

Estrategias Transversales en las Encuestas de Hogares

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

① Muestreo aleatorio simple en dos etapas estratificado

Motivación

Desde que se popularizaron las encuestas de hogares en 1940, se ha hecho evidente algunas tendencias que están ligadas a los avances tecnológicos en las agencias estadísticas y en la sociedad y se han acelerado con la introducción del computador.

Gambino & Silva (2009)

Bibliografía y referencias

- Kish, L. (1965) *Survey Sampling*. John Wiley and Sons.
- Cochran, W. G. (1977) *Sampling Techniques*. John Wiley and Sons.
- Särndal, et. al. (2003) *Model-assisted Survey Sampling*. Springer.
- Gutiérrez, H. A. (2016) *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U.
- Gutiérrez, H. A. (2017) *TeachingSampling*. *R package*.

Muestreo aleatorio simple en dos etapas estratificado

Muestreo en dos etapas estratificado

- La teoría discutida en las secciones anteriores es aplicable cuando las unidades primarias de muestreo son seleccionadas dentro de un estrato.
- No hay nuevos principios de estimación o diseño involucrado en el desarrollo de esta estrategia de muestreo.

Muestreo en dos etapas estratificado

- Se supone que el muestreo en cada estrato respeta el principio de la independencia.
- Las estimaciones del total, así como el cálculo y estimación de la varianza son simplemente resultado de añadir o sumar para cada estrato la respectiva cantidad.

Muestreo en dos etapas estratificado

- Dentro de cada estrato U_h $h = 1, \dots, H$ existen N_{lh} unidades primarias de muestreo, de las cuales se selecciona una muestra s_{lh} de n_{lh} unidades mediante un diseño de muestreo aleatorio simple.
- Suponga, además que el sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple.
- Para cada unidad primaria de muestreo seleccionada $i \in s_{lh}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i .

Muestreo en dos etapas estratificado

Para utilizar los principios de estimación del último conglomerado en este diseño particular se definen las siguientes cantidades:

- 1 $d_{I_i} = \frac{N_{Ih}}{n_{Ih}}$, que es el factor de expansión de la i -ésima UPM en el estrato h .
- 2 $d_{k|i} = \frac{N_i}{n_i}$, que es el factor de expansión del k -ésimo hogar para la i -ésima UPM.
- 3 $d_k = d_{I_i} \times d_{k|i} = \frac{N_{Ih}}{n_{Ih}} \times \frac{N_i}{n_i}$, que es el factor de expansión final del k -ésimo elemento para toda la población U .

Práctica en R

```
data('BigCity')

FrameI <- BigCity %>% group_by(PSU) %>%
  summarise(Stratum = unique(Stratum),
            Persons = n(),
            Income = sum(Income),
            Expenditure = sum(Expenditure))

attach(FrameI)
```

Práctica en R

```
head(FrameI, 10)
```

PSU	Stratum	Persons	Income	Expenditure
PSU0001	idStrt001	118	70912	44232
PSU0002	idStrt001	136	68887	38382
PSU0003	idStrt001	96	37213	19495
PSU0004	idStrt001	88	36926	24031
PSU0005	idStrt001	110	57494	31142
PSU0006	idStrt001	116	75272	43473
PSU0007	idStrt001	68	33028	21833
PSU0008	idStrt001	136	64293	47660
PSU0009	idStrt001	122	33156	23292
PSU0010	idStrt002	70	65254	37115

Práctica en R

```
sizes = FrameI %>% group_by(Stratum) %>%  
  summarise(NIh = n(),  
    nIh = 2,  
    dI = NIh/nIh)  
  
NIh <- sizes$NIh  
nIh <- sizes$nIh
```

Práctica en R

```
head(sizes, 10)
```

Stratum	Nlh	nlh	dl
idStrt001	9	2	4.5
idStrt002	11	2	5.5
idStrt003	7	2	3.5
idStrt004	13	2	6.5
idStrt005	11	2	5.5
idStrt006	5	2	2.5
idStrt007	14	2	7.0
idStrt008	7	2	3.5
idStrt009	8	2	4.0
idStrt010	8	2	4.0

Práctica en R

```
samI <- S.STSI(Stratum, NIh, nIh)
UI <- levels(as.factor(FrameI$PSU))
sampleI <- UI[samI]

FrameII <- left_join(sizes,
                     BigCity[which(BigCity$PSU %in% sampleI), ])
attach(FrameII)
```

Práctica en R

```
head(FrameII, 10)
```

Stratum	Unh nh dl	HHID	Person	PSU	Zone	Sex	Age	Marital	HS	TE	Expend	Employ	Poverty
idStrt	9012	4.5	idHH	0009	PSU	0008	Male	27	Married	13	74667	Employed	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Female	26	Married	13	74667	Employed	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Female	21	NA	13	74667	NA	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Male	44	Married	42	5592	Employed	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Female	40	Married	42	5592	Inactive	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Male	20	Single	42	5592	Unemployed	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Female	18	Single	42	5592	Inactive	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Male	30	Married	39	0356	Employed	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Female	35	Married	39	0356	Inactive	NotPoor
idStrt	9012	4.5	idHH	0009	PSU	0008	Male	12	Single	39	0356	NA	NotPoor

Práctica en R

```
HHdb <- FrameII %>%  
  group_by(PSU) %>%  
  summarise(Ni = length(unique(HHID)))
```

```
Ni <- as.numeric(HHdb$Ni)  
ni <- ceiling(Ni * 0.1)  
sum(ni)
```

```
## [1] 680
```


Práctica en R

```
sam = S.SI(Ni[1], ni[1])
clusterII = FrameII[which(FrameII$PSU == sampleI[1]), ]
sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])
clusterHH <- left_join(sam.HH, clusterII, by = "HHID")
clusterHH$dki <- Ni[1]/ni[1]
clusterHH$dk <- clusterHH$dI * clusterHH$dki
sam_data = clusterHH
```

Práctica en R

```
head(sam_data, 10)
```

HHID	Stratum	HH Income	HH Indl	Person	PSID	Zone	Sex	Age	Marital	HS Exp	Emp	Relig	Boval	dk
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	32	Widow	243	207	Employ	Relig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	18	Singl	243	207	NA	Relig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Male	9	NA	243	207	NA	Relig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	31	NA	243	207	NA	Relig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Male	58	Married	228	87	Employ	NotRelig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	58	Married	228	87	Inact	NotRelig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Male	36	Married	893	494	Employ	NotRelig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	35	Married	893	494	Employ	NotRelig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	81	NA	893	494	NA	NotRelig	Boval	40	
idHH100000002	idHH100000002	4.5	idPer100000002	PSID100000002	Female	41	NA	893	494	NA	NotRelig	Boval	40	

Práctica en R

```
for (i in 2:length(Ni)) {  
  sam = S.SI(Ni[i], ni[i])  
  clusterII = FrameII[which(FrameII$PSU == sampleI[i]), ]  
  sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])  
  clusterHH <- left_join(sam.HH, clusterII, by = "HHID")  
  clusterHH$dki <- Ni[i]/ni[i]  
  clusterHH$dk <- clusterHH$dI * clusterHH$dki  
  data1 = clusterHH  
  sam_data = rbind(sam_data, data1)  
}  
encuesta <- sam_data
```

Práctica en R

```
dim(encuesta)
```

```
## [1] 2422  17
```

```
sum(encuesta$dk)
```

```
## [1] 144817
```

```
nrow(BigCity)
```

```
## [1] 150266
```

```
attach(encuesta)
```

Práctica en R

Definir diseño muestral con la librería srvyr

```
library(srvyr)

diseno <- encuesta %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = dk,
    nest = T
  )

sum(weights(diseno))

## [1] 144817
```

Práctica en R

Calibrando los pesos muestrales, para ello empleamos la función `calibrate` de la librería `survey`

```
library(survey)
totales <- colSums(model.matrix(~ -1 + Zone:Sex, BigCity))
diseno_cal <- calibrate(diseno, ~-1 + Zone:Sex, totales, ca

sum(weights(diseno))

## [1] 144817

sum(weights(diseno_cal))

## [1] 150266

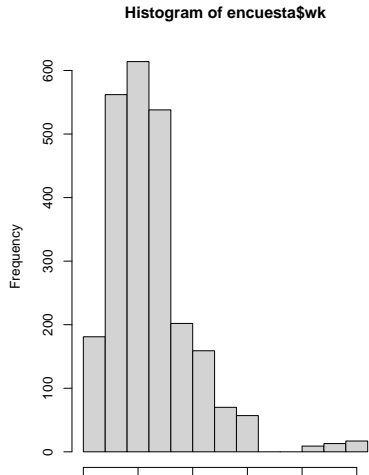
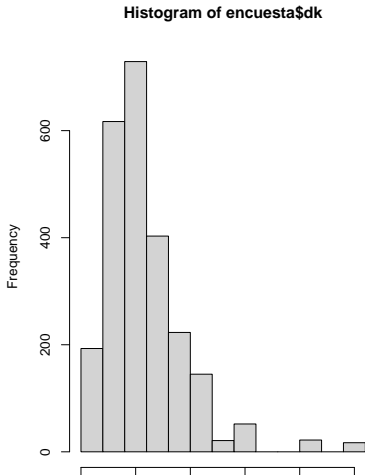
nrow(BigCity)

## [1] 150266

encuesta$wk <- weights(diseno_cal)
```

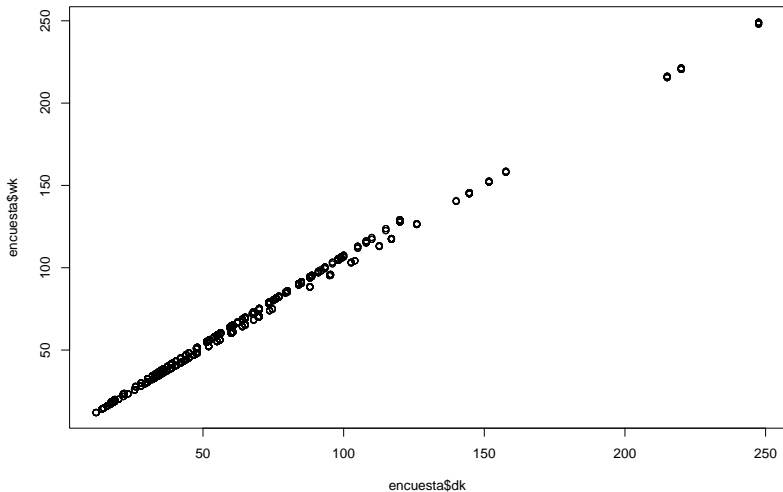
Práctica en R

```
par(mfrow = c(1,2))  
hist(encuesta$dk)  
hist(encuesta$wk)
```



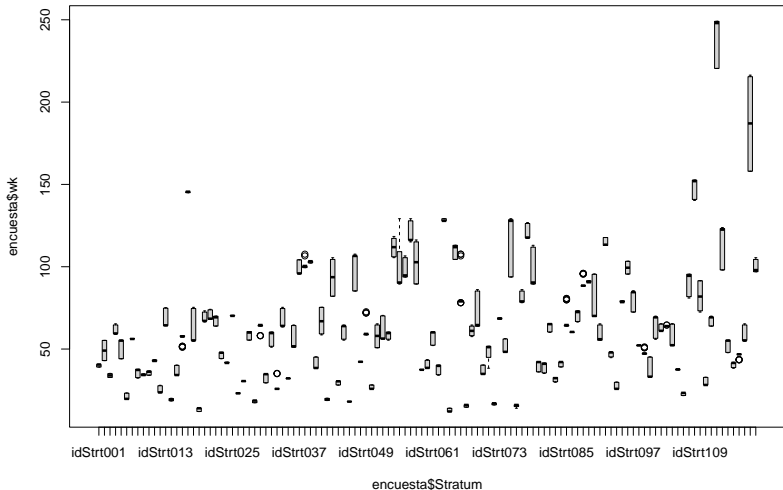
Práctica en R

```
plot(encuesta$dk, encuesta$wk)
```



Práctica en R

```
boxplot(encuesta$wk ~ encuesta$Stratum)
```



Práctica en R

```
Region <- as.numeric(gsub(pattern = "\\D", replacement = "", x = Region))

encuesta$Region <- cut(Region, breaks = 5, labels = c("North", "South", "East", "West", "Central"))

encuesta %<>% mutate(CatAge = case_when(Age <= 5 ~ "0-5",
                                         Age <= 15 ~ "6-15",
                                         Age <= 30 ~ "16-30",
                                         Age <= 45 ~ "31-45",
                                         Age <= 60 ~ "46-60",
                                         TRUE ~ "Más de 60"),
                    CatAge = factor(CatAge,
                                     levels = c("0-5", "6-15", "16-30", "31-45", "46-60", "Más de 60"),
                                     ordered = TRUE))

saveRDS(object = encuesta, file = "../Data/encuesta.rds")
```