Estrategias Transversales en las Encuestas de Hogares

Curso de Muestreo Probabilístico en Encuestas de Hogares

Andrés Gutiérrez, Ph.D.

CEPAL - Unidad de Estadísticas Sociales

1. Estimación de la varianza
2. Muestreo aleatorio simple en dos etapas estratificado
3. Muestreo autoponderado en dos etapas estratificado

Motivación

Desde que se popularizaron las encuestas de hogares en 1940, se ha hecho evidente algunas tendencias que están ligadas a los avances tecnológicos en las agencias estadísticas y en la sociedad y se han acelerado con la introducción del computador.

Gambino & Silva (2009)

Bibliografía y referencias

- Kish, L. (1965) *Survey Sampling*. John Wiley and Sons.
- Cochran, W. G. (1977) *Sampling Techniques*. John Wiley and Sons.
- Särndal, et. al. (2003) *Model-assisted Survey Sampling*. Springer.
- Gutiérrez, H. A. (2016) *Estrategias de muestreo: diseño de encuestas y estimación de parámetros*. Ediciones de la U.
- Gutiérrez, H. A. (2017) `TeachingSampling`. *R package*.

Estimación de la varianza

Desventajas

- Como se vio anteriormente, si el diseño sin reemplazo tiene r etapas es necesario hacer r cálculos de varianza.
- Además es necesario escribir las expresiones algebraicas correspondientes en cada diseño particular.
- Lo anterior en una encuesta a gran escala puede llegar a ser muy tedioso, costoso y además muy demorado.
- Este proceso no es sistemático y puede estar sujeto a errores matemáticos y computacionales.

Aproximaciones (1)

Una posible solución al problema es mantener la primera parte del estimador de la varianza como estimador general de la misma.

$$\widehat{Var}_2(\hat{t}_\pi) = \sum\sum_{sl} \frac{\Delta_{lij}}{\pi_{li}} \frac{\hat{t}_i}{\pi_{li}} \frac{\hat{t}_j}{\pi_{lj}} \tag{1}$$

- Este estimador sobre-estima la varianza para las unidades primarias de muestreo.

Aproximaciones (2)

Otra posible solución para estimar la varianza del estimador de Horvitz-Thompson, es asumir que el muestreo en la primera etapa se llevó a cabo con reemplazo. Así, la estimación (sesgada) de la varianza estaría dada por:

$$\widehat{Var}_3(\hat{t}_\pi) = \frac{1}{m_I(m_I-1)} \sum_{i=1}^{m_I} \left(\frac{\hat{t}_i}{p_{Ii}} - \hat{t}_\pi \right)^2 \qquad (2)$$

Aproximaciones (2)

Un caso especial del anterior término se tiene suponiendo que $\pi_{Ii} = n_I p_{Ii}$, si el muestreo en la primera etapa fue aleatorio simple, entonces $p_{Ii} = \frac{1}{N}$. El estimador de la varianza, bajo la anterior condición es:

$$\widehat{Var}(\hat{t}_\pi) = \frac{N_I^2}{m_I(m_I - 1)} \sum_{i=1}^{m_I} \left(\hat{t}_i - \frac{\sum_{i=1}^{m_I} \hat{t}_i}{m_I} \right)^2 = \frac{N_I^2}{m_I} S^2_{\hat{t}_i}$$

La técnica del último conglomerado

- Para la estimación de la varianza de los estimadores de interés en encuestas multi-etápicas, los programas computacionales existentes utilizan una aproximación conocida como la técnica del último conglomerado.
- Esta aproximación, que sólo tiene en cuenta la varianza de los estimadores en la primera etapa, supone que ese muestreo fue realizado con reemplazo.
- Los procedimientos de muestreo en etapas posteriores de la selección son ignorados a menos que el factor de corrección para poblaciones finitas sea importante a nivel municipal.

La técnica del último conglomerado

Se supone un diseño de muestreo en varias etapas (dos o más) en donde la primera etapa supone la selección de una muestra s_I de m_I unidades primarias de muestreo (UPM) U_i ($i \in s_I$) de tal forma que

- Si la selección se realizó con reeemplazo, la i-ésima UPM tiene probabhilidad de selección p_{I_i}.
- Si la selección se realizó sin reeemplazo, la i-ésima UPM tiene probabilidad de inclusión π_{I_i}.

La técnica del último conglomerado

En las subsiguientes etapas de muestreo, se procede a seleccionar una muestra de elementos para cada una de las UPM seleccionadas en la primera etapa de muestreo. Dentro de la i-ésima UPM se selecciona una muestra s_i de elementos.

En particular la probabilidad condicional de que el k-ésimo elemento pertenzca a la muestra dada que la UPM que la contiene ha sido seleccionada en la muestra de la primera etapa está dada por la siguiente expresión:

$$\pi_{k|i} = Pr(k \in s_i | i \in s_I)$$

La técnica del último conglomerado

Por ejemplo, si el muestreo es sin reemplazo en todas sus etapas, la probabilidad de inclusión del k-ésimo elemento a la muestra s está dada por

$$\begin{aligned}\pi_k &= Pr(k \in s) \\ &= Pr(k \in s_i, i \in s_I) \\ &= Pr(k \in s_i | i \in s_I) Pr(i \in s_I) = \pi_{k|i} \times \pi_{I_i}\end{aligned}$$

La técnica del último conglomerado

Dado que el inverso de las probabilidades de inclusión son un ponderador natural, entonces se definen las siguientes cantidades:

1. $d_{I_i} = \frac{1}{\pi_I}$, que es el factor de expansión de la i-ésima UPM.
2. $d_{k|i} = \frac{1}{\pi_{k|i}}$, que es el factor de expansión del k-ésimo elemento dentro de la i-ésima UPM.
3. $d_k = d_{I_i} \times d_{k|i}$, que es el factor de expansión final del k-ésimo elemento para toda la población U.

La técnica del último conglomerado

En general, el estimador del total toma la siguiente forma:

$$\hat{t}_y = \sum_k d_k y_k = \sum_h \sum_i \sum_k d_k y_k = \sum_h \sum_i \sum_k d_{I_i} \; d_{k|i} y_k$$

Y la aproximación al último conglomerado es:

$$\widehat{Var}(\hat{t}_{y,p}) = \sum_h \frac{m_{Ih}}{m_{Ih} - 1} \sum_i \left(\breve{t}_{y_{ih}} - \bar{\breve{t}}_{y_h} \right)^2 \tag{3}$$

En donde $\breve{t}_{y_i} = \sum_{k \in s_i} d_k y_k$ y $\bar{\breve{t}}_y = \frac{1}{m_I} \sum_{i=1}^{m_I} \breve{t}_{y_i}$

La técnica del último conglomerado

- Este procedimiento tiende a sobrestimar la varianza verdadera.
- Resulta ser una técnica apetecida por los investigadores puesto que utiliza directamente los pesos finales de muestreo o factores de expansión que son publicados por los INE.
- Si la fracción de muestreo de UPM en los estratos es significativa, es posible ajustar la varianza añadiendo un término de ajuste dentro de cada estrato.

Muestreo aleatorio simple en dos etapas estratificado

Muestreo en dos etapas estratificado

- La teoría discutida en las secciones anteriores es aplicable cuando las unidades primarias de muestreo son seleccionadas dentro de un estrato.
- No hay nuevos principios de estimación o diseño involucrado en el desarrollo de esta estrategia de muestreo.

Muestreo en dos etapas estratificado

- Se supone que el muestreo en cada estrato respeta el principio de la independencia.
- Las estimaciones del total, así como el cálculo y estimación de la varianza son simplemente resultado de añadir o sumar para cada estrato la respectiva cantidad.

Muestreo en dos etapas estratificado

- Dentro de cada estrato U_h $h = 1, \ldots, H$ existen N_{lh} unidades primarias de muestreo, de las cuales se selecciona una muestra s_{lh} de n_{lh} unidades mediante un diseño de muestreo aleatorio simple.
- Suponga, además que el sub-muestreo dentro de cada unidad primaria seleccionada es también aleatorio simple.
- Para cada unidad primaria de muestreo seleccionada $i \in s_{lh}$ de tamaño N_i se selecciona una muestra s_i de elementos de tamaño n_i.

Muestreo en dos etapas estratificado

Para utilizar los prinicpios de estimación del último conglomerado en este diseño particular se definen las siguientes cantidades:

1. $d_{I_i} = \frac{N_{Ih}}{n_{Ih}}$, que es el factor de expansión de la i-ésima UPM en el estrato h.
2. $d_{k|i} = \frac{N_i}{n_i}$, que es el factor de expansión del k-ésimo hogar para la i-ésima UPM.
3. $d_k = d_{I_i} \times d_{k|i} = \frac{N_{Ih}}{n_{Ih}} \times \frac{N_i}{n_i}$, que es el factor de expansión final del k-ésimo elemento para toda la población U.

Práctica en R

```
data('BigCity')

 FrameI <- BigCity %>% group_by(PSU) %>%
 summarise(Stratum = unique(Stratum),
           Persons = n(),
           Income = sum(Income),
           Expenditure = sum(Expenditure))

attach(FrameI)
```

Práctica en R

```
head(FrameI, 10)
```

PSU	Stratum	Persons	Income	Expenditure
PSU0001	idStrt001	118	70912	44232
PSU0002	idStrt001	136	68887	38382
PSU0003	idStrt001	96	37213	19495
PSU0004	idStrt001	88	36926	24031
PSU0005	idStrt001	110	57494	31142
PSU0006	idStrt001	116	75272	43473
PSU0007	idStrt001	68	33028	21833
PSU0008	idStrt001	136	64293	47660
PSU0009	idStrt001	122	33156	23292
PSU0010	idStrt002	70	65254	37115

Práctica en R

```
sizes = FrameI %>% group_by(Stratum) %>%
        summarise(NIh = n(),
        nIh = 2,
        dI = NIh/nIh)

NIh <- sizes$NIh
nIh <- sizes$nIh
```

Práctica en R

```
head(sizes, 10)
```

Stratum	Nlh	nlh	dl
idStrt001	9	2	4.5
idStrt002	11	2	5.5
idStrt003	7	2	3.5
idStrt004	13	2	6.5
idStrt005	11	2	5.5
idStrt006	5	2	2.5
idStrt007	14	2	7.0
idStrt008	7	2	3.5
idStrt009	8	2	4.0
idStrt010	8	2	4.0

Práctica en R

```
samI <- S.STSI(Stratum, NIh, nIh)
UI <- levels(as.factor(FrameI$PSU))
sampleI <- UI[samI]

FrameII <- left_join(sizes,
             BigCity[which(BigCity$PSU %in% sampleI), ])
attach(FrameII)
```

Práctica en R

```
head(FrameII, 10)
```

Stratum	NIh	nIh	dI	HHID	PersonID	PSU	Zone
idStrt001	9	2	4.5	idHH00001	idPer01	PSU0001	Rural
idStrt001	9	2	4.5	idHH00001	idPer02	PSU0001	Rural
idStrt001	9	2	4.5	idHH00001	idPer03	PSU0001	Rural
idStrt001	9	2	4.5	idHH00001	idPer04	PSU0001	Rural
idStrt001	9	2	4.5	idHH00001	idPer05	PSU0001	Rural
idStrt001	9	2	4.5	idHH00002	idPer01	PSU0001	Rural
idStrt001	9	2	4.5	idHH00002	idPer02	PSU0001	Rural
idStrt001	9	2	4.5	idHH00002	idPer03	PSU0001	Rural
idStrt001	9	2	4.5	idHH00002	idPer04	PSU0001	Rural
idStrt001	9	2	4.5	idHH00002	idPer05	PSU0001	Rural

Práctica en R

```
HHdb <- FrameII %>%
        group_by(PSU) %>%
        summarise(Ni = length(unique(HHID)))

Ni <- as.numeric(HHdb$Ni)
ni <- ceiling(Ni * 0.1)
sum(ni)
```

```
## [1] 693
```

Práctica en R

```
sam = S.SI(Ni[1], ni[1])
clusterII = FrameII[which(FrameII$PSU == sampleI[1]), ]
sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])
clusterHH <- left_join(sam.HH, clusterII, by = "HHID")
clusterHH$dki <- Ni[1]/ni[1]
clusterHH$dk <- clusterHH$dI * clusterHH$dki
data = clusterHH
```

Práctica en R

```
head(data, 10)
```

HHID	Stratum	Nlh	nlh	dl	PersonID	PSU	Zone
idHH00002	idStrt001	9	2	4.5	idPer01	PSU0001	Rural
idHH00002	idStrt001	9	2	4.5	idPer02	PSU0001	Rural
idHH00002	idStrt001	9	2	4.5	idPer03	PSU0001	Rural
idHH00002	idStrt001	9	2	4.5	idPer04	PSU0001	Rural
idHH00002	idStrt001	9	2	4.5	idPer05	PSU0001	Rural
idHH00005	idStrt001	9	2	4.5	idPer01	PSU0001	Rural
idHH00005	idStrt001	9	2	4.5	idPer02	PSU0001	Rural
idHH00005	idStrt001	9	2	4.5	idPer03	PSU0001	Rural
idHH00005	idStrt001	9	2	4.5	idPer04	PSU0001	Rural
idHH00005	idStrt001	9	2	4.5	idPer05	PSU0001	Rural

Práctica en R

```
for (i in 2:length(Ni)) {
      sam = S.SI(Ni[i], ni[i])
      clusterII = FrameII[which(FrameII$PSU == sampleI[i]), ]
      sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])
      clusterHH <- left_join(sam.HH, clusterII, by = "HHID")
      clusterHH$dki <- Ni[i]/ni[i]
      clusterHH$dk <- clusterHH$dI * clusterHH$dki
      data1 = clusterHH
      data = rbind(data, data1)
}
```

Práctica en R

```r
dim(data)
```

```
## [1] 2601   17
```

```r
sum(data$dk)
```

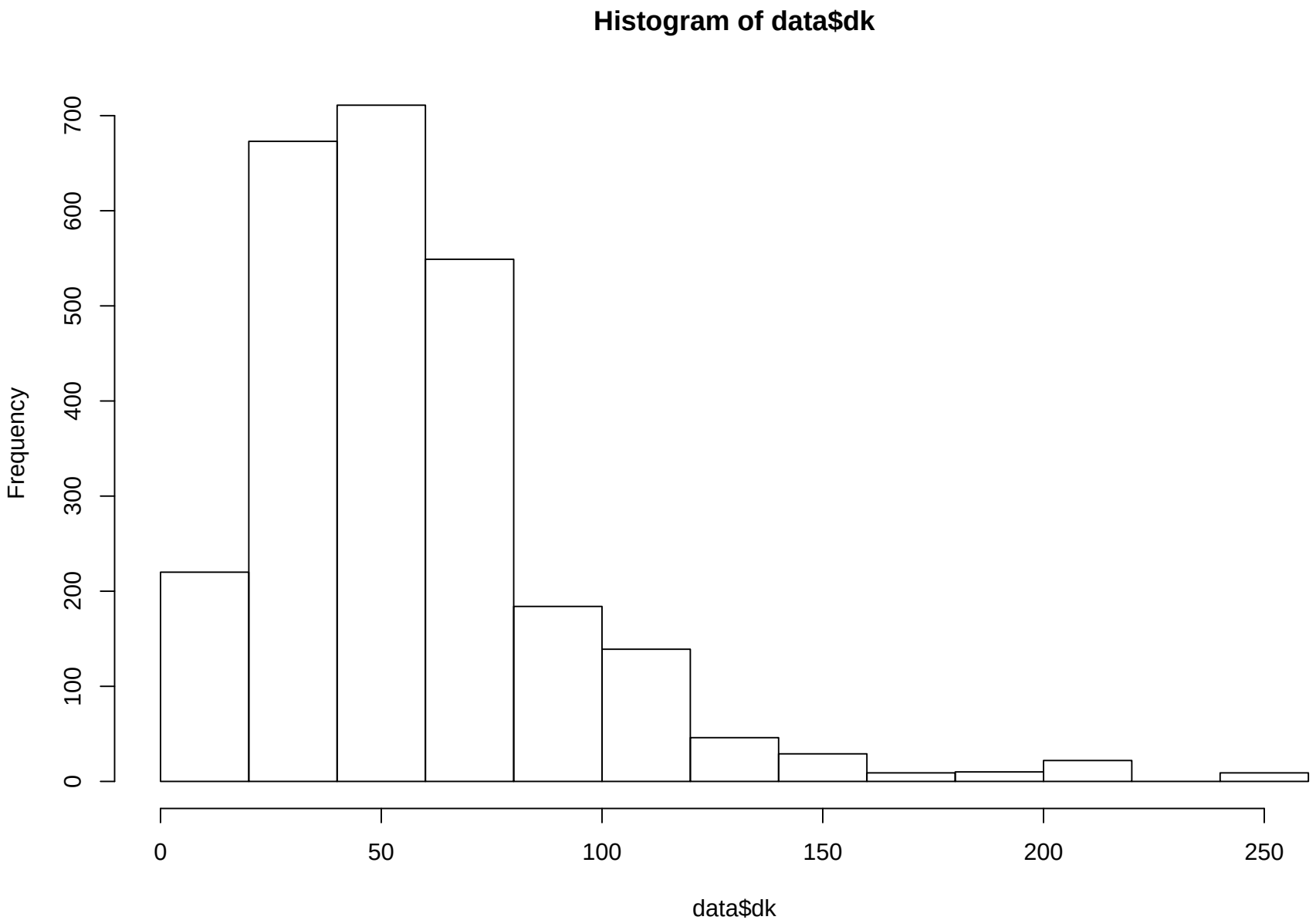
```
## [1] 152495
```

```r
attach(data)

estima <- data.frame(Income, Expenditure)
area <- as.factor(PSU)
stratum <- as.factor(Stratum)
```

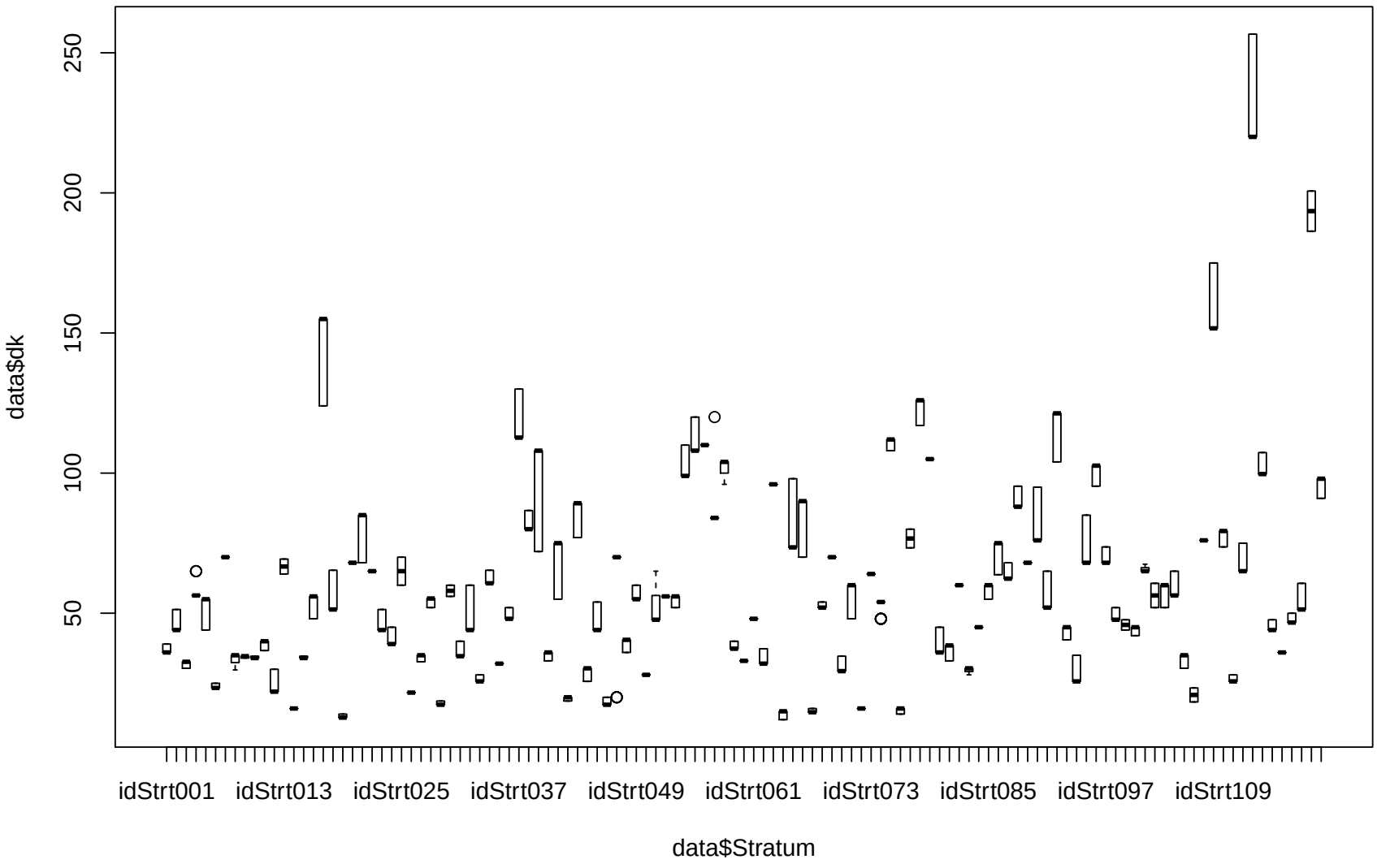
Práctica en R

```
hist(data$dk)
```



Práctica en R

```
boxplot(data$dk ~ data$Stratum)
```



Práctica en R

```
E.UC(stratum, area, dk, estima)
```

	N	Income	Expenditure
Estimation	152494.7	88884741.1	55821115.3
Standard Error	4005.4	5142678.5	2895120.5
CVE	2.6	5.8	5.2
DEFF	Inf	14.3	15.6

Muestreo autoponderado en dos etapas estratificado

Muestreo autoponderado en dos etapas estratificado

En muchas encuestas de dos etapas es común asumir que en la primera etapa de muestreo se selecciona una muestra S_I de unidades primarias de muestreo en cada estrato h cuyas probabilidades de inclusión son proporcionales al tamaño de las mismas (número de hogares o personas).

$$\pi_{Ii} = \frac{N_i}{N} n_I \qquad i \in U_I \tag{4}$$

Muestreo autoponderado en dos etapas estratificado

Más adelante, en la segunda etapa de muestreo, se seleccionan muestras s_i $i \in S_I$ de hogares de tamaño constante $n_i = n_0$ para cada unidad primaria incluida en la muestra.

Por lo tanto, la probabilidad de inclusión de las unidades secundarias será

$$\pi_{k|i} = \frac{n_0}{N_i} \qquad i \in S_I \tag{5}$$

Muestreo autoponderado en dos etapas estratificado

La probabilidad de inclusión general del k-ésimo elemento en el estrato h es constante y está dada por

$$\pi_k = \pi_{Ii}\pi_{k|i} = n_{Ih}\frac{N_i}{N_h}\frac{n_0}{N_i} = n_{Ih}\frac{n_0}{N_h} = \frac{n_h}{N_h} = c_h \qquad k \in U_i \qquad (6)$$

Muestreo autoponderado en dos etapas estratificado

El estimador de Horvitz-Thompson toma la siguiente forma

$$\hat{t}_{y,\pi} = \sum_h \sum_i \sum_k \frac{y_k}{\pi_k} = \sum_h \frac{N_h}{n_h} \sum_i \sum_k y_k \tag{7}$$

Muestreo estratificado en dos etapas autoponderado

1. Nótese la facilidad de cálculo del estimador.
2. Esta clase de diseños auto-ponderados se utilizan cuando se desea controlar el trabajo de campo.
3. El número de entrevistas en cada unidad primaria incluida en la muestra será constante.

Muestreo estratificado en dos etapas autoponderado

Para utilizar los prinicpios de estimación del último conglomerado en este diseño particular se definen las siguientes cantidades:

1. $d_{I_i} = \frac{N_{Ih}}{n_{Ih} N_i}$, que es el factor de expansión de la i-ésima UPM en el estrato h.
2. $d_{k|i} = \frac{N_i}{n_0}$, que es el factor de expansión del k-ésimo hogar para la i-ésima UPM.
3. $d_k = d_{I_i} \times d_{k|i} = \frac{N_{Ih}}{n_{Ih} N_i} \times \frac{N_i}{n_0} = \frac{N_h}{n_h}$, que es el factor de expansión final del k-ésimo elemento para toda la población U.

Práctica en R

```
data('BigCity')
FrameI <- BigCity %>% group_by(PSU) %>%
summarise(Stratum = unique(Stratum),
          Households = length(unique(HHID)),
          Income = sum(Income),
          Expenditure = sum(Expenditure))

attach(FrameI)
```

Práctica en R

```
head(FrameI, 10)
```

PSU	Stratum	Households	Income	Expenditure
PSU0001	idStrt001	26	70912	44232
PSU0002	idStrt001	32	68887	38382
PSU0003	idStrt001	24	37213	19495
PSU0004	idStrt001	22	36926	24031
PSU0005	idStrt001	28	57494	31142
PSU0006	idStrt001	30	75272	43473
PSU0007	idStrt001	24	33028	21833
PSU0008	idStrt001	36	64293	47660
PSU0009	idStrt001	26	33156	23292
PSU0010	idStrt002	22	65254	37115

Práctica en R

```
sizes = FrameI %>% group_by(Stratum) %>%
        summarise(NIh = n(), nIh = 2)

NIh <- sizes$NIh
nIh <- sizes$nIh
```

Práctica en R

```
head(sizes, 10)
```

Stratum	Nlh	nlh
idStrt001	9	2
idStrt002	11	2
idStrt003	7	2
idStrt004	13	2
idStrt005	11	2
idStrt006	5	2
idStrt007	14	2
idStrt008	7	2
idStrt009	8	2
idStrt010	8	2

Práctica en R

```
resI <- S.STpiPS(Stratum, Households, nIh)
head(resI, 10)
```

4	0.18
9	0.21
15	0.23
17	0.16
25	0.30
27	0.28
29	0.15
30	0.16
41	0.21
42	0.18

Práctica en R

```
samI <- resI[, 1]
piI <- resI[, 2]
UI <- levels(as.factor(FrameI$PSU))
sampleI <- data.frame(PSU = UI[samI], dI = 1/piI)

FrameII <- left_join(sampleI,
           BigCity[which(BigCity$PSU %in% sampleI[,1]), ])

attach(FrameII)
```

Práctica en R

```
head(FrameII, 10)
```

PSU	dI	HHID	PersonID	Stratum	Zone	Sex	A
PSU0004	5.6	idHH00042	idPer01	idStrt001	Rural	Male	
PSU0004	5.6	idHH00042	idPer02	idStrt001	Rural	Female	
PSU0004	5.6	idHH00042	idPer03	idStrt001	Rural	Male	
PSU0004	5.6	idHH00042	idPer04	idStrt001	Rural	Female	
PSU0004	5.6	idHH00042	idPer05	idStrt001	Rural	Male	
PSU0004	5.6	idHH00042	idPer06	idStrt001	Rural	Female	
PSU0004	5.6	idHH00043	idPer01	idStrt001	Rural	Male	
PSU0004	5.6	idHH00043	idPer02	idStrt001	Rural	Female	
PSU0004	5.6	idHH00043	idPer03	idStrt001	Rural	Female	
PSU0004	5.6	idHH00044	idPer01	idStrt001	Rural	Male	

Práctica en R

```
HHdb <- FrameII %>%
        group_by(PSU) %>%
        summarise(Ni = length(unique(HHID)),
                  ni = 4)
Ni <- as.numeric(HHdb$Ni)
ni <- 4
```

Práctica en R

```
head(HHdb, 10)
```

PSU	Ni	ni
PSU0004	22	4
PSU0009	26	4
PSU0015	38	4
PSU0017	26	4
PSU0025	28	4
PSU0027	26	4
PSU0029	26	4
PSU0030	28	4
PSU0041	30	4
PSU0042	26	4

Práctica en R

```
sam = S.SI(Ni[1], ni)
clusterII = FrameII[which(FrameII$PSU == sampleI$PSU[1]), ]
sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])
clusterHH <- left_join(sam.HH, clusterII, by = "HHID")
clusterHH$dki <- Ni[1]/ni
clusterHH$dk <- clusterHH$dI * clusterHH$dki
data = clusterHH
```

Práctica en R

```r
head(data, 10)
```

HHID	PSU	dI	PersonID	Stratum	Zone	Sex	A
idHH00044	PSU0004	5.6	idPer01	idStrt001	Rural	Male	
idHH00044	PSU0004	5.6	idPer02	idStrt001	Rural	Female	
idHH00044	PSU0004	5.6	idPer03	idStrt001	Rural	Female	
idHH00044	PSU0004	5.6	idPer04	idStrt001	Rural	Male	
idHH00044	PSU0004	5.6	idPer05	idStrt001	Rural	Male	
idHH00044	PSU0004	5.6	idPer06	idStrt001	Rural	Male	
idHH00051	PSU0004	5.6	idPer01	idStrt001	Rural	Female	
idHH00051	PSU0004	5.6	idPer02	idStrt001	Rural	Male	
idHH20689	PSU0004	5.6	idPer01	idStrt001	Rural	Male	
idHH20689	PSU0004	5.6	idPer02	idStrt001	Rural	Female	

Práctica en R

```
for (i in 2:length(Ni)) {
      sam = S.SI(Ni[i], ni)
      clusterII = FrameII[which(FrameII$PSU == sampleI$PSU[i]), ]
      sam.HH <- data.frame(HHID = unique(clusterII$HHID)[sam])
      clusterHH <- left_join(sam.HH, clusterII, by = "HHID")
      clusterHH$dki <- Ni[i]/ni
      clusterHH$dk <- clusterHH$dI * clusterHH$dki
      data1 = clusterHH
      data = rbind(data, data1)
}
```

Práctica en R

```
sum(data$dk)
```

```
## [1] 148654
```

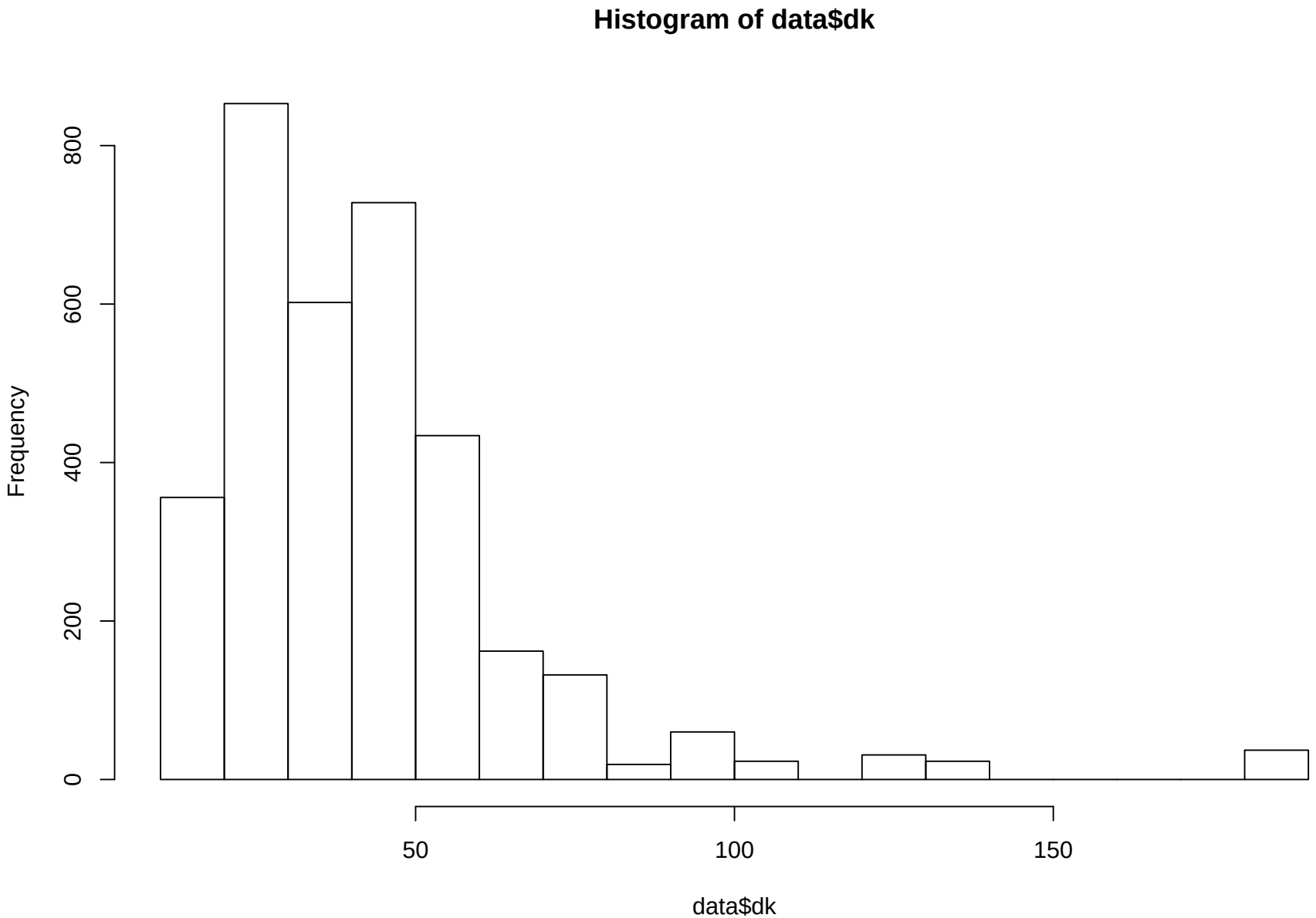
```
dim(data)
```

```
## [1] 3460   15
```

```
attach(data)
estima <- data.frame(Income, Expenditure)
area <- as.factor(PSU)
stratum <- as.factor(Stratum)
```

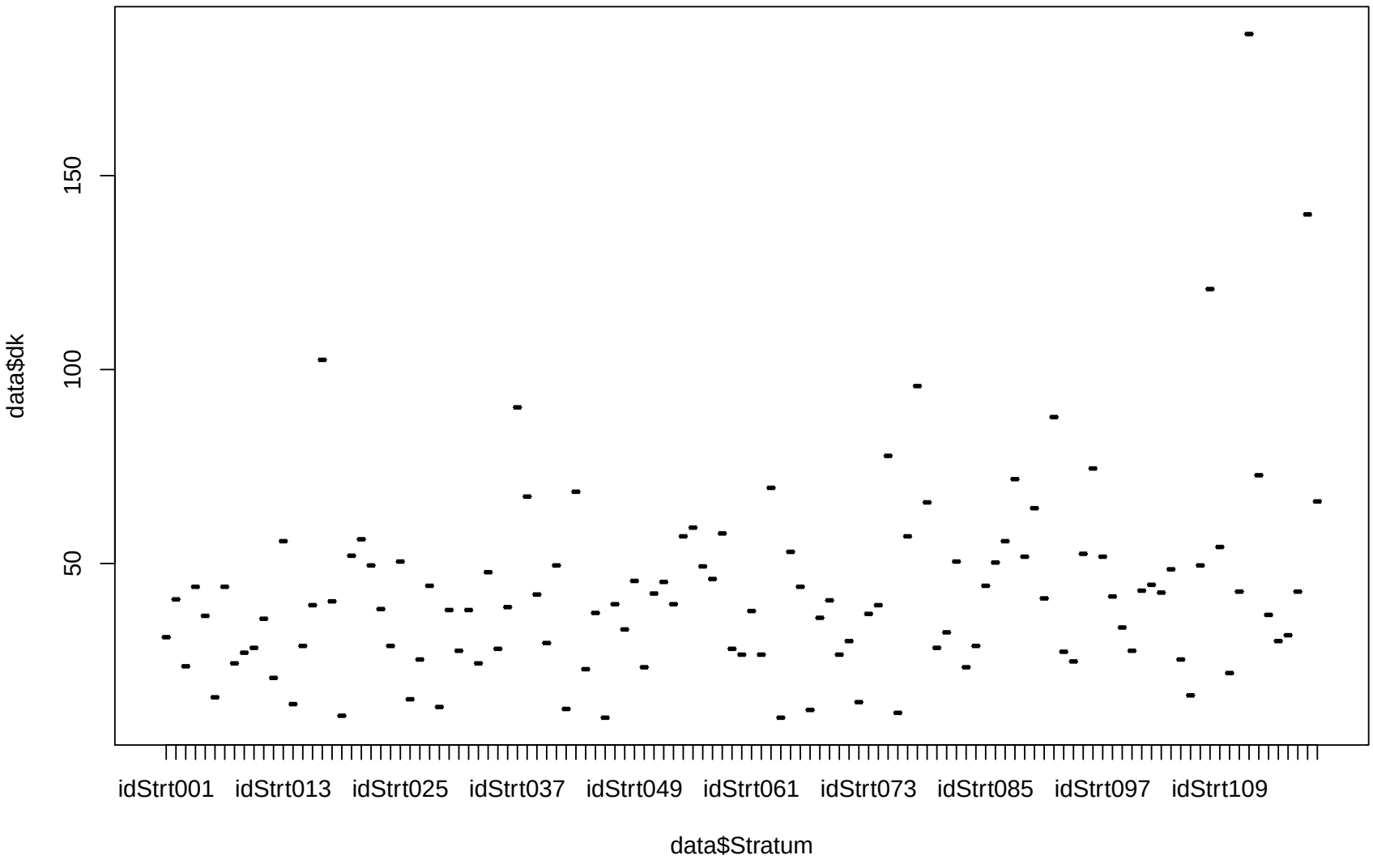
Práctica en R

```
hist(data$dk)
```



Práctica en R

```
boxplot(data$dk ~ data$Stratum)
```



Práctica en R

```
E.UC(stratum, area, dk, estima)
```

	N	Income	Expenditure
Estimation	148654.5	86097076.6	55409532.1
Standard Error	3353.0	3951640.0	2350503.9
CVE	2.3	4.6	4.2
DEFF	Inf	9.0	9.7

¡Gracias!

Andrés Gutiérrez

Experto Regional en Estadísticas Sociales

Division de Estadísticas

Email: andres.GUTIERREZ@cepal.org