Mis apuntes de muestreo

Robert Romero J.

DEFINICIONES PREVIAS

- Población objetivo: Todas las unidades de interés
- Muestra: Subconjunto de la población (aleatoria / No Aleatoria)
- Unidad de Muestreo: Objeto a ser seleccionado y q' permite el acceso a la información.
- Unidad de observación: Personas: → y_1: Sexo, y_2: Educación, y_3: NJE, ... y_k.
  Hogares:
  Viviendas:
- Operación Estadística: Hay 3

① Censos
② Registros Administrativos: Documentación auxiliar proveniente de una fuente (Ej: Registros de Clientes de un banco). Al tomar una muestra de estos se tiene un marco muestral.
③ Encuestas: Estimaciones de θ q' son funciones del total (t_y)
Se accesa a los elementos a través de un algoritmo de selección.
Cada uno de los elementos de la muestra son observados.

	y_1	y_2	...	y_k
1				
2				
⋮				
n				

Cada uno de los valores grabados de la variable son usados para estimar los parámetros.
④ Marcos: De listas / De áreas (cartográfico)

FICHA TECNICA: [DISEÑO MUESTRAL]

Resume la calidad de la encuesta.

Contiene:

① Población objetivo
② Muestra (características) ?
③ Unidad de muestreo
④ Unidad de observación
⑤ Variables de interés (Breve descripción)
⑥ Descripción del marco de muestreo
⑦ Parámetros de interés
⑧ Diseño de muestreo (si la encuesta es probabilística)
Si no es probabilística entonces la técnica de recolección de datos.
⑨ Tamaño de la muestra
⑩ Nivel de Confianza
⑪ Error de muestreo:
- Error Estandar: $\sqrt{V(\hat{\theta})}$
- Margen de Error: $z_{1-\alpha/2}\sqrt{V(\hat{\theta})}$
- Coeficiente de Variación: $\frac{\sqrt{V(\hat{\theta})}}{\theta}$ ⇒ No aplica para proporciones $p<0{,}3$; $p>0{,}7$

▶ POBLACION FINITA:

Un conjunto de n elementos $\{e_1, e_2, \dots, e_N\}$ donde N no es necesariamente conocido y además N es finito ($N < \infty$) y $U = \{1, 2, \dots, N\}$ es el conjunto de rótulos

▶ MUESTRA ALEATORIA SIN REEMPLAZO:

Subconjunto de la población, extraído mediante un mecanismo de selección estadístico.

$S \rightarrow$ variable aleatoria
$s \rightarrow$ realización de S

Ejemplo:

Muestreo aleatorio y sin reemplazamiento
U = {Cristiano, Messi, Falcao, Iniesta, Xavi}

Una muestra s de $n=2$ puede ser:

$s_1 = \{$Falcao, Xavi$\}$ ⇝ lo vamos a representar por un vector columna de 1's y 0's.

$$s_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}$$ ⇝ Esto es una realización de S.

Entonces una muestra aleatoria sin reemplazo se denota como:

$$S = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{pmatrix} \in \{0,1\}^{N \text{dimensional}}.$$

donde cada componente $S_k \begin{cases} 1 & \text{si } k \in s_i \\ 0 & \text{si } k \notin s_i \end{cases}$ > En muestreo sin reemplazo

↳ Quién sería S?:

Retomando nuestro ejemplo de $n=2$

$$S = \left[\underset{s_1}{\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}} \underset{s_2}{\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}} \underset{s_3}{\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}} \cdots\cdots \underset{s_N}{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}} \right] \Rightarrow S = \binom{5}{2} = 10$$

El conjunto S tiene 10 vectores ó 10 elementos y cada vector tiene 5 componentes

▶ TAMAÑO DE LA MUESTRA: Denotamos con $n(s)$ el tamaño muestral de una muestra en particular s, donde s es una realización de S.

$$n(s) = \sum_{k \in s} 1$$

entonces

$$s = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \Rightarrow n(s) = 2.$$

▶ MUESTREO ALEATORIO CON REEMPLAZAMIENTO: Identificado mediante el vector columna

$$s = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{N-1} \\ s_N \end{pmatrix} \in \mathbb{N}^N$$

donde s_k represent. el número de veces q' el elemento k se encuentra contenido en la muestra.

Ej: Continuando con los jugadores

con $m=3$

$$s_1 = \begin{pmatrix} 0 \\ 2 \\ 0 \\ 0 \\ 1 \end{pmatrix} \Rightarrow s_1 = \{ \text{messi, messi, xavi} \}$$

con $m=7$

$$s_2 = \begin{pmatrix} 2 \\ 3 \\ 0 \\ 2 \\ 0 \end{pmatrix} \Rightarrow s_2 = \underbrace{\{ \text{cristiano, cristiano, messi, messi, messi, iniesta, iniesta} \}}_{m = \text{tamaño de la muestra para muestreo con reemplazamiento}}$$

$m > n$ se puede dar, es decir, puede haber mayor muestra q' elementos incluidos en la misma.

En este caso el tamaño de muestra efectivo es $n=3$, en cada elemento seleccionado le aplico solo una vez la medición y la repito k veces según esté seleccionado en la muestra m.

▶ TIPOS DE MUESTREO

↳ Con reemplazo

↳ Sin reemplazo

▶ TAMAÑO DE LA MUESTRA

↳ Fijo

↳ Aleatorio

→ SOPORTE (Q):

Se trata de un conjunto de todas las muestras posibles y es lo q' en teoría de probabilidad conocemos como el Álgebra de Eventos.

Ej: $U = \{mov, pep, del\ bosque\}$, el soporte sería:

$$Q: \left\{ \begin{pmatrix}0\\0\\0\end{pmatrix} \begin{pmatrix}1\\0\\0\end{pmatrix} \begin{pmatrix}0\\1\\0\end{pmatrix} \begin{pmatrix}0\\0\\1\end{pmatrix} \begin{pmatrix}1\\1\\0\end{pmatrix} \begin{pmatrix}1\\0\\1\end{pmatrix} \begin{pmatrix}0\\1\\1\end{pmatrix} \begin{pmatrix}1\\1\\1\end{pmatrix} \right\}$$

⇒ en este soporte tenemos todas las posibles muestras sin reemplazo $n=0,\ n=1,\ n=2,\ n=3.$

Lexicográficamente, este soporte es:

$Q = \{\phi, (mov), (pep), (del\ bosque), (mov, pep), \ldots, (mov, pep, del\ bosque)\}$

Este soporte q' vimos se caracteriza por contener todas las combinaciones posibles entre elementos.

↳ Soporte Simétrico:

Se dice simétrico un soporte, si para cualquier $s \in Q$ todas las permutaciones de los elementos q' componen a s están contenidos en Q, es decir, todas las posibles muestras de tamaño n, están contenidas en el soporte. Hay varios tipos:

① Soporte simétrico sin reemplazo:

Se define como

$$Q = \left[\begin{pmatrix}S_{11}\\S_{21}\\\vdots\\S_{N1}\end{pmatrix} \begin{pmatrix}S_{12}\\S_{22}\\\vdots\\S_{N2}\end{pmatrix} \cdots \begin{pmatrix}S_{1j}\\S_{2j}\\\vdots\\S_{Nj}\end{pmatrix} \right] \in [0,1]^N$$

y su tamaño se calcula como $\boxed{\#Q = 2^N}$

Ej: $U = \{mov, pep, del\ bosque\}$, entonces $\#Q = 2^3 = 8$ (Verificar según ejemplo anterior)

② Soporte simétrico sin reemplazo de tamaño fijo: se define como:

$$S_n = \{ s \in S \mid \underbrace{\textstyle\sum_{k \in U} S_k = n}_{\text{que cumple}} \}$$

y su tamaño se define como $\#(S_n) = \binom{N}{n}$

Ejemplo:

$N=3$ y $n=2$

$$S_n = \left\{ \begin{pmatrix}1\\1\\0\end{pmatrix} \begin{pmatrix}1\\0\\1\end{pmatrix} \begin{pmatrix}0\\1\\1\end{pmatrix} \right\} \Rightarrow \#(S_n) = \binom{3}{2} = 3$$

$N=3$ $n=1$

$$S_n = \left\{ \begin{pmatrix}1\\0\\0\end{pmatrix} \begin{pmatrix}0\\1\\0\end{pmatrix} \begin{pmatrix}0\\0\\1\end{pmatrix} \right\}$$

$\#(S_n) = \binom{3}{1} = 3$

① Soporte simétrico con reemplazo: Se define por: $R = \mathbb{N}^N$ $\mathbb{N} \in$ Naturales.

donde $\#(R) = \infty$

② Soporte simétrico con reemplazo de tamaño fijo:

Se define por: $R_n = \{ s \in R \mid \sum_{k \in U} s_k = m \}$

y su tamaño se calcula como:

$$\#(R_n) = \binom{N+M-1}{M} = \binom{N+M-1}{N-1}$$

ya q' si T es el total de posibles soluciones no negativas ($x_i \geq 0$) de la ecuación $x_1 + x_2 + \cdots + x_r = n$, entonces

$$T = \binom{n+r-1}{r-1} = \binom{n+r-1}{n}$$

Ej:

Si $N=3$ → $N = \{e_1, e_2, e_3\}$

para $m=4$: $\{e_1, e_2, e_3, e_3\}$

Los cajones q' dividen son $N-1$!

e_1 e_2 e_3 e_3

$* \mid * \mid * \quad *$

1 2 3 4 5 6

$\Rightarrow \binom{m+N-1}{m} = \binom{6}{4} = 15$

para $m=3 \Rightarrow \{e_3, e_3, e_3\}$

e_1 e_2 e_3

$\emptyset \mid \emptyset \mid ***$

1 2 3 4 5

$\Rightarrow \binom{m+N-1}{m} = \binom{5}{3} = 10$

y viene del principio básico de conteo: De cuantas formas puedo organizar las estrellas en los cajones?

▸ RESUMEN

↳ Muestreo con reemplazo (El orden importa)

N^k ✗

↳ Muestreo con reemplazo (El orden no importa

$$\binom{m+N-1}{m} = \binom{M+N-1}{N-1}$$ ✓ USAMOS.

↳ Muestreo sin reemplazo (orden importa)

$$_nP_k = \frac{n!}{(n-k)!}$$ X

↳ Muestreo sin reemplazo (orden no importa)

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$ ✓ USAMOS.

Ejemplo:
Si $N=3$ $U=\{$mov, pep, delbosque$\}$ y $m=4$

$$\#(R_4) = \binom{4+3-1}{4} = \binom{6}{4} = \binom{6}{2} = 15$$

$S=\{$(mov, mov, mov, mov), (pep, pep, pep, pep), (db, db, db, db), (mov, mov, mov, pep), (mov, mov, mov, db), (mov, mov, pep, pep), (mov, mov, db, db), (mov, mov, pep, db), (mov, pep, pep, pep), (mov, db, db, db), (mov, pep, pep, db), (mov, pep, db, db), (pep, pep, pep, db), (pep, pep, db, db), (pep, db, db, db)$\}$

▶ DISEÑO DE MUESTREO: $p(\cdot)$

Teniendo un soporte de muestreo, es posible establecer la probabilidad de selección de una muestra dada s $[p(s)]$.

A la larga, esto quiere decir q' asumimos la existencia de una función $p(\cdot)$ q' asigna una probabilidad de selección $p(s)$ de una muestra s contenida en el soporte.

$$Q \xrightarrow{p(\cdot)} (0,1]$$
con $p(s)>0$ $\forall s \in Q$

Ejemplo:
Para $N=5$ $n=2$ ⇒ sin reemplazo ⇒ $\#Q=\binom{5}{2}=10$

	I_1	I_2	I_3	I_4	I_5	$p(s)$
1	1	1	0	0	0	0,13
2	1	0	1	0	0	0,2
3	1	0	0	1	0	0,15
4	1	0	0	0	1	0,1
5	0	1	1	0	0	0,15
6	0	1	0	1	0	0,04
7	0	1	0	0	1	0,02
8	0	0	1	1	0	0,06
9	0	0	1	0	1	0,07
10	0	0	0	1	1	0,08

- donde $\sum_{s\in Q} p(s) = 1$.
- y $p(s)>0$ para todo $s\in Q$

son características q' SIEMPRE se deben cumplir.

En muestreo la variable aleatoria son vectores y cada vector tiene una probabilidad conocida y asignada.

(*) Pg 26/27 del libro Estrategias de Muestreo (Andres Gutierrez) replicar los ejercicios de teaching sampling.

EJERCICIO

1. Considere $U = \{mou, pep, del\ bosque\}$ y un muestreo con reemplazamiento de tamaño $m=2$

   a. Escriba el soporte R de manera vectorial

$$R = \left\{ \begin{pmatrix}2\\0\\0\end{pmatrix} \begin{pmatrix}0\\2\\0\end{pmatrix} \begin{pmatrix}0\\0\\2\end{pmatrix} \begin{pmatrix}1\\1\\0\end{pmatrix} \begin{pmatrix}1\\0\\1\end{pmatrix} \begin{pmatrix}0\\1\\1\end{pmatrix} \right\}$$

   b. Escriba de manera lexicográfica el soporte.

$R = \{$(mou,mou), (pep,pep), (delbosque, del bosque), (mou,pep), (mou, del bosque), (pep, del bosque)$\}$

2. Considere $U = \{mou, pep, del\ bosque, pekerman, Bielsa, simeone\}$ y un muestreo sin reemplazamiento de $n=3$

   a. Escriba el soporte Q de manera vectorial

$$Q = \left\{ \begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix} \begin{pmatrix}1\\1\\0\\1\\0\\0\end{pmatrix} \begin{pmatrix}1\\1\\0\\0\\1\\0\end{pmatrix} \begin{pmatrix}1\\1\\0\\0\\0\\1\end{pmatrix} \begin{pmatrix}1\\0\\1\\1\\0\\0\end{pmatrix} \begin{pmatrix}1\\0\\1\\0\\1\\0\end{pmatrix} \begin{pmatrix}1\\0\\1\\0\\0\\1\end{pmatrix} \begin{pmatrix}1\\0\\0\\1\\1\\0\end{pmatrix} \begin{pmatrix}1\\0\\0\\1\\0\\1\end{pmatrix} \begin{pmatrix}1\\0\\0\\0\\1\\1\end{pmatrix} \begin{pmatrix}0\\1\\1\\1\\0\\0\end{pmatrix} \begin{pmatrix}0\\1\\1\\0\\1\\0\end{pmatrix} \begin{pmatrix}0\\1\\1\\0\\0\\1\end{pmatrix} \begin{pmatrix}0\\0\\1\\1\\1\\0\end{pmatrix} \begin{pmatrix}0\\0\\1\\1\\0\\1\end{pmatrix} \begin{pmatrix}0\\1\\0\\0\\1\\1\end{pmatrix} \begin{pmatrix}0\\0\\1\\0\\1\\1\end{pmatrix} \begin{pmatrix}0\\0\\0\\1\\1\\1\end{pmatrix} \begin{pmatrix}0\\1\\0\\1\\0\\1\end{pmatrix} \begin{pmatrix}0\\1\\0\\1\\1\\0\end{pmatrix} \right\}$$

   b. Escriba de manera lexicográfica el soporte

$Q = \{$(mou, pep, del bosque), (mou, pep, pekerman), (mou, pep, Bielsa), (mou, pep, simeone) (mou, delbosque, Pekerman), (mou, del bosque, Bielsa), (mou, del bosque, simeone), (mou, pekerman, Bielsa) (mou, pekerman, simeone) (mou, bielsa, simeone), (pep, del bosque, pekerman), (pep, del bosque, Bielsa) (pep, del bosque, simeones), (del bosque, pekerman, bielsa), (del bosque, pekerman, simeone), (pep, bielsa, simeone) (del bosque, Bielsa, simeone), (pekerman, Bielsa, simeone), (pep, pekerman, simeone), (pep, pekerman, biel$\}$

3. Considere una población de tamaño $N=10$ y con una muestra de tamaño 5 calcule la cardinalidad de los respectivos soportes para la muestra:

   - Con reemplazo

$$\#R = \binom{m+N-1}{m} \Rightarrow \begin{matrix} m=5 \\ N=10 \end{matrix} \Rightarrow \#R = \binom{5+10-1}{5} = \underline{\underline{2002}}$$

   - Sin reemplazo

$$\#Q = \binom{N}{n} = \binom{10}{5} = \underline{252}$$

▶ DISEÑO DE MUESTREO $p(\cdot)$:

Es una función f

$$Q \xrightarrow{f} [0,1]$$

Es una función de probabilidad multivariada sobre Q. Las propiedades del diseño de muestreo son:

i). $\sum_{s \in Q} p(s) = 1$

ii) $p(s) > 0$

si $p(s)=0$ no es un diseño de muestreo probabilístico para algún s

Según como sea el soporte tendremos:

- Muestreo con reemplazo
  - ↳ Tamaño fijo $n(s)=n$
  - ↳ De tamaño aleatorio
- Muestreo sin reemplazo
  - ↳ De tamaño fijo $n(s)=n$
  - ↳ De tamaño aleatorio

↳ Algoritmo de selección de la muestra: Es un mecanismo estadístico utilizado para seleccionar una muestra probabilística.

▶ PROBABILIDAD DE INCLUSIÓN: la inclusión es un evento aleatorio el cual se define por la función indicadora

$$I_k(s) = \begin{cases} 1 & \text{si } k \in s \\ 0 & \text{si } k \notin s \end{cases}$$

Donde la función $I_k(s)$ es una función de la variable aleatoria S.

Las probabilidades de inclusión pueden ser:

↳ De Primer Orden: (π_k) Esto se define como la probabilidad de que el k-ésimo elemento de la población $\{e_1, e_2, \dots, e_N\}$ pertenezca a la muestra aleatoria

$$\pi_k = P(k \in S) = P(I_k = 1) = \boxed{\sum_{s \in Q} I_k(s)p(s) = \sum_{k \ni s} p(s)}$$

los k que hay en S.

Ejemplo: (Sin reemplazo).
para $U = \{$Cristiano, Messi, Falcao, Iniesta, Xavi$\}$ $\rightarrow N=5$ con $n=2$

$\rightarrow$ El tamaño del soporte es $\#Q = \binom{5}{2} = 10$

id	I_1	I_2	I_3	I_4	I_5	lexicográficamente	$p(s)$	$I_1 p(s)$	$I_2 p(s)$	$I_3 p(s)$	$I_4 p(s)$	$I_5 p(s)$
1	1	1	0	0	0	(Cristiano, Messi)	0,11	0,11	0,11	0	0	0
2	1	0	1	0	0	(Cristiano, Falcao)	0,1	0,1	0	0,1	0	0
3	1	0	0	1	0	(Cristiano, Iniesta)	0,08	0,08	0	0	0,08	0
4	1	0	0	0	1	(Cristiano, Xavi)	0,2	0,2	0	0	0	0,2
5	0	1	1	0	0	(Messi, Falcao)	0,01	0	0,01	0,01	0	0
6	0	1	0	1	0	(Messi, Iniesta)	0,18	0	0,18	0	0,18	0
7	0	1	0	0	1	(Messi, Xavi)	0,12	0	0,12	0	0	0,12
8	0	0	1	1	0	(Falcao, Iniesta)	0,03	0	0	0,03	0,03	0
9	0	0	1	0	1	(Falcao, Xavi)	0,15	0	0	0,15	0	0,15
10	0	0	0	1	1	(Iniesta, Xavi)	0,02	0	0	0	0,02	0,02
							$\sum_s p(s) = 1$	0,49	0,42	0,29	0,31	0,49

$E(I) = (\pi_1 \quad \pi_2 \quad \pi_3 \quad \pi_4 \quad \pi_5)'$

Para Falcao: $I_3(s) = \begin{cases} 1 & \text{si Falcao} \in S \\ 0 & \text{si Falcao} \notin S \end{cases}$

$\hookrightarrow$ realización de una variable aleatoria S.

$$\pi_3(\omega) = \sum_{3 \ni s} p(s) = 0,29$$

En un diseño muestral sin reemplazo $\mu = \pi$, donde $\pi = (\pi_1, \pi_2, \dots, \pi_k)'$ es el vector de probabilidades de inclusión resultado del diseño de muestreo.

Entonces $\mu = E(S) = \sum_{s \in Q} p(s) s$

$\hookrightarrow$ Esperanza de la probabilidad de Inclusión de primer orden:

$$E(I_k) = \pi_k \rightsquigarrow E(I_k) = \sum_{k \in S} p(s) I_k(s) \Rightarrow E(I_k(S))$$

$$E(I_k) = (1)\, p(I_k = 1) + (0)\, p(I_k = 0)$$

$$E(I_k) = p(I_k = 1)$$

$$E(I_k) = p(I_k \in S) = \sum_{k \ni S} p(s)$$

$$\boxed{E(I_k) = \pi_k}$$

Scanned with CamScanner

→ Esperanza de la muestra aleatoria de primer orden:

$$E(S) = \sum_{s \in Q} \vec{s_i}\, p(s_i) = n(s) \qquad \text{Magnitud del vector???}$$

$$= \vec{S_1} p(s_1) + \vec{S_2} p(s_2) + \dots + \vec{S}_{n(q)}\, p(S_{n(q)})$$

La esperanza de la muestra es el vector de probabilidades de inclusión de primer orden.

$$E(S) = (\pi_1, \pi_2, \dots, \pi_N)'_{N \times 1}$$

▶ PROBABILIDAD DE INCLUSIÓN DE SEGUNDO ORDEN (π_{kl})

Acá vamos a determinar cuál es la probabilidad de inclusión de dos o más elementos a la vez en la muestra.

$$\pi_{kl} = P(k \in S \wedge l \in S)$$

$$= P(I_k(s)\, I_l(s))$$

$$= \sum_{s \ni kl} p(s) \longrightarrow E(I_k(s)\, I_l(s)) = \pi_{kl}$$

$$= \sum_{s \in Q} I_k(s)\, I_l(s)\, p(s)$$

Ej: En el soporte anterior

$$\pi_{\text{moisés, ernesto}} = \pi_{2,4} = 0{,}18$$

▶ MATRIZ DE PROBABILIDADES DE INCLUSIÓN:

$$\begin{pmatrix} \pi_{11} & \pi_{12} & \pi_{13} & \cdots & \pi_{1N} \\ \pi_{21} & \pi_{22} & \pi_{23} & \cdots & \pi_{2N} \\ \pi_{31} & \pi_{32} & \pi_{33} & \cdots & \pi_{3N} \\ \vdots & \vdots & & & \\ \pi_{N1} & \pi_{N2} & \cdots & \cdots & \pi_{NN} \end{pmatrix}$$

Es una matriz simétrica con las probabilidades de inclusión de primer orden en su trazo y las probabilidades de inclusión de segundo orden en el resto de sus posiciones.

El número de las probabilidades de inclusión se calculan como:

$$\overbrace{\frac{(N^2-N)}{2}}^{\text{2do orden}} + \overbrace{N}^{\text{1er orden}} = \frac{N^2-N}{2} + \frac{2N}{2} = \boxed{\frac{N(N+1)}{2}}$$

LA ESTADISTICA I_k.

Esta estadística toma valores aleatorios según si el diseño de muestreo es con reemplazamiento o sin reemplazamiento.

Propiedades.

- $\boxed{E(I_k) = \pi_k}$
- $V(I_k) = \pi_k(1-\pi_k)$; demostración.

$$V(I_k(S)) = E[I_k(S) - E(I_k(S))]^2$$
$$= E(I_k(S)^2) - [E(I_k(S))]^2$$

Como:

$$I_k = \begin{cases} 1 & si\ k \in S \\ 0 & si\ k \notin S \end{cases}$$

$$I_k^2 = \begin{cases} 1^2 & si\ k \in S \\ 0^2 & si\ k \notin S \end{cases}$$

entonces

$$I_k^2 = I_k \text{ , por lo cual.}$$

$$V(I_k(S)) = E(I_k) - [E(I_k)]^2$$
$$V(I_k) = \boxed{= \pi_k - \pi_k^2 = \pi_k(1-\pi_k)}$$

- $cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ → ojo, usualmente $\pi_{kl} \neq \pi_k\pi_l$

Demostración

$$Cov(I_k, I_l) = E(I_k . I_l) - E(I_k)E(I_l)$$
$$= (1)\ P(I_k I_l = 1) + (0)\ \cancel{P(I_k I_l = 0)} - \pi_k\pi_l$$
$$\boxed{cov(I_k, I_l) = \pi_{kl} - \pi_k\pi_l} = \underline{\Delta_{kl}}$$

PARAMETROS POBLACIONALES:

Un parámetro poblacional es una función de los valores correspondientes a todos los elementos de la población

$$f(y_1, y_2, \ldots, y_N) = \theta$$

donde y es la variable de interés.

Los parámetros de interés q' nos interesan son:

- Total Poblacional (t_y):

$$t_y = \sum_{k \in U} y_k$$

- Promedio Poblacional:

$$\bar{y}_U = \sum_{k \in U} \frac{y_k}{N} = \frac{t_y}{N}$$

- Varianza Poblacional:

$$S_y^2 = \frac{\sum_U (y_k - \bar{y}_U)^2}{N-1}$$

▶ DOMINIOS:

Un dominio es una subpoblación denotada U_d, tal que:

i). $U_{univ} = \bigcup_{d=1}^{D} U_d$ → un dominio es una partición del universo, entonces el universo se conforma de la unión de los dominios contenidos en el.

unión

↳ Esto quiere decir q' no pueden existir elementos q' pertenezcan a más de un dominio.

ii) Si $k \in U_i$, entonces $k \notin U_q$, donde U_q y U_i son dominios mutuamente excluyentes y $U_q \cup U_i = U_{univ}$ y $U_q \cap U_i = \phi$.

iii) El número de elementos pertenecientes a un dominio U_d es N_d, con lo cual llamamos a N_d = Tamaño del Dominio.

iv) $P_d = \frac{N_d}{N}$ ⇒ Tamaño Relativo del dominio (P_d).

↳ Parámetros de Interés de los dominios:

- Total de un dominio (t_{yd}):

$$t_{yd} = \sum_U y_{dk} = \sum_U y_k z_{dk} \quad \text{donde} \quad z_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{si } k \notin U_d \end{cases}$$

- Tamaño del dominio (N_d)

$$N_d = \sum_{U_d} z_{dk}$$

- Promedio de la característica de interés en el dominio

$$\bar{y}_{U_d} = \frac{t_{yd}}{N_d} = \frac{\sum_U y_{dk}}{N_d}$$

▶ ESTADÍSTICA $n(s)$: TAMAÑO DE LA MUESTRA :

Es una cantidad aleatoria q' puede ser expresada en función de una estadística de inclusión:

$$n(S) = \sum_U I_k(S)$$

↳ Propiedades de Inferencia:

- $E(n(S)) = \sum_U \pi_k$

ya que:

$$E[n(S)] = E\left[\sum_U I_k\right] = \sum_U \underbrace{E(I_k)}_{\pi_k} = \sum_U \pi_k$$

- $V[n(S)] = \sum_U \pi_k - \left(\sum_U \pi_k\right)^2 + \sum\sum_{k \neq l} \pi_{kl}$

ya que:

$$V[n(S)] = V\left[\sum_U I_k\right]$$ → la varianza de una suma es igual a la suma de las varianzas más su covarianza

$$= \sum_U \underbrace{V(I_k)}_{\pi_k(1-\pi_k)} + \sum\sum_{k\neq l} Cov(I_k, I_l)$$

$$= \sum_U \pi_k - \sum_U \pi_k^2 + \sum\sum_{k\neq l} \underbrace{cov(I_k, I_l)}_{\pi_{kl} - \pi_k\pi_l}$$

$$= \sum_U \pi_k - \left[\sum_U \pi_k\right]^2 + \sum\sum_{k\neq l} \pi_{kl} - \underbrace{\sum\sum_{k\neq l} \pi_k\pi_l}_{0}$$

$$V[n(S)] = \sum_U \pi_k - \left[\sum_U \pi_k\right]^2 + \sum\sum_{k\neq l} \pi_{kl}$$

—//— RECORDANDO —//—

$$VAR(x_i + x_j) = V(x_i) + V(x_j) + cov(x_i, x_j) + cov(x_j, x_i)$$

$$= V(x_i) + V(x_j) + 2\,cov(x_i, x_j)$$, utilizando la matriz de varianzas y covarianzas:

$$\begin{array}{c|c} & \begin{array}{ccccc} 1 & 2 & 3 & \cdots & p \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ p \end{array} & \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & & \cdots & \sigma_{2p} \\ \sigma_{31} & & \sigma_3^2 & & \\ \vdots & & & \ddots & \\ \sigma_{p1} & & & & \sigma_{pp} \end{pmatrix} \end{array} \Rightarrow V(x_i + x_j) = \sum V(x_i) + \sum\sum_{i\neq j} Cov(x_i, x_j)$$

↳ Cuando trabajamos un diseño de muestreo fijo, tenemos estas propiedades adicionales:

- $E[n(S)] = n$

$$E[n(S)] = E(n) = n \rightarrow$$ la esperanza de una constante es ella misma.

- $\sum_{k \in U} \pi_{kl} = n\pi_l$

$$\sum_{k \in U} \pi_{kl} = \sum_{k \in U} E[I_k(S) I_l(S)]$$
$$= \sum_{k \in U} \left[\sum_{S \in Q} I_k(S) I_l(S) p(S)\right]$$
$$= \sum_{S \in Q} \left[\sum_{k \in U} I_k(S) I_l(S) p(S)\right]$$
$$= \sum_{S \in Q} I_l(S) p(S) \underbrace{\sum_{k \in U} I_k(S)}_{n}$$
$$= n \underbrace{\sum_{S \in Q} I_l(S) p(S)}_{\pi_l}$$

$$\boxed{\sum_{k \in U} \pi_{kl} = n\pi_l}$$

- $\sum_{l \in U} \pi_{kl} = n\pi_k \Rightarrow$

$$\sum_{l \in U} \pi_{kl} = \sum_{l \in U} E(I_k(S) I_l(S))$$
$$= \sum_{l \in U} \left[\sum_{S \in Q} I_k(S) I_l(S) p(S)\right]$$
$$= \sum_{S \in Q} \left[\sum_{l \in U} I_k(S) I_l(S) p(S)\right]$$
$$= \sum_{S \in Q} I_k(S) p(S) \underbrace{\sum_{l \in U} I_l(S)}_{n}$$
$$= n \underbrace{\sum_{S \in Q} I_k(S) p(S)}_{\pi_k}$$

$$\boxed{\sum_{l \in U} \pi_{kl} = n\pi_k}$$

- $\sum_{k \in U} \Delta_{kl} = 0 \Rightarrow$

$$\sum_{k \in U} \pi_{kl} - \pi_k \pi_l = 0$$
$$\sum_{k \in U} \pi_{kl} - \pi_l \underbrace{\sum_{k \in U} \pi_k}_{n} = 0$$
$$\underbrace{\sum_{k \in U} \pi_{kl}}_{n\pi_l} - n\pi_l = 0$$ → (ver propiedad anterior)

$$\boxed{n\pi_l - n\pi_l = 0}$$

$$\sum_{l\in U} \Delta_{kl} = 0 \Rightarrow \sum_{l\in U} \pi_{kl} - \pi_k\pi_l = 0$$

$$\underbrace{\sum_{l\in U} \pi_{kl}}_{\text{con } \pi_k} - \pi_k \underbrace{\sum_{l} \pi_l}_{\text{con}} = 0$$

$$n\pi_k - n\pi_k = 0$$

- $\pi_k(1-\pi_k) = \sum_{k\neq l} \pi_k\pi_l - \pi_{kl} = -\sum_{k\neq l} cov(I_k, I_l)$

Para demostrar esta propiedad, es necesario redefinir el tamaño de la muestra

si $n = \sum_U I_k \Rightarrow n = I_l + \sum_{k\neq l} I_k$

si $\pi_k(1-\pi_k) \Rightarrow VAR(I_k) \Rightarrow cov(I_k, I_k) \Rightarrow \pi_k(1-\pi_k) = cov\left(I_k, n - \sum_{k\neq l} I_k\right)$

ya q' $n = I_l + \sum_{k\neq l} I_k \Rightarrow I_l = n - \sum_{k\neq l} I_k$

entonces:

Recordando q':

$$Cov\left(x_k, \sum_{i\neq k} x_i\right) = cov(x_k, x_1 x_2 + \dots + x_{k-1} + x_k + \dots)$$
$$= \sum cov(x_k, x_i)$$

y recordando q':

$$cov\left(x, a - \sum y_i\right) = \sum cov(x, -y_i)$$
$$= -\sum cov(x, y_i)$$

tendriamos q'

$$cov\left(I_k, n - \sum_{k\neq l} I_k\right) = -\sum_{k\neq l} cov(I_k, I_l)$$
$$= -\sum_{k\neq l} \pi_{kl} - \pi_k\pi_l$$
$$= \sum_{k\neq l} \pi_k\pi_l - \pi_{kl}$$

- ESTIMADOR DE HORVITZ-THOMPSON: Se utiliza para calcular totales de población aplicando un factor de expansión al total muestral.

$$\hat{t}_{y\pi} = \sum_S \frac{y_k}{\pi_k} \Rightarrow \boxed{\hat{t}_{y\pi} = \sum_S d_k y_k} \quad \text{con } d_k = \frac{1}{\pi_k}$$

Teniendo en cuenta que $\pi_k = \sum_S p(s)$ y $p(s) = \frac{n}{N}$ entonces

$$\hat{t}_{y\pi} = \sum_S \sum_S \frac{N}{n} y_k \Rightarrow \hat{t}_{yk} = \frac{N}{n} \sum_S y_k \Rightarrow \boxed{\hat{t}_{yk} = N\bar{y}_k}$$

↳ Propiedades del Estimador Horvitz-Thompson:

• Esperanza del estimador: Teniendo en cuenta q'

$$\hat{t}_{yk} = \sum_{s} \frac{y_k}{\pi_k} = \sum_{U} I_k \frac{y_k}{\pi_k}$$

entonces

$$E(\hat{t}_{yk}) = E\left(\sum_{U} I_k \frac{y_k}{\pi_k}\right) = \sum_{U} \frac{y_k}{\pi_k} E(I_k) = \sum_{U} \frac{y_k}{\cancel{\pi_k}} \cancel{\pi_k}$$

$$\boxed{E(\hat{t}_{yk}) = \sum_{U} y_k} = t_y$$

• Varianza del estimador:

$$V(\hat{t}_{yk}) = V\left(\sum_{U} I_k \frac{y_k}{\pi_k}\right)$$

$$= \underline{\sum_{U} \frac{y_k^2}{\pi_k^2} V(I_k)} + \sum_{k \neq l}\sum \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \operatorname{cov}(I_k, I_l)$$

$: \sum_{U} \frac{y_k^2}{\pi_k^2} \operatorname{cov}(I_k, I_k)$

$: \sum_{k=l}\sum \frac{y_k}{\pi_k} \frac{y_k}{\pi_k} \Delta_{kl}$

$$= \sum_{U} \frac{y_k^2}{\pi_k^2}\left[\pi_k(1-\pi_k)\right] + \sum_{k \neq l}\sum \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\left[\pi_{kl} - \pi_k \pi_l\right]$$

$$= \sum_{U} \frac{y_k^2}{\pi_k^2}\left[\pi_k - \pi_k^2\right] + \sum_{k \neq l}\sum \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}$$

$$= \sum_{U}\left[\frac{y_k^2 \cancel{\pi_k}}{\cancel{\pi_k^2}} - \frac{y_k^2 \cancel{\pi_k^2}}{\cancel{\pi_k^2}}\right] + \sum_{k \neq l}\sum \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}$$

?? $\Leftarrow$ $= \sum_{U}\left[\frac{y_k^2}{\pi_k} - y_k^2\right] + \sum_{k \neq l}\sum \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \Delta_{kl}$

$$V_1(\hat{t}_{yk}) = \boxed{= \sum_{U}\sum \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}} \quad \text{ó} \quad V_1(\hat{t}_{y\pi}) = \sum_{U}\sum \left(\frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l - \left(\sum_{U} y_k\right)^2\right)$$

• Si el diseño $p(\cdot)$ es de tamaño fijo, entonces la varianza del estimador H-T se escribe.

$$V_2(\hat{t}_{yk}) = -\frac{1}{2} \sum_{U}\sum \Delta_{kl}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2$$

↳ Estimador Alternativo de Sen-Yates-Grundy.

Nombre: Robert Romero   Código: 2113124   ESTADISTICA-MUESTREO

Ejercicio: Para $N=5$ y $n=3$ calcular el total de la población mediante el estimador Horvitz Thompson

$l_i \rightarrow e_i$

$U = \{$ Messi, Beckham, Cristiano, Eto'o, Rooney $\}$

	Messi	Beckham	Cristiano	Eto'o	Rooney
$y =$	33	31,5	29,2	23,3	20,6
$\pi_k =$	0,68	0,58	0,52	0,78	0,44
$d_k =$	1,471	1,724	1,923	1,282	1,273

$\Rightarrow y_k = 137,6$

$V(\hat{t}_{y\pi}) = \sum_{i=1}^{10} p(s)\left[\hat{t}_{y\pi} - E(\hat{t}_{y\pi})\right]^2$

SOPORTE:

S_i	I_1	I_2	I_3	I_4	I_5	lexicográficamente	$p(s)$	$\hat{t}_{y\pi} = \sum_U I_k(s) d_k y_k$	$E(\hat{t}_{y\pi}) = \sum_{i=1}^{10} \hat{t}_{y\pi} p(s)$	$V(\hat{t}_{y\pi})$
S_1	1	1	1	0	0	(Messi, Beckham, Cristiano)	0,11	158,99	17,49	50,51
S_2	1	1	0	1	0	(Messi, Beckham, Eto'o)	0,1	132,71	13,27	2,35
S_3	1	1	0	0	1	(Messi, Beckham, Rooney)	0,08	149,66	11,97	1,70
S_4	1	0	1	1	0	(Messi, Cristiano, Eto'o)	0,2	134,55	26,91	1,81
S_5	1	0	1	0	1	(Messi, Cristiano, Rooney)	0,01	151,50	1,51	1,94
S_6	1	0	0	1	1	(Messi, Eto'o, Rooney)	0,18	125,22	22,54	27,43
S_7	0	1	0	1	1	(Beckham, Eto'o, Rooney)	0,12	140,33	16,84	0,92
S_8	0	0	1	1	1	(Cristiano, Eto'o, Rooney)	0,03	157,28	4,72	11,66
S_9	0	1	1	1	0	(Beckham, Cristiano, Eto'o)	0,15	131,00	19,65	6,46
S_{10}	0	1	1	0	1	(Beckham, Cristiano, Rooney)	0,02	132,84	2,66	0,45

$\boxed{\sum_{i=1}^{10} p(s) = 1}$

$\pi_k = \sum_{s \ni k} I_k(s) p(s)$

$\pi_1 = 0,68$

$\pi_2 = 0,58$

$\pi_3 = 0,52$

$\pi_4 = 0,78$

$\pi_5 = 0,44$

$\boxed{\sum_U \pi_k = n = 3}$

$\boxed{E(\hat{t}_{y\pi}) = 137,56}$

$\boxed{V(\hat{t}_{y\pi}) = 117,25}$

→ Estimación de la Varianza del Estimador Horvitz-Thompson:

Si todas las $\pi_{kl} > 0$ entonces los estimadores insesgados de $V_1(\hat{t}_{y\pi})$ y $V_2(\hat{t}_{y\pi})$ son los siguientes:

$$\hat{V}_1(\hat{t}_{y\pi}) = \sum_S\sum \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

y

$$\hat{V}_2(\hat{t}_{y\pi}) = \sum_S\sum \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \left(-\frac{1}{2}\right)$$

la estimación se hace a través de un total expandido y la varianza se va a calcular en la muestra sobre la varianza expandida.

Vamos a ver q' el estimador de la varianza Horvitz Thompson es insesgado para esto:

$$\hat{V}_1(\hat{t}_{y\pi}) = \sum_S\sum \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$V_1(\hat{t}_{y\pi}) = \sum_U\sum I_k I_l \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \quad \text{con} \quad I_k I_l = \begin{cases} 1 & \text{si } k \wedge l \in S \\ 0 & \text{en otro caso.} \end{cases}$$

entonces el valor esperado de $V_1(\hat{t}_{y\pi})$:

$$E(V_1(\hat{t}_{y\pi})) = E\left(\sum_U\sum I_k I_l \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\right)$$

$$= \sum_U\sum E(I_k I_l) \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$= (1)\sum_U\sum P(I_k=1 \wedge I_l=1) \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} + \underbrace{(0)\sum_U\sum P(I_k=0 \wedge I_l=0) \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}}_{0}$$

$$= \sum_U\sum P(I_k=1 \wedge I_l=1) \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$= \sum_U\sum \cancel{\pi_{kl}} \frac{\Delta_{kl}}{\cancel{\pi_{kl}}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$\boxed{E[V_1(\hat{t}_{y\pi})] = \sum\sum \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}}$$

y análogamente se hace para $V_2(\hat{t}_{y\pi})$

Bautista (1998) resalta que:

- Si $\pi_{kl} > 0 \; \forall k \neq l \in S$ pero si $\pi_{kl} = 0 \; \forall k \neq l \notin S$, entonces no se puede garantizar el insesgamiento de las anteriores expresiones.
- Es posible q' $\hat{V}_1(\hat{t}_{y\pi}) < 0 \vee \hat{V}_2(\hat{t}_{y\pi}) < 0$, resultados q' no pueden ser interpretados o usados. Para evitar esto es necesario q' $\Delta_{kl} < 0 \; \forall k \neq l$
- No necesariamente las estimaciones arrojadas por las anteriores expresiones coinciden en todos los casos.

↳ $\widehat{CVE}$:

$$\widehat{CVE} = \frac{\sqrt{\hat{V}(t_{\hat{y}\pi})}}{t_{\hat{y}\pi}} * 100 \quad \text{y con lo estimo} \quad CV = \frac{\sqrt{V(t_{\hat{y}\pi})}}{t} * 100$$

↳ Estimador alternativo para la varianza del Total Poblacional Estimado

$$V_2(t_{\hat{y}\pi}) = -\frac{1}{2} \sum_U \sum \Delta_{kl} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2 \rightsquigarrow \text{Solo para diseños de tamaño fijo.}$$

$$\hat{V}_2(t_{\hat{y}\pi}) = -\frac{1}{2} \sum_s \sum \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l} \right)^2$$

▶ ESTIMACION DE OTROS PARÁMETROS:

↳ Estimación del Tamaño de la Población:

En general cuando nos emprendemos a una investigación por muestreo, se desconoce el total de la población. En este caso podemos utilizar el estimador de Horvitz-Thompson para realizar dicha estimación, puesto q' N puede ser escrito como:

$$N = \sum_U 1 \quad ó \quad N = \sum_U y_k^* \quad \text{con } y_k^* = \begin{cases} 1 & \text{para todo } k \in U \\ 0 & \text{en otro caso} \end{cases}$$

por lo cual

$$\boxed{\hat{N}_\pi = \sum_s \frac{y_k^*}{\pi_k} \quad ó \quad \hat{N}_\pi = \sum_s \frac{1}{\pi_k}}$$

donde $E(\hat{N}_\pi) = N$

$$V(\hat{N}_\pi) = \sum_U \sum \Delta_{kl} \frac{1}{\pi_k} \frac{1}{\pi_l}$$

$$\hat{V}(\hat{N}_\pi) = \sum_s \sum \frac{\Delta_{kl}}{\pi_{kl}} \frac{1}{\pi_k} \frac{1}{\pi_l}$$

③ Tomar todas las muestras generadas de un espacio con $N=7$ y $n=4$ y estime $\hat{N}$ para cada una de las 35 muestras.

↳ Estimación de un Promedio con N Conocido:

En el espacio poblacional: $\bar{y}_U = \frac{1}{N} \sum_U y_k \Rightarrow \bar{y}_U = \frac{1}{N} t_y$

Entonces:

En el espacio muestral:

$$\boxed{\hat{\bar{y}}_\pi = \frac{1}{N} \sum_s \frac{y_k}{\pi_k}}$$

↳ Estimación de un Promedio con N Desconocido:

$$\hat{\bar{Y}}_{Hk} = \frac{1}{\hat{N}_\pi} \sum_s \frac{y_k}{\pi_k} \Rightarrow \hat{\bar{Y}}_{Hk} = \frac{\sum_s y_k/\pi_k}{\sum_s 1/\pi_k} \rightsquigarrow \text{Estimador de Hájek}$$

↳ Varianza del Promedio Muestral Estimado con N conocido:

$$\hat{\bar{y}}_{\pi} = \frac{1}{N} \sum_{s} \frac{y_k}{\pi_k}$$

$$V(\hat{\bar{y}}_{\pi}) = V\left(\frac{1}{N} \sum_{s} \frac{y_k}{\pi_k}\right)$$

$$V(\hat{\bar{y}}_{\pi}) = \frac{1}{N^2} V\left(\sum_{s} \frac{y_k}{\pi_k}\right)$$

$$\boxed{V(\hat{\bar{y}}_{\pi}) = \frac{1}{N^2}\left[\sum\sum_{U} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\right]}$$

↳ Varianza Estimada del Promedio Muestral Estimado con N conocido.

$$\hat{V}(\hat{\bar{y}}_{\pi}) = \frac{1}{N^2}\left[\sum\sum_{U} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}\right]$$

↳ Estimación de una Proporción con N conocido:

En el espacio poblacional entonces: $p = \frac{1}{N} \sum_{U} z_k$ con $z_k = \begin{cases} 1 & \text{si tiene la característica} \\ 0 & \text{en otro caso} \end{cases}$

En el espacio muestral

$$\hat{p} = \frac{1}{N} \sum_{s} \frac{z_k}{\pi_k}$$

donde
- $E(\hat{p}) = p$
- $V(\hat{p}) = V\left(\frac{1}{N} \sum_{s} \frac{z_k}{\pi_k}\right)$

$$V(\hat{p}) = \frac{1}{N^2} V\left(\sum_{s} \frac{z_k}{\pi_k}\right)$$

$$V(\hat{p}) = \frac{1}{N^2}\left[\sum\sum_{U} \Delta_{kl} \frac{z_k}{\pi_k} \frac{z_l}{\pi_l}\right]$$

- $\hat{V}(\hat{p}) = \frac{1}{N^2}\left[\sum\sum_{U} \frac{\Delta_{kl}}{\pi_{kl}} \frac{z_k}{\pi_k} \frac{z_l}{\pi_l}\right]$

↳ Intervalos de Confianza:

$$t \in \hat{t}_{y\pi} \pm \sqrt{\hat{V}(\hat{t}_{y\pi})} \cdot t_{v,1-\alpha/2}$$

Como es engorroso encontrar los v grados de libertad de t, por lo cual se aproxima a z. Por lo cual:

$$t \in \hat{t}_{y\pi} \pm z_{1-\alpha/2}\sqrt{\hat{V}(\hat{t}_{y\pi})}$$

Nótese q' si saco $\hat{t}_{y\pi}$ factor común, entonces.

$$t \in \hat{t}_{y\pi}\left[1 \pm z_{1-\alpha/2}\, CVE(\hat{t}_{y\pi})\right]$$

Así mismo:

$$t \in \hat{t}_{y\pi} \pm \underbrace{z_{1-\alpha/2}\underbrace{\sqrt{\hat{V}(\hat{t}_{y\pi})}}_{\text{Error Estándar de } \hat{t}_{y\pi}}}_{\text{Márgen de Error.}}$$

▶ Muestreo con Reemplazamiento

Una muestra con reemplazo se denota como:

$$\vec{S} = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} \in \mathbb{N}^{N+} \quad \text{con} \quad \#Q = \binom{M+N-1}{m} = \binom{m+N-1}{N-1}$$

$\binom{N+m-1}{m}$

De manera general un diseño de muestreo con reemplazo se define como:

$$p(s) = \begin{cases} \dfrac{m!}{n_1(s)!\,n_2(s)!\cdots n_N(s)!}\, p_1^{n_1(s)}\, p_2^{n_2(s)} \cdots p_N^{n_N(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso} \end{cases}$$

Ejemplo: Si tenemos un experimento con m ensayos independientes ~~Si tenemos m experimentos independientes~~, con posibles resultados $e_1, e_2, \dots, e_N$, entonces.

1	Pr	2	Pr	3	Pr	.....	m	Pr
e_1	P_1	e_1	P_1	e_1	P_1		e_1	P_1
e_2	P_2	e_2	P_2	e_2	P_2		e_2	P_2
e_3	P_3	e_3	P_3	e_3	P_3		e_3	P_3
e_4	P_4	e_4	P_4	e_4	P_4		e_4	P_4
e_5	P_k	e_k	P_k	e_5	P_k		e_5	P_k
$\vdots$	$\vdots$	e_i	P_i	$\vdots$	$\vdots$		$\vdots$	$\vdots$
e_N	P_N	e_N	P_N	e_N	P_N		e_N	P_N
	$\sum_U p(k) = 1$		1		1			1

La probabilidad de inclusión del elemento $e_k = P_k$, es decir la probabilidad de no inclusión del e_k $= 1 - P_k$

Probabilidad de selección del elemento k

Y a cada uno de estos posibles resultados en cada experimento se tiene asociada su respectiva probabilidad p_k

$1 - P_k - P_l =$ la Probabilidad de q' k y l no estén en la muestra.

↳ Probabilidad de Inclusión de Primer Orden

$$\pi_k = 1-(1-p_k)^m$$

Ya q' cada evento independiente tiene asociada una probabilidad de éxito "p_k" cuando $k \in S$, entonces cada uno de estos sorteos aleatorios está determinado por una probabilidad tipo Bernoulli. Por lo tanto, cuando se realizan m ensayos independientes, se utiliza la distribución de probabilidad binomial para hallar las probabilidades de inclusión de primer orden.

$$\pi_k = P(k \in S) = 1 - P(k \notin S)$$

$$= 1 - \cancelto{1}{\binom{m}{m}}(1-p_k)^m \cancelto{1}{(p_k)^{m-m}}$$

$$\boxed{\pi_k = 1-(1-p_k)^m}$$

↳ Probabilidad de Inclusión de Segundo Orden:

$$\pi_{kl} = P(k\in S \wedge l \in S)$$

Recordando las leyes de Morgan donde $A\cap B = (A^c \cup B^c)^c$ entonces tenemos que:

$$\pi_{kl} = 1 - P(k\notin S \cup l \notin S)$$

$$\pi_{kl} = 1 - [P(k\notin S) + P(l\notin S) - P(k\notin S \cap l\notin S)]$$

$$P(k\notin S) = (1-p_k)^m$$
$$P(l\notin S) = (1-p_l)^m$$
$$P(k\notin S\cap l\notin S) = (1-p_k-p_l)^m$$

Por lo cual

$$\boxed{\pi_{kl} = 1-(1-p_k)^m - (1-p_l)^m + (1-p_k-p_l)^m}$$

Ejercicio de Särndal

▸ RECORDATORIO

Sea $x_1, x_2, \dots, x_N$ una m.a iid, con esto tenemos q'

$$E(x_i) = \mu$$
$$V(x_i) = \sigma^2$$
$$COV(x_i, x_j) = 0$$

Entonces

i) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ Insesgado para μ.

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

iii) $\hat{V}(x_i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$

lo llamamos con σ^2

iv). Estimador para la varianza de $\bar{x}$

$$\hat{V}(\bar{x}) = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}(x_i-\bar{x})^2$$

... entonces retomando

Si m es una v.a $\rightarrow z_k = \frac{y_{ki}}{p_{ki}} \in U$ con $i = 1, 2, \ldots, m$.

entonces se puede definir un espacio de experimentos:

1			2			3			...	M		
e_i	z_i	P_r	e_i	z_i	P_r	e_i	z_i	P_r		e_i	z_i	P_r
e_1	$y_{11}/p_{11} = \frac{y_1}{p_1}$	$p_{11} = p_1$	e_1	y_{12}/p_{12}	p_{12}	e_1	$\cdots$			e_1	y_{1m}/p_{1m}	p_{1m}
e_2	$y_{21}/p_{21} = \frac{y_2}{p_2}$	$p_{22} = p_2$	e_2	$y_{22}/p_{22} = \frac{y_2}{p_2}$	$p_{22} = p_2$	e_2	$\cdots$			e_2	y_{2m}/p_{2m}	p_{2m}
$\vdots$	$\vdots$		$\vdots$	$\vdots$		$\vdots$	$\ddots$			$\vdots$	$\vdots$	$\vdots$
e_N	$y_{N1}/p_{N1} = \frac{y_N}{p_N}$	$p_{NN} = p_N$	e_N	y_{N2}/p_{N2}	p_{N2}	e_N	$\cdots$			e_N	y_{Nm}/p_{Nm}	p_{Nm}

Entonces se deduce de esta tabla que:

i) $P\left(z_i = \frac{y_{ki}}{p_{ki}}\right) = p_{ki}$

Además que:

ii) $E(z_i) = t_y \Rightarrow E(z_i) = \sum_U \frac{y_k}{\cancel{p_k}} \cancel{p_k} = \sum_U y_k = t_y$

z_i

iii) $V(z_i) = \sum_U \left(\frac{y_k}{p_k} - t_y\right)^2 \cdot p_k \Rightarrow$ ó $\Rightarrow \sum_U (z_i - E(z_i))^2 \cdot p_k$

↳ Estimador de Hansen Horwitz:

Un estimador "natural" para t_y es

$$\bar{z}_i = \frac{1}{m}\sum_{i=1}^{m} z_i \Rightarrow \bar{z} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_k}{p_k}$$

entonces:

$$\boxed{\hat{t}_{yp} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_k}{p_k}}$$

- El estimador de Hansen Horwitz tiene las siguientes propiedades.

i)
$$\begin{aligned} E(\hat{t}_{yp}) &= E\left(\frac{1}{m}\sum_{i=1}^{m}\frac{y_k}{p_k}\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} E\left(\frac{y_k}{p_k}\right) \\ &= \frac{1}{m}\sum_{i=1}^{m} E(z_i) \\ &= \frac{1}{m}\sum_{i=1}^{m} t_y \\ &= \frac{\cancel{m}}{\cancel{m}} t_y \Rightarrow \boxed{E(\hat{t}_{yp}) = t_y} \end{aligned}$$

Scanned with CamScanner

ii) $V(\hat{t}_{yp}) = V\left(\frac{1}{m}\sum_{i=1}^{m}\frac{y_k}{p_k}\right)$

$= \frac{1}{m^2}\sum_{i=1}^{m} V\left(\frac{y_k}{p_k}\right)$

$= \frac{1}{m^2}\sum_{i=1}^{m} E(z_i - E(z_i))^2$

$= \frac{1}{m^2}\, m\, E(z_i - E(z_i))^2$

$= \frac{1}{m} E(z_i - E(z_i))^2$

$= \frac{1}{m}\sum_{i=1}^{N}{}_{U} (z_i - t_y)^2 p_k$

Ejercicio de demostración

$$\boxed{V(\hat{t}_{yp}) = \frac{1}{m}\sum_U\left(\frac{y_k}{p_k} - t_y\right)^2 \cdot p_k}$$

↳ Varianza Estimada de $V(\hat{t}_{yp})$:

$$\hat{V}(\hat{t}_{yp}) = \frac{1}{m(m-1)}\sum_s\left(\frac{y_k}{p_k} - \hat{t}_{yp}\right)^2$$

→ xq' tengo q' calcular un parámetro.

Ejemplo:

	Exp 1	Exp 2	Exp 3	Exp 4
y_k	172	182	175	157
p_k	2/60	3/60	1/60	1/70
$z_i = y_k/p_k$	172(30)	182(20)	175(60)	157(70)

** → las probabilidades no suman 1 porq' s una muestra k

→ Factor de Expansión.

→ En Excel la varianza de $\hat{t}_{yp}$ es:

$V(\hat{t}_{yp}) = \frac{1}{4} * VAR(\quad)$

** COMENTARIO:

p_k debe ser proporcional a y_k ($p_k \propto y_k$), es decir

$p_k = C\, y_k$, tal que

$t_y = \sum_U y_k$, entonces

$\boxed{t_y = \sum_U \frac{p_k}{C}}$

y $\sum_U p_k = 1$, entonces

$\boxed{t_y = \frac{1}{C}}$

Si $p_k = C y_k$, entonces

$V(\hat{t}_{yp}) = \frac{1}{m}\sum_U\left(\frac{y_k}{C y_k} - t_y\right)^2 p_k$

$V(\hat{t}_{yp}) = \frac{1}{m}\sum_U\cancel{\left(\frac{1}{C} - \frac{1}{C}\right)^2} p_k$

$\boxed{V(\hat{t}_{yp}) = 0}$

→ Como elegimos p_k?? Los elementos q' contribuyen más al total tienen mayor probabilidad de selección; pero como no conozco muchas veces los valores de y_k, entonces busco una variable auxiliar altamente correlacionada con y_k y así poder dar correspondencia a la relación $y_k \sim p_k$. Esto se utiliza para muestreo PPT.

COMENTARIO: Una muestra no necesariamente se tiene q' parecer a la población.

→ Estimación del Tamaño en muestreo con reemplazo: $N = \sum_U y_k^*$ donde $y_k^* = 1 \ \forall \ k \in U$

entonces; $\hat{t}_{yp} = \frac{1}{m} \sum_s \frac{y_k}{p_k}$ → Recordar q' $p_k \propto y_k$ por lo q' $p_k = C y_k$

entonces:

$$= \frac{1}{m} \sum_s \frac{\cancel{y_k}}{C \cancel{y_k}}$$

entonces

$$= \frac{1}{m} \sum_s \frac{1}{C} \Rightarrow \text{como } \sum_U p_k = 1 = C = \frac{1}{N} \Rightarrow p_k = \frac{y_k}{N}.$$

En pocas palabras

$$\hat{t}_{yp} = \frac{1}{m} \sum_s \frac{1}{\frac{1}{N}} \Rightarrow \hat{t}_{yp} = \frac{1}{m} \sum_s N \Rightarrow \hat{t}_{yp} = \frac{1}{\cancel{m}} \cancel{m} N$$

$$\boxed{\hat{t}_{yp} = N}$$

Esto quiere decir q' si no se conoce N realmente, no se debe utilizar el estimador de Hansen Horwitz

Esto implica q' lo q' es $\overset{??}{\bar{X}}, \hat{P}$ y N no es adecuado estimarlo con $\hat{t}_{yp}$, a no ser q' se conozca el tamaño real de la población.

$\hat{t}_{yp}$ puede superar ampliamente a $\hat{t}_{y\pi}$ si $p_k \propto y_k$, como y_k generalmente no se conoce utilizamos una variable auxiliar x_k, tal q' $|cor(x_k, y_k)| \approx 1$ y $p_k \propto x_k$ igualmente.

▶ MUESTREO ALEATORIO SIMPLE SIN REEMPLAZO: Tiene las siguientes características:

- Sin reemplazo
- De tamaño fijo
- Sirve para poblaciones homogeneas en su variable a medir $CV_y = \frac{S_Y}{\bar{Y}} \cdot 100$ ojalá $< 20\%$
- Se seleccionan los elementos directamente del marco muestral

- El tamaño del soporte es $\boxed{\binom{N}{n} = n_Q}$. Cada una de las muestras del soporte tiene la misma probabilidad $p(s) = 1/k$ de ser seleccionada, donde k es el número de posibles muestras.

$$p(s) = \frac{1}{\binom{N}{n}}$$

- Todos los diseños de muestreo se comparan con este, ya q' el MAS es el diseño de referencia. Si mi diseño pierde con el MAS entonces el diseño planeado no sirve.
- Cuando las poblaciones no son homogéneos ste diseño no es eficiente para estimar parámetros $(t_y, \bar{x}, \ldots)$
- Esta comparación se hace a través del efecto de diseño (deff)

$$deff = \frac{V_{diseño\, Q}(\hat{\theta})}{V_{MAS}(\hat{\theta})} \qquad deff \begin{cases} < 1 & \text{Diseño Q más eficiente} \\ = 1 & \text{Igual} \\ > 1 & \text{MAS más eficiente.} \end{cases}$$

Si deff < 1 significa q' el otro diseño es más eficiente.

Ejemplo: Si $V_{MAS} = 200$, $V_{otro\,diseño} = 100$ $\Rightarrow$ deff $= 2$ → Esto quiere decir q' el otro diseño es 2 veces más eficiente.

- Si no hay información auxiliar de la variable a estimar, use MAS.
- Dentro de los diseños q' no usan información auxiliar es el mejor.

↳ Algoritmos de selección:

El objetivo de un algoritmo de selección es otorgar una probabilidad de selección a cada elemento del soporte:

⇒ Coordinado Negativo: [Sunter (1977)]

e_k	U_k (aleatorio)
1	U_1
2	U_2
3	U_3
⋮	⋮
N	U_N

con $U_k \sim U(0,1)$ ⇒

e_k	U_k (ordenado)
18	—
4	—
15	—
⋮	⋮
⋮	U_N

→ la muestra se constituye por los primeros n elementos.

Pasos:

1. generar n valores U_K, $K=1,\dots,N$ donde $U_K \sim U(0,1)$
2. Asociar a cada u_K un valor U_K
3. Ordenar de manera ascendente / descendente los valores U_K
4. Los primeros / últimos n elementos son los q constituyen la muestra

⇒ Fan-Muller-Rechuzo (1962).

1. generar $\xi \sim U(0,1)$
2. Calcular $C_K = \frac{n-n_K}{N-K+1}$ donde n_K = tamaño de la muestra antes del paso K
3. Si $\xi < C_K$ se selecciona el elemento, si no NO.
4. Esto se hace desde $K=1$ hasta obtener la muestra de tamaño n.

Ej: $N=30$ $n=7$

- Con $K=1$

$\xi = 0{,}5099$ $\quad C_1 = \frac{7-0}{30-1+1} = \frac{7}{30} = 0{,}23$

$C_1 < \xi$ NO SE SELECCIONA

- $K=2$

$\xi = 0{,}034$ $\quad C_2 = \frac{7-0}{30-2+1} = \frac{7}{29} = 0{,}24$

$C_2 > \xi$ → SE SELECCIONA

- $K=3$

$\xi = 0{,}542$ $\quad C_3 = \frac{7-1}{30-3+1} = \frac{6}{28} = 0{,}21$

$C_3 < \xi$ → NO SE SELECCIONA

y así sucesivamente hasta seleccionar $n=7$.

OJO: Este algoritmo garantiza $\frac{1}{\binom{N}{n}}$ probabilidad a cada elemento de ser seleccionado.

↳ PROBABILIDAD DE INCLUSION DE PRIMER ORDEN EN MAS:

$$\pi_K = P(K \in S) = P(I_K(S)=1) = \sum_{S \ni K} p(S) = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}}$$

$$= \frac{\frac{(N-1)!}{(N-1-n+1)!\,(n-1)!}}{\frac{N!}{(N-n)!\,n!}} = \frac{(N-1)!\,n!}{(n-1)!\,N!} = \frac{n!}{(n-1)!}\,\frac{(N-1)!}{N!} = n \cdot \frac{1}{N} \Rightarrow$$

$$\boxed{\Rightarrow \pi_K = \frac{n}{N}}$$

Scanned with CamScanner

↳ PROBABILIDAD DE INCLUSIÓN DE SEGUNDO ORDEN EN MAS:

$$\pi_{kl} = P(k \wedge l \in S) = P(I_k(s) I_l(s) = 1)$$

$$= \sum_{s} p(s) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{\frac{(N-2)!}{(\cancel{N-2-n+2})!(n-2)!}}{\frac{N!}{(N-n)!\,n!}}$$

$$= \frac{n!\,(N-2)!}{(n-2)!\,N!} = \frac{n(n-1)}{N(N-1)} \Rightarrow \boxed{\pi_{kl} = \frac{n(n-1)}{N(N-1)}}$$

↳ Δ_{kl} de MAS

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

Si $k = l$

$$\Delta_{kk} = \pi_k (1 - \pi_k)$$

$$\Delta_k = \frac{n}{N}\Big[1 - \underbrace{\frac{n}{N}}_{}\Big] = \frac{n(N-n)}{N^2}$$

↳ Corrector de finitud para poblaciones finitas.

Si $k \neq l$

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

$$\Delta_{kl} = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = \frac{-n(N-n)}{N^2(N-1)}$$

↳ Estimador de Horvitz - Thompson:

$$\hat{t}_{y\pi} = \sum_{s} \frac{y_k}{\pi_k} \quad \text{como } \pi_k = \frac{n}{N}, \text{ entonces:}$$

$$\hat{t}_{y\pi} = \frac{N}{n} \sum_{s} y_k \quad \text{donde } \frac{N}{n} \sum_{s} y_k \to \bar{y}_s$$

↳ Factor de expansión.

entonces

$$\hat{t}_{y\pi} = N \bar{y}_s$$

↳ Varianza de $\hat{t}_{y\pi}$.

$$V(\hat{t}_{y\pi}) = V(N\bar{y}_s)$$

$$= N^2 V(\bar{y}_s)$$

$$= \frac{N^2}{n^2} V\Big(\sum_{s} y_k\Big)$$

$$= \frac{N^2}{n^2} V\Big(\sum_{U} I_k y_k\Big)$$

donde I_k es la variable aleatoria, entonces. teniendo en cuenta q' es la varianza de una suma

Y dado q' $V\big(\sum_{n} a_i x_i\big) = \sum_{n} V(a_i x_i) + \sum\sum_{1 \le i \le j \le n} Cov(a_i x_i, a_j x_j)$

entonces

$$= \frac{N^2}{n^2}\Big[\sum_{N} V(I_k y_k) + \sum\sum_{k \neq l} Cov(y_k I_k, y_l I_l)\Big]$$

$$= \frac{N^2}{n^2}\left[\sum_N y_k^2 \underbrace{V(I_k)}_{ⓐ} + \sum_{k\neq l}\sum y_k y_l \underbrace{cov(I_k, I_l)}_{ⓑ}\right]$$

$\Rightarrow$ Dado q' ⓐ $\Rightarrow$ $V(I_k) = \begin{cases} \pi_k - \pi_k^2 & \text{si} \quad k = k \\ \pi_{kl} - \pi_k \pi_l & \text{si} \quad k \neq l \end{cases}$ entonces.

$$= \frac{N^2}{n^2}\left[\frac{n(N-n)}{N^2}\sum_N y_k^2 - \frac{n(N-n)}{N^2(N-1)}\sum_{k\neq l}\sum y_k y_l\right]$$

$$= \frac{N^2}{n^2}\left(\frac{n(N-n)}{N^2}\right)\left[\sum_N y_k^2 - \frac{1}{N-1}\sum_{k\neq l}\sum y_k y_l\right]$$

$$= \frac{N}{N}\frac{N-n}{n}\left[\sum_N y_k^2 - \frac{1}{N-1}\sum_{k\neq l}\sum y_k y_l\right]$$

Usando la identidad.

$$\sum_N (y_k - \bar{y}_U)^2 = \sum_N y_k^2 - \frac{\left(\sum_N y_k\right)^2}{N} = \frac{1}{N}\left[(N-1)\sum_N y_k^2 - \sum_{k\neq l}\sum y_k y_l\right]$$

se tiene q'

$$\sum_N y_k^2 - \frac{1}{N-1}\sum_{k\neq l}\sum y_k y_l = \frac{N}{N-1}\sum_N (y_k - \bar{y}_U)^2$$

entonces.

$$= \frac{N}{N}\frac{(N-n)}{n}\left[\frac{N}{N-1}\sum_N (y_k - \bar{y}_U)^2\right] = \frac{N^2}{N}\frac{(N-n)}{n}\left[\underbrace{\frac{1}{N-1}\sum_N (y_k - \bar{y}_U)^2}_{S^2_{y_U}}\right]$$

$$= \frac{N^3 - N^2 n}{Nn} S^2_{y_U} = \frac{N^3}{Nn}\left[1 - \frac{n}{N}\right] S^2_{y_U}$$

$$\boxed{= V(\hat{t}_{y\pi})_{MAS} = \frac{N^2}{n}\underbrace{\left(1 - \frac{n}{N}\right)}_{\text{Corrector de finitud.}} S^2_{y_U}}$$ donde $\frac{n}{N}$ = Fracción de muestreo.

$$\boxed{\hat{V}(\hat{t}_{y\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_S}}$$ ok con $S^2_{y_S} = \frac{\sum_S (y_k - \bar{y}_S)^2}{n-1}$

Para demostrar su insesgamiento bastaría con hacer

$$E(\hat{V}(\hat{t}_{y\pi})) = E\left(\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{y_S}\right)$$

$$= \frac{N^2}{n}\left(1 - \frac{n}{N}\right) E(S^2_{y_S})$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right) E\left(\frac{1}{n-1}\sum_s (y_k-\bar{y}_k)^2\right)$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right) E\left(\sum_s y_k^2 - n\bar{y}_k^2\right)$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left[E\left(\sum_s y_k^2\right) - n E\left(\bar{y}_k^2\right)\right]$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left[E\left(\sum_U I_k y_k^2\right) - n E\left[\frac{\hat{t}_{y\pi}^2}{N^2}\right]\right]$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left[\frac{n}{N}\sum y_k^2 - \frac{n}{N^2} E(\hat{t}_{y\pi})^2\right]$$

donde $E(\hat{t}_{y\pi})^2 = V(\hat{t}_{y\pi}) + t_y^2$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left[\frac{n}{N}\sum_U y_k^2 - \frac{n}{N^2}\left(\frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{yu}^2 + t_y^2\right)\right]$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{n-1}\right)\left[\frac{n}{N}\sum_U y_k^2 - \frac{n}{N^2}\left(\frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{yu}^2 + t_y^2\right)\right]$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left[\frac{1}{N}\sum_U y_k^2 - \frac{1}{n}\left(1-\frac{n}{N}\right)S_{yu}^2 - \frac{1}{N^2}t_y^2\right]$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{n}{n-1}\right)\left(\frac{1}{N}\left(\sum_U y_k^2 - \frac{\left[\sum_U y_k\right]^2}{N}\right) - \frac{N-n}{n} S_{yu}^2\right)$$

$$= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\frac{n}{n-1}\left[\frac{N-1}{N}S_{yu}^2 - \frac{N-n}{n}S_{yu}^2\right] \Rightarrow = \frac{N^2}{n}\left(1-\frac{n}{N}\right)S_{yu}^2 = E(\hat{V}(\hat{t}_{y\pi})) = V(t$$

Libro andrés Gutiérrez

- Estimación de la media Poblacional

$$\hat{\bar{y}}_\pi = \frac{\hat{t}_{y\pi}}{N} = \frac{\sum_s y_k}{n} = \bar{y}_s$$

$$V_{mas}(\hat{\bar{y}}_\pi) = \frac{1}{n}\left(1-\frac{n}{N}\right)S_{yu}^2$$

$$\hat{V}_{mas}(\hat{\bar{y}}_\pi) = \frac{1}{n}\left(1-\frac{n}{N}\right)S_{ys}^2$$

- Intervalo de confianza

$IC(\bar{y}) = \bar{y} \pm t_{n-1,1-\alpha/2}\sqrt{V(\bar{y})}$ → Cuando el tamaño de la muestra es suficientemente grande (≥ 30) el percentil t puede cambiarse por el correspondiente valor del percentil Z.

$IC(\hat{t}_{y\pi}) = \hat{t}_{y\pi} \pm t_{n-1,1-\alpha/2}\sqrt{V(\hat{t}_{y\pi})}$ →

- Coeficiente de Variación.

$$CV = \frac{\sqrt{V(\hat{t}_{y\pi})}}{t_y} \qquad \hat{CVE} = \frac{\sqrt{\hat{V}(\hat{t}_{y\pi})}}{\hat{t}_{y\pi}}$$

- La varianza del estimador es proporcional a la varianza de los datos.
- La precisión depende de S^2_{yu}
- Si hago censo no hay varianza.

- TAMAÑO DE LA MUESTRA PARA MEDIAS Y TOTALES:

Bajo MAS sin reemplazo, tenemos q':

$$IC(\bar{y}_s) = \bar{y}_s \pm Z_{1-\alpha/2} \frac{S_{yu}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

y como generalmente no se conoce S_{yu} se sustituye por S_{ys}.

la expresión $Z_{1-\alpha/2} \frac{S_{yu}}{\sqrt{n}} \sqrt{1-\frac{n}{N}}$ = Márgen de Error $= e$

entonces para decidir sobre la precisión mínima de la investigación tenemos que

$$P_r(|\bar{y}_s - \bar{y}_u| \leq e) = 1-\alpha$$
$$P_r(-e \leq y_s - y_u \leq e) = 1-\alpha$$

Por lo tanto la cantidad a minimizar es e:

$$e = Z_{1-\alpha/2} \frac{S_{yu}}{\sqrt{n}} \sqrt{1-\frac{n}{N}}$$

$$\sqrt{n} = Z_{1-\alpha/2} \frac{S_{yu}}{e} \sqrt{1 - \frac{n}{N}} \Rightarrow n = z^2 \frac{S^2_{yu}}{e^2}\left(1 - \frac{n}{N}\right)$$

$$n = \frac{z^2 S^2_{yu}}{e^2 N} - \frac{n}{N} \frac{z^2 S^2_{yu}}{e^2}$$

$$n + \frac{n}{N} \frac{z^2 S^2_{yu}}{e^2} = \frac{z^2 S^2_{yu}}{e^2 N} = n\left(1 + \frac{z^2 S^2_{yu}}{N e^2}\right) = \frac{z^2 S^2_{yu}}{e^2 N}$$

si $n_0 = \frac{z^2 S^2_{yu}}{e^2}$ tenemos que

$$\boxed{n = \frac{n_0}{1 + \frac{n_0}{N}}}$$

$\Rightarrow$ Tamaño de muestra cuando se desea controlar el Error máximo absoluto.

y donde las cantidades Z y e son previamente establecidas por el investigador.

→ Tamaño de la Muestra Controlando el Error Máximo Relativo:

En algunas ocasiones se quiere lograr una precisión relativa, aunque en estos casos el tamaño de muestra es frecuentemente mayor q' cuando se desea controlar el error máximo absoluto.

El error máximo relativo se define como.

$$\varepsilon = \frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U} = \frac{e}{\bar{y}_U}$$ entonces.

$$P_r\left(\left|\frac{\bar{y}_s - \bar{y}_U}{\bar{y}_U}\right| \leq e\right) = 1-\alpha$$

$$P_r\left(|\bar{y}_s - \bar{y}_U| \leq e|\bar{y}_U|\right) = 1-\alpha$$

entonces:

$$e|\bar{y}_U| = Z_{1-\alpha/2}\frac{S_{y_U}}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

y despejando n se tiene.

$$\boxed{n \geq \frac{K_0}{1+\frac{K_0}{N}}}$$

con $K_0 = \dfrac{Z^2_{1-\alpha/2}\, S^2_{y_U}}{e^2\, \bar{y}_U^2} \rightarrow CV^2$

entonces $K_0 = \dfrac{Z_{1-\alpha/2}\, CV^2}{e^2}$ ⇒ si la población es muy grande y n/N es despreciable

por lo cual si N es muy grande podremos escribir:

$$\boxed{K_0 = \frac{Z^2\, CV^2}{e^2}}$$

después de un poco de algebra se puede aproximar por

$$K_0 = \frac{CV^2}{CV(\bar{y})^2}$$

donde $CV[\bar{y}]$ es el CV deseado para la media muestral.

• Consideración:
El S_{yu} lo puedo calcular de otros estudios, una prueba piloto o median te simulaciones según la distribución q' se cree tiene la variable de interés.

• generalmente $e = 1\%$
$CV = 5\%$

↳ Tamaño de muestra según CV:

$$CV = \frac{\sqrt{\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{ys}}}{t_y}$$

despejando n tenemos q':

$$n = \frac{N^2}{\frac{CV^2}{CV_y^2} + N}$$

• Consideración:
• Si la muestra piloto para estimar $S^2_{y_1}$ no es tomada intencionalmente atendiendo a los criterios del experto, entonces

$$n = \frac{Z^2 S^2_{y_1}}{e^2}\left(1 + \frac{2}{n_1}\right)$$

donde n_1 = tamaño de la prueba piloto

y $\frac{2}{n_1}$ es el precio q' se paga por el desconocimiento de $S^2_{y_U}$.

Ejemplo:
Se quiere estimar el ingreso mensual de una población y para determinar el tamaño muestral se tomó una muestra piloto con la siguiente información:

$N = 10000$ $S^2_{y_1} = 6{,}059$ (Varianza de la muestra Piloto)
$n_1 = 30$ (tamaño de la muestra piloto). $\bar{X} = 2{,}45$

a). Si la muestra piloto fue seleccionada por un experto, y se desea conocer el tamaño de muestra mínimo para estimar el ingreso promedio y total de la comunidad con un ERROR ABSOLUTO no mayor a 0,2 salarios mínimos (2000 en el caso del total) y una confiabilidad mínima del 95%, como se realizaría dicho calculo?

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde} \quad n_0 = \frac{Z^2_{1-\alpha/2}\, S^2_{ys}}{e^2}$$

$$n = \frac{\frac{1{,}96^2 \cdot 6{,}059}{0{,}2^2}}{1 + \frac{1}{10000}\left(\frac{1{,}96^2 \cdot 6{,}059}{0{,}2^2}\right)} = 549{,}91 \approx 550$$

es el número de personas q' se deben seleccionar sin incluir a la muestra piloto.

b) Considere q' la muestra piloto es una MAS. Bajo las mismas condiciones, cuál sería el tamaño de la muestra final?

$$n = \frac{z^2_{1-\alpha/2} S^2_{y_U}}{e^2}\left(1+\frac{2}{n_1}\right) \Rightarrow n = \frac{1{,}96^2\ 6{,}059}{0{,}2^2}\cdot\left(1+\frac{2}{30}\right) = 620{,}70 \approx 621$$

lo cual indica q' hay q' escoger 591 personas adicionales a las 30 seleccionadas.

c) Si lo q' se desea es controlar el ERROR RELATIVO antes q' el absoluto y se busca q' este sea como máximo del 5% y adicionalmente se asume q' se conoce una estimación a priori de la media y la varianza q' serán dadas por la muestra piloto, cuál sería el tamaño de la muestra?

$$n = \frac{k_0}{1+\frac{k_0}{N}} \qquad \text{con } k_0 = \frac{z^2 CV^2}{e^2}$$

$$k_0 = \frac{1{,}96^2\left(\frac{\sqrt{6{,}059}}{2{,}45}\right)^2}{0{,}05^2} = 290{,}961784$$

$$n = \frac{290{,}961784}{1+\frac{290{,}961784}{10000}} = 282{,}05 \approx \boxed{283}$$ → siempre se aproxima al mayor!

▶ DISEÑO MUESTRAL BERNOULLI:

Se utiliza para revisar por ejemplo cuestiones de narcotráfico o incautaciones → se requiere usar para trabajos q' requieren una decisión en el momento mismo.

En este existen N experimentos en los cuales se decide la selección o no selección de cada elemento de U.

↳ Diseño de muestreo:

$$p(s) = \begin{cases} \pi^{n(s)}(1-\pi)^{N-n(s)} & \text{si } s \text{ tiene tamaño igual a } n(s) \\ 0 & \text{en otro caso.} \end{cases}$$

> donde $n(s)$ es la variable aleatoria.

donde $\pi = \pi_k \quad \forall\ k \in U \qquad$ y $\sum_{s \in Q} p(s) = 1$

↳ Algoritmo de selección:

Pasos

1) Fijar el valor de π tal q' $0 < \pi < 1$
2) Obtener $\varepsilon_k \sim U(0,1)\ \forall\ k = 1, 2, \ldots, N$
3) El elemento k-ésimo pertenece a la muestra con $P_k = \pi$. Es decir si $\varepsilon_k < \pi$ el individuo k-ésimo es seleccionado.

↳ Tamaño del soporte:

- Definiendo a Q_r como el soporte q' contiene todas las posibles muestras de tamaño $r=n$, tenemos

$$\# Q_r = \binom{N}{r}$$

- Definiendo a Q como el soporte q' contiene todas las posibles muestras de tamaño entre $r=0$ y $r=N$

$$\# Q = \sum_{i=0}^{N} \binom{N}{r_i} = 2^N$$

↳ Distribución del tamaño de la muestra:
Bajo Bernoulli, $n(s) \sim Bin(r, \pi)$

$$P(n(s)=r) = \sum_{s \in Q_r} p(s) = \binom{N}{r} \pi^r (1-\pi)^{N-r}$$

con $r=1,2,\dots,N$ y Q_r el soporte q' contiene todas las posibles muestras de tamaño r, donde $Q_r \subset Q$

- En $p(s)$ Bernoulli no hay manera de calcular el tamaño de la muestra.

↳

$$E(n(s)) = N\pi$$

$$V(n(s)) = N\pi(1-\pi)$$

↳ Probabilidad de Inclusión de primer orden

Teniendo en cuenta q' $\pi_k = P(k \in S)$, y existen $\binom{N-1}{r-1}$ muestras de tamaño r que contienen $\ni k$, tenemos q'

$$\pi_k = \sum_{s \in Q} I_k(s)\, p(s)$$

$$\pi_k = \cancel{0 \cdot p(s_0)}^{0} + \binom{N-1}{0} p(s_1) + \binom{N-1}{1} p(s_2) + \dots + \binom{N-1}{r-1} p(s_r) + \binom{N-1}{N-1} p(s_N)$$

$$= \binom{N-1}{0} \pi^1 (1-\pi)^{N-1} + \binom{N-1}{1} \pi^2 (1-\pi)^{N-2} + \dots + \binom{N-1}{r-1} \pi^r (1-\pi)^{N-r} + \binom{N-1}{N-1} \pi^N (1-\pi)^{N-N}$$

$$\pi_k = \sum_{r=1}^{N} \binom{N-1}{r-1} \pi^r (1-\pi)^{N-r}$$

Si voy un paso atrás tengo:

$$= \sum_{r=0}^{N-1} \binom{N-1}{r} \pi^{r+1} (1-\pi)^{N-(r+1)} = \pi \underbrace{\sum_{r=0}^{N-1} \binom{N-1}{r} \pi^r (1-\pi)^{N-r-1}}_{\sim Bin(N-1,\pi)}$$

entonces $\boxed{\pi_k = \pi}$.

→ Probabilidad de Inclusión de 2^{do} Orden.

$$\pi_{kl} = P(k \wedge l \in S) = \underbrace{0 \cdot P(s_0)}_{n=0} + \underbrace{0 \cdot P(s_1)}_{n=1} + \underbrace{1 \, P(s_2)}_{n=2} + \overbrace{\binom{N-2}{1} P(s_3)}^{n=3} +$$

$$\binom{N-2}{2} P(s_3) + \cdots + \binom{N-2}{r-2} P(s_r) + \binom{N-2}{N-2} P(\omega_N)$$

entonces

$$\pi_{kl} = \sum_{r=2}^{N} \binom{N-2}{r-2} \pi^r (1-\pi)^{N-r}$$

$$\pi_{kl} = \sum_{r=0}^{N-2} \binom{N-2}{r} \pi^{r+2} (1-\pi)^{N-(r+2)}$$

$$\pi_{kl} = \pi^2 \underbrace{\sum_{r=0}^{N-2} \binom{N-2}{r} \pi^r (1-\pi)^{N-2-r}}_{}$$

$$\pi_{kl} = \pi^2$$

→ Estimador de Horvitz-Thompson

- $\hat{t}_{y_B} = \sum_s \frac{y_k}{\pi_k}$

- $V(\hat{t}_{y_B}) = \sum_U \sum \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$ con $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$

  y $\Delta_{kl} = \begin{cases} 0 & \text{si } k \neq l \\ \pi(1-\pi) & \text{si } k = k = l \end{cases}$

entonces con $k=l$

$$V(\hat{t}_{y_B}) = \sum_U \sum \Delta_{kk} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}$$

$$= \frac{\pi(1-\pi)}{\pi^2} \sum_U y_k^2 \Rightarrow \boxed{\left(\frac{1}{\pi} - 1\right) \sum_U y_k^2 = V(\hat{t}_{y_B})}$$

- $\hat{V}(\hat{t}_{y_B})$ con $k=l$

$$\hat{V}(\hat{t}_{y_B}) = \frac{1}{\pi}\left(\frac{1}{\pi} - 1\right) \sum_s y_k^2$$

- Comentario: El estimador de H.T para Diseño Bernoulli es inapropiado ya q' sobreestima al total ya q' si $n=N$

$$\hat{t}_{y_\pi} = \frac{1}{\pi} \sum_U y_k = \frac{t_y}{\pi}$$

Deducir: $\sum$ conocer el total de impuesto pagado por cada empresa (dato)

impuesto como auxiliar

S.PiPS(n, datos$impuesto)

datos[res[,1],]

(Sam)

res[,2]

Sam[, colnames(Sam) %in% c('subsidio','Empleado','Impuesto')]

PiPS(Yk, Pik)

$$\sum_S \sum \frac{\pi(1-\pi)}{\pi^4} J^2$$

$$\frac{1-\pi}{\pi^3} \sum_S J^2$$

$$\frac{1}{\pi^3} - \frac{1}{\pi^2}$$

$$\frac{\pi(1-\pi)}{\pi^2} \sum_S 1$$

$$\frac{1-\pi}{\pi}$$

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

$$\hat{P}_\pi = \frac{1}{N} \sum_S \frac{z_{dk}}{\pi_k} \Rightarrow V(\hat{P}_\pi) = V\left(\frac{1}{N}\sum_S \frac{z_{dk}}{\pi_k}\right)$$

$$= \frac{1}{N^2} \sum_S \frac{z_{dk}^2}{\pi_k^2} V(I_k)$$ $\Rightarrow$ se había visto que $Cov(I_k, I_l) = 0$ si $k \neq l$

entonces

$$= \frac{1}{N^2} \sum_S \frac{z_{dk}^2}{\pi_k^2} \pi(1-\pi)$$

$$V(\hat{P}_\pi) = \frac{1}{N^2} \frac{\pi(1-\pi)}{\pi^2} \sum_S z_{dk}^2$$

$$\hat{V}(\hat{P}_\pi) = \frac{1}{N^2} \frac{1}{\pi_{k}} \pi(1-\pi) \sum_S z_{dk}^2$$

$$= \frac{1}{N^2} \frac{\pi(1-\pi)}{\pi^2} \sum_S z_{dk}^2$$

$$= \frac{(1-\pi)}{\pi^2} \sum_S z_{dk}^2$$

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

- Estimador alternativo del total.

El estimador alternativo es el estimador del MAS

$$\hat{t}_{y\,ALT} = N\bar{y}_s = N \frac{\sum_s y_k}{n}$$

- Comentario:

Diseños de tamaño de muestra aleatorios suelen ser más ineficientes q' los diseños de tamaño fijo

↳ Efecto Diseño.

$$deff(\hat{t}_B) = \frac{V(\hat{\theta}_{BERN})}{V(\hat{\theta}_{MAS})}$$

$$deff(\hat{t}_B) = \frac{\left(\frac{1}{\pi}-1\right)\sum_U y_k}{N^2\left(1-\frac{n}{N}\right)S^2_{y_U}}$$

dado q' $p(\cdot)$ Bernoulli no tiene tamaño de muestra fijo hacemos de π la fracción de muestra fijo : $\pi = n/N$

$$deff(\hat{t}_B) = \frac{\left(\frac{N}{n}-1\right)\sum_U y_k}{N^2\left(1-\frac{n}{N}\right)S^2_{y_U}} = \frac{\frac{(N-n)\sum_U y_k}{n}}{\frac{N^2(N-n)S^2_{y_U}}{N}} = \frac{N(N-n)\sum_U y_k}{N^2 n(N-n)S^2_y}$$

$$= \frac{(N-n)\sum_U y_k}{S^2_{y_U}(N-n)\, n\, N} = \frac{\bar{y}_k}{S^2_{y_U} N}$$

▶ ESTIMACION EN DOMINIO. (Andres Gutierrez).

Un dominio es una subpoblación específica o subgrupo poblacional q' cumple las siguientes condiciones.

1) $U_d \subset U$ tal q' $U = \bigcup_{d=1}^{D} U_d$

2) Si $k \in U_L$ entonces $k \notin U_d$ para $d \neq L$

3) El número de elementos en el dominio U_d es N_d y es llamado Tamaño Absoluto del Dominio.

4) El tamaño Relativo de U_d está dado por $P_d = \frac{N_d}{N}$

- Total de un Dominio.

$$t_{yd} = \sum_{U_d} y_k$$

- Total de la variable de Interés en el dominio

$$t_{yd} = \sum_{U} y_{dk}$$ donde $y_{dk} = z_{dk} * y_k$ siendo

$$z_{dk} = \begin{cases} 1 & \text{si } k \in U_d \\ 0 & \text{en otro caso.} \end{cases}$$

- Tamaño del Dominio:

$$N_d = \sum_{U} z_{dk}$$

- Media de la Variable de Interés en el Dominio

$$\bar{y}_{Ud} = \frac{t_{yd}}{N_d} = \frac{\sum_U y_{dk}}{N_d}$$

↳ Estimación del Total de Un Dominio. (de Horvitz Thompson)

- $$\hat{t}_{yd\pi} = \frac{N}{n} \sum_{s} y_{dk} = \frac{N}{n} \sum_{s_d} y_k$$

- $$V(\hat{t}_{yd\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{ydU}$$

- $$\hat{V}(\hat{t}_{yd\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yds}$$

Con S^2_{ydU} y S^2_{yds} el estimador de la varianza de los valores de la característica de interés. y_d en el universo y en la muestra.

(!) Nótese q' en la expresión S^2_{ydU} los valores q' intervienen son los de la variable de interés si el elemento pertenece al dominio y ceros si el elemento no pertenece al dominio. Por tanto, las anteriores expresiones van a tomar valores grandes por la inclusión de ceros. Este es el precio q' se debe pagar por el desconocimiento de la pertenencia de los elementos a los dominios.

↳ Estimación del tamaño absoluto de un dominio.

- $$\hat{N}_{d\pi} = \frac{N}{n} \sum_{s} z_{dk} = \frac{N}{n} \sum_{s_d} z_k$$

- $$V(\hat{N}_{d\pi}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{zdU}$$

- $\hat{V}(N\hat{d}_\pi) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{zds}$ $\quad$ (**) $S^2_{zds} = \frac{n}{n-1}\hat{p}_{s1}(1-\hat{p}_{s1})$ donde $\hat{p} = \frac{nd}{n}$

En las expresiones S^2_{zdu} y S^2_{zds} los valores q' intervienen son 1's si $k \in Ud$ y 0's si $k \notin Ud$.

→ Estimación del Tamaño Relativo de un Dominio

- $\hat{P}_{d\pi} = \frac{1}{N}\sum_s \frac{N}{n} z_{dk} = \frac{1}{n}\sum_s z_{dk} = \frac{n_d}{n}$

- $V(\hat{P}_{d\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{zdu}$ $\quad$ * * vu si $S^2_{zdu} = \frac{N}{N-1} P(1-P)$ ✓

- $\hat{V}(\hat{P}_{d\pi}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S^2_{zds}$ $\quad$ ** vu si $S^2_{zds} = \frac{n}{n-1}\hat{P}_{s1}(1-\hat{P}_{s1})$ ✓

con S^2_{zdu} y S^2_{zds} el estimador de la varianza de los valores de la característica de interés y_d en el universo y en la muestra.

→ Estimación de la media de un Dominio:

- $\hat{\bar{y}}_{Ud\pi} = \frac{N/n \sum_s y_{dk}}{N_d}$

- $V(\hat{\bar{y}}_{Ud\pi}) = \frac{1}{N_d^2}\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{ydu}$

- $\hat{V}(\hat{\bar{y}}_{Ud\pi}) = \frac{1}{N_d^2}\frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{yds}$

Para poder utilizar este estimador es necesario conocer N_d. En la práctica pocas veces se conoce, por lo cual hay q' utilizar una estimación alternativa

$$\hat{\bar{y}}_{sd} = \frac{\hat{t}_{yd\pi}}{\hat{N}_{d\pi}} = \frac{\sum_s y_{dk}}{z_{dk}} = \frac{\sum_{sd} y_k}{n_d}$$

* ⇒ Por el momento V y $\hat{V}$ no se calculan.

$$= N_d \bar{y}_{sd}$$

Ipsos

MUESTREO ALEATORIO SIMPLE CON REEMPLAZO.

Sea $p_1, p_2, \dots, p_N$ probabilidades de selección con $p_k = \frac{1}{N}$

"en este caso el tamaño de la muestra puede ser mayor a N."

y $$p\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_N \end{pmatrix} = \frac{m!}{s_1! s_2! \dots s_N!} p_k^{\sum s_i} = \frac{m!}{s_1! s_2! \dots s_N!} \left(\frac{1}{N}\right)^{\sum s_i}$$

→ los elementos tienen la misma probabilidad de selección pero las muestras no tienen la misma probabilidad $p(s)$.

Entonces:

- $$p(s) = \begin{cases} \frac{m!}{n_1(s)! n_2(s)! \dots n_N(s)!} \prod_U \left(\frac{1}{N}\right)^{n_k(s)} & \text{si } \sum_U n_k(s) = m \\ 0 & \text{en otro caso.} \end{cases}$$

- Cardinalidad:

$$\#Q = \binom{N+m-1}{m}$$

- Probabilidad de inclusión de Primer orden

$$\pi_k = 1 - \left(1 - \frac{1}{N}\right)^m$$

la $\sum_U \pi_k$ nos da el tamaño efectivo $= m$ ya q' en algunos casos se da q' $m \neq n$.

- pwr estimador

$$t_{\hat{y}p} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k} = \frac{N}{m} \sum_{i=1}^{m} y_k$$

Nota: El pwr estimador y el de H-T son iguales en MASR?

NO!!! ... Ya q' acá se toman en cuenta los elementos repetidos

- Varianza

$$V(t_{\hat{y}p}) = V\left(\frac{1}{m} \sum_{i=1}^{m} \frac{y_k}{p_k}\right), \text{ haciendo } \frac{y_k}{p_k} = z_i$$

$$= \frac{1}{m^2} \sum_{i=1}^{m} V(z_i)$$

$$= \frac{1}{m^2} \sum_{i=1}^{m} E(z_i - E(z_i))^2$$

$$= \frac{1}{m^{\cancel{2}}} \cancel{m}\, E(z_i - E(z_i))^2$$

$$= \frac{1}{m} E(z_i - t_y)^2$$

$$= \frac{1}{m} E\left(\frac{y_k}{p_k} - t_y\right)^2$$

$$= \frac{1}{m} \sum_U \left(\frac{y_k}{p_k} - t_y\right)^2 p_k$$

$$= \frac{1}{Nm} \sum_U (Ny_k - t_y)^2$$

$$= \frac{1}{Nm} \sum_U (N^2 y_k^2 - 2Ny_k t_y + t_y^2)$$

$$= \frac{1}{Nm} N^2 \sum_U \left(y_k^2 - \frac{2y_k t_y}{N} + \frac{t_y^2}{N^2}\right)$$

$$= \frac{N}{m} \sum_U (y_k^2 - 2y_k \bar{y}_U + \bar{y}_U^2)$$

$$= \frac{N}{m} \sum_U (y_k - \bar{y}_U)^2$$

$$\longrightarrow \boxed{o^2 = \frac{N}{m} \sum_{k=1}^{N} y_k^2 - t_y^2}$$

$$\boxed{V(\hat{t_{yp}}) = \frac{N(N-1)}{m} S^2_{y_U}} \longrightarrow \text{con } S^2_{y_U} = \frac{1}{N-1} \sum_U (y_k - t_y)^2$$

- Estimación de la varianza del estimador pwr

$$\hat{V}(\hat{t_{yp}}) = \frac{1}{m(m-1)} \sum_{Sr} \left(\frac{y_k}{p_k} - \hat{t_{yp}}\right)^2$$

$$= \frac{1}{m(m-1)} \sum_{Sr} (Ny_k - \hat{t_{yp}})^2$$

$$= \frac{N^2}{m(m-1)} \sum_{Sr} (y_k - \bar{y}_{Sr})^2$$

$$= \frac{N^2 \cancel{(m-1)}}{m \cancel{(m-1)}} S^2_{y_{Sr}} \Rightarrow \boxed{\hat{V}(\hat{t_{yp}}) = \frac{N^2}{m} S^2_{y_{Sr}}}$$

$$\text{con } S^2_{y_{Sr}} = \frac{1}{m-1} \sum_{Sr} (y_k - \bar{y}_{Sr})^2$$

- Algoritmo de selección:
  - Método secuencial de Tillé:

Seleccionar n_k veces el elemento k-ésimo de acuerdo a una distribución binomial:

$$Bin(n, p).$$

$$Bin\left(m - \sum_{i=1}^{k-1} n_i \,,\, \frac{1}{N-k+1}\right)$$

donde
m = tamaño de la muestra deseado
k = elemento k-ésimo del paso
N = Universo

• Efecto Diseño:

$$deff = \frac{V(\hat{ty}_e)}{V(\hat{ty}_\pi)} \quad si \quad deff \begin{cases} >1 & \text{Más preciso el MAS} \\ <1 & \text{Más preciso MASR} \end{cases}$$

Para hacerlos comparables vamos a hacer una restricción $m=n$, es decir, tomamos la muestra efectiva

$$deff = \frac{\frac{N(N-1)}{m}\cancel{S^2_{yu}}}{\frac{N^2}{n}\left(1-\frac{n}{N}\right)\cancel{S^2_{yu}}} = \frac{\frac{\cancel{N}(N-1)}{\cancel{n}}}{\frac{N^{\cancel{2}}}{\cancel{n}}\left(1-\frac{n}{N}\right)} = \frac{\cancel{N}(N-1)}{\cancel{N}(N-n)} \boxed{= \frac{N-1}{N-n} = deff}$$

cuando $n \to N$ $\quad deff > 1$

y si $N \to \infty$:

$$deff = \frac{N-1}{N-n} = \frac{\cancel{\left(1-\frac{1}{N}\right)} \to 1 \text{ (si } N\to\infty)}{\left(1-\frac{n}{N}\right)} = \frac{1}{1-\frac{n}{N}} \boxed{= \frac{1}{1-f} = deff}$$

con f = fracción de muestreo.

Porqué los diseños de muestra aleatoria pierden con los de tamaño fijo?
→ Porque en el MAS si aplicamos el factor de corrección por finitud.

- MUESTREO SISTEMATICO

$N \to$ Universo
$n \to$ Muestra
$a \to$ # arranques
$c \to$ residuo.

$a \to$	1	2	3	4	5
	y_1	y_2	y_3	y_4	y_5
	y_6	y_7	y_8	y_9	y_{10}
	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}
	y_{16}	y_{17}	y_{18}	y_{19}	

Ej: Si tengo un universo con 19 elementos y quiero seleccionar una muestra entonces:

$a \to$ Defino el número de grupos
$n \to$ mi muestra

en este ejemplo

$a=5$
$n=3$

$\Rightarrow$ $19 \,|\, \underline{5}$, 4, $3 \to n$ $\quad$ → 4 muestras con $n=4$; $5-4=1$ muestra con $n=3$

entonces $\quad N = an + c \quad$ con $\quad \Rightarrow$ # muestra: $\{n, n+1\}$

$\downarrow \; \downarrow \; \downarrow \; \downarrow$
$19 \quad 5 \quad 3 \quad 4$

Con cuál arranque me quedo? $\Rightarrow$ Hago MAS para n fijo igual a 1.

entonces

$$C = N - \left\| \frac{N}{a} \right\| a \qquad \text{donde } \left\| \frac{N}{a} \right\| \text{ es la parte entera de la división } \frac{N}{a}$$

- $\# Q_r = a$ donde $a = \frac{N}{n}$ cuando $C = 0$ ó $a = \left\| \frac{N}{n} \right\|$ cuando $C \neq 0$

- $p(s) = \begin{cases} 1/a & \text{si } s \in Q_r \\ 0 & \text{en otro caso} \end{cases}$

- $\pi_k = \frac{1}{a} \approx \frac{N}{n}$

- $\pi_{kl} = \begin{cases} 1/a & \text{si } k \wedge l \in S_r \\ 0 & \text{en otro caso.} \end{cases}$

- Estimador de Horvitz Thompson

$$\hat{t_y}_\pi = \sum_{S_r} \frac{y_k}{\pi_k} = a \sum_{S_r} y_k = \boxed{a \cdot t_{sr} = \hat{t_y}_\pi}$$

$$\text{ó} = a \sum_{S_r} y_k \approx \frac{N}{n} \sum_{S_r} y_k \; \boxed{\approx N \bar{y}_{sr} = \hat{t_y}_\pi}$$

- Varianza del estimador

$$V(\hat{t_y}_\pi) = \sum_U \sum \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} = \sum_U \sum (\pi_{kl} - \pi_k \pi_l) \frac{y_k y_l}{\pi_k \pi_l}$$

$$= \sum_U \sum \frac{\pi_{kl}\, y_k y_l}{\pi_k \pi_l} - y_k y_l$$ como $k \wedge l$ tienen q' estar en el mismo grupo:

$$= \sum_{r=1}^{a} \left[\sum_{S_r} \sum \frac{\pi_{kl}}{\pi_k \pi_l} y_k y_l \right] - \left(\sum_U y_k \right)^2$$

$$= \sum_{r=1}^{a} \left[\sum_{S_r} \sum \frac{\frac{1}{a}}{\frac{1}{a^2}} y_k y_l \right] - \left(\sum_U y_k \right)^2$$

$$= \sum_{r=1}^{a} \left[a \left(\sum_{S_r} y_k \right)^2 \right] - \left(\sum_U y_k \right)^2$$

$$= a \sum_{r=1}^{a} \left[t_{y_{Sr}}^2 \right] - t_{y_U}^2$$, vimos anteriormente q' el estimador insesgado de $t_{y_U} = N \bar{y}_{sr}$, por lo cual

$$= a \sum_{r=1}^{a} \left[t_{y_{Sr}}^2 \right] - (a \bar{y}_U)^2$$

$$= a \sum_{r=1}^{a} \left[t_{y_{Sr}}^2 \right] - \left[a \sum_{r=1}^{a} \frac{y_k}{a} \right]^2 = a \sum_{r=1}^{a} \left[t_{y_{Sr}}^2 \right] - \left[\sum_{r=1}^{a} y_k \right]^2$$