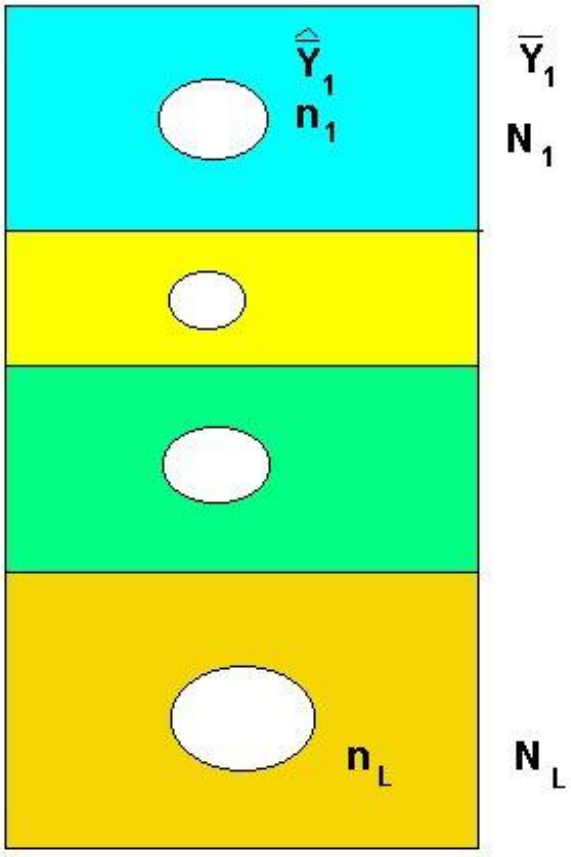
3. MUESTREO ESTRATIFICADO

3.1 Definición y Notación

Hasta este momento se ha considerado a la población y a la muestra como conjuntos de elementos con cierta homogeneidad, sin embargo, en ocasiones es conveniente dividir a la población en subpoblaciones o estratos. Los estratos se forman en función de variables altamente correlacionadas con las variables en estudio, como nivel socioeconómico, tamaño de la localidad, giro de empresas, etc.



Los elementos que se incluyen en cada estrato, se procura que sean homogéneos con respecto a las características que se investigan para obtener mayor eficiencia en el diseño.

Las principales ventajas que se tienen con la estratificación son las siguientes:

a) Utilizar información previa a la población para reducir el error de muestreo, esto es, ganar precisión en las estimaciones debido a que los elementos en cada estrato tienen cierto grado de homogeneidad.

b) Es posible dividir la población en estratos que coincidan con divisiones geográficas o administrativas para las cuales se requieren estimaciones separadas del total, esto es que los estratos pueden ser dominios de estudio.

c) La estratificación permite hacer compensaciones en diseños de muestreo menos eficientes, como por ejemplo el muestreo por conglomerados.

d) Desde el punto de vista logístico se pueden designar delegados que supervisen y controlen la encuesta en cada región ó estrato.

La notación correspondiente al muestreo estratificado es la siguiente:

$N = \sum_{h=1}^{L} N_h$ Total de unidades en la población

L Número de estratos

N_h Total de unidades en el estrato h

$n = \sum_{h=1}^{L} n_h$ Tamaño total de la muestra

n_h Total de unidades en la muestra del estrato h

y_{hi} El valor de la característica investigada en la i-ésima unidad del estrato h

$W_h = \frac{N_h}{N}$ El peso ó ponderación del estrato h

$f_h = \frac{n_h}{N_h}$ Fracción de muestreo en el estrato h

$Y_h = \sum_{i=1}^{Nh} y_{hi}$ Total del estrato h

$\overline{Y} = \frac{\sum_{i=1}^{Nh} y_{hi}}{N_h}$ Media del estrato h

$Y_{st} = \sum_{h=1}^{L} \sum_{i=1}^{Nh} y_{hi}$ Total de la población

$\overline{Y}_{st} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{Nh} y_{hi}}{N}$ Media de la población total

$$yh = \frac{\sum_{i=1}^{nh} y_{hi}}{n_h}$$ Media muestral del estrato h.

$$S_h^2 = \frac{\sum_{i=1}^{Nh} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$ Medida de variación del h-ésimo estrato

$$s_h^2 = \frac{\sum_{i=1}^{nh} (y_{hi} - y_h)^2}{n_h - 1}$$ Medida de variación de la muestra en el h-ésimo estrato

3.2 Estimadores en Muestreo Estratificado para Medias y Totales.

La media de la población total $\bar{Y}$ se puede expresar como la suma ponderada de las medias de todos los estratos:

$$\begin{aligned} \bar{Y}_{st} &= \frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh} y_{hi}}{N} \\ &= \sum_{h=1}^{L} \frac{1}{N} N_h \sum_{i=1}^{Nh} \frac{y_{hi}}{N_h} \\ &= \sum_{h=1}^{L} \frac{N_h}{N} \bar{Y}_h \end{aligned}$$

$$\boxed{\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h}$$

Para obtener el estimador de $\bar{Y}$ se sustituyen $\bar{Y}_h$ por sus correspondientes estimadores y_h

$$\boxed{\hat{\bar{Y}}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h}$$

En ocasiones tomaremos la notación alternativa de $\hat{\bar{Y}}_{st}$, como $\bar{y}_{st}$. Debe notar que el estimador anterior, en general, no coincide con la media de la muestra total, la cual tendría la siguiente expresión:

$$\boxed{y = \frac{1}{n} \sum_{h=1}^{L} \sum_{i=1}^{n_h} y_{hi}}$$

El estimador $\bar{y}$ coincidiría con $\bar{y}_{st}$ sólo en el caso de que se cumpla la siguiente relación de proporcionalidad para todos los estratos.

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

Si en cada uno de los L estratos $\bar{y}h$ es un estimador insesgado de $\overline{Y}h$ entonces $\bar{y}_{st}$ es un estimador insesgado de $\overline{Y}_{st}$.

$$\begin{aligned} E(y_{st}) &= E\sum_{h=1}^{L}\frac{N_h}{N}y_h \\ &= \sum_{h=1}^{L}\frac{N_h}{N}E(y_h) \\ &= \sum_{h=1}^{L}\frac{N_h}{N}\overline{Y}_h \\ &= \sum_{h=1}^{L}W_h\overline{Y}_h \\ &= \overline{Y}_{st} \end{aligned}$$

Como las muestras se obtienen de manera independiente en cada estrato, entonces la varianza del estimador de la media total se obtiene mediante la suma de los ponderadores de los estratos al cuadrado por las varianzas de los estimadores de las medias en los estratos.

$$\begin{aligned} V(y_{st}) &= V\left(\sum_{h=1}^{L}\frac{N_h}{N}y_h\right) \\ &= \sum_{h=1}^{L}\frac{N_h{}^2}{N^2}V(y_h) \\ &= \sum_{h=1}^{L}W_h{}^2V(y_h) \end{aligned}$$

Si se utiliza muestreo aleatorio simple en todos los estratos, la varianza del estimador de la media total $\bar{y}_{st}$ tiene la expresión siguiente, a la cual se designará como Forma General:

$$\boxed{V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{Nh^2Sh^2}{nh} - \frac{1}{N^2}\sum_{h=1}^{L}NhSh^2}$$

$$\begin{aligned} V(y_{st}) &= \sum_{h=1}^{L}W_h^2\,V(y_h) \\ &= \sum_{h=1}^{L}W_h^2\left\{1-\frac{\mathrm{n_h}}{\mathrm{N_h}}\right\}\frac{S_h^2}{n_h} \\ &= \sum_{h=1}^{L}\frac{N_h^2}{N^2}\left\{1-\frac{\mathrm{n_h}}{\mathrm{N_h}}\right\}\frac{S_h^2}{n_h} \\ &= \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2S_h^2}{n_h} - \frac{1}{N^2}\sum_{h=1}^{L}N_hS_h^2 \end{aligned}$$

Si se ignora el factor de corrección por finitud $\left\{1-\frac{n_h}{N_h}\right\}$ la expresión se simplifica.

$$\boxed{V(y_{st})=\frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h}}$$

Como en general Sh^2 es desconocido, para estimar la varianza se le sustituye por sh^2 muestral en las fórmulas mencionadas.

El error estándar se obtiene como la raíz cuadrada de la varianza y los intervalos de confianza para la media estratificada se calculan en forma análoga al caso de muestreo aleatorio simple.

$$\bar{y}_{st} \pm Z_{1-\alpha/2}\hat{E}E(\bar{y}_{st})$$

El estimador del total Y_{st} se obtiene mediante la multiplicación del estimador de la media total por el tamaño de la población.

$$\boxed{\hat{Y}_{st} = N\bar{y}_{st}}$$

La varianza del estimador del total se calcula mediante el producto del cuadrado del tamaño de la población total por la varianza de la media.

$$\boxed{V(\hat{Y}) = N^2 V(\bar{y}_{st})}$$

Ejemplo 3.1

En una región hay 12,789 productores de cereal. Los predios han sido divididos en función del uso de tecnología dominante en 3 estratos: Uso intensivo de tecnología, Uso medio de tecnología y Uso bajo de tecnología. Se tomó una muestra de 31 predios divididos como lo indica la tabla. En cada predio se midió el rendimiento en Toneladas por Ha.

a) Estimar el rendimiento medio por Ha. en la región y construir un intervalo de 90% de confianza.
b) Estimar la producción total en la región y construir un intervalo de 90%.

Número	Estrato 1	Estrato 2	Estrato 3
1	**5.06**	**1.6**	**1.28**
2	**3.66**	**2.33**	**2.43**
3	**4.02**	**3.46**	**4.48**
4	**4.82**	**3.67**	**1.19**
5	**4.27**	**1.93**	**4.23**
6	**3.32**	**2.55**	**0.23**
7	**2.19**	**1.58**	**4.10**
8	**4.1**	**4.09**	**2.99**
9	**1.93**	**2.26**	**3.62**
10		**4.35**	**6.36**
11			**2.82**
12			**2.27**

Cuadro de Cálculos

Estrato	N_h	W_h	n_h	y_h	$\hat{S}_h^2$	$W_h y_h$	$N_h^2 \hat{S}_h^2 / n_h$	$N_h \hat{S}_h^2$
Alto	**3,253**	**0.2543592**	**9**	**4.020**	**0.47920**	**1.0225**	**563,433.19**	**1,558.84**
Medio	**4,234**	**0.3310658**	**10**	**3.436**	**0.66632**	**1.1375**	**1,194,487.64**	**2,821.18**
Bajo	**5,302**	**0.4145750**	**12**	**2.543**	**0.88204**	**1.0544**	**2,066,272.88**	**4,676.59**
Total	**12,789**	**1.0000000**	**31**			**3.2145**	**3,824,193.71**	**9,056.61**

La media se obtiene mediante la fórmula $\hat{Y}_{st} = \sum_{h=1}^{L} W_h y_h$ = **3.2145**

Como la asignación de muestra en los estratos es arbitraria, se utiliza la fórmula general de la varianza.

$$V(y_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{Nh^2 Sh^2}{nh} - \frac{1}{N^2}\sum_{h=1}^{L} NhSh^2 = \frac{1}{(12{,}789)^2}(3{,}824{,}193{,}71) - \frac{1}{(12{,}789)^2}(9{,}056.61) = 0.02333$$

El error estándar se calcula al aplicar raíz cuadrada a la varianza estimada.

$$\hat{EE}(y_{st}) = \sqrt{0.02333} = 0.1527279$$

Para construir el intervalo de 90% de confianza se utiliza el valor percentilar 1.645 de la distribución normal considerando que **n** tiene el mínimo de unidades necesario para la aplicación de la normal estándar.

$y_{st} \pm Z_{1-\alpha/2}\hat{EE}(y_{st})$ Al sustituir se obtiene el intervalo 3.2145 ± 1.645(0.1527279)

El intervalo solicitado para el rendimiento tiene los límites: **(2.9632 , 3.4657)**

La estimación del total se obtiene al multiplicar la media general estimada por el tamaño de la población.

$\hat{Y} = N\bar{y}_{est}$ 12,789(3.2145) = 41,110.24

Los límites del intervalo se obtienen al sumar y restar al total estimado el producto del error estándar de la media por la población y el coeficiente de confianza

$\hat{Y} \pm Z_{1-\alpha/2}N(\hat{EE}(y_{st}))$ 41,110.24 ± 3,213.08

3.3 Fuentes de variación en muestreo estratificado.

Al tener a la población dividida es estratos, la variación total de la característica de interés se puede atribuir a dos fuentes: la variación dentro de estratos y la variación entre estratos.
Esto se puede observar mediante el siguiente análisis:

$$\sigma_{st}^2 = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{Nh}(y_{hi}-\bar{Y})^2$$

$$= \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{Nh}[(y_{hi}-\bar{Y}_h)+(\bar{Y}_h-\bar{Y})]^2$$

$$= \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{Nh}\left[(y_{hi}-\bar{Y}_h)^2+2(y_{hi}-\bar{Y}_h)(\bar{Y}_h-\bar{Y})+(\bar{Y}_h-\bar{Y})^2\right]$$

$$= \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{Nh}(y_{hi}-\bar{Y}_h)^2+\frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{Nh}(\bar{Y}_h-\bar{Y})^2$$

$$= \frac{1}{N}\sum_{h=1}^{L}N_h\sum_{i=1}^{Nh}\frac{(y_{hi}-\bar{Y}h)^2}{N_h}+\frac{1}{N}\sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2$$

$$= \underbrace{\frac{1}{N}\sum_{h=1}^{L}N_h\sigma_h^2}_{\text{Dentro de Estratos}}+\underbrace{\frac{1}{N}\sum_{h=1}^{L}N_h(\bar{Y}_h-\bar{Y})^2}_{\text{Entre Estratos}}$$

3.4 Afijación de la Muestra.

El objetivo siguiente es calcular el tamaño total de muestra **n** y distribuir este tamaño de muestra entre los L estratos. A este proceso se le conoce también como afijación de la muestra.

Si se supone que **n** es conocida, la pregunta que surge es: ¿Qué es una buena afijación?. Se entenderá por una buena afijación, aquella que proporcione máxima precisión para un nivel de confianza dado y de ser posible con el mínimo costo. Como la precisión está relacionada con la varianza, el se buscará minimizar la varianza.

Afijación de Igual Número en cada Estrato.

La forma más simple para asignar el tamaño de muestra correspondiente a cada estrato, es dividir el tamaño total de la muestra entre los L estratos. De este modo la expresión de $\mathbf{n_h}$ sería la siguiente:

$$n_h = \frac{n}{L}$$

La asignación de igual número en cada estrato es ineficiente, pero puede haber razones de otro tipo para su empleo.

Si se considera que la muestra total n se asigna según éste criterio, la fórmula de la varianza del estimador de la media total toma una expresión particular.

Se parte de la Forma General:

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

Al sustituir $n_h = \frac{n}{L}$, se tendrá la fórmula particular para la varianza:

$$\boxed{V(y_{st}) = \frac{L}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^{\ 2}}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2}$$

Determinación del tamaño de Muestra Total n para una Varianza fija D^2

Hasta este punto se ha supuesto que se conoce el tamaño de muestra n y no se ha mencionado la forma de obtenerlo. Para ello, se parte de la fórmula que relaciona precisión, confianza y varianza:

$$d^2 = Z^2_{(1-\alpha/2)} V(\bar{y}st)$$

Se despeja la varianza y se asignan valores a la precisión d y al coeficiente de confianza $Z^2_{(1-\alpha/2)}$. La varianza se iguala a una constante que se llamará D^2, la varianza deseada.

$$V(\bar{y}_{st}) = \frac{d^2}{Z^2_{(1-\alpha/2)}} = D^2$$

Se despejará n al sustituir $V(\bar{y}_{st})$ por D^2 según el criterio de afijación igual para cada estrato.

$$\mathrm{D}^2 = \frac{L}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2$$

$$\therefore \qquad \mathrm{N}^2\mathrm{D}^2 + \sum_{h=1}^{L} N_h S_h^2 = \frac{L}{n}\sum_{h=1}^{L} N_h^2 S_h^2$$

$$\boxed{\mathrm{n} = \frac{L\sum_{h=1}^{L} N_h^2 S_h^2}{N^2 D^2 + \sum_{h=1}^{L} N_h S_h^2}}$$

Afijación Proporcional al Tamaño del Estrato.

Se parte de una relación de proporcionalidad que iguala la razón del tamaño del estrato respecto al tamaño de la población con la razón del tamaño de muestra en el estrato respecto al tamaño total de la muestra.

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

De donde, al despejar n_h se tiene la fórmula de afijación: $n_h = \frac{N_h}{N} n$

A continuación se obtendrá la expresión correspondiente a la varianza de $\bar{y}_{st}$ al suponer afijación proporcional de la muestra.

Nuevamente se parte de la Forma General de la varianza de $\bar{y}_{st}$.

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

En ella se sustituye:

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

$$V(y_{st}) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{\frac{N_h}{N} n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2$$

$$\boxed{V(y_{st}) = \frac{1}{N}\sum_{h=1}^{L}\frac{N_h S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2}$$

El tamaño de muestra total n, se obtiene en forma semejante al caso anterior mediante una varianza deseada D^2 y se despeja n de la fórmula de la varianza de $\bar{y}_{st}$ con el criterio de afijación proporcional.

$$D^2 = \frac{1}{N}\sum_{h=1}^{L}\frac{N_h S_h^2}{n} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2$$

$$\therefore \quad N^2 D^2 + \sum_{h=1}^{L} N_h S_h^2 = N\sum_{h=1}^{L}\frac{N_h S_h^2}{n}$$

$$\boxed{n = \frac{N\sum_{h=1}^{L} N_h S_h^2}{N^2 D^2 + \sum_{h=1}^{L} N_h S_h^2}}$$

Ejemplo 3.2
Una fábrica de productos alimenticios tiene 1,921 empleados en sus plantas y oficinas en todo el país y requiere estimar la antigüedad de sus empleados de las diferentes áreas. Se ha estimado la varianza a partir de archivos incompletos. Calcule el tamaño de muestra necesario para estimar la antigüedad promedio de los empleados toda la fábrica con una precisión de 0.5 años y un nivel de confianza de 95%. Utilice afijación proporcional al tamaño del estrato y distribuya la muestra resultante.

Estrato	N_h	$\hat{S}_h^2$
Producción	**1,250**	**16.0000**
Distribución	**531**	**49.0000**
Administración	**140**	**36.0000**
Total	**1,921**	

Como se desea calcular inicialmente el tamaño total de la muestra se aplicará la siguiente fórmula y para ello se complementará el cuadro anterior con cálculos adicionales.

$$n = \frac{N\sum_{h=1}^{L} N_h S_h^2}{N^2 D^2 + \sum_{h=1}^{L} N_h S_h^2}$$

Estrato	N_h	$\hat{S}_h^2$	$N_h \hat{S}_h^2$
Producción	**1,250**	**16.0000**	**20,000.000**
Distribución	**531**	**49.0000**	**26,019.000**
Administración	**140**	**36.0000**	**5,040.000**
Total	**1,921**		**51,059.000**

La varianza deseada D2 se obtiene con el cociente del cuadrado de la precisión deseada d = 0.5 entre el cuadrado del coeficiente de confianza Z = 1.96:

$$D^2 = \frac{d^2}{Z^2} = \left(\frac{0.5}{1.96}\right)^2 = \mathbf{0.067057}$$

Se sustituye en la fórmula y se obtiene el tamaño de muestra requerido.

$$n = \frac{1921(51{,}059.00)}{(1{,}921)^2(0.067057)^2 + 51{,}059.00} = 336.817$$

Se redondea al entero mayor y por tanto n = 337. La distribución proporcional al tamaño del estrato se presenta en la siguiente tabla.

Estrato	N_h	$\hat{S}_h^2$	$N_h \hat{S}_h^2$	W_h	n_h
Producción	**1,250**	**16.0000**	**20,000.000**	**0.650703**	**219**
Distribución	**531**	**49.0000**	**26,019.000**	**0.276419**	**93**
Administración	**140**	**36.0000**	**5,040.000**	**0.072879**	**25**
Total	**1,921**		**51,059.000**	**1.000000**	**337**

Afijación Óptima.

Es posible incorporar una función de costos cuando se conoce la cantidad que cuesta levantar un cuestionario en cada uno de los estratos. Una función de uso frecuente es la siguiente:

$$C = C_0 + \sum_{h=1}^{L} Ch\ nh$$

Donde Co representa la suma de costos fijos y Ch el costo de levantar un cuestionario en el estrato h.

Se considera que C_1 es el costo total de levantamiento de cuestionarios
:

$$C_1 = \sum_{h=1}^{L} Ch\ nh$$

Dado un presupuesto fijo C_1 se pretende distribuir n en los L estratos de manera que la varianza de la media poblacional $V(\bar{y}_{st})$ sea mínima. Se supone muestreo aleatorio simple en cada estrato.

Se parte de la expresión general:

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

Se define una función $\phi(n_h)$ y utilizando multiplicadores de Lagrange se minimizará $\phi(n_h)$, esto es encontrar la n_h que minimice la varianza sujeta a la restricción de costos. Estrictamente n_h es entero, pero para aplicar la técnica, se supone que n_h es cualquier real y así se tendrá $\phi(n_h)$ continua.

$$\phi\ (n_h) = \frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2 S_h^2}{n_h} - \frac{1}{N^2}\sum_{h=1}^{L} N_h S_h^2 + \lambda\left(\sum_{h=1}^{L} C_h\ n_h - C_1\right)$$

Se deriva $\phi(n_h)$ respecto a n_h (una en especial y el caso se generaliza para cualquier h) y se iguala a cero:

$$\frac{\partial \phi(n_h)}{\partial\ n_h} = -\frac{N_h^2 S_h^2}{N^2 n_h^2} + \lambda\ C_h = 0$$

$$\therefore \quad \lambda\ C_h = \frac{N_h^2 S_h^2}{N^2 n_h^2}$$

$$\mathrm{n}_h^2 = \frac{N_h^2 S_h^2}{\lambda\ C_h\ N^2}$$

$$n_h = \frac{N_h S_h}{\sqrt{\lambda\ C_h} N} \qquad \text{(a)}$$

Esta expresión está en función de h a la cual se ha dado una equivalencia en término de elementos conocidos.

Si se suma (a) para toda h hasta L:

$$\sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{n}_{\mathrm{h}} = \frac{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}}{\sqrt{\lambda}\ N}$$

Pero recuérdese que: $\sum_{h=1}^{L} n_h = n$

$$n = \frac{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}}{\sqrt{\lambda}\ N}$$

De donde:

$$\sqrt{\lambda} = \frac{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}}{nN} \qquad (\mathrm{b})$$

Se sustituye (b) en (a):

$$n_h = \frac{N_h S_h \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}} \mathrm{n}$$

Se observa que el tamaño de la muestra es proporcional directamente al producto NhSh e inversamente proporcional a $\sqrt{Ch}$

Varianza del estimador de la Media con Afijación Optima

Sabemos que: $V(y_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 S_h^2}{n_h} - \frac{1}{N^2} \sum_{h=1}^{L} N_h S_h^2$

En afijación óptima tenemos la siguiente expresión para n_h:

$$n_h = \frac{N_h S_h \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}} \mathrm{n}$$

Al sustituir en la expresión general de $V\left(\bar{y}_{st}\right)$

$$V\left(\begin{matrix} y_{st} \\ op \end{matrix}\right) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{\dfrac{N_h^2 S_h^2}{N_h S_h \Big/ \sqrt{C_h}}}{\dfrac{\sum_{h=1}^{L} N_h S_h}{\Big/ \sqrt{C_h}}} n - \frac{1}{N^2} \sum_{h=1}^{L} N_h S_h^2$$

$$\boxed{V\left(\begin{matrix} y_{st} \\ op \end{matrix}\right) = \frac{1}{N^2} \left\{ \sum_{h=1}^{L} \frac{N_h S_h \sqrt{C_h}}{n} \right\} \left\{ \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h} \right\} - \frac{1}{N^2} \sum_{h=1}^{L} N_h S_h^2}$$

Tamaño de Muestra Total para una Varianza deseada D^2 en Afijación Optima.

Se sustituye en la fórmula anterior la varianza deseada:

$$D^2 = \frac{1}{N^2} \left\{ \sum_{h=1}^{L} \frac{N_h S_h \sqrt{C_h}}{n} \right\} \left\{ \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h} \right\} - \frac{1}{N^2} \sum_{h=1}^{L} N_h S_h^2$$

$$D^2 N^2 n = \left\{ \sum_{h=1}^{L} N_h S_h \sqrt{C_h} \right\} \left\{ \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h} \right\} - n \sum_{h=1}^{L} N_h S_h^2$$

$$n \left\{ D^2 N^2 + \sum_{h=1}^{L} N_h S_h^2 \right\} = \left\{ \sum_{h=1}^{L} N_h S_h \sqrt{C_h} \right\} \left\{ \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h} \right\}$$

$$n = \frac{\left\{ \sum_{h=1}^{L} N_h S_h \sqrt{C_h} \right\} \left\{ \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h} \right\}}{D^2 N^2 + \sum_{h=1}^{L} N_h S_h^2}$$

Tamaño de Muestra para un presupuesto fijo C_1

Se parte de considerar que el costo de operación está limitado a un presupuesto C_1 y a partir de ello calcular el tamaño total de muestra.

$$C_1 = \sum_{h=1}^{L} C_h n_h$$

Se sustituye la expresión de n_h correspondiente a la afijación óptima:

$$C_1 = \sum_{h=1}^{L} C_h \frac{N_h S_h \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}} n$$

$$C_1 = \left\{ \frac{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}} \right\} n$$

$$\therefore$$

$$n = \frac{C_1 \sum_{h=1}^{L} N_h S_h \Big/ \sqrt{C_h}}{\sum_{h=1}^{L} N_h S_h \sqrt{C_h}}$$

Afijación de Neyman y otros criterios como casos particulares de la afijación óptima.
La afijación de Neyman se puede considerar un caso particular de la afijación óptima en el que el costo de cada entrevista es igual en todos los estratos, es decir C_h = K1 para toda h. Se parte de la fórmula de nh para la afijación óptima.

$$n_h = \frac{N_h S_h \Big/ \sqrt{K1}}{\sum_{h=1}^{L} N_h S_h \Big/ \sqrt{K1}} \mathrm{n}$$

Se obtiene la fórmula de Neyman $$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \mathrm{n}$$

La fórmula para la varianza del estimador de la media se obtiene mediante un procedimiento análogo de sustitución.

$$V\left(\bar{y}_{\substack{st \\ ney}} \right) = \frac{1}{N^2} \frac{\left(\sum_{h=1}^{L} NhSh \right)^2}{n} - \frac{1}{N^2} \sum_{h=1}^{L} NhSh^2$$

Finalmente la fórmula para el tamaño general de muestra con afijación de Neyman.:

$$n = \frac{\left(\sum_{h=1}^{L} NhSh \right)^2}{D^2 N^2 + \sum_{h=1}^{L} NhSh^2}$$

Ahora se parte de la fórmula de afijación de Neyman y considérese el caso de desviaciones estándares iguales en todos los estratos. Esto es $S_h = K2$

$$n_h = \frac{N_h K2}{\sum_{h=1}^{L} N_h K2} \mathrm{n}$$

$$n_h = \frac{N_h}{\sum_{h=1}^{L} N_h} \mathrm{n}$$

$$\mathrm{n_h} = \frac{N_h}{N} \mathrm{n}$$

Pero esta última expresión corresponde a la afijación proporcional. Así la afijación proporcional se considera un caso particular de la afijacion óptima en el que tanto los costos, como las varianzas en los estratos son iguales.

Finalmente se parte de la fórmula de afijación proporcional y ahora considérese que los tamaños de los estratos son iguales, Nh = K3.

$$n_h = \frac{K3}{\sum_{h=1}^{L} K3} \mathrm{n}$$

$$n_h = \frac{1}{L} \mathrm{n}$$

Resulta la afijación de muestra igual en todos los estratos. En conclusión el asignar tamaños de muestra iguales a todos los estratos resulta en varianza mínima, solamente que costos, varianzas y tamaños de los estratos sean homogéneos. En la medida que uno o más supuestos no se cumplan, la afijación igual en cada estrato resulta menos eficiente para estimar el parámetro global. Hay que insistir en que con frecuencia hay conflicto de intereses entre la estimación del parámetros global y las estimaciones para los estratos cuando éstos son dominios de estudio. En estos casos se sacrifica un poco la precisión en la estimación global para lograr precisiones homogéneas en los estratos.

3.5 Comparación del Muestreo Estratificado con Afijación Proporcional y el Muestreo Aleatorio Simple.

Al inicio de este capítulo se argumentó que el muestreo estratificado brinda estimadores más eficientes que los que se obtienen por muestreo aleatorio simple. Se verificará a continuación que la varianza del estimador de la media es menor mediante el muestreo estratificado con afijación proporcional al compararla con la varianza del estimador de la media resultante del muestreo aleatorio simple.

Si el factor de corrección por finitud es ignorado se pueden tomar las siguientes expresiones de la varianza:

$$V\left(\bar{y}_{m.a.s.}\right) = \frac{S^2}{n}$$

$$V\left(\bar{y}_{st}\right) = \frac{\sum_{h=1}^{L} NhSh^2}{nN}$$

Se parte de la expresión que corresponde a S^2 para toda la población.

$$S^2 = \frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(yh_i - \bar{Y}\right)^2}{N-1}$$

$$\therefore \quad (N-1)S^2 = \sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(yh_i - \bar{Y}\right)^2$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(yh_i - \bar{Y}h + \bar{Y}h - \bar{Y}\right)^2$$

$$= \sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(yh_i - \bar{Y}h\right)^2 + \sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(\bar{Y}h - \bar{Y}\right)^2$$

Si se considera válida la aproximación Nh ≈ Nh-1 y Nh ≈ Nh-1 para valores grandes de Nh, se tendrá:

$$NS^2 = \sum_{h=1}^{L} NhSh^2 + \sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(\bar{Y}h - \bar{Y}\right)^2$$

$$\therefore \quad \frac{S^2}{n} = \frac{\sum_{h=1}^{L} NhSh^2}{nN} + \frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(\bar{Y}h - \bar{Y}\right)^2}{nN}$$

$$V\left(\bar{y}_{m.a.s.}\right) = V\left(\bar{y}_{\substack{st \\ prop}}\right) + \frac{\sum_{h=1}^{L}\sum_{i=1}^{Nh}\left(\bar{Y}h - \bar{Y}\right)^2}{nN}$$

La varianza del muestreo aleatorio simple es entonces igual a la del muestreo con afijación proporcional más la variación entre estratos. Ello ilustra la mayor eficiencia del muestreo estratificado en la medida en que se logran estratos que maximicen la varianza entre estratos y minimicen la varianza dentro de estratos.

$$\therefore \quad V\left(\bar{y}_{m.a.s.}\right) \geq V\left(\bar{y}_{\substack{st \\ prop}}\right)$$

3.6 Comparación del Muestreo Estratificado con Afijación Proporcional y el Muestreo Estratificado con Afijación de Neyman

Nuevamente, si se ignora el factor de corrección por finitud las expresiones de la varianza de la media se toman de la manera siguiente:

$$V\left(\bar{y}_{\substack{st \\ prop}}\right) = \frac{\sum_{h=1}^{L} NhSh^2}{nN}$$

$$V\left(\bar{y}_{\substack{st \\ ney}}\right) = \frac{\left(\sum_{h=1}^{L} NhSh\right)^2}{nN^2}$$

Se parte de la expresión de la varianza de la media proporcionala la cual se le suma y resta la varianza de Neyman.

$$V\left(\bar{y}_{\substack{st \\ prop}}\right) = \frac{1}{N}\frac{\sum_{h=1}^{L} NhSh^2}{n} + \frac{1}{N^2}\frac{\left(\sum_{h=1}^{L} NhSh\right)^2}{n} - \frac{1}{N^2}\frac{\left(\sum_{h=1}^{L} NhSh\right)^2}{n}$$

$$= V\left(\bar{y}_{\substack{st \\ ney}}\right) + \left\{\frac{\sum_{h=1}^{L} NhSh^2}{nN} - \frac{1}{N^2 n}\left(\sum_{h=1}^{L} NhSh\right)^2\right\}$$

$$= V\left(\bar{y}_{\substack{st \\ ney}}\right) + \frac{1}{nN}\left\{\sum_{h=1}^{L} NhSh^2 - \frac{1}{N}\left(\sum_{h=1}^{L} NhSh\right)^2\right\}$$

$$= V\left(y_{\substack{st \\ ney}} \right) + \frac{1}{nN}\left\{ \sum_{h=1}^{L} NhSh^2 - \frac{2}{N}\left(\sum_{h=1}^{L} NhSh \right)^2 + \frac{1}{N}\left(\sum_{h=1}^{L} NhSh \right)^2 \right\}$$

$$= V\left(y_{\substack{st \\ ney}} \right) + \frac{1}{nN}\left\{ \sum_{h=1}^{L} NhSh^2 - \frac{2}{N}\left(\sum_{h=1}^{L} NhSh \right)^2 + \frac{\sum_{h=1}^{L} Nh}{N^2}\left(\sum_{h=1}^{L} NhSh \right)^2 \right\}$$

$$= V\left(y_{\substack{st \\ ney}} \right) + \frac{1}{nN}\left\{ \sum_{h=1}^{L} NhSh^2 - \frac{2\left(\sum_{h=1}^{L} NhSh \right)\left(\sum_{h=1}^{L} NhSh \right)}{N} + \frac{\sum_{h=1}^{L} Nh\left(\sum_{h=1}^{L} NhSh \right)^2}{N^2} \right\}$$

$$= V\left(\bar{y}_{\substack{st \\ ney}} \right) + \frac{\sum_{h=1}^{L} Nh}{nN}\left\{ Sh^2 - \frac{2Sh\sum_{h=1}^{L} NhSh}{N} + \frac{\left(\sum_{h=1}^{L} NhSh \right)^2}{N^2} \right\}$$

$$= V\left(\bar{y}_{\substack{st \\ ney}} \right) + \frac{\sum_{h=1}^{L} Nh}{nN}\left\{ Sh - \frac{\sum_{h=1}^{L} NhSh}{N} \right\}^2$$

Por lo tanto

$$V\left(y_{\substack{st \\ prop}} \right) \geq V\left(y_{\substack{st \\ ney}} \right)$$

Se concluye que la varianza del estimador de la media con afijación proporcional es igual a la varianza del estimador con afijación de Neyman más una cantidad que solamente se anula cuando las desviaciones estándares de los estratos son todas iguales.

Se cumple entonces la triple desigualdad:

$$V(y_{mas}) \geq V(y_{prop}) \geq V(y_{Ney})$$

Ejemplo 3.3
Utilice los datos por estratos del Ejemplo 3.2 y calcule el tamaño de muestra necesario para alcanzar una precisión d = 0.5 en la estimación de la media global, con una confianza del 95% . Considere en este caso afijación de Neyman.

Se aplicarán las fórmulas siguientes

$$n = \frac{\left(\sum_{h=1}^{L} NhSh\right)^2}{D^2 N^2 + \sum_{h=1}^{L} NhSh^2} \qquad n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \text{n}$$

Estrato	N_h	$\hat{S}_h^2$	$N_h \hat{S}_h^2$	W_h	$N_h \hat{S}_h$	n_h
Producción	**1,250**	**16.000**	**20,000.000**	**0.6507**	**5,000.000**	**164**
Distribución	**531**	**49.000**	**26,019.000**	**0.2764**	**3,717.000**	**122**
Administración	**140**	**36.000**	**5,040.000**	**0.0729**	**840.000**	**28**
Total	**1,921**		**51,059.000**	**1.0000**	**9,557.000**	**314**

Z = 1.96
d = 0.5
D^2 = 0.067057 **n = 313.64**

Note que el tamaño de muestra obtenido mediante Neyman es más pequeño en 23 unidades comparado con el calculado para afijación proporcional al tamaño con las mismas condiciones de precisión y nivel de confianza.

3.7 Muestreo Estratificado para Proporciones

En el muestreo estatificado para proporciones procedemos de manera similar que con el muestreo aleatorio simple. El parámetro de la proporción P, para la población y su estimador a partir de la muestra. En ambos casos el parámetro se considera la media de una variable que adopta solamente dos valores 1 y 0 en cada estrato.

$$\text{Ph} = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi} \qquad \hat{\text{P}}\text{h} = \frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}$$

El estimador de P global se obtiene como la suma ponderada de las estimaciones de las proporciones en cada estrato.

$$\hat{\text{P}} = \sum_{i=1}^{n_h} W_h \hat{P}_h$$

Para cálculo de varianzas se considera en forma simplificada la siguiente igualdad:

$$S_{\text{h}}^2 = P_h Q_h$$

La forma general para la varianza de la proporción de la población adopta la forma:

$$V\left(\hat{P}_{st}\right)=\frac{1}{N^2}\sum_{h=1}^{L}\frac{N_h^2P_hQ_h}{n_h}-\frac{1}{N^2}\sum_{h=1}^{L}N_hP_hQ_h$$

A continuación se exponen las fórmulas de varianza y cálculo de tamaño de muestra total en función de una varianza deseada.

Asignación igual en cada estrato.

$$V\left(\hat{P}_{=}\right)=\frac{L}{N^2}\sum_{h=1}^{L}\frac{N_h^2P_hQ_h}{n}-\frac{1}{N^2}\sum_{h=1}^{L}N_hP_hQ_h$$

$$\mathrm{n}=\frac{L\sum_{h=1}^{L}N_h^2P_hQ_h}{N^2D^2+\sum_{h=1}^{L}N_hP_hQ_h} \qquad \mathrm{n_h}=\frac{n}{L}$$

Asignación proporcional al tamaño del estrato.

$$V\left(\hat{P}_{PT}\right)=\frac{1}{N}\sum_{h=1}^{L}\frac{N_hP_hQ_h}{n}-\frac{1}{N^2}\sum_{h=1}^{L}N_hP_hQ_h$$

$$\mathrm{n}=\frac{N\sum_{h=1}^{L}N_hP_hQ_h}{N^2D^2+\sum_{h=1}^{L}N_hP_hQ_h} \qquad \mathrm{n_h}=\frac{N_h}{N}n$$

Asignación de Neyman.

$$V\left(\hat{P}_{Ney}\right)=\frac{1}{N^2}\frac{\left(\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\right)^2}{n}-\frac{1}{N^2}\sum_{h=1}^{L}N_hP_hQ_h$$

$$n=\frac{\left(\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\right)^2}{D^2N^2+\sum_{h=1}^{L}N_hP_hQ_h} \qquad n_h=\frac{N_h\sqrt{P_hQ_h}}{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}}\mathrm{n}$$

Asignación Óptima considerando costos diferenciales por estrato.

$$V\left(\hat{P}_{Op}\right)=\frac{1}{N^2}\left\{\sum_{h=1}^{L}\frac{N_h\sqrt{P_hQ_h}\sqrt{C_h}}{n}\right\}\left\{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\Big/\sqrt{C_h}\right\}-\frac{1}{N^2}\sum_{h=1}^{L}N_hP_hQ_h$$

$$n=\frac{\left\{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\sqrt{C_h}\right\}\left\{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\Big/\sqrt{C_h}\right\}}{D^2N^2+\sum_{h=1}^{L}N_hP_hQ_h}$$

$$n_h=\frac{N_h\sqrt{P_hQ_h}\Big/\sqrt{C_h}}{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}\Big/\sqrt{C_h}}\mathrm{n}$$

$$n_h=\frac{N_h\sqrt{P_hQ_h}}{\sum_{h=1}^{L}N_h\sqrt{P_hQ_h}}\mathrm{n}$$