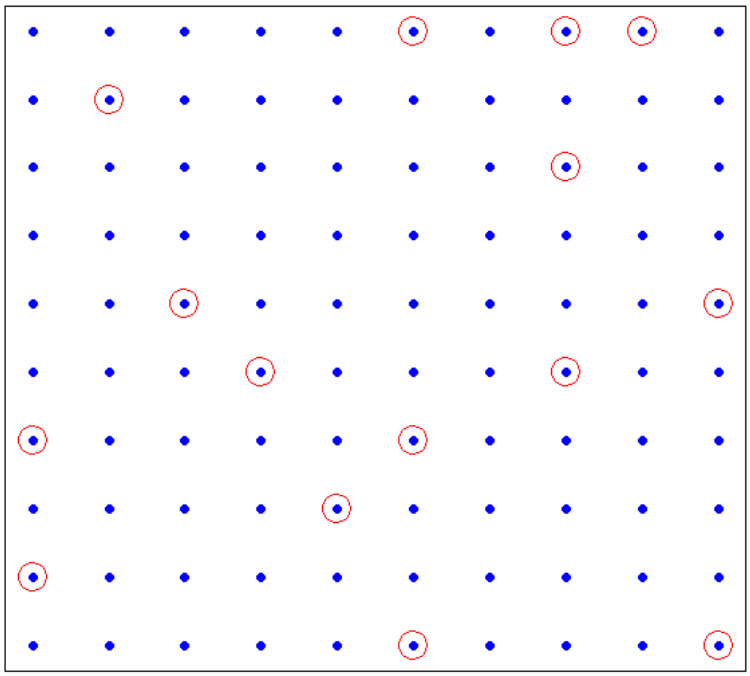
Muestreo Aleatorio Simple

Introducción

El muestreo aleatorio simple (*m.a.s.*) es un diseño muestral muy popular, dadas las propiedades que posee con respecto a las estimaciones de parámetros y de errores de muestreo. Es un diseño de tamaño fijo y exige disponer de un marco poblacional con sus elementos muy bien identificados, por lo que su uso es frecuente junto con otras técnicas.

Se puede usar la librería `animation` para mostrar la idea básica de este tipo de muestreo:

```
# Animación con R
library(animation)
sample.simple(nrow=10, ncol=10, size=15, p.col=c("blue", "red"),
p.cex = c(1,3))
```



Definición del diseño m.a.s. y parámetros asociados

Es un diseño de tamaño fijo n y exige disponer de un marco poblacional con sus elementos bien identificados.

A partir de una población $U = \{u_1, \ldots, u_N\}$ se toma una muestra de tamaño n: $\{u_{i_1}, \ldots, u_{i_n}\}$

Así,

$S = \{$Todos los subconjuntos de tamaño n **sin elementos repetidos**, donde el orden de los elementos **no** importa$\}$

El diseño muestral es $(S, P(\cdot))$ donde

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

En el tema anterior ya se vieron algunas propiedades de este diseño:

– La probabilidad de inclusión de un elemento de la población es:

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

para $k = 1, \ldots, N$, donde f se denomina **fracción de muestreo**.

Del mismo modo,

$$\pi_{kl} = \frac{n(n-1)}{N(N-1)}$$

para $k \neq l$.

$$\pi_{kk} = \pi_k = \frac{n}{N}$$

Por otro lado

$$\begin{aligned} \Delta_{kl} &= Cov(\mathbf{I}_k, \mathbf{I}_l) = \pi_{kl} - \pi_k\pi_l = -\frac{f(1-f)}{N-1} \\ \Delta_{kk} &= \frac{n}{N} - \left(\frac{n}{N}\right)^2 = f(1-f) \end{aligned}$$

– Es un diseño muestral *cuantificable*, π_k y $\pi_{kl} > 0$, y dado que es de tamaño fijo y

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l < 0$$

para $k \neq l = 1, \ldots, N$, se pueden obtener estimadores no negativos del error de muestreo.

En general, una muestra se extrae de forma secuencial, es decir se obtiene elemento a elemento. Aunque con R es trivial hacerlo usando el comando `sample`.

```r
U = c("Manolo", "Luisa", "Pedro", "Eva", "Juan")
N = length(U)

# Mira la ayuda de sample
help(sample)

sam = sample(N, 2, replace=FALSE)
U[sam]
```

```
[1] "Eva"  "Juan"
```

Estimación de parámetros poblacionales

Dada U población finita de tamaño N, consideramos la variable aleatoria en estudio X y deseamos estimar la media y el total poblacional usando $m.a.s.(N, n)$ mediante el estimador de Horvitz-Thompson.

Se tienen los siguientes resultados:

– **Media**

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Ya se vio en el tema anterior que:

i)

$$\widehat{\overline{X}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

es un estimador insesgado de $\overline{X}$

ii) Además,

$$V(\widehat{\overline{X}}) = \frac{1-f}{n} S_X^2$$

donde

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$$

y

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

iii) El estimador de la varianza insesgado para $V(\widehat{\overline{X}})$ es

$$\widehat{V}(\widehat{\overline{X}}) = \frac{1-f}{n}\widehat{S}_X^2$$

tal que

$$\widehat{S}_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\overline{X}})^2$$

es decir, la cuasivarianza muestral.

– **Total**

$$X = \sum_{i=1}^{N} X_i$$

i)

$$\widehat{X} = N\widehat{\overline{X}}$$

es un estimador insesgado de X.

ii)

$$V(\widehat{X}) = N^2\frac{(1-f)}{n}S_X^2$$

iii) Un estimador insesgado de $V(\widehat{X})$ es

$$\widehat{V}(\widehat{X}) = N^2\frac{(1-f)}{n}\widehat{S}_X^2$$

Observación:

En una población infinita la varianza de la media muestral es

$$\frac{\sigma^2}{n} \simeq \frac{S^2}{n}$$

si la población es finita, el único cambio que presenta dicha varianza es la introducción del factor

$$\frac{N-n}{N} = 1 - f < 1$$

Este factor $(1-f)$ se denomina **factor de corrección para poblaciones finitas** (*finite population correction*) **fpc**. Resalta la alteración que se produce al trabajar con poblaciones finitas.

En la varianza de la media muestral se ve que a medida que crece la fracción de muestreo f, disminuye la varianza, o lo que es lo mismo: a medida que aumenta el tamaño muestral la fracción de muestreo crece y disminuye la varianza del estimador de la media poblacional.

En la práctica el *fpc* se puede ignorar siempre y cuando la fracción de muestreo no exceda del 5 % (es decir sea menor que 0.05). Pero, ignorar el *fpc* equivale a sobrestimar el error estándar del estimador de la media poblacional $\overline{X}$.

Intervalos de Confianza

Si se asume que las estimaciones de $\widehat{\overline{X}}$ y de $\widehat{X}$ se distribuyen como una normal entonces los intervalos de confianza al $(1-\alpha)\,\%$ son

I) Para la media $\overline{X}$:

$$\widehat{\overline{X}} \mp z_{\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} \widehat{S}_X^2}$$

donde

$$\widehat{S}_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\overline{X}})^2$$

II) Para el total X:

$$\widehat{X} \mp z_{\frac{\alpha}{2}} \sqrt{\frac{N^2(1-f)}{n} \widehat{S}_X^2}$$

Si el tamaño muestral es menor que 50, es más correcto utilizar una distribución t-Student con $(n-1)$ grados de libertad.

Determinación tamaño muestral n

Una cuestión importante es decidir cuantos elementos deben formar parte de la muestra. Este tamaño muestral dependerá del error que estemos dispuestos a asumir al obtener estimaciones.

Pero además, habrá que tener en cuenta el presupuesto disponible y los costes que se deriven del muestreo.

La forma estándar de proceder es fijar un error de muestreo que deberá ser el mínimo posible, aunque hay varios criterios para fijarlo.

- Fijar $\sigma(\widehat{\theta})$
- Fijar el error máximo admisible
- Fijar el error de muestreo relativo

$$\frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})}$$

Vamos a considerar los casos de la estimación de la media $\overline{X}$ y el total X.

Obtención de n al estimar $\overline{X}$

Fijar $\sigma(\widehat{\theta})$

Sabemos que

$$\sigma\left(\widehat{\overline{X}}\right) = \sqrt{(1-f)\frac{S_X^2}{n}}$$

Si consideramos un error de muestreo e fijo como máximo, debemos encontrar n tal que

$$e = \sqrt{(1-f)\frac{S_X^2}{n}} \Rightarrow \frac{S_X^2}{n} = e^2 + \frac{S_X^2}{N}$$

Despejando

$$n = \frac{S_X^2}{e^2 + \frac{S_X^2}{N}} = \frac{NS_X^2}{Ne^2 + S_X^2}$$

Observaciones:

La expresión obtenida depende del tamaño poblacional N.

Si el tamaño de la población aumenta indefinidamente, en el límite

$$n \underset{N\to\infty}{\longrightarrow} \frac{S_X^2}{e^2}$$

que es un valor constante. Es decir, no hace falta tomar una muestra más grande.

Un problema que se plantea es que, para obtener n, se necesita calcular S_X^2. La solución consiste en usar una muestra *piloto* de tamaño n'.

Fijar el error máximo admisible

El error máximo admisible d es la precisión mínima exigible y es la diferencia en valor absoluto entre el verdadero valor del parámetro y su estimación.

Fijado d y un nivel de confianza $1-\alpha$ se busca el tamaño muestral n tal que en el $(1-\alpha)\,\%$ de las muestras se vaya a obtener un error máximo de d, es decir

$$P\left(|\widehat{\overline{X}} - \overline{X}| \leq d\right) \;\geq\; 1-\alpha \Leftrightarrow$$

$$P\left(\frac{|\widehat{\overline{X}} - \overline{X}|}{e} \leq \frac{d}{e}\right) \;\geq\; 1-\alpha$$

Si asumimos normalidad de $\widehat{\overline{X}}$ se obtiene que

$$\frac{d}{e} = z_{\frac{\alpha}{2}} \Rightarrow d = z_{\frac{\alpha}{2}} e$$

Si consideramos el apartado anterior y observando que fijar un error máximo admisible d es equivalente a fijar un error de muestreo $e = \frac{d}{z_{\frac{\alpha}{2}}}$ el tamaño muestral a utilizar será

$$n = \frac{S_X^2}{e^2 + \frac{S_X^2}{N}} = \frac{S_X^2}{\left(\frac{d}{z_{\frac{\alpha}{2}}}\right)^2 + \frac{S_X^2}{N}} \Rightarrow$$

$$n = \frac{S_X^2 \left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2}{1 + \frac{S_X^2}{N}\left(\frac{z_{\frac{\alpha}{2}}}{d}\right)^2}$$

Fijar un error relativo de muestreo

Sea

$$e_r = \frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})}$$

un error de muestreo dado (equivalentemente coeficiente de variación).

Se trata de buscar n para un valor fijo de e_r:

$$\begin{aligned} e_r &= \frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})} = \frac{\sigma(\widehat{\overline{X}})}{E(\widehat{\overline{X}})} = \frac{\sqrt{(1-f)\frac{S_X^2}{n}}}{\overline{X}} \Rightarrow \\ n &= \frac{\left(\frac{S_X}{\overline{X}}\right)^2}{e_r^2 + \frac{\left(\frac{S_X}{\overline{X}}\right)^2}{N}} \Rightarrow \\ n &= \frac{CV^2}{e_r^2 + \frac{CV^2}{N}} \end{aligned}$$

donde el coeficiente de variación $CV = \frac{S_X}{\overline{X}}$ y se debe estimar a partir de una muestra *piloto*.

Se puede observar que se obtiene la misma expresión que en el caso de fijar $\sigma(\widehat{\theta})$ sustituyendo S_X por CV.

Obtención de n al estimar el total

Fijar $\sigma(\widehat{\theta})$.

Dado que

$$\widehat{X} = N\widehat{\overline{X}}$$

entonces se trata de fijar

$$e\left(\widehat{X}\right) = \sigma\left(\widehat{X}\right) = N\sqrt{(1-f)\frac{S_X^2}{n}}$$

luego aplicando el resultado anterior, se obtiene que

$$n = \frac{N^2\left(\frac{S_X^2}{e\left(\widehat{X}\right)^2}\right)}{1+\left(\frac{S_X^2}{e\left(\widehat{X}\right)^2}\right)N}$$

En este caso se observa que cuando aumenta indefinidamente el tamaño de la población hay que aumentar también el tamaño de la muestra.

Fijando el error máximo admisible

Se fijan d y α tales que

$$P\left(|\widehat{X} - X| \leq d\right) \geq 1-\alpha$$

Se sabe que fijar d equivale a fijar

$$\frac{d}{e} = z_{\frac{\alpha}{2}} \Rightarrow d = z_{\frac{\alpha}{s}}e$$

y se obtiene como en el caso de la media

$$n = \frac{N^2 S_X^2\left(\frac{z_{\frac{\alpha}{s}}}{d}\right)^2}{1+NS_X^2\left(\frac{z_{\frac{\alpha}{s}}}{d}\right)^2}$$

Fijando un error relativo de muestreo

Si fijamos

$$e_r(\widehat{X}) = \frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})} = \frac{\sigma(\widehat{X})}{E(\widehat{X})} = \frac{N\sigma(\widehat{\overline{X}})}{NE(\widehat{\overline{X}})} = e_r(\widehat{\overline{X}})$$

Es decir, coincide con el caso de la estimación de la media

Observaciones

En general los pasos a seguir para elegir el tamaño de la muestra pueden resumirse en:

i) Se debe preguntar: ¿Cuánta precisión se necesita? ¿en términos de qué se define una precisión?

II) Determinar una ecuación que relacione el tamaño de la muestra n y las expectativas que se tienen con respecto a la muestra.

III) Estimar normalmente mediante una muestra piloto las cantidades desconocidas para determinar n.

IV) Si el tamaño muestral obtenido es demasiado grande, replantear la precisión inicial fijada.

Muestreo aleatorio simple con reemplazamiento

Un muestreo aleatorio simple con reemplazamiento es simplemente un m.a.s.(N, n) en el que devolvemos, en cada extracción, el elemento obtenido al conjunto original.

El m.a.s. con reemplazamiento consiste en seleccionar en cada extracción un elemento de U con probabilidad constante igual a $\frac{1}{N}$, es decir

$$p_k = \frac{1}{N},$$

para todo $k = 1, \ldots, N$.

Si consideramos las muestras ordenadas de tamaño n, el espacio muestral está formado por N^n muestras, todas ellas equiprobables:

$$P(s) = \frac{1}{N^n}$$

para todo $s \in S$ donde $S = \{\text{Muestras ordenadas de } n \text{ elementos con reposición}\}$

En una muestra de tamaño n un elemento dado puede aparecer un cierto número de veces.

Si denominamos $N_r =$ número de veces que aparece en una muestra la unidad r

Entonces

$$P\{N_r = k\} = \binom{n}{k} \left(\frac{1}{N}\right)^k \left(1 - \frac{1}{N}\right)^{n-k}$$

Así $N_r \sim Bin(n, \frac{1}{N})$.

Las probabilidades de inclusión (que está al menos en una muestra s) son

$$\begin{aligned} P\{k \in s\} &= \pi_k = 1 - P\{k \notin s\} = \\ &\quad 1 - P\{\text{en todas las extracciones no salga } k\} \\ &= 1 - \left(1 - \frac{1}{N}\right)^n \end{aligned}$$

para $k = 1, \ldots, N$.

Por otro lado, para $k \neq l = 1, \ldots, N$

$$\pi_{kl} = P\{k, l \in s\} = 1 - P\{k \text{ ó } l \notin s\} =$$

$$1 - [P\{k \notin s\} + P\{l \notin s\} - P\{k, l \notin s\}] =$$

$$1 - [2(1 - \pi_k) - P\{\text{en las } n \text{ extracciones no salgan } k, l\}] =$$

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

$$1 - 2\left(1 - \frac{1}{N}\right)^n + \left(1 - \frac{2}{N}\right)^n$$

Estimadores de $\overline{X}$ y X

Basándonos en el estimador de Hansen-Hurwitz se obtiene que:

Estimación de $\overline{X}$:

Media muestral

$$\widehat{\overline{X}} = \frac{1}{n}\sum_{k \in s} X_k$$

$$V\left(\widehat{\overline{X}}\right) = \frac{1}{n}\sum_{k=1}^{N}\left(\frac{\frac{1}{N}X_k}{\frac{1}{N}} - \overline{X}\right)^2 \frac{1}{N} =$$

$$\frac{1}{nN}\sum_{k=1}^{N}\left(X_k - \overline{X}\right)^2 = \frac{1}{n}\sigma_X^2 = \frac{1}{n}\frac{N-1}{N}S_X^2$$

donde σ_X^2 es la varianza poblacional y S_X^2 es la cuasivarianza poblacional.

El estimador es

$$\widehat{V}\left(\widehat{\overline{X}}\right) = \frac{1}{n(n-1)}\sum_{k \in s}\left(X_k - \widehat{\overline{X}}\right)^2 = \frac{1}{n}\widehat{S}_X^2$$

donde $\widehat{S}_X^2$ es la cuasivarianza muestral.

Estimación del total X:

Como $X = N\overline{X}$, entonces

$$\widehat{X} = N\widehat{\overline{X}}$$

$$V\left(\widehat{X}\right) = \frac{N^2}{n}\sigma_X^2$$

$$\widehat{V}\left(\widehat{X}\right) = \frac{N^2}{n}\widehat{S}_X^2$$

Comparación entre el muestreo m.a.s con y sin reposición

La pregunta que se plantea es cuál de los dos tipos de muestreo es más preciso, o tiene menor error.

Si se comparan las varianzas

$$\frac{V_{masr}\left(\widehat{\overline{X}}\right)}{V_{mas}\left(\widehat{\overline{X}}\right)} = \frac{\frac{1}{n}\frac{N-1}{N}S_X^2}{\frac{1}{n}(1-f)S_X^2} = \frac{1-\frac{1}{N}}{1-f} \simeq \frac{1}{1-f}$$

dado que N es grande.

Así, si f es pequeña las dos muestras son igual de eficientes, es decir, si se muestrean pocos elementos de una población grande, es lo mismo elegirlos con o sin reemplazamiento.

Si f es grande, entonces $V_{masr}\left(\widehat{\overline{X}}\right) \geq V_{mas}\left(\widehat{\overline{X}}\right)$.

Por ejemplo, si $f = 0{,}5$ entonces $V_{masr}\left(\widehat{\overline{X}}\right) \approx 2V_{mas}\left(\widehat{\overline{X}}\right)$.

Ejemplo con R

```
# Generas unos datos artificiales
set.seed(666)
unosdatos = rbind(matrix(rep("nc",165),165,1,byrow=TRUE),
                  matrix(rep("sc",70),70,1,byrow=TRUE))
genero = rbinom(235, 1, 0.43)
genero[genero==1]="M"
genero[genero==0]="H"
unosdatos = cbind.data.frame(unosdatos,c(rep(1,100),rep(2,50),
            rep(3,15),rep(1,30),rep(2,40)),genero,100*runif(235))

dim(unosdatos)
names(unosdatos) = c("provincia", "region", "genero", "ingresos")
head(unosdatos)
```

```
  provincia region genero  ingresos
1        nc      1      M 61.050833
2        nc      1      H 41.378524
3        nc      1      M 63.162693
4        nc      1      H  2.017675
5        nc      1      H 56.045284
6        nc      1      M 21.955715
```

```
N = dim(unosdatos)[[1]]
n = 50

# Selección de una muestra aleatoria
srs_cuales = sample(N,n)
srs = unosdatos[srs_cuales, ]

library(survey)
srs$popcuanto = N
srs$pesos = N/n

dsrs = svydesign(id=~1, fpc=~popcuanto, data=srs)
summary(dsrs)
```

```
Independent Sampling design
svydesign(id = ~1, fpc = ~popcuanto, data = srs)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2128  0.2128  0.2128  0.2128  0.2128  0.2128
Population size (PSUs): 235
Data variables:
[1] "provincia" "region" "genero" "ingresos" "popcuanto" "pesos"
```

```
svytotal(~ingresos, dsrs, na.rm=TRUE)
```

```
          total     SE
ingresos 14180 810.37
```

```
svymean(~ingresos , dsrs , na.rm=TRUE)
```

```
           mean     SE
ingresos 60.339 3.4484
```

```
svyvar(~ingresos , dsrs , na.rm=TRUE)
```

```
          variance     SE
ingresos   755.25 99.552
```

```
svyquantile(~ingresos , quantile=c(0.25,0.5,0.75), design=dsrs ,
na.rm=TRUE, ci=TRUE)
```

```
$quantiles
             0.25      0.5     0.75
ingresos 40.50154 65.42247 84.27074

$CIs
, , ingresos

           0.25      0.5     0.75
(lower 31.15592 46.49287 76.89246
upper) 44.89626 76.53099 88.61348
```

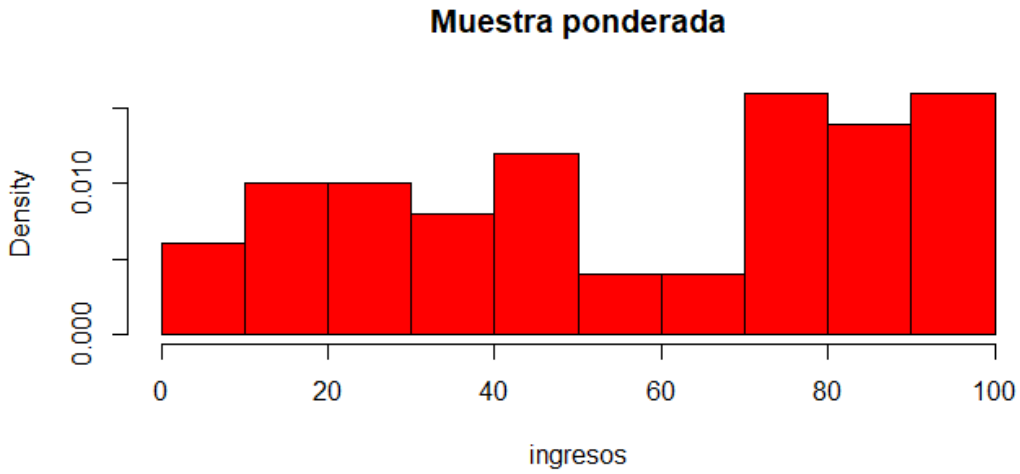
```
confint(svytotal(~ingresos , dsrs , na.rm=TRUE))
```

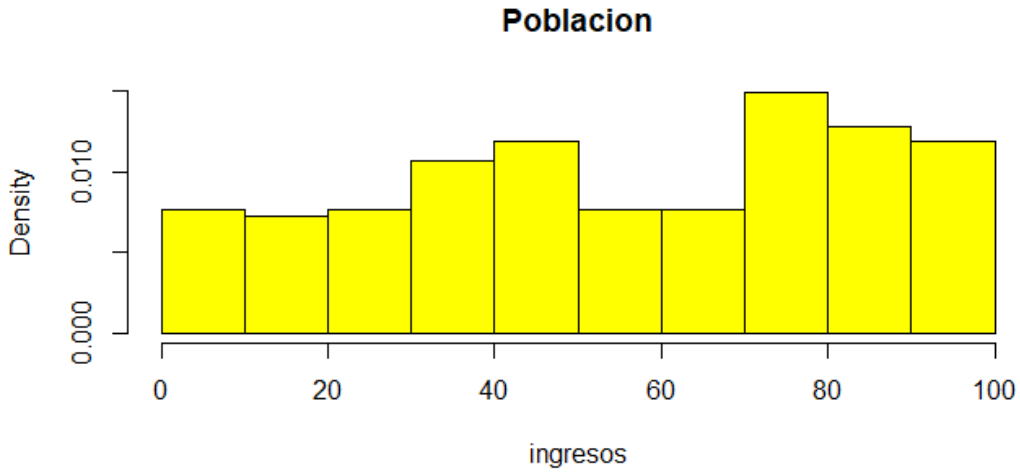
```
            2.5 %   97.5 %
ingresos 12591.27 15767.84
```

```
confint(svymean(~ingresos , dsrs , na.rm=TRUE))
```

```
            2.5 %   97.5 %
ingresos 53.57988 67.09721
```

```
par(mfrow=c(2,1))
svyhist(~ingresos , dsrs , main="Muestra ponderada", col="red")
hist(unosdatos$ingresos , main="Poblacion", xlab="ingresos",
col="yellow", prob=TRUE)
```





Si no se especifica el tamaño de la población, es necesario especificar las probabilidades de muestreo o pesos de muestreo.

La variable `pesos` en el conjunto de datos contiene el peso muestral que es igual a $235/50 = 4{,}7$.

El efecto de omitir el tamaño de la población aparece como `(with replacement)` en la salida. La media estimada y el total son los mismos, pero los errores estándar son un poco más grandes que en el caso de *sin reemplazamiento*.

```
dsrsR = svydesign(id=~1, weights=~srs$pesos, data=srs)
summary(dsrsR)
```

```
Independent Sampling design (with replacement)
svydesign(id = ~1, weights = ~srs_pesos, data = srs)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2128  0.2128  0.2128  0.2128  0.2128  0.2128
Data variables:
[1] "provincia" "region" "genero" "ingresos" "popcuanto" "pesos"
```

```
svytotal(~ingresos, dsrsR)
```

```
          total     SE
ingresos 14180 913.33
```

```
svymean(~ingresos, dsrsR)
```

```
            mean     SE
ingresos 60.339 3.8865
```

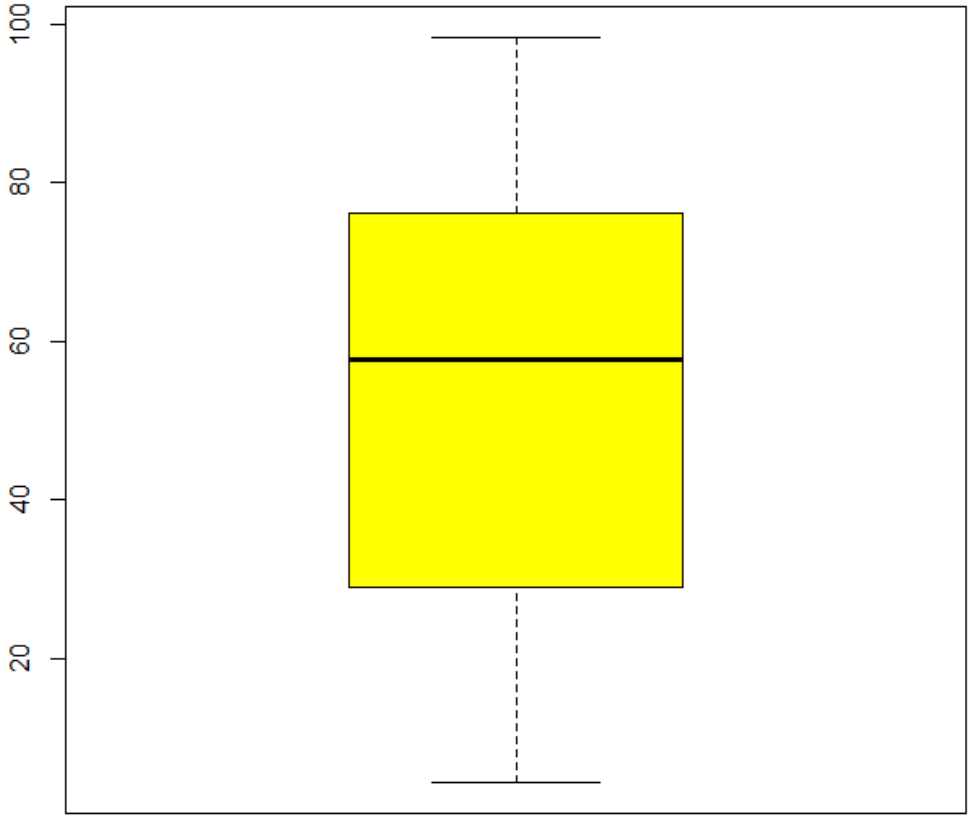
```
confint(svytotal(~ingresos, dsrsR, na.rm=TRUE))
```

```
             2.5 %   97.5 %
ingresos 12389.46 15969.66
```

```
confint(svymean(~ingresos, dsrsR, na.rm=TRUE))
```

```
             2.5 %   97.5 %
ingresos 52.72111 67.95599
```

```
svyboxplot(ingresos~1, design=dsrsR, col="yellow")
```



Introducción al estudio de variables cualitativas

En una población finita es frecuente querer estimar la proporción de unidades que poseen cierta característica, o el número total de unidades de la población que la poseen.

Por ello se puede definir para todo $i = 1, \ldots, N$

$$\mathbf{I}_i = \begin{cases} 1 & \text{si} \quad i \text{ posee la característica} \\ 0 & \text{si} \quad i \text{ no posee la característica} \end{cases}$$

Así los parámetros de interés se pueden escribir como

$$P = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}_i \qquad T = \sum_{i=1}^{N} \mathbf{I}_i$$

dado que P es simplemente una $\overline{X}$ de variables dicotómicas.

Para calcular los estimadores y sus varianzas, entonces se aplican las fórmulas de los estimadores de Horwitz-Thompson.

Así, la proporción de valores muestrales

$$\widehat{P} = \frac{1}{n}\sum_{k\in s} \mathbf{I}_k$$

y como $\widehat{T} = N\widehat{P}$

$$\widehat{T} = \frac{N}{n}\sum_{k\in s} \mathbf{I}_k$$

Sus varianzas respectivas son

$$V\left(\widehat{P}\right) = \frac{(N-n)}{N-1}\frac{P(1-P)}{n}$$

$$V\left(\widehat{T}\right) = \frac{N^2(N-n)}{N-1}\frac{P(1-P)}{n}$$

Demostración:

Como

$$V\left(\widehat{\overline{X}}\right) = \frac{1-f}{n}S_X^2$$

donde

$$S_X^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_k - \overline{X}\right)^2 =$$

$$\frac{1}{N-1}\left[\sum_{k=1}^{N} X_k^2 - N\underbrace{\overline{X}^2}_{=P^2}\right] = \frac{1}{N-1}\left[NP - NP^2\right] = \frac{NP(1-P)}{N-1}$$

de modo que

$$V(\widehat{P}) = \frac{\frac{N-n}{N}}{n} \frac{NP(1-P)}{N-1} = \frac{(N-n)}{n(N-1)} P(1-P)$$

Del mismo modo se demuestra, como $\widehat{T} = N\widehat{P}$, la expresión para $\widehat{T}$.

De manera análoga, sustituyendo las expresiones de $\widehat{\overline{X}}$ por $\widehat{P}$ en las fórmulas generales de la media, se pueden obtener los resultados para los estimadores de las varianzas:

$$\widehat{V}\left(\widehat{P}\right) = \frac{(1-f)}{n-1} \widehat{P}(1-\widehat{P})$$

$$\widehat{V}\left(\widehat{T}\right) = N^2 \frac{(1-f)}{n-1} \widehat{P}(1-\widehat{P})$$

donde

$$\widehat{P} = \frac{1}{n} \sum_{k \in s} \mathbf{I}_k$$

Muestreo con reemplazamiento

En este caso, sustituyendo las expresiones de $\widehat{\overline{X}}$ por $\widehat{P}$ en las fórmulas generales de la media, para muestreo con reemplazamiento, se obtienen los resultados para las varianzas:

$$V\left(\widehat{P}\right) = \frac{P(1-P)}{n}$$

$$V\left(\widehat{T}\right) = N^2 \frac{P(1-P)}{n}$$

y los estimadores de las varianzas:

$$\widehat{V}\left(\widehat{P}\right) = \frac{1}{n-1} \widehat{P}(1-\widehat{P})$$

$$\widehat{V}\left(\widehat{T}\right) = N^2 \frac{1}{n-1} \widehat{P}(1-\widehat{P})$$

donde

$$\widehat{P} = \frac{1}{n} \sum_{k \in s} \mathbf{I}_k$$

Determinación del tamaño muestral

Se trata ahora de determinar el tamaño muestral necesario cuando se fija el error de muestreo e o para obtener una precisión determinada d.

Fijando el error de muestreo $\sigma(\widehat{\theta}) = e$

En este caso las variables X_i son dicotómicas por lo que

$$S_X^2 = \frac{NP(1-P)}{N-1}$$

Sustituyendo en la fórmula general y tomando S_X^2 en lugar de su estima $\widehat{S}_X^2$, dado que depende solo de P:

$$n = \frac{NS_X^2}{Ne^2 + S_X^2} \Rightarrow$$

$$n = \frac{NP(1-P)}{(N-1)e^2 + P(1-P)}$$

Del mismo modo para estimar el total, aplicando la fórmula general, se obtiene

$$n = \frac{N^3P(1-P)}{(N-1)e^2 + N^2P(1-P)}$$

Se puede observar que para calcular n se debería conocer P.

Pero es inmediato comprobar (calculando el máximo de la función, tomado derivadas) que el valor de n es máximo cuando se toma $P = \frac{1}{2}$. Lo cual lleva a una estima de n *conservadora*.

Ejemplo

En un distrito donde hay 4000 casas se quiere estimar el porcentaje de propietarios de 2 vehículos, con un error de muestreo no mayor que 0.01.

El porcentaje verdadero de propietarios de dos vehículos se piensa que está entre el 5 % y 10 %. ¿Cuál debe ser el tamaño de la muestra para conseguir el error de muestro fijado?

Sea $U =$ "4000 casas" $N = 4000$

Se quiere estimar una proporción P. Para este caso, se tiene que

$$n = \frac{NP(1-P)}{e^2(N-1) + P(1-P)}$$

Para estimar P nos basamos en la información recogida, en lugar de tomar $P = 1/2$ ya que se conseguiría una n demasiado grande. Como se sabe que $0{,}05 < P < 0{,}10$ entonces

$$n = \frac{4000 \cdot 0{,}1 \cdot 0{,}9}{0{,}01^2 \cdot 3999 + 0{,}1 \cdot 0{,}9} = 734{,}8$$

Es decir, se requieren 735 casas. ■

Fijando el error máximo admisible d

Por otro lado, si queremos fijar un error máximo admisible d, usando la fórmula general, se obtiene que

$$n = \frac{NP(1-P)z^2_{\frac{\alpha}{2}}}{P(1-P)z^2_{\frac{\alpha}{2}} + (N-1)d^2}$$

Dado que $P \in (0,1)$ el máximo de la expresión se alcanza en $P = 1/2$, de lo que se deduce que $P(1-P) \leq \frac{1}{4}$.

Luego

$$n = \frac{Nz^2_{\frac{\alpha}{2}}}{z^2_{\frac{\alpha}{2}} + 4(N-1)d^2}$$

Si se busca el valor de n, correspondiente al total, se obtiene, entonces,

$$n = \frac{N^3P(1-P)z^2_{\frac{\alpha}{2}}}{N^2P(1-P)z^2_{\frac{\alpha}{2}} + (N-1)d^2}$$

Intervalos de Confianza

Para construir intervalos de confianza para P se puede usar la aproximación a la normal (cuando n es grande)

$$\widehat{P} \sim N\left(P, \sqrt{\frac{1-f}{n}\frac{N}{N-1}P(1-P)}\right)$$

de modo que el intervalo de confianza es

$$\left[\widehat{P} \mp z_{\frac{\alpha}{2}}\sqrt{\widehat{V}(\widehat{P})}\right] =$$

$$\left[\widehat{P} \mp z_{\frac{\alpha}{2}}\sqrt{\frac{1-f}{n-1}\widehat{P}(1-\widehat{P})}\right]$$

Intervalos de confianza simultáneos para proporciones para características no dicotómicas

Supongamos que se quiere estudiar una variable cualitativa que presenta k modalidades $(k > 2)$

Sea P_j la proporción de individuos de la población con la modalidad j tal que $j = 1, \ldots, k$.

Se quieren estimar dichas proporciones y construir intervalos de confianza. Se utiliza un muestreo aleatorio con reemplazamiento (para simplificar los cálculos) de tamaño n.

Sea n_j el número de elementos de la muestra donde se obtiene la respuesta j y sea

$$\widehat{P}_j = \frac{n_j}{n}$$

la proporción muestral de la respuesta j.

Entonces como se tiene un muestreo con reemplazamiento

$$n_j \sim B(n, P_j)$$

y

$$E\left(\widehat{P}_j\right) = E\left(\frac{n_j}{n}\right) = \frac{nP_j}{n} = P_j$$

$$V\left(\widehat{P}_j\right) = \frac{1}{n^2}nP_j(1-P_j) = \frac{P_j(1-P_j)}{n}$$

siendo un estimador insesgado de la varianza

$$\widehat{V(\widehat{P}_j)} = \frac{1}{n-1}\widehat{P}_j(1-\widehat{P}_j)$$

Utilizando los intervalos de confianza simultáneos de *Bonferroni* con un nivel de confianza $1-\alpha$:

$$\left[\widehat{P}_j \mp z_{\frac{\alpha}{2k}}\sqrt{\frac{1}{n-1}\widehat{P}_j(1-\widehat{P}_j)}\right]$$

Si fijamos un error máximo admisible d para cada P_j, el tamaño de la muestra viene dado por:

$$n \simeq z^2_{\frac{\alpha}{2k}}\frac{\text{máx}\left\{\widehat{P}_j(1-\widehat{P}_j)\right\}}{d^2}$$

y como para todo j

$$P_j(1-P_j) \leq \frac{1}{4}$$

siendo $P_j \in (0,1)$ se obtiene que

$$n \simeq \frac{z^2_{\frac{\alpha}{2k}}}{4d^2}$$

Observación: Si $k=2$ no son válidos los cálculos anteriores.

Ejemplo

En una población de 1000 personas mayores de 25 años se desea estimar las proporciones de personas solteras, emparejadas y en otras circunstancias.

Dada una *m.a.s.* con reemplazamiento de 500 personas se obtiene 355 emparejadas, 112 solteras y 33 en otro estado.

a) Estimar las proporciones y construir I.C. simultáneos al 95 %

b) Obtener n tal que se obtenga un error máximo admisible del 4 %

a)

$$\hat{p}_1 = \frac{355}{500} = 0{,}71; \quad \hat{p}_2 = \frac{112}{500} = 0{,}224; \quad \hat{p}_3 = \frac{33}{500} = 0{,}0066$$

Siendo

$$z_{\frac{\alpha}{3\cdot 2}} = z_{\frac{0{,}05}{6}} = z_{0{,}0083} = 2{,}40$$

Entonces

$$IC_1 = \left[\frac{355}{500} \mp 2{,}40\frac{\sqrt{\frac{355}{500}\cdot\frac{145}{500}}}{499}\right] = [0{,}6; 0{,}75]$$

$$IC2_1 = \left[\frac{112}{500} \mp 2{,}40\frac{\sqrt{\frac{112}{500}\cdot\frac{388}{500}}}{499}\right] = [0{,}179; 0{,}268]$$

$$IC_1 = \left[\frac{33}{500} \mp 2{,}40\frac{\sqrt{\frac{33}{500}\cdot\frac{467}{500}}}{499}\right] = [0{,}039; 0{,}092]$$

b)

$$n = \frac{2{,}40^2}{4 \cdot 0{,}04^2} \approx 900$$

Ejemplo con R

```
prop.table(svytable(~genero, design=dsrs))
```

```
genero
   H    M
0.56 0.44
```

Alternativamente se puede escribir:

```
svymean(~genero, design=dsrsR)
```

```
        mean     SE
generoH 0.56 0.0709
generoM 0.44 0.0709
```