Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Muestreo Estratificado

Andrés Peña M.

a.pena@rusersgroup.com

Diciembre 2021

X Seminario Internacional de Estadística Aplicada

Tabla de contenidos

Muestreo Estratificado

Muestreo Estratificado

- **Estrato** es un subconjunto de unidades muestrales de la población.

Muestreo Estratificado

- **Estrato** es un subconjunto de unidades muestrales de la población.
- Cada estrato se muestrea por separado y se obtienen los estimadores de parametros (media,total, proporción) para cada estrato, luego se combinan para tener los estimadores de toda la población.

Muestreo Estratificado

- **Estrato** es un subconjunto de unidades muestrales de la población.
- Cada estrato se muestrea por separado y se obtienen los estimadores de parametros (media,total, proporción) para cada estrato, luego se combinan para tener los estimadores de toda la población.
- Los estratos forman una particion de la población y se selecciona **muestra en cada estrato en forma independiente**.

Muestreo Estratificado

Razones para utilizar este tipo de diseño de muestra:

1. Estadística.- Para *reducir la varianza* de los estimadores, es decir, tener mas precisión.

Cuando la población esta constituida por unidades heterogéneas y tenemos una idea previa de los grupos de unidades mas homogéneas entre sí, entonces es conveniente agruparlas en estratos.

Muestreo Estratificado

Considere una población de 20 elementos en los cuales Y toma los valores:

$$\{6, 3, 4, 4, 5, 3, 6, 2, 3, 2, 2, 6, 5, 3, 5, 2, 4, 6, 4, 5\}$$

$$\bar{Y} = 4 \quad S^2 = \frac{\sum_{i=1}^{20} \left(Y_i - \bar{Y}\right)^2}{19} = \frac{40}{19} = 2.11$$

Si tomamos una muestra aleatoria simple de tamaño 5 y usamos $\bar{y}$ como estimador de $\bar{Y}$, tenemos:

$$V(\bar{y}) = \left(1 - \frac{n}{N}\right) \frac{S^2}{n} = \left(1 - \frac{5}{20}\right) \frac{2.11}{5} = 0.316$$

Muestreo Estratificado

Dada la estructura de la población, se puede ordenar como:

$$\underbrace{2,2,2,2}\ \underbrace{3,3,3,3}\ \underbrace{4,4,4,4}\ \underbrace{5,5,5,5}\ \underbrace{6,6,6,6}$$
$$\text{ESTRATOS}$$

Suponga que tenemos un mecanismo por el cual podemos seleccionar un elemento al azar de cada grupo para formar nuestra muestra de tamaño 5. Obtenemos, en cada una de las posibles muestras, los valores:

$$\{2,3,4,5,6\} \text{ cuya } \bar{y} = 4 = \bar{Y}$$

Este estimador tendría varianza **cero** ya que la varianza dentro de cada estrato es cero y no hay fluctuaciones muestrales y, además, el estimador siempre sería igual al parámetro.

Muestreo Estratificado

Ejemplo de uso de muestreo estratificado:

- Suponga un estudio donde interesa conocer alguna característica de los hogares en alguna ciudad de latinoamérica.

Muestreo Estratificado

Ejemplo de uso de muestreo estratificado:

- Suponga un estudio donde interesa conocer alguna característica de los hogares en alguna ciudad de latinoamérica.
- Se sabe que esa característica depende fuertemente del nivel socioeconómico de las familias.

Muestreo Estratificado

Ejemplo de uso de muestreo estratificado:

- Suponga un estudio donde interesa conocer alguna característica de los hogares en alguna ciudad de latinoamérica.
- Se sabe que esa característica depende fuertemente del nivel socioeconómico de las familias.
- Se construyen estratos considerando áreas de la ciudad con niveles socioeconomicos semejantes. Así las colonias/barrios se pueden clasificar en relacion al nivel socioeconómico como: muy alto, alto, medio, medio-bajo y bajo, formando 5 estratos.

Muestreo Estratificado

Ejemplo de uso de muestreo estratificado:

- Suponga un estudio donde interesa conocer alguna característica de los hogares en alguna ciudad de latinoamérica.
- Se sabe que esa característica depende fuertemente del nivel socioeconómico de las familias.
- Se construyen estratos considerando áreas de la ciudad con niveles socioeconomicos semejantes. Así las colonias/barrios se pueden clasificar en relacion al nivel socioeconómico como: muy alto, alto, medio, medio-bajo y bajo, formando 5 estratos.
- La encuesta se planea para cada estrato por separado.

Muestreo Estratificado

2. Disponibilidad de marcos.- Si la población está identificada a través de dos o más marcos, cada marco define un estrato.

Si para una parte de la población se tiene un buen marco, éste se usa para el muestreo de ese estrato; y las otras partes de la población se muestrean usando otros marcos, tal vez más imprecisos, y posiblemente con otros diseños de muestra.

Muestreo Estratificado

3. Costo.- Cuando hay diferentes costos de localizar y levantar la información de las unidades muestrales.

Por ejemplo, en una encuesta en predios agrícolas hay una región cuyo acceso es difícil (solo por avioneta o a caballo).

Esta región puede constituir un estrato, que será muestreado con un tamaño de muestra más pequeño.

Muestreo Estratificado

Algunas consideraciones importantes:

- El efecto de la formación de estratos es **reducir la variabilidad** de los estimadores, en la medida en que las unidades dentro de cada estrato sean homogeneas.

Muestreo Estratificado

Algunas consideraciones importantes:

- El efecto de la formación de estratos es **reducir la variabilidad** de los estimadores, en la medida en que las unidades dentro de cada estrato sean homogeneas.
- Se pueden usar **diferentes diseños** de muestra en cada estrato.

Muestreo Estratificado

Algunas consideraciones importantes:

- El efecto de la formación de estratos es **reducir la variabilidad** de los estimadores, en la medida en que las unidades dentro de cada estrato sean homogeneas.
- Se pueden usar **diferentes diseños** de muestra en cada estrato.
- **No interesa** tener estimaciones por estrato.

Notación

A nivel **poblacional**:

L No. de estratos

N_h No. de unidades muestrales estrato h, $h = 1, \ldots, L$

$N = \sum_{h=1}^{L} N_h$ No. de unidades en la población

Y_{hi} valor de la medición en U_{hi},
$i = 1, \ldots, N_h, h = 1, \ldots, L$

$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} Y_{hi}}{N_h}$ media poblacional estrato h

Notación

$$Y_h = \sum_{i=1}^{N_h} Y_{hi} = N_h \bar{Y}_h \quad \text{total poblacional estrato } h$$

$$Y = \sum_{h=1}^{L} Y_h = \sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi} \quad \text{total poblacional}$$

$$\bar{Y} = \frac{Y}{N} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} Y_{hi}}{\sum_{h=1}^{N} N_h} \quad \text{media poblacional}$$

$$S_h^2 = \frac{\sum_{i=1}^{N_h} \left(Y_{hi} - \bar{Y}_h\right)^2}{N_h - 1} \quad \text{varianza poblacional estrato } h$$

$$W_h = \frac{N_h}{N} \quad \text{peso del estrato}$$

$$\sum_{h=1}^{L} W_h = 1$$

Notación

Consideremos que tenemos una M.A.S. en cada estrato.

A nivel **muestral**:

n_h tamaño de muestra estrato h

$n = \sum_{h=1}^{L} n_h$ tamaño de muestra

$\hat{\bar{Y}}_h = \bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$ estimador media estrato h

$\hat{Y}_h = N_h \bar{y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}$ estimador total estrato h

Estimadores

Estimador del total

El estimador del total poblacional es:

$$\begin{aligned}\hat{Y} &= \sum_{h=1}^{L} \hat{Y}_h = \sum_{h=1}^{L} N_h \bar{y}_h \\ &= \sum_{h=1}^{L} N_h \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h} \\ &= \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{N_h}{n_h} y_{hi}\end{aligned}$$

Donde $\frac{N_h}{n_h}$ es el factor de expansión.

Estimador del total

La varianza del estimador del total es:

$$V(\hat{Y}) = \sum_{h=1}^{L} V\left(\hat{Y}_h\right) \text{ muestras independientes en c/estrato}$$

$$= \sum_{h=1}^{L} V\left(N_h \bar{y}_h\right)$$

$$= \sum_{h=1}^{L} N_h^2 V\left(\bar{y}_h\right)$$

Como tenemos una M.A.S. en cada estrato,

$$V(\hat{Y}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

Estimador del total

El estimador de la varianza del estimador del total es:

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}$$

donde,

$$\hat{S}_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador del total, el intervalo aproximado del $(1 - \alpha) \times 100\%$ de confianza para el total poblacional es:

$$\hat{Y} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{Y})}$$

Estimador de la media

El estimador de la **media** poblacional es:

$$\hat{\bar{Y}} = \frac{\hat{Y}}{N} = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{N}$$
$$= \sum_{h=1}^{L} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{L} W_h \bar{y}_h$$

$\hat{\bar{Y}}$ es una suma ponderada de los promedios muestrales en cada estrato.

Estimador de la media

La varianza del estimador de la media es:

$$V(\hat{\bar{Y}}) = V\left(\sum_{h=1}^{L} W_h \bar{y}_h\right)$$

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

El estimador de la varianza del estimador de la media es:

$$\hat{V}(\hat{\bar{Y}}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{S}_h^2}{n_h}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador de la media, el intervalo aproximado del $(1-\alpha) \times 100\%$ de confianza para la media poblacional es:

$$\hat{\bar{Y}} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{\bar{Y}})}$$

Estimador de una proporción

Sea $Y_{hi} = \begin{cases} 1 & U_{hi} \text{ tiene la característica} \\ 0 & U_{hi} \text{ no tiene la característica} \end{cases}$

El estimador de la proporción P de unidades que tienen cierta característica es:

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{p}_h \quad \text{con} \quad \hat{p}_h = \sum_{i=1}^{n_h} \frac{y_{hi}}{n_h}$$

Estimador de una proporción

La varianza de este estimador:

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

con estimador:

$$\hat{V}(\hat{P}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}$$

Si el tamaño de muestra en cada estrato es grande y podemos hacer la aproximación a la normal del estimador de la proporción, el intervalo aproximado del $(1 - \alpha) \times 100\%$ de confianza para la proporción poblacional es:

$$\hat{P} \pm z_{1-\alpha/2} \sqrt{\hat{V}(\hat{P})}$$

Estimadores de Razón en muestreo estratificado

Estimador de razón combinado R_c

Combina la información de los estratos y después hace el cociente.

$$\hat{R}_c = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{L} \hat{Y}_h}{\sum_{h=1}^{L} \hat{X}_h}$$

En caso de tener M.A.S. en cada estrato

$$\hat{R}_c = \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h}$$

Estimador de razon combinado

La varianza, y su estimador, en caso de tener una M.A.S. en cada estrato:

$$V\left(\hat{R}_c\right) = \frac{1}{X^2}\sum_{h=1}^{L} N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\times \frac{1}{N_h-1}\sum_{i=1}^{N_h}\left[\left(Y_{hi}-\bar{Y}_h\right)-R_c\left(X_{hi}-\bar{X}_h\right)\right]^2$$

$$\hat{V}\left(\hat{R}_c\right) = \frac{1}{\hat{X}^2}\sum_{h=1}^{L} N_h^2\left(\frac{1}{n_h}-\frac{1}{N_h}\right)\times \frac{1}{n_h-1}\sum_{i=1}^{n_h}\left[y_{hi}-\hat{R}_c x_{hi}-\frac{\sum_{j=1}^{n_h}\left(y_{hj}-\hat{R}_c x_{hj}\right)}{n_h}\right]^2$$

Estimador de la media poblacional con razón combinado

$$\hat{\bar{Y}}_c = \hat{R}_c \bar{X} \quad \{\bar{X} \text{ conocida } \}$$

Con varianza y estimador de varianza:

$$V\left(\hat{Y}_c\right) = \bar{X}^2 V\left(\hat{R}_c\right)$$
$$\hat{V}\left(\hat{Y}_c\right) = \bar{X}^2 \hat{V}\left(\hat{R}_c\right)$$

Estimador del total poblacional con razón combinado

$$\hat{Y}_c = \hat{R}_c X \quad \{X \text{ conocido } \}$$

Con varianza y estimador de varianza:

$$V\left(\hat{Y}_c\right) = X^2 V\left(\hat{R}_c\right)$$
$$\hat{V}\left(\hat{Y}_c\right) = X^2 \hat{V}\left(\hat{R}_c\right)$$

El estimador de razón combinado se usa cuando se tienen muchos estratos y/o los tamaños de muestra en cada estrato son pequeños. Supone que las razones en cada estrato son similares.

Distribución de la muestra a los estratos

Distribución de la muestra a los estratos

Suponga que se tiene un tamaño de muestra n determinado.

Cómo se reparte n entre los L estratos?

1. Distribución óptima

Sea C_h el costo de obtener información de una unidad en el estrato h. Se tiene una función de costo de la forma:

$$\text{costo } = C = C_0 + \sum_h C_h n_h$$

La varianza del estimador $\hat{Y}$ se minimiza, con un costo total fijo, cuando:

$$n_h = n \frac{N_h S_h}{\sqrt{C_h}} \left[\sum_{k=1}^{L} \frac{N_k S_k}{\sqrt{C_k}} \right]^{-1}$$

Distribución de la muestra a los estratos

Observe que,

$$n_h \propto \frac{N_h S_h}{\sqrt{C_h}}$$

Esto quiere decir que en un estrato dado, se toma más muestra si:

- El estrato es más grande
- El estrato es más variable
- El costo es menor

Distribución de la muestra a los estratos

2. Distribución de Neyman Si se considera que los costos C_h son constantes en todos los estratos:

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

3. Distribución proporcional Si se considera que tanto los costos como las varianzas S_h son constantes en todos los estratos, entonces:

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

Esta distribución produce muestras autoponderadas:

$$\frac{n_h}{N_h} = \frac{n}{N} \Rightarrow \frac{N_h}{n_h} = \frac{N}{n} \text{ factor de expansión}$$

Determinación del tamaño de muestra

n asignación de Neyman

Considerando la asignación de Neyman (costos C_h constantes):

$$n_h = n\frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i}$$

Para estimar la media:

$$n = \frac{\left[\sum_{h=1}^{L} N_h S_h\right]^2}{N^2 \frac{\delta^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{L} N_h S_h^2}$$

Para estimar el total:

$$n = \frac{\left[\sum_{h=1}^{L} N_h S_h\right]^2}{\frac{\delta^2}{z^2_{1-\alpha/2}} + \sum_{h=1}^{L} N_h S_h^2}$$

n distribución proporcional

Si consideramos la distribución proporcional:

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

Para estimar la media:

$$V(\hat{\bar{Y}}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

$$\delta = z_{1-\alpha/2}\sqrt{V(\hat{Y})} \Rightarrow V(\hat{Y}) = \frac{\delta^2}{z_{1-\alpha/2}^2}$$

$$n = \frac{N\sum_{h=1}^{L} N_h S_h^2}{N^2\frac{\delta^2}{z_{1-\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}$$

n distribución proporcional

Para estimar el total:

$$n = \frac{N \sum_{h=1}^{L} N_h S_h^2}{\frac{\delta^2}{z_{1-\alpha/2}^2} + \sum_{h=1}^{L} N_h S_h^2}$$

Nota: Para estimar proporciones utilice las expresiones de tamaño de muestra para estimar la media con $S_h^2 = P_h\,(1 - P_h)$.

Comparación de varianzas

Se puede demostrar (Cochran) que:

$$V_{opt}(\hat{Y}) \leq V_{prop}(\hat{Y}) \leq V_{\text{m.a.s.}}\ (\hat{\bar{Y}})$$

Gracias!!!