UNTREF

Teoría y Técnicas de Muestreo
Trabajo 1

Miller, Manuel
Parrino Augusto

Repositorio:
https://github.com/miller862/Tecnicas_de_Muestreo

Fecha de entrega: 11/11/2024

Para replicabilidad, se utilizó el seed 999

Ejercicio I

Se desea estimar mediante una Muestra Aleatoria Simple de 9 alumnos la media de X, la media de Y y la
razón R = $\bar{Y} / \bar{X}$, mediante los estimadores usuales

Selección de una muestra aleatoria de 9 casos

```
###MAS
N=nrow(df)
k=9
s_mas <- df[sample(nrow(df), size = 9, replace = FALSE), ]
```

Media de x

```
x_media= mean(s_mas$X)
> x_media
7
```

Media de y

```
> y_media= mean(s_mas$Y)
> y_media
13.88889
```

Media de r

```
 r=y_media/x_media
 r
 1.984127
```

¿Cuántas muestras posibles hay?

```
#Alternativa 1
choose(18, 9)

48620

#Alternativa 2
sposibles <- factorial(N) / (factorial(k) * factorial(N - k))
posibles
```

48620

• Hallar la varianza y CV de los estimadores $\bar{y}$ y $\bar{x}$

Varianza del estimador de X media

```
varMAS_xmedia <- (1-k/N)*VARX/k
[1] 0.5455701

varMAS_x <- N^2*varMAS_xmedia   #Recordar: Var(k*X)= k^2*Var(X)
[1] 176.7647
```

CV del estimador de X media

```
CVMAS_xmedia <- 100*sqrt(varMAS_x)/N
[1] 73.86272
```

Varianza del estimador de Y media

```
 varMAS_ymedia <- (1-k/N)*VARY/k
[1] 2.147467
> varMAS_y <- N^2*varMAS_ymedia   #Recordar: Var(k*X)= k^2*Var(X)
[1] 695.7794
```

CV del estimador de Y media

```
> CVMAS_ymedia <- 100*sqrt(varMAS_y)/N
[1] 146.5424
```

• Calcular la varianza aproximada y CV aproximado del estimador r

Varianza de r

```
VARr <- r^2 * (varMAS_ymedia / (y_media^2) + varMAS_xmedia / (x_media^2))
[1] 0.08765808
```

CV de r

```
CVr <- (sqrt(VARr) / r) * 100
[1] 14.92198
```

• Seleccionando 10,000 muestras aleatorias simples, estimar la varianza y CV de r. Comparar con el resultado obtenido en el punto anterior.

Creación de la función

```
estimo_R_MAS <- function(x){
 muestra <- df[sample(nrow(df),x, replace=FALSE),]
 x_media <- mean(muestra$X)
 y_media <- mean(muestra$Y)
 r=y_media/x_media
varMAS_xmedia <- (1-k/N)*VARX/k
varMAS_x <- N^2*varMAS_xmedia
varMAS_ymedia <- (1-k/N)*VARY/k
varMAS_y <- N^2*varMAS_ymedia
 VARr <- r^2 * (varMAS_ymedia / (y_media^2) + varMAS_xmedia / (x_media^2))
CVr <- (sqrt(VARr) / r) * 100

  a <- c(x,x_media,y_media,r,VARr,CVr)
  return(a)
}

mean(df_estimNUEVO$x_media)
mean(df_estimNUEVO$y_media)
mean(df_estimNUEVO$r)
vargeneral<-var(df_estimNUEVO$r)
cvgeneral<-(sqrt(vargeneral) / mean(r)) * 10
```

Cálculo para las 10000 muestras

```
lista <- rep(9,10000)
lista_estim <- lapply(lista, estimo_R_MAS)

df_estim10000 <- data.frame(matrix(unlist(lista_estim),
                        nrow=length(lista_estim), byrow=TRUE))
```

Head (df_estim10000)

	n	x_media	y_media	r	VARr	CVr
1	9	6.444444	13.61111	2.112069	0.11030750	15.72514
2	9	7.333333	13.61111	1.856061	0.07488102	14.74327
3	9	7.444444	15.27778	2.052239	0.08021032	13.80026
4	9	7.333333	14.05556	1.916667	0.07720065	14.49652
5	9	8.777778	17.00000	1.936709	0.05443021	12.04636
6	9	8.444444	16.11111	1.907895	0.05796459	12.61906

Comparación

	10000 muestras	muestra aleatoria	Parámetro poblacional
X media	6.955922	7	6.944444
Y media	13.43949	13.88889	13.41667
r	1.933213	1.984127	1.932
VARr	0.00883059	0.08765808	-
CVr	4.73615	14.92198	-

Ejercicio II

La tabla tabla_muestras_posibles.xlsx contiene 20 unidades, a las que se le midieron una variable Y. Será
nuestro universo/marco de muestreo. Se desea estimar la media de Y mediante una MAS(10). Se compararon
estos estimadores:

• Media muestral
• Media muestral truncada, eliminando 10% inferior y 10% superior (en nuestro caso resulta el menor
valor de la muestra y el mayor valor de la muestra)
• Mediana

1. Listar con R todas las muestras posibles y calcular para cada una de ellas media, media truncada y
mediana y 2. Agregar a cada muestra la media, media truncada y mediana de los diez valores.

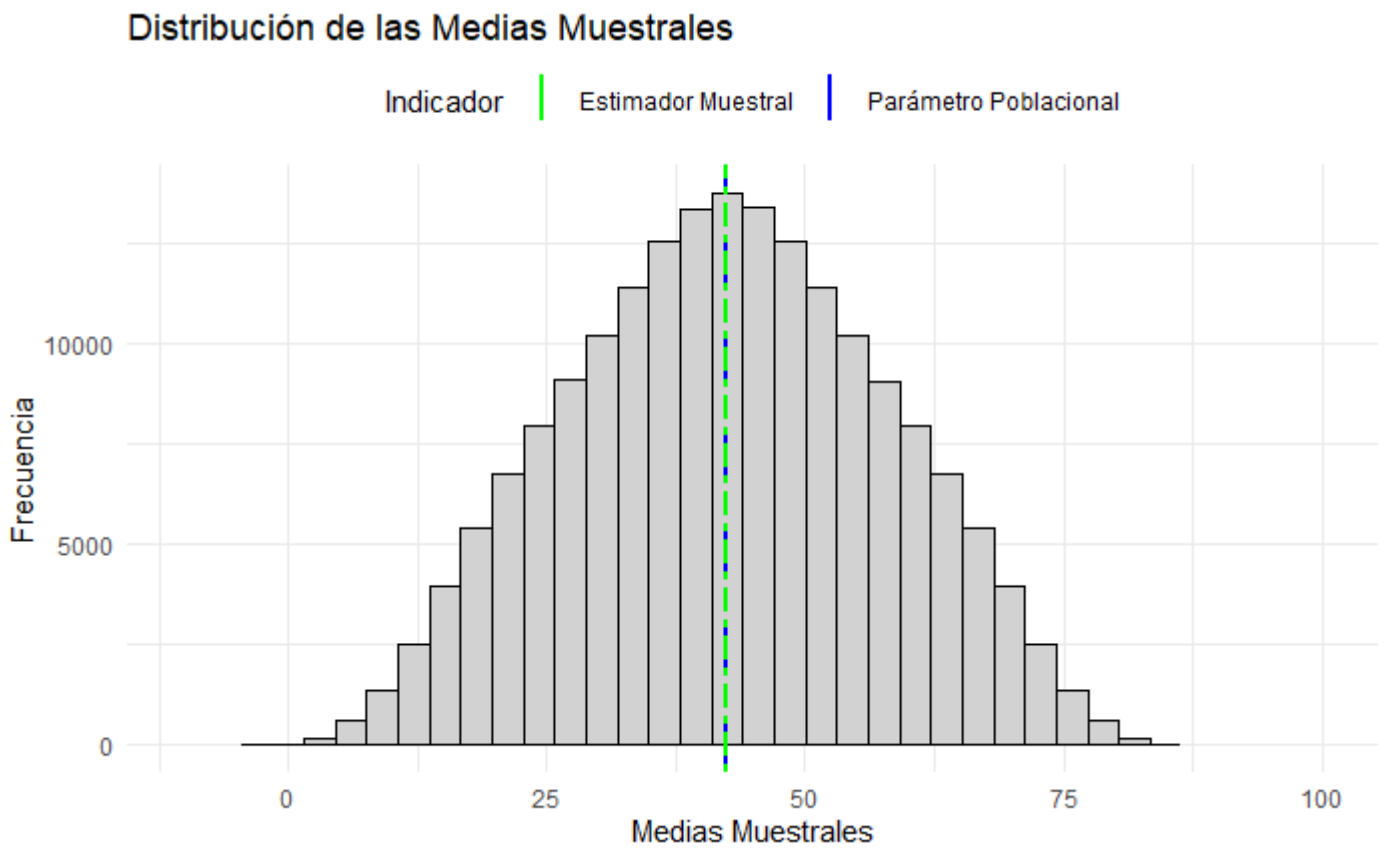
- Pueden verse todas las muestras posibles en el repositorio

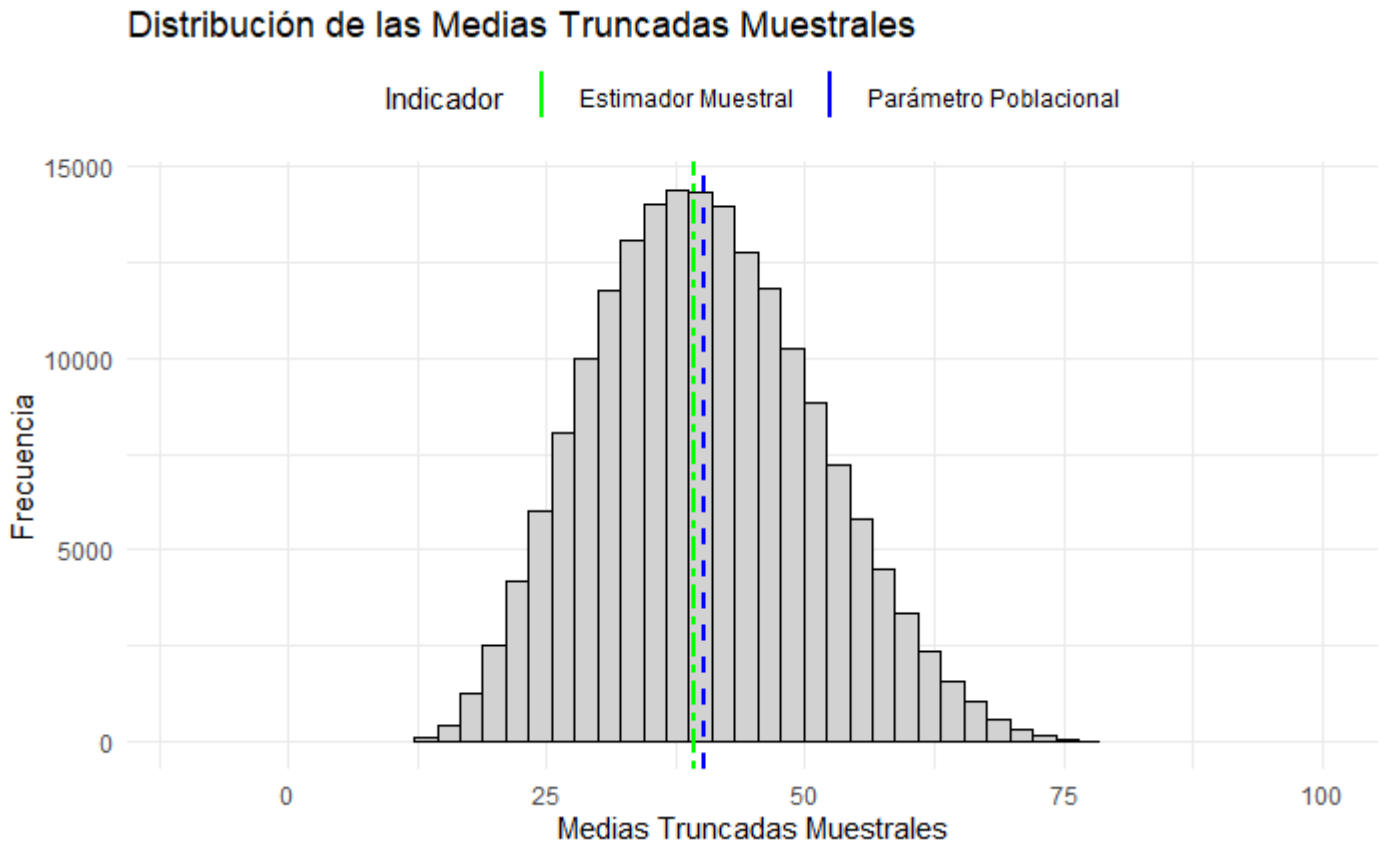
Y media	Y mediana	Y media truncada
Y_mean=mean(df_tabla$Y)	Y_median<-median(df_tabla$Y)	Y_truncmean<-mean(sort(df_tabla$Y)[3:(length(df_tabla$Y) - 2)])
42.435	28.37879	39.34896

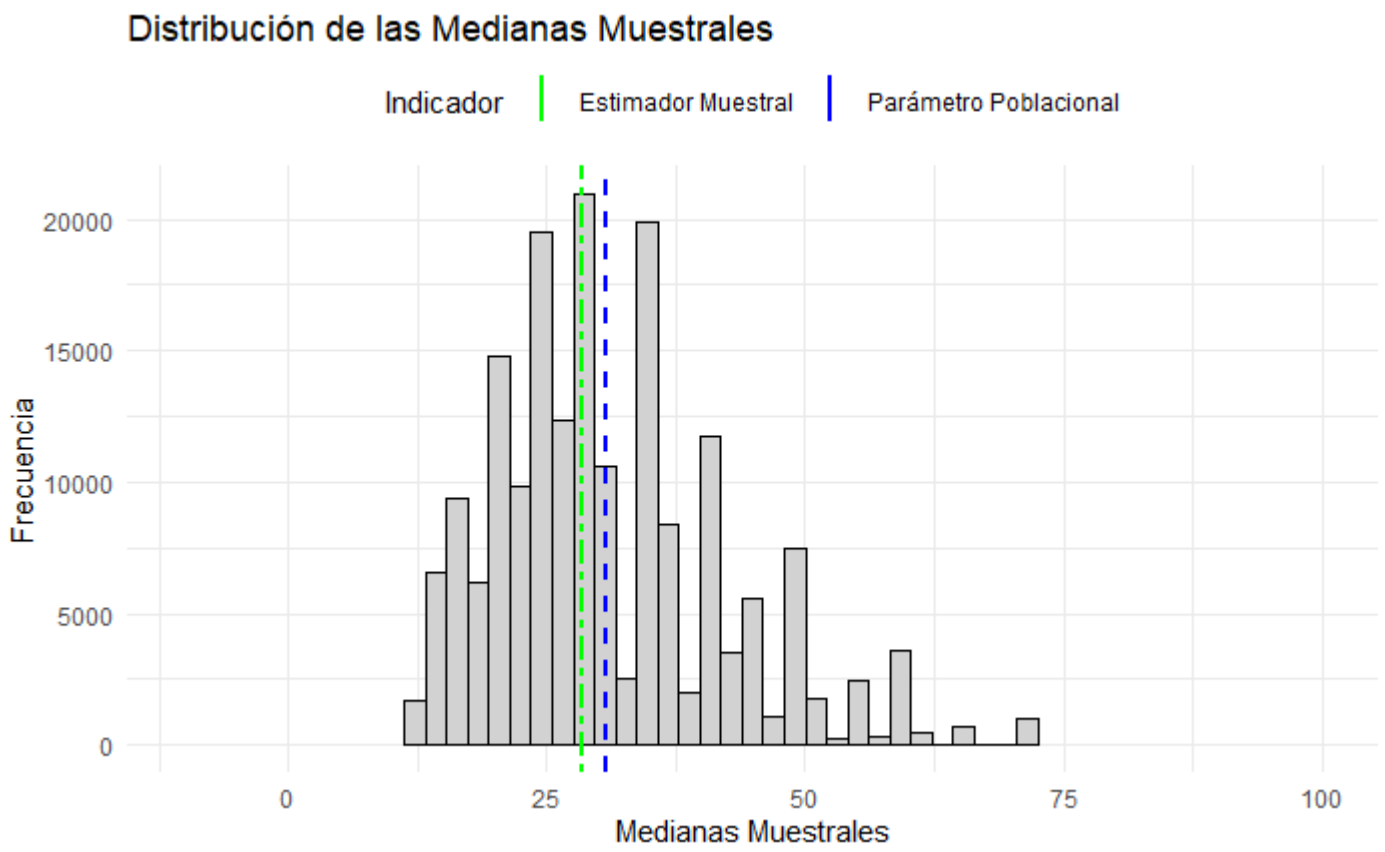
3. Verificar que la media muestral es un estimador insesgado de la media poblacional, lo que no se cumple para la mediana y la media truncada

Media
mean(df_muestras$Media)== Y_mean #Resultado TRUE --> ES INSESGADO [1] TRUE
Mediana
mean(df_muestras$Mediana)== Y_median #Resultado FALSE --> ES SESGADO [1] FALSE
Media truncada
mean(df_muestras$Mediatrunc)== Y_truncmean #Resultado FALSE --> ES SESGADO [1] FALSE

4. Graficar mediante tres histogramas las tres series de estimaciones. Tienen una distribución aproximadamente
normal? Incluir en los gráficos una línea vertical de referencia que indique la ubicación del parámetro a estimar.




Distribución de las Medias Truncadas Muestrales
Indicador
Estimador Muestral
Parámetro Poblacional
15000
10000
5000
0
Frecuencia
0
25
50
75
100
Medias Truncadas Muestrales



5. Tabular CV y Error Medio Cuadrático de los tres estimadores

Estimador	CV	EMC
Media Muestral	36.24570	236.5710
Media Truncada	26.55862	114.5424
Mediana	36.73640	128.1706

6. ¿Qué estimador le parece preferible?

La media truncada parece ser el mejor estimador en este caso, al presentar un menor CV y EMC, lo que se da a raíz de que por esa vía eliminamos valores que son muy extremos e influyen altamente en el cálculo de otros estimadores.

Ejercicio III

Supongamos que la tabla de radios censales 2010 es nuestro universo bajo estudio.
Deseamos estimar,
encuestando en su totalidad una Muestra Aleatoria Simple de n=240 radios censales:
• Total de población

• Total de hogares que habitan en viviendas tipo Casa
• Total de hogares que habitan en viviendas rancho/ casilla
• Proporción de hogares que habitan en viviendas rancho/ casilla
Los tres primeros parámetros los estimaremos con el estimador usual del total en un MAS ($N \cdot \bar{y}$), que
sabemos que es el estimador de Horvitz-Thompson. El cuarto parámetro es una razón, lo estimaremos como es
usual en el MAS con la razón muestral.

1. Hallar los cuatro parámetros (o sea los cuatro valores poblacionales)
    - Pueden verse en el punto 2

2. Hallar los CV de los cuatro estimadores (el del estimador (d) será una aproximación)

```
#Calculamos estimadores, estimaciones y parámetros
media_muestral_pob <- mean(muestra_radios$Pob_radio)
Pob_est_ht <- media_muestral_pob * N
Pob_param <- sum(censoviv$Pob_radio)
var_poblacion <- var(muestra_radios$Pob_radio)
var_mediaPobMAS<- (1 - n / N) * (var_poblacion / n)
var_total_poblacion <- N^2 * var_mediaPobMAS
cv_total_poblacion <- (sqrt(var_total_poblacion) / Pob_est_ht) * 100

media_muestral_casa <- mean(muestra_radios$Casa)
Casa_est_ht <- media_muestral_casa * N
Casa_param <- sum(censoviv$Casa)
var_casa <- var(muestra_radios$Casa)
var_mediaCasaMAS<- (1 - n / N) * (var_casa / n)
var_total_casa <- N^2 * var_mediaCasaMAS
cv_total_casa <- (sqrt(var_total_casa) / Casa_est_ht) * 100
cv_total_casa

media_muestral_RC<-mean(muestra_radios$Rancho_Casilla)
RC_est_ht<-media_muestral_RC*N
RC_param<-sum(censoviv$Rancho_Casilla)
var_ranchocasilla<-var(muestra_radios$Rancho_Casilla)
var_mediaRCMAS<-(1-n/N) * (var_ranchocasilla/n)
var_total_RC<-var_total_casa <- N^2 * var_mediaRCMAS
cv_total_RC <- (sqrt(var_total_RC) / RC_est_ht) * 100
cv_total_RC

propRC_est<-mean(muestra_radios$prop_RC)
propRC_Param<-sum(censoviv$Rancho_Casilla)/sum(censoviv$Viv_radio)
varpropRC <- (media_muestral_RC)^2 *
  (var_mediaRCMAS / (media_muestral_RC^2) + var_mediavivMAS /
(media_muestral_viv^2))
```

```
CVr <- (sqrt(varpropRC) / (media_muestral_RC) * 100)

muestra_radios$pondera<-N/n
muestra_radios$fpc<-N
```

	Estimación	Parámetro	CV
Población	39084822	40115211	4.508381
Casa	10274994	10620866	4.185823
Rancho Casilla (c)	425942.7	461725	13.18223
Proporción Rancho Casilla (d)	0.04960193	0.03348484	13.67792

3. ¿Qué estimador tiene el CV más alto?
   - La proporción rancho casilla

4. ¿Cómo son los CV de los estimadores (c) y (d)?
   - Son CV más elevados que los de las otras variables. Lo atribuimos a estas son medias calculadas

5. Qué tamaño de muestra se necesitaría para que el estimador del total de población sea (aproximadamente..)
2%?.

Se necesitan aproximadamente 1000 casos para reducir el coeficiente de variabilidad en torno al 2.01%

6. Qué tamaño de muestra se necesitaría para que el estimador del total de hogares que habitan en
rancho/casilla sea (aproximadamente..) 2%?.

Se necesitan aproximadamente 10000 casos para reducir el coeficiente de variabilidad en torno al 2.10%

7. Seleccionar una MAS de tamaño n=240 y con survey estimar total de población, total de hogares que habitan en rancho/casilla y proporción de hogares que habitan en rancho/casilla, y los respectivos CV e intervalos de confianza con un nivel de confianza de 90%.

Survey Total Poblacion	Survey Total Rancho Casilla	Survey Ratio Rancho Casilla/Viviendas

disenopob <- svydesign(id= ~1,weights=~pondera, data=muestra_radios, fpc=~fpc) EstPob<-svytotal(~Pob_radio, design = disenopob, deff=T, cv=T, ci=T) EstPob deff(EstPob) diseno SE(EstPob) estandar confint(EstPob, level=0.9) cv(EstPob)	disenoRC <- svydesign(id= ~1,weights=~pondera, data=muestra_radios, fpc=~fpc) EstRC<-svytotal(~Rancho_Casilla, design = disenoRC, deff=T, cv=T, ci=T) EstRC deff(EstRC) SE(EstRC) confint(EstRC, level=0.9) cv(EstRC)	disenopropRC <- svydesign(id= ~1,weights=~pondera, data=muestra_radios, fpc=~fpc) EstRatio <- svyratio(~Rancho_Casilla, ~Viv_radio, disenopropRC, deff=TRUE, cv=TRUE, ci=TRUE) EstRatio cv(EstRatio) deff(EstRatio) confint(EstRatio, level=0.9) cv(EstRatio)

	Total	SE	DEff	Intervalo de confianza	CV
T Poblacion	39084822	1762093	1	5 % 95 % 36186437 41983207	4,50%
T Ranchos+ Casillas	425943	56149	1	5 % 95 % 333586.2 518299.1	13,18%
Prop. Ranchos+Casillas/ T viviendas	0.03226606	0.004191183	1	5 % 95 % 0.02537218 0.03915995	12,98%

- Los intervalos contienen el parámetro en cuestión?
- Si, los intervalos incluyen a los parámetros

8. Comentar los resultados hallados

Como se mencionó en el punto anterior, se puede notar en los estimadores de medidas calculadas un mayor coeficiente de variación que en las variables que ya se encontraban presentes como totales originariamente. El Deff es igual a 1, ya que trabajamos con un muestreo aleatorio sistemático.

Ejercicio 4

Supongamos que la tabla de radios censales 2010 es nuestro universo bajo estudio.
Deseamos estimar,
encuestando en su totalidad una Muestra Sistemática de n=240 radios censales:
• Total de población
• Total de hogares que habitan en viviendas tipo Casa
• Total de hogares que habitan en viviendas rancho/ casilla

Se desea comparar dos estrategias (las dos utilizando como estimador la media muestral):

1. Muestreo sistemático, ordenando la tabla por Provincia-Total de viviendas del radio

Estrategia 1: Ordenado por Provincia-Total de Viviendas

```
censoviv <- censoviv[order(censoviv$Provincia, censoviv$Viv_radio), ]
s <- censoviv[seq(aa, N, l), ]
```

2. Muestreo sistemático, ordenando la tabla por un número pseudo aleatorio

Estrategia 2: Ordenado por número aleatorio

```
censoviv$aleatorio <- runif(nrow(censoviv))
censoviv_al <- censoviv[order(censoviv$aleatorio), ]
s_al <- censoviv_al[seq(aa, N, l), ]
```

• Hallar los tres parámetros (o sea los tres valores poblacionales)

Población	Casa	Rancho Casilla
40115211	10620866	461725

• Las estrategias 1 y 2 son insesgadas?

Ambas estrategias resultan inseguras ya que en los datos no hay ningún patrón cíclico que coincida con los intervalos de muestreo, por ende, ambas alternativas son insesgadas, lo que se corrobora con sus sesgos y sesgos relativos. Además, el estimador de orbis-thomson para los totales es un estimador insesgado de los parámetros poblacionales.

• Hallar CV, deff, sesgo relativo y EMC de cada estrategia, seleccionando todas las muestras posibles

- Código disponible en repositorio.

• Presentar en una tabla los resultados

Estimador	CV.Estrategia.1	CV.Estrategia.2	deff.Estrategia.1	deff.Estrategia.2	Sesgo.Relativo. Estrategia.1	Sesgo.Relativo .Estrategia.2	EMC.Estrategia.1	EMC.Estrategia.2
Población	2,749771327	4,103278836	0,436381366	0,971706481	0	0	1,21678E+12	2,70944E+12
Casa	2,271101674	3,916368912	0,302774439	0,900353885	0	0	58182585145	1,73016E+11
Rancho/Casilla	15,03218564	14,62054979	1,047801144	0,991201647	0	0	4817381506	4557159067

• Comentar los resultados

Se puede observar una menor varianza en la Estrategia 1 que en la Estrategia 2. Al tratarse de un ordenamiento por un valor de interés, se traslada ello a un muestreo con mejores valores que en la estrategia aleatoria. La Estrategia 1 cuenta con un efecto de diseño muy menor al de un MAS para los valores no calculados,, mientras que en la Estrategia 2 este es cercano a 1. También se observa un sesgo nulo atento a lo mencionado en el punto anterior. Para el cálculo de “Rancho Casilla”, bien se podría haber utilizado un MAS para estimar su valor.

Ejercicio 5

Deseamos ahora probar otra estrategia para estimar los parámetros del ejercicio IV:
selección de la muestra
mediante Madow, con la cantidad de viviendas del radio como variable auxiliar
1. Seleccionar mediante sampling una muestra de n=240 radios mediante Madow, con total de viviendas del radio como variable auxiliar. Ordenando la tabla según código de radio (jurisdicción departamento+fracción+radio)

```
censovivmad<-censoviv[order(censoviv$Codigo),]
censovivmad$pi_i<-n*censovivmad$Viv_radio/sum(censovivmad$Viv_radio)
pik<-censovivmad$pi_i
s<-sampling:: UPsystematic(pik)
muestra_radios = censovivmad[s==1,]
muestra_radios$pondera <- 1/muestra_radios$pi_i
```

2. Con survey estimar los parámetros, CV y deff correspondiente

	Estimación	SE	deff	cv
Rancho Casilla	395834,9	56096,98	0,727306	14,17181
Población	40157890	831878	0,351973	2,071518
Casa	10760689	259968,1	0,480711	2,415906

3. Repetir los pasos 1 y 2 diez veces

	Estimacion_Pob	SE_Pob	deff_Pob	cv_Pob
1	39967471	876385,9	0,260126	2,192748
2	40558239	917334,8	0,43911	2,261772
3	39854992	866604,3	0,331248	2,174393
4	39711590	874462,2	0,323212	2,202033

5	40921296	888414,8	0,343487	2,171033
6	40231161	883465,8	0,175265	2,195974
7	40203435	935311,8	0,179902	2,326447
8	39118100	831955	0,393926	2,126778
9	39638884	858507,4	0,359247	2,165821
10	39809092	926946,7	0,260526	2,32848

	Estimacion_Casa	SE_Casa	deff_Casa	cv_Casa
1	10762924	257169,6	0,323395	2,389403
2	10633220	265687,5	0,556003	2,498655
3	10435976	267781,6	0,476555	2,565947
4	10419187	272382,8	0,43265	2,614243
5	10901790	233593,3	0,378432	2,142706
6	10428155	254367,9	0,220145	2,439242
7	10340205	263630,5	0,224406	2,549568
8	10526385	255239,7	0,498718	2,424761
9	10876953	251963,6	0,459243	2,31649
10	10537353	267108,5	0,3164	2,534873

	Estimacion_Rancho	SE_Rancho	deff_Rancho	cv_Rancho
1	436800,5	64146,39	1,021926	14,68551
2	436246,8	70520,27	1,080007	16,16522
3	499141	81106,34	0,823561	16,24918
4	460989,3	78486,39	1,106898	17,02564
5	521208,6	76686,26	1,171307	14,71316
6	624841,3	95404,74	1,088126	15,26864
7	614833,7	94775,42	1,369969	15,41481

8	412368,8	68185,73	0,929363	16,53513
9	449449,2	63980,72	0,946095	14,23536
10	469214,1	75212,63	1,19795	16,02949

4. En una tabla resumir los resultados del ejercicio anterior y lo hallado en este ejercicio

Estimador	CV.Estrategia.1	CV.Estrategia.2	CV.Madow	deff.Estrategia.1	deff.Estrategia.2	deff.M
Población	2,74977133	4,103279	1,265163	0,436381	0,971706	0,306
Casa	2,27110167	3,916369	1,887918	0,302774	0,900354	0,388
Rancho/Casilla	15,0321856	14,62055	15,04649	1,047801	0,991202	1,073

Los resultados indican que es posible obtener mejor CV y deff con 10 muestras distintas mediante Madow que repitiendo muestreos sistemáticos para todas las muestras posibles y con estos dos criterios de ordenamiento distintos. Es ampliamente mejor que un MAS para estimar el total de población y de casas, no así para calcular la proporción de Ranchos y casillas sobre el total de viviendas, donde el muestreo sistemático en base a un número pseudoaleatorio performa mejor.

Ejercicio 6

En un ballotage, un candidato X encarga una estadístico una muestra aleatoria simple de electores para saber
si gana o pierde la elección. Supongamos que la gente no miente y que no cambia el voto luego de la encuesta.
Por motivos de costo se encuestaron a 400 personas. De ellas, 212 afirman que votarán por X.

```
n=400
xi=212
p=xi/n
q=1-p
VarP= p*q/n
CVP= sqrt(q/(p*n))
binom.confint(212,400,conf.level=0.95,method=c('all'))
```

1. ¿Qué le informa el estadístico al candidato?. ¿La información que da la encuesta es útil?

	method	x	n	mean	lower	upper
1	agresti-coull	212	400	0.5300000	0.4810353	0.5783939
2	asymptotic	212	400	0.5300000	0.4810892	0.5789108
3	bayes	212	400	0.5299252	0.4811203	0.5786466
4	cloglog	212	400	0.5300000	0.4798858	0.5775344
5	exact	212	400	0.5300000	0.4797721	0.5797809
6	logit	212	400	0.5300000	0.4809578	0.5784697
7	probit	212	400	0.5300000	0.4809860	0.5785631
8	profile	212	400	0.5300000	0.4810147	0.5786023
9	lrt	212	400	0.5300000	0.4810329	0.5785807
10	prop.test	212	400	0.5300000	0.4797909	0.5796238
11	wilson	212	400	0.5300000	0.4810362	0.5783931

Todos los intervalos de confianza incluyen menos y más de 50%, por lo cual, incluye el caso más desfavorable. En el caso de un ballotage, significa que los datos no permiten concluir si el candidato ganará o perderá la elección.

2. ¿Cuál es el CV del estimador?
4.708483

Ejercicio 7

En un ballotage, un candidato X encarga una estadístico una muestra aleatoria simple de electores para saber si gana o pierde la elección. El candidato afirma que supone que obtendrá un porcentaje de votos cercano al 50%. Que es consciente que con eso no se puede determinar un tamaño de muestra para garantizar si gana o pierde la elección, pero que está conforme si el intervalo de confianza al 95% de la estimación tiene una amplitud total de 1%. ¿Qué tamaño de muestra sería necesario? (como suponemos que se trata de una gran ciudad, podemos obviar el factor de corrección por población finita)

```
confianza <- 0.95
p <- 0.5
E <- 0.005
```

```
# Calcular Z para el nivel de confianza
Z <- qnorm((1 + confianza) / 2)

# Calcular el tamaño de muestra
n <- (Z^2 * p * (1 - p)) / (E^2)
n <- ceiling(n)  # Redondeo hacia arriba
n
```

Respuesta : 38415 casos

Explicación:

En primer lugar, definimos la confianza buscada, definimos la cantidad de casos favorables que en este caso es 0.5 para reflejar el porcentaje de votos esperado en torno al 50%. Finalmente, para establecer en 1% la amplitud del intervalo de confianza, definimos el error en 0.005 (Porque es el margen tolerado a cada lado de la distribución).
Luego calculamos el valor del estadístico Z para la confianza definida y por último aplicamos la fórmula.

$$n = \frac{z^2\, p\, q}{e^2}$$

Así obtenemos que harían falta 38415 casos para poder brindar al candidato un pronóstico con un intervalo de confianza de amplitud en torno al 1%

Ejercicio 8

Se selecciona una muestra aleatoria simple de 24 hogares de una localidad pequeña para indagar cierta
característica rara. En la muestra ningún hogar la presenta. ¿Puede dar un intervalo de confianza al 90% para la proporción de hogares con esa característica? (puede utilizarse el paquete de R binom)

```
binom.confint(0,24,conf.level = .9, method= c('all'))
```

	method	x	n	mean	lower	upper
1	agresti-coull	0	24	0.00	-0.01914406	0.12045424
2	asymptotic	0	24	0.00	0.00000000	0.00000000

3	bayes	0	24	0.02	0.00000000	0.05425166
4	cloglog	0	24	0.00	0.00000000	0.11734616
5	exact	0	24	0.00	0.00000000	0.11734616
6	logit	0	24	0.00	0.00000000	0.11734616
7	probit	0	24	0.00	0.00000000	0.11734616
8	profile	0	24	0.00	0.00000000	0.10322831
9	lrt	0	24	0.00	0.00000000	0.05480639
10	prop.test	0	24	0.00	0.00000000	0.17171501
11	wilson	0	24	0.00	0.00000000	0.10131018

Con el 90% de confianza se puede concluir que la condición estará presente en entre el 0 y el 5%, 11% o 17% de los hogares según el método escogido.