Análisis de Encuestas Complejas (usando R)

Lectura 3 - Estimación e Inferencia en Poblaciones Finitas: Muestreo Probabilístico

Universidad de Santiago de Chile

Miguel Alvarado

November 4, 2020

Outline

Inferencia: Model-based vs Design-based

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

- En la medida que el modelo de probabilidad representa el proceso que generó los datos, es posible extraer conclusiones que pueden ser generalizadas hacia otras situaciones donde opera el mismo proceso que generó los datos.

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

- En la medida que el modelo de probabilidad representa el proceso que generó los datos, es posible extraer conclusiones que pueden ser generalizadas hacia otras situaciones donde opera el mismo proceso que generó los datos.
- Los datos observados corresponden a realizaciones de una variable aleatoria que sigue alguna distribución de probabilidad.

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

- En la medida que el modelo de probabilidad representa el proceso que generó los datos, es posible extraer conclusiones que pueden ser generalizadas hacia otras situaciones donde opera el mismo proceso que generó los datos.
- Los datos observados corresponden a realizaciones de una variable aleatoria que sigue alguna distribución de probabilidad.

Design-based Inference: el investigador especifica una población (fija), donde cuyas características son desconocidas pero se asumen fijos: no aleatorios. Sin embargo, la muestra observada es aleatoria, pues depende de la selección aleatoria ($p(\cdot)$) de elementos que provienen de esta población.

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

- En la medida que el modelo de probabilidad representa el proceso que generó los datos, es posible extraer conclusiones que pueden ser generalizadas hacia otras situaciones donde opera el mismo proceso que generó los datos.
- Los datos observados corresponden a realizaciones de una variable aleatoria que sigue alguna distribución de probabilidad.

Design-based Inference: el investigador especifica una población (fija), donde cuyas características son desconocidas pero se asumen fijos: no aleatorios. Sin embargo, la muestra observada es aleatoria, pues depende de la selección aleatoria $(p(\cdot))$ de elementos que provienen de esta población.

- Su propósito es estimar características de una población fija y la inferencia no permite generalizar los resultados hacia otras poblaciones.

Inferencia: Model-based vs Design-based

Model-based Inference: el investigador especifica un modelo de probabilidad para el proceso *aleatorio* que genera los datos.

- En la medida que el modelo de probabilidad representa el proceso que generó los datos, es posible extraer conclusiones que pueden ser generalizadas hacia otras situaciones donde opera el mismo proceso que generó los datos.
- Los datos observados corresponden a realizaciones de una variable aleatoria que sigue alguna distribución de probabilidad.

Design-based Inference: el investigador especifica una población (fija), donde cuyas características son desconocidas pero se asumen fijos: no aleatorios. Sin embargo, la muestra observada es aleatoria, pues depende de la selección aleatoria $(p(\cdot))$ de elementos que provienen de esta población.

- Su propósito es estimar características de una población fija y la inferencia no permite generalizar los resultados hacia otras poblaciones.
- Los datos observados corresponden a parámetros poblacionales fijos (no aleatorios).

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

UdeSantiago de Chile

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.
- $\forall s \in Q$: $p(\cdot)$ (probabilidades de selección) son conocidas.

UdeSantiago de Chile

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.
- $\forall s \in Q$: $p(\cdot)$ (probabilidades de selección) son conocidas.

Muestreo Aleatorio Simple: Cualquier muestrea $(s \subset U)$ de n elementos desde una población de tamaño N tiene igual probabilidad de selección: $\forall s \in Q : p(s) = p$.

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.
- $\forall s \in Q$: $p(\cdot)$ (probabilidades de selección) son conocidas.

Muestreo Aleatorio Simple: Cualquier muestrea ($s \subset U$) de n elementos desde una población de tamaño N tiene igual probabilidad de selección: $\forall s \in Q : p(s) = p$.

Las propiedades que se requieren del método de muestreo para *Design-based Inference*:

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.
- $\forall s \in Q$: $p(\cdot)$ (probabilidades de selección) son conocidas.

Muestreo Aleatorio Simple: Cualquier muestrea ($s \subset U$) de n elementos desde una población de tamaño N tiene igual probabilidad de selección: $\forall s \in Q : p(s) = p$.

Las propiedades que se requieren del método de muestreo para *Design-based Inference*:

- $\pi_k > 0, \forall k \in s$.
- π_k debe ser conocido $\forall k \in s$.

Muestra Probabilística: Muestreo Aleatorio Simple

El concepto estadístico fundamental en *Design-based Inference* es el de *muestra probabilística* o *muestra aleatoria*.

- Soporte Q.
- $\forall s \in Q$: $p(\cdot)$ (probabilidades de selección) son conocidas.

Muestreo Aleatorio Simple: Cualquier muestrea $(s \subset U)$ de n elementos desde una población de tamaño N tiene igual probabilidad de selección: $\forall s \in Q : p(s) = p$.

Las propiedades que se requieren del método de muestreo para *Design-based Inference*:

- $\pi_k > 0, \forall k \in s$.
- π_k debe ser conocido $\forall k \in s$.
- $\pi_{kl} > 0, \forall (k, l) \in s$.
- π_{kl} debe ser conocido $\forall (k, l) \in s$.

UdeSantiago de Chile

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

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

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

En nuestro ejemplo: $n = 175740$ y $N = 17574003$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

En nuestro ejemplo: $n = 175740$ y $N = 17574003$.

- $\pi_k = \frac{n}{N} = \frac{175740}{17574003} = 0.009999998$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

En nuestro ejemplo: $n = 175740$ y $N = 17574003$.

- $\pi_k = \frac{n}{N} = \frac{175740}{17574003} = 0.009999998$.
- $d_k = \frac{1}{\pi_k} = \frac{17574003}{175740} = 100$.

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

En nuestro ejemplo: $n = 175740$ y $N = 17574003$.

- $\pi_k = \frac{n}{N} = \frac{175740}{17574003} = 0.009999998$.
- $d_k = \frac{1}{\pi_k} = \frac{17574003}{175740} = 100$.

d_k señala que cada elemento $k \in s$ representa a 100 elementos de la población (a el mismo y a otros 99 que no estan en s).

Ponderadores (Pesos) de Diseño (Básico)

El *ponderador de diseño*: $d_k = \frac{1}{\pi_k}$.

Si tomamos un MAS de $n > 0$ elementos de una población de tamaño $N > n$: $d_k = \frac{n}{N}; \forall k$.

- $\pi_k = \frac{n}{N}$.
- $d_k = \frac{1}{\pi_k} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$.

En nuestro ejemplo: $n = 175740$ y $N = 17574003$.

- $\pi_k = \frac{n}{N} = \frac{175740}{17574003} = 0.009999998$.
- $d_k = \frac{1}{\pi_k} = \frac{17574003}{175740} = 100$.

d_k señala que cada elemento $k \in s$ representa a 100 elementos de la población (a el mismo y a otros 99 que no estan en s).

El fundamento estadístico destrás de *toda* la *Design-based Inference* es que un elemento de la muestra s, que es seleccionado con probabilidad π_k representa a $\frac{1}{\pi_k}$ elementos de la población.

UdeSantiago de Chile

Ponderadores (Pesos) de Diseño (Básico)

```
# Nuestro ejemplo
rm(list=objects())
load("PoblacionSimulada.RData")
load("myFuns.RData")

# N
(N <- dim(C17)[1])

## [1] 17574003

# n
(n <- round(dim(C17)[1]*0.01, 0))

## [1] 175740

# pi_k
(pi_k <- n/N)

## [1] 0.009999998

# d_k
(d_k <- 1/pi_k)

## [1] 100

# En nuestra población
C17$pi_k <- n/N
C17$d_k <- 1/C17$pi_k
```

Ponderadores (Pesos) de Diseño (Básico)

```
# MAS
set.seed(1123)
muestra <- ma_muestra(C17, n)

# Población
dim(C17)

## [1] 17574003        8

# Muestra
dim(muestra)

## [1] 175740      8
```

Ponderadores (Pesos) de Diseño (Básico)

```r
# Población
head(C17, n = 9)

##            id region sexo gedad pet ocu        pi_k d_k
## 7265616   1      8    2     7   1   0 0.009999998 100
## 2610920   2      5    1     5   1   0 0.009999998 100
## 3382716   3      5    2     4   1   1 0.009999998 100
## 12842040  4     13    2     1   0  NA 0.009999998 100
## 9694759   5     13    1     1   0  NA 0.009999998 100
## 9958254   6     13    1     1   0  NA 0.009999998 100
## 5759220   7      8    1     1   0  NA 0.009999998 100
## 2782507   8      5    1     7   1   1 0.009999998 100
## 7857447   9      9    2     1   0  NA 0.009999998 100

# Muestra
head(muestra, n = 9)

##                 id region sexo gedad pet ocu        pi_k d_k
## 13519202  7330918     13    2     2   1   0 0.009999998 100
## 4185966   6338207      6    1     6   1   1 0.009999998 100
## 17448718 11163856     16    2     4   1   1 0.009999998 100
## 9509974   5932911     13    1     1   0  NA 0.009999998 100
## 2197546   5606305      5    1     2   1   0 0.009999998 100
## 12561213  8299022     13    1     7   1   0 0.009999998 100
## 11170529  9694055     13    1     3   1   1 0.009999998 100
## 7983993  14730298      9    2     3   1   1 0.009999998 100
## 15173605 10123174     13    2     5   1   1 0.009999998 100
```

Estimadores: Estimador de Horvitz–Thompson

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

Estimadores: Estimador de Horvitz–Thompson

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k$$

Estimadores: Estimador de Horvitz–Thompson

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k$$

Donde π_k corresponde a la *probabilidad de inclusión* del k-ésimo elemento en la muestra s y d_k es el *ponderador básico* o *ponderador de diseño* y corresponde al inverso de la probabilidad de inclusión π_k.

Estimadores: Estimador de Horvitz–Thompson

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

$$\hat{t}_{y,\pi} \quad = \quad \sum_{k \in S} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k$$

  Donde π_k corresponde a la *probabilidad de inclusión* del k-ésimo elemento en la muestra s y d_k es el *ponderador básico* o *ponderador de diseño* y corresponde al inverso de la probabilidad de inclusión π_k.

- Resultado: Si todas las probabilidades de inclusión de primer orden son mayores que cero $(\pi_k > 0, \forall k)$, el Estimador de Horvitz–Thompson es insesgado para el total poblacional de la característica de interés y. Por tanto:

Estimadores: Estimador de Horvitz–Thompson

- Si para la población U se quiere estimar el total poblacional de la característica de interés y, denotado por t_y, el *Estimador de Horvitz–Thompson* para t_y, denotado por $\hat{t}_{y,\pi}$, se define como:

$$\hat{t}_{y,\pi} = \sum_{k \in S} \frac{1}{\pi_k} y_k = \sum_{k \in S} d_k y_k$$

  Donde π_k corresponde a la *probabilidad de inclusión* del k-ésimo elemento en la muestra s y d_k es el *ponderador básico* o *ponderador de diseño* y corresponde al inverso de la probabilidad de inclusión π_k.

- Resultado: Si todas las probabilidades de inclusión de primer orden son mayores que cero ($\pi_k > 0, \forall k$), el Estimador de Horvitz–Thompson es insesgado para el total poblacional de la característica de interés y. Por tanto:

$$E\left(\hat{t}_{y,\pi}\right) = t_y$$

Estimadores: Estimador de la Varianza

- El estimador de t_y es $\hat{t}_{y,\pi}$. El *Estimador de la Varianza* de $\hat{t}_{y,\pi}$, viene dado por:

$$Var(\hat{t}_{y,\pi}) = \sum_{k,l \in S} \left(\frac{y_k y_l}{\pi_{kl}} - \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right)$$

UdeSantiago de Chile

Estimadores: Estimador de la Varianza

- El estimador de t_y es $\hat{t}_{y,\pi}$. El *Estimador de la Varianza* de $\hat{t}_{y,\pi}$, viene dado por:

$$Var(\hat{t}_{y,\pi}) = \sum_{k,l \in S} \left(\frac{y_k y_l}{\pi_{kl}} - \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right)$$

- La formula aplica para cualquier diseño muestral; por complicado que este sea.

Estimadores: Estimador de la Varianza

- El estimador de t_y es $\hat{t}_{y,\pi}$. El *Estimador de la Varianza* de $\hat{t}_{y,\pi}$, viene dado por:

$$Var(\hat{t}_{y,\pi}) = \sum_{k,l \in S} \left(\frac{y_k y_l}{\pi_{kl}} - \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \right)$$

- La formula aplica para cualquier diseño muestral; por complicado que este sea.
- La formula depende de π_{kl} y no solo de π_k.

Estimadores: Estimador del Total y su varianza

Definimos el diseño muestra de un MAS en la libreria "survey"

```
# Diseño Muestral MAS sin reemplazo
muestra.dsg <- svydesign(id=~1, weights=~d_k, data=muestra)
```

Estimamos el total para la PET y su medida de precisión (varianza).

```
# Estimamos el total de la PET y su varianza
svytotal(~pet, muestra.dsg, na.rm = TRUE)

##        total    SE
## pet 14089002 16715

# PET (parámetro)
par_pet

## [1] 14050253
```

Estimadores: Estimador del Total (factor) y su varianza

Definido el diseño muestra, no es necesario volver a definirlo. Estimamos el total por sexo y su medida de precisión (varianza).

```
# Estimamos el total por sexo y su varianza
svytotal(~ factor(sexo), muestra.dsg, na.rm = TRUE)

##                  total    SE
## factor(sexo)1 8598202 20956
## factor(sexo)2 8975802 20956

# Totales por sexo (parámetro)
table(C17$sexo)

##
##       1       2
## 8601989 8972014
```