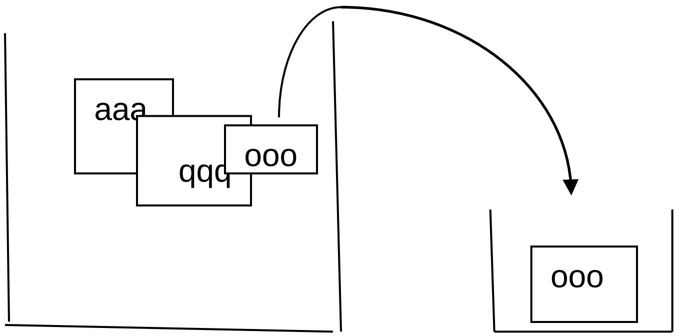
Capítulo 3

DISEÑO SIMPLE

Índice

3.1 Diseño simple sin reposición

Principio



Universo

N

Muestra

n

Probabilidad de pertenecer a la muestra

$$\pi_\alpha = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Tasa de muestreo

$$f = \frac{n}{N}$$

Estimación de la media

- Estimador: $$\hat{\bar{Y}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad E(\bar{y}) = \bar{Y}$$
- Varianza del estimador $$V(\bar{y}) = (1-f)\frac{S^2}{n}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad E(s^2) = S^2$$

- Estimación de la varianza del estimador $$\hat{V}(\bar{y}) = (1-f)\frac{s^2}{n}$$

Se asume la normalidad del estimador

- Intervalo de confianza para $\bar{Y}$ $$\left[\bar{y} - 2\sqrt{\hat{V}(\bar{y})}, \bar{y} + 2\sqrt{\hat{V}(\bar{y})}\right]$$

Estimación del total

- Estimador: $$\hat{T} = N\overline{y} = \sum_{i=1}^{n} \frac{N}{n} y_i = \sum_{i=1}^{n} \frac{1}{\pi_i} y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$$

- Varianza del estimador $$V\left(\hat{T}(Y)\right) = N^2 V\left(\overline{y}\right) = N^2 (1-f) \frac{S^2}{n}$$

- Estimación de la varianza del estimador $$\hat{V}\left(\hat{T}(Y)\right) = N^2 \hat{V}\left(\overline{y}\right) = N^2 (1-f) \frac{s^2}{n}$$

Se asume la normalidad del estimador

- Intervalo de confianza para *T* $$\left[\hat{T} - 2\sqrt{\hat{V}\left(\hat{T}\right)}, \hat{T} + 2\sqrt{\hat{V}\left(\hat{T}\right)}\right]$$

Estimador de Horvitz-Thompson

Estimador de Horvitz-Thompson, π-estimador, estimador de las sumas dilatadas

$$\hat{T} = \frac{N}{n}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} w_i y_i$$

“ponderación” $w_i = N/n$ de cada unidad

La "ponderación” *N/n* de cada unidad de la muestra, permite “extender al universo” el dato observado sobre dicha unidad).

3. 2 Notas

- La precisión depende principalmente de *n* y no de *N*

- Encuesta similar en dos países de tamaño muy distinto: si se desea la misma precisión, se debe escoger un mismo tamaño de muestra en los dos países

- Muestra/ submuestras correspondientes a partes de la población: ¿Problema?

- Varianza del estimador proporcional a 1/n. Desviación tipo, proporcional a $1/\sqrt{n}$: para reducir el intervalo de confianza a la mitad, se debe multiplicar el tamaño de la muestra por 4.

- ¿Cuántas unidades observar para obtener una precisión dada? Lo veremos en problemas <u>MUY IMPORTANTE</u>

3.3 Ejemplo

5 individuos tienen en el bolsillo una suma de dinero $Y_{a.}$

Indiv	Y	$\left(R-\overline{R}\right)^2$
1	100	4
2	80	324
3	100	4
4	120	484
5	90	64

Calculen $\overline{Y}, \sigma^2 y\ S^2$

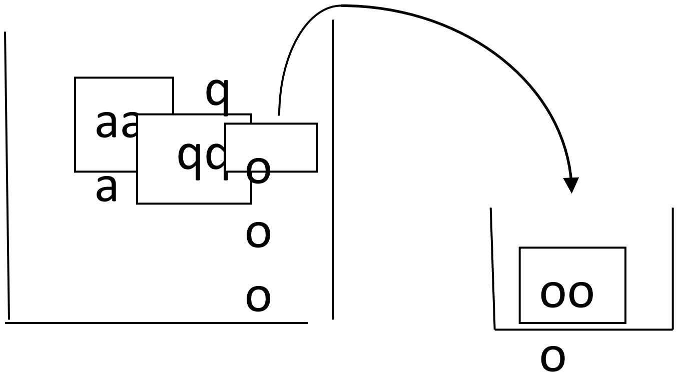
Estudiar el estimador de la media: Muestra de tamaño 2

Comprueben que:

1. la media simple $\bar{y}$ es un estimador sin sesgo de la media poblacional $\bar{Y}$
2. s^2 estima S^2 sin sesgo
3. el estimador de la varianza del estimador de la media es sin sesgo
4. el error cuadrático medio ECM es, en este caso, igual a la varianza del estimador

Muestra	$\bar{y}$	$p(s)$	$(\bar{y}-\bar{Y})^2$	s^2	Intervalo "exacto"	Intervalo estimado
1,2	90	0.1				
1,3	100	0.1				
1,4	110	0.1				
1,5	95	0.1				
2,3	90	0.1				
2,4	100	0.1				
2,5	85	0.1				
3,4	110	0.1				
3,5	95	0.1				
4,5	105	0.1				

3.4 Nota sobre el Muestreo aleatorio simple con reposición



Universo Muestra

N m

Probabilidad de pertenecer a la muestra

$$\Pi_\alpha = 1 - (1-1/N)^m = \mathbf{1-(1-m/N+0(1/m)) \cong m/N}$$

- Estimador: $$\hat{\bar{Y}} = \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

- Varianza del estimador

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

- Estimación de la varianza del estimador

estimación de σ^2 $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$ sin sesgo (ASCR)

estimación de $V(\bar{y})$ $$\hat{V}(\bar{y}) = \frac{s^2}{n}$$ sin sesgo (ASCR)

- Distribución del estimador

Normal, TLC

Intervalo de confianza $$\left[\bar{y} - 2\sqrt{\hat{V}(\bar{y})}, \bar{y} + 2\sqrt{\hat{V}(\bar{y})}\right]$$

3.5 Cuantas unidades observar en la muestra?

¿Cómo escoger n?

a. Fuerte restricción presupuestaria

Sea *C* el presupuesto global para la encuesta, y *c* el coste unitario correspondiente a una entrevista:

$$n = \frac{C}{c}$$

c integra todos los costes de la encuesta

b. Débil restricción de coste

¿Cómo escoger n para satisfacer una precisión prefijada? Recordar:

$$\left[\bar{y}-2\sqrt{(1-f)\frac{\sigma'^2}{n}},\underbrace{\bar{y}+2\sqrt{(1-f)\frac{\sigma'^2}{n}}}_{\Uparrow\ \textbf{\textit{d fijado}}}\right]$$

$$\boxed{\boldsymbol{n}=\boldsymbol{N}\frac{1}{1+\left(\boldsymbol{N}\boldsymbol{d}^2 / 4\boldsymbol{S}^2\right)}}$$

Si se "desprecia" (1-*f*)

$$\boxed{\boldsymbol{n}\approx 4\cdot\frac{\boldsymbol{S}^2}{\boldsymbol{d}^2}}$$

Problema: el valor de S^2 no se conoce antes de la encuesta

3.6 Cuando la media es una proporción

Estimación de la media de la variable indicadora Y:

Variable de interés Y. Para la unidad α

- $Y_\alpha = 1$ si dicha unidad pertenece al dominio;
- $Y_\alpha = 0$ en caso contrario.

media $$P = \frac{\sum_{\alpha} Y_\alpha}{N}$$

total $$N_D = \sum_{\alpha} Y_\alpha$$

$$\sigma^2 = \mathrm{V}(\mathrm{Y}) = \mathrm{P}(1-\mathrm{P})$$

$$S^2 = \cancel{\frac{N}{N-1}} \cdot \mathrm{P}(1-\mathrm{P})$$

- Estimador de la media

$$\hat{p} = \frac{\sum_i y_i}{n}$$

- Varianza del estimador

$$V(\hat{p}) = (1 - f) \cdot \cancel{\frac{N}{N-1}} \cdot \frac{P(1-P)}{n}$$

- Estimación de la varianza del estimador

estimación de σ^2

$$s^2 = \frac{n}{n-1} \cdot \hat{p}(1-\hat{p})$$

estimación de $V(\hat{p})$

$$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \cdot \frac{1}{n} \cdot \frac{n}{n-1} \cdot \hat{p}(1-\hat{p})$$

$$= (1 - f) \cdot \frac{\hat{p}(1-\hat{p})}{n-1}.$$

- Intervalo de confianza

$$\left[\hat{p} - 2\sqrt{\hat{V}(\hat{p})}, \hat{p} + 2\sqrt{\hat{V}(\hat{p})}\right]$$

- Estimador del total del dominio $\hat{N}_D = N.\hat{p}$

- Varianza del estimador

$$V\left(\hat{N}_D\right) = N^2 V(\hat{p}) = N^2 (1-f) \cdot \frac{P(1-P)}{n}$$

- Estimación de la varianza del estimador

$$\hat{V}\left(\hat{N}_D\right) = N^2 \hat{V}(\hat{p}) = N^2 (1-f) \cdot \frac{\hat{p}(1-\hat{p})}{n-1}$$

- Intervalo de confianza

$$\left[\hat{N}_D - 2\sqrt{\hat{V}\left(\hat{N}_D\right)},\ \hat{N}_D + 2\sqrt{\hat{V}\left(\hat{N}_D\right)}\right]$$

Ejemplo de aplicación

Se desea conocer la proporción P de alumnos del instituto AAA que desean proseguir estudios superiores. Hay 1000 alumnos en este instituto.

Entre los 200 alumnos escogidos al azar, se constata que 80 declaran desear proseguir estudios superiores.

Estimar P por punto y por intervalo.

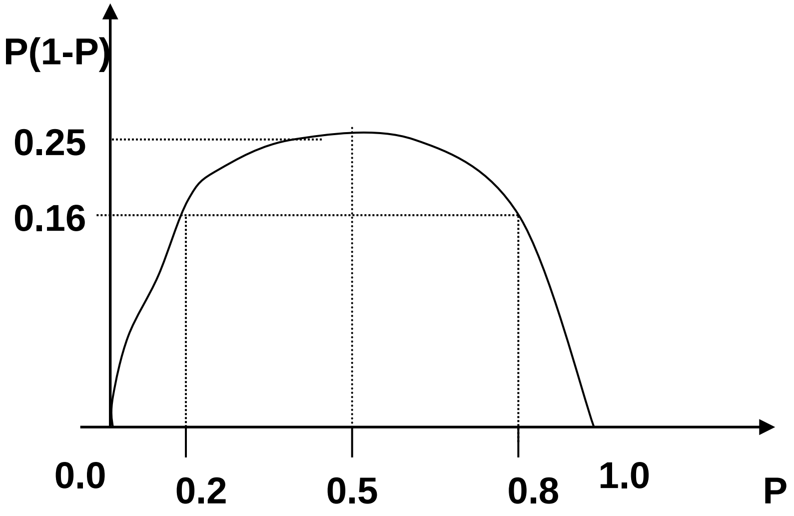
Tamaño de la muestra para estimar una proporción

Precisión absoluta prefijad *d*

Se ha establecido:

$$n = N \frac{1}{1 + \left({Nd^2}/{4S^2} \right)}$$

En caso de una proporción, es bastante usual tener una idea del orden de magnitud de P y, dado que P(1-P) corresponde a la siguiente curva



Precisión absoluta.

$$\boxed{n = N\frac{1}{1+\left(Nd^2 / 4S^2 \right)}}$$

con $S^2 = \cancel{\frac{N}{N-1}} \cdot P(1-P)$

Si se desprecia *(1-f)*

$$\boxed{n = \frac{1}{\left(d^2 / 4S^2 \right)}}$$

Caso particular: peor valor para S^2

$$\boxed{n = N\frac{1}{1+\left(Nd^2 \right)}}$$

con $S^2 = 0.25 = \frac{1}{4}$

Si se desprecia *(1-f)*

$$\boxed{n \approx \frac{1}{d^2}}$$

Ejemplo

¿Cuál será el tamaño de muestra para estimar la proporción de mujeres en Cataluña con una precisión del 2% con una confianza de 95%?

3.7 Métodos de extracción

El problema que interesa aquí es el siguiente: extracción de grandes muestras a partir de marcos muestrales de gran tamaño, disponibles bajo la forma de ficheros informáticos que se desea leer de forma secuencial

Se puede utilizar un generador de números pseudo-aleatorios que simula la extracción de observaciones de una variable aleatoria uniforme entre 0 y 1

Muestreo ASSR

Recordar:

$$\pi_{\alpha} = \frac{n}{N}$$

Objetivos

Un muestreo consiste a extraer, de un universo de N unidades, una muestra de tamaño n que repecte:

- las probabilidades de inclusión π_a
- a veces, que permita calcular las probabilidades de inclusión de segundo orden π_{ab} para poder calcular la varianza del estimador: dado que las observaciones no son independientes intervienen las covarianzas entre las observaciones y por tanto estos π_{ab}

Actualización de las probabilidades

A1 generar u según una ley uniforme sobre (0,1)
A2 Si ($u>n/N$) hacer
  $N:=N-1$
  Avanzar hasta el siguiente registro
  Ir a A1
A3 sino hacer
  Seleccionar el siguiente registro
  $N:=N-1$
  $N:=n-1$
  Si $n=0$, fin (la muestra está completada)
  Si no ir a A1

Actualización de la muestra

B1 Inicializar la muestra con los *n* primeros elementos

B2 Para $a > n$

generar *u* según una ley uniforme sobre (0,1)

Si ($u>n/a$), el registro *a* no está seleccionado

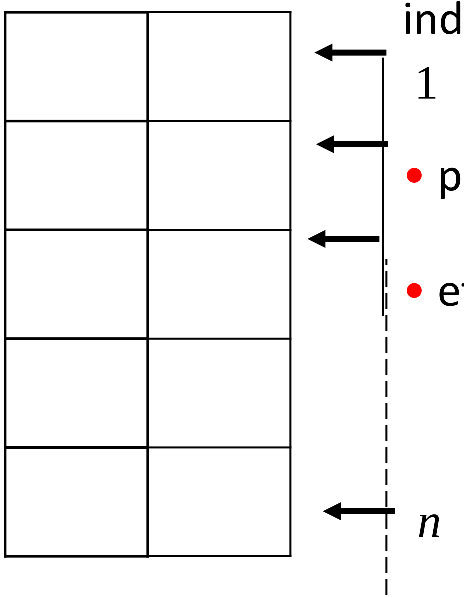
Si no el registro a está seleccionado

y **sustituye una unidad que figuraba** en la muestra: cada uno de los "antiguos" tiene la misma probabilidad de salir de la muestra (por ejemplo, se hace salir a la unidad que ocupa el rango $[n.u] + 1$ en la muestra)

fin para

ASSR mediante una extracción sistemática

Fichero-marco muestral en orden aleatorio



Se divide el fichero en *N/n* bloques y se observa un individuo en cada bloque, de la manera siguiente:

- primero escogido al azar
- efectuar saltos de longitud igual a PASO=*N/n*

Algoritmo

Paso=N/n (no siempre entero)
Extracción al azar de un número *X* entre 0 y 1

Selección del individuo de rango 1 + INT [X*PASO]
Para I variando entre 1 y (n-1) seleccionar los individuos de rango
1 + INT [(X+I).PASO]
Fpara

Así, se obtiene la muestra de tamaño *n*