_j de $n = 2$ unidades sin reemplazamiento y probabilidades P_i desiguales. Se pide:

(a) La probabilidad $P(S_j)$ para todos los valores j.

(b) La probabilidad que cada U_i tiene de ser elegida.

(a) Hay tres muestras posibles,

$$S_1 = \{U_1, U_2\}\ ,\ S_2 = \{U_1, U_3\}\ ,\ S_3 = \{U_2, U_3\}$$

con probabilidades respectivas,

$$P(S_1) = P(U_1) \cdot P(U_2/U_1) + P(U_2) \cdot P(U_1/U_2) = \frac{1}{6} \cdot \frac{2}{5} + \frac{2}{6} \cdot \frac{1}{4} = \frac{3}{20}$$

al tener que mantenerse las probabilidades asociadas a los elementos de la población en cada extracción, lo cual significa, por ejemplo, que $P(U_3/U_1) = (3/2)\,P(U_2/U_1)$, y ser además $P(U_3/U_1) + P(U_2/U_1) = 1$; de donde se deduce que $P(U_2/U_1) = 2/5$. Las demás probabilidades condicionadas se calculan de igual manera.

$$P(S_2) = P(U_1) \cdot P(U_3/U_1) + P(U_3) \cdot P(U_1/U_3) = \frac{1}{6} \cdot \frac{3}{5} + \frac{3}{6} \cdot \frac{1}{3} = \frac{8}{30}$$

$$P(S_3) = P(U_2) \cdot P(U_3/U_2) + P(U_3) \cdot P(U_2/U_3) = \frac{2}{6} \cdot \frac{3}{4} + \frac{3}{6} \cdot \frac{2}{3} = \frac{7}{12}$$

(b) Las probabilidades pedidas serán,

$$P(U_1 \text{ pertenezca a la muestra}) = P(S_1) + P(S_2) = \frac{5}{12}$$

$$P(U_2 \text{ pertenezca a la muestra}) = P(S_1) + P(S_3) = \frac{11}{15}$$

$$P(U_3 \text{ pertenezca a la muestra}) = P(S_2) + P(S_3) = \frac{17}{20}$$

Problema 7.3.- Una población finita consta de seis elementos. Se divide en tres grupos: $G_1 = \{A_1, A_2\}$, $G_2 = \{A_3, A_4\}$ y $G_3 = \{A_5, A_6\}$. Se extrae una muestra de tamaño tres eligiendo un elemento de cada grupo, con probabilidades: P(A_1)=1/3 y P(A_2)=2/3; P(A_3)=1/2 y P(A_4)=1/2; P(A_5)=2/5 y P(A_6)=3/5.

Supongamos se ha obtenido en particular la muestra $\{A_2, A_3, A_6\}$ y que los correspondientes valores de la variable en estudio son 100, 80 y 120. Se pide:

(a) Una estimación centrada del total de la población a partir de esa muestra.

(b) Una estimación centrada de la varianza del estimador utilizado.

(a) Si supuesto se seleciona la terna $\omega = (A_{i_1}, A_{i_2}, A_{i_3})$ llamamos $Y(\omega) = (Y_1, Y_2, Y_3)$ al valor de la variable en la terna selecionada, con probabilidades respectivas (π_1, π_2, π_3), un estimador insesgado del total poblacional es

$$\hat{Y} = \sum_{i=1}^{3} \frac{Y_i}{\pi_i}.$$

Una estimación insesgada para el problema que nos ocupa será, por tanto,

$$\hat{Y} = 100 \cdot \frac{3}{2} + 80 \cdot 2 + 120 \cdot \frac{5}{3} = 510$$

(b) El estimador pedido es

$$\hat{V}(\hat{Y}) = \sum_{i=1}^{3} \left(\frac{Y_i}{\pi_i}\right)^2 (1 - \pi_i) + \sum_{i \neq j} \frac{Y_i}{\pi_i} \cdot \frac{Y_j}{\pi_j} \cdot \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_{ij}}$$

Como en nuestro caso es $\pi_{ij} = \pi_i \cdot \pi_j$ se anulará el segundo sumando, quedando, por tanto

$$\hat{V}(\hat{Y}) = \left(100 \cdot \frac{3}{2}\right)^2 \cdot \frac{1}{3} + (80 \cdot 2)^2 \cdot \frac{1}{2} + \left(120 \cdot \frac{5}{3}\right)^2 \cdot \frac{2}{5} = 36.300$$

Problema 7.4.- Sea P la proporción de elementos de una determinada clase y P_h la correspondiente al estratato h, de una población estratificada. Se extrae una muestra de tamaño 2 por el siguiente procedimiento aleatorio:

1) Se elige un estrato aleatoriamente, el h, con probabilidad π_h. $(0 < \pi_h < 1, \sum_h \pi_h = 1)$.

2) En el estrato seleccionado se extrae una muestra aleatoria simple de tamaño 2.

Se pide:

(a) Calcular la ley de probabilidad de la muestra. (Es decir, si A es pertenecer a la clase y A^* no pertenecer, determinar la probabilidad de los posibles resultados de la muestra).

(b) Calcular la esperanza del estadístico $\hat{P}$=*proporción de elementos del tipo A en la muestra.*

(c) Si en la fase 1) de selección de la muestra, se asigna a los estratos probabilidades proporcionales a los tamaños, estudiar si el resultante $\hat{P}$ es insesgado o no.

(a) Las muestras posibles son

$$\left.\begin{matrix}(A,A)\\(A^*,A^*)\\(A,A^*)\\(A^*,A)\end{matrix}\right\} \text{ con probabilidades } \begin{cases} \sum_h \pi_h \cdot P_h \cdot \dfrac{P_h N_h - 1}{N_h - 1} \\ \sum_h \pi_h \cdot (1-P_h) \cdot \dfrac{(1-P_h) N_h - 1}{N_h - 1} \\ \sum_h \pi_h \cdot P_h \cdot \dfrac{(1-P_h) N_h}{N_h - 1} \\ \sum_h \pi_h \cdot (1-P_h) \cdot \dfrac{P_h N_h}{N_h - 1} \end{cases}$$

(b) $\hat{P}$ es una variable aleatoria con distribución

$$\hat{P} = \begin{cases} 0 & \text{con probabilidad } \sum_h \pi_h \cdot (1-P_h) \cdot \dfrac{(1-P_h)\,N_h - 1}{N_h - 1} \\ 1/2 & \text{con probabilidad } 2\sum_h \pi_h \cdot P_h \cdot \dfrac{(1-P_h)\,N_h}{N_h - 1} \\ 1 & \text{con probabilidad } \sum_h \pi_h \cdot P_h \cdot \dfrac{P_h\,N_h - 1}{N_h - 1} \end{cases}$$

Su esperanza será, por tanto,

$$E[\hat{P}] = 0 + \frac{1}{2} \cdot 2 \sum_h \pi_h \cdot P_h \cdot \frac{(1-P_h)\,N_h}{N_h - 1} + 1 \cdot \sum_h \pi_h \cdot P_h \cdot \frac{P_h\,N_h - 1}{N_h - 1} = \sum_h \pi_h\,P_h$$

(c) Si ahora es $\pi_h = N_h/N$, será $E[\hat{P}] = \sum_h W_h\,P_h = P$, es decir, centrado.

CAPÍTULO 8

Muestreo polifásico

Problema 8.1.- 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 obtuvo una muestra aleatoria simple de $n = 200$ establecimientos, a los que se solicitó por correo este dato. Después de varios recordatorios se consiguió una respuesta del 75% con

$$\overline{y}_1 = 20.000 \text{ Pts} \qquad \hat{S}_1^2 = 1.400.000$$

Con posterioriddad se enviaron agentes entrevistadores a una submuestra de 20 establecimientos elegidos por el mismo procedimiento anterior, de la lista formada por los establecimientos que no contestaron, obteniéndose los siguientes resultados:

$$\overline{y}_2 = 16.000 \text{ Pts} \qquad \hat{S}_2^2 = 1.600.000$$

Se pide:

(a) Estimación del salario medio mensual por establecimiento.

(b) Varianza estimada para el estimador de la media antes considerado, $\overline{y}$, siendo la cuasivarianza muestral total

$$\hat{S}^2 = \sum_{h=1}^{2} \hat{W}_h (\overline{y}_h - \overline{y})^2 + \sum_{h=1}^{2} \hat{W}_h \hat{S}_h^2$$

y siendo respectivamente $\overline{y}_h$ y $\hat{S}_h^2$ las medias y cuasivarianzas muestrales antes mencionadas.

Es un caso de muestreo doble estratificado, en donde en una primera fase se estiman los estratos y a continuación se realiza un muestreo estratificado

dentro de la muestra inicial. En nuestro caso, solo en el segundo estrato (el de los que no respondieron).

(a) El estimador de la media es,

$$\overline{y} = \sum_{h=1}^{2} \hat{W}_h \overline{y}_h$$

siendo en nuestro caso,

$$\hat{W}_1 = \frac{n_1}{n} = \frac{150}{200} = 0'75 \ , \ \overline{y}_1 = 20.000$$

$$\hat{W}_2 = \frac{n_2}{n} = \frac{50}{200} = 0'25 \ , \ \overline{y}_2 = 16.000$$

con lo que será

$$\overline{y} = 0'75 \cdot 20.000 + 0'25 \cdot 16.000 = 19.000$$

(b) La varianza muestral total es,

$$\hat{S}^2 = \sum_{h=1}^{2} \hat{W}_h (\overline{y}_h - \overline{y})^2 + \sum_{h=1}^{2} \hat{W}_h \hat{S}_h^2 =$$

$$= 0'75(20.000 - 19.000)^2 + 0'25(16.000 - 19.000)^2 + 0'75 \cdot 1.400.000 +$$

$$+0'25 \cdot 1.600.000 = 4.450.000$$

con lo que la varianza estimada para $\overline{y}$ será,

$$\hat{V}(\overline{y}) = \sum_{h=1}^{2} \frac{1 - \lambda_h}{n\lambda_h} \hat{W}_h \hat{S}_h^2 + \frac{N - n}{nN} \hat{S}^2$$

siendo $\lambda_h = n_{h1}/n_h$ la fracción de muestreo para cada estrato en la segunda fase. En nuestro caso es $\lambda_1 = 1$ y $\lambda_2 = 20/50$, con lo que la varianza estimada para $\overline{y}$ quedará

$$= \frac{1 - 2/5}{200 \cdot 2/5} \cdot 0'25 \cdot 1.600.000 + \frac{2.000 - 200}{200 \cdot 2.000} \cdot 4.450.000 = 23.025$$

Problema 8.2.- Para estimar la media de una población se realizó una encuesta por correo seleccionando una muestra aleatoria simple de n unidades elementales.

De las $n_2 = n - n_1$ unidades que no contestaron se eligió, siguiendo el mismo procedimiento, una muestra de n_{21} $(= f_{21} \cdot n_2)$ unidades de las que se obtuvo la información requerida mediante entrevista.

Un estudio piloto previo proporcionó la siguiente información:

1. El cociente entre la varianza de las unidades que no respondieron y la varianza total fue

$$\frac{\sigma_2^2}{\sigma^2} = 4'65$$

2. El coste unitario de envío por correo fue $c_0 = 25$ pts. El de proceso para cada una de las n_1 unidades que contestaron fue de $c_1 = 20$ pts y el de cada unidad entrevistada $c_2 = 540$ pts.

3. La tasa de no respuesta en el envío por correo fue del 10%, $(W_2 = 1 - W_1 = 0'1)$.

Se pide:

(a) Demostrar que el coste total esperado de la encuesta es

$$E[C] = n\,(c_0 + c_1 \cdot W_1 + c_2 \cdot f_{21} \cdot W_2)$$

(b) Calcular los valores óptimos de n y f_{21} que minimizarían este coste total esperado 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 con el mismo procedimiendo en el caso de no existir falta de respuesta.

La función de costo será

$$C = c_0\, n + c_1\, n_1 + c_2\, n_{21}$$

la cual puede escribirse de la forma

$$C = n\left(c_0 + c_1\,\frac{n_1}{n} + c_2\,\frac{n_{21}}{n_2}\cdot\frac{n_2}{n}\right)$$

siendo su coste esperado,

$$E[C] = n\,(c_0 + c_1\,W_1 + c_2\,f_{21}\,W_2)$$

con $W_i = N_i/N$.

Por otro lado, la varianza del estimador de la media es

$$V(\hat{\bar{X}}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n} + W_2\left(\frac{1}{f_{21}} - 1\right)\frac{S_2^2}{n}$$

que para el caso de muestreo con reemplazamiento queda,

$$V(\hat{\bar{X}}) = \frac{\sigma^2}{n} + W_2\left(\frac{1}{f_{21}} - 1\right)\frac{\sigma_2^2}{n}$$

Queremos determinar los valores de n y f_{21} que minimicen el coste esperado sujeto a la restricción de ser $V(\hat{\bar{X}}) = V_0(\hat{\bar{X}})$, con

$$V_0(\hat{\bar{X}}) = \frac{\sigma^2}{100}$$

Aplicando el método de los multiplicadores de Lagrange a la función

$$\phi = n\,(c_0 + c_1\,W_1 + c_2\,f_{21}\,W_2) + \lambda\left[\frac{\sigma^2}{n} + W_2\left(\frac{1}{f_{21}} - 1\right)\frac{\sigma_2^2}{n} - V_0(\hat{\bar{X}})\right]$$

y derivando ahora esta función respecto a n, y a f_{21} e igualando ambas a cero, se obtiene el sistema

$$\frac{\partial\phi}{\partial n} = c_0 + c_1\,W_1 + c_2\,f_{21}\,W_2 - \lambda\left[\frac{\sigma^2}{n^2} + W_2\left(\frac{1}{f_{21}} - 1\right)\frac{\sigma_2^2}{n^2}\right] = 0$$

$$\frac{\partial\phi}{\partial f_{21}} = n\,c_2\,W_2 - \lambda\frac{W_2}{f_{21}^2}\frac{\sigma_2^2}{n} = 0$$

De este par de ecuaciones se obtiene la solución

$$f_{21} = \sqrt{\frac{c_0 + c_1\,W_1}{c_2\left[\frac{1}{4'65} - W_2\right]}}$$

Y de la restricción igualada a cero

$$\frac{\sigma^2}{n} + W_2 \left(\frac{1}{f_{21}} - 1 \right) \frac{4'65 \cdot \sigma^2}{n} = \frac{\sigma^2}{100}$$

un valor para n,

$$n = 100 \left[1 + 4'65\, W_2 \left(\frac{1}{f_{21}} - 1 \right) \right]$$

Aplicando estas dos fórmulas a los datos del problema se obtienen los valores solicitados, $f_{21}' = 0'8319$ y $n = 109'396$.