Técnicas de Muestreo

Lista de Ejercicios N° 1

Nombre: Justo Andrés Manrique Urbina
Código: 20091107

Pregunta N°3

a) Halle la probabilidad de que la media del nivel de hemoglobina de las tres personas seleccionadas supere los 14 gramos por decilitro.

Resolución para muestra MASs – Código en R

```
rm(list = ls())

# 1. Creación del vector poblacional
pob <- c(13.9,11.5,16.7,14.4, 14.6, 15.1)

# 2. Carga de la librería combinat
library(combinat)

# 3.1 Creación de la matriz de la combinatoria, y generación de la media por fila
samples_pob <- t(as.matrix(combn(pob,3)))
spobmean <- cbind(samples_pob,apply(samples_pob,1,mean))

# 4.1 Cálculo de probabilidades mayor a 14 g/dl

sum(spobmean[,4] > 14)/length(spobmean[,4])
```

Resultado: 0.7

Resolución para muestra MASc – Código en R:

```
# 1. Creación del vector poblacional
pob <- c(13.9,11.5,16.7,14.4, 14.6, 15.1)

# 2. Carga de la librería combinat
library(combinat)

#3.2 Creación de la matriz y generación de la media por fila
samples_pob_c <- expand.grid(rep(list(pob),3))
spobmean_c <- cbind(samples_pob_c,apply(samples_pob_c,1,mean))

#4,2 Cálculo de probabilidades mayor a 14g/dl
sum(spobmean_c[,4] > 14)/length(spobmean_c[,4])
```

Resultado: 0.699

b) Suponga que para estimar el nivel promedio de hemoglobina en estas 6 personas se propone la mediana de los valores observados en la muestra, ¿sería este un estimador insesgado? ¿tiene este una menor varianza que la media muestral?

Resolución para muestra MASs – Código en R:

```
# Se utilizan las variables de la pregunta 3a

# 1. Añadir las probabilidades de cada muestra
probs <- rep(1/length(samples_pob[,1]),length(samples_pob[,1]))

# 2. Añadir mediana, mediana, y probabilidad a samples_pob
spobmeanvarmed <-
cbind(samples_pob,apply(samples_pob,1,mean),apply(samples_pob,1,median),apply(sampl
es_pob,1,var),probs)

# 3. Determinar la probabilidad de la media, medianta y varianza
meanaggr <- aggregate(spobmeanvarmed[,7],by=list(spobmeanvarmed[,4]),sum)
colnames(meanaggr) <- c("Media Muestral","Probabilidad")
medaggr <-  aggregate(spobmeanvarmed[,7],by=list(spobmeanvarmed[,5]),sum)
colnames(medaggr) <- c("Mediana Muestral", "Probabilidad")
varaggr <- aggregate(spobmeanvarmed[,7],by=list(spobmeanvarmed[,6]),sum)
colnames(varaggr) <- c("Varianza Muestral", "Probabilidad")

# 4. Determinar la media, varianza y mediana
rep3b <-
c(sum(meanaggr[,1]*meanaggr[,2]),sum(varaggr[,1]*varaggr[,2]),sum(medaggr[,1]*medagg
r[,2]))

# 5, Determinar la varianza de la media/mediana
sum(((meanaggr[,1]-sum(meanaggr[,1]*meanaggr[,2]))^2)*meanaggr[,2])
sum(((medaggr[,1]-sum(medaggr[,1]*medaggr[,2]))^2)*medaggr[,2])
```

Resultado:

	Estimador puntual	Varianza	Conclusión
Media Muestral	14.37	0.482	El estimador de la media muestral es insesgado, puesto que es igual a la media poblacional: 14.37.
Mediana Muestral	14.50	0.15	El estimador es sesgado, puesto que no es igual a la media poblacional. Sin embargo, presenta una menor varianza que la media muestral.

Resolución para muestra MASc – Código en R

```
# Se utilizan las variables de la pregunta 3a

# 1. Añadir las probabilidades de cada muestra
probs_c <- rep(1/length(samples_pob_c[,1]),length(samples_pob_c[,1]))

# 2. Añadir mediana, mediana, y probabilidad a samples_pob
spobmeanvarmed_c <- cbind(samples_pob_c,apply(samples_pob_c,1,mean),apply(samples_pob_c,1,median),apply(samples_pob_c,1,var),probs_c)

# 3. Determinar la probabilidad de la media, medianta y varianza
meanaggr_c <- aggregate(spobmeanvarmed_c[,7],by=list(spobmeanvarmed_c[,4]),sum)
colnames(meanaggr) <- c("Media Muestral","Probabilidad")
medaggr_c <-  aggregate(spobmeanvarmed_c[,7],by=list(spobmeanvarmed_c[,5]),sum)
colnames(medaggr) <- c("Mediana Muestral", "Probabilidad")
varaggr_c <- aggregate(spobmeanvarmed_c[,7],by=list(spobmeanvarmed_c[,6]),sum)
colnames(varaggr) <- c("Varianza Muestral", "Probabilidad")

# 4. Determinar la media, varianza y mediana
rep3b_c <- c(sum(meanaggr_c[,1]*meanaggr_c[,2]),sum(varaggr_c[,1]*varaggr_c[,2]),sum(medaggr_c[,1]*medaggr_c[,2]))

# 5, Determinar la varianza de la media/mediana
sum(((meanaggr_c[,1]-sum(meanaggr_c[,1]*meanaggr_c[,2]))^2)*meanaggr_c[,2])
sum(((medaggr_c[,1]-sum(medaggr_c[,1]*medaggr_c[,2]))^2)*medaggr_c[,2])
```

	Estimador puntual	Varianza	Conclusión
Media Muestral	14.37	0.804	El estimador de la media muestral es insesgado, puesto que es igual a la media poblacional: 14.37.
Mediana Muestral	14.44	1.16	El estimador es sesgado, puesto que no es igual a la media poblacional. Sin embargo, presenta una menor varianza que la media muestral

c) Usando los números aleatorios 0.018, 0.310 y 0.549 tome las muestras requeridas y estime la media del nivel de hemoglobina de las 6 personas.

Resolviendo con MASc:

N°	Valor	Probabilidad	Acumulada
1	13.9	0.167	0.167
2	11.5	0.167	0.333
3	16.7	0.167	0.500
4	14.4	0.167	0.667
5	14.6	0.167	0.833
6	15.1	0.167	1.000

Media Muestral: (13.9+11.5+14.4)/3 = 13.27

Resolviendo con MASs:

N°	Valor	Probabilidad	Acumulada
1	13.9	0.167	0.167
2	11.5	0.167	0.333
3	16.7	0.167	0.500
4	14.4	0.167	0.667
5	14.6	0.167	0.833
6	15.1	0.167	1.000

N°	Valor	Probabilidad	Acumulada
2	11.5	0.2	0.2
3	16.7	0.2	0.4
4	14.4	0.2	0.6
5	14.6	0.2	0.8
6	15.1	0.2	1.000

N°	Valor	Probabilidad	Acumulada
3	16.7	0.25	0.25
4	14.4	0.25	0.5
5	14.6	0.25	0.75
6	15.1	0.25	1.000

Media Muestral: (13.9+16.7 +14.6)/3 = 15.07

Pregunta N°9

a) Halle, con la fórmula estándar, el tamaño de muestra para este estudio.

Resolución de la pregunta mediante código en R:

```
z = qnorm(1-0.05/2)
p = 2/30
error = 3/100
N = 200

## Pregunta 9.a ##
((z^2*p*(1-p))*N) / ((z^2*p*(1-p)) + error^2*(N-1))
```

Resultado: El tamaño de muestra para el estudio es de 114,

Pregunta N°18

f) Obtenga el tamaño de muestra que se necesitaría en una encuesta futura para medir el consumo de agua en la zona rural, si es que se deseara estimar T_d con un margen de error no mayor a 950,000 litros con una confianza al 95%. Suponga que en la encuesta se encontró

que el 65% de los hogares contaban con servicios de agua y desague y en promedio ellos consumieron en el mes 12,000 litros con una desviación estándar de 1,540 litros. ¿Qué estimación de T_d le da esta encuesta?

Resolución en R:

```
o2_d <- 1540 ^ 2
z <- qnorm(1-0.05/2)
N <- 3000
qd <- 1 - pd
pd <- 0.65
Nd <- N*pd
ud <- 12000
e <- 950000

# Cálculo de la muestra
n <- (((Nd-1)*o2_d + qd*Nd*ud^2)*z^2*(N^2))/((((Nd-
1)*o2_d+qd*Nd*ud^2)*z^2*N)+e^2*(N-1))

consumo <- N * ud
```

Resultado: La muestra necesaria es de 914. La estimación de T_d da 36,000,000 litros.

Pregunta N° 25

a) Realice una muestra piloto de 12 películas y a partir de ella obtenga las estimaciones necesarias para el tamaño final de muestra.

Resolución en R:

```
rm(list=ls())
set.seed(123456)
sort(sample(250,12), decreasing=FALSE)
```

Resultado: [1] 24 41 49 85 89 98 131 142 188 192 200 240

Análisis mediante Excel:

N°	Título	Media[1]	SD[2]
24	It's a Wonderful Life	8.5	1.755
41	Terminator 2	8.4	1.476
49	Memento	8.4	1.534
85	Lawrence of Arabia	8.3	1.919
89	Singin' in the Rain	8.3	1.734
98	The Kid	8.3	1.673
131	Incendies	8.2	1.591

[1] Fuente: https://www.imdb.com/chart/top
[2] Fuente: IMDb; por otro lado, para extraer la desviación estándar se hizo uso de la web http://knowpapa.com/sd-freq/

N°	Título	Media[1]	SD[2]
142	Lock, Stock and Two Smoking Barrels	8.2	1.454
188	En el nombre del padre	8.1	1.41
192	The Grand Budapest Hotel	8.1	1.51
200	Memories of Murder	8.1	1.551
240	The Best Years of Our Lives	8	1.941

Con el objetivo de obtener la medida de controversia, se extrajo la desviación estándar muestral de la cuadra columna del Cuadro N°1 a través de la fórmula DESVEST.M en Excel, obteniéndose 0.177 cómo resultado.

Posteriormente, se sacó la nueva muestra a través del programa R, conforme se ve a continuación:

```
rm(list=ls())
set.seed(123456)
sort(sample(250,12), decreasing=FALSE)

z = qnorm(1-0.05/2)
o2 = 0.177^2
N = 250
e = 0.1

(z^2*o2*N)/((z^2*o2)+(e^2*N))
```

Resultado: 11.5 (redondeado a 12)

Con dicho resultado, se extrajo nuevamente una muestra a través del programa R:

```
set.seed(6553)
sort(sample(250,12),decreasing=FALSE)
```

Resultado: [1] 12 13 34 36 82 133 144 168 206 222 236 245

Análisis en Excel:

N°	Título	Media[3]	SD[4]
12	Forrest Gump (1994)	8.7	1.457
13	Star Wars: Episode V - The Empire Strikes Back(1980)	8.7	1.591
34	Once Upon a Time in the West (1968)	8.5	1.665
36	Casablanca (1942)	8.5	1.718
82	Like Stars on Earth (2007)	8.3	1.543

[3] Fuente: https://www.imdb.com/chart/top
[4] Fuente: IMDb; por otro lado, para extraer la desviación estándar se hizo uso de la web http://knowpapa.com/sd-freq/

N°	Título	Media[3]	SD[4]
133	Judgment at Nuremberg (1961)	8.2	1.368
144	Casino (1995)	8.2	1.377
168	Sunrise (1927)	8.1	1.186
206	Logan (2017)	8.1	1.502
222	Monsters, Inc. (2001)	8	1.353
236	8½ (1963)	8	2.006
245	Gangs of Wasseypur (2012)	8	1.432

Resolución en R:

```
controversia_s <- c(1.457,1.591,1.665,1.718,1.543,1.368,1.377,1.186,1.502,1.353,2.006,1.432)
media_cont <- mean(controversia_s)
sd_cont <- sd(controversia_s)
z <- qnorm(1-0.05/2)
n <- 12
N <- 250
r <- sqrt(1-n/N)
L1 <- media_cont - z*r*sd_cont/sqrt(n)
L2 <- media_cont + z*r*sd_cont/sqrt(n)
```

Resultado: [1.4; 1.63] ó 1.52 +- 0.117

c) Según los resultados, ¿podría decir que Whiplash - Música y Obsesión (2014) es una película de calificación controversial?

Título	Media[5]	SD[6]
Whiplash	8.5	1.771

Resultado: Se observa que la desviación estándar supera el resultado de la 3b) por lo que se considera la película como controversial.

[5] Fuente: https://www.imdb.com/chart/top

[6] Fuente: IMDb; por otro lado, para extraer la desviación estándar se hizo uso de la web http://knowpapa.com/sd-freq/

Pregunta N° 30

a) Tome en esta base de datos un MASs de tamaño n = 500 y estime con la librería survey la diferencia de medias del índice api para estos años.

Resolución en R:

```
library("survey")
set.seed(5253)
data(api)

# Obtención de la población general
N = dim(apipop)[1]

# Declaración del tamaño de muestra
n = 500

# Generación de la muestra
muestra = sample(N,n)

# Obtención de la muestra
sample_pob = apipop[muestra,]
disMASs = svydesign(id=~1,fpc=rep(N,n),data = sample_pob)

# obtención de las medias
means = svymean(~api00+api99,disMASs)

# Contraste entre medias
contr = svycontrast(means,c(api00=1,api99=-1))
```

Resultados:

	Contrast	SE
Contraste	32.4	1.27

b) Obtenga, con la librería survey un intervalo de confianza al 95% para la diferencia anterior.

Resolución en R:

```
confint(contr)
```

Resultado:

	2.5%	97.5%
Contraste	29.9	34.9

c) Con la misma muestra tomada en a) obtenga el IC en b) sin utilizar el paquete survey

Resolución en R:

```
diferencia = mean(sample$api00 - sample$api99)
var_api99 = var(sample$api99)
```

```
var_api00 = var(sample$api00)
cov_api9900 = cov(sample$api99,sample$api00)
e = qnorm(1-0.05/2)*sqrt((1 - n/N)/n)*sqrt(var_api99+var_api00-2*cov_api9900)
c(diferencia-e,diferencia+e)
```

Resultado: 29.9, 34.9g