matemática: Una publicación del ICM – ESPOL
2013, Vol. 11, No. 1

DISEÑO DE MUESTRA PARA CONTEO RÁPIDO

Vera Francisco[1]

Resumen. El presente artículo discute la estimación del margen de error en la proyección de proporciones a partir de una muestra, utilizando una combinación de muestreo y censo, El censo observa todas las unidades de una cierta categoría, y la muestra escoge estratificadamente por conglomerados los de otra categoría. La estimación de los márgenes de error se realiza con la técnica del bootstrap.

Palabras Clave: muestreo estratificado, muestreo por conglomerados, bootstrap, proyección electoral.

Abstract. This article discusses the estimation of the margin of error in the estimation of proportions from a sample, using a combination of a census and a sample. The census observes all units from certain category, while the sample chooses clusters in a stratified way from another category. The estimation of the margins of error is done using bootstrap.

Key Words: stratified sampling, cluster sampling, bootstrap, electoral projection.

Recibido: Febrero 2013
Aceptado: Marzo 2013

1. INTRODUCCIÓN

El propósito de la muestra para conteo rápido es poder estimar el porcentaje de votos válidos que tendrá cada candidato presidencial en la primera vuelta de las elecciones ecuatorianas, llevada a cabo el 17 del presente mes.

Este documento explica la técnica de selección de la muestra para el conteo rápido, presenta las expresiones matemáticas que se deben usar para la estimación de las proporciones, y un análisis del margen de error de muestreo de la estimación.

2. POBLACIÓN

La población consiste en los electores ecuatorianos a nivel nacional y no incluye a los electores del extranjero. Estos están agrupados en juntas electorales, la mayoría de las cuales tiene 300 electores, las otras tienen menos de 300 electores. De esta manera los 11,000,000 de electores están agrupados en 39700 juntas.

Las juntas a su vez están agrupadas por recintos electorales, el más grande de los cuales tiene 123 juntas, mientras que los más pequeños tienen 2 juntas. El mayor costo de tomar una observación es el de trasladarse al recinto electoral. Una vez en el recinto, el costo de observar una junta o todas las juntas es similar, por lo que se puede observar todas las juntas en un recinto.

Hemos dividido los recintos electorales en dos grupos. Los recintos grandes y los recintos pequeños. En el grupo de los recintos grandes se realizará un censo, es decir, no habrá error de muestreo en este grupo. En el grupo de recintos pequeños se tomará una muestra, la cual tendrá un margen de error de muestreo.

[1] *Vera Francisco, Ph. D., Profesor de la Escuela Superior Politécnica del Litoral (ESPOL). (e_mail: fvera@espol.edu.ec).*

Al final se combina la información de ambos grupos para obtener un solo estimador, y se calcula el error de este estimador.

3. SELECCIÓN DE LA MUESTRA

La muestra se seleccionó usando muestreo por conglomerados estratificado (Thompson, 2012). En cada estrato (provincias y circunscripciones) se toma una muestra aleatoria de los recintos electorales. En cada conglomerado (recinto), se realiza un censo (no una muestra).

La muestra es tomada solo del grupo de recintos pequeños. Los recintos grandes no son considerados al seleccionar la muestra.

Nomenclatura

A continuación describimos los símbolos empleados en las expresiones matemáticas y algunas de las relaciones entre los mismo.

i: Grupo (1: recintos grandes, 2: recintos pequeños).

N_i: Total de electores en grupo i

M_i: Total de recintos en grupo i

j: Provincia o circunscripción (1 a 31)

N_{ij}: Total de electores en grupo i provincia o circunscripción j.

M_{ij}: Total de recintos en grupo i provincia o circunscripción j.

$$N_i = \sum_{j=1}^{31} Nij, i = 1,2$$

$$M_i = \sum_{j=1}^{31} Mij, i = 1,2$$

N_{ijk}: Total de electores en grupo i provincia o circunscripción j recinto $k (k = 1,...,M_{ij})$

$$N_{ij} = \sum_{k=1}^{M_{ij}} N_{ijk}$$

N_{ijkc}: Total de electores en grupo i provincia o circunscripción j recinto k que votan por candidato c ($c = 1,...,11$, incluyendo 8 candidatos, votos nulos, blancos y ausentes)

$N_{ij\cdot c}$: Total de electores en grupo i provincia o circunscripción j que votan por candidato c

$$N_{ij\cdot c} = \sum_{k=1}^{M_{ij}} N_{ijkc}$$

$N_{i\cdot\cdot c}$: Total de electores en grupo i que votan por candidato c

$$N_{i\cdot c} = \sum_{j=1}^{31} N_{ij\cdot c}$$

V_{ijk}: Total de votos válidos en grupo i provincia o circunscripción j recinto k

$$V_{ijk} = \sum_{c=1}^{8} N_{ijkc}$$

V_{ij}: Total de votos válidos en grupo i provincia o circunscripción j

$$V_{ij} = \sum_{c=1}^{8} N_{ij\cdot c}$$

V_i: Total de votos válidos en grupo i

$$V_i = \sum_{c=1}^{8} N_{i\cdot\cdot c}$$

V: Total de votos válidos

$$V = V_1 + V_2$$

T_c: Total de electores que votan por candidato c

$$T_c = N_{1\cdot\cdot c} + N_{2\cdot\cdot c}$$

p_c: Proporción de electores (de los votos válidos) que votan por candidato c

$$p_c = \frac{T_c}{V}$$

m: Total de recintos en muestra de grupo 2 (los recintos pequeños)

n: Total de electores en muestra de grupo 2

m_j: Total de recintos en muestra de grupo 2 provincia o circunscripción j

n_j: Total de electores en muestra de grupo 2 provincia o circunscripción j

n_{jc}: Total de electores en muestra de grupo 2 provincia o circunscripción j que votan por candidato c

v_j: Total de votos válidos en muestra de grupo 2 provincia o circunscripción j

$$v_j = \sum_{c=1}^{8} n_{jc}$$

4. ESTIMACIÓN DE TOTALES Y PROPORCIONES POR CANDIDATO

Para el cálculo de T_c necesitamos $N_{1\cdot\cdot c}$ y $N_{2\cdot\cdot c}$. El primero se obtendrá su valor exacto por censo, mientras que el segundo se estimará a partir de una muestra. El estimador del total sería

$$\hat{T}c = N_{1\cdot\cdot c} + \hat{N}_{2\cdot\cdot c}$$

El sombrero denota que no es el valor exacto sino el estimado a partir de la muestra. Por esa razón el primer término del lado derecho de la igualdad no tiene sombrero.

Para estimar el segundo total se utiliza la expresión matemática $\hat{N}_{2\cdot c} = \sum_{j=1}^{31} N_{2j}\left(\frac{n_{jc}}{n_j}\right)$

Para estimar V necesitamos V_1 y V_2. El primero se obtendrá su valor exacto por censo, mientras que el segundo se estimará a partir de una muestra. El estimador del total de votos válidos sería $\widehat{V} = V_1 + \widehat{V_2}$. Para estimar el segundo total se utiliza

$$\widehat{V_2} = \sum_{j=1}^{31} N_{2j}\left(\frac{v_j}{n_j}\right)$$

La estimación de la proporción de electores que votan por el candidato c es

$$\hat{p}_c = \frac{\hat{T}_c}{\hat{V}}$$

5. ANÁLISIS DEL MARGEN DE ERROR DE LA PROPORCIÓN POR CANDIDATO

En el muestreo por conglomerados es difícil obtener el margen de error del estimador seleccionado cuando los conglomerados (los recintos) tienen distintos tamaños. Así que procedimos a estimar el margen de error usando la técnica de simulación de Monte Carlo. En esta

estimación suponemos un desconocimiento del comportamiento del electorado.

Una vez que se tenga tanto la información del censo de los recintos grandes, como de la muestra de los recintos pequeños, se procede a estimar el margen de error utilizando la técnica de bootstrap (Efron, 1979).

Para la estimación previa a la elección, se supuso que todas las categorías, los 8 candidatos más los votos nulos y blancos, tenían igual probabilidad en cada voto. La muestra fue estratificada por provincias, y en cada provincia se procedió a escoger un número de conglomerados. Se repitió la simulación mil veces, y se obtuvieron mil posibles errores de muestreo. El error estándar de los errores fue de 0.006, mientras que el máximo fue 0.025. El percentil 95 fue de 0.012, indicando que la probabilidad que el error de muestreo sea menor a 1.2 puntos porcentuales es 0.95. Este error es para la estimación a nivel nacional

La noche de las elecciones, al tener los datos reales que iban llegando del conteo rápido, se procedió a realizar las simulaciones basadas en bootstrap. Por cuestiones logísticas no se pudo obtener todos los recintos en la muestra seleccionada, pero se obtuvo un censo para los recintos con más de 53 mesas.

En la muestra final, el margen de error a nivel nacional fue ahora mucho más bajo, 0.0038, es decir, 0.38 puntos porcentuales. Se proyectó con este margen de error que la proporción del candidato ganador era 56.47% de los votos.

Se obtuvo proyecciones a nivel provincial, obteniendo los mayores márgenes de error en las provincias de Bolívar, Morona Santiago, Napo, Orellana, Sucumbíos y Zamora Chinchipe, las cuales tuvieron márgenes de error superiores a 4 puntos porcentuales, pero menores a 10.

Las provincias de Guayas, Manabí, Pichincha y Azuay, tuvieron márgenes de error de 0.62, 1.04, 0.82 y 1.56 puntos porcentuales, respectivamente.

6. CONCLUSIONES

La estimación del margen de error exacta en muestreo por conglomerados, cuando estos son de distintos tamaños, es un problema np-hard, que resulta casi imposible calcularlos. La técnica del bootstrap es una buena alternativa y da estimaciones razonables.

ANEXO: CÓDIGO EN R

```
#Parámetros
corte=53 #Mayores a este número son grandes
basefile="Recintos.xlsx"
muestrafile="CalculoError.xlsx"
varnames=c("EMITIDOS","VALIDOS","NULOS","BLANCOS",
          "LUCIO.GUTIERREZ","ALVARO.NOBOA",
          "NELSON.ZAVALA","ALBERTO.ACOSTA",
          "GUILLERMO.LASSO","MAURICIO.RODAS",
          "NORMAN.WRAY","RAFAEL.CORREA")

#Trabajo
library(xlsx)

remuestreo=function(pob,mues,index,pmuestra)
{
 res=matrix(0,nrow=nrow(pob),ncol=ncol(pmuestra))
 colnames(res)=pvarnames
 res=as.data.frame(res)
 if(nrow(pmuestra)>0)
 {
  res[index,]=pmuestra
  res[!index,]=pmuestra[sample.int(nrow(pmuestra),nrow(pob)-nrow(mues),TRUE),]
 }
 res
}
muestrear=function(res,grandes,ng,np)
{
 i=1:nrow(res)
 ig=i[grandes]
 mg=res[ig[sample.int(length(ig),ng)],]
 ip=i[!grandes]
 mp=res[ip[sample.int(length(ip),np)],]
 rbind(mg,mp)
}

estimar=function(pob,mues,grandes)
{
 ig=mues$COD_RECINTO %in% pob[grandes,]$COD_RECINTO
 rg=apply(mues[ig,varnames],2,sum)
 sg=sum(mues[ig,"TOT_INSCRITOS"])
 if(sg>0)
  rg=(rg/sg)*sum(pob[grandes,"TOT_INSCRITOS"])
 rp=apply(mues[!ig,varnames],2,sum)
 sp=sum(mues[!ig,"TOT_INSCRITOS"])
 if(sp>0)
  rp=(rp/sp)*sum(pob[!grandes,"TOT_INSCRITOS"])
 rp+rg
}

estprop=function()
{
 rescir=t(mapply(estimar,basecir,muestracir,grandescir))
 pestcir=cbind(rescir[,1:4]/Ncir,rescir[,-(1:4)]/rescir[,"VALIDOS"])
 totest=colSums(rescir)
 N=sum(Ncir)
 pest=c(totest[1:4]/N,totest[-(1:4)]/totest["VALIDOS"])
```

```
 rbind(Nacional=pest,pestcir)
}

error=function(i)
{
 remuestra=mapply(remuestreo,basecir,muestracir,indexcir,pmuestracir,SIMPLIFY=FALSE)
 rescir=mapply(function(pob,remues)
   {
     r=pob$TOT_INSCRITOS*remues
     colnames(r)=varnames
     r
   },basecir,remuestra,SIMPLIFY=FALSE)
 rescir=mapply(function(pob,res)cbind(pob,res),basecir,rescir,SIMPLIFY=FALSE)
 muescir=mapply(muestrear,rescir,grandescir,ngrandescir,npequeñoscir,SIMPLIFY=FALSE)
 rescir=t(mapply(estimar,basecir,muescir,grandescir))
 colnames(rescir)=varnames
 pestcir=cbind(rescir[,1:4]/Ncir,rescir[,-(1:4)]/rescir[,"VALIDOS"])
 rcir=t(mapply(function(pob,remues)colSums(pob$TOT_INSCRITOS*remues),basecir,remuestra))
 pcir=cbind(rcir[,1:4]/Ncir,rcir[,-(1:4)]/rcir[,"pVALIDOS"])
 ecir=pcir-pestcir
 totest=colSums(rescir,na.rm=TRUE)
 tot=colSums(rcir,na.rm=TRUE)
 N=sum(Ncir)
 pest=c(totest[1:4]/N,totest[-(1:4)]/totest["VALIDOS"])
 p=c(tot[1:4]/N,tot[-(1:4)]/tot["pVALIDOS"])
 e=pest-p
 rbind(Nacional=e,ecir)
}

transf=function(cir)
{
 cir=as.character(cir)
 i1=cir=="Guayas Circ. 1"
 i2=cir=="Guayas Circ. 2"
 i3=cir=="Guayas Circ. 3"
 i4=cir=="Guayas Circ. 4"
 i=i1|i2|i3|i4
 cir[i]="Guayas"
 i1=cir=="Manabi Circ. 1"
 i2=cir=="Manabi Circ. 2"
 i=i1|i2
 cir[i]="Manabí"
 i1=cir=="Pichincha Circ. 1"
 i2=cir=="Pichincha Circ. 2"
 i3=cir=="Pichincha Circ. 3"
 i4=cir=="Pichincha Resto"
 i=i1|i2|i3|i4
 cir[i]="Pichincha"
 cir=as.factor(cir)
}

base=read.xlsx(basefile,1,encoding="UTF-8")
base$CIRCUNSCRIPCION=transf(base$CIRCUNSCRIPCION)

#Correr desde aquí cada vez que se actualiza la muestra

muestra=read.xlsx(muestrafile,1,encoding="UTF-8")
muestra$CIRCUNSCRIPCION=transf(muestra$CIRCUNSCRIPCION)

basecir=split(base,base$CIRCUNSCRIPCION)
```

```
muestracir=split(muestra,muestra$CIRCUNSCRIPCION)
indexcir=mapply(function(pob,mues)pob$COD_RECINTO %in%
mues$COD_RECINTO,basecir,muestracir)
pmuestracir=lapply(muestracir,function(x)x[,varnames]/x$TOT_INSCRITOS)
pvarnames=paste("p",varnames,sep="")
grandescir=lapply(basecir,function(x)x$JUN_JUNR>corte)
ngrandescir=mapply(function(i,j) sum(i&j),indexcir,grandescir)
npequeñoscir=mapply(function(i,j) sum(i&!j),indexcir,grandescir)
Ncir=sapply(basecir,function(x)sum(x$TOT_INSCRITOS))

ep=estprop()
es<-lapply(1:200,error)
efin=array(dim=c(dim(es[[1]]),length(es)))
for(i in 1:length(es)) efin[,,i]=es[[i]]
dimnames(efin)=c(dimnames(es[[1]]),list(1:length(es)))
marg=apply(efin,c(1,2),quantile,probs=0.95,na.rm=TRUE)

#wb=createWorkbook()
wb=loadWorkbook("Análisis de Error.xlsx")
wbSheets=getSheets(wb)
#sheetEst=createSheet(wb,sheetName="Proyección")
cs1=CellStyle(wb,dataFormat=DataFormat("0,0%"))
l1=NULL
for(i in 1:ncol(ep))l1=c(l1,list(cs1))
#addDataFrame(ep,sheetEst,colStyle=l1)
addDataFrame(ep,wbSheets[[1]])
#sheetMar=createSheet(wb,sheetName="Margen de Error")
cs2=CellStyle(wb,dataFormat=DataFormat("0,00%"))
l2=NULL
for(i in 1:ncol(marg))l2=c(l2,list(cs2))
#addDataFrame(marg,sheetMar,colStyle=l2)
addDataFrame(marg,wbSheets[[2]])
saveWorkbook(wb,"Análisis de Error.xlsx")
```

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. **THOMPSON, S.** (2012). Sampling, 3rd edition. Wiley series in probability and statistics, Wiley, New Jersey.

[2]. **EFRON, B.** (1979). "*Bootstrap methods: another look at the jackknife*". The annals of Statistics, 7, 1-26.