Trabajo Práctico Muestreo Nro 2

Felipe Gonzalez

9 de diciembre de 2015

SELECCIÓN DE LA MUESTRA

PRIMERA ETAPA

Para la primera etapa se procede a seleccionar una cantidad de UPM (m) en cada estrato (6, 9 y 3 respectivamente).

```
#Carga de librerías y data original
library(sampling)
library(survey)
library(dplyr)
library(knitr)
set.seed(1)

data = read.csv('radiosACT2.csv')
data = data[order(data$estrato),]
```

```
#Totales de UPM a seleccionar por estrato
M1 = nrow(subset(data,estrato == 1))
M2 = nrow(subset(data,estrato == 2))
M3 = nrow(subset(data,estrato == 3))

#Cantidades de UPM a seleccionar  por estrato
m1 = 6
m2 = 9
m3 = 3
```

```
#a) Probabilidades de inclusión de primer y segundo orden para la primer etapa
##Primer orden
pii1 = m1/M1
pii2 = m2/M2
pii3 = m3/M3

#Segundo orden
piij1 = pii1 * (m1-1) / (M1-1)
piij2 = pii2 * (m2-1) / (M2-1)
piij3 = pii3 * (m3-1) / (M3-1)
```

Cuadro 1 - Probabilidades de 1er y 2do (orden π_{hi}^{UP} y π_{hij}^{UP}) de primera etapa para las UPM

Estrato	π_{hi}^{UP}	π_{hij}^{UP}
Estrato 1	0.2308	0.0462
Estrato 2	0.3	0.0828
Estrato 3	0.375	0.1071

A continuación se procedera aeleccionar la muestra estratificada de 1er etapa bajo MSA según los tamaños definidos en los párrafos anteriores.

```
#b) Selección de muestra estratificada de 1er etapa bajo MSA
muestraResumen = strata(data = data, stratanames = 'estrato',
                        size = c(m1,m2,m3), #se utilizan para la seleccion los parámetros previamente e
                        method = 'srswor',
                        description = F)

muestraResumen = getdata(data, muestraResumen)
```

Cuadro2 - Selección de elementos de la muestra para la primera etapa

SEGUNDA ETAPA

Se procederá a crear una tabla o archivo con la información resumen por Radio o UPM de las probabilidades de 1er orden (π_{hi}^{UP}) y de la 2da etapa ($\pi_{k/hi}$) de las viviendas de la muestra seleccionada en el paso anterior, con los siguientes tamaños de muestras de viviendas (n_i): 90 viviendas del estrato 1, 50 viviendas estrato 2 y 80 viviendas estrato 3. A su vez, se incorpora la ponderación final de una vivienda como resultado de aplicar el diseño muestral empleado (w).

```
#Todal de viviendas a seleccionar cada radio de acuerdo al estrato (n.i)
muestraResumen$n.i = ifelse(muestraResumen$estrato==1, 90,
                        ifelse(muestraResumen$estrato==2,50,80)
                        )

#Probabilidad de las viviendas de ser seleccionadas dentro de cada conglomerado (pikhi)
muestraResumen$pikhi = round(muestraResumen$n.i / muestraResumen$tviv,4)

#Probabilidad final de ser incluida esa vivienda en la muestra (pik)
muestraResumen$pik = round(muestraResumen$pikhi * muestraResumen$Prob,4)

#Pesos (w) - Inversa de la probabilidad final de ser incluida esa vivienda en la muestra
muestraResumen$w = round(1/muestraResumen$pikhi,2)

#
muestraResumen$M = ifelse(muestraResumen$estrato==1, M1,
                              ifelse(muestraResumen$estrato==2,M2,M3)
)
```

Cuadro 3 - Tabla resumen de probabilidades de inclusión por Radio o UPM

	radio	tviv	estrato	Prob	n.i	pikhi	pik	w	M
8	8	389	1	0.2308	90	0.2314	0.0534	4.32	26
11	11	422	1	0.2308	90	0.2133	0.0492	4.69	26
14	14	416	1	0.2308	90	0.2163	0.0499	4.62	26
29	29	375	1	0.2308	90	0.2400	0.0554	4.17	26
48	48	381	1	0.2308	90	0.2362	0.0545	4.23	26
53	53	284	1	0.2308	90	0.3169	0.0731	3.16	26
2	2	363	2	0.3000	50	0.1377	0.0413	7.26	30
16	16	380	2	0.3000	50	0.1316	0.0395	7.60	30
19	19	407	2	0.3000	50	0.1229	0.0369	8.14	30

```
23    23  422       2 0.3000  50 0.1185 0.0356 8.44 30
40    40  321       2 0.3000  50 0.1558 0.0467 6.42 30
41    41  285       2 0.3000  50 0.1754 0.0526 5.70 30
43    43  379       2 0.3000  50 0.1319 0.0396 7.58 30
50    50  322       2 0.3000  50 0.1553 0.0466 6.44 30
57    57  427       2 0.3000  50 0.1171 0.0351 8.54 30
26    26  338       3 0.3750  80 0.2367 0.0888 4.22  8
36    36  288       3 0.3750  80 0.2778 0.1042 3.60  8
44    44  306       3 0.3750  80 0.2614 0.0980 3.83  8
```

```
#Se selecciona la muestra del estrato 3
muestraResumenEst3 = muestraResumen[muestraResumen$estrato == 3,]
```

Por último se presentan las r las probabilidades de 1er y 2do orden finales (π_k y π_{kl}) teniendo en cuenta las 2 etapas utilizando una vivienda del Estrato 3 como ejemplo. Las probabilidades 1er finales (π_k) ya fueron calculadas para cada vivienda y presentadas en el cuadro 3. De este modo el vector *pik* consigna estas probabilidades para cada vivienda de acuerdo al radio al cual pertenezcan. Así, una vivienda que pertenezca al radio 26 tendrá una π_k igual a 0.0888

Para las probabilidades de inclusion de segundo orden, tambien es necesario saber el conglomerado al que pertenece cada vivienda que se empareje. En primer lugar, para las probabilidades de segundo orden de la primera tapa, en la medida en que siempre se selecciona 3 conglomerados de 8, tal como se onbserva en el **Cuadro 1** las probabilidades de segundo orden estan dadas por:

$$\pi_{hij}^{UP} = \frac{m_3}{M_3}\frac{m_3 - 1}{M_3 - 1}$$

$$\pi_{hij}^{UP} = \frac{3}{8}\frac{2}{7} = 0.1071$$

Sin embargo, para las probabilidades de segundo orden de la segunda tapa es necesario considerar los conglomerados de las viviendas del estrato 3 que se emparejan. Por haber solo 3 conglomerados en el estrato 3 las combinaciones pueden ser solo las que figuran en el siguiente cuadro.

```
congComb = combn(3,2)
congComb = t(cbind(congComb,rep(1,2),rep(2,2),rep(3,2)))
congComb = as.data.frame(congComb)
names(congComb) = c('i','j')
congComb
```

```
  i j
1 1 2
2 1 3
3 2 3
4 1 1
5 2 2
6 3 3
```

El paso siguiente es determinar el total de viviendas dentro de cada radio, al momento de seleccionar la primer unidad como así también al momento de seleccionar la segunda unidad, existiendo la posibilidad de que se seleccionen 2 unidades del mismo conglomerado. Con esos datos se puede proceder a determinar las π_{kl} de acuerdo a las fórmulas siguientes:

$$\pi_{kl} = \pi^{UP}_{hij} \frac{n_i}{N_i} \frac{n_j}{N_j} \text{ si k } \in U_i, l \in U_j$$

$$\pi_{kl} = \pi^{UP}_{hij} \frac{n_i}{N_i} \frac{n_i - 1}{N_i - 1} \text{ si k,l } \in U_i$$

Para eso es necesario determinar el total de viviendas (*N3*) de cada radio al momento de seleccionar la primer unidad y la segunda, como así también el total de viviendas a ser seleccionadas en cada momento (*n3*). Se ordena para este proceso una matrix con cada combinación posible para el primer y segundo elemento (i y j) de acuerdo a la procedencia de cada uno de los radios (el 1, 2 o 3 según el orden que tomen en la tabla resumen de las unidades primarias de muestreo)

```
congComb$N3.i = NA
congComb$N3.j = NA
congComb$n3.i = 80 #en la primera seleccion, siempre se seleccionan 80 viviendas en el estrato 3, sin i
congComb$n3.j = NA

for(i in 1:nrow(congComb)){
  #para la selección del primer elemento se toma el total de viviendas del radio tal cual aparece en la
  congComb$N3.i[i]=muestraResumenEst3$tviv[congComb$i[i]]
  #Para la selección del segundo elemento, depende si este procede del mismo radio o no
  if(congComb$j[i]==congComb$i[i]){ #Mismo radio,
    congComb$N3.j[i]=congComb$N3.i[i]-1 #cambia el total (N3) del que se selecciona
    congComb$n3.j[i]=congComb$n3.i[i]-1 #la cantidad a seleccionar (n3)
  }else{# si no son los mismos, se toman los parametros dados
    congComb$N3.j[i]=muestraResumenEst3$tviv[congComb$j[i]] #para el total de viviendas en cada radio
    congComb$n3.j[i]=80 #y el total a seleccionar
  }
}
congComb
```

```
  i j N3.i N3.j n3.i n3.j
1 1 2  338  288   80   80
2 1 3  338  306   80   80
3 2 3  288  306   80   80
4 1 1  338  337   80   79
5 2 2  288  287   80   79
6 3 3  306  305   80   79
```

Establecidos estos valores para cada escenario posible, se procede a aplicar las fórmulas vectorialmente:

```
#segundo orden primer etapa
congComb$pSegOrd.PrimEtap = piij3

#segundo orden segunda etapa
congComb$pSegOrd.SegEtap = congComb$n3.i/congComb$N3.i * congComb$n3.j/congComb$N3.j

#finales
congComb$pSegOrd = congComb$pSegOrd.PrimEtap * congComb$pSegOrd.SegEtap
congComb
```

```
  i j N3.i N3.j n3.i n3.j pSegOrd.PrimEtap pSegOrd.SegEtap     pSegOrd
1 1 2  338  288   80   80        0.1071429      0.06574622 0.007044238
```

```
2 1 3  338  306   80   80       0.1071429      0.06187879 0.006629871
3 2 3  288  306   80   80       0.1071429      0.07262164 0.007780890
4 1 1  338  337   80   79       0.1071429      0.05548435 0.005944751
5 2 2  288  287   80   79       0.1071429      0.07646148 0.008192301
6 3 3  306  305   80   79       0.1071429      0.06771670 0.007255361
```

De este modo se puede obtener en un cuadro las π_k y π_{kl} para cada vivienda en función del radio del cual procedan las mismas.

Cuadro 4a - Probabilidades de inclusión de primer orden finales π_k para el Estrato 3

Radio	π_k
1- 26	0.0888
2- 36	0.1042
3- 44	0.098

Cuadro 4b - Probabilidades de inclusión de segundo orden finales $\pi_k l$ para el Estrato 3

Radio elemento k	Radio elemento l	$\pi_k l$
1	2	0.0070442
1	3	0.0066299
2	3	0.0077809
1	1	0.0059448
2	2	0.0081923
3	3	0.0072554

ESTIMACIONES

Para poder realizar las estimaciones es necesario proveer a la tabla de datos suministrada en el archivo *muestraACT2.csv* (que contiene la muestra seleccionada de viviendas y personas) las probabilidades de inclusión obtenidas en la primer parte del trabajo. Estas varían de acuerdo a cada Estrato y Radio, con lo cual se debe realizar un *merge* utilizando esas variables como claves.

```
#Orden de cuadro resumen
muestraResumen =
  muestraResumen %>%
  group_by(estrato) %>%
  mutate(Cong = row_number()) %>%
  select(estrato,Cong,radio,tviv,M,Prob,n.i,pikhi,pik,w)

#Carga y orden de muestra de viviendas seleccionadas
muestra = read.csv('muestraACT2.csv')
names(muestra)[1]='estrato'
muestra$Sexo = factor(muestra$Sexo,levels = 1:2,labels = c('Masculino','Femenino'))

#Merge por Estrato y Conglomerado
muestra = merge(muestra,muestraResumen,all.x = T,by = c('estrato','Cong'))
```

Una vez que la tabla de datos cuente con las variables relevadas como así también los vectores con las probabilidades de inclusión de ambas etapas de nuestro diseño, entonces puede procederse con las estimaciones.

Para ello es necesario conformar el objeto de diseño que utilizará el paquete **survey**, mediante la función *svydesign()*, especificando los vectores de probabilidades mencionados.

```
muestraEstim = svydesign(id=~Cong,#vector de clusters
                         strata = ~estrato,
                         probs = ~Prob + pikhi,
                         #weights = ~w,
                         data=muestra, #tabla de datos
                         nest=TRUE) #el id de los cluster no es unico, anidados para cada estrato
```

A partir de este punto es posible avanzar con las estimaciones:

```
#a) la proporción de individuos con cobertura en salud por sexo, un coeficiente de variación (CV) para
tablaCobSalud = svyby(~Salud,~Sexo, muestraEstim, svymean, keep.var=F,deff = TRUE)
tablaCobSalud$CV = round(cv(svyby(~Salud,~Sexo, muestraEstim, svymean))*100,2)
tablaCobSalud$statistic.Salud =  round(tablaCobSalud$statistic.Salud*100,2)
tablaCobSalud$DEff.Salud =  round(tablaCobSalud$DEff.Salud,2)
tablaCobSalud = tablaCobSalud[,-1]
names(tablaCobSalud) = c("Cobertura","ED","CV")
```

En el cuadro se puede observar las estimaciones para el parámetro de interés, con un leve aumento de la cobertura para las mujeres, en ambos casos un coeficiente de variación apenas por debajo del 5% y un efecto de diseño que aumenta la varianza del estimador 5 veces más que si se hubiese realizado por muestreo aleatorio simple.

Cuadro 5 - Proporción de individuos con cobertura en salud por sexo, y el efecto de diseño (ED) y coeficiente de variación (CV)

	Cobertura	ED	CV
Masculino	63.84	5.67	4.60
Femenino	66.14	5.65	4.19

```
#b) la proporción de mayores de 55 años con cobertura dentro, un CV para la estimación y el efecto de d

#Se crea un vector booleano para mayores de 55 y se multiplica como numerico con el de cobertura
muestraEstim$variables$Salud2 = muestraEstim$variables$Salud * (muestraEstim$variables$Edad>55)

tablaCobSalud2 = as.data.frame(svymean(~Salud2,muestraEstim, deff = TRUE))
tablaCobSalud2$CV = round(cv(svymean(~Salud2,muestraEstim))*100,2)
tablaCobSalud2 = tablaCobSalud2[,-2]
names(tablaCobSalud2) = c("Cobertura","ED","CV")
tablaCobSalud2$ED = round(tablaCobSalud2$ED,2)
tablaCobSalud2$Cobertura = round(tablaCobSalud2$Cobertura*100,2)
rownames(tablaCobSalud2) = 'Cobertura > 55'
```

El **Cuadro 6** muestra que la cobertura para esta subpoblación es del orden del 15.18%. El coeficiente de variación aumenta levemente en relación a la estimación previa aunque el efecto diseño permanece relativamente constante.

Cuadro 6 - Proporción de individuos mayores de 55 años con cobertura en salud por sexo, y el efecto de diseño (ED) y coeficiente de variación (CV)

	Cobertura	ED	CV
Cobertura > 55	15.18	5.48	8.36

```r
#c) el total de la PEA, el total de desocupados, y la tasa de desocupación (desocupados/PEA) para la lo

muestraEstim$variables$Desocup = as.numeric(muestraEstim$variables$Ocup != 1 & muestraEstim$variables$PI
pea = svytotal(~PEA,muestraEstim,deff=TRUE)

PEA = data.frame(Total = round(pea[[1]]),
                 ED = round(deff(pea)[[1]],2),
                 CV = round(cv(pea)[[1]]*100,2))
rownames(PEA) = 'PEA'



desocup.tot =  svytotal(~Desocup,muestraEstim,deff=TRUE)

DesocupTot =  data.frame(Total = round(desocup.tot[[1]]),
                 ED = round(deff(desocup.tot)[[1]],2),
                 CV = round(cv(desocup.tot)[[1]]*100,2))
rownames(DesocupTot) = 'Desocupacion'

desocup.tasa = svyratio(numerator = ~Desocup,
         denominator = ~ PEA == 1 ,
         design=muestraEstim,
         se = FALSE,
         deff = TRUE)

DesocupTasa =  data.frame(Tasa = round(desocup.tasa$ratio[1]*100,1),
                 ED = round(deff(desocup.tasa)[[1]],2),
                 CV = round(cv(desocup.tasa)[[1]]*100,2))
rownames(DesocupTasa) = 'Desocupacion'
```

Cuadro 7 - Total de población económicamente activa, Efecto de diseño (ED) y coeficiente de variación (CV).

	Total	ED	CV
PEA	45388	3.9	5.77

Cuadro 8 - Total de desocupados, Efecto de diseño (ED) y coeficiente de variación (CV).

	Total	ED	CV
Desocupacion	6736	2.6	8

Cuadro 9 - Tasa de desocupados, Efecto de diseño (ED) y coeficiente de variación (CV).

	Tasa	ED	CV
Desocupacion	18.9	1.47	5.66

```r
#d) la tasa de desocupados por sexo, los CV para cada estimación y los efectos de diseños respectivos.

desocup.sexo =
  svyby(
    ~ Desocup ,
    by = ~ Sexo ,
    denominator = ~ PEA == 1 ,
    design = muestraEstim ,
```

```
    na.rm = TRUE ,
    deff = TRUE,
    svyratio
  )

DesocupSexo = as.data.frame(desocup.sexo)
DesocupSexo = DesocupSexo[,c(2,4)]
names(DesocupSexo) = c('Desocupacion','ED')
DesocupSexo$CV = round(cv(desocup.sexo)*100,2)
DesocupSexo$Desocupacion = round(DesocupSexo$Desocupacion*100,2)
DesocupSexo$ED = round(DesocupSexo$ED,2)
```

Como se puede observar, a la hora de analizar la tasa de desempleo, se observa que la misma es superior a la media (18.9 %) entre las mujeres (23.41 %) que entre los hombres (15.32 %). En ninguna de las estimaciones realizadas, el Coeficiente de Variación supera el 10%.

Cuadro 10 - Tasa de desocupados por sexo, Efecto de diseño (ED) y coeficiente de variación (CV).

	Desocupacion	ED	CV
Masculino	15.32	1.57	8.87
Femenino	23.41	1.27	6.94