ITAM

Laboratorio #9: Muestreo Aleatorio

Luis Martinez

Otoño 2021

.

Ejercicio 1

1 Let $N = 6$ and $n = 3$. For purposes of studying sampling distributions, assume that all population values are known.

$$y_1 = 98 \qquad y_2 = 102 \qquad y_3 = 154$$
$$y_4 = 133 \qquad y_5 = 190 \qquad y_6 = 175$$

We are interested in $\bar{y}_U$, the population mean. Two sampling plans are proposed.

- Plan 1. Eight possible samples may be chosen.

Sample Number	Sample, $\mathcal{S}$	$P(\mathcal{S})$
1	{1,3,5}	1/8
2	{1,3,6}	1/8
3	{1,4,5}	1/8
4	{1,4,6}	1/8
5	{2,3,5}	1/8
6	{2,3,6}	1/8
7	{2,4,5}	1/8
8	{2,4,6}	1/8

- Plan 2. Three possible samples may be chosen.

Sample Number	Sample, $\mathcal{S}$	$P(\mathcal{S})$
1	{1,4,6}	1/4
2	{2,3,6}	1/2
3	{1,3,5}	1/4

a What is the value of $\bar{y}_U$?

b Let $\bar{y}$ be the mean of the sample values. For each sampling plan, find

(i) $E[\bar{y}]$; (ii) $V[\bar{y}]$; (iii) Bias$(\bar{y})$; (iv) MSE$(\bar{y})$.

c Which sampling plan do you think is better? Why?

(a)

$$\bar{y}_{\mathcal{U}} = \frac{\sum_1^N y_i}{N}$$

```
#Creamos nuestro vector
N<- 6 # Valores de la poblacion
y<- c(98, 102, 154, 133, 190, 175)

y.barra<- sum(y)/N
paste0("La media poblacional es: ", y.barra)
```

```
## [1] "La media poblacional es: 142"
```

(b)

Tenemos que usar la siguiente expresión:

$$\mathbb{E}[\bar{y}] = \sum_{1}^{N} \bar{y}_i \mathbb{P}(S_i)$$

```
#Creamos nuestro vector
Muestras<- 8
ProbaSi<- 1/8
#Muestrearon {1,3,5}
y_1<- c(98,154,190)
y.barra_1<- sum(y_1)/3

#Muestrearon {1,3,6}
y_2<- c(98,154,175)
y.barra_2<- sum(y_2)/3

#Muestrearon {1,4,5}
y_3<- c(98,133,190)
y.barra_3<- sum(y_3)/3


#Muestrearon {1,4,6}
y_4<- c(98,133,175)
y.barra_4<- sum(y_4)/3

#Muestrearon {2,3,5}
y_5<- c(102,154,190)
y.barra_5<- sum(y_5)/3

#Muestrearon {2,3,6}
y_6<- c(102,154,175)
y.barra_6<- sum(y_6)/3
#Muestrearon {2,4,5}
y_7<- c(102,133,190)
y.barra_7<- sum(y_7)/3

#Muestrearon {2,4,6}
y_8<- c(102,133,175)
y.barra_8<- sum(y_8)/3

y_ProbaSi<- round(c(y.barra_1,y.barra_2,y.barra_3,y.barra_4,y.barra_5,y.barra_6,
                    y.barra_7,y.barra_8)*ProbaSi,2)
y_ProbaSi
```

```
## [1] 18.42 17.79 17.54 16.92 18.58 17.96 17.71 17.08
```

```
media.muestral<- round(sum(y_ProbaSi),0)
paste0("La media muestral es: ", media.muestral )
```

```
## [1] "La media muestral es: 142"
```

Ahora vamos con la varianza:

$$Var\left[\bar{y}\right] = \sum_{1}^{N}[\bar{y}_i - \mathbb{E}(\bar{y})]^2\mathbb{P}(S_i)$$

```
#Para la Varianza
y.barra_index<- c(y.barra_1,y.barra_2,y.barra_3,y.barra_4,y.barra_5,y.barra_6,
                  y.barra_7,y.barra_8)
varianza<- rep(NA,Muestras)
for(i in 1:Muestras){
  varianza[i]<- ((y.barra_index[i]-media.muestral)^2)*ProbaSi
}

varianza<- round(varianza,2)
varianza<- sum(varianza)

paste0("La varianza es: ", varianza)
```

```
## [1] "La varianza es: 18.96"
```

```
#Checamos el sesgo
sesgo<- media.muestral-y.barra
paste0("El sesgo es: ", sesgo)
```

```
## [1] "El sesgo es: 0"
```

```
#Checamos MSE
MSE<- varianza + (sesgo)^2
paste0("El MSE es: ", MSE)
```

```
## [1] "El MSE es: 18.96"
```

(c)

Ahora, consideramos el muestreo 2 y hacemos exactamente lo mismo pero con estos datos.

```
#Ahora tenemos diferentes probas en las muestras
ProbasSi<- c(1/4, 1/2, 1/4)
Muestras2<- 3
#Muestrearon {1,4,6}
y_1<- c(98,133,175)
y.barra_1<- sum(y_1)/3

#Muestrearon {2,3,6}
y_2<- c(102,154,175)
y.barra_2<- sum(y_2)/3

#Muestrearon {1,3,5}
y_3<- c(98,154,190)
y.barra_3<- sum(y_3)/3

y.barras<- c(y.barra_1,y.barra_2,y.barra_3)
y_ProbaSi2<- rep(NA,Muestras2)
for(i in 1:Muestras2){
  y_ProbaSi2[i]<- y.barras[i]*ProbasSi[i]
```

```
}
media.muestral2<- sum(y_ProbaSi2)
paste0("La media muestral para el esquema 2 es: ", media.muestral2)
```

```
## [1] "La media muestral para el esquema 2 es: 142.5"
```

```
#Varianzas
y.barra_index2<- c(y.barra_1,y.barra_2,y.barra_3)
varianza2<- rep(NA,Muestras2)
for(i in 1:Muestras2){
  varianza2[i]<- ((y.barra_index2[i]-media.muestral2)^2)*ProbasSi
}

varianza2<- round(varianza2,2)
varianza2<- sum(varianza2)

paste0("La varianza es: ", varianza2)
```

```
## [1] "La varianza es: 19.02"
```

```
#Checamos el sesgo esquema 2
sesgo2<- media.muestral2-y.barra
paste0("El sesgo es: ", sesgo2)
```

```
## [1] "El sesgo es: 0.5"
```

```
#Checamos MSE esquema 2
MSE2<- varianza2 + (sesgo2)^2
paste0("El MSE es: ", MSE2)
```

```
## [1] "El MSE es: 19.27"
```

Después de esto.... comenten cuál es el mejor esquema de muestreo (piénsenlo en términos de costos)

Ejercicio 2

A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. This number is not directly available on the database, so requires the investigator to examine each record separately. A frequency table for number of refereed publications is given below for an SRS of 50 faculty members.

Refereed Publications	0	1	2	3	4	5	6	7	8	9	10
Faculty Members	28	4	3	4	4	2	1	0	2	1	1

a Plot the data using a histogram. Describe the shape of the data.

b Estimate the mean number of publications per faculty member, and give the SE for your estimate.

c Do you think that $\bar{y}$ from (b) will be approximately normally distributed? Why or why not?

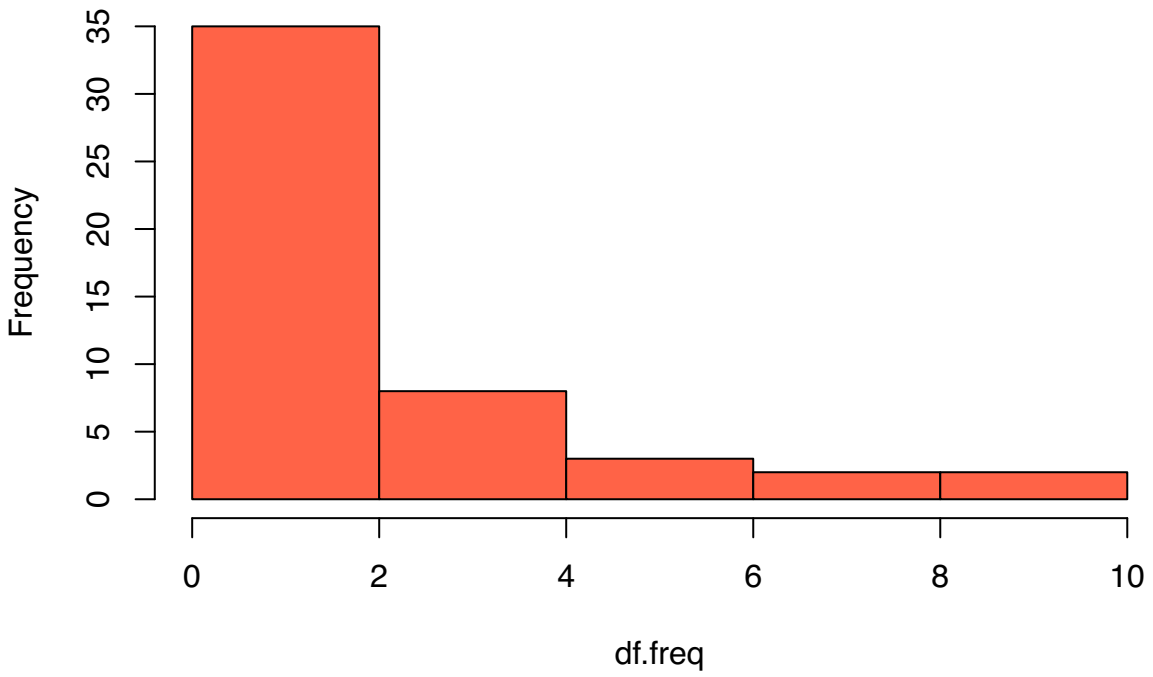
d Estimate the proportion of faculty members with no publications and give a 95% CI.

(a) y (b)

```
N<- 807
df<- as.data.frame(cbind(Refereed.Pubs= 0:10,
                         FacultyMembers=c(28,4,3,4,4,2,1,0,2,1,1)))
head(df)
```

```
##   Refereed.Pubs FacultyMembers
## 1             0             28
## 2             1              4
## 3             2              3
## 4             3              4
## 5             4              4
## 6             5              2
```

```
df.freq<- as.vector(rep(df$Refereed.Pubs, df$FacultyMembers))
hist(df.freq, col='tomato')
```




```
FacMemb<- as.vector(df$FacultyMembers)
pubsxmiembro<- as.vector(df$FacultyMembers*df$Refereed.Pubs)

media.publicaciones<- sum(pubsxmiembro)/(sum(FacMemb))
paste0("La media de publicaciones por miembro es: ", media.publicaciones)
```

```
## [1] "La media de publicaciones por miembro es: 1.78"
```

```
pubsxmiembro_2<- as.vector(df$FacultyMembers*(df$Refereed.Pubs)^2)

#Desviacion Estandar
desvest<- sqrt((sum(pubsxmiembro_2)/sum(FacMemb))
               -(media.publicaciones)^2)

#Error Estandar
EE<- (desvest/sqrt(sum(FacMemb)))*sqrt(1-(sum(FacMemb)/N))
paste0("El EE es: ", EE)
```

```
## [1] "El EE es: 0.363722363718055"
```

(c)

Ustedes opinan que $\bar{y}$ se distribuye normalmente??? Comenten.

(d)

```
#La proporcion de miembros sin publicaciones es:
N<- 807
pgorro<- 28/sum(FacMemb)
```

```r
#El Error Estandar de esta proporcion esta dado por:
EE_0pubs<- sqrt((pgorro*(1-pgorro)/(sum(FacMemb)-1))*(1-(sum(FacMemb)/N)))

# IC's al 95%
alfa<- .05
z<- qnorm(.025,0,1,lower.tail = FALSE)

confint.low<- round(pgorro-z*EE_0pubs,3)
confint.upp<- round(pgorro+z*EE_0pubs,3)

paste0("El IC al 95% es: [", confint.low ,", ", confint.upp, "]" )
```

```
## [1] "El IC al 95% es: [0.425, 0.695]"
```

Decision theoretic approach for sample size estimation. (Requires calculus.) In a decision theoretic approach, two functions are specified:

$$L(n) = \text{ Loss or "cost" of a bad estimate}$$
$$C(n) = \text{ Cost of taking the sample}$$

Suppose that for some constants c_0, c_1, and k,

$$L(n) = k\,V(\bar{y}_S) = k\left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$
$$C(n) = c_0 + c_1 n.$$

What sample size n minimizes the total cost $L(n) + C(n)$?

Definimos $F(n) = L(n) + C(n) = k\left(1 - \frac{n}{N}\right)\frac{S^2}{n} + c_0 + c_1 n$

Básicamente es un problema de minimización sin restricciones sobre n

$$\min_{\{n\}} F(n) \equiv \min_{\{n\}} k\left(1 - \frac{n}{N}\right)\frac{S^2}{n} + c_0 + c_1 n$$
$$\equiv \min_{\{n\}} \frac{kS^2}{n} - \frac{kS^2}{N} + c_0 + c_1 n$$

FOC:

$$[n]: \quad -\frac{kS^2}{n^2} + c_1 = 0 \iff \sqrt{\frac{kS^2}{c_1}} = n$$

Verificamos mínimo $\quad \frac{\partial^2 F(n)}{\partial n^2} = \frac{\partial}{\partial n}\left[-\frac{kS^2}{n^2} + c_1\right] = \frac{2kS^2}{n^3} > 0$

Dado a que $L(n) = k\left(1 - \frac{n}{N}\right)\frac{S^2}{n} > 0 \quad \because \quad S^2, n, N > 0$

$\frac{n}{N} < 1 \quad \therefore \quad 1 - \frac{n}{N} > 0$

$\therefore \quad k > 0$

$\therefore \quad n = \sqrt{\frac{kS^2}{c_1}}$ es el tamaño de muestra que minimiza $L(n) + C(n)$

(Requires probability.) A typical opinion poll surveys about 1000 adults. Suppose that the sampling frame contains 100 million adults including yourself, and that an SRS of 1000 adults is chosen from the frame.

a What is the probability that you are selected to be in the sample?

b Now suppose that 2000 such samples are selected, each sample selected independently of the others. What is the probability that you will *not* be in any of the samples?

c How many samples must be selected for you to have a 0.5 probability of being in at least one sample?

(a) $N = 100\,000\,000$ de una MAS con $n = 1000$

Queremos $P\{j \in S\} \equiv$ Probabilidad de que la persona j sea seleccionada

Si utilizamos MAS c/R $\Rightarrow P\{i_k = j\} = 1/N \quad \forall k = 1, \dots, n$

$\Rightarrow P\{j \in S\} = P\{i_1 = j\} + P\{i_2 = j\} + \dots + P\{i_n = j\}$

$= \sum_{1}^{n} \frac{1}{N} = \frac{n}{N}$

$= \frac{1000}{100\,000\,000} = .00001$

Recuerden MAS c/R

$\{i_1, \dots, i_n\}$, $1 \le i_t \le N$

$\Rightarrow P\{i_t \in S_i\} = 1/\binom{N}{n}$

$\therefore P\{i_t = j\} = 1/N$

↓

Proba ser seleccionada

(b) $|S| = 2000$

$\Rightarrow$ Queremos $P\{j \in S_1\} + P\{j \in S_2\} + \dots + P\{j \in S_{2000}\}$

$= .00001\,(2000) = .02$

$\therefore$ La proba de que la persona j ∉ en estas 2000 muestras

$1 - \sum_{1}^{2000} P\{j \in S_i\} = 1 - .02 = .98$

(c) Deseamos

$.5 = P\{j \in S_1\} + \dots + P\{j \in S_k\}$

$= .00001\,k \quad \Rightarrow \quad k = .5/.00001 = 50\,000$

$\therefore$ Necesitamos 50000 muestras de tamaño 1000

28 (Requires probability.) In an SRSWR, a population unit can appear in the sample anywhere between 0 and n times. Let

$$Q_i = \text{number of times unit } i \text{ appears in the sample,}$$

and

$$\hat{t} = \frac{N}{n} \sum_{i=1}^{N} Q_i y_i.$$

Extra Side Tiempo

a Argue that the joint distribution of $Q_1, Q_2, \ldots, Q_N$ is multinomial with n trials and $p_1 = p_2 = \cdots = p_N = 1/N$.

b Using (a) and properties of the multinomial distribution, show that $E[\hat{t}] = t$.

c Using (a) and properties of the multinomial distribution, find $V[\hat{t}]$.

Recordemos la distribución multinomial:

$$P\{X_1 = x_1, \ldots, X_k = x_k\} = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \cdots p_k^{x_k}$$

$$E[X_i] = np_i \,, \quad V[X_i] = np_i(1-p_i) \,, \quad \mathrm{Cov}(X_i, X_j) = -np_i p_j \,, \; i \neq j$$

Vamos a considerar $X_1, \ldots, X_n$ v.a.i's t.q $X_i \sim F_X(\mu_i, \sigma_i^2)$

Usamos el estadístico $W = \sum_1^n a_i X_i$ con $a_i > 0 \; \forall i$

$\therefore$ $\mu_W = \sum_1^n a_i \mu_i$ ——> Por Qué?

$$\sigma_W^2 = \sum_1^n a_i \sigma_i^2 + \sum_{i=1}^{N} \sum_{i \neq j = 2}^{n} a_i a_j \, \mathrm{Cov}(X_i, X_j)$$

Sea $Q_i \equiv$ # de veces que la unidad i aparece en la muestra

$\Rightarrow$ $0 \leq Q_i \leq n$. Tomamos $y_i \equiv$ medida de interés de la i-ésima unidad

$\Rightarrow$ $\hat{t} = \frac{N}{n} \sum_1^N Q_i y_i$

(a) Es fácil ver que la proba de seleccionar 1 unidad es $1/N$

$\therefore$ $p_1 = \ldots = p_N = 1/N$ $\therefore$ $\sum_1^N p_i = 1$

Como estamos muestreando n unidades tenemos $\sum_1^N Q_i = n$

$\Rightarrow$ Para seleccionar la unidad i Q_i veces $\forall i = 1, \ldots, N$ $\exists$

$$\binom{n}{Q_1, Q_2, \ldots, Q_N} \text{ formas}$$

$\therefore$ $Q_1, \ldots, Q_N \sim \text{Multinomial}\left(n, p_1 = \ldots = p_N = 1/N\right)$

(b) p.d] $E[\hat{t}] = t$ → ¿Qué significa esto?

$$E[\hat{t}] = E\left[\frac{N}{n}\sum_1^n Q_i y_i\right] = \frac{N}{n}\sum_1^N E[Q_i y_i]$$

$$= \frac{\cancel{N}}{\cancel{n}}\sum_1^N \frac{\cancel{n}}{\cancel{N}} E[y_i]$$

Porque $E[y_i] = y_i$ $\longrightarrow$ $= \sum_1^N y_i = t$ ✓

Sabemos que
$E[Q_i] = np_i = \frac{n}{N}$
$\sum_1^N y_i = t$
y por MAS

(c) $\mathrm{Var}(\hat{t}) = \mathrm{Var}\left[\frac{N}{n}\sum_1^n Q_i y_i\right]$

$$= \left(\frac{N}{n}\right)^2 \mathrm{Var}\left(\sum_1^n Q_i y_i\right) = \left(\frac{N}{n}\right)^2 \left[\sum_1^n y_i^2 V(Q_i) + \sum_{i=1}^N \sum_{i\neq j=2}^n y_i y_i \,\mathrm{Cov}(Q_i, Q_j)\right]$$

$$= (*)$$

Vemos que $V(Q_i) = np_i(1-p_i) = \frac{n(N-1)}{N^2}$ ∴ $p_i = 1/N$; $1-p_i = 1-\frac{1}{N} = \frac{N-1}{N}$

$$\mathrm{Cov}(Q_i, Q_j) = -np_ip_j = -\frac{n}{N^2}$$

$$\Rightarrow (*) = \left(\frac{N}{n}\right)^2 \left[\frac{n(N-1)}{N^2}\sum_1^n y_i^2 - \frac{n}{N^2}\sum_{i=1}^N \sum_{i\neq j=2}^n y_i y_i\right]$$

$$= \left(\frac{N}{n}\right)^2 \left(\frac{\cancel{n}}{\cancel{N^2}}\right)\left[\left(1-\frac{1}{N}\right)\sum_1^N y_i^2 - \frac{1}{N}\sum_{i=1}^N \sum_{i\neq j=2}^n y_i y_j\right]$$

$$= \left(\frac{N}{n}\right)\left[\sum_1^N y_i^2 - \frac{1}{N}\left(\sum_1^N y_i^2 + \sum_{i=1}^N \sum_{i\neq j=2}^n y_i y_j\right)\right]$$

$$= \left(\frac{N}{n}\right)\left[\sum_1^N y_i^2 - N\bar{y}^2\right]$$

$$\sum_1^N (y_i - \bar{Y})^2 = \sum_1^N y_i^2 - 2\sum_1^N y_i \bar{y} + N\bar{y}^2$$

$$= \sum_1^N y_i^2 - 2N\bar{y}^2 + N\bar{y}^2$$

$$= \sum_1^N y_i^2 - N\bar{y}^2$$