Capítulo 3 - Ejercicio 2

Suponga que al tomarse un **MAE** de tamaño n con asignación proporcional de una variable estadística y se ha estimado la varianza poblacional σ^2_{N-1} de esta variable mediante la varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2\delta_{hi} \tag{1}$$

con

$$\bar{Y} = \frac{1}{n}\sum_{h=1}^{H}\sum_{i=1}^{N_h}y_{hi}\delta_{hi} \tag{2}$$

es decir, como si se tratase a la data como proveniente de un **MASs**.

a) Muestre que S^2 es un estimador sesgado de σ^2_{N-1} ¿sobreestima o subestima S^2 a σ^2_{N-1} ?

Solución. Tenemos definida la variable aleatoria siguiente:

$$\delta_{hi}: \text{Variable indicadora , con } E[\delta_{hi}] = \frac{n}{N} \tag{3}$$

Además nos indican utilizar las fórmulas se usan como un MASs y el MAE (asignación proporcional). Entonces:

$$W_h = \frac{N_h}{N} = \frac{n_h}{n} \quad \Rightarrow \quad \frac{n}{N} = \frac{n_h}{N_h} \tag{4}$$

Además en el MAE la variable indicadora sería f_{hi}, donde

$$E(f_{hi}) = \frac{n_h}{N_h} \tag{5}$$

Calculemos $E[S^2]$, siendo S^2 definida en (1):

$$E[S^2] = \frac{1}{n-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2E[\delta_{hi}] \overset{\text{por (3)}}{=} \frac{1}{n-1}\frac{n}{N}\sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2 \tag{6}$$

Desarrollando el binomio cuadrado de la expresión (6):

$$\begin{aligned}
\sum_{i=1}^{N_h}(y_{hi}-\bar{Y})^2 &= \sum_{i=1}^{N_h}(y_{hi}-\mu_h+\mu_h-\bar{Y})^2 \\
&= \underbrace{\sum_{i=1}^{N_h}(y_{hi}-\mu_h)^2}_{N_h\sigma^2_{h(N)}} + \sum_{i=1}^{N_h}(\mu_h-\bar{Y})^2 + 2(\mu_h-\bar{Y})\underbrace{\sum_{i=1}^{N_h}(y_{hi}-\mu_h)}_{0} \\
&= N_h\sigma^2_{h(N)} + N_h(\mu_h-\bar{Y})^2
\end{aligned} \tag{7}$$

Reemplazando (4) y (7) en (6):

$$\begin{aligned}
E[S^2] &= \frac{1}{n-1}\frac{n}{N}\sum_{h=1}^{H}\left(N_h\sigma^2_{h(N)} + N_h(\mu_h-\bar{Y})^2\right) \\
&= \frac{n}{n-1}\left(\underbrace{\sum_{h=1}^{H}W_h\sigma^2_{h(N)}}_{\sigma^2_N} + \sum_{h=1}^{H}W_h(\mu_h-\bar{Y})^2\right) \\
&= \frac{n}{n-1}\left(\sigma^2_N + \sum_{h=1}^{H}W_h(\mu_h-\bar{Y})^2\right)
\end{aligned} \tag{8}$$

Sabemos que

$$\sigma_N^2 = \frac{N-1}{N}\sigma_{N-1}^2 \tag{9}$$

Por tanto, (8) tomaría la siguiente forma:

$$\begin{aligned} E[S^2] &= \frac{n}{n-1}\left(\frac{N-1}{N}\sigma_{N-1}^2 + \sum_{h=1}^{H} W_h(\mu_h - \bar{Y})^2\right) \\ &= \frac{n}{n-1}\frac{N-1}{N}\sigma_{N-1}^2 + \underbrace{\frac{n}{n-1}\sum_{h=1}^{H} W_h(\mu_h - \bar{Y})^2}_{\text{Si Consideramos despreciable este término}} \\ &= \frac{n}{n-1}\frac{N-1}{N}\sigma_{N-1}^2 \end{aligned} \tag{10}$$

□

b) Muestre sin embargo que S^2 es una estimador asintóticamente insesgado de σ_{N-1}^2.

Solución.

$$E[S^2] = \frac{1-\frac{1}{N}}{1-\frac{1}{n}}\sigma_{N-1}^2 \tag{11}$$

Pero si $n \to \infty$ entonces $N \to \infty$:

$$\lim_{n\to\infty}\left(\lim_{N\to\infty}\frac{1-\frac{1}{N}}{1-\frac{1}{n}}\right) = \lim_{n\to\infty}\frac{1}{1-\frac{1}{n}} = 1 \quad \Rightarrow \quad E[S^2] = \sigma_{N-1}^2 \text{ si } n \to \infty \tag{12}$$

□

c) ¿Qué repercusión tiene el resultado en b) con respecto al efecto de diseño del **MAE** para estimar la media de la población?

Solución. La evaluación del efecto del diseño permite medir el grado de distorsión que sufren las varianzas debido al diseño muestral empleado y, por lo tanto, proporciona una valoración directa de la alteración que sufren los intervalos de confianza estimados cuando el diseño muestral se aparta del caso aleatorio simple.
Conclusiones: Para el muestreo estratificado el efecto del diseño en el análisis de los datos obtenidos por muestreo esta incluyendo ponderaciones en los análisis estadísticos.
La cual subestima el verdadero valor de σ_{N-1}^2

□

Capítulo 3 - Ejercicio 4

Considere una población de $N = 20$ domicilios, donde son conocidas las variable y = renta familiar mensual en miles de soles y la variable estrato socioeconómico (A = alto, M = medio y B = bajo) al cual pertenecen. Los valores de estas variables se resumen en la tabla:

Unidad	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
renta	13	17	6	5	9	12	19	6	14	12	8	5	11	20	6	18	10	9	12	8
estrato	M	A	B	B	B	M	A	B	M	M	B	B	M	A	B	A	M	B	B	B

Con la finalidad de estimar la renta familiar media, se tienen las alternativas de efectuar un **MAE** de tamaño 10 con afijación proporcional, un *MASs* con 10 observaciones o un **MASc** con 10 observaciones.

a) Determine las varianzas de estos diseños e indique cuál es más eficiente y porqué.

Parámetros poblacionales

Estrato	A	M	B	Total
Media	18.5	12.0	7.4	11
Varianza	1.667	2.0	4.933	22.105
Tamaño	4	6	10	20
Cantidad Muestra	2	3	5	10

Determinación de las varianzas según el diseño

Muestreo Aleatorio Estratificado

Muestra MAE	
Var(Ybar) =	0.17

Muestreo Aleatorio Simple sin reposición

Muestra MASs	
Var(Ybar) =	1.1053

Muestreo Aleatorio Simple con reposición

Muestra MASc	
Var(Ybar) =	2.1

Efecto (MAE)=0.1538
Efecto (MASc)=1.90

Dado que el Efecto MAE es menor que el Efecto MASc entonces el diseño de MAE es más eficiente y esto básicamente porque tienen una varianza menor

b) Usando los números aleatorios

$$0.91, 0.02, 0.7, 0.35, 0.1, 0.96, 0.51, 0.46, 0.23, 0.87$$

tome las muestras requeridas para estos diseños y estime la renta familiar media bajo cada uno.

MAE:

Determinación de la muestra seleccionada bajo MAE

Para A

Distribución empirica acumulada

1	0.25	
2	0.50	
3	0.75	
4	1.00	X

1	0.333	X
2	0.667	
3	1.000	

Para M

Distribución empirica acumulada

1	0.167	
2	0.333	
3	0.500	
4	0.667	
5	0.833	X
6	1.000	

1	0.2	
2	0.4	X
3	0.6	
4	0.8	
6	1.0	

1	0.25	X
3	0.50	
4	0.75	
6	1.00	

Para B

Distribución empirica acumulada

1	0.1	
2	0.2	
3	0.3	
4	0.4	
5	0.5	
6	0.6	
7	0.7	
8	0.8	
9	0.9	
10	1.0	X

1	0.111	
2	0.222	
3	0.333	
4	0.444	
5	0.556	X
6	0.667	
7	0.778	
8	0.889	
9	1.000	

1	0.125	
2	0.250	
3	0.375	
4	0.500	X
6	0.625	
7	0.750	
8	0.875	
9	1.000	

1	0.143	
2	0.286	X
3	0.429	
6	0.571	
7	0.714	
8	0.857	
9	1.000	

1	0.167	
3	0.333	
6	0.500	
7	0.667	
8	0.833	
9	1.000	X

Elección de la muestra seleccionada bajo MAE

MAE	selección aleatoria por estrato:					Muestra de 10 unidades en total					Muestra seleccionada					Media muestral
A	0.91	0.02				4	1				18	17				17.50
M	0.7	0.35	0.1			5	2	1			11	12	13			12.00
B	0.96	0.51	0.46	0.23	0.87	10	5	4	2	9	8	8	6	5	12	7.80

Estimación de la renta familiar medio bajo MAE:

$$\bar{Y}(\text{MAE}) = 11.00$$

MASs:

Determinación de la muestra seleccionada bajo MASs

Para B

Distribución empírica acumulada

1	0.05		1	0.053	X	2	0.056		2	0.059		2	0.063	
2	0.10		2	0.105		3	0.111		3	0.118		3	0.125	X
3	0.15		3	0.158		4	0.167		4	0.176		4	0.188	
4	0.20		4	0.211		5	0.222		5	0.235		5	0.250	
5	0.25		5	0.263		6	0.278		6	0.294		6	0.313	
6	0.30		6	0.316		7	0.333		7	0.353	X	8	0.375	
7	0.35		7	0.368		8	0.389		8	0.412		9	0.438	
8	0.40		8	0.421		9	0.444		9	0.471		10	0.500	
9	0.45		9	0.474		10	0.500		10	0.529		11	0.563	
10	0.50		10	0.526		11	0.556		11	0.588		12	0.625	
11	0.55		11	0.579		12	0.611		12	0.647		13	0.688	
12	0.60		12	0.632		13	0.667		13	0.706		15	0.750	
13	0.65		13	0.684		14	0.722	X	15	0.765		16	0.813	
14	0.70		14	0.737		15	0.778		16	0.824		17	0.875	
15	0.75		15	0.789		16	0.833		17	0.882		18	0.938	
16	0.80		16	0.842		17	0.889		18	0.941		20	1.000	
17	0.85		17	0.895		18	0.944		20	1.000				
18	0.90		18	0.947		20	1.000							
19	0.95	X	20	1.000										
20	1.00													

2	0.067		2	0.071		2	0.077		2	0.083		2	0.091	
4	0.133		4	0.143		4	0.154		4	0.167		4	0.182	
5	0.200		5	0.214		5	0.231		5	0.250	X	6	0.273	
6	0.267		6	0.286		6	0.308		6	0.333		8	0.364	
8	0.333		8	0.357		8	0.385		8	0.417		10	0.455	
9	0.400		9	0.429		9	0.462	X	10	0.500		12	0.545	
10	0.467		10	0.500		10	0.538		12	0.583		13	0.636	
11	0.533		11	0.571	X	12	0.615		13	0.667		15	0.727	
12	0.600		12	0.643		13	0.692		15	0.750		16	0.818	
13	0.667		13	0.714		15	0.769		16	0.833		17	0.909	X
15	0.733		15	0.786		16	0.846		17	0.917		18	1.000	
16	0.800		16	0.857		17	0.923		18	1.000				
17	0.867		17	0.929		18	1.000							
18	0.933		18	1.000										
20	1.000	X												

Elección de la muestra seleccionada bajo MASs

Números aleatorios										Muestra de 10 unidades										Media muestral
0.91	0.02	0.7	0.35	0.1	0.96	0.51	0.46	0.23	0.87	19	1	14	7	3	20	11	9	5	17	11.90

Renta familar medio bajo el Muestreo Aleatorio Simple sin reposición:

$$\bar{Y}(\text{MAE}) = 11.90$$

MASc:

Elección de la muestra seleccionada bajo MASc

Números aleatorios										Muestra de 10 unidades										Media muestral
0.91	0.02	0.7	0.35	0.1	0.96	0.51	0.46	0.23	0.87	19	1	14	7	2	20	11	10	5	18	12.70

UNIDAD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Distribución empírica acumulada	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
RENTA	13	17	6	5	9	12	19	6	14	12	8	5	11	20	6	18	10	9	12	8

Renta familar medio bajo el Muestreo Aleatorio Simple con reposición

$$\bar{Y}(\text{MASc}) = 12.70$$

Capítulo 3 - Ejercicio 11

Considere la base de datos poblacional Province 91 del capítulo 2 y la variable Stratum que identifica si la municipalidad de la provincia en estudio es urbana o rural. Usando esta última como variable de estratificación y la variable número de personas desempleadas como variable de investigación tome un **MAE** de 8 municipalidades y responda a lo siguiente:

a) Halle los tamaños de muestra por estrato, usando tanto la asignación proporcional como la de Neyman.

Solución.

Listing 1: Cargando los datos.

```
setwd(wd)
library(data.table)

p91<-fread("province91.csv")

#Tamanos de poblacion
N<-nrow(p91)
Nh<-p91[,.N,stratum]$N

#Tamano de muestra (dato de problema)
n<-8
```

Los tamaños de muestra para cada estrato h pueden expresarse como fracciones de la muestra total:

$$n_h = a_h n$$

Para la asignación proporcional, se toma $a_h = \frac{N_h}{N}$.
Calculamos los n_h resultantes:

Listing 2: Calculando nh (Asignación proporcional).

```
nh_proporcional<-round(n*Nh/N)
cat("n estrato 1: ",nh_proporcional[1])
cat("n estrato 2: ",nh_proporcional[2])
```

Resultado: $n_{\text{estrato 1}}$ =2, $n_{\text{estrato 2}}$ =6.

En el caso de la asignación de Neyman, tenemos $a_h = \frac{N_h S_h}{\sum_{j=1}^{H} N_j S_j}$.
Asumimos que algún estudio previo nos ha provisto de valores certeros para los S_h requeridos en el cálculo y procedemos a obtener los n_h correspondientes:

Listing 3: Calculando nh (Asignación de Neyman).

```
Sh<-p91[,sd(ue91)**2,stratum]$V1
nh_neyman<-round(n*Nh*Sh/sum(Nh*Sh))

cat("n estrato 1: ",nh_neyman[1])
cat("n estrato 2: ",nh_neyman[2])
nh_neyman<-c(6,2)
```

Resultado: $n_{\text{estrato 1}}$ =7, $n_{\text{estrato 2}}$ =1.

No es útil tomar el resultado de la fórmula directamente, ya que con $n_2 = 1$ no sería posible obtener un estimado de la varianza en ese estrato o para la población total. Entonces para la asignación de Neyman, trabajaremos con $n_1 = 6, n_2 = 2$. □

b) Halle para los dos esquemas anteriores los efectos de diseño de la estimación del total de personas desempleadas en la provincia.

Solución. El efecto de diseño está dado por:

$$D_{eff} = \frac{V_{MAE}(\bar{Y})}{V_{MASs}(\bar{Y})}$$

Procedemos a realizar el cálculo:

Listing 4: Efecto diseño.

```
S<-sd(p91$ue91)**2

VMASs<-(1-n/N)*(S)/n
VMAE_proporcional<-sum(((Nh/N)**2)*(1-nh_proporcional/Nh)*(Sh/nh_proporcional))
VMAE_neyman<-sum(((Nh/N)**2)*(1-nh_neyman/Nh)*(Sh/nh_neyman))

Deff_proporcional<-VMAE_proporcional/VMASs
Deff_neyman<-VMAE_neyman/VMASs

cat("Deff proporcional: ",Deff_proporcional)
cat("Deff Neyman: ",Deff_neyman)
```

Resultado: Deff proporcional: 0.7369311, Deff Neyman: 0.6335367

□

c) Tome las muestras requeridas bajos los dos esquemas anteriores y estime tanto el total de personas desempleadas en la provincia como los efectos de diseño de estas estimaciones. Para esto último puede utilizar, como es usual en muchos softwares estadísticos, la metodología planteada en el ejercicio 2.

Solución.

Listing 5: Generamos las muestras

```
sel_neyman<-mapply(sample,Nh,nh_neyman)
p91_neyman<-rbind(p91[stratum==1][sel_neyman[[1]]],p91[stratum==2][sel_neyman
    [[2]]])

sel_proporcional<-mapply(sample,Nh,nh_proporcional)
p91_proporcional<-rbind(p91[stratum==1][sel_proporcional[[1]]],p91[stratum==2][
    sel_proporcional[[2]]])
```

El total de personas desempleadas en la provincia es la suma de personas desempleadas en cada municipalidad. Podemos usar el número promedio de desempleados por estrato ponderado por el número de municipalidades en cada estrato:

$$D = \sum_{1}^{H} D_i = \sum_{1}^{H} \frac{N_i}{N_i} D_i = \sum_{1}^{H} N_i \bar{D}_i$$

Calculamos los estimadores correspondientes en las muestras generadas:

Listing 6: Estimadores.

```
d_neyman<-ceiling(sum(((p91_neyman[,sum(ue91),stratum]$V1)/nh_neyman)*Nh))
d_proporcional<-ceiling(sum(((p91_proporcional[,sum(ue91),stratum]$V1)/
    nh_proporcional)*Nh))

cat("Total desempleados proporcional: ",d_proporcional)
cat("Total desempleados Neyman: ",d_neyman)
```

Resultado: Total desempleados proporcional: 16933, Total desempleados Neyman: 12206

Para calcular el efecto de diseño, estimamos las varianzas requeridas a partir de nuestra muestra:

Listing 7: Efecto diseño.

```
s_neyman<-sd(p91_neyman$ue91)**2
sh_neyman<-p91_neyman[,sd(ue91)**2,stratum]$V1
s_proporcional<-sd(p91_proporcional$ue91)**2
sh_proporcional<-p91_proporcional[,sd(ue91)**2,stratum]$V1

vMASs_neyman<-(1-n/N)*(s_neyman)/n
vMASs_proporcional<-(1-n/N)*(s_proporcional)/n
vMAE_neyman<-sum(((Nh/N)**2)*(1-nh_neyman/Nh)*(sh_neyman/nh_neyman))
vMAE_proporcional<-sum(((Nh/N)**2)*(1-nh_proporcional/Nh)*(sh_proporcional/
    nh_proporcional))

Deff_proporcional<-VMAE_proporcional/vMASs_proporcional
Deff_neyman<-VMAE_neyman/vMASs_neyman
cat("Deff proporcional: ",Deff_proporcional)
cat("Deff Neyman: ",Deff_neyman)
```

Deff proporcional: 1.572082, Deff Neyman: 0.2053843

□

Capítulo 3 - Ejercicio 12

Considere la base de datos apipop vista en clase y supongamos que estemos interesados en estimar el número total de alumnos matriculados en esta población mediante un **MAE**, donde el criterio de estratificación es nuevamente el tipo de colegio. Se desea estimar este número con un error de estimación no mayor a los 150,000 alumnos y un nivel de confianza del 95 %. Para lograr ello,

a) Tome un **MAE** piloto de sólo 30 escuelas.
Se tomará una muestra de la base de datos apipop de forma aleatoria y estratificada por escuelas

Solución.

Listing 8: MAE Piloto.

```
library(survey)
data(api)
attach(apipop)
# Tamano de cada estrato
table(stype)

# Generando la MAE piloto 21 de E , 4 de H, 5 de M
index=c(sample(which(stype=="E"),21),sample(which(stype=="H"),4),sample(which(
    stype=="M"),5))
muestra=apipop[index ,]

# Calculando las desviaciones de la muestra en cada estrato
v=aggregate(x = muestra["enroll"], by = muestra["stype"], FUN = var)
Sh=sqrt(v[2])
```

Los resultados obtenidos son los siguientes:

Estrato 1 (Escuelas elementales)	:	$N_1 = 4421$	
Estrato 2 (Escuelas superiores)	:	$N_2 = 755$	
Estrato 3 (Escuelas medias)	:	$N_3 = 1018$	
Tamaño total	:	$N = 6194$	
Muestra en el estrato 1 (Escuelas elementales)	:	$n_1 = 21,$	$S_1 = 189.8864$
Muestra en el estrato 2 (Escuelas superiores)	:	$n_2 = 4,$	$S_2 = 397.3219$
Muestra en el estrato 3 (Escuelas medias)	:	$n_3 = 5,$	$S_3 = 724.8191$
Tamaño total de la muestra	:	$n = 30$	

□

b) Halle los tamaños de muestra requeridos mediante una asignación óptima. Considere costos de muestreo iguales y utilice las estimaciones necesarias de la muestra piloto tomada en a).

Solución. Si se consideran costos de muestreo iguales entonces mediante un caso particular de asignación óptima, denominada *asignación de Neyman*, las fracciones a_h de n para la muestra en cada estrato serían:

$$a_h = \frac{N_h S_h}{\sum_{j=1}^{3} N_j S_j}$$

donde los S_h son tomadas de la muestra piloto.
Por lo que el tamaño de muestra óptimo para estimar el número total de alumnos matriculados vendría

dado por

$$n = \frac{\sum_{h=1}^{3} \frac{N_h^2}{a_h} S_h^2}{\left(\frac{e}{z_{1-\frac{\alpha}{2}}}\right)^2 + \sum_{h=1}^{3} N_h S_h^2} = \frac{\left(\sum_{h=1}^{3} N_h S_h\right)^2}{\left(\frac{e}{z_{1-\frac{\alpha}{2}}}\right)^2 + \sum_{h=1}^{3} N_h S_h^2}$$

siendo $e = 150000$ y $\alpha = 0.05$.

Listing 9: Tamaños de muestra requeridos.

```
Nh=c(table(stype)[1],table(stype)[2],table(stype)[3])
ah=Nh*Sh/sum(Nh*Sh)

# Calculando n
n=sum(Nh^2*Sh^2/ah)/((150000/qnorm(1-0.05/2))^2+sum(Nh*Sh^2))

# Calculando nh por cada estrato
nh=n*ah
# Redondeando valores
nh=c(ceiling(nh[1,]),ceiling(nh[2,]),ceiling(nh[3,]))
```

□

Los resultados que obtenemos son los siguientes:

$$n_1 = 237 \quad n_2 = 85 \quad n_3 = 208$$

c) Realice el **MAE** y reporte el intervalo de confianza al 95 % para el número de matriculados en esta población.

Listing 10: Estimación del Total de alumnos.

```
index = c(sample(which(stype=="E"),nh[1]),sample(which(stype=="H") ,nh[2]),
sample(which(stype=="M") ,nh[3]))
sample1 = apipop[index ,]
summary(sample1$stype)
aux = data.frame(pw = c(rep(4421/nh[1],nh[1]), rep(755/nh[2],nh[2]), rep(1018/nh
    [3],nh[3]))
, fpc = c(rep(4421,nh[1]),rep(755,nh[2]), rep(1018,nh[3])))
sampleMAE = cbind(num = 1:sum(nh),sample1 ,aux)
disenhoMAE = svydesign(id=~1,strata=~stype ,fpc = ~fpc, data = sampleMAE)
svytotal(~enroll ,disenhoMAE,na.rm=T)
```

Resultado:

$$\text{Total} = 3958584 \quad , \quad SE = 85615$$

El intervalo de confianza al %95, viene dado por:

$$IC = [3958584 - 1.96 \times 85615, 3958584 + 1.96 \times 85615] = [3790779, 4126389]$$

d) Suponga que en la muestra anterior era también de interés estimar la proporción de escuelas en esta población que recibieron un premio (awards). Estime tal proporción y reporte su máximo error de estimación.

Solución.

Listing 11: Estimación de la proporción.

```
1
2  # Agrupando datos respecto a variable awards
3  library(sqldf)
4  da=sqldf('select stype,awards from sample1 where "awards"=="Yes"')
5  summary(da$stype)
6
7  # Calculando datos especificos para el calculo
8  n1h=c(557,288,143)
9  s4=sum(n1h)
10 ph=c(557/s4,143/s4,288/s4)
11 nh=c(727,251,755)
12 n=sum(nh)
13
14 # Estimando p
15 p=sum(Nh*ph)/N
16
17 # SE(p)
18 s5=0
19 for (i in 1:3){
20 s5=s5+((Nh[i]/N)**2)*(1-nh[i]/Nh[i])*ph[i]*(1-ph[i])/(nh[i]-1)
21 }
22 sqrt(s5)
```

Resultado:

$$\bar{p} = 0.263377 \quad , \quad SE(\bar{p}) = 0.01157953$$

□

Capítulo 4 - Ejercicio 1

Muestre que el estimador $\hat{\tau}_\psi$ definido como

$$\hat{\tau}_\psi = \frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i} \quad , \quad \text{donde } \delta_i \sim \text{Binomial}(n.\psi_i): \begin{array}{l}\text{Número de veces que la unidad i}\\ \text{es seleccionada en la muestra.}\end{array} \tag{13}$$

es un estimador insesgado del total poblacional. Pruebe también que la varianza de este estimador viene dada por:

$$V(\hat{\tau}_\psi) = \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{\tau_i}{\psi_i}-\tau\right)^2 + \frac{1}{n}\sum_{i=1}^{N}\frac{V(\hat{\tau_{ij}})}{\psi_i} \tag{14}$$

y que

$$\hat{V}(\hat{\tau}_\psi) = \frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\left(\frac{\hat{\tau}_{ij}}{\psi_i}-\hat{\tau}_\psi\right)^2 \tag{15}$$

es una estimador insesgado de esta varianza.

Solución.

a) Probar que el estimador $\hat{\tau}_\psi$ es un estimador insesgado del total poblacional.

$$\begin{aligned}
E[\hat{\tau}_\psi] &= E\left[E[\hat{\tau}_\psi|\delta_i]\right] = E\left[E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i}\right]\right] = E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{E\left[\hat{\tau}_{ij}\right]}{\psi_i}\right]\\
&= E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\tau_i}{\psi_i}\right] = E\left[\frac{1}{n}\sum_{i=1}^{N}\delta_i\frac{\tau_i}{\psi_i}\right] = \frac{1}{n}\sum_{i=1}^{N}E[\delta_i]\frac{\tau_i}{\psi_i}\\
&= \frac{1}{n}\sum_{i=1}^{N}n\psi_i\frac{\tau_i}{\psi_i} = \sum_{i=1}^{N}\tau_i\\
E[\hat{\tau}_\psi] &= \tau_\psi
\end{aligned} \tag{16}$$

Por tanto $\hat{\tau}_\psi$ es un estimador insesgado.

b) Determinar $\hat{\tau}_\psi$.
Primero veamos lo siguiente

$$V[\hat{\tau}_\psi] = \underbrace{E[V[\hat{\tau}_\psi|\delta_i]]}_{\text{PARTE I}} + \underbrace{V[E[\hat{\tau}_\psi|\delta_i]]}_{\text{PARTE II}} \tag{17}$$

Desarrollando PARTE I:

$$\begin{aligned}
E[V[\hat{\tau}_\psi|\delta_i]] &= E\left[V\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i}\right]\right] = E\left[\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i^2}\right] = E\left[\frac{1}{n^2}\sum_{i=1}^{N}\delta_i\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i^2}\right]\\
&= \frac{1}{n^2}\sum_{i=1}^{N}E[\delta_i]\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i^2} = \frac{1}{n^2}\sum_{i=1}^{N}E[\delta_i]\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i^2} = \frac{1}{n^2}\sum_{i=1}^{N}n\psi_i\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i^2}\\
E[V[\hat{\tau}_\psi|\delta_i]] &= \frac{1}{n}\sum_{i=1}^{N}\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i}
\end{aligned} \tag{18}$$

Desarrollando PARTE II:

$$
\begin{aligned}
V[E[\hat{\tau}_\psi|\delta_i]] &= V\left[E\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i}\right]\right] = V\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{E\left[\hat{\tau}_{ij}\right]}{\psi_i}\right] \\
&= V\left[\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\tau_i}{\psi_i}\right] = V\left[\frac{1}{n}\sum_{i=1}^{N}\delta_i\frac{\tau_i}{\psi_i}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{N}V[\delta_i]\left(\frac{\tau_i}{\psi_i}\right)^2+\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}cov[\delta_i,\delta_j]\frac{\tau_i\tau_j}{\psi_i\psi_j} \\
&= \frac{1}{n^2}\sum_{i=1}^{N}n\psi_i(1-\psi_i)\left(\frac{\tau_i}{\psi_i}\right)^2+\frac{1}{n^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(-n\psi_i\psi_j)\frac{\tau_i\tau_j}{\psi_i\psi_j} \\
&= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{1}{\psi_i}-1\right)\tau_i^2-\frac{1}{n}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\tau_i\tau_j \\
&= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}-\left(\sum_{i=1}^{N}\tau_i^2+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\tau_i\tau_j\right)\right) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}-\left(\sum_{i=1}^{N}\tau_i\right)^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}-\tau_\psi^2\right) \quad : \text{ Considerado para el item c)} \qquad (19) \\
&= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}-2\tau_\psi\tau_\psi+\tau_\psi^2\right) = \frac{1}{n}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}-2\tau_\psi\left(\sum_{i=1}^{N}\tau_i\right)+\tau^2\left(\sum_{i=1}^{N}\psi_i\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\left(\frac{\tau_i}{\psi_i}\right)^2-2\tau_\psi\frac{\tau_i}{\psi_i}+\tau_\psi^2\right) \\
V[E[\hat{\tau}_\psi|\delta_i]] &= \frac{1}{n}\sum_{i=1}^{N}\psi_i\left(\frac{\tau_i}{\psi_i}-\tau_\psi\right)^2 \qquad (20)
\end{aligned}
$$

Reemplazando (18) y (20) en la ecuación (17), obtenemos la ecuación (14) para $V[\hat{\tau}_\psi]$.

c) Probar que $\hat{V}(\hat{\tau}_\psi)$ es un estimador insesgado de $V(\hat{\tau}_\psi)$, es decir que $E[\hat{V}(\hat{\tau}_\psi)] = V(\hat{\tau}_\psi)$.
Primero reformulemos $\hat{V}(\hat{\tau}_\psi)$:

$$
\begin{aligned}
\hat{V}(\hat{\tau}_\psi) &= \frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\left(\frac{\hat{\tau}_{ij}}{\psi_i}-\hat{\tau}_\psi\right)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\left(\left(\frac{\hat{\tau}_{ij}}{\psi_i}\right)^2-2\left(\frac{\hat{\tau}_{ij}}{\psi_i}\right)\hat{\tau}_\psi+\hat{\tau}_\psi^2\right) \\
&= \frac{1}{n(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}-2\hat{\tau}_\psi\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}}{\psi_i}}_{n\hat{\tau}_\psi}+\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\hat{\tau}_\psi^2\right) \\
&= \frac{1}{n(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}-2\hat{\tau}_\psi n\hat{\tau}_\psi+\hat{\tau}_\psi^2\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}}_{n}\right) \\
&= \frac{1}{n(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}-n\hat{\tau}_\psi^2\right) \qquad (21)
\end{aligned}
$$

Ahora determinemos su esperanza:

$$E\left[\hat{V}(\hat{\tau}_\psi)\right] = \frac{1}{n(n-1)}\left(\underbrace{E\left[E\left[\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}\right]\right]}_{\text{PARTE I}} - n \times \underbrace{E\left[\hat{\tau}_\psi^2\right]}_{\text{PARTE II}}\right) \tag{22}$$

Desarrollando PARTE I:

$$\begin{aligned}
E\left[E\left[\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{\hat{\tau}_{ij}^2}{\psi_i^2}\right]\right] &= E\left[\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{E\left[\hat{\tau}_{ij}^2\right]}{\psi_i^2}\right] = E\left[\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{V\left[\hat{\tau}_{ij}\right]+E\left[\hat{\tau}_{ij}\right]^2}{\psi_i^2}\right] \\
&= E\left[\sum_{i=1}^{N}\sum_{j=1}^{\delta_i}\frac{V\left[\hat{\tau}_{ij}\right]+\tau_i^2}{\psi_i^2}\right] = E\left[\sum_{i=1}^{N}\frac{\delta_i\left(V\left[\hat{\tau}_{ij}\right]+\tau_i^2\right)}{\psi_i^2}\right] \\
&= \sum_{i=1}^{N}\frac{E\left[\delta_i\right]\left(V\left[\hat{\tau}_{ij}\right]+\tau_i^2\right)}{\psi_i^2} = \sum_{i=1}^{N}\frac{n\psi_i\left(V\left[\hat{\tau}_{ij}\right]+\tau_i^2\right)}{\psi_i^2} \\
&= n\sum_{i=1}^{N}\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i} + n\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i}
\end{aligned} \tag{23}$$

Desarrollando PARTE II:

$$E\left[\hat{\tau}_\psi^2\right] = V\left[\hat{\tau}_\psi\right] + E\left[\hat{\tau}_\psi\right]^2 = Var\left[\hat{\tau}_\psi\right] + \tau_\psi^2 \tag{24}$$

Reemplazando (23) y (24) en (22), tenemos lo siguiente

$$\begin{aligned}
E\left[\hat{V}(\hat{\tau}_\psi)\right] &= \frac{1}{n(n-1)}\left(n\sum_{i=1}^{N}\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i} + n\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i} - n\left(V\left[\hat{\tau}_\psi\right]+\tau_\psi^2\right)\right) \\
&= \frac{1}{(n-1)}\left(\sum_{i=1}^{N}\frac{\tau_i^2}{\psi_i} - \tau_\psi^2 + \sum_{i=1}^{N}\frac{V\left[\hat{\tau}_{ij}\right]}{\psi_i} - V\left[\hat{\tau}_\psi\right]\right) \\
&\overset{\text{Por equivalencia de (19) y (20)}}{=} \frac{1}{(n-1)}\left(\underbrace{\sum_{i=1}^{N}\psi_i\left(\frac{\tau_i}{\psi_i}-\tau_\psi\right)^2 + \sum_{i=1}^{N}\frac{V(\hat{\tau_{ij}})}{\psi_i}}_{nV\left[\hat{\tau}_\psi\right]} - V\left[\hat{\tau}_\psi\right]\right) \\
&= \frac{1}{(n-1)}(n-1)V\left[\hat{\tau}_\psi\right] = V\left[\hat{\tau}_\psi\right]
\end{aligned} \tag{25}$$

□

Capítulo 4 - Ejercicio 6

Muestre que la correlación intraclase definida en (4.3) puede escribirse como

$$\rho = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(NM-1)(M-1)\sigma^2} \tag{26}$$

donde μ y σ^2 son respectivamente la media y varianza poblacionales de la variable de investigación y.

Solución. Asumiendo que los tamaños M_i de los conglomerados son todos iguales, es decir M entonces la correlación intraclase estaría definida de la siguiente manera:

$$\rho = 1-\left(\frac{M}{M-1}\right)\frac{SCE}{SCT} \tag{27}$$

Además se tiene que:

$$\begin{aligned} &\frac{SCC}{SCT} = 1-\frac{SCE}{SCT} = 1-\frac{M-1}{M}(1-\rho) = \frac{1+\rho(M-1)}{M} \\ \Rightarrow\quad &\rho = \frac{M\times SSC - SCT}{SCT(M-1)} \end{aligned} \tag{28}$$

Asímismo, tendríamos las siguientes relaciones

$$SCC = \sum_{i=1}^{N} M(\mu_i-\mu)^2 \tag{29}$$

$$SCT = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2 \tag{30}$$

$$SCT = (K-1)\sigma^2 = (NM-1)\sigma^2 \tag{31}$$

Reemplazando (29), (30) y (31) en (28) convenientemente

$$\begin{aligned} \rho &= \frac{M\times\sum_{i=1}^{N}M(\mu_i-\mu)^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2}{(NM-1)(M-1)\sigma^2} \\ &= \frac{\sum_{i=1}^{N}\left(M(\mu_i-\mu)\right)^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2}{(NM-1)(M-1)\sigma^2} \\ &= \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{M}(y_{ij}-\mu)\right)^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2}{(NM-1)(M-1)\sigma^2} \\ &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)\left(\sum_{k=1}^{M}(y_{ik}-\mu)-(y_{ij}-\mu)\right)}{(NM-1)(M-1)\sigma^2} \\ &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)\sum_{k\neq j}^{M}(y_{ik}-\mu)}{(NM-1)(M-1)\sigma^2} \\ &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k\neq j}^{M}(y_{ij}-\mu)(y_{ik}-\mu)}{(NM-1)(M-1)\sigma^2} \end{aligned} \tag{32}$$

□

Capítulo 4 - Ejercicio 13

Un estudiante perteneciente a un internado desea estimar el promedio final medio que alcanzaron él y sus compañeros en un curso de la institución. En lugar de obtener un listado de todos sus compañeros y realizar un **MASs**, él se da cuenta que los alumnos de su institución están distribuidos en 100 cuartos de 4 alumnos cada uno; por lo que decide seleccionar al azar 5 de estos cuartos y preguntarles a todos los estudiantes en ellos, el puntaje promedio que obtuvieron en el curso. Los resultados son los siguientes

	Cuarto				
Alumno No.	1	2	3	4	5
1	15.4	11.8	10	15	13.4
2	13	15.2	12.8	14.4	9.6
3	17.2	16.4	12.6	17.2	16.4
4	15.2	13.4	9.4	18.2	16

a) Obtenga la estimación buscada, junto con su error estándar de estimación.

Solución. La expresión general para el estimador de la media es:

$$\bar{Y} = \frac{1}{K}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\frac{NM_i}{nm_i}Y_{ij}$$

Procedemos a cargar los datos y calcular el estimador.

Listing 12: Cargando datos y estimando.

```
library(data.table)
p14<-data.table(puntaje=c
    (15.4,13,17.2,15.2,11.8,15.2,16.4,13.4,10,12.8,12.6,9.4,15,14.4,
        17.2,18.2,13.4,9.6,16.4,16),cuarto=rep(1:5,each=4))

K <- 400
N <- 100
M <- rep(4,5)
n <- length(p14[,unique(cuarto)])
m <- p14[,.N,cuarto]$N

sum(N*M*p14$puntaje/(n*m))/K
```

Resultado: 14.13.

La varianza del estimador de la media poblacional está dada por:

$$V(\bar{Y}) = (1 - \frac{n}{N})\frac{\sigma_m^2}{n}$$

Entonces, calculamos el error estándar:

Listing 13: Error estándar.

```
vartot<-sd(p14[,mean(puntaje),cuarto]$V1)**2
varest<-(1-n/N)*vartot/n

sqrt(varest)
```

Resultado: 0.8183245.

□

b) Halle una estimación de la correlación intraclase.

Solución. La correlación intraclase para conglomerados del mismo tamaño está definida por:

$$\rho = 1 - \frac{M}{M-1} \times \frac{\text{SCC}}{\text{SCT}}$$

Calculamos:

Listing 14: Correlación intraclase.

```
M<-unique(M)
a_res<-anova(object = lm(puntaje~as.factor(cuarto),data = p14))
scc<-a_res$`Sum Sq`[1]
sct<-sum(a_res$`Sum Sq`)

rho<-1-(M/(M-1))*scc/sct
rho
```

Resultado: 0.402225

Vemos que hay un grado importante de correlación en las observaciones al interior de cada cuarto o conglomerado. Entonces la estimación será menos eficiente (en términos de varianza del estimador) que un MASs.

□

c) Estime el efecto de este diseño.

Solución. Empleamos la forma de D_{eff} que emplea ρ en el cálculo:

Listing 15: Efecto diseño.

```
(N*M-1)*(1+rho*(M-1))/(M*(N-1))
```

Resultado: 2.223392

Cómo se anticipaba por el valor de ρ, el diseño es menos eficiente que un MASs.

□