Teoría del Muestreo

Eloy Alvarado Narváez

Instituto de Estadística
Universidad de Valparaíso

Muestreo por Conglomerado

- Recuerdo de Muestreo I
- Unidad primaria de muestreo
- Estratos vs Conglomerados

Introducción

¿Cuándo se usa un muestreo por conglomerado?

- Construir un marco de muestreo en donde las unidades observables son difícil, caro o imposible de muestrear. Así, un muestreo por conglomerado nos permitiría facilitar el proceso.
- La población está ya organizada en conglomerados, por lo que obtener información de ellos será más sencillo que otras técnicas de muestreo.

Si bien, es posible encontrar semejanzas entre muestro por conglomerado y muestreo estratificado, pues en ambas técnicas se selecciona un *grupo* de elementos de la población, pero su proceso de selección son **distintos**.

Ejemplo

Asumamos que tenemos un población de H estratos, y cada estrato tiene n_h elementos.

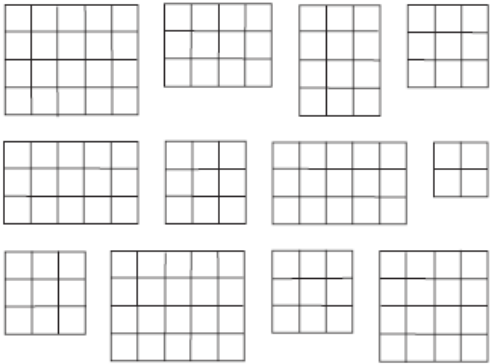


Figure 1: Población

Si realizamos un muestreo estratificado:

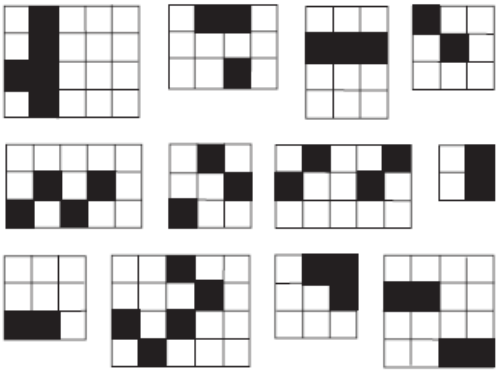


Figure 2: Estratos

En este caso, la varianza del estimador $\overline{y}_U$ depende de la variabilidad de los valores **DENTRO** de los estratos.

Si realizamos un muestreo por conglomerado:

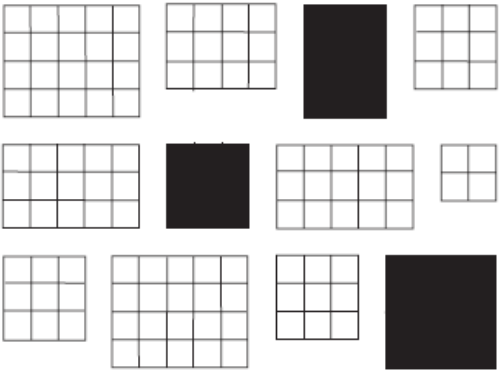


Figure 3: Conglomerados

En este caso, los estratos son la unidad de muestreo. Mientras más conglomerados muestreamos, más pequeña nuestra varianza. La varianza del estimador $\overline{y}_U$ depende principalmente de la variablidad **ENTRE** las medias de los conglomerados.

Notación utilizada

Bajo un **muestreo aleatorio simple**, las unidades muestrales son también los elementos observados. Bajo un **muestreo por conglomerado**, las unidades muestrales son los **conglomerados** (unidad primaria de muestreo) y los elementos observados son las **unidades secundarias de muestreo** en cada conglomerado.

El universo $\mathcal{U}$ es la población de N *unidades primarias de muestreo*. (UPM) $\mathcal{S}$ refiere a las muestras de unidades primarias de muestreo escogidas desde la población de N *unidades primarias de muestreo*.

$\mathcal{S}_i$ refiere a la muestra de *unidades secundarias de muestreo* escogidas desde la $i-$ésima unidad primaria de muestreo.

Las medidas son:
y_{ij} : medición del elemento $j-$ésimo en la $i-$ésima unidada primaria de muestreo.

Es claro ver que la notación es algo engorrosa debido a que se debe especificar las unidades primarias y secundarias, esto es, 2 subíndices.

Nivel: UPM - Cantidades poblacionales

$N =$ Número de **UPM** en la población.

$M_i =$ Número de **USM** en la **UPM** i-ésima.

$$M_0 = \sum_{i=1}^{N} M_i = \text{Número total de } \textbf{USM} \text{ en la población.}$$

$$t_i = \sum_{j=1}^{M_i} y_{ij} = \text{Total en la } \textbf{UPM} \ i\text{-ésima.}$$

$$t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} = \text{total poblacional.}$$

$$\mathcal{S}_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2 = \text{varianza poblacional del total de } \textbf{UPM}.$$

Nivel: USM - Cantidades poblacionales

$$\overline{y}_u = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0} = \text{media poblacional.}$$

$$\overline{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i} = \text{media poblacional en la } \textbf{UPM} \ i\text{-ésima.}$$

$$\mathcal{S}^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \overline{y}_U)^2}{M_0 - 1} = \text{varianza poblacional.}$$

$$\mathcal{S}_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \overline{y}_{iU})^2}{M_i - 1} = \text{varianza poblacional dentro de cada } \textbf{UPM} \ i\text{-ésima.}$$

Cantidades muestrales

n= número de **UPM** en la muestra.

m_i= número de **USM** en la muestra desde la **UPM** i-ésima.

$\overline{y}_i = \sum_{j \in \mathcal{S}_i} \frac{y_{ij}}{m_i}$ = media muestral para la **UPM** i-ésima.

$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij}$ = estimación del total para la **UPM** i-ésima.

$\hat{t}_{unb} = \sum_{i \in S} \frac{N}{n} \hat{t}_i$ = estimador insesgado del total de la población.

$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2$ = varianza muestral total.

$s_i^2 = \sum_{j \in \mathcal{S}_i} \frac{(y_{ij} - \overline{y}_i)^2}{m_i - 1}$ = varianza muestral dentro de cada **UPM** i-ésima.

w_{ij} = Pesos muestrales (ponderaciones muestrales) para la **USM** en la **UPM** i-ésima.

Muestreo por conglomerado unietápico

En un muestreo por conglomerado unietápico se tienen dos posibilidades:

- **Todos** los elementos que componen un conglomerado son muestreados.
- **Ninguno** de los elementos que componen un conglomerado son muestreados.

¿Cuándo utilizamos un muestreo por conglomerado unietápico?

Este tipo de muestreo es usualmente utilizado cuando el costo de muestrear las **USM** es despreciable comparado con el costo de muestrear las **UPM**.

Ejemplo:

Para encuentas educacionales, la **UPM** natural es son los cursos; todos los estudiantes en una clase son usualmente incluidos como **USM** ya que el costo de agregarlos -entregándoles el cuestionario- es despreciable.

En una población de N **UPM**, la i-ésima **UPM** contiene M_i **USM** (elementos). En el diseño más simple, se realiza un **M.A.S** de n **UPM** desde la población y se mide la variable de interés en **CADA** elementos de las **UPM** muestreadas. Así, para un muestreo por conglomerado unietápico $M_i = m_i$.

Conglomerados de igual tamaño: Estimación

Consideremos el caso más sencillo: cada **UPM** tiene la misma cantidad de elementos, con $M_i = m_i = M$. Naturalmente, este caso no es usual en conglomerados de personas pero puede ocurrir en muestreos en el área de agricultura e industria.
La estimación de la media poblacional y totales es sencilla: tratamos cada media o total de las **UPM** como las observaciones e ignoramos los elementos individuales.

Así, tenemos un **M.A.S** de n datos, $\{t_i, i \in \mathcal{S}\}$: t_i es el total de todos los elementos en la **UPM** $i-$ésima. Entonces,

$$\overline{t}_{\mathcal{S}} = \sum_{i \in \mathcal{S}} \frac{t_i}{n}$$

estima el promedio total de los conglomerados. Por ejemplo:
En una encuesta de hogares que desea estimar el salario de casas para dos personas, t_i es el total del salario para el hogar $i-$ésimo. $\overline{t}_U$ es el salario promedio por hogar, y $\overline{y}_U$ es el salario promedio por persona. Para estimar el salario total t, podemos utilizar el estimador:

$$\hat{t} = \frac{N}{n} \sum_{i \in S} t_i$$

y los resultados vistos bajo un **M.A.S.** son válidos para $\hat{t}$ pues tenemos un muestreo aleatorio simple de n unidades desde una población de tamaño N. Como resultado, $\hat{t}$ es un estimador insesgado de t, con varianza dada por:

$$\mathbb{V}[\hat{t}] = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n}$$

y error estándar,

$$SE[\hat{t}] = N\sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}$$

donde S_t^2 y s_t^2 son la varianza poblaciones y muestral, respectivamente, del total de las **UPM**.

Estas varianzas definidas por:

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2$$

y,

$$s_t^2 = \frac{1}{n-1} \sum_{i \in \mathcal{S}} \left(t_i - \frac{\hat{t}}{N} \right)^2$$

Para estimar $\overline{y}_U$, se divide la estimación total por el número de personas, obteniendo:

$$\hat{\overline{y}} = \frac{\hat{t}}{NM}$$

con,

$$\mathbb{V}[\hat{\overline{y}}] = \left(1 - \frac{n}{N}\right) \frac{S_t^2}{nM^2}$$

y,

$$SE[\hat{\overline{y}}] = \frac{1}{M}\sqrt{\left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}}$$

Si se dan cuenta, no hay nada nuevo en aplicar un muestreo por conglomerado unietápico: simplemente usamos los resultados de un **M.A.S** con los totales por **UPM** como las observaciones.

Ejemplo

Un estudiante desea estimar las notas promedios (GPA) de los alumnos que se hospedan en los dormitorios universitarios. En vez de obtener una lista de todos los estudiantes de su dormitorio y realizar un **M.A.S.**, él nota que los dormitorios consisten en 100 habitaciones, cada una con 4 estudiantes; escoge 5 de esas habitaciones al azar y luego consulta a cada persona en aquellas 5 habitaciones cual es su GPA. Los resultados son los siguientes:

Person Number	Suite (psu) 1	2	3	4	5
1	3.08	2.36	2.00	3.00	2.68
2	2.60	3.04	2.56	2.88	1.92
3	3.44	3.28	2.52	3.44	3.28
4	3.04	2.68	1.88	3.64	3.20
Total	12.16	11.36	8.96	12.96	11.08

- Calcular $\hat{t}, s_{\hat{t}}^2, \hat{\bar{y}}$ y $SE[\hat{\bar{y}}]$

Un muestreo unietápico con un **M.A.S.** de **UPM** produce una muestra auto-ponderada. El peso o ponderación para cada observación es:

$$w_{ij} = \frac{1}{\mathbb{P}\{j\text{-ésima } \textbf{USM} \text{ de la } i\text{-ésima } \textbf{UPM} \text{ esté en la muestra}\}} = \frac{N}{n}$$

Para el ejemplo anterior, se tiene:

$$\begin{aligned}\hat{t} &= \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij} \\ &= \frac{N}{n}(3.08 + 2.6 + \cdots + 3.28 + 3.2) \\ &= \frac{100}{5}(56.52) = 1130.4\end{aligned}$$

Así, al igual que en un muestreo estratificado, podemos estimar el total poblacional sumando el producto de los valores observados y las ponderaciones muestrales. La media poblacional es estimada por:

$$\begin{aligned}\hat{\bar{y}} &= \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}} \\ &= \frac{1130.4}{NM} \\ &= 2.826\end{aligned}$$

Conglomerados de igual tamaño: Teoría

En un muestreo por conglomerado unietápico, cuando cada **UPM** tiene M **USM**, la variabilidad del estimador insesgado de t depende en su totalidad de la variabilidad **ENTRE** las **UPM**, debido a que:

$$S_t^2 = \sum_{i=1}^{N} \frac{(t_i - \overline{t}_U)^2}{N-1} = \sum_{i=1}^{N} \frac{M^2(\overline{y}_{iU} - \overline{y}_U)^2}{N-1} = M(MSB)$$

En donde MSB, hace referencia a *between mean square* o media cuadrática **ENTRE** conglomerados. Así, para un muestreo por conglomerado, se tiene:

$$\mathbb{V}[\hat{t}_{cluster}] = N^2 \left(1 - \frac{n}{N}\right) \frac{M(MSB)}{n}$$

Population ANOVA Table—Cluster Sampling

Source	df	Sum of Squares	Mean Square
Between psus	$N-1$	$\text{SSB} = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{iU}-\bar{y}_U)^2$	MSB
Within psus	$N(M-1)$	$\text{SSW} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{iU})^2$	MSW
Total, about $\bar{y}_U$	$NM-1$	$\text{SSTO} = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_U)^2$	S^2

Podemos notar que:

- Si MSB/MSW es grande bajo un muestreo por conglomerado, entonces este muestreo disminuirá su precisión. En esta situación, MSB es relativamente grande debido a que mide la variable *ENTRE* conglomerados: elementos en diferentes conglomerados usualmente varían más que los elementos en el mismo conglomerado, debido a que los conglomerados tienen diferentes medias.
- Si MSW es bajo, entonces la varianza elemento a elemento *DENTRO* de los conglomerados es baja y en consecuencia, los elementos son relativamente homogéneos.

Comparemos el muestreo por conglomerado con un muestreo aleatorio simple. Si, en vez de tomar una muestra de conglomerados de M elementos cada uno de los n elementos, tomamos un **M.A.S.** con nM observaciones, entonces la varianza de la estimación total sería:

$$\mathbb{V}[\hat{t}_{M.A.S}] = (NM)^2 \left(1 - \frac{nM}{NM}\right) \frac{S^2}{nM} = N^2 \left(1 - \frac{n}{M}\right) \frac{MS^2}{n}$$

Si comparamos este valor con el equivalente bajo un muestreo por conglomerado, vemos que si $MSB > S^2$, entonces el muestreo por conglomerado es menos eficiente que un **M.A.S.**.

Coeficiente Correlación intraclase

El coeficiente de correlación intraclase (ICC en inglés), nos cuantifica cuan similares son los elementos en el mismo conglomerado. Nos da una **medida de homogeneidad** dentro de los conglomerados. Este coeficiente es definido como el coeficiente de correlación de Pearson para los $NM(M-1)$ pares (y_{ij}, y_{ik}) para i entre 1 y N y $j \neq k$, puede ser escrito en términos de los valores poblaciones de la tabla ANOVA:

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SSTO}$$

En donde, SSW y $STTO$ hace referencia a la suma cuadrada dentro y total de los conglomerados, respectivamente, por sus siglas en inglés.

Debido a que $0 \leq SSW/SSTO \leq 1$, sigue de lo anterior que:

$$-\frac{1}{M-1} \leq ICC \leq 1$$

Si los conglomerados son perfectamente homogéneos ($SSW = 0$), entonces $ICC = 1$. Asimismo, la ecuación anterior implica que:

$$MSB = \frac{NM-1}{M(N-1)} S^2[1 + (M-1)ICC]$$

¿Cuánta precisión perdemos al tomar un muestreo por conglomerado?

$$\frac{\mathbb{V}[\hat{t}_{cluster}]}{\mathbb{V}[\hat{t}_{M.A.S.}]} = \frac{MSB}{S^2} = \frac{NM-1}{M(N-1)}[1+(M+1)ICC]$$

Si N (el número de **UPM**) es grande, entonces $NM-1 \approx M(N-1)$, así la razón anterior es aproximadamente $1+(M-1)ICC$.
Si fijamos $ICC = 1/2, M = 5$ entonces $1+(M-1)ICC = 3$, por lo que necesitaríamos medir 300 elementos usando un muestreo por conglomerado para obtener la misma precisión que con un **M.A.S.** de 100 elementos.
Debido a que usualmente es más barato recolectar datos bajo un **M.C.**, esperamos tener mayor precisión por unidad monetaria utilizada en este muestreo.

Con respecto al coeficiente de correlación intraclase:

- Si los conglomerados ocurren naturalmente en la población, el coeficiente **ICC** es usualmente positivo, esto es: elementos dentro de un mismo conglomerado tienden a ser más similares entre sí que elementos seleccionados aleatoriamente desde la población.
- El coeficiente **ICC** es negativo si los elementos dentro de un conglomerado están *más* dispersos que los de un grupo escogido aleatoriamente. Esto fuerza a las medias de los conglomerados a ser casi iguales, pues $SSTO = SSW + SSB$, si la suma cuadrada total ($SSTO$) es tomada fija y SSW es grande, entonces SSB debe ser pequeña.
- Si $ICC < 0$, un muestreo por conglomerado es más eficiente que un **M.A.S** de elementos. Tener un coeficiente negativo es raro para conglomerados formados naturalmente.

El coeficiente **ICC** está sólo definido para conglomerados de igual tamaño. Una medida alternativa de homogeneidad en poblaciones generales es el R^2 ajustado, definido como:

$$R_a^2 = 1 - \frac{MSW}{S^2}$$

Si todos las **UPM** son del mismo tamaño, entonces el incremento en la varianza debido a un muestreo por conglomerado es:

$$\frac{\mathbb{V}[\hat{t}_{cluster}]}{\mathbb{V}[\hat{t}_{M.A.S.}]} = \frac{MSB}{S^2} = \frac{MSB}{S^2} = 1 + \frac{N(M-1)}{N-1} R_a^2$$

Es claro ver que R_a^2 es cercano al **ICC** para muchas poblaciones. R_a^2 es la cantidad relativa de variabilidad en la población explicada por las medias de las **UPM**, ajustada por los grados de libertad.

Ejemplo

Consideremos dos poblaciones artificiales, cada una teniendo 3 **UPM** con 3 elementos por **UPM**.

	Population A			Population B		
psu 1	10	20	30	9	10	11
psu 2	11	20	32	17	20	20
psu 3	9	17	31	31	32	30

Los elementos son los mismo en las dos poblaciones, por lo que comparten los valores $\overline{y}_U = 20$ y $S^2 = 84.5$. En la población A, las **UPM** son similares y la mayor parte de la variabilidad ocurre dentro de las **UPM**; en la población B, la mayor parte de la variablidad ocurre entre las **UPM**.

	Population A		Population B	
	$\bar{y}_{iU}$	S_i^2	$\bar{y}_{iU}$	S_i^2
psu 1	20	100	10	1
psu 2	21	111	19	3
psu 3	19	124	31	1

- Calcular las tablas de ANOVA para ambas poblaciones.

ANOVA Table for Population A:

Source	df	SS	MS
Between psus	2	6	3
Within psus	6	670	111.67
Total, about mean	8	676	84.5

For population A:

$$R_a^2 = 1 - \frac{111.67}{84.5} = -0.3215$$

$$\text{ICC} = 1 - \left(\frac{3}{2}\right)\frac{670}{676} = -0.4867$$

ANOVA Table for Population B:

Source	df	SS	MS
Between psus	2	666	333
Within psus	6	10	1.67
Total, about mean	8	676	84.5

For population B:

$$R_a^2 = 1 - \frac{1.67}{84.5} = 0.9803$$

$$\text{ICC} = 1 - \left(\frac{3}{2}\right)\frac{10}{676} = 0.9778$$

- La población A tiene mayor variación dentro de las **UPM**, pero poca variación entre las medias de las **UPM**, esto se ve reflejado en los valores negativos del **ICC** y R_a^2. Elementos en el mismo conglomerado son menos similares entre si que los de un grup de elementos escogidos al azar desde la población entera. En este caso, un muestreo por conglomerado es más eficiente.
- Caso contrario sucede en la población B: la mayor variabilidad ocurre entre las **UPM**, y los conglomerados son relativamente homogéneos. Así, el **ICC** y R_a^2 son muy cercanos a 1, indicando que muy poca nueva información será recogida al muestrear más de un elemento por **UPM**. En este caso, un muestreo por conglomerado unietápico es mucho menos eficiente que un **M.A.S.**.

Conglomerados de distinto tamaño: Estimación

Estimador insesgado

Bajo un muestreo por conglomerado unietápico de tamaño n de N **UPM**, sabamos como estimar el total y media poblacional de dos maneras: usando estimación insesgada y usando estimación por razón. Un estimador insesgado de t es calculado exactamente como antes, esto es:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} t_i$$

y,

$$SE[\hat{t}_{unb}] = N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n}}$$

La diferencia entre conglomerados de igual y distinto tamaño es que la variación entre los totales de los conglomerados individuales t_i es más probable que sea grande cuando los conglomerados tienen diferente tamaño.

La probabilidad que una **UPM** esté en la muestra es n/N, al igual que en un **M.A.S.** de tamaño n desde N **UPM**. Debido a que un **M.C** unietápico es utilizado, una **USM** es incluida en la muestra cada vez que su **UPM** es incluida en la muestra. Así,

$$w_{ij} = \frac{1}{\mathbb{P}\{j\text{-ésima } \textbf{USM} \text{ de la } i\text{-ésima } \textbf{UPM} \text{ esté en la muestra}\}} = \frac{N}{n}$$

Un muestreo por conglomerado unietápico procude una muestra autoponderada cuando las **UPM** son seleccionadas con igual probabilidades. Así, usando las ponderaciones se puede reescribir el estimador insesgado como:

$$\hat{t}_{unb} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_j} w_{ij} y_{ij}$$

Podemos utilizar las ecuaciones anteriores para obtener un estimador insesgado para $\overline{y}_U$ y para encontrar su error estándar. Definimos:

$$M_0 = \sum_{i=1}^{N} M_i$$

como el número total de **USM** en la población; entonces $\widehat{\overline{y}_{unb}} = \widehat{t}_{unb}/M_0$ y $SE[\widehat{\overline{y}_{unb}}] = SE[\hat{t}_{unb}]/M_0$.
El estimador insesgado de la media $\widehat{\overline{y}_{unb}}$ puede ser ineficiente cuando los valores de M_i son desiguales, pues dependen de la variabilidad de los totales por conglomerado. Adicionalmente, se requiere conocimiento de M_0.

Conglomerados de distinto tamaño: Estimador de razón

Usualmente esperamos que t_i esté positivamente correlacionado con M_i. Si las **UPM** son comunas, esperaremos que el número total de hogares viviendo en pobreza en una comuna $i(t_i)$ sea aproximadamente proporcional al número total de hogares en la comuna $i(M_i)$. La media poblacional $\overline{y}_U$ es la razón:

$$\overline{y}_U = \frac{\sum_i^N t_i}{\sum_{i=1}^N M_i} = \frac{t}{M_0}$$

donde t_i y M_i son usualmente positivamente correlacionados. Así,

$$\widehat{\overline{y}}_r = \frac{\widehat{t}_{unb}}{\widehat{M}_0} = \frac{\sum_{i\in\mathcal{S}} t_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}} M_i\overline{y}_i}{\sum_{i\in\mathcal{S}} M_i} = \frac{\sum_{i\in\mathcal{S}}\sum_{i\in\mathcal{S}_j} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{i\in\mathcal{S}_j} w_{ij}}$$

Dado que un **M.A.S.** de conglomerados es seleccionado, todos las ponderaciones son las mismas, esto es: $w_{ij} = N/n$.

Si las M_i son desiguales y un tamaño de muestra diferente de n es obtenido, el denominador será diferente. Así,

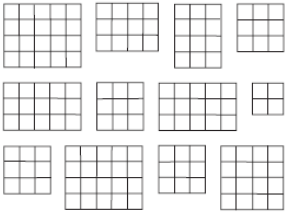
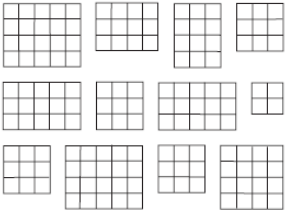
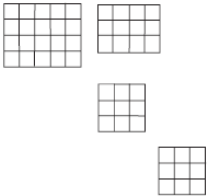
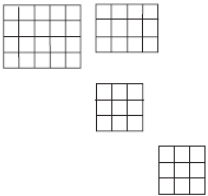
$$SE[\widehat{\overline{y}}_r] = \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\overline{M}^2}\frac{\sum_{i\in\mathcal{S}(t_i-\widehat{\overline{y}}_r M_i)^2}}{n-1}}$$
$$= \sqrt{\left(1-\frac{n}{N}\right)\frac{1}{n\overline{M}^2}\frac{\sum_{i\in\mathcal{S}} M_i^2(\overline{y}_i-\widehat{\overline{y}}_r)^2}{n-1}}$$

La varianza del estimador de razón depende de la variabiidad de las medias por elemento en los conglomerados, y puede ser mucho menor que la del estimador insesgado. Si el número total de elementos en la población, $M_0 = \sum_{i=1}^{N} M_i$ es conocido, podemos utilizar el estimador de razón para estimar el total poblacional: $\hat{t}_r = M_0\hat{\overline{y}}_r$ con $SE[\hat{t}_r] = M_0 SE[\hat{\overline{y}}_r]$.

Muestreo por conglomerado bietápico

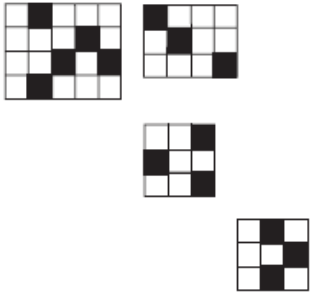
Como vimos antes, bajo un muestreo por conglomera unietápico se observan todas las **USM** dentro de las **UPM** seleccionadas. En muchas ocaciones, los elementos dentro de los conglomerados son bastante similares entre si, por lo que medir todos los elementos dentro de cada uno de ellos se vuelve un pérdida de recursos.
Alternativamente, es posible que sea costoso la medición de las **USM** respecto al costo de muestrear las **UPM**. En aquellas situaciones, puede ser más barato tomar una submuestra dentro de cada **UPM**. Así, las etapas dentro de un muestreo por conglomerado bietápico son:

- Seleccionar un **M.A.S.** $\mathcal{S}$ de n **UPM** de la población de N conglomerados.
- Seleccionar un **M.A.S.** de **USM** para cada **UPM** seleccionada. El **M.A.S.** de m_i elementos de la $i-$ésima **UPM** se denota por $\mathcal{S}_i$.

One-Stage	Two-Stage
Population of N psufs:	Population of N psu's:
	
Take an SRS of n psu's:	Take an SRS of n psu's:
	

Sample all ssu's in sampled psu's:

Take an SRS of m_i ssu's in sampled psu i:



Notamos que la diferencia entre un muestreo por conglomerado unietápico y un bietápico está en la última etapa. El muestreo de **USM** requiere la indexación doble que hablamos al inicio del curso.
Las estimaciones puntuales de t y $\overline{y}_U$ son análogas a las vistas para un muestreo por conglomerado unietápico, pero las fórmulas de varianza se vuelven más complicadas.

Bajo un muestreo por conglomerado unietápico, podemos estimar la población total mediante:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} t_i$$

El total por conglomerado t_i era conocido debido a que muestreamos cada **USM** en las **UPM** seleccionadas.
Bajo un muestreo por conglomerado *bietápico*, debido a que no muestramos todas las **USM** en las **UPM** obtenidas, debemos estimar los totales para cada uno de los conglomerados mediante:

$$\widehat{t_i} = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = M_i \overline{y}_i$$

y un estimador insesgado del total poblacional es:

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i\in\mathcal{S}}\hat{t}_i = \frac{N}{n}\sum_{i\in\mathcal{S}} M_i\overline{y}_i = \sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i}\frac{N}{n}\frac{M_i}{m_i}y_{ij}$$

Para estimar los totales y medias muestrales por conglomerado, la mayoría de estadísticos utiliza ponderaciones muestrales. La ecuación anterior sugiere que el peso (ponderación) de la $j-$ésima **USM** de la $i-$ésima **UPM** es

$$\frac{N}{n}\frac{M_i}{m_i},$$

y podemos comprobar que esto es así al calcular la probabilidad que un elemento particular pertenezca a la muestra, esto es:

$\mathbb{P}\{j\text{-ésima } \textbf{USM} \text{ de la } i\text{-ésima } \textbf{UPM} \text{ esté en la muestra}\} =$

$\mathbb{P}\{i\text{-ésima } \textbf{UPM} \text{ seleccionada}\} \times \mathbb{P}\{j\text{-ésima}\textbf{USM} \text{ seleccionada} | i\text{-ésima } \textbf{UPM} \text{ seleccionada}\}$

$= \frac{n}{N}\frac{m_i}{M_i}$

Así, el peso de que un elemento pertenezca a la muestra es el recíproco de su probabilidad de selección:

$$w_{ij} = \frac{NM_i}{nm_i}$$

Bajo un muestreo por conglomerado bietápico, los $\hat{t}_i$ son **variables aleatorias**, por lo que la varianza de $\hat{t}_{unb}$ tiene dos componentes:

- La variabilidad entre las **UPM**.
- La variabilidad dentro de las **UPM**.

En el caso unietápico, sólo nos preocupabamos por la primera componente pues la varianza sólo dependía de ello. En el caso bietápico, la varianza de $\hat{t}_i$ será la varianza utietápico más un término adicional que toma en cuanta la estimación de $\hat{t}_i$ en vez de su medición directa. Así,

$$\mathbb{V}[\hat{t}_{unb}] = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i}$$

en donde S_t^2 es la varianza población del total de conglomerados, y S_i^2 es la varianza poblacional de los elementos dentro del conglomerado $i-$ésimo.

¿Qué sucede en el caso $m_i = M_i$?
Ahora, al igual que siempre, debemos estimar la varianza de nuestra estimación $\hat{t}_{unb}$. Por lo que, sea:

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{N} \right)^2$$

la varianza muestral entre las estimaciones totales de los conglomerados y sea:

$$s_i^2 = \frac{1}{m_i - 1} \sum_{j \in S_i} (y_{ij} - \overline{y}_i)^2$$

la varianza muestral de las **USM** dentro del conglomerado (**UPM**) $i-$ésima.

El estimados insesgado de la varianza $\mathbb{V}[\hat{t}_{unb}]$ estará dado por:

$$\widehat{\mathbb{V}[\hat{t}_{unb}]} = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in \mathcal{S}} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$$

y su error estándar estimado, será la raíz cuadrada de la expresión anterior.

- En el caso en que N sea grande, la contribución del segundo término de la varianza es despreciable.

Al igual que en el caso unietápico, utilizamos estimación por razón para la media poblacional. Así, nuestro estimador de razón estará dado por:

$$\widehat{\overline{y}}_r = \frac{\sum_{i \in \mathcal{S}} \hat{t}_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\sum_{i \in \mathcal{S}} M_i \overline{y}_i}{\sum_{i \in \mathcal{S}} M_i}$$

en donde, la única diferencia con el caso unietápico es el reemplazo de t_i por $\hat{t}_i$. Y su cálculo, utilizando sus ponderaciones:

$$\widehat{\overline{y}}_r = \frac{\widehat{t}_{unb}}{\widehat{M}_0} = \frac{\sum_{i \in \mathcal{S}} \sum_{i \in \mathcal{S}_j} w_{ij} y_{ij}}{\sum_{i \in \mathcal{S}} \sum_{i \in \mathcal{S}_j} w_{ij}}$$

Los pesos son distintos, pero la forma del estimador es la misma que antes. Así, la varianza del estimador es:

$$\widehat{\mathbb{V}[\hat{\overline{y}}_r]} = \frac{1}{\overline{M}^2}\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN\overline{M}^2}\sum_{i\in\mathcal{S}} M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

en donde s_i^2 está dado por:

$$s_i^2 = \frac{1}{m_i-1}\sum_{j\in S_i}(y_{ij}-\overline{y}_i)^2$$

y,

$$s_r^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}(M_i\overline{y}_i - M_i\hat{\overline{y}}_r)^2$$

En donde $\overline{M}$ es el tamaño promedio de los conglomerados. Y al igual que antes, el segundo término en la estimación de la varianza conforme N es grande.

Tarea: Estudiar ejemplo 5.7, página 186. Sampling: Design and analysis, Lohr.

Ejemplo

Supongamos que queremos estimar el número promedio de patas de los perros de hogares de rescate en una ciudad. Esta ciudad tiene dos hogares: **Puppy Palace** con 30 cachorros y **Dog's Life** con 10 cachorros.
Seleccionamos un hogar con probabilidad **1/2**. Luego que el hogar es seleccionado, escogemos dos cachorros aleatoriamente de aquel hogar y usamos $\hat{\bar{y}}_{unb}$ para estimar el número promedio de patas de los cachorros.

Supongamos que seleccionamos el hogar **Puppy Palace**, sin mayor sorpresa, ambos cachorros observados tienen 4 patas, por lo que $\hat{t}_{PP} = 30 \times 4 = 120$. Luego usamos:

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} \hat{t}_i = \frac{N}{n} \sum_{i \in \mathcal{S}} M_i \overline{y}_i = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} \frac{N}{n} \frac{M_i}{m_i} y_{ij}$$

un estimador insesgado para el número total de patas en ambos hogares es:

$$\hat{t}_{unb} = \frac{2}{1} \hat{t}_{PP} = 240$$

Luego, dividimos el número total de patas estimadas por el número total de cachorros/perros para estimar el número medio de patas por perro, $240/40 = 6$.

Si ahora seleccionamos el hogar **Dog's Life**, $\hat{t}_{DL} = 10 \times 4 = 40$ y

$$\hat{t}_{unb} = \frac{2}{1}\hat{t}_{DL} = 80$$

al seleccionar el hogar **Dog's Life**, el estimador insesgado del número medio de patas por cachorro/perro es $80/40 = 2$.
Claramente, estos no son un buen estimador del número de patas por cachorro. Pero el estimar es matemáticamente insesgado: $(6 + 2)/2 = 4$, por lo que promediar todos los posibles resultados muestrales nos da el número correcto.

La mala calidad del estimador se ve reflejado en la gran magnitud de la varianza del estimador, ya que:

$$\begin{aligned}\mathbb{V}[\hat{t}_{unb}] &= N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i} \\ &= 2^2\left(1-\frac{1}{2}\right)\frac{S_t^2}{1} + \frac{2}{1}\sum_{i=1}^{2}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{S_i^2}{m_i} \\ &= \frac{1}{2}(4)(3200) = 6400\end{aligned}$$

En cambio, el estimador de razón es mejor debido a que: Si **Puppy palace** es seleccionado, $\hat{\bar{y}}_r = 120/30 = 4$; si **Dog's Life** es seleccionado, $\hat{\bar{y}}_r = 40/10 = 4$. Debido a que la estimación es la misma para todas las posibles muestras: $\mathbb{V}[\hat{\bar{y}}_r] = 0$

- En general, el estimador insesgado del total poblacional es ineficiente si los conglomerados son desiguales y t_i es aproximadamente proporcional a M_i. La varianza de $\hat{t}_{unb}$ depende de la varianza de los t_i y la varianza puede ser de gran magnitud si los M_i son desiguales.
- El estimador de razón, generalmente funciona bien cuando t_i es aproximadamente proporcional a M_i. Acá consideramos t_i la variable de interés y M_i la variable auxiliar. (Recordar Muestreo I)

Diseñando un muestreo por conglomerado

- Discusión diseño de muestreos
- Pasos para diseñar un muestreo por conglomerado

Problemáticas en un muestreo

- ¿Cuál es la precisión necesitada?
- ¿De qué tamaño deberían ser las **UPM**?
- ¿Cuántas **USM** deberían ser muestreadas de cada **UPM** seleccionada de la muestra?
- ¿Cuántas **UPM** deberían ser muestreadas?

La primera pregunta debe ser respondida en **cualquier** tipo de diseño de muestreo. Para responder la segunda y cuarta pregunta. se necesita saber el costo de muestrear una **UPM** para posibles tamaños de los mismos, el costo de muestrear una **USM** y la medida de homogeneidad (R_a^2 o ICC) para los posibles tamaños de conglomerados.

Escogiendo el tamaño de las UPM

Usualmente el tamaño de las **UPM** son una *unidad natural*, pero existen muestreos en los cuales el investigador puede escoger una gran gama de tamaños para las **UPM**, por ejemplo:

- grupo de individuos
- áreas de distintos tamaños

Un principio general para muestreos sobre *áreas* es que a mayor tamaño de las **UPM**, mayor será la variabilidad esperada dentro de cada **UPM**. Por lo que se espera tener un R_a^2 e ICC menores en magnitud con tamaños de conglomerado grandes que pequeños.
Sin embargo, si el tamaño de las **UPM** es muy grande, se podría perder la disminución de costos de un muestreo por conglomerado.

Ejemplo

En un estudio de cierto tipo de escarabajos en un determinado sitio, se evalúa cuantos tallos/ramas han de ser muestreadas por sitio bajo un muestreo por conglomerado. Así se presentan los siguientes datos:

Number of Stems Sampled per Site	$\hat{\bar{y}}$	SE($\hat{\bar{y}}$)	Cost to Sample One Field	Relative Net Precision
1	1.12	0.15	31.67	0.24
2	1.01	0.10	33.33	0.30
3	0.96	0.08	35.00	0.34
4	0.91	0.07	36.67	0.35
5	0.91	0.06	38.33	0.40

En donde el precisión relativa es calculada como:

$$1000/\text{costo} * CV(\hat{\bar{y}})$$

¿Qué configuración sería la más adecuada?

Escogiendo los tamaños de las USM

El objetivo de un diseño muestral es generalmente obtener la mayor información posible por el mínimo costo e inconveniencias. Ahora nos concentraremos en diseñar un muestreo por conglomerado bietápico en donde las **UPM** tienen el mismo número de **USM**. Un enfoque para iguales tamaños de conglomerados, es **minimizar la varianza**:

$$\mathbb{V}[\hat{t}_{unb}] = N^2 \left(1 - \frac{n}{N}\right) \frac{S_t^2}{n} + \frac{N}{n} \sum_{i=1}^{N} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{S_i^2}{m_i}$$

para un costo fijo

Si definimos $M_i = M$ y $m_i = m$ para todos los conglomerados, entonces $\mathbb{V}[\hat{\bar{y}}]$ puede ser escrito como:

$$\mathbb{V}[\hat{\bar{y}}_{unb}] = \left(1 - \frac{n}{N}\right) \frac{MSB}{nM} + \left(1 - \frac{m}{M}\right) \frac{MSW}{nm}$$

En donde **MSB** e **MSW** son los mismos que definidos en la tabla anova anteriormente. Podemos analizar varios casos:

- Si **MSW**=0 entonces $R_a^2 = 1$, y cada elemento de los conglomerados es igual a la media del conglomerado. Por lo que tomar más de un dato por **UPM** no entrega información adicional y aumenta el costo global.

Para otros valores de R_a^2, el óptimo depende del costo relativo de muestrear las **UPM** y **USM**. Consideremos una función simple de costo:

$$\text{Costo total} = C = c_1 n + c_2 nm$$

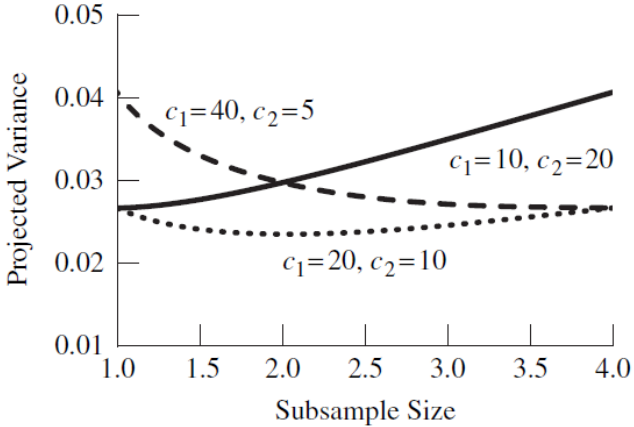
donde c_1 es el costo por **UPM** (sin incluir es costo de medir las **USM**) y c_2 es el costo de medir cada **USM**. Así,

$$n_{opt} = \frac{C}{c_1 + c_2 m_{opt}}$$

y,

$$m_{opt} = \sqrt{\frac{c_1 M(N-1)(1-R_a^2)}{c_2(NM-1)R_a^2}}$$

minimizan la varianza para un costo total fijo C. Por lo que, varios valores pueden satisfacer nuestras condiciones, por lo que compararlo con la varianza proyectada del estimar nos dará más información para poder determinar los tamaños. Una forma rápida de determinar los tamaños que buscamos es mediante gráficos.



Escogiendo el tamaño de muestra (Número de UPM)

Luego de que el tamaño de los conglomerados es determinado y la fracción de elementos a muestrear de ellos, nos concentramos en el número de conglomerados a muestrear. Como cualquier diseño de muestreo, diseñar un muestreo por conglomerado es un proceso iterativo:

- Determinar la precisión deseada
- Escoger los tamaños de los conglomerados y subelementos a muestrear.
- Estimar la varianza que se podrá logar con aquella configuración.
- Escoger el tamaño n para alcanzar la precisión deseada
- iterar. (agregar estratificaciones/variables auxiliares/etc) hasta que el costo del diseño esté dentro del presupuesto.

Si los conglomerados son de igual tamaño e ignoramos el factor de corrección, se tiene que:

$$\mathbb{V}[\widehat{\overline{y}_{unb}}] \leq \frac{1}{n}\left[\frac{MSB}{M} + \left(1 - \frac{m}{M}\right)\frac{MSW}{M}\right] = \frac{1}{n}v$$

y un intervalo de confianza $100(1-\alpha)\%$ aproximado será:

$$\widehat{\overline{y}_{unb}} \pm z_{\alpha/2}\sqrt{\frac{1}{n}v}$$

Si deseamos alcanzar el error de estimación e, fijamos $n = z^2_{\alpha/2}v/e^2$. Esto presupone que se tiene conocimiento anterior de término v, usualmento obtenido de muestreos anteriores.

Muestreo con Probabilidades desiguales:

Introducción

Hasta ahora, hemos visto sólo tipos de muestreo en donde se tiene la misma probabilidad de escoger las unidades muestrales. Estos diseños son bastante sencillos de diseñar y explicar, sin embargo, no son siempre viables en la práctica, y de serlo pueden no ser tan eficiente como un diseño usando probabilidades desiguales.

Lectura Complementaria

Sección 6.2.1: Selecting primary sampling units: Caso particular en donde sólo seleccionamos una sola UPM. Sampling: Design and Analysis, Lohr.

Supongamos que $n > 1$, y que el muestreo considera reemplazo. Un diseño con reemplazo significa que las probabilidades de selección no cambian luego de obtener la primera unidad. Sea:

$$\psi_i = \mathbb{P}(\text{Seleccionar la unidad } i \text{ en la primera selección})$$

Si muestreamos con reemplazo, entonces ψ_i es también la probabilidad que la unidad i sea seleccionada en la segunda selección o cualquiera subsiguiente.
La idea de un muestreo con probabilidades desiguales es sencilla: Seleccionar n **UPM** con reemplazo. Luego estimar el total poblacional, usando los estimados antes vistos, de forma separada para cada **UPM** obtenida.

Algunas **UPM** podrían ser seleccionadas más de una vez; las estimaciones del total poblacional, calculadas usando una **UPM** en particular, es incluida cuantas veces ese **UPM** fue seleccionada.
Debido a que las **UPM** son seleccionadas con reemplazo, tenemos n estimaciones del total poblacional independientes. Usando estos valores podemos estimar el total poblacional al calcular el promedio de esas n estimaciones de t. Finalmente, su varianza estimada será la varianza muestral de esas n estimaciones independientes de t, divida por n.

Selección de UPM

Existen varias formas de muestrear las **UPM** con probabilidades desiguales. Todas requieren que se tenga una medida del tamaño para cada **UPM** en la población. El **Método cumulativo de tamaño** extiende el proceso en el cual se generan números aleatorios y las **UPM** correspondientes a esos números son seleccionadas en la muestra.

Ejemplo

Consideremos la población de clases de introducción a la estadística en alguna universidad, como en la tabla adjunta.

Population of Introductory Statistics Classes

Class Number	M_i	ψ_i	Cumulative M_i Range	
1	44	0.068006	1	44
2	33	0.051005	45	77
3	26	0.040185	78	103
4	22	0.034003	104	125
5	76	0.117465	126	201
6	63	0.097372	202	264
7	20	0.030912	265	284
8	44	0.068006	285	328
9	54	0.083462	329	382
10	34	0.052550	383	416
11	46	0.071097	417	462
12	24	0.037094	463	486
13	46	0.071097	487	532
14	100	0.154560	533	632
15	15	0.023184	633	647
Total	647	1		

La universidad tiene 15 de tales clases; la clase i tiene M_i elementos, para un total de 647 estudiantes en los cursos de introducción a la estadística. Decidimos muestrear 5 clases con reemplazo y probabilidad proporcional a M_i, y luego recolectar los cuestionarios para cada estudiante muestreado. Para este ejemplo, se tiene que $\psi_i = M_i/647$

Para seleccionar la muestra, generamos 5 enteros aleatorios con reemplazo entre 1 y 647. Luego, las **UPM** a escoger en la muestra serán aquellas cuyos rangos acumulados M_i incluyen los números generados aleatoriamente.
Supongamos que los números generados son:

$$\{487, 369, 221, 326, 282\}$$

Por lo que las unidades muestrales serán:

$$\{13, 9, 6, 8, 7\}$$

Este método permite que la misma unidad pueda aparecer más de una vez en la muestra, por ejemplo: para números aleatorios:

$$\{553, 082, 245, 594, 150\}$$

la muestra será:

$$\{14, 3, 6, 14, 5\}$$

Otra forma de hacerlo, cuando los ψ_i no son proporcionales a M_i es hacer un rango acumulado ψ_i en vez de M_i, y generar números aleatorios entre 0 y 1.

Método de Lahiri

El **Método de Lahiri** puede ser más trazable que el método cumulativo cuando los tamaños de las **UPM** son grandes. Es un ejemplo de un método de *rechazo*, pues genera un par de números aleatorios para seleccionar **UPM** y luego rechaza algunas de ellas si el tamaño de la **UPM** es muy pequeño.
Sea $N =$ número de **UPM** en una población y $\max M_i =$ máximo tamaño de **UPM**. El método procede como:

- Obtener un número aleatorio entre 1 y N. Este indica que **UPM** se está considerando.
- Obtener un número aleatorio entre 1 y $\max M_i$. Si este número aleatorio es menor o igual a M_i, entonces se incluye la $i-$ésima **UPM** en la muestra; caso contrario, se vuelve al paso 1.
- Repetir hasta alcanzar el tamaño de muestra deseado.

Ejemplo

Utilizando el ejemplo anterior, aplicaremos el método de Lahiri.

Lahiri's Method, for Example 6.3

First Random Number (psu i)	Second Random Number	M_i	Action
12	6	24	$6 < 24$; include psu 12 in sample
14	24	100	Include in sample
1	65	44	$65 > 44$; discard pair of numbers and try again
7	84	20	$84 > 20$; try again
10	49	34	Try again
14	47	100	Include
15	43	15	Try again
5	24	76	Include
11	87	46	Try again
1	36	44	Include

Se tiene que la clase más grande tiene $\max M_i = 100$ estudiantes, por lo que generamos un par de enteros aleatorios, el primero entre 1 y 15, y el segundo entre 1 y 100, hasta que la muestra tiene 5 elementos.Así, las **UPM** a muestrear son:

$$\{12, 14, 14, 5, 1\}$$

Teoría de Estimación

Debido a que estamos muestreando con reemplazo, la muestra puede contener la misma **UPM** más de una vez. Sea $\mathcal{R}$ el conjunto de n unidades en la muestra, incluidas las repetidas. Por ejemplo:

$$\mathcal{R} = \{12, 14, 14, 5, 1\}$$

La **UPM** está incluida dos veces en el conjunto. Cuando muestreamos n **UPM** con reemplazo, tenemos n estimados independientes de t, por lo que los promediamos:

$$\widehat{t_\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}} \frac{t_i}{\psi_i} = \frac{1}{n}\sum_{i\in\mathcal{R}} u_i = \overline{u}$$

Estimamos $\mathbb{V}[\widehat{t_\psi}]$ como:

$$\widehat{\mathbb{V}[\widehat{t_\psi}]} = \frac{S_u^2}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}(u_i - \overline{u})^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \widehat{t_\psi}\right)^2$$

El estimador $\widehat{t_\psi}$ es llamado el estimador **Hansen-Hurwitz**. La estimación $\widehat{\mathbb{V}[\widehat{t_\psi}]}$ es la varianza estimada de los promedios $\overline{u}$ desde un muestreo aleatorio simple con reemplazo.

¿Son $\widehat{t_\psi}$ y $\widehat{\mathbb{V}(\widehat{t_\psi})}$ estimadores insesgados para t y $\mathbb{V}[\widehat{t_\psi}]$, respectivamente?
La respuesta es afirmativa. Para demostrar esta propiedad necesitamos variables aleatorios que hagan el seguimiento de cuales **UPM** aparecen múltiples veces en la muestra. Definamos:

$$Q_i = \text{ Número de veces que la unidad } i \text{ aparece en la muestra;}$$

Q_i es un análogo -con reemplazo- de la variable aleatoria Z_i que indica la inclusión de una muestra bajo un muestreo aleatorio simple sin reemplazo que vimos en el curso anterior.

Estimador Hansen-Hurwitz

Así, $\widehat{t_\psi}$ es el promedio de todos los $\dfrac{t_i}{\psi_i}$ de las unidades escogidas para pertenecer en la muestra, incluyendo las unidades cuantas veces hayan aparecido en la muestra:

$$\widehat{t_\psi} = \frac{1}{n}\sum_{i\in\mathcal{R}}\frac{t_i}{\psi_i} = \frac{1}{n}\sum_{i=1}^{N} Q_i\frac{t_i}{\psi_i}$$

Si una unidad aparece k veces en la muestra, esta es contada k veces en el estimador. Notar que: $\sum_{i=1}^{N} Q_i = n$ y $\mathbb{E}(Q_i) = n\psi_i$, por lo que $\widehat{t_\psi}$ es un estimador insesgado de t.

Para calcular la varianza, notamos que $\widehat{t_\psi}$ es el promedio de n observaciones independientes, cada una con varianza

$$\sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi} - t \right)^2$$

Por lo que,

$$\mathbb{V}[\widehat{t_\psi}] = \frac{1}{n} \sum_{i=1}^{N} \psi_i \left(\frac{t_i}{\psi_i} - t \right)^2$$

Esta ecuación incluye un promedio ponderado de los N valores de $\frac{t_i}{\psi_i} - t$, ponderamos por las probabilidades desiguales de selección ψ_i.

Para mostrar que la estimador de la varianza es insesgado para $\mathbb{V}[\widehat{t_\psi}]$, la escribimos en términos de la variable aleatoria Q_i:

$$\widehat{\mathbb{V}[\widehat{t_\psi}]} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{R}}\left(\frac{t_i}{\psi_i} - \widehat{t}_\psi\right)^2 = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{N} Q_i\left(\frac{t_i}{\psi_i} - \widehat{t_\psi}\right)^2$$

Ahora como es usual, para mostrar que un estimador es insesgado debemos obtener su esperanza.

$$
\begin{aligned}
\mathbb{E}[\widehat{\mathbb{V}(\widehat{t_\psi})}] &= \frac{1}{n(n-1)} \sum_{i=1}^{N} \mathbb{E}\left[Q_i \left(\frac{t_i}{\psi_i} - \widehat{t_\psi} \right)^2 \right] \\
&= \frac{1}{n(n-1)} \mathbb{E}\left[\sum_{i=1}^{N} Q_i \left(\frac{t_i}{\psi_i} - t + t - \widehat{t_\psi} \right)^2 \right] \\
&= \frac{1}{n(n-1)} \mathbb{E}[\sum_{i=1}^{N} Q_i \left(\frac{t_i}{\psi_i} - t \right)^2 + \sum_{i=1}^{N} Q_i (\widehat{t_\psi} - t)^2 \\
&- 2 \sum_{i=1}^{N} Q_i \left(\frac{t_i}{\psi_i} - t \right) (\widehat{t_\psi} - t)] \\
&= \frac{1}{n(n-1)} \mathbb{E}\left[\sum_{i=1}^{N} Q_i \left(\frac{t_i}{\psi_i} - t \right)^2 + n(\widehat{t_\psi} - t)^2 - 2n(\widehat{t_\psi} - t)^2 \right] \\
&= \frac{1}{n(n-1)} \left[\sum_{i=1}^{N} n\psi_i \left(\frac{t_i}{\psi_i} - t \right)^2 - n\mathbb{V}(\widehat{t_\psi}) \right] = \mathbb{V}(\widehat{t_\psi})
\end{aligned}
$$

En el caso en que N sea pequeño o algún ψ_i sea muy grande, es posible que la muestra consista en sólo un mismo elemento muestreado n veces. En tal caso, la varianza estimada será cero; de ocurrir es mejor realizar un muestreo sin reemplazo. Estimamos la media poblacional $\overline{y}_U$ como:

$$\widehat{\overline{y}_\psi} = \frac{\widehat{t_\psi}}{\widehat{M_{0\psi}}}$$

donde,

$$\widehat{M_{0\psi}} = \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{M_i}{\psi_i}$$

estima el número total de elementos en la población; $\widehat{\overline{y}_\psi}$ es un cuociente, por lo mediante métodos indirectos podemos obtener:

$$\widehat{\mathbb{V}[\widehat{\overline{y}_\psi}]} = \frac{1}{(\widehat{M_{0\psi}})^2} \frac{1}{n} \frac{1}{n-1} \sum_{i \in \mathcal{R}} \left(\frac{t_i}{\psi_i} - \frac{\widehat{\overline{y}_\psi} M_i}{\psi_i} \right)^2$$

Ejercicio

Utilizando el ejemplo de las clases introductorias de estadística, anteriormente al usar el método de Lahiri se obtuvo la muestra $\{12, 14, 14, 5, 1\}$. La respuesta t_i es el número total de horas que los estudiantes en la clase i estudiaron estadística la semana pasada, con los siguientes datos:

Class	ψ_i	t_i	$\frac{t_i}{\psi_i}$
12	$\frac{24}{647}$	75	2021.875
14	$\frac{100}{647}$	203	1313.410
14	$\frac{100}{647}$	203	1313.410
5	$\frac{76}{647}$	191	1626.013
1	$\frac{44}{647}$	168	2470.364

Obtener las siguientes cantidades:

- $\widehat{t}_{\psi}$
- $SE(\widehat{t}_{\psi})$
- $\widehat{\overline{y}_{\psi}}$
- $\widehat{\mathbb{V}[\overline{y}_{\psi}]}$

Diseñando las probabilidades de selección

En general, queremos escoger los ψ_i tales que la varianza de $\widehat{t}_\psi$ sea lo más pequeña posible. Idealmente, escogeríamos $\psi_i = t_i$ (así $\widehat{t}_\psi = t$ para todas las muestras y $\mathbb{V}[\widehat{t}_\psi] = 0$, pero esto no es práctico debido a que no se sabe apriori los t_i hasta después de la muestra.

Debido a que muchos de los totales en una **UPM** están relacionados con el número de elementos en el mismo, usualmente tomamos los ψ_i como la proporción de elementos en la **UPM** $i-$ésima o el tamaño relativo del mismo. Así, **UPM** más grandes tendrán mayor probabilidad de ser escogidos que sus pares con menos elementos.

Sea M_i el número de elementos en la **UPM** $i-$ésima y $M_0 = \sum_{i=1}^{N} M_i$ el número de elementos en la población, establecimos $\psi_i = M_i/M_0$.

Bajo esta elección, tendremos un muestreo con **probabilidades proporcionales al tamaño**. (Al igual que en el ejemplo anterior).

Lectura complementaria

Sección 6.2.4 ***Weights in Unequal-Probability Sampling with Replacement.*** Forma alternativa de estimación de los parámetros poblacionales utilizando los pesos w_{ij}.

Muestreo bietápico con probabilidades desiguales

Los estimadores para un muestreo bietápico con probabilidades desiguales son similares a los del caso unietápico. El procedimiento es como sigue:

- Tomar una muestra de i **UPM** con reemplazo, escogiendo la $i-$ésima **UPM** con probabilidad conocida ψ_i.
- Q_i cuenta el número de veces que la **UPM** $i-$ésima es escogida.
- Tomamos una **muestra probabilística** de m_i subunidades de la **UPM** $i-$ésima.

La única diferencia del caso biétapico con su versión unietápica, es que en el actual, debemos estimar t_i.
Otro punto a notar, es que el paso 3 requiere un **muestreo probabilístico**, este siendo usualmente un **M.A.S.** o **M.S.**

Este submuestreo debe cumplir dos condiciones:

- Cada vez que la **UPM** $i-$ésima es seleccionada para pertenecer a la muestra, el mismo diseño de submuestreo es usado para seleccionar la **USM** de esa **UPM**. Diferentes submuestreos de la misma **UPM**, sin embargo, deben ser muestreados independientemente, sino la **estimación de la varianza no será insesgada**.
- La $j-$ésima submuestra tomada desde la $i-$ésima **UPM** es seleccionada de tal manera que $\mathbb{E}[\widehat{t}_{ij}] = t_i$. Debido a que el mismo procedimiento es usado cada vez que la **UPM** $i-$ésima es seleccionada para pertenecer a la muestra. Podemos definir $\mathbb{V}[\widehat{t}_{ij}] = V_i \ \ \forall j$.

Los estimadores del caso unietápico son modificados para que permitan la selección de algunos elementos más de una sola vez. Así,

$$\widehat{t_\psi} = \frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \frac{\widehat{t_{ij}}}{\psi_i}$$

y,

$$\widehat{\mathbb{V}[\widehat{t_\psi}]} = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{N} \sum_{j=1}^{Q_i} \left(\frac{\widehat{t_{ij}}}{\psi_i} - \widehat{t_\psi} \right)^2$$

En resumen, los pasos para realizar un **muestreo bietápico con probabilidades desiguales con reemplazo**:

- Determinar las probabilidades de selección ψ_i, el número n de **UPM** a muestrear y el procedimiento de submuestreo a usar en cada **UPM** escogida.
- Seleccionar las n **UPM** con probabilidaes ψ_i y con reemplazo. Usando alguno de los métodos estudiados: Método cumulativa o de Lahiri (existen otros).
- Realizar submuestreo como se estipulo en el paso 1, si se repite alguna **UPM**, las muestras han de ser independientes.
- Estimar el total poblacional t para cada **UPM** en la muestra. El resultado serán n estimaciones en la forma $\widehat{t_{ij}}/\psi_i$.
- $\widehat{t_\psi}$ es el promedio de esas n estimaciones.
- Finalmente, la desviación estándar de nuestra estimación será la raíz de la varianza de los valores calculados en el paso 4.

Ejemplo

El ejemplo es similar al que ya hemos visto antes, pero ahora se desea estimar los t_i.

Class	M_i	ψ_i	y_{ij}	$\bar{y}_i$	$\hat{t}_i$	$\hat{t}_i/\psi_i$
12	24	0.0371	2, 3, 2.5, 3, 1.5	2.4	57.6	1552.8
14	100	0.1546	2.5, 2, 3, 0, 0.5	1.6	160.0	1035.2
14	100	0.1546	3, 0.5, 1.5, 2, 3	2.0	200.0	1294.0
5	76	0.1175	1, 2.5, 3, 5, 2.5	2.8	212.8	1811.6
1	44	0.0680	4, 4.5, 3, 2, 5	3.7	162.8	2393.9
			average			1617.5
			std. dev.			521.628

Calcule:

- $\widehat{\overline{y}_\psi}$ y su desviación estándar.

Muestreo con probabilidades desiguales sin reemplazo

Generalmente, muestrear con reemplazo es menos eficiente que muestrear sin reemplazo. El primer caso es más fácil de estudiar y análizar las muestrear. Sin embargo, en muestreos grandes con estratos/grupos pequeños, las ineficiencias pueden sobrepasar nuestras ganancias. Es claro que la complejidad de los muestreos sin reemplazo viene del hecho que las probabilidades de selección cambian al escoger un elemento y sus sucesiones.

Ejemplo

Consideremos la siguiente información:

Store	Size (m^2)	ψ_i	t_i (in Thousands)
A	100	$\frac{1}{16}$	11
B	200	$\frac{2}{16}$	20
C	300	$\frac{3}{16}$	24
D	1000	$\frac{10}{16}$	245
Total	1600	1	300

Seleccionemos 2 **UPM** sin reemplazo y probabilidades desiguales. Como antes, se tiene que:

$$\psi_i = \mathbb{P}(\text{ Seleccionar la unidad i-ésima en la primera elección })$$

Supongamos que escogemos la tienda A, por lo que:

$$\mathbb{P}(\text{ tienda A es escogida en la primera elección }) = \psi_A = \frac{1}{16}$$

Así, la probabilidad para la segunda elección en el caso que esta sea la tienda B:

$\mathbb{P}$(Tienda B es escogida en la 2da elección|Tienda A es escogida en la 1ra elección)

será $= \dfrac{\dfrac{2}{16}}{1 - \dfrac{1}{16}} = \dfrac{\psi_B}{1 - \psi_A}$

El denominador es la suma de los ψ_i para las tiendas B,C y D. En general se tiene:

$$\begin{aligned}
&\mathbb{P}(\text{ unidad i sea escogida primero, unidad k sea escogida segunda }) \\
&= \mathbb{P}(\text{unidad i sea escogida 1ro})\mathbb{P}(\text{unidad k escogida 2do}|\text{unidad i escogida 1ra}) \\
&= \psi_i \frac{\psi_k}{1-\psi_i}
\end{aligned}$$

De forma similar:

$$\begin{aligned}
&\mathbb{P}(\text{ unidad k sea escogida primero, unidad i sea escogida segunda }) \\
&= \psi_k \frac{\psi_i}{1-\psi_k}
\end{aligned}$$

Notar que estas probabilidades no son las mismas, pues el orden de selección hace diferencia.

Al sumar las probabilidades de las dos opciones, podemos encontrar la probabilidad que la muestra de tamaño 2 consista en las **UPM** i y k.
Para $n = 2$,

$$\mathbb{P}(\text{ unidad i y k en la muestra }) = \psi_i \frac{\psi_k}{1-\psi_i} + \psi_k \frac{\psi_i}{1-\psi_k} = \pi_{ik}$$

La probabilidad que la **UPM** $i-$ésima esté en la muestra es entonces:

$$\pi_i = \sum_{S: i \in \mathcal{S}} P(\mathcal{S})$$

La siguiente tabla muestras las probabilidades π_i y π_{ik} para las tiendas:

Store i	Store k: A	B	C	D	π_i
A	—	0.0173	0.0269	0.1458	0.1900
B	0.0173	—	0.0556	0.2976	0.3705
C	0.0269	0.0556	—	0.4567	0.5393
D	0.1458	0.2976	0.4567	—	0.9002
π_k	0.1900	0.3705	0.5393	0.9002	2.0000

Estimador Horvitz-Thompson para Muestreo unietápico

Asumamos que tenemos una muestra sin reemplazo de n **UPM**, y sabemos la **probabilidad de inclusión**:

$$\pi_{ik} = \mathbb{P}(\text{unidades i y k están en la muestra})$$

La probabilidad de inclusión π_i puede ser calculada como la suma de las probabilidades de todas las muestras que contienen la unidad $i-$ésima y tiene la propiedad que:

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

Y para los π_{ik} se tiene:

$$\sum_{k=1, k\neq i}^{N} = (n-1)\pi_i$$

Dentro de este contexto, tenemos el estimador **Horvitz-Thompson (HT)** para el total poblacional, definido como:

$$\widehat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}$$

Donde $Z_i = 1$ si la **UPM** $i-$ésima está en la muestra, y 0 en caso contrario.
El estimador **HT** es insesgado* para t. Ya que $\mathbb{P}(Z_i = 1) = \pi_i$ se tiene:

$$\mathbb{E}[\widehat{t}_{HT}] = \sum_{i=1}^{N} \pi_i \frac{t_i}{\pi_i} = t$$

Revisar Teorema 6.2, Sampling: Design and Analysis, Lohr.

Es posible mostrar que la varianza del estimador **HT**, está dada por: (HT y Sen-Yates-Grundy)

$$\begin{aligned}\mathbb{V}[\widehat{t_{HT}}] &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1, k \neq i}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2\end{aligned}$$

Es claro notar que la varianza del estimador **HT** es 0 si t_i es proporcional a π_i. **Cuando las probabilidades de inclusión π_i o las probabilidades conjuntas de inclusión π_{ik} son desiguales, al sustituir las cantidades muestrales se llegan a estimadores diferentes de la varianza.**

Ejemplo

Utilizando los mismos datos del ejemplo de las tiendas. Para seleccionar la primera **UPM**, primero generamos un número aleatorio entre $\{1, \ldots, 16\}$; Supongamos que salió el 12, el cual nos dice que la tienda D fue escogida, luego generamos un segundo número aleatorio entre $\{1, ..., 6\}$; supongamos que salió el 6, el cual nos dice que la tienda C es escogida 2da.
Así, el estimador **Horvitz-Thompson** para el total poblacional (ventas totales) para la muestra $\{C, D\}$ es:

$$\widehat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \frac{245}{0.9002} + \frac{24}{0.5393} = 316.6639$$

Debido a que en este ejemplo sabemos la población entera, podemos calcular la varianza teórica de $\widehat{t}_{HT}$ como:

$$\mathbb{V}[\widehat{t}_{HT}] = \frac{1}{2}\sum_{i=1}^{N}\sum_{k\neq i}^{N}(\pi_i\pi_k - \pi_{ik})\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2 = 4383,6$$

Como se dijo antes, las varianzas dan distinto si utilizamos el estimador HT o su equivalente SYG (tras reemplazar por las cantidades muestrales), respectivamente:

$$\widehat{\mathbb{V}_{HT}[\widehat{t}_{HT}]} = \sum_{i\in\mathcal{S}}\frac{1-\pi_i}{{\pi_i}^2}t_i^2 + \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},k\neq i}\frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}}\frac{t_i}{\pi_i}\frac{t_k}{\pi_k}$$

$$\widehat{\mathbb{V}_{SYG}[\widehat{t}_{HT}]} = \frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},k\neq i}\frac{(\pi_i\pi_k-\pi_{ik})}{\pi_{ik}}\left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k}\right)^2$$

Luego, reemplazando estos valores se obtiene para la muestra obtenida:

$$\widehat{\mathbb{V}_{HT}[\widehat{t}_{HT}]} = 6782.8$$

y,

$$\widehat{\mathbb{V}_{SYG}[\widehat{t}_{HT}]} = 3259.8$$

Debido a que todos los valores en esta población son conocidos, podemos examinar los estimadores para todas las muestras posibles. Las cuales se resumen en la siguiente tabla.

Variance estimates for all possible without-replacement samples of size 2, for the supermarket example

Sample, $\mathcal{S}$	$P(\mathcal{S})$	$\hat{t}_{\text{HT}}$	$\hat{V}_{\text{HT}}(\hat{t}_{\text{HT}})$	$\hat{V}_{\text{SYG}}(\hat{t}_{\text{HT}})$
{A, B}	0.01726	111.87	−14,691.5	47.1
{A, C}	0.02692	102.39	−10,832.1	502.8
{A, D}	0.14583	330.06	4,659.3	7,939.8
{B, C}	0.05563	98.48	−9,705.1	232.7
{B, D}	0.29762	326.15	5,682.8	5,744.1
{C, D}	0.45673	316.67	6,782.8	3,259.8

Notamos que para el caso de la varianza HT, 3 elementos dan **varianza negativa**. Esto a pesar de que ambas estimaciones de la varianza son **estimaciones insesgadas** para la varianza teórica.

Es fácil comprobar que:

$$\sum_{\text{posibles muestras } \mathcal{S}} \mathbb{P}(\mathcal{S})\mathbb{V}_{HT}\widehat{(\widehat{t}_{HT}}, \mathcal{S}) = \sum_{\text{posibles muestras } \mathcal{S}} \mathbb{P}(\mathcal{S})\mathbb{V}_{HT}\widehat{(\widehat{t}_{SYG}}, \mathcal{S}) = 4383.6$$

Este ejemplo muestra las posibles complicaciones que aparecen al estimar la varianza de $\widehat{t}_{HT}$. Los estimadores insesgado para la varianza de $\widehat{t}_{HT}$, pueden tomar valores negativos en ambos casos (HT y SYG) bajo ciertos **diseños con probabilidades desiguales**.

- Bajo distintas muestras la varianza podría escaparse demasiado.
- La estabilidad de la misma -aveces- puede ser mejorada mediante una cuidadosa elección del plan de muestreo, pero en general, los cálculos son bastante incómodos.
- Adicionalmente, en la práctica, ambos estimadores son díficiles de implementar debido a que requiere el conocimiento de la probabilidad de inclusión conjunto π_{ik}.

Alternativa

La alternativa a los métodos anteriores, propuesto por Durbin (1953).

$$\widehat{\mathbb{V}_{WR}(\widehat{t}_{HT})} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{t_i}{\psi_i} - \widehat{t}_{HT}\right)^2 = \frac{n}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{t_i}{\pi_i} - \frac{\widehat{t}_{HT}}{n}\right)^2$$

Este estimador soluciona los problemas de inestabilidad y costo computacional, pero fundamentalmente es errado, debido a que **propone no tomar en cuenta que la muestra no contempla reemplazo**, y utiliza los resultados de su análogo con reemplazo.

Selección de UPM

Existen bastantes métodos para seleccionar las **UPM** sin reemplazo tal que las probabilidades deseadas de inclusión se cumplan. En la práctica, un muestreo sistemático es el más usado para la selección de las **USM**. De ser este el caso, para la estimación de la varianza HT, se debe utilizar el estimador propuesto por Durbin. El estudio de los procedimiento para seleccionar las **UPM** es un tema en sí, y escapan de lo que estudiaremos en el curso. Pero se pondrá a disposición un libro especializado en el tema.

Estimador Horvitz-Thompson para Muestreo Bietápico

El estimador HT para muestreo bietápico es similar al caso unietápico:

$$\widehat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{t_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{t_i}{\pi_i}$$

En donde reemplazado t_i por un estimador insesgado $\widehat{t}_i$ de los totales por **UPM** para el valor desconocido de t_i, esto es:

$$\widehat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{\widehat{t}_i}{\pi_i} = \sum_{i=1}^{N} Z_i \frac{\widehat{t}_i}{\pi_i}$$

donde $Z_i = 1$ si la **UPM** $i-$ésima pertenece a la muestra, y 0 en caso contrario. El estimador HT para muestreo con probabilidades desiguales bietápico es insesgado para t siempre y cuando $\mathbb{E}[\widehat{t}_i] = t_i$ para cada **UPM**.

Es posible mostrar que la varianza de este estimador es:

$$\begin{aligned}\mathbb{V}[\widehat{t}_{HT}] &= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{k \neq i}^{N} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_i \pi_k} t_i t_k + \sum_{i=1}^{N} \frac{\mathbb{V}[\widehat{t}_i]}{\pi_i} \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1, k \neq i}^{N} (\pi_i \pi_k - \pi_{ik}) \left(\frac{t_i}{\pi_i} - \frac{t_k}{\pi_k} \right)^2 + \sum_{i=1}^{N} \frac{\mathbb{V}[\widehat{t}_i]}{\pi_i}\end{aligned}$$

En donde nuevamente, la varianza se expresa en sus formas HT y SYG, respectivamente. El último término representa la variabilidad adicional debido a la estimación de los t_i.

Así, el estimador HT para las varianzas anteriores bajo un muestreo por conglomerado (y prob. desiguales) bietápico son:

$$\widehat{\mathbb{V}_{HT}[\widehat{t}_{HT}]} = \sum_{i\in\mathcal{S}} \frac{1-\pi_i}{{\pi_i}^2}\widetilde{t}_i^2 + \sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},k\neq i} \frac{\pi_{ik}-\pi_i\pi_k}{\pi_{ik}} \frac{\widehat{t}_i}{\pi_i}\frac{\widehat{t}_k}{\pi_k} + \sum_{i\in\mathcal{S}} \frac{\widehat{\mathbb{V}[\widehat{t}_i]}}{\pi_i}$$

$$\widehat{\mathbb{V}_{SYG}[\widehat{t}_{HT}]} = \frac{1}{2}\sum_{i\in\mathcal{S}}\sum_{k\in\mathcal{S},k\neq i} \frac{(\pi_i\pi_k-\pi_{ik})}{\pi_{ik}} \left(\frac{\widehat{t}_i}{\pi_i} - \frac{\widehat{t}_k}{\pi_k}\right)^2 + \sum_{i\in\mathcal{S}} \frac{\widehat{\mathbb{V}[\widehat{t}_i]}}{\pi_i}$$

Al igual como discutimos antes, estas varianzas no son estables por lo que es recomendable utilizar el propuesto por Durbin.

$$\widehat{\mathbb{V}_{WR}(\widehat{t}_{HT})} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{n\widehat{t}_i}{\pi_i} - \widehat{t}_{HT}\right)^2 = \frac{n}{n-1}\sum_{i\in\mathcal{S}}\left(\frac{\widehat{t}_i}{\pi_i} - \frac{\widehat{t}_{HT}}{n}\right)^2$$

Introducción

A lo largo del curso (y el anterior) hemos vistos los diseños muestrales por separado pero en la práctica, muchos muestreos utilizan muestreos complejos, esto es, un diseño muestral que utiliza -en varias etapas- los componentes singulares que hemos estudiado.
Bajo este contexto, la estación de desviaciones típicas se vuelven complejas sobre todo si estamos en presencia de un muestreo sin reemplazo. Para solucionar lo anterior, pesos muestrales (factores de expansión) y efectos de diseño son utilizados.

Ensamblaje

Hemos visto la mayoría delos componentes utilizados en un muestreo complejo: Muestreo aleatorio simple, estimación de razón, estratos (y estratificación) y conglomerados. Ahora, nos concentraremos en estudiar como juntar estos componentes en un diseño. Haremos un pequeño repaso de los componentes que utilizaremos.

Bloques muestrales

Muestreo por conglomerado con reemplazo:
Seleccionamos una muestra de n conglomerados con reemplazo; **UPM** $i-$ésima es seleccionada con probabilidad ψ_i en la selección. Estimamos el total para la **UPM** $i-$ésima usando el estimador insesgado $\widehat{t}_i$. Luego, consideramos los n valores de $u_i = \hat{t}_i/\psi_i$ como observaciones. Estimamos la población total como $\overline{u}$, y estimamos la varianza del total poblacional como s_u^2/n.

Muestreo por conglomerado sin reemplazo:
Seleccionamos una muestra de n **UPM** sin reemplazo; π_i es la probabilidad que la $i-$ésima **UPM** sea incluida en la muestra. Estimamos el total para la **UPM** $i-$ésima usando el estimador insesgado $\widehat{t_i}$ y calculamos el estimador insesgado de la varianza $\widehat{\mathbb{V}(\hat{t}_i)}$.
Luego estimamos la población total con el estimador **HT**:

$$\hat{t}_{HT} = \sum_{i \in \mathcal{S}} \frac{\hat{t}_i}{\pi_i}$$

Estratificación:
Sean $\hat{t}_1, \ldots, \hat{t}_H$ los estimadores insesgados de los totales por estrato $t_1, \ldots, t_H$ y sean $\widehat{\mathbb{V}(\hat{t}_1)}, \ldots, \widehat{\mathbb{V}(\hat{t}_H)}$ los estimadores insesgados de la varianza. Luego estimamos el total poblacional como:

$$\hat{t} = \sum_{h=1}^{H} \hat{t}_h$$

y su varianza por:

$$\widehat{\mathbb{V}(\hat{t})} = \sum_{h=1}^{H} \widehat{\mathbb{V}(\hat{t}_h)}$$

Ejemplo Muestreo Complejo

La malaria ha sido un serio problema de salud en Gambia. Su prevalencia se ve reducida con el uso de redes impregnadas con insecticida en las camas de los hogares, pero sólo es efectivo si las redes tienen un uso masivo.

En 1991, se realizó una encuenta nacional diseñada para estimar la prevalencia de redes en área rurales. El marco muestral consistió en todas las villas rurales de menos de 3000 personas en Gambia. Las villas fueron estratificadas en tres regiones geográficas. (Oeste, Centro y Este) y si la villa en cuestión disponía de una clínica de salud pública.

En cada región 5 distritos fueron escogidos con probabilidad proporcional a la población del distrito usando datos censales. %>% En cada distrito 4 villas fueron escogidas, nuevamente con probabilidad proporcional a la población censal: 2 villas con Clínicas y 2 villas sin clínicas.

Finalmente, 6 recintos fueron escogidos -más o menos- aleatoriamente desde cada villa, y el investigador obtuvo el número de camas y redes, junto con otra información, para cada recinto.
En resumen:

Etapa	Unidad Muestral	Estratificación
1	Distrito	Región
2	Villa	Clínica/No-Clínica
3	Recinto	

Para calcular las estimaciones de las desviaciones típicas usando las fórmulas que hemos visto, se debiese empezar en la etapa 3 y luego avanzar hasta la 1ra etapa.

El procedimiento para estimar el número total de redes (sin usar estimación de razón):

- Registrar el número total de redes para cada recinto.
- Estimar el número total de redes para cada villa como (número total de recintos en la villa) $\times$ (número promedio de redes por recinto). Estimar la varianza del número total de redes, por villa.
- Estimar el número total de redes por villa con clínica en cada distrito y su varianza (usando fórmulas de un muestro con probabilidades desiguales). Repetir lo mismo para villas sin clínica.
- Sumar las estimaciones de cada estrato (con y sin clínica) para estimar el número de redes en cada distrito; sumar las varianzas estimadas de los dos estratos para estimar la varianza del distrito.
- Estimar el número total de redes y su varianza, para cada distrito. Luego, utilizar un muestreo por conglomerado bietápico para estimar el número total de redes por región.
- Finalmente, sumar las estimaciones totales de cada región para estimar el número total de redes en Gambia. Sumar las varianzas regionales para proceder como en un muestreo estratificado.

Estimación de razón en muestras complejas

La estimación por razón, puede ser usada en cualquier nivel de la encuesta, pero usualmente es usada en las últimas etapas. Los principios de esta técnica de estimación son los mismo para cualquier diseño probabilístico usado dentro de cada estrato en un muestreo estratificado a etapas.
Supongamos que la población total t_x es conocida para la variable auxiliar x, y que $\hat{t}_y$ y $\hat{t}_x$ son estimadores insesgado para sus equivalentes poblacionales.
El **Estimador de razón combinado** de la población total para la variable y es:

$$\hat{t}_{yrc} = \hat{B} t_x$$

donde,

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x}$$

Es posible mostrar que el error medio cuadrático de $\hat{t}_{yrc}$ puede ser estimado como:

$$\widehat{\mathbb{V}(\hat{t}_{yrc})} = \left(\frac{t_x}{\hat{t}_x}\right)^2 \left[\widehat{\mathbb{V}(\hat{t}_y)} + \hat{B}^2\widehat{\mathbb{V}(\hat{t}_x)} - 2\hat{B}\widehat{COV(\hat{t}_y, \hat{t}_x)}\right]$$

El **Estimador de razón separado** aplica una estimación de razón dentro de cada estrato primero, luego combina los estratos:

$$\hat{t}_{yrs} = \sum_{h=1}^{H} \hat{t}_{yhr} = \sum_{h=1}^{H} t_{xh}\frac{\hat{t}_{yh}}{\hat{t}_{xh}}$$

con

$$\widehat{\mathbb{V}(\hat{t}_{yrs})} = \sum_{h=1}^{H} \widehat{\mathbb{V}(\hat{t}_{yhc})}$$

Simplicidad en los diseños

Todos estos componentes del diseño han mostrado ser más eficientes en una encuesta tras realizarla. A veces, es posible tentarse con la utilización de un muestreo complejo, pero se debe investigar si ha habido estudios anteriores y asegurarse que efectivamente la implementación de un muestreo complejo es más eficiente.
En muchos casos, un muestreo más simple puede dar la misma información por unidad monetaria gastada que un muestreo complejo: es más fácil de analizar, administrar y los datos son menos propensos a ser interpretados erróneamente en investigaciones posteriores.

Construcción de los pesos muestrales

En la mayoría de encuestas a gran escala, los pesos son usados para calcular las estimaciones puntuales. Hasta ahora hemos visto este uso en muestreo aleatorio simple, estratificado y por conglomerado. Bajo un muestreo sin reemplazo, los pesos muestrales para una unidad observada es siempre el recíproce de la probabilidad que esa unidad sea incluida en la muestra. Recordemos que para una muestreo estratificado,

$$\hat{t}_{est} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}$$

donde los pesos muestrales $w_{hj} = (N_h/n_h)$ pueden ser vistos como el número de observaciones en la población representadas por la unidad muestral y_{hj}

La probabilidad de seleccionar la $j-$ésima unidad en el $h-$ésimo estrato para pertenecer a la muestra es $\pi_{hj} = n_h/N_h$, así el peso w_{hj} es el inverso de π_{hj}.
La suma de los pesos muestrales en un muestreo estratificado es igual al tamaño poblacional N; cada unidad muestreada *representa* cierto número de unidades en la población, así, la muestra completa *representa* la población total. La estimación de $\overline{y}_U$ bajo un muestreo estratificado está dado por:

$$\overline{y}_{str} = \frac{\sum_{h=1}^{L} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\sum_{h=1}^{L} \sum_{j \in \mathcal{S}_h} w_{hj}}$$

La misma forma de los estimadores es usado bajo un muestreo por conglomerado y en su forma general bajo un muestreo con probabilidades desiguales. Bajo un muestreo por conglomerado con igual probabilidades, por ejemplo:

$$w_{ij} = \frac{NM_i}{nm_i} = \frac{1}{\text{Prob. j-ésima USM en la i-ésima UPM esté en la muestra}}$$

nuevamente,

$$\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}$$

y el estimador de la media poblacional es:

$$\frac{\hat{t}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}}$$

Para un muestreo por conglomerado con probabilidades desiguales, cuando π_i es la probabilidad que la $i-$ésima **UPM** esté en la muestra y $\pi_{j|i}$ es la probabilidad que la $j-$ésima **USM esté en la muestra dado que la** $i-$**ésima** UPM** está en la muestra, los pesos muestrales son:

$$w_{ij} = \frac{1}{(\pi_i \pi_{j|i})}$$

Para un muestreo por conglomerado a tres etapas, el principio se extiende: Sea w_p los pesos para las **UPM**, $w_{s|p}$ es el peso para la **USM**, y $w_{t|s,p}$ el peso asociado a la unidad terciaria muestral **UTM**. Luego, el peso muestral total para una unidad observada es:

$$w = w_p \times w_{s|p} \times w_{t|s,p}$$

Toda la información necesaria para construir las estimaciones puntuales está contenida en los pesos muestrales, pero estos no dan información on como encontrar los errores estándar de las estimaciones, por lo que saber sólo los pesos muestrales no servirá para hacer estadística inferencial.

Las varianzas dependen de las probabilidad que cualquier par de observaciones sean seleccionadas para estar en la muestra, y requiere más conocimiento del diseño muestral que el dado por los pesos muestrales.
En lo que sigue del curso, consideraremos diseños estratificados multietápico, por lo que especificaremos la notación a utilizar.

Consideramos y_i la medición de la unidad $i-$ésima y w_i su peso asociado. Así, para una muestra aleatoria estratificada, y_i es una observación dentro de un estrato en particular, y $w_i = N_h/n_h$, donde la unidad $i-$ésima está en el estrato h. Esta notación nos permite escribir el estimador general para el total poblacional como:

$$\hat{t}_y = \sum_{i \in \mathcal{S}} w_i y_i$$

donde todas las mediciones están al nivel de la observación. El estimador general para la media poblacional está dado por:

$$\hat{\bar{y}} = \frac{\hat{t}_y}{\sum_{i \in \mathcal{S}} w_i}$$

El denominador estima el número de observaciones en la población.

Ejemplo

En el ejemplo de Gambia, el muestreo fue diseñado de tal forma que dentro de cada región cada recinto tendría la misma probabilidad de ser incluidos en la muestra; las probabilidades variaron sólo debido a que los diferentes distritos tienen diferentes números de personas en villas con clínica y debido a que el número de recintos podría no ser siempre exactamente proporcional a la población de la villa. Por ejemplo, para villas con clínicas de la región central, la probabilidad que un recinto dado sea incluido en la muestra fue:

$$\begin{aligned}
&\mathbb{P}(\text{distrito seleccionado}) \times \mathbb{P}(\text{villa seleccionada}|\text{distrito seleccionado}) \\
&\times \mathbb{P}(\text{recinto seleccionado}|\text{distrito y villa seleccionada}) \\
&\propto \frac{D1}{R} \times \frac{V}{D2} \times \frac{1}{C}
\end{aligned}$$

donde,

- C = número de recintos en la villa
- V = número de personas en la villa
- $D1$ = número de personas en el distrito
- $D2$ = número de personas en el distrito en villas con clínicas
- R = número de personas en villas con clínicas en todos los distritos centrales.

Debido a que el número de recintos en una villa será aproximadamente proporcional al número de personas en la villa, V/C debería ser aproximadamente el mismo para todos los recintos.
El valor de R es también el mismo para todos los recintos dentro de una región.
Los pesos para cada región, los recíprocos de las probabilidades de inclusión, difieren bastante debido a la variabilidad del término $D1/D2$.
Conforme R varía de estrato en estrato, recintos en estratos con más población tienen mayor pesos que en los mismos en estratos con menos gente.

Muestras autoponderadas y no autoponderadas

Bajo una muestra autoponderada, los pesos muestrales son todos iguales. Este tipo de muestras, bajo la ausencia de error no muestrales, puede ser considerada representativa de la población debido a que cada observación representa la misma cantidad de unidades no observadas en la población.
En este contexto, método estadísticos clásicos pueden ser utilizados para obtener las estimaciones puntuales.

La mayoría de las muestras autoponderadas usadas en la práctica no son M.A.S., sin embargo, estratificación es usada -comúnmente- para reducir la varianza y obtener estimaciones separadas sobre particularidades de interés; agrupamiento por conglomerados, usualmente con probabilidades desiguales, es usada -comúnmente- para reducir costos.

Estimación de la función de distribución

Hasta ahora nos hemos concentrado en estimar los parámetros poblacionales de usual interés: media, total y proporciones.
Históricamente, la teoría de muestreo fue desarrollada principalmente para encontrar estos 3 estadísticos básicos y responder preguntas básicas como: **"¿Qué porcentaje de...?"**
En la práctica, otros estadísticos podrían de ser interés, como por ejemplo la mediana y percentiles en general.

Podemos estimar estos estadísticos (pero no sus desviaciones estándar) mediante los pesos muestrales. Estos pesos nos permiten construir una **distribución empírica** para la población.
Supongamos que los valores para la población de N unidades son conocidos. Entonces, cualquier cantidad de interés puede ser calculada a partir de la **función de masa de probabilidad**,

$$f(y) = \frac{\text{número de unidades cuyos valores son } y}{N}$$

o la **función de distribución acumulada** (cdf),

$$F(y) = \frac{\text{número de unidades con valor} \leq y}{N} = \sum_{x \leq y} f(x)$$

En teoría de probabilidad, estas son las pmf y cdf para una variable aleatoria Y, donde Y es el valor obtenido desde una muestra aleatoria de tamaño uno desde la población. Así, $f(y) = \mathbb{P}(Y = y)$ y, $F(y) = \mathbb{P}(Y \leq y)$. Claramente, $\sum f(y) = F(\infty) = 1$

Cualquier cantidad poblacional puede ser calculada desde la masa la función de masa de probabilidad o cdf.

La media poblacional es:

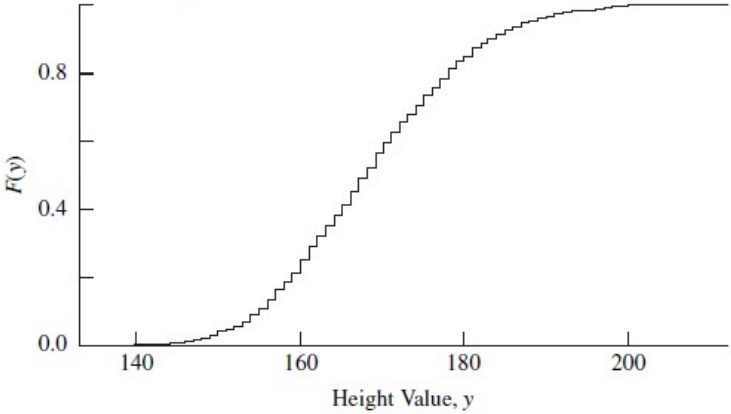
$$\overline{y}_U = \sum_{\text{valores de y en la pob.}} yf(y)$$

de igual manera la varianza poblacional, puede ser escrita usando la función de masa de probabilidad:

$$\begin{aligned} S^2 &= \frac{1}{N-A}\sum_{i=1}^{N}(y_i - \overline{y}_U)^2 \\ &= \frac{N}{N-1}\sum_{y} f(y)\left[y - \sum_{x} xf(x)\right]^2 \\ &= \frac{N}{N-1}\left[\sum_{y} y^2 f(y) - \left(\sum_{x} xf(x)\right)^2\right] \end{aligned}$$

Si la **cdf** F fuese continua, la mediana poblacional como el valor m que satisface $F(m) = 1/2$. Pero debido a que F salta en los valores de y en la población, es posible que la función $F(y)$ no contenga al valor $1/2$.
Para solucionar lo anterior, definimos la **mediana de población finita** como el valor m que satisface $F(m) = 1/2$ si es que ese valor existe; en caso contrario, la mediana poblacional es cualquier valor en el intervalo $[m_1, m_2]$, donde m_1 es el mayor valor de y en la población tal que $F(y) < 1/2$ y m_2 es el menor valor de y tal que $F(y) > 1/2$.

The function $F(y)$ for the population of heights.



En general, θ_q es un $q(100\%)$ cuantíl si $F(\theta_q) = q$ si es que ese valor existe; en caso contrario, $\theta_q \in [a, b]$ donde a es el mayor valor de y en la población tal que $F(y) < q$ y b es el menor valor de y tal que $F(y) > q$.
Si es que $q < 1/N, \theta_q$ es el menor valor de y y si $q > 1 - 1/N, \theta_q$ es el mayor valor de y.

Los pesos muestrales nos permiten construir la función de cuantía empírica y cdfs para los datos. Así, definimos la **función de cuantía empírica (epmf)** como:

$$\hat{f}(y) = \frac{\sum_{i \in \mathcal{S}; y_i = y} w_i}{\sum_{i \in \mathcal{S}} w_i}$$

y la **función de distribución empírica** como:

$$\hat{F}(y) = \sum_{x \leq y} \hat{f}(x)$$

Para una muestra auto-ponderada, $\hat{f}(y)$ se reduce a la frecuencia relativa de y en la muestra.
Para una muestra no auto-ponderada, $\hat{f}(y)$ y $\hat{F}(y)$ son intentos de reconstruir las funciones poblaciones f y F desde la muestra.
Tarea: Leer sección 7.4 Plotting Data from a Complex Survey

Efectos de diseño

¿Cómo podemos saber si un plan de muestreo es *mejor* que simplemente hacer un M.A.S?

En 1951, Cornfield sugirió medir la efectividad de un plan de muestreo como el cociente entre las estimaciones de las varianzas dadas por un M.A.S y un muestreo complejo. Luego en 1965, Kish llamó al reciproco de este valor **efectos de diseño (deff)** y lo utilizó para cualquier diseño muestral.

Así,

$$deff(\text{Plan,Estadístico}) = \frac{\mathbb{V}(\textbf{Estimador del plan de muestreo})}{\mathbb{V}(\textbf{Estimador bajo un M.A.S.})}$$

cabe destacar que el contraste se hace con un M.A.S. con el mismo número de observaciones. Por ejemplo, para una muestra de tamaño n, tendríamos:

$$deff(\text{plan}, \hat{\bar{y}}) = \frac{\mathbb{V}(\hat{\bar{y}})}{\left(1 - \frac{n}{N}\right)\frac{S^2}{n}}$$

Si bien los efectos de diseño, nos sirven para calcular la precisión ganada (o pérdida) al usar un muestreo más *complicado* que un M.A.S.; no nos evita el tener que calcular varianzas.
Es esperable que para distintas cantidades poblaciones a estimar, tengamos diferentes efectos de diseño. Kish mostró, como los **efectos de diseño** nos permiten utilizar conocimiento previo para nuestreo plan actual de muestreo.

En general la varianza bajo un M.A.S. es más fácil de obtener que $\mathbb{V}(\hat{\bar{y}})$:

- Si estamos estimando una proporción, la varianza bajo un M.A.S. es aproximadamente $p(1-p)/n$.
- Si estamos estimando otro tipo de media, la varianza bajo un M.A.S. es aproximadamente S^2/n.

Así, si el efecto de diseño es aproximadamente conocido, la varianza del estimador bajo un muestreo complejo puede ser estimada mediante $deff \times \mathbb{V}_{M.A.S.}$.

Por ejemplo, podemos estimar la varianza de una proporción estimada como:

$$\widehat{\mathbb{V}(\hat{p})} = deff \times \frac{\hat{p}(1-\hat{p})}{n}$$

Hasta ahora hemos visto los efectos de diseño de varios muestreos pero no los hemos nombrado como tal.
En terminos generales, queremos que nuestros efecto de diseño , para un estimación en particular sea menor a 1, pero otras consideraciones deben ser evaluadas.
Recordar, que al realizar un plan de muestreo en la práctica: los costos asociados al plan de muestreo influyen en la precisión a obtener.

Efectos de diseño e I.C.

Si los efectos de diseño para cada estadístico son conocidos, es posible construir intervalos de confianza estándar para la media y el total. Considerando una muestra de n unidades desde una población de tamaño N y $\hat{p}$ es la estimación del parámetro de interés, entonces. un I.C. del 95% para p (asumiendo factor de corrección ≈ 1), está dado por:

$$\hat{p} \mp 1.96\sqrt{deff}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Cuando estimamos la media en vez de la proporción, si la muestra es lo suficientemente grande para utilizar el teorema del límite central, un I.C. del 95% está dado por:

$$\hat{\bar{y}} \mp 1.96\sqrt{deff}\sqrt{\frac{S^2}{n}}$$

Ciertos autores (y softwares), a veces, utilizan en vez **deff** el concepto de **deft** que son bastante similares, pero este último en vez de dividir las varianzas divide las desviaciones típicas y asume un factor de corrección igual a 1.

Efectos de diseño y tamaños de muestra

Los efectos de diseño son extremadamente útiles para estimar el tamaño de muestra necesario en un estudio. Este fue el objetivo principal presentado por Cornfield. En su problemática, el máximo error permitido fue especificado en un 20% del valor real de la proporción, esto es: $0.2 \times p$. Así, el tamaño de muestra necesario bajo un M.A.S. estará dado por:

$$n = \frac{1.96^2 p(1-p)}{(0.2p)^2}$$

por lo que si la proporción real es 1%, $n = 9508$. En su problema, se deseaba analizar la alternativa de muestreo por bloques o individualmente; tras obtener los efectos de diseño por estudios pasados: $deff = 7.4$.
El estudio por bloques promediaba 4600 individuos (U.M), por lo que sugierió que 140.000 UM era lo ideal.

Si se conoce el efecto de diseño para un estudio similar (conocimientos previos), sólo se necesita estimar el tamaño de muestra que se necesitaria bajo un M.A.S. y luego multiplicar aquel tamaño por **deff** para obtener el número de observaciones necesarias bajo un muestreo complejo.
Se sugiere separar los efectos de diseño por estrato.

Tarea: Leer caso estudio: Sección 7.6 The National Crime Victimization Survey

Estimación de la varianza en muestreos complejos

La media y total poblacional son fácilmente estimados usando los pesos muestrales, no así estimar sus varianzas; en el ejemplo del país Gambia esbozamos el procedimiento para estimar la varianza bajos distintos niveles de estratificación (y de conglomerado), en donde empezamos en los niveles más bajo y luego combinamos los resultados para llegar al nivel global.

A lo largo de los muestreos estudiados, hemos explicitado las fórmulas de las varianzas; siendo unas más complejas que otras, como es el caso bietápico en un muestreo por conglomerado (Sin reemplazo).

Ahora nos concentraremos en describir varios métodos para estimar la varianza de la estimación del total y otros estadísticos bajo un muestreo complejo.

Métodos de linealización

Todas las fórmulas de varianza que hemos visto (desde M.A.S. hasta muestreo con probabilidades desiguales) eran para estimadores de media y totales. Estas fórmulas pueden ser usadas para encontrar la varianza para cualquier combinación lineal de medias y totales estimados.

Sea y_{ij} la respuesta de la unidad i al item j. Supongamos que $\hat{t}_1, \ldots, \hat{t}_k$ son estimadores insesgados de los k totales poblacionales $t_1, \ldots, t_k$ con $\hat{t}_j = \sum_{i \in \mathcal{S}} w_i y_{ij}$. Entonces, para cualquier constante $a_1, \ldots, a_k$ podemos definir una nueva variable:

$$q_i = \sum_{j=1}^{k} a_j y_{ij}$$

tal que:

$$\hat{t}_q = \sum_{i \in \mathcal{S}} w_i q_i = \sum_{j=1}^{k} a_j \hat{t}_j$$

$$\mathbb{V}\left(\sum_{j=1}^{k} a_j \hat{t}_j\right) = \mathbb{V}(\hat{t}_q) = \sum_{j=1}^{k} a_j^2 \mathbb{V}(\hat{t}_j) + 2 \sum_{j=1}^{k-1} \sum_{l=j+1}^{k} a_j a_l COV(\hat{t}_j, \hat{t}_l)$$

Así, si t_1 es el número total de dólares reportados por víctimas de robo, t_2 es el número de días laborales de ausencia debido al delito, y t_3 es el total de sus costos médicos, una medida para las consecuencia financieras del robo (asumiendo \$150 por día no trabajado) podría ser $\hat{t}_1 + 150\hat{t}_2 + \hat{t}_3$, y su varianza estaría dada por:

$$\begin{aligned}\mathbb{V}(\hat{t}_1 + 150\hat{t}_2 + \hat{t}_3) &= \mathbb{V}(\hat{t}_q) \\ &= \mathbb{V}(\hat{t}_1) + 150^2\mathbb{V}(\hat{t}_2) + \mathbb{V}(\hat{t}_3) \\ &+ 300COV(\hat{t}_1, \hat{t}_2) + 2COV(\hat{t}_1, \hat{t}_3) + 300COV(\hat{t}_2, \hat{t}_3)\end{aligned}$$

donde $q_i = y_{i1} + 150y_{i2} + y_{i3}$ es la pérdida financiera debido al robo para la persona $i-$ésima.

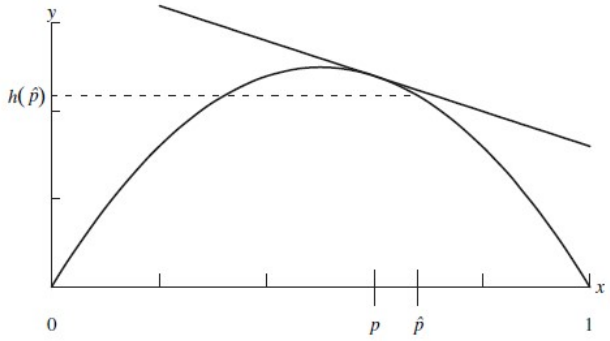
Supongamos ahora que estamos interesados en la proporción de la pérdida total contabilizada por el robo de propiedad, t_1/t_q. Este no es un estadístico lineal, ya que t_1/t_q no puede ser expresado en la forma $a_1 t_1 + a_2 t_2$ para constantes a_1, a_2.

Pero el teorema de Taylor (de cálculo) nos permite **linealizar** a función suave no lineal $h(t_1, t_2, \ldots, t_k)$ del total poblacional; el teorema de Taylor nos entrega las constantes $a_0, a_1, \ldots, a_k$ tal que:

$$h(t_1, \ldots, t_k) \approx a_0 + \sum_{j=1}^{k} a_j t_j$$

Así $\mathbb{V}[h(\hat{t}_1, \ldots, \hat{t}_k)]$ puede ser aproximado por $\mathbb{V}(\sum_{j=1}^{k} a_j \hat{t}_j)$, el cual sabemos calcular por la fórmula antes dada.

La cantidad $\theta = p(1-p)$, donde p es la proporción poblacional, puede ser estimada por $\hat{\theta} = \hat{p}(1-\hat{p})$. Asumiendo que $\hat{p}$ es un estimador insesgado de p y que $\mathbb{V}(\hat{p})$ es conocido. Sea $h(x) = x(1-x)$, entonces $\theta = h(p)$ y $\hat{\theta} = hat(\hat{p})$. Obviamente h es una función no-lineal de x, pero la función puede ser aproximada en cualquier punto cercado a por la línea tangente a la función:



Ejemplo

La versión -de primer orden- del teorema de Taylor, establece que si la segunda derivada de h es continua, entonces:

$$h(x) = h(a) + h'(a)(x-a) + \int_a^x (x-t)h''(t)dt$$

bajo condiciones regulares -usualmente satisfechas dentro del contexto de estadística-, el último término es relativamente pequeño a los dos primeros, usando la aproximación:

$$\begin{aligned} h(\hat{p}) &\approx h(p) + h'(p)(\hat{p}-p) \\ &= p(1-p) + (1-2p)(\hat{p}-p) \end{aligned}$$

Así,

$$\mathbb{V}[h(\hat{p})] \approx (1-2p)^2 \mathbb{V}(\hat{p}-p)$$

como $\mathbb{V}(\hat{p})$ es conocido, la varianza aproximada de $h(\hat{p})$ puede ser estimada por:

$$\widehat{\mathbb{V}[h(\hat{p})]} = (1-2\hat{p})^2 \widehat{V}(\hat{p})$$

Procedimiento general

El procedimiento para linealizar un estimador de la varianza de un estimador no-lineal de medias o totales:

- Expresar la cantidad de interés como una función de medias o totales de variables medidas o calculadas a partir de la muestra. En general, $\theta = h(t_1, t_2, \ldots, t_k)$ o $\theta = h(\overline{y}_{1U}, \ldots, \overline{y}_{kU})$.
- Encontrar las derivadas parciales de h con respecto a cada argumento. Las derivadas parciales, evaluadas en las cantidades poblaciones, desde la constantes de linealización a_j.
- Aplicar el teorema de Taylor para linealizar la estimación:

$$h(\hat{t}_1, \ldots, \hat{t}_k) \approx h(t_1, \ldots, t_k) + \sum_{j=1}^{k} a_j(\hat{t}_j - t_j)$$

donde,

$$a_j = \left.\frac{\partial h(c_1, c2, \ldots, c_k)}{\partial c_j}\right|_{t_1,\ldots,t_k}$$

- Definir la nueva variable q por

$$q_i = \sum_{j=1}^{k} a_j y_{ij}$$

Ahora encontramos las varianzas estimadas de $\hat{t}_q = \sum_{i \in \mathcal{S}} w_i q_i$, sustituyendo estimadores por las cantidades poblaciones desconocidas. Esto generalmente aproxima la varianza de $\hat{\theta} = h(\hat{t}_1, \ldots, \hat{t}_k)$

Ejemplo: Estimador de Razón

En el curso pasado vimos los estimadores de razón: estimaciones puntuales, varianzas y su sesgos asociados, en donde bajo un M.A.S. el estimador:

$$\hat{B} = \overline{y}/\overline{x} = \hat{t}_y/\hat{t}_x$$

Que nosotros notamos por la letra R en Muestreo I, y su varianza estimada:

$$\widehat{\mathbb{V}(\hat{B})} = \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\overline{x}^2}$$

donde s_e^2 es la varianza muestral de los residuos $e_i = y_i - \hat{B}x_i$. Para llegar a esta aproximación de la varianza, lo que utilizamos (implícitamente) fue el teorema de Taylor.

Siguiendo los pasos que antes establecimos:

- Expresamos el estimador B como función de los totales poblacionales. Sea $h(c, d) = d/c$, así

$$B = h(t_x, t_y) = \frac{t_y}{t_x}$$

y,

$$\hat{B} = h(\hat{t}_x, \hat{t}_y) = \frac{\hat{t}_y}{\hat{t}_x}$$

Asumiendo que los estimadores $\hat{t}_x$ y $\hat{t}_y$ son insesgados.

- Las derivadas parciales son:

$$\frac{\partial h(c, d)}{\partial c} = -\frac{d}{c^2}$$

y,

$$\frac{\partial h(c, d)}{\partial d} = \frac{1}{c}$$

evaluadas en $c = t_x$ y $d = t_y$, los valores son $-t_y/t_x^2$ y $1/t_x$.

- Por teorema de Taylor:

$$\begin{aligned}\hat{B} &= h(\hat{t}_x, \hat{t}_y) \\ &\approx h(t_x, t_y) + \left.\frac{\partial h(c,d)}{\partial c}\right|_{t_x,t_y} (\hat{t}_x - t_x) + \left.\frac{\partial h(c,d)}{\partial d}\right|_{t_x,t_y} (\hat{t}_t - t_t)\end{aligned}$$

Luego, usando las derivadas parciales se tiene que:

$$\hat{B} - B \approx -\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y)$$

- El error cuadrático medio (mse) de $\hat{B}$ es:

$$\begin{aligned}\mathbb{E}[(\hat{B}-B)^2] &\approx \mathbb{E}\left[\left(-\frac{t_y}{t_x^2}(\hat{t}_x - t_x) + \frac{1}{t_x}(\hat{t}_y - t_y)\right)\right] \\ &= \frac{1}{t_x^2}(B^2\mathbb{V}(\hat{t}_x) + \mathbb{V}(\hat{t}_y) - 2BCOV(\hat{t}_x, \hat{t}_y))\end{aligned}$$

- Sustituyendo estimadores de las cantidades desconocidas, definimos:

$$q_i = \frac{1}{\hat{t}_x}[t_i - \hat{B}x_i] = \frac{1}{\hat{t}_x}e_i$$

Finalmente, encontramos $\widehat{\mathbb{V}(\hat{B})} = \widehat{\mathbb{V}(\hat{t}_q)} = \widehat{\mathbb{V}(\hat{t}_e)}/\hat{t}_x^2$

Ventajas y desventajas

- **Ventajas:** Si las derivadas parciales son conocidas, el procedimiento de linialización casi siempre da una varianza estimada para un estadístico dado, y este puede ser utilizado en diseños muestrales generales. La teoría respecto a este tipo de aproximación está bastante desarrollada y por ende, implementado en varios softwares.
- **Desventajas:** Los cálculos no son sencillos, y el método es difícil de aplicar con función complejas que incorporen ponderaciones. Se debe encontrar una expresión analítica para las derivadas parciales de h o calcular las derivadas parciales **numéricamente**. Un fórmula de varianza es necesaria para cada estadístico a estimar, por lo que requiere mayor tiempo de programación; un método diferente es necesario para cada estadístico. **No todos los estadísticos pueden ser expresados como función suave de los totales poblacionales** (por ejemplo los cuantiles). La precisión depende del tamaño de muestra (la estimación de la varianza suele subestimar si la varianza no es lo suficientemente grande).

Replicación de un diseño muestral.

Supongamos que un diseño muestral básico es **replicado independientemente** R veces.

Independientemente hace referencia a que cada uno de los R conjuntos de variables aleatorias usadas para seleccionar la muestra es independiente de los otros conjuntos. *Después de cada muestra, las unidades muestrales son reemplazadas en la población, estando disponible para futuras muestras.*

Las R muestras replicadas producen R estimaciones independientes de la cantidad de interés; la variabilidad dentro de esas estimaciones pueden ser usadas para estimar la varianza de $\hat{\theta}$. Mahalanobis (1939,1946) describe los primeros usos de este método, que el llama *"redes replicadas de unidades muestrales"*

Sea

$$\begin{aligned}\theta &= \text{ parámetro de interés}\\ \hat{\theta}_r &= \text{ estimación de } \theta \text{ calculada desde la } r\text{-ésima réplica}\\ \tilde{\theta} &= \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}_i\end{aligned}$$

Si $\hat{\theta}_r$ es un estimador insesgado de θ, también lo es $\tilde{\theta}$, y:

$$\widehat{\mathbb{V}_1(\tilde{\theta})} = \frac{1}{R}\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \tilde{\theta})^2$$

es un estimador insesgado de $\mathbb{V}(\tilde{\theta})$. Notar que $\widehat{\mathbb{V}_1(\tilde{\theta})}$ es la varianza muestral de las R estimaciones independientes de θ dividida por R (estimador usual de la varianza de una media muestral).

Dividiendo la muestra en grupos aleatorios

En la práctica, las submuestras no son usualmente obtenidas independientemente, pero la muestra completa es seleccionada de acuerdo al diseño muestral. La muestra completa es dividida en R grupos, por lo que cada grupo forma una versión miniatura del muestreo, reflejando el diseño muestral. Los grupos son tratados como si fueran réplicas independientes del diseño muestral básico.

Si la muestra es un **M.A.S.** de tamaño n, los grupos son formados de forma aleatoria repartiendo las n observaciones entre R grupos, cada uno de tamaño n/R.

Estos pseudo grupo no son del todo replicadas independientes debido a que una unidad muestral puede sólo ser incluida en uno de los grupos; si la población es relativamente grande al tamaño muestral, es posible tratar los grupos como si estos fueran réplicas independientes.

En un **muestreo por conglomerado**, las **UPM** son divididas aleatoriamente entre los R grupos. Las **UPM** llevan consigo todas las unidades dentro de sí al grupo asignado, por lo que cada grupo aleatorio es una muestra por conglomerado

En un **muestreo estratificado multietápico**, un grupo aleatorio contiene una muestra de **UPM** de cada estrato. Notamos que si k **UPM** son muestreadas en el estrato más pequeño, a lo más k grupo aleatorios pueden ser formados.

Si θ es una cantidad no lineal, $\tilde{\theta}$ -en general- no será igual a $\hat{\theta}$(el estimador calculado directamente desde la muestra completa).

Ventajas y desventajas

- **Ventajas:** Ningún software especial es necesario para estimar la varianza, y es bastante sencillo calcular la estimación de la varianza. El método está bien definido para problemas multiparamétricos o no-paramétricos. Puede ser utilizado para estimar varianzas para percentiles y funciones no suaves.
- **Desventajas:** El número de grupos aleatorios es usualmente pequeño, esto provoca estimaciones imprecisas de la varianza. Si $\hat{\theta}$ es un estadístico no lineal, $\tilde{\theta}$ puede tener un sesgo grande si el número de observaciones en cada grupo es pequeño. Generalmente, son necesarios 10 grupos para obtener un estimador estable de la varianza y evitar inflar el intervalo de confianza (debido al cuantíl t y un n bajo para g.l). Ajustar los grupos aleatorios puede ser complicado bajo diseños muestrales complicados, debido a que cada grupo debe tener la misma estructura de diseño que el muestro completo.

Métodos de Remuestreo y replicación:

Debido a que éstas técnicas pueden ser estudiadas en extenso, nos concentraremos -brevemente- en las dos técnicas más utilizadas:

- Jacknife
- Bootstrap

pero existen **bastantes** otros métodos.

Jackknife

El método **jackknife** extiende el método de agrupación aleatoria al permitir que los grupos replicados se superpongan. Inicialmente, este método fue introducido como un método para reducir el sesgo (Quenouille, 1956); en 1958, Tukey propuso utilizarlo para estimar varianzas y calcular intervalos de confianza.

Estudiaremos el método *delete-1 jackknife*, pero existen otras formas para este método.

Para un M.A.S., sea $\hat{\theta}_{(j)}$ el estimador con la misma forma que $\hat{\theta}$, pero sin usar la observación $j-$ésima, así, si $\hat{\theta} = \overline{y}$ entonces,

$$\hat{\theta}_{(j)} = \overline{y}_{(j)} = \sum_{i \neq j} y_i/(n-1)$$

Para un M.A.S., definimos el estimador *delete-1 jackknife* como:

$$\widehat{\mathbb{V}_{JK}(\hat{\theta})} = \frac{n-1}{n} \sum_{j=1}^{n} (\hat{\theta}_{(j)} - \hat{\theta})^2$$

Ventajas y desventajas

- **Ventajas:** El método Jackknife es multipropósito. El mismo procedimiento es usado para estimar la varianza para cada estadístico en el cual jackknife puede ser usado. Jackknife entrega un estimador consistente de la varianza cuando θ es una función suave de totales poblacionales.
- **Desventajas:** Para algunos diseños muestrales, jackknife puede requerir una cantidad grande de cálculos (haciéndolo computacionalmente costoso). Jackknife no funciona muy bien para estimar la varianza de algunos estadísticos que no son funciones suaves de los totales poblacionales (por ejemplo, jackknife no entrega un estimador consistente para la varianza de cuantiles en un M.A.S.).

Bootstrap

Al igual que Jackknife, los resultados teóricos de la técnica bootstrap fueron primeramente desarrollados para áreas de estadísticas que no eran la teoría de muestreo.

Supongamos que $\mathcal{S}$ es una muestra desde un M.A.S. con reemplazo de tamaño n. Nosotros esperamos que al obtener la muestra, esta reproduzca las propiedades de la población completa. Luego, consideramos la muestra $\mathcal{S}$ como si fuera la población entera, y tomamos **remuestras** desde $\mathcal{S}$.

Si la muestra realmente es similar a la población -**Si la función de masa de probabilidad empírica de la muestra es similar a la función de masa de probabilidad de la población**-, entonces las muestras generadas desde la función de masa de probabilidad empírica deberían comportarse como muestras obtenidas desde la población.

- **Ventajas:** Bootstrap funciona con funciones suaves de medias poblacionales y para algunas función no-suaves como los cuantiles para diseños muestrales generales. Bootstrap permite encontrar intervalos de confianza directamente.
- **Desventajas:** Bajo ciertas configuraciones, bootstrap podría requerir más cálculos que Jackknife u otros métodos, ya que R (las réplicas) es usualmente un número grande. Las estimaciones bajo este método difieren cuando un conjunto diferente de muestras obtenidas por bootstrap son utilizadas.

Funciones de varianza generalizadas

Las funciones de varianza generalizadas son un modelo matemático que describe la **relación entre los valores poblacionales a ser estimados (como los totales poblacionales) y la varianza de su estimador**.

Por ejemplo, en el caso *The National Crime Victimization Survey*, los investigadores postulan que:

$$\widehat{\mathbb{V}(\hat{t})} = a\hat{t}^2 + b\hat{t}$$

En donde $\hat{t}$ es el número estimado de personas afectadas por un tipo de crimen en particular.

¿Cómo fue encontrada esta expresión?

El procedimiento general para construir una función de varianza generalizada es:

- Usando replicación (u otro método), estimar las varianzas de las k totales poblaciones de especial interés, $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_k$. Sean v_i las varianzas relativas para $\hat{t}_i$.

$$v_i = \frac{\widehat{\mathbb{V}(\hat{t}_i)}}{\hat{t}_i^2} \qquad i = 1, 2, \ldots, k$$

- Postular un modelo que relacione v_i con $\hat{t}_i$. Un ejemplo de modelo es:

$$v_i = \alpha + \frac{\beta}{\hat{t}_i}$$

  Esto es un modelo de regresión lineal en donde la variable respuesta es v_i y la variable explicativa es $\frac{1}{\hat{t}_i}$.
- Usar métodos de estimación usuales para encontrar las estimaciones de α y β: a y b, respectivamente.

- Usar la ecuación de regresión para predecir la varianza relativa de un total estimado

$$\hat{t}_{nuevo} : \hat{v}_{nuevo} = a + b/\hat{t}_{nuevo}$$

Debido a que $\hat{v}_{nuevo}$ es la predicción del valor de la varianza relativa $\frac{\widehat{\mathbb{V}(\hat{t}_{nuevo})}}{\hat{t}^2_{nuevo}}$, la función de varianza generalizada estimada de $\mathbb{V}(\hat{t}_{nuevo})$ es:

$$\widehat{\mathbb{V}(\hat{t}_{nuevo})} = a\hat{t}^2_{nuevo} + b\hat{t}_{nuevo}$$

Ventajas y desventajas

- **Ventajas:** Las F.V.G. pueden ser usadas cuando no se entrega suficiente información en conjuntos de datos públicos que nos permitan calcular directamente los errores estándar. Usualmente quienes dirigen los muestreos, pueden calcular las F.V.G. y además tienen mayor información para estimar las varianzas que la entrega al público general. Las F.V.G. ayudan a ahorrar tiempo en la confección de reportes anuales y sirven para diseñar estudios similares en el futuro.
- **Desventajas:** El modelo que relaciona v_i con $\hat{t}_i$ puede no ser apropiado para la cantidad de interés, lo que puede resultar en estimaciones de variable no confiables; todas las propiedades de predicción de regresiones lineales se heredan. **Si una subpoblación tiene un número alto de conglomerados (y por consecuencia un alto deff), la estimación por F.G.V puede ser muy pequeña.**

Intervalos de Confianza

Muchos de los resultados teóricos de los métodos de estimación de varianza que hemos vistos, asumen que:

$$\frac{(\hat{\theta} - \theta)}{\sqrt{\widehat{\mathbb{V}(\hat{\theta})}}} \longrightarrow N(\mu, \sigma^2)$$

Lo que nos permite construir intervalos de confianza usuales, de la forma:

$$\hat{\theta} \pm 1.96\sqrt{\widehat{\mathbb{V}(\hat{\theta})}}$$

Alternativamente, el cuantil normal puede ser cambiado por un cuantil t-student con g.d.l = número de grupos - 1 para métodos de agrupación aleatoria y, g.d.l= número de **UPM** - número de **estratos**

El intervalo de confianza calculado directamente por bootstrap también es válido.

A modo general, los supuestos de los métodos que hemos vistos son:

- La cantidad de interés θ puede ser expresada como una función suave de los totales poblacionales; más precisamente, $\theta = h(t_1, t_2, \ldots, t_k)$ donde las segundas derivadas parciales de h son continuas.
- Los tamaños de muestra son relativemente grandes: ya sea el número de *UPM* muestradas en cada estrato es grande, o el muestreo contiene un gran número de estratos. Además, para construir un I.C. usando una distribución normal, los tamaños de muestras deben ser lo suficientemente grandes para que la distribución muestral de $\hat{\theta}$ sea aproximadamente normal.

Pero, ¿Qué sucede en los casos en donde no tenemos los supuestos anteriores? Por ejemplo en la mediana, ya que F^{-1} y $\hat{F}^{-1}$ no son funciones suaves.

Si asumimos que las poblaciones y muestras son lo suficientemente grandes tal que estas cantidades pueden ser aproximadas por funciones continuas, tenemos algunos resultados clásicos en los métodos estudiados.

Los métodos de agrupación aleatoria funcionan bien, si el número de grupos aleatorios R es *moderado*. Sea $\hat{\theta}_q(r)$ el cuantil estimado para el grupo aleatorio r. Entonces, un I.C. para θ_q es:

$$\hat{\theta}_q \pm t \sqrt{\frac{1}{R}\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_q(r) - \theta_q)^2}$$

donde t es el cuantil apropiado de una distribución t con $R-1$ grados de libertad.

Un forma alternativa para construir los I.C. propuesto por Woodruff en 1952, basa su construcción en utilizar la función inversa de la distribución de probabilidad y su equivalente empírico. Primeramente, es claro notar que un intervalo de confianza del 95% para $F(y)$ está dado por:

$$\hat{F}(y) \pm 1.96\sqrt{\widehat{\mathbb{V}[\hat{F}(y)]}}$$

Luego, tras trabajando con la argumento de la probabilidad convenientemente, es posible establecer que un I.C. del 95% para el cuantil θ_q está dado por:

$$\left[\hat{F}^{-1}\{q - 1.96\sqrt{\widehat{\mathbb{V}[\hat{F}(y)]}}\}, \hat{F}^{-1}\{q + 1.96\sqrt{\widehat{\mathbb{V}[\hat{F}(y)]}}\right]$$

Hay bastantes detalles técnicos con este tipo de confección de intervalos de confianza, debido que F y $\hat{F}$ son funciones escalonadas; tienen saltos en los valores de y en la población y muestra.

A modo general, para que el método funcione, los saltos deben ser *pequeños* y la distribución muestral de $\hat{F}(y)$ debe ser aproximadamente normal.

Software estadístico para teoría de muestreo

En lo que sigue del curso con focalizaremos en la aplicación computacional de la teoría que hemos aprendido a lo largo de Muestreo I y II. Aplicando los diseños muestrales simples individualmente, hasta llegar a una configuración de muestreo complejo. Utilizaremos **R** como principal herramienta.

Muestreo Aleatorio Simple

Como lo hemos hecho antes, utilizaremos el paquete **survey** disponible en el CRAN.

```
library(survey)
data(api)
```

Primero cargamos la librería y los datos que utilizaremos. Los datos describen el desempeño de un conjunto de estudiantes en las escuelas del Estado de California, Estados Unidos. (*API, por academic performance index*)

Primero debemos describir el conjunto de dato a R, para ello utilizamos el el comando **svydesign**

```
srs_design<-svydesign(id=~1, fpc=~fpc, data=apisrs)
srs_design
```

```
## Independent Sampling design
## svydesign(id = ~1, fpc = ~fpc, data = apisrs)
```

El argumento **id=~1** indica que se tomaron muestras de escuelas individuales. El argumento **fpc=~fpc** indica que la variable **fpc** en el conjunto de datos contiene el tamaño poblacional. La notación ~ indica que la variable está contenida en el conjunto de datos especificado.

Las funciones **svymean() y svytotal()** estiman la media y total poblacional, respectivamente.

```
svytotal(~enroll, srs_design)
```

```
##           total     SE
## enroll 3621074 169520
```

```
svymean(~enroll, srs_design)
```

```
##           mean     SE
## enroll 584.61 27.368
```

En este caso, el factor de corrección tiene muy poco impacto en las estimaciones, y podría haber sido ignorado.

Si el tamaño de la población no se especifica, es necesario especificar las probabilidades de inclusión (o probabilidades muestrales) o los pesos muestrales (factores de expansión o ponderadores). La variable **pw** en el conjunto de datos hace referencia a los pesos muestrales $6194/200 = 30.97$

```
nofpc<- svydesign(id=~1, weights=~pw,data=apisrs)
nofpc
```

```
## Independent Sampling design (with replacement)
## svydesign(id = ~1, weights = ~pw, data = apisrs)
```

```
svytotal(~enroll, nofpc)
```

```
##           total     SE
## enroll 3621074 172325
```

```
svymean(~enroll, nofpc)
```

```
##           mean     SE
## enroll 584.61 27.821
```

Las funciones **svymean() y svytotal()** también pueden ser aplicados a variables categóricas (o *factores*). En este caso, se creará una tabla de estimaciones poblaciones para cada categoría del factor. La variable **stype** indica que tipo de escuela es la observación: *Elementary, Middle School o High School*.

```
svytotal(~stype, srs_design)
```

```
##          total     SE
## stypeE 4397.74 196.00
## stypeH  774.25 142.85
## stypeM 1022.01 160.33
```

Múltiples variables puede ser analizadas en la misma función **svymean() y svytotal()**, y contrastes puede ser obtenidos desde los resultados.

```
means<- svymean(~api00+api99, srs_design)
means
```

```
##          mean     SE
## api00 656.58 9.2497
## api99 624.68 9.5003
```

```
svycontrast(means, c(api00=1, api99=-1))
```

```
##          contrast     SE
## contrast     31.9 2.0905
```

```
srs_design<-update(srs_design, apidiff=api00-api99)
srs_design<-update(srs_design, apipct=apidiff/api99)
svymean(~apidiff+apipct, srs_design)
```

```
##               mean     SE
## apidiff 31.900000 2.0905
## apipct   0.056087 0.0041
```

En lo anterior, la función **svycontrast()** calcula la diferencia entre los medias de los años 1999 y 2000:

$$(1 - \text{media 2000}) + (-1 \times \text{media 1999})$$

Una notación alternativa es:

```
svycontrast(means, quote(api00-api99))
```

```
##           nlcon     SE
## contrast  31.9 2.0905
```

La función **update()** crea variables adicionales en el objeto del diseño muestral (srs_design). Notar que estas funciones entregan un nuevo objeto del diseño muestral que se asignar a alguna variable.

Muestreo Estratificado

Como ya sabemos, los **M.A.S.** no son generalmente utilizados a gran escala, pues otros diseños muestrales dan el mismo nivel de precisión a un menor costo. Para este diseño muestral, utilizaremos un muestreo estratificado aleatorio de 200 escuelas desde los datos **API**. La muestra es estratificada por tipo de escuela, con $n_E = 100, n_M = 50, n_H = 50$ por *Elementary Schools*, *Middle Schools* y *High Schools*.

```r
strat_design<-svydesign(id=~1, strata=~stype, fpc=~fpc, data=apistrat)
strat_design
```

```
## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~stype, fpc = ~fpc, data = apistrat)
```

```r
svytotal(~enroll, strat_design)
```

```
##           total     SE
## enroll 3687178 114642
```

```r
svymean(~enroll, strat_design)
```

```
##           mean     SE
## enroll 595.28 18.509
```

```r
svytotal(~stype, strat_design)
```

```
##        total SE
## stypeE  4421  0
## stypeH   755  0
## stypeM  1018  0
```

En lo anterior, el argumento **strata=~stype** especifica la variable estratificadora. En este caso, la variable **fpc** indica el tamaño poblacional **por estrato**, y no la población entera como su equivalente bajo un **M.A.S.**, esto es: 4421 *Elementary Schools*,1018 *Middle Schools* y 755 *High Schools*.

La estratificación ha reducido los desviaciones estándar significativemente. La estratificación por tipo de escuela es sólo posible si el tipo para cada escuela en la población es conocido de antemano, esta información extra de la población es la fuente del aumento en la precisión de las estimaciones.

No existe una ganancia en el desempeño académico en sí, que tiene una distribución similar en todas las escuelas. En el otro extremo, la estimación del número de escuelas para cada tipo, calculada por **svytotal(~stype, strat_design)**, no tiene dispersión, y **R** entrega desviación estándar 0, pues saber el tipo de cada escuela en la población implica saber el número de escuelas de cada tipo.

Pesos muestrales replicados

Las desviaciones estándar de la media u otras cantidades poblacionales, son por definición, la desviación estándar de aquellas cantidades poblaciones a lo largo de muchas muestras independientes de los datos. Como ya hemos estudiados, podemos realizar métodos de replicación o remuestreo para estimar las desviaciones estándar de cantidades poblacionales de interés, en lo que sigue mostraremos como crear los pesos muestrales replicados en R, para ello utilizamos la función **as.svrepdesign()**

```r
boot_design<-as.svrepdesign(strat_design, type="bootstrap",
                            replicates=100)
jk_design<-as.svrepdesign(strat_design)
boot_design
```

```
## Call: as.svrepdesign.default(strat_design, type = "bootstrap", repli
## Survey bootstrap with 100 replicates.
```

```r
jk_design
```

```
## Call: as.svrepdesign.default(strat_design)
## Stratified cluster jackknife (JKn) with 200 replicates.
```

```r
svymean(~enroll,boot_design)
```

```
##           mean     SE
## enroll 595.28 18.468
```

```r
svymean(~enroll,jk_design)
```

```
##           mean     SE
## enroll 595.28 18.509
```

Existen varias ventajas de los diseños con pesos muestrales replicados, la principal es poder determinar la estimación de varianza dentro de la misma especificación del diseño, además de poder calcular desviaciones estándar para diferencias entre distintas subpoblaciones para estadísticos arbitrarios. Comparación de medias y proporciones es posible para cualquier objeto de diseño muestral usando regresión.

Otras cantidades poblacionales

La meadiana es un ejemplo de un estadístico definido implícitamente en vez de explicitamente en término de medias o totales poblacionales. Las medianas y otras cantidades presentan algunas dificultades técnicas en la estimación. Incluso bajo un *M.A.S* la mediana no es única bajo un tamaño de muestra par: cualquier número entre las observaciones de al medio es una válida opción.
La función **svyquantile()** interpola linealmente entre las dos observaciones adyacentes cuando el cuantíl no es único.

```r
svyquantile(~api00, strat_design, c(0.25, 0.5, 0.75), ci=TRUE)
```

```
## $api00
##      quantile ci.2.5 ci.97.5       se
## 0.25      565    535     597 15.71945
## 0.5       668    642     694 13.18406
## 0.75      756    726     778 13.18406
##
## attr(,"hasci")
## [1] TRUE
## attr(,"class")
## [1] "newsvyquantile"
```

Estimaciones en subpoblaciones

Dado que cada estrato en un muestreo estratificado es un muestreo aleatorio simple separado, es fácil entregar estimaciones para las medias, totales y otros estadísticos para cada estrato. Para subpoblaciones que no son los estratos, la situación es algo más complicada. El paquete survey entrega un forma fácil de resolver este problema.

En los datos **API** la variable **emer** es el porcentaje de profesores con sólo certificación de enseñanza de emergencia, un indicador de la dificultad en contratar profesores en la escuela.

Alrededor de 20% de las escuelas no tienen profesores con una certificación de emergencia y alrededor del mismo número tiene más de 20% de sus profesores con certificación de emergencia. Podemos estimar la media en el *Academic Performance Index* y el número total de estudiantes en ambos subconjuntos.

```r
emerg_high<-subset(strat_design, emer>20)
emerg_low<-subset(strat_design, emer==0)
svymean(~api00+api99, emerg_high)
```

```
##          mean     SE
## api00 558.52 21.708
## api99 523.99 21.584
```

```r
svymean(~api00+api99, emerg_low)
```

```
##          mean     SE
## api00 749.09 17.516
## api99 720.07 19.061
```

```r
svytotal(~enroll,emerg_high)
```

```
##         total     SE
## enroll 762132 128674
```

```r
svytotal(~enroll,emerg_low)
```

```
##         total    SE
## enroll 461690 75813
```

Muestreo por conglomerado / multietápicos

Para el caso de muestreo por conglomerado en donde tenemos varios etapas (o niveles), la única diferencia que debemos hacer es en el argumento **id**, en donde este se reemplaza por la columna que identifica la unidad muestral, en los distintos niveles.

Por ejemplo, consideremos un muestreo por conglomerado bietápico (o a dos etapas) de la población de los datos **API**, en donde se tienen 40 distritos escolares, luego 5 escuelas dentro de cada distrito (o para todas las escuelas si es que hay menos de 5)

El diseño tiene un tamaño poblacional de 757 en la primera etapa, para el número de distritos escolares en California. En el segundo nivel, el muestreo de escuelas es **dentro** de cada distrito, por lo que el tamaño poblacional es el número de escuelas en aquel distrito. Cada argumento debe ser agregado al igual que al realizar modelos en R, en este caso, uno para cada etapa. Al igual que en el M.A.S., los pesos pueden no ser agregados si estos pueden ser obtenidos a partir de los tamaños poblacionales.

```
clus2_design<-svydesign(id=~dnum+snum, fpc=~fpc1+fpc2, data=apiclus2)
clus2_design

## 2 - level Cluster Sampling design
## With (40, 126) clusters.
## svydesign(id = ~dnum + snum, fpc = ~fpc1 + fpc2, data = apiclus2)
```

Estratos con una sola UPM

Cuando un estrato en la población tiene solo una potencial **UPM**, el fracción de muestreo para este estrato debe ser del 100%, sino sería 0%. El estrato puede no contribuir a la varianza en el primer nivel, pero puede contribuir a la varianza en las siguientes etapas del muestreo.

La mejor forma de tratar este tipo de estratos, es combinarlos con otro que sea lo más similar posible. En caso de que lo anterior no es posible o no fue realizado, el comportamiento del paquete **survey** es entregar un error, pero existen dos aproximaciones que pueden resolver la problemática. Estas son:

```
options(survey.lonely.psu = "adjust")
options(survey.lonely.psu = "average")
```

La primera entrega una estimación conservadora de la varianza que usa los residuos de la media poblacional en vez de los de la media del estrato. La segunda, asigna la contribución a la varianza como el promedio de todos los estratos con una **UPM** adicional. La opción **adjust** es conservadora, y la opción **average** tiene por objetivo los diseños donde **UPM** unitarias son el resultado de observaciones sin respuesta o cuando los estratos son relativamente comparables.

Otros tópicos en muestreo

Con lo que hemos visto en curso, ya hemos completado los tópicos elementales de la teoría clásica de muestreo, pero es posible seguir estudiando estos tópicos en detalle. En lo que sigue, mencionaremos tópicos relevantes dentro del área y referimos a libros específicos.

- Un revisión detallada de lo que hemos visto en esta sección, junto con tópicos relacionados a **gráficos** en muestreo complejo usando R, puede ser estudiada en: *Complex survey: A Guide to Analysis Using R, Thomas Lumley*.
- Para una revisión más exhaustiva y con detalles sobre los algoritmos utilizados en teoría de muestreo, implementados en el paquete *sampling* del CRAN, puede ser estudiada en: *Sampling algorithms, Yves Tillé*.
- Para una revisión de la teoría de muestreo con miras en aplicación en industria (Control de calidad), referimos a *Theory of Sampling and Sampling Practice, Francis F. Pitard*.
- Para una revisión exhaustiva de los tópicos que hemos visto a lo largo del curso -centrado en la teoría-, referimos a *Survey Sampling, Theory and Applications, Raghunath Arnab*.
- Para la teoría de muestreo de áreas, referimos a *Survey Sampling, Leslie Kish*.
- Para la teoría de muestreo aplicados a *Remote Sensing* y *GIS (Geographic Information Systems)*, referimos a *Sampling Methods, Remote Sensing and GIS Multiresource Forest Inventory, Kohl, Magnussen, Marchetti*.