ANÁLISIS DE MUESTRAS COMPLEJAS EN R

María Eugenia Riaño

1. ¿Qué es una muestra compleja?

Una muestra compleja es una muestra obtenida mediante un diseño que involucre **estratos, conglomerados, y/o etapas de selección**.

Un diseño complejo implica, en la mayoría de los casos, probabilidades de selección diferentes entre las unidades de muestreo.

¿Cómo analizar datos provenientes de una muestra compleja?

Los métodos estadísticos a utilizar deben de tener en cuenta el **diseño muestral:**

- Los errores estándar usuales, que asumen un muestreo simple con reposición, serán incorrectos si los datos provienen de una muestra compleja.

Ejemplo:

En términos de varianza, un diseño por conglomerados puede ser menos eficiente que un diseño simple. Si la variable de interés es homogénea dentro del conglomerado, y si los errores estándar se calculan asumiendo un diseño simple, se subestimará la verdadera varianza poblacional, pudiendo llevar a conclusiones erróneas sobre el comportamiento de los parámetros de interés.

2. El paquete Survey de R

Creado por **Thomas Lumley** en el 2003

http://r-survey.r-forge.r-project.org/survey/

Versión actual **3.32** (ha tenido 97 actualizaciones!!)

Journal of Statistical Software, 2004 (versión 2.2)

Complex Surveys: A guide to Analysis using R, Wiley 2010

Características del paquete Survey

- Cálculo de promedios, totales, razones, cuantiles, tablas de contingencia, modelos de regresión, entre otros para la muestra completa y para dominios.
- Las varianzas se calculan utilizando linearización de Taylor o con técnicas de remuestreo (Bootstrap, Jackknife)
- Post Estratificación, estimadores de raking generalizado, calibración.
- Diseños en dos fases.

Utilizando el paquete Survey de R

`svydesign` especifica el diseño muestral

Argumentos

`id` indica las unidades de muestreo (PSUs, muestreo directo de elementos)

`data` base de datos con la que se va a trabajar

`strata` indica los estratos

`weights` indica los pesos muestrales

`fpc` indica si se deben realizar correcciones por poblaciones finitas

Utilizando el paquete Survey de R

Ejemplo: Encuesta Continua de Hogares

- Estratos: geográficos.

- Dos etapas de selección: unidades primarias de muestreo, zonas censales. Unidades de segunda etapa, viviendas particulares ocupadas. Probabilidades proporcionales al tamaño en la primer etapa de selección.

- Se utilizan estimadores de raking generalizado que ajustan los pesos muestrales a totales poblacionales por sexo y tramo etario.

Utilizando el paquete Survey de R

Diseño estratificado con pesos muestrales diferentes

```
p.s=svydesign(id=~1, strata=~region_3, weights=~pesomen,
data= hog)
```

Se crea un objeto que contiene los datos y la información del diseño de la muestra.

Utilizando el paquete Survey de R

Diseño estratificado en etapas con pesos muestrales diferentes

```
p.c=svydesign(id=~codsegm+numero, strata=~region_3,
weights=~pesomen, data= hog, nest=TRUE)
```

La función **svydesign** genera un "entorno" con funciones propias:

sum() es **svytotal()**
mean() es **svymean()**
glm() es **svyglm()**

Utilizando el paquete Survey de R

Estimación de promedios

```
>svymean(~HT11,p.s)
       mean     SE
HT11 69536 890.36

>svymean(~HT11,p.c)
      mean     SE
HT11 69536  1374.2
```

Utilizando el paquete Survey de R

Tablas

```
> svytable(~pobre06+dpto,p.s)

          dpto
pobre06 Montevideo Artigas Canelones Cerro Largo
No pobre 478597     21995   187008     29068
Pobre    32196      1703    9916       2258

..............
```

Utilizando el paquete Survey de R

Dominios

```
> svyby(~pobre06,~dpto,p.s,svytotal)

dpto      pobre06Nopobre pobre06Pobre se.pobre06Nopobre se.pobre06Pobre
Montevideo      478597         32196       3943.301         3372.3688
Artigas          21995          1703       2688.271          852.5694
Canelones       187008          9916       7034.014         1911.2359
Cerro Largo      29068          2258       3303.456         1015.5243
```

.....

Utilizando el paquete Survey de R

Modelos Lineales

```
> m=svyglm(HT11~d21_14_1+d21_15_4+d21_18_1,p.s)
> summary(m)
Call:
svyglm(formula = HT11 ~ d21_14_1 + d21_15_4 + d21_18_1, p.s)

Survey design:
svydesign(id = ~1, strata = ~region_3, weights = ~pesomen, data = hog)

Coefficients:

             Estimate Std. Error  t value Pr(>|t|)
(Intercept)   34272.9      899.6   38.099 <2e-16 ***
d21_14_1      10929.7     1302.5  8.392  <2e-16 ***
d21_15_4      22309.0     1183.9  18.844 <2e-16 ***
d21_18_1      24432.1     1684.3  14.506 <2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1653659940)

Number of Fisher Scoring iterations: 2
```

GRACIAS!!