Técnicas de Muestreo

Lista de Ejercicios 2

Justo Andrés Manrique Urbina – 20091107

Pregunta 1: Ver al final del informe.

Pregunta 2:

11.- En el MAE hemos tomado siempre la estrategia de obtener los tamaños de muestra de acuerdo a las especificaciones del máximo error de estimación tolerable para estimar un parámetro poblacional a un nivel de confianza dado. En ciertas situaciones sin embargo, el investigador pudiera estar interesado en tratar de estimar el parámetro de interés para cada estrato con un máximo error de estimación prefijado en él a un nivel de confianza dado. La pregunta entonces es ¿cuál es el máximo error de estimación que se estaría cometiendo al estimar con este procedimiento el parámetro en toda la población para el nivel de confianza dado? Resuelva este problema para el caso del ejercicio 6, asumiendo que en él se desee estimar el número total de horas de trabajo perdidas al interior de cada estrato con un error no mayor a las 100 horas y una confianza del 95 %.

Resolución en R:

```
# Limpieza del ambiente en R para correr el flujo
rm(list=ls())

# Carga de librería y excel que contiene la tabla del ejercicio
library(readxl)
pba <- read_xlsx("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Inputs/P2_db.xlsx")

# Identificación del número de la muestra por cada estrato (Obreros, Técnicos, Administradores)
z <- qnorm(1-0.05/2)
n_ob <- ceiling(z^2 * pba[1,1]*pba[2,1]/((z^2*pba[1,1])+(100/pba[2,1])^2*pba[2,1]))
n_tec <- ceiling(z^2 * pba[1,2]*pba[2,2]/((z^2*pba[1,2])+(100/pba[2,2])^2*pba[2,2]))
n_adm <- ceiling(z^2 * pba[1,3]*pba[2,3]/((z^2*pba[1,3])+(100/pba[2,3])^2*pba[2,3]))

# Identificación individual de los errores por cada estrato
e_ob <- (1-(n_ob/pba[2,1]))*(pba[2,1])^2*(pba[1,1]/n_ob)
e_tec <- (1-(n_tec/pba[2,2]))*(pba[2,2])^2*(pba[1,2]/n_tec)
e_adm <- (1-(n_adm/pba[2,3]))*(pba[2,3])^2*(pba[1,3]/n_adm)

# Identificación del error estándar global
sub_et <- (e_ob+e_tec+e_adm)^(0.5)
et <- z * sub_et
et
```

Resultados:

El error de estimación para el nivel de confianza y error dados es de ± 163.9221

Pregunta 3:

12.- Suponga que en el MAE de la subsección 3.6.2 para el CES 2016 de Amazonas, le piden a usted que reporte las estimaciones del rendimiento medio en Matemáticas por sexo.
a) De estas estimaciones, junto con sus errores estándar de estimación estimados.
b) ¿Cómo haría para comparar el rendimiento medio de las estudiantes mujeres que pertenecen a colegios estatales y no estatales? ¿Se podría concluir, con una confianza del 95 %, que hay diferencias entre estos rendimientos medios?
c) responda b) para el caso de los estudiantes hombres.

Resolución en R:

```
# Limpieza de ambiente para ejecución de script
rm(list=ls())

# Carga de datos y seteo de semilla
library(haven)
load("D:/ce2s16Cz.rdata")
set.seed(12329)

#Afijación de muestra según asignación de Neyman
Pop = ce2s16Cz
Pop$Estrato=interaction(Pop$Area,Pop$Gestion)
Pop = Pop[order(Pop$Estrato),]
table(Pop$Estrato)
Nh = as.vector(table(Pop$Estrato))
sigmah = sd(Pop$M500_M[Pop$Estrato=="Urbana.Estatal"][sample(Nh[1],10)])
sigmah[2] = sd(Pop$M500_M[Pop$Estrato=="Rural.Estatal"][sample(Nh[2],10)])
sigmah[3] = sd(Pop$M500_M[Pop$Estrato=="Urbana.No estatal"][sample(Nh[3],10)])
sigmah[4] = sd(Pop$M500_M[Pop$Estrato=="Rural.No estatal"][sample(Nh[4],10)])
ah = Nh*sigmah/sum(Nh*sigmah)
d = dim(Pop)[1]*5/qnorm(0.975)
n = sum(((Nh*sigmah)^2)/ah)/(d^2 + sum(Nh*sigmah^2))
nh = round(ah*n)

# Determinación de la muestra
library(sampling)
set.seed(12345)
m=strata(Pop,c("Estrato"),size=nh,method="srswor")
me16Am = getdata(Pop,m)
table(is.na(me16Am$M500_M))
me16Am = me16Am[is.na(me16Am$M500_M)==0,]
nh = as.vector(table(me16Am$Estrato))
nh
me16Am = cbind(me16Am,fpc = rep(Nh,nh))
save(me16Am,file="D:/me16Am.RData")

#### Pregunta 12.a ####
library(survey)
dis16MAE = svydesign(id=~1,strata=~Estrato,fpc=~fpc,data=me16Am)
svyby(~M500_M,~Sexo,design=dis16MAE,svymean)

#### Pregunta 12.b y c####
svyby(~M500_M,~Gestion+~Sexo,design=dis16MAE,svymean,vartype = "var")
```

Resultados:

Pregunta a)

Sexo	Rendimiento medio M500_M	Error estándar
Hombre	541.2206	4.141779
Mujer	533.0612	4.084729

Pregunta b) y c)

Gestión	Sexo	Rendimiento medio M500_M	Varianza
Estatal	Hombre	536.8491	18.54166
No estatal	Hombre	565.1348	156.33394
Estatal	Mujer	528.2654	18.46652
No estatal	Mujer	566.9353	141.55921

Posteriormente, con dicho cuadro se determinan las diferencias entre las medias de cada sexo, por gestión escolar:

- Diferencia entre medias de las mujeres: $528.2654 - 566.9353 = -38.6699$
- Diferencia entre medias de los hombres: $536.8491 - 565.1348 = -28.2857$

Luego se halla el intervalo de confianza de dicha diferencia en relación a la siguiente fórmula:

$$\overline{D} \pm z_{1-\frac{a}{2}}\sqrt{\hat{v}(\bar{y}_{m_E}) + \hat{v}(\bar{y}_{m_{NE}})}$$

- Diferencias de medias con el intervalo de confianza, sector mujeres:

$$-38.6669 \pm 1.96\sqrt{18.46652 + 141.55921}$$

$$-38.6699 \pm 24.794$$

- Diferencias de medias con el intervalo de confianza, sector hombre:

$$-28.2857 \pm 1.96\sqrt{18.54166 + 156.33394}$$

$$-28.2857 \pm 25.919$$

En ambos sexos, existen diferencias significativas al 95% entre el rendimiento medio de la sección Matemática de las gestiones educativas del sector estatal y no estatal. Se observa que el sector no estatal tiene un mayor rendimiento que el estatal.

Pregunta 4:

17.- Considere el ejercicio 26 del capítulo anterior y replique este estudio, pero ahora utilizando un MAE con asignación de Neyman, donde su variable de estratificación será el año de la película. Concretamente considere en un primer estrato a aquellas películas que sean anteriores a 1970, otro estrato con películas entre los 70 y anteriores a los 80, un tercer estrato con películas de entre los 80 y anteriores a los 90 y un estrato final con películas de los 90 hasta la actualidad.

Resolución en R:

```
rm(list=ls())
library(readxl)
IMDb_Work <- read_excel("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Inputs/IMDb_Work.xlsx", col_types = c("text", "text", "numeric", "numeric", "blank"))
IMDb_WorkingDB <- IMDb_Work
IMDb_WorkingDB <- IMDb_WorkingDB[order(IMDb_WorkingDB$Estrato),]
set.seed(9875)
IMDb_Summary <- as.vector(table(IMDb_WorkingDB$Estrato))
IMDb_PreSample <- sample(IMDb_Summary[1],5)
IMDb_PreSample <- append(IMDb_PreSample,sample(IMDb_Summary[2],5))
IMDb_PreSample <- append(IMDb_PreSample,sample(IMDb_Summary[3],5))
IMDb_PreSample <- append(IMDb_PreSample,sample(IMDb_Summary[4],5))

library(xlsx)
write.xlsx (IMDb_PreSample,"D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Outputs/PreSample.xlsx")
write.xlsx (IMDb_WorkingDB,"D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Outputs/IMDb_WorkingDB.xlsx")

library(readxl)
sigmah <- read_excel("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Outputs/3. Strata_Piloto.xlsx")
sigmah <- as.vector(sigmah$SD)

ah <- IMDb_Summary*sigmah/sum(IMDb_Summary*sigmah)
d = dim(IMDb_WorkingDB)[1]*0.1/qnorm(0.975)
n = sum(((IMDb_Summary*sigmah)^2)/ah)/(d^2 + sum(IMDb_Summary*sigmah^2))
nh = ceiling(ah*n)
## Se suma +1 a todos los estratos para obtener variabilidad
nh = nh+1


library(sampling)
m <- strata(IMDb_WorkingDB,c("Estrato"),size = nh,method = "srswor")
mdata <- getdata(IMDb_WorkingDB,m)

write.xlsx(mdata,"D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Outputs/FinalSample.xlsx")

FinalSample <- read_excel("D:/Justo Andrés/Dropbox/Maestría en Estadística/2018 - 1/Técnicas de Muestreo/Tarea - Lista 2/Proyecto_Lista2/Outputs/4. FinalSample.xlsx")
mn_FS <- mean(FinalSample$SD)
FinalSample_1 <- FinalSample[FinalSample$Estrato=="1",]
V_Y1 <-
(dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="1",])[1]/dim(IMDb_WorkingDB)[1])^2*(1-
```

```
dim(FinalSample[FinalSample$Estrato=="1",])[1]/dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="1",])[1])*(var(FinalSample_1$SD)/(dim(FinalSample[FinalSample$Estrato=="1",])[1]))

  FinalSample_2 <- FinalSample[FinalSample$Estrato=="2",]
  V_Y2 <-
(dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="2",])[1]/dim(IMDb_WorkingDB)[1])^2*(1-dim(FinalSample[FinalSample$Estrato=="2",])[1]/dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="2",])[1])*(var(FinalSample_2$SD)/(dim(FinalSample[FinalSample$Estrato=="2",])[1]))

  FinalSample_3 <- FinalSample[FinalSample$Estrato=="3",]
  V_Y3 <-
(dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="3",])[1]/dim(IMDb_WorkingDB)[1])^2*(1-dim(FinalSample[FinalSample$Estrato=="3",])[1]/dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="3",])[1])*(var(FinalSample_3$SD)/(dim(FinalSample[FinalSample$Estrato=="3",])[1]))

  FinalSample_4 <- FinalSample[FinalSample$Estrato=="4",]
  V_Y4 <-
(dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="4",])[1]/dim(IMDb_WorkingDB)[1])^2*(1-dim(FinalSample[FinalSample$Estrato=="4",])[1]/dim(IMDb_WorkingDB[IMDb_WorkingDB$Estrato=="4",])[1])*(var(FinalSample_4$SD)/(dim(FinalSample[FinalSample$Estrato=="4",])[1]))

  sd_est <- (V_Y1+V_Y2+V_Y3+V_Y4)^(0.5)

  z <- qnorm(1-0.05/2)

  IC <- sd_est*z
```

Resolución en R:

1. Con el propósito de realizar la afijación de Neyman, se genera una muestra piloto de 5 películas por estrato, con el fin de obtener desviaciones estándar iniciales. Ver tabla a continuación:

Rank & Title	Año	Estrato	Media	SD
25. ¡Qué bello es vivir! (1946)	1946	1	8.48	1.76
134. Judgment at Nuremberg (1961)	1961	1	8.06	1.78
171. Rebecca (1940)	1940	1	8.05	1.68
197. Persona (1966)	1966	1	8.00	1.97
226. Les diaboliques (1955)	1955	1	7.87	1.85
2. El padrino (1972)	1972	2	8.93	1.81
22. La guerra de las galaxias (1977)	1977	2	8.54	1.64
52. Alien - El octavo pasajero (1979)	1979	2	8.38	1.44
104. Monty Python and the Holy Grail (1975)	1975	2	8.24	1.71
240. Dos extraños amantes (1977)	1977	2	7.96	1.81
13. El imperio contraataca (1980)	1980	3	8.63	1.62
44. Volver al futuro (1985)	1985	3	8.45	1.41
70. Érase una vez en América (1984)	1984	3	8.35	1.60
148. El hombre elefante (1980)	1980	3	8.13	1.52
165. La cosa (1982)	1982	3	8.13	1.57
162. Petróleo sangriento (2007)	2007	4	8.06	1.79
167. Identidad peligrosa (1998)	1998	4	8.12	1.73
176. Mary and Max (2009)	2009	4	8.12	1.63
201. 12 años de esclavitud (2013)	2013	4	8.07	1.50

237. The Bourne Ultimatum (2007)	2007	4	8.09	1.44

2. En base a la muestra piloto, se halló la medida de controversia por cada uno de los estratos, a fin de que esto pueda utilizarse en la afijación de Neyman.

Estrato	SD
1	0.11
2	0.15
3	0.08
4	0.15

En base a ello, se tienen los siguientes tamaños de muestra iniciales:

Estrato 1	Estrato 2	Estrato 3	Estrato 4
2	1	1	4

Como se verá más adelante, se requiere obtener la varianza muestral para reportar los límites de confianza. En aquellos estratos cuyo tamaño de muestra es 1 tendrá varianza 0, por lo que no se podrá computar el error de estimación adecuadamente. Por lo tanto, se subió una unidad a todos los estratos, teniendo así el siguiente tamaño de muestra:

Estrato 1	Estrato 2	Estrato 3	Estrato 4
3	2	2	5

3. En base a lo indicado anteriormente, se extrae la muestra. Ver tabla a continación:

Rank & Title	Año	Estrato	Mean	SD
97. Ladri di biciclette (1948)	1948	1	8.174394	1.77
109. Metropolis (1927)	1927	1	7.757188	2.51
125. Ikiru (1952)	1952	1	7.761202	2.38
51. Apocalypse Now (1979)	1979	2	8.342868	1.67
89. Taxi Driver (1976)	1976	2	8.331265	1.52
43. Los cazadores del arca perdida (1981)	1981	3	8.424583	1.44
214. Nausicaä del Valle del Viento (1984)	1984	3	8.167336	1.63
21. Ciudad de Dios (2002)	2002	4	8.540778	1.56
170. Kill Bill: Vol. 1 (2003)	2003	4	8.241051	1.51
179. Eskiya (1996)	1996	4	8.555396	1.98
232. Hechizo del tiempo (1993)	1993	4	8.013363	1.48
243. Guardianes de la Galaxia (2014)	2014	4	7.866823	1.73

4. Una vez obtenida la muestra, se obtiene la media muestral de la columna SD la cual es 1.7658. El límite de confianza se estima mediante las siguientes fórmulas:

$$\hat{SE}(\bar{Y}) = \sqrt{\hat{V}(\bar{Y})} = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}}.$$

$$[\bar{Y} - z_{1-\frac{\alpha}{2}} \hat{SE}(\bar{Y}), \bar{Y} + z_{1-\frac{\alpha}{2}} \hat{SE}(\bar{Y})]$$

Se obtiene el siguiente resultado:

$$1.7658 \pm 0.1528$$

5. Posteriormente, se obtiene los datos de la película Whiplash. Ver tabla a continuación:

Rank & Title	Año	Mean	SD
46. Whiplash: Música y obsesión (2014)	2014	8.596576	1.57

Dado que el límite inferior del límite de confianza es de 1.613, podríamos calificar la película Whiplash como controversial en tanto este término se entienda como que esté fuera de los límites de confianza de nuestra estimación.