TP1 - Muestreo

Felipe Gonzalez

16 de octubre de 2015

1- Diseño muestral

```
marco.personas = spss.get("marco_personas.sav", use.value.labels=TRUE)
n = 1100
N = nrow(marco.personas)
```

Se presenta un marco muestral compuesto por un listado de personas (N =82019) y su *Condición de actividad*. El objetivo es seleccionar una muestra de tamaño 1100 que permita estimar la proporción de población económicamente activa (PEA) para el mismo trimestre del año 2015.

```
head(marco.personas)
```

```
            id estado.2014
1 380051 1 1      Inactivo
2 380087 1 2       Ocupado
3 380087 1 1      Inactivo
4 380100 1 1       Ocupado
5 380100 1 4      Inactivo
6 380100 1 3      Inactivo
```

a) Determinar el margen de error

Se solicita determinar el margen de error obtenido a partir de un Muestreo Simple al Azar (MSA) asumiendo un nivel de confianza del 95% y la proporción poblacional la que se puede obtener del marco en base a la información del año 2014, brindar una idea del margen de error con el cual se trabajará para ese tamaño de muestra.

El margen de error se calcula siguiendo la siguiente fórmula:

$$c = Z_{\alpha/2}\sqrt{(1-\frac{n}{N})}\frac{S}{\sqrt{n}}$$

Para poder calcular dicho margen es necesario obtener el valor de la distribución normal para ese nivel de confianza (Z) y una medida de dispersión de la variable a analizar (S).

Para obtener Z solo hay que obtener dicho valor de la tabla correspondiente a partir de los parámetros dados:

```
confianza = 0.95
riesgo = (1 - confianza)/2
z = qnorm(1 - riesgo)
```

De este modo obtenemos para **Z** el siguiente valor: **1.96**

Para la medida de dispersión de una proporción se obtiene de acuerdo a la siguiente fórmula.

$$S = p(1-p)$$

Para lograr dicha medida, es necesario obtener la distribución de frecuencias de la condición de actividad para observar,de este modo, en ese año qué proporción pertenecía a la población económicamente activa (PEA).

El comportamiento de la variable para el mismo trimestre de 2014 se caracteriza por la siguiente a la distribución de frecuencias:

```
cuadro = prop.table(table(marco.personas$estado.2014))
#se calcula la PEA como la suma de la población ocupada y desocupada
p = as.numeric(cuadro[names(cuadro)=="Ocupado"] + cuadro[names(cuadro)=="Desocupado"])
s = p * (1 - p)
cuadro
```

```
Entrevista individual no realizada (no respuesta al cuestion
                                                 0.00000000
                                                    Ocupado
                                                 0.47130543
                                                 Desocupado
                                                 0.03345566
                                                   Inactivo
                                                 0.49523891
                                          Menor de 10 años
                                                 0.00000000
```

De este modo obtenemos a su vez los valores de **p** (**0.5**) y de **S** (**0.25**). Con estos valores, sumados al valor de **Z** (**1.96**) se puede obtener el valor del margen de error

```
margen.error = z * sqrt(1-(n/N)) * s/sqrt(n)
```

En conclusión, el margen de error obtenido es del orden del **1.47%**

b) Coeficiente de variación

El organismo solicitante, tiene como parámetros de calidad, que el coeficiente de variación de las estimaciones que surjan de estudios por muestreo no deben superar el 5% en el caso de las estimaciones de parámetros principales (en este caso, la proporción de PEA), y no debe ser superior al 20% para cualquier otra estimación.

El coeficiente de variación de una estimación para una proporción a partir de una muestra se obtiene de acuerdo a la siguiente fórmula:

$$CV_{MSA}(\hat{P}_{\pi y}) = \frac{\sqrt{V_{MSA}(\hat{P}_{\pi y})}}{E_{MSA}(\hat{P}_{\pi y})}$$

Bajo un diseño de muestreo aleatorio simple para una proporción, la fórmula para un estimador de la varianza es la siguiente:

$$V_{MSA}(\hat{p}_{y\pi}) = (\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}P_y(1-P_y)$$

Por lo tanto, reemplazando, se obtiene:

$$CV_{MSA}(\hat{P}_{\pi y}) = \frac{\sqrt{(\frac{1}{n} - \frac{1}{N})\frac{N}{N-1}P_y(1-P_y)}}{P_y}$$

```
obtenerCV = function(n,N,s,p){
  return(sqrt( (1/n - 1/N) * (N/(N-1)) * s)/p)
}

cv = obtenerCV(n,N,s,p)
```

Al calcular el coeficiente de variación con N = 82019, n = 1100, p = 0.5047611 y s = 0.2499773, se obtiene un **CV** de 2.97%

c) Qué tamaño de muestra hubiera sido suficiente para cumplir con el requerimiento de un CV del 5%?

Con un tamaño de muestra de 1100 casos se obtuvo un coeficiente de variación del 2.97%. Por lo tanto, un tamaño de muestra más pequeño podría arrojar un coeficiente menor de 5% a menor costo. Para llegar a dicho tamaño, existen dos caminos. Dada la fórmula anterior para calcular el coeficiente de variación, puede minimizarse el valor de **n** para que cumpla con la restricción impuesta por el organismo. Por otro lado, existe una fórmula de trabajo que puede acercar el valor de **n** a partir de los valores del coeficiente de variación dado y el buscado.

Minimizar n

```
#Se establece un valor para el CV deseado
cv.buscado = 0.05
#Se fija un valor para el nuevo tamaño de muestra a partir del existente
n.minim = n
#Se inicia un while loop comparando el CV obtenido con el deseado, mientras aquel sea menor a este
while(cv<cv.buscado){
  cv = obtenerCV(n.minim,N,s,p)
  #Se puede disminuir hasta alcanzar el nuevo tamaño de muestra (n.minim)
  n.minim = n.minim-1
}
```

Este método arroja un nuevo tamaño de muestra de n = 389

Formula de trabajo

Asumiendo un tamaño "grande" de N (en este caso N es de 82019 casos) se puede derivar el tamaño de muestra n a partir de la siguiente fórmula de trabajo dada un coeficiente de variación deseado (CV_0) y el observado en el marco (CV_y):

$$n \simeq (\frac{CV_y}{CV_0})^2 \approx (\frac{1-P}{P})\frac{1}{CV_0^2}$$

Reemplazando para los valores del problema:

$$n \simeq \left(\frac{0.05}{0.05}\right)^2 \approx \left(\frac{1-0.5}{0.5}\right)\frac{1}{0.05^2} \approx 392$$

```
#Se calcula n de acuerdo a la formula de trabajo
n.form = ((1-p)/p)/(0.05*0.05)
```

Como se puede observar, los tamaños de muestra calculados a partir de un coeficiente de variación objetivo por los diferentes métodos son similares: minimizando n se obtiene una muestra de tamaño de muestra de 389 casos mientras que con la aproximación de la fórmula de trabajo se obtiene un tamaño de 392 casos.

d) ¿Cuáles son las probabilidades de inclusión de primer y segundo orden para el diseño planteado?

Se parte del supuesto de un diseño de muestreo aleatorio simple con tamaño de muestra fijo. En este diseño, las probabilidades de inclusión de primer orden están dadas por la siguiente ecuación:

```
pik = n/N
pikl = (n*(n-1))/(N*(N-1))
```

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

$$\pi_k = \frac{1100}{82019} = 0.0134$$

A su vez, las probabilidades de inclusión de segundo orden se encuentran dadas por:

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

$$\pi_{kl} = \frac{1100(1099)}{82019(8.2018 \times 10^4)} = 2 \times 10^{-4}$$

e) En el caso que se quisiera utilizar un muestreo sistemático tradicional

a. ¿Cuántas muestras posibles existen?

Dados los tamaños de N (82019) y n (1100), se puede calcular el intervalo de selección k:

```
k = as.integer(N/n)
```

Con este intervalo $k = 74$, bajo un muestreo sistemático la cantidad de muestras posibles ($\Omega_{SISn} = k$) son 74. Sin embargo este procedimiento implicaría dejar 619 casos nula probabilidad de inclusión. Estos podrían incluirse en una muestra adicional, lo que implicaría que haya un total de 75 muestras, aunque con probabilidades de inclusión diferentes. Este es uno de los inconveniente con el muestreo sistemático tradicional.

b. ¿Son todas de igual tamaño?

La única manera que haya tamaños de muestra iguales es que el tamaño de n sea un factor de N. Por lo tanto para N = 82019 solo existen 2 tamaños de muestra que den muestras iguales: 7 o 11717. Para el resto de los tamaños de muestra no habrá muestras de igual tamaño

c. ¿Cuáles son las probabilidades de inclusión de primer y segundo orden?

Las probabilidades de inclusión varían notablemente de acuerdo al intervalo elegido y al tamaño de muestra. Para n factor de N $\pi_k = n/N = 1/k$ y $\pi_{kl} = n/N$ si tanto el elemento k como el elemento l pertenecen ambos a la muestra. En caso contrario $\pi_{kl} = 0$

d. En el caso de que así no fueran ¿cómo podrían solucionarlo?

Pueden utilizarse diferentes cantidades m de arranques. En caso de arranques múltiples cada arranque genera una submuestra de tamaño n/m. Por ende, n debe ser factor de N y a su vez m de n. Para solucionar este último problema, puede realizarse un muestreo sistemático circular en lugar de un muestreo sistemático tradicional.

e. Para la solución en *d* ¿cuáles son las probabilidades de inclusión de primer y segundo orden?

Para un muestreo sistemático tradicional donde n es factor de N y a su vez m de n $\pi_k = m/mk = n/N$ y $\pi_{kl} = n/N$ si tanto el elemento k como el elemento l pertenecen ambos a la misma submuestra. En cambio si pertenecen a submuestras distintas $\pi_{kl} = \frac{n}{N}\frac{m-1}{mk-1}$.

f. ¿Cuál sería el problema adicional de estos métodos de selección y cómo podría solucionarlos?

El problema adicional es el aumento de la varianza del estimador en comparación con la varianza del estimador bajo muestreo aleatorio simple.

2- Selección de la muestra

En **R** el proceso de selección de una muestra de acuerdo al diseño de muestreo aleatorio simple sin reposición puede hacerse de dos maneras. En primer lugar puede generarse un vector de números pseudo-aleatorios dentro del marco muestral para luego ordenar el mismo y seleccionar los primeros casos. En segundo lugar, puede hacerse uso de la función *mstage()* del paquete *sampling* que implementa un muestreo multietápico con determinadas probabilidades de inclusión.

Generación de números pseudo-aleatorios

```
#Establecer la semilla para el proceso psuedo-aleatorio
set.seed(1)
#Generar vector de números entre 0 y 1 con largo igual al tamaño del marco muestral
marco.personas$numeroRandom = runif(N, min = 0, max = 1)
#Ordenar el marco por este numero
marco.personas = marco.personas[order(marco.personas$numeroRandom),]
#Seleccionar una cantidad de filas igual al tamaño de muestra n
muestra1 = marco.personas[1:n,1:ncol(marco.personas)-1]
write.xlsx(muestra1,"punto2_muestra1.xls",row.names = F)
```

Posterior a esta selección, queda establecida la muestra seleccionada de la cual se muestran sus primeros 5 registros:

```
head(muestra1)
```

```
                id estado.2014
68378 E189197 1 4   Desocupado
21679 382387 1 1       Ocupado
46408 424007 1 1       Ocupado
75362 E178952 1 2      Ocupado
77690 E184503 1 1      Ocupado
60925 E182079 1 4     Inactivo
```

La función *mstage()* del paquete *sampling*

La función *mstage()* toma como parámetros el marco muestral y el tamaño de muestra deseado. Pueden especificarse diferentes etapas con estratos y clusters, métodos de selección unitarios y un vector de probabilidades de inclusión. Ante la ausencia de los mismos, no se realizan etapas previas, el método de selección de unidades por defecto seleccionado es el de muestreo aleatorio simple sin reemplazo y se asignan las mismas probabilidades de inclusión para todos los casos.

```
#Se construye la muestra
set.seed(1)
m = mstage(data = marco.personas, size = n)
#Se extrae del marco muestral la muestra seleccionada con la función getdata()
muestra2 = getdata(marco.personas,m)[[1]]
names(muestra2)[ncol(muestra2)]="probInc"
write.xlsx(muestra2,"punto2_muestra2.xls",row.names = F)
```

De igual modo, posterior a esta selección, queda establecida la muestra seleccionada de la cual se muestran sus primeros registros:

```
head(muestra2)
```

```
                id estado.2014 numeroRandom ID_unit    probInc
76255 E190942 1 1     Inactivo  0.001298270     107 0.01341153
80258 424754 1 3      Inactivo  0.001774297     150 0.01341153
31018 383184 1 2      Inactivo  0.001867227     158 0.01341153
52459 E184994 1 1     Inactivo  0.001889492     164 0.01341153
67445 E169728 2 2   Desocupado  0.004613234     388 0.01341153
30537 421617 1 2      Inactivo  0.005738971     464 0.01341153
```

3- Estimación

```
listaBases = read.xlsx("Distribucion datos tp1.xlsx",1,endRow = 30)
casoFelipe = listaBases$"datos.TP1"[listaBases$Nombre=="Felipe" & listaBases$Apellido=="Gonzalez"]
muestra = spss.get(paste("bases muestras para los alumnos/muestra_personas_",casoFelipe,
                         ".sav",sep=""), use.value.labels=TRUE)
```

Se parte de una muestra con 1100 casos y 7 variables. El encabezado de la tabla es el siguiente:

```
head(muestra)
```

```
     provincia  sexo edad sit.cony                          nivel.ed
1 Buenos Aires Mujer   23  Soltero Superior Universitaria Incompleta
2 Buenos Aires Mujer   16  Soltero                 Primaria Completa
3 Buenos Aires Varón   28  Soltero Superior Universitaria Incompleta
4 Buenos Aires Varón   19  Soltero             Secundaria Incompleta
5 Buenos Aires Mujer   34    Unido             Secundaria Incompleta
6 Buenos Aires Varón   24  Soltero               Secundaria Completa
    estado ingreso
1 Inactivo       0
2 Inactivo       0
3  Ocupado    2000
4  Ocupado    6000
5 Inactivo       0
6  Ocupado    5400
```

a) Estimar el total y la proporción de PEA y brindar un Coeficiente De Variación (CV) para dicha estimación. Según la estimación, ¿la proporción de PEA creció o disminuyó en relación al año 2014?

Para realizar estimaciones el paquete de software *survey* permite obtener los estimadores y sus desvíos estándar. De todos modos, para esta primer estimación se hará también paso a paso siguiendo las fórmulas. El resto se procederá a estimar utilizando las diferentes funciones del paquete *survey*.

Para poder estimar el total poblacional de la población económicamente activa (PEA) se utilizará el estimador Horvitz Thompson, de acuerdo a la siguiente fórmula:

$$\hat{T_{Y-HT}} = \hat{T_{\pi y}} = \sum_{k \in s} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{I_k Y_k}{\pi_k} = \sum_{k \in U} w_k I_k Y_k$$

Para poder realizar dicha estimación es necesario introducir los pesos (w_k) calculados a partir de las probabilidades de inclusión de primer orden (π_k). Para esto se asume que el modelo de selección es el de muestreo aleatorio simple sin reposición en función del marco muestral provisto en los pasos anteriores.

Por lo tanto es posible asumir que las probabilidades de inclusión (de primer y de segundo orden) continúan siendo las mismas.

$$\pi_k = 0.0134$$

$$\pi_{kl} = 2 \times 10^{-4}$$

En función de estos π_k pueden calcularse los pesos o ponderadores del siguiente modo:

$$w = \frac{1}{\pi_k} = 74.56$$

```
#Se crean estos vectores en la tabla de datos
muestra$pik = pik
muestra$w = 1/pik
```

A su vez, es necesario establecer un vector *booleano* o *dummy* que permita identificar a los casos pertenecientes a la *PEA* (I_k).

```
muestra$ik = as.numeric(muestra$estado == "Ocupado" | muestra$estado == "Desocupado" )
```

Con estos insumos ya es posible calcular el total.

```
#Se calcula el total a partir de la sumatoria a lo largo del vector resultante de multipliar ik por los
total.ht = round(sum(muestra$w * muestra$ik))
```

El total para la población económicamente activa de 2015 es de 43321 personas.

Para obtener el coeficiente se puede repetir los procedimientos realizados en punto **b** con la misma fórmula, aunque en esta ocasión para un total:

$$CV_{MSA}(\hat{T}_{\pi y}) = \frac{\sqrt{V_{MSA}(\hat{T}_{\pi y})}}{E_{MSA}(\hat{T}_{\pi y})}$$

Para poder continuar es necesario calcular un estimador de la varianza $V_{MSA}(\hat{T}_{\pi y})$. La fórmula bajo muestro simple al azar para un estimador de la varianza es:

$$V_{MSA}(\hat{T}_{\pi y}) = N^2(\frac{1}{n} - \frac{1}{N})s_y^2$$

Por lo tanto, reemplazando, se obtiene:

$$CV_{MSA}(\hat{P}_{\pi y}) = \frac{\sqrt{N^2(\frac{1}{n} - \frac{1}{N})s_y^2}}{\hat{T_{\pi y}}}$$

Habiendo obtenido el estimador $\hat{T_{\pi y}}$ queda solo calcular ese estimador de la varianza a partir de la varianza de la variable I_k creada que determina si cada elemento pertenece a la PEA.

```
#Se estima la varianza de la muestra
var.ik = var(muestra$ik)
#se calcula la varianza del estimador
var.total.ht = N^2*((1/n)-(1/N)) * var.ik
#Se calcula el coeficiente de variación
cv.total.ht = sqrt(var.total.ht)/total.ht
```

De este modo se obtiene un coeficiente de variación del 2.83

Por otro lado, el paquete *survey* permite construir un objeto de tipo específico que replica un diseño muestral complejo. En primer lugar, mediante la función *svydesign()* se construye dicho objeto utilizando como parámetros la tabla de datos, un vector con los pesos (w_k) y un vector de identificación de clusters que por tratarse de un muestreo simple al azar se establece como igual a 1.

```
muestra2 = svydesign(id=~1,         #vector de clusters (fijado en 1)
                     weights=~w,    #vector de pesos definidos previamente
                     data=muestra) #tabla de datos
```

Una vez construido dicho objeto, *survey* cuenta con una serie de funciones que permite extraer del objeto con la tabla de datos y la información del diseño los estimadores junto con el desvío estándar de los mismos.

De este modo se puede calcular un total:

```
round(svytotal(~ik, muestra2)[[1]])
```

```
[1] 43321
```

Así como también su coeficiente de variación:

```
round(cv(svytotal(~ik, muestra2))[[1]]*100,2)
```

```
[1] 2.85
```

```
tabla.resumen = as.data.frame(svytotal(~ik, muestra2))
names(tabla.resumen)=c("total","desvio")
tabla.resumen$cv=round(tabla.resumen$desvio / tabla.resumen$total * 100,2)
tabla.resumen$total[1]=round(tabla.resumen$total[1])
tabla.resumen[2,] = c(total.ht,
                      sqrt(var.total.ht),
                      round(cv.total.ht*100,2))
row.names(tabla.resumen) = c("survey","formulas")
```

Como se puede observar en la tabla resumen, los resultados a los que se llegan son similares. Las pequeñas diferencias del 0,01% se deben a que los números reales en los sistemas informáticos al no ser de base decimal sino binario se almacenan como aproximaciones.

Cuadro 3.a.1 - Estimaciones del total de la PEA, desvío estándar y coeficiente de variación

```
         total   desvio   cv
survey   43321 1235.079 2.85
formulas 43321 1226.769 2.83
```

Por lo tanto, se estima para la población económicamente activa un total de 43321 con un coeficiente de variación del 2.83%

Del mismo modo que se calculó un total con la función *svytotal()*, el paquete *survey* también provee de una función para calcular una proporción, aunque mediante un rodeo. La función a utilizar es *svymean()* y tal cual su nombre lo indica, lo que efectivamente hace es calcular un promedio. Al utilizar vectores *booleanos* o *dummies* con valor 1 para el elemento deseado, calcular una proporción no es diferente a calcular un promedio de una variable que toma valores 1 y 0 y dividir por el total de registros. Obviamente, ese total también es un estimado e impacta

Por lo tanto, para establecer una proporción utilizando el paquete *survey* conviene transformar la variable en un factor con valores *Activo* para quienes forman parte de la PEA ($I_k = 1$) e *Inactivos* para los que no. Esta factorización de las variables permite que el paquete asimile rápidamente qué subgrupos debe conformar y asignarles 1s y 0s automáticamente.

```
muestra2$variables$pea = muestra2$variables$ik
muestra2$variables$pea[muestra2$variables$pea==1] = "Activos"
muestra2$variables$pea[muestra2$variables$pea==0] = "Inactivos"
```

Una vez factorizada la variable se puede calcular la proporción de la PEA junto con su coeficiente de variación.

```
round(svymean(~pea, muestra2)*100,2)
```

```
              mean     SE
peaActivos   52.82 0.0151
peaInactivos 47.18 0.0151
```

```
round(cv(svymean(~pea, muestra2))*100,2)
```

```
  peaActivos peaInactivos
       2.85         3.19
```

De este modo se puede observar que existe una proporción de la población perteneciente a la PEA del orden del 52.82% con un CV del 2.85%. En **2014** se registraba una proporción de la PEA del orden del 50.48. Sin embargo, se se incluyen los intervalos de confianza para la estimación de **2015** los resultados no son convincentes en la medida en que el parámetro de 2014 queda comprendido dentro de los intervalos de confianza del estimador de **2014**:

```
round(confint(svymean(~pea,muestra2))[1,]*100,2)
```

```
 2.5 % 97.5 %
 49.87  55.77
```

b) Estimar la proporción de PEA por sexo, grupo de edad (10-17, 18-40, 41-59, 60 y más), nivel educativo, situación conyugal, y regiones (NEA, NOA, Pampeana, Cuyo, Patagonia, CABA). Teniendo en cuenta el límite del 20% de CV para publicar resultados, indicar si es posible publicar todas las estimaciones, y en el caso de que así no fuera, indicar cómo podría solucionarlo.

A continuación se calcula la proporción de PEA cruzadas por diferentes variables mediante la función *svymean()* del paquete *survey*.

```
#Sexo
tablaSexo = svyby(~pea,~sexo, muestra2, svymean, keep.var=F)
tablaSexo = tablaSexo[,2:3]
names(tablaSexo) = c("Activos","Inactivos")
tablaSexo = round(100 * tablaSexo, 2)
tablaSexoCV = round(cv(svyby(~pea,~sexo, muestra2, svymean))*100,2)
names(tablaSexoCV) = c("CV Activos","CV Inactivos")
```

Al ver la proporción de PEA por sexo se observa que dentro de los varones, la proporción de población económicamente activa es mayor.

Cuadro 3.b.1 - Proporción de PEA por sexo en %

```
      Activos Inactivos
Varón   63.55     36.45
Mujer   43.06     56.94
```

Como se puede observar en las tablas, ninguno de los coeficientes supera la restricción del 20%.

Cuadro 3.b.2 - Coeficiente de variación de las estimaciones en %

```
      CV Activos CV Inactivos
Varón       3.31         5.77
Mujer       4.79         3.62
```

```
#Edad
tablaEdad = svyby(~pea,~edadG, muestra2, svymean, keep.var=F)
tablaEdad = tablaEdad[,2:3]
names(tablaEdad) = c("Activos","Inactivos")
tablaEdad = round(100 * tablaEdad, 2)
tablaEdadCV = round(cv(svyby(~pea,~edadG, muestra2, svymean))*100,2)
names(tablaEdadCV) = c("CV Activos","CV Inactivos")
```

En el caso de la proporción de PEA por edad, se observa que las edades de entre 18 y 59 años registran mayor proporción de PEA (entre 18 y 40 años es menor, probablemente debido a ls jóvenes estudiantes que son *Inactivos*)

Cuadro 3.b.3 - Proporción de PEA por edad en %

```
         Activos Inactivos
10-17       0.64     99.36
18-40      68.57     31.43
41-59      78.95     21.05
60 y más   25.00     75.00
```

Como se puede observar en las tablas de los coeficientes, solo la celda de Activos de entre 10 y 17 supera la restricción del 20%. En ese caso, si se desea realizar estimaciones para esa subpoblación se podría realizar un muestreo estratificado que ofrezca representatividad pare la misma.

Cuadro 3.b.4 - Coeficiente de variación de las estimaciones en %

```
         CV Activos CV Inactivos
10-17         99.73         0.64
18-40          3.06         6.68
41-59          3.29        12.33
60 y más      12.77         4.26
```

```
#nivel educativo
tablaEduc = svyby(~pea,~nivel.ed, muestra2, svymean, keep.var=F)
tablaEduc = tablaEduc[,2:3]
names(tablaEduc) = c("Activos","Inactivos")
tablaEduc = round(100 * tablaEduc, 2)
tablaEducCV = round(cv(svyby(~pea,~nivel.ed, muestra2, svymean))*100,2)
names(tablaEducCV) = c("CV Activos","CV Inactivos")
```

A la hora d observar la proporción de población económicamente activa de acuerdo al nivel educativo, se observa que los niveles educativos con mayor proporción son *Superior Universitaria Completa*, *Secundaria completa* y *Primaria completa*. Seguramente hacia el interior se observen diferencias vinculadas con la condición de ocupación y el ingreso medio.

Cuadro 3.b.5 - Proporción de PEA por nivel educativo en %

```
                                                Activos Inactivos
Primaria Incompleta (incluye educación especial)   20.00     80.00
```

Primaria Completa	58.21	41.79
Secundaria Incompleta	37.78	62.22
Secundaria Completa	71.56	28.44
Superior Universitaria Incompleta	54.62	45.38
Superior Universitaria Completa	84.50	15.50
Sin instrucción	16.67	83.33

Como se puede observar en las tablas de los coeficientes los casos que superan la restricción del 20% son aquellos *Activos sin instrucción* y apenas por encima del límite los *Inactivos Superior Universitaria Completa*. Estas son poblaciones no tradicionales, más aquella que ésta y seguramente sean pocos casos en la muestra. Puede realizarse un muestreo estratificado que apunte a lograr los casos necesarios de esas subpoblaciones o puede agruparse *Sin instrucción* con *Primaria incompleta*.

Cuadro 3.b.6 - Coeficiente de variación de las estimaciones en %

	CV Activos	CV Inactivos
Primaria Incompleta (incluye educación especial)	16.62	4.15
Primaria Completa	5.98	8.33
Secundaria Incompleta	7.81	4.74
Secundaria Completa	4.27	10.75
Superior Universitaria Incompleta	8.36	10.06
Superior Universitaria Completa	3.77	20.56
Sin instrucción	52.73	10.55

```
#situación conyugal
tablaConyu = svyby(~pea,~sit.cony, muestra2, svymean, keep.var=F)
tablaConyu = tablaConyu[,2:3]
names(tablaConyu) = c("Activos","Inactivos")
tablaConyu = round(100 * tablaConyu, 2)
tablaConyuCV = round(cv(svyby(~pea,~sit.cony, muestra2, svymean))*100,2)
names(tablaConyuCV) = c("CV Activos","CV Inactivos")
```

El corte por situación conyugal no ofrece otra información que una merma en la proporción de la población económicamente activa entre los *Viudos* y los *Solteros*, seguramente vinculado con la edad. Es plausible de esperar (dada además la distribución previa por edad) que los *Solteros* sean inactivos por ser estudiantes y los *Viudos* por retirados. Esta categoría, a su vez, seguramente este relacionada con el sexo en la medida en que las mujeres tienen mayor sobrevida media que los hombres.

Cuadro 3.b.7 - Proporción de PEA por sitación conyugal en %

	Activos	Inactivos
Unido	71.71	28.29
Casado	64.04	35.96
Separado o divorciado	79.66	20.34
Viudo	20.63	79.37
Soltero	38.88	61.12

Como se puede observar en las tablas nuevamente en las poblaciones atípicas los coeficientes de variación superan la restricción del 20%. Las soluciones pueden ir en el sentido de las previamente mencionadas.

Cuadro 3.b.8 - Coeficiente de variación de las estimaciones en %

	CV Activos	CV Inactivos
Unido	4.39	11.12

```
Casado                     4.39         7.81
Separado o divorciado      6.58        25.78
Viudo                     24.72         6.43
Soltero                    5.72         3.64
```

```
#Region
tablaRegion = svyby(~pea,~region, muestra2, svymean, keep.var=F)
tablaRegion = tablaRegion[,2:3]
names(tablaRegion) = c("Activos","Inactivos")
tablaRegion = round(100 * tablaRegion, 2)
tablaRegionCV = round(cv(svyby(~pea,~region, muestra2, svymean))*100,2)
names(tablaRegionCV) = c("CV Activos","CV Inactivos")
```

A la hora de observar la proporción por región se observa que las regiones con mayor proporción de población económicamente activa son la región Patagónica y CABA.

Cuadro 3.b.9 - Proporción de PEA por región en %

```
          Activos Inactivos
NEA         46.90     53.10
NOA         49.64     50.36
Pampeana    53.87     46.13
Cuyo        49.47     50.53
Patagonia   61.25     38.75
CABA        64.00     36.00
```

Finalmente, el coeficiente de variación solamente supera la restricción impuesta para la proporción de *Inactivos* en la Ciudad de Buenos Aires.

Cuadro 3.b.10 - Coeficiente de variación de las estimaciones en %

```
          CV Activos CV Inactivos
NEA             8.84         7.81
NOA             6.09         6.00
Pampeana        4.62         5.40
Cuyo           10.37        10.16
Patagonia       6.29         9.94
CABA           15.01        26.68
```

c) Estimar la tasa de desocupación. ¿Está estimando una proporción o una razón? ¿Por qué? ¿Es posible utilizar el estimador de HT en este caso? ¿Por qué?

La tasa de desocupación se calcula como la **razón** entre la población ocupada y la población económicamente activa. Ambas cantidades, tanto el nominador como el denominador, son estimaciones. Por lo tanto no es posible utilizar el estimador HT en la medida en que el estimador de la tasa de desocupación no es una combinación lineal de variables aleatorias Por más que ambos sean estimadores insesgados, el estimador para la razón no lo será. Es decir, por más que el estimador de la población económicamente activa y el de la población desocupada sean insesgados, no lo será el estimador para la tasa de desocupación. Al no ser una función lineal en los parámetros, no puede calcularse una varianza.[1]

[1] Sarndal, Swenson, and Wretman (1992), Model Assisted Survey Sampling, Springer-Verlag, pag 162

d) Estimar el ingreso promedio para la población en general y para cada una de las categorías ocupacionales. ¿Cuál es la diferencia entre ambos estimadores? ¿Es posible publicar todas las estimaciones teniendo en cuenta el límite del 20% para los CV?

```
#Ingreso
tablaIngreso = svyby(~ingreso,~estado, muestra2, svymean, keep.var=F)
tablaIngreso$CV = round(cv(svyby(~ingreso,~estado, muestra2, svymean))*100,2)
tablaIngreso = tablaIngreso[,2:3]
tablaIngreso[4,] = c(svymean(~ingreso, muestra2)[1],
                     round(cv(svymean(~ingreso, muestra2))*100,2))

rownames(tablaIngreso) = c("Ocupado","Desocupado","Inactivo","Total")
names(tablaIngreso) = c("Ingreso medio","CV")
```

El ingreso medio para la población general es del orden de los 4132.61$. Mientras que para los *Ocupados* aumenta a 6880.23$. De las estimaciones, la única que presenta un coeficiente de variación por encima de la restricción impuesta del 20% es para la categoría *Desocupado*.

Cuadro 3.d.1 - Estimación del ingreso medio (en $) y Coeficiente de variación (en %) en la población total y por condición de actividad

	Ingreso medio	CV
Ocupado	6880.232	3.78
Desocupado	1302.000	72.34
Inactivo	1465.171	9.03
Total	4132.605	4.06