Un ejemplo de muestreo equilibrado

Yves Tillé
University of Neuchâtel

Euskal Estatistika Erakundea
XXIII Seminario Internacional de Estadística
November 2010

Los datos

- Hemos aplicado el método del cubo sobre une base de datos del Instituto Vasco de Estadística. Las unidades estadísticas son las secciones o barrios.
- Las variables son topo demográficas.
- Las tres provincias del país vasco contienen 1717 secciones en las cuales viven 2036795 habitantes.
- En nuestro ejemplo, hemos seleccionado 80 secciones con probabilidades desiguales proporcionales al número de habitantes en las secciones.
- El código en lenguaje R, muestra la agregación de variables que hemos usado para equilibrar la muestra. Las variables usadas son presentadas.
- La variable 'provincia' contiene tres columnas con las variables indicadoras de las provincias vascas.
- La variable 'provinciapop' también contiene tres columnas.
- Cada columna corresponde a una provincia y contiene un cero si la sección no está en la provincia o el número de habitantes si la sección está en la provincia.

Lista de los variables para equilibrar

UNO	variable constante que vale 1 sobre todas las secciones
total	Número de habitantes en la sección
Varones	Número de varones
Mujeres	Número de mujeres
Edad24	Número personne de menos de 25 años
Edad2565	Número personne de menos entre 25 y 65 años
Edad65m=	Número personne de mas de 65 años
Soltero	Número de solteros
Casado	Número de casados
resto	Número otro estado civil
ocupados	Número de ocupados
parados	Número de parados
inactivos	Número de inactivos
estudiosin	Sin estudio
estudioprim	Nivel primario
estudiosec	Nivel secundario
estudiosup	Nivel superior
provincia	Tres variables indicatores (0 o 1) de las provincias
provinciapop	Tres variables total $\times$ provincia

Lista de los variables disponibles

Sexo y edad	
edad0	15 años y menos, varones
edad1	16-24 años, varones
edad2	25-44 años, varones
edad3	45-64 años, varones
edad4	65 años y más, varones
edad5	15 años y menos, mujeres
edad6	16-24 años, mujeres
edad7	25-44años , mujeres
edad8	45-64 años, mujeres
edad9	65 años y más, mujeres

Lista de los variables disponibles

Estado civil y edad	
ECIV1	Solteros
ECIV2	Solteras
ECIV3	Casados
ECIV4	Casadas
ECIV5	resto varones
ECIV6	resto mujeres

Lista de los variables disponibles

Nivel de instrucción, o grado más elevado de estudios realizados o en curso,	
nivi0	menores no clasificables
nivi1	analfabetos
nivi2	sin estudios
nivi3	preescolar-primarios
nivi4	formación profesional
nivi5	secundarios
nivi6	medio-superiores
nivi7	superiores

Lista de los variables disponibles

Profesión	
prof0	No trabaja, ni ha trabajado
prof1	Director Gerente
prof2	Profesional Técnico
prof3	Técnico de Apoyo
prof4	Empleado Administrativo
prof5	Comerciante, Camarero
prof6	Agricultor, Pescador
prof7	Trabajador cualificado
prof8	Operador maquinaria
prof9	Trabajador no Cualificado

Lista de los variables disponibles

Relación con la actividad por sexo	
rel1	ocupados
rel2	ocupadas
rel3	parados
rel4	paradas
rel5	inactivos
rel6	inactivas

Lista de los variables disponibles

Rama de actividad	
ract0	no trabaja ni ha trabajado
ract1	Agricultura, ganadería, silvicultura y pesca
ract2	Industria y energía
ract3	Construcción
ract4	Comercio, reparación, hostelería, transporte y comunicaciones
ract5	Actividades financieras e inmobiliarias y servicios a empresas
ract6	Otras actividades de servicios

Lista de los variables disponibles

Situación profesional	
spr0	no trabaja ni ha trabajado
spr1	empresario
spr2	autónomo
spr3	cooperativista
spr4	asalariado fijo
spr5	asalariado eventual
spr6	ayuda familiar

Lista de los variables disponibles

Nivel global de euskera	
ekn0	menores no clasificados (menos de 2 años)
ekn1	euskaldun entienden y hablan bien euskera
ekn2	cuasi-euskaldun entienden bien o con dificultad el euskera
ekn3	erdaldun no entienden ni hablan euskera

Lista de los variables disponibles

Lengua materna	
len1	euskera
len2	castellano
len3	las dos lenguas
len4	otra lengua

Lista de los variables disponibles

Comunidad Autónoma de nacimiento	
can1	Euskadi
can2	Navarra
can3	Andalucía
can4	Castilla y León
can5	Extremadura
can6	Galicia
can7	La Rioja
can8	Resto de las comunidades autónomas
can9	Nacionalidad extranjera (esta modalidad es nueva respecto al 2001)

Lista de los variables disponibles

Año de construcción de la vivienda	
acon1	Antes de 1900
acon2	Entre 1901 y 1940
acon3	Entre 1941 y 1950
acon4	Entre 1951 y 1960
acon5	Entre 1961 y 1970
acon6	Entre 1971 y 1980
acon7	Entre 1981 y 1990
acon8	En 1991 o posterior

Lista de los variables disponibles

Superficie útil de la vivienda	
supf1	Igual o menos de 60 m2
supf2	Entre 61 y 90 m2
supf3	Entre 91 y 120 m2
supf4	Entre 121 y 150 m2
supf5	Entre 151 y 180 m2
supf6	181 m2 o más

Lista de los variables disponibles

Número de personas residentes en la vivienda	
tafam1	1 persona
tafam2	2 personas
tafam3	3-5 personas
tafam4	6 y más personas

```
#
# Hay que cambiar el directorio corriente
# Lectura de los datos
D=read.table("a.txt",header = TRUE)
attach(D)
#
Varones=edad0+edad1+edad2+edad3+edad4
Mujeres=edad5+edad6+edad7+edad8+edad9
Edad24=edad0+edad1+edad5+edad6
Edad2565=edad2+edad3+edad7+edad8
Edad65m=edad4+edad9
Soltero=eciv1+eciv2
Casado=eciv3+eciv4
resto=eciv5+eciv6
ocupados=rel1+rel2
parados =rel3+rel4
inactivos=rel5+rel6
estudiosin= nivi0+nivi1+nivi2
estudioprim=nivi3+nivi4
estudiosec=nivi5+nivi6
estudiosup=nivi7
provincia=disjunctive(as.integer(seccion/100000000))
colnames(provincia)<-c("secpro1","secpro2","secpro3")
provinciapop=provincia*total
colnames(provinciapop)<-c("pro1pop","pro2pop","pro3pop")
```

```
#
# calculo de los variables
#
pik=inclusionprobabilities(total,80)
#
# creacion de la matriz de los variables de equilibrio
#
UNO=rep(1,length(total))
X=cbind(UNO,provincia,total,provinciapop,Varones,Mujeres,Edad2
Soltero,Casado,resto,ocupados,parados,inactivos,estudiosin,
estudioprim,estudiosec,estudiosup)
```

Selección de la muestra

- Para seleccionar la muestra, hay que instalar el paquete 'sampling'.
- También, hay que cargar este paquete con el comando 'library(sampling)'.
- Los datos son cargados en la matriz D. Las nuevas variables son calculadas a partir de los variables de la base de datos.
- Después hemos calculado las probabilidades de inclusión proporcionales a la variable total para una muestra de tamaño igual a 80.
- Hemos creado un variable constante que vale 1 en todas partes.
- La matriz X contiene las 23 variables de equilibrio. Algunas variables son redundantes, lo que no es un problema para usar la función 'samplecube'.
- La función 'samplecube' selecciona un muestra equilibrada sobre los variables X con probabilidades de inclusión π_k.

```
#
# Cargar el paquete sampling
#
library(sampling)
#
# Selecccion de une muestra equilibrada
#
s=samplecube(X,pik,method=1)
```

El resultado de la función 'samplecube' describe la población (23 variables de equilibrio y 1717 secciones) y el vector de probabilidades de inclusión.

```
> s=samplecube(X,pik,method=1)

BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 23  variables and  1717  units
The size of the inclusion probability vector is  1717
The sum of the inclusion probability vector is  80
The inclusion probability vector has  1717  non-integer elements
Step 1  Step  2,


BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain  17 non integer probabi
The sum of these probabilities is  7
This sum is  integer
The linear program will consider  19448  possible samples
The mean cost is  0.03246569
The smallest cost is  0.003944129
The largest cost is  0.07663838
The cost of the selected sample is 0.005887202
```

```
QUALITY OF BALANCING
             TOTALS HorvitzThompson_estimators Relative_deviation
UNO            1717                1694.6036       -1.304392357
secpro1         248                 256.1786        3.297830061
secpro2         544                 507.2046       -6.763864729
secpro3         925                 931.2204        0.672474464
total       2036795             2036795.0000        0.000000000
pro1pop      281205              280059.3125       -0.407420743
pro2pop      648219              636498.4375       -1.808117704
pro3pop     1107371             1120237.2500        1.161873482
Varones      997334              996655.1796       -0.068063496
Mujeres     1039461             1040139.8204        0.065305036
Edad24       494021              494345.4598        0.065677337
Edad2565    1183026             1181787.7680       -0.104666505
Edad65m      359748              360661.7721        0.254003393
Soltero      881212              880188.3255       -0.116166658
Casado       966664              966750.7062        0.008969635
resto        188919              189855.9683        0.495962979
ocupados     846020              843975.6520       -0.241642991
parados      111620              111658.8241        0.034782418
inactivos   1079155             1081160.5239        0.185842062
estudiosin   246500              247256.4153        0.306862188
estudioprim 1051551             1052369.1752        0.077806516
estudiosec   490458              490298.7225       -0.032475263
estudiosup   248275              246870.6870       -0.565628020
```

- La función 'balancedstratification' aplica una fase de vuelo en cada estrato.
- Después, una fase global de estratificación es aplicada sobre todos los estratos.
- Al final, la fase de aterrizaje es aplicada sobre toda la población.

```
#
# Selecccion de une muestra equilibrada
# estratificada sobre los provincias
#
pro=cleanstrata(as.integer(seccion/1000000000))
s=balancedstratification(X,pro,pik)
```

```
> s=balancedstratification(X,pro,pik)

FLIGHT PHASE OF STRATUM 1
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 24  variables and  248  units
The size of the inclusion probability vector is  248
The sum of the inclusion probability vector is  11.045
The inclusion probability vector has  248  non-integer elements
Step 1  Step  2,

FLIGHT PHASE OF STRATUM 2
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 24  variables and  544  units
The size of the inclusion probability vector is  544
The sum of the inclusion probability vector is  25.46035
The inclusion probability vector has  544  non-integer elements
Step 1  Step  2,

FLIGHT PHASE OF STRATUM 3
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 24  variables and  925  units
The size of the inclusion probability vector is  925
The sum of the inclusion probability vector is  43.49465
The inclusion probability vector has  925  non-integer elements
Step 1  Step  2,
```

```
FINAL TREATMENT
BEGINNING OF THE FLIGHT PHASE
The matrix of balanced variable has 26  variables and  1717  units
The size of the inclusion probability vector is  1717
The sum of the inclusion probability vector is  80
The inclusion probability vector has  39  non-integer elements
Step 1  Step  2,


BEGINNING OF THE LANDING PHASE
At the end of the flight phase, there remain  17 non integer probab
The sum of these probabilities is  9
This sum is  integer
The linear program will consider  24310  possible samples
The mean cost is  0.9535204
The smallest cost is  0.1108615
The largest cost is  2.545963
The cost of the selected sample is 0.1162985
```

```
QUALITY OF BALANCING
                TOTALS  HT_estimators Relative_deviation
Stratum1         11.05          11.00           -0.4074
Stratum2         25.46          25.00           -1.8081
Stratum3         43.49          44.00            1.1619
UNO               1717        1737.39            1.1881
secpro1            248         275.05           10.9079
secpro2            544         537.24           -1.2413
secpro3            925         925.09            0.0108
total          2036795     2036795.00            0.0000
pro1pop         281205      280059.31           -0.4074
pro2pop         648219      636498.43           -1.8081
pro3pop        1107371     1120237.25            1.1619
Varones         997334      997914.73            0.0582
Mujeres        1039461     1038880.26           -0.0559
Edad24          494021      494628.40            0.1230
Edad2565       1183026     1184683.36            0.1401
Edad65m         359748      357483.22           -0.6295
Soltero         881212      882495.63            0.1457
Casado          966664      965694.73           -0.1003
resto           188919      188604.62           -0.1664
ocupados        846020      848311.10            0.2708
parados         111620      111184.40           -0.3903
inactivos      1079155     1077299.49           -0.1719
estudiosin      246500      245757.25           -0.3013
estudioprim    1051551     1047897.83           -0.3474
estudiosec      490458      492006.43            0.3157
estudiosup      248275      251133.47            1.1513
```