Ejercicio 1

Considere n unidades, n_1 de los cuales son indistinguibles de un tipo A_1, n_2 de otro tipo A_2 y así sucesivamente hasta n_k de otro tipo A_k.

a) Muestre que el número de maneras de ordenar estas n unidades viene dado por

$$\frac{n!}{n_1!n_2!\ldots n_k!}$$

Demostración. Para ubicar las n unidades, situamos en primer lugar las n_1 unidades del tipo A_1. Para esto, únicamente hay que elegir el lugar en que van a situarse estas unidades y eso puede hacerse de $C_{n_1}^n$ maneras.
Razonando de esta forma, el número total de ordenaciones posibles es:

$$C_{n_1}^n \times C_{n_2}^{n-n_1} \times C_{n_3}^{n-n_1-n_2} \times \ldots \times C_{n_k}^{n-n_1-n_2-\ldots-n_k} \tag{1}$$

donde $n = n_1 + n_2 + \ldots + n_k$. Desarrollando (1) se obtiene:

$$\frac{n!}{n_1!(n-n_1)!} \times \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \times \frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \times \ldots \times \frac{(n-n_1-n_2-\ldots-n_k)!}{n_k!(n-n)!}$$
$$= \frac{n!}{n_1!n_2!\ldots n_k!} = PR_n^{n_1,n_2,\ldots,n_k} \tag{2}$$

□

b) Para una entrevista de trabajo se han presentado 12 estadísticos, 3 de la UNI, 4 de San Marcos, 3 de la Agraria y 2 de la PUCP. Si ellos son llamados para la entrevista en un orden aleatorio ¿con qué probabilidad los tres primeros entrevistados serán de San Marcos?

Demostración. Método 1:

Número de estadísticos en la población $- \rightarrow$ $N = 12$
Número de estadísticos en la población que pertenecen a la San Marcos $- - \rightarrow$ $M = 4$
Tamaño de la muestra $- \rightarrow$ $n = 3$

Definimos la v.a.:

$$X = \text{Número de estadísticos de la San Marcos en la muestra.}$$

entonces $X \sim H(N, M, n)$. El problema nos pide $P(X = 3)$.

$$P(X = x) = \frac{C_x^M C_{n-x}^{N-M}}{C_n^N} \quad \Rightarrow \quad P(X = 3) = \frac{C_3^4 C_{3-3}^{12-4}}{C_3^{12}} = 0.018182 \tag{3}$$

Método 2:
Podemos hacer uso del item a), definiendo la siguiente v.a.:

$Y =$ Número de estadísticos de la San Marcos que fueron los primeros entrevistados en la muestra.

Siendo las unidades de cada tipo indistinguibles entonces se presenta un total de ordenaciones:

$$PR_{12}^{3,4,3,2} = \frac{12!}{3!4!3!2!} \tag{4}$$

Sin embargo el problema nos pide $P(Y=3)$, entonces

$$P(Y=y) \quad = \quad \begin{cases} \dfrac{PR_{12-y}^{3,4-y,3,2}}{PR_{12}^{3,4,3,2}} & y \leq \text{mín}\{4,n\} \\ 0 & \text{en otro caso.} \end{cases} \tag{5}$$

Esto es debido a que si es y la cantidad de estadísticos de San Marcos que fueron primeros entrevistados entonces las ordenaciones van restringidas a los $12-y$ estadísticos restantes es decir a 3 de UNI, $4-y$ de San Marcos, 3 de Agraria y 2 de PUCP. Por otro lado, la condición $y < \text{mín}\{4,y\}$ es a causa de que la cantidad y es parte de la muestra ($y<n$) además que son de la San Marcos ($y<4$). Luego,

$$P(Y=3) \quad = \quad \frac{PR_{9}^{3,1,3,2}}{PR_{12}^{3,4,3,2}} = \frac{\dfrac{9!}{3!1!3!2!}}{\dfrac{12!}{3!4!3!2!}} = 0.018182 \tag{6}$$

El resultado es el mismo que en (3). □

Ejercicio 3

Considere una pequeña población conformada por 6 personas, a las que se les ha medido su nivel de hemoglobina en gramos por decilitro, encontrándose:

$$13.9, 11.5, 16.7, 14.4, 14.6, 15.1$$

Mediante un MASc y un MASCs de tamaño $n=3$:

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

Demostración.

Listing 1: Definiendo la población.

```
library(prob)
options(digits=3)
ypop= c(13.9,11.5,16.7,14.4,14.6,15.1)
```

Listing 2: MASs.

```
samplesMASs=urnsamples(ypop,3)
ybars=apply(samplesMASs,1,mean)
S2s=apply(samplesMASs,1,var)
Probs=rep(1/length(ybars),length(ybars))
bsamplesMASs=cbind(samplesMASs,ybars,S2s,Probs)
pp1s=aggregate(bsamplesMASs[,6],by=list(bsamplesMASs[,4]),sum)
colnames(pp1s)=c("Media","Probabilidad")
pp1s[pp1s$Media>14,]
pp1s[pp1s$Media>14,"Probabilidad"]
sum(pp1s[pp1s$Media>14,"Probabilidad"])
```

Resultado: 0.70.

Listing 3: MASc.

```
samplesMASc=urnsamples(ypop,3,replace=TRUE,ordered=TRUE)
ybarc=apply(samplesMASc,1,mean)
S2c=apply(samplesMASc,1,var)
Probc=rep(1/length(ybarc),length(ybarc))
bsamplesMASc=cbind(samplesMASc,ybarc,S2c,Probc)
pp1c=aggregate(bsamplesMASc[,6],by=list(bsamplesMASc[,4]),sum)
colnames(pp1c)=c("Media","Probabilidad")
pp1c[pp1c$Media>14,]
pp1c[pp1c$Media>14,"Probabilidad"]
sum(pp1c[pp1c$Media>14,"Probabilidad"])
```

Resultado: 0.699. □

b) Halle la varianza de la media anterior y compruebe que se cumple la proposición 2.2.

Demostración.

Listing 4: Media y varianza.

```
>mean(ypop)
[1] 14.4
>var(ypop)
[1] 2.89
```

Listing 5: MASs.

```
samplesMASs=urnsamples(ypop,3)
sum(((pp1s[,1]-sum(pp1s[,1]*pp1s[,2]))^2)*pp1s[,2])
```

Resultado: 0.482.
Comprobación: Varianza (Media Muestral) $= (1 - 3/6) \times 2.89/6 = 0.482$ L.q.q.d.

Listing 6: MASc.

```
samplesMASc=urnsamples(ypop,3,replace=TRUE,ordered=TRUE)
sum(((pp1c[,1]-sum(pp1c[,1]*pp1c[,2]))^2)*pp1c[,2])
```

Resultado: 0.804.
Comprobación: Varianza (Media Muestral) $= 2.89 \times (5/6) \times (1/3) = 0.804$ L.q.q.d. □

c) Suponga que para estimar la media del nivel de hemoglobina en estos 6 pacientes se propusiera la mediana de los valores observados en la muestra ¿sería este un estimador insesgado? ¿tiene este una menor varianza que la media muestral?

Demostración.

Listing 7: MASc.

```
> ybarsm=apply(samplesMASs,1,median)
> Probsm=rep(1/length(ybarsm),length(ybarsm))
> bsamplesMASsm=cbind(samplesMASs,ybarsm,S2s,Probsm)
> pp1sm=aggregate(bsamplesMASsm[,6],by=list(bsamplesMASsm[,4]),sum)
> colnames(pp1sm)=c("Mediana","Probabilidad")
> sum(pp1s[,1]*pp1s[,2])
[1] 14.4
> sum(pp1sm[,1]*pp1sm[,2])
[1] 14.5
```

Esto nos indica que la Mediana Muestral es un estimador insesgado de μ.

Listing 8: MASs.

```
> sum(((pp1sm[,1]-sum(pp1sm[,1]*pp1sm[,2]))^2)*pp1sm[,2])
[1] 0.15
```

Además tiene una menor varianza que la media muestral: $0.15 < 0.482$

Listing 9: MASc.

```
> ybarcm=apply(samplesMASc,1,median)
> Probcm=rep(1/length(ybarcm),length(ybarcm))
> bsamplesMAScm=cbind(samplesMASc,ybarcm,S2c,Probcm)
> pp1cm=aggregate(bsamplesMAScm[,6],by=list(bsamplesMAScm[,4]),sum)
> colnames(pp1cm)=c("Mediana","Probabilidad")
> sum(pp1c[,1]*pp1c[,2])
[1] 14.4
> sum(pp1cm[,1]*pp1cm[,2])
[1] 14.4
```

Esto nos indica que la Mediana Muestral es un estimador insesgado de μ.

Listing 10: MASc.

```
> sum(((pp1cm[,1]-sum(pp1cm[,1]*pp1cm[,2]))^2)*pp1cm[,2])
[1] 1.16
```

Además tiene una mayor varianza que la media muestral: $1.16 > 0.804$ □

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

Demostración.

Listing 11: Selección con MASc

```
sapply(r,function(x){
        p <- 1/6*1:6
        min(which(p>=x))
})
```

Resultado: $1, 2, 4$.
Las observaciones elegidas son: 13.9, 11.5, 14.4 generándose una media muestral de 13.26667.

Listing 12: Selección con MASs

```
for(i in 1:3){
        p <- (1/(7-i))*(1:(7-i))
        print(min(which(p>=r[i])))
}
```

Resultado: $1, 2, 3$.
Como el orden cambia al retirar observaciones, esto corresponde a los números de orden 1, 3 y 5. Las observaciones elegidas son: $13.9, 16.7, 14.6$ obteniendo una media muestral de 15.06667. □

Ejercicio 11

Suponga que en una zona rural de 3000 viviendas se ha planifiado una encuesta por muestreo tomándose un MASs de 100 viviendas. Un interés de la encuesta es estimar el consumo total mensual de agua para los hogares que cuentan con servicio de agua y desague.

a) Tomada la muestra, proponga una estimación de este total. Asuma, como es natural, que antes de tomarse la muestra no es posible identificar de antemano si un hogar de la zona posee o no servicio de agua y desague, pero que si conocemos cuantas vivivendas en la zona tienen este servicio.

Demostración. Sea y la variable estadística que representa el consumo mensual de cada familia. Teniendo una población de $N = 3000$ viviendas, obtendremos N valores $y_1, y_2, \ldots, y_N$ para y:

$$\mathcal{P} = \{y_1, y_2, \ldots, y_N\} \tag{7}$$

generándose un consumo total mensual como sigue:

$$\tau = \sum_{i=1}^{N} y_i = N\left(\frac{\sum_{i=1}^{N} y_i}{N}\right) = N\mu \tag{8}$$

donde μ es la media poblacional.
Además sabemos que un estimador de μ es la media muestral $\bar{Y}$. Veamos lo siguiente:

$$E[N\bar{Y}] = NE[\bar{Y}] = N\mu = \tau \tag{9}$$

Es decir $\hat{\tau} = N\bar{Y}$ es un estimador de τ. □

b) De manera general, dada una población de tamaño N y una MASs en ella de tamaño n muestre que para una variable estadística y y cierto subconjunto de esta población (dominio d) cuyas membresías se desconocen, $\hat{\tau}_d = \frac{N}{n}\sum_{i=1}^{n} Y_i\delta_{di}$, donde Y_i es el valor de y en la i-ésima vivienda seleccionada y δ_{di} es una variable indicadora que vale 1 si, y solamente si, la i-ésima vivienda seleccionada pertenece al dominio d, es un estimador insesgado del total τ_d de y para este dominio.

Demostración. Sean $u_1, u_2, \ldots, u_{N_d}$ los N_d elementos del subconjunto de $\mathcal{P}$ de dominio d, llamemos $\mathcal{P}_d$ a este subconjunto:

$$\mathcal{P}_d = \{u_1, u_2, \ldots, u_{N_d}\} \tag{10}$$

Observamos entonces que el consumo total mensual en $\mathcal{P}_d$, vendría determinado como:

$$\tau_d = \sum_{i=1}^{N_d} u_i \tag{11}$$

El objetivo es determinar que $E[\hat{\tau}_d] = \tau_d$, para concluir que $\hat{\tau}_d$ es un estimador insesgado de τ_d. Veamos entonces:

$$\begin{aligned} E[\hat{\tau}_d] &= E\left[\frac{N}{n}\sum_{i=1}^{n} Y_i\delta_{di}\right] \\ &= \frac{N}{n}\sum_{i=1}^{n} E\left[Y_i\delta_{di}\right] \end{aligned} \tag{12}$$

Por definición, sabemos que

$$\delta_{di} = \begin{cases} 1 & \text{si } Y_i \in \mathcal{P}_d \\ 0 & \text{si } Y_i \notin \mathcal{P}_d \end{cases} \quad \text{ó} \quad \delta_{di} = \begin{cases} 1 & \text{si } Y_i \in \{u_1, \ldots, u_{N_d}\} \\ 0 & \text{si } Y_i \notin \mathcal{P}_d \end{cases} \tag{13}$$

Analicemos el siguiente componente de (12):

$$\begin{aligned}
E\left[Y_i \delta_{di}\right] &= \sum_{\substack{y_k \in \mathcal{P} \\ l \in \{0,1\}}} y_k l P(Y_i = y_k, \delta_{di} = l) \\
&= \sum_{\substack{y_k \in \mathcal{P} \\ l \in \{0,1\}}} y_k l P(Y_i = y_k | \delta_{di} = l) P(\delta_{di} = l) \\
&= \underbrace{\sum_{y_k \in \mathcal{P}} y_k(0) P(Y_i = y_k | \delta_{di} = 0) P(\delta_{di} = 0)}_{=0} + \sum_{y_k \in \mathcal{P}} y_k P(Y_i = y_k | \delta_{di} = 1) P(\delta_{di} = 1) \\
&= \sum_{y_k \in \mathcal{P}_d} y_k P(Y_i = y_k | \delta_{di} = 1) P(\delta_{di} = 1) + \sum_{y_k \notin \mathcal{P}_d} y_k \underbrace{P(Y_i = y_k | \delta_{di} = 1)}_{=0} P(\delta_{di} = 1) \\
&= \sum_{j=1}^{N_d} u_j P(Y_i = u_j | \delta_{di} = 1) P(\delta_{di} = 1) \\
&= \sum_{j=1}^{N_d} u_j \left(\frac{1}{N_d}\right) \left(\frac{N_d}{N}\right) \\
&= \frac{1}{N} \sum_{j=1}^{N_d} u_j
\end{aligned} \tag{14}$$

Reemplazando (14) en (12) tenemos lo siguiente

$$E[\hat{\tau}_d] = \frac{N}{n} \sum_{i=1}^{n} \frac{1}{N} \sum_{j=1}^{N_d} u_j = \frac{N}{n} \frac{n}{N} \sum_{j=1}^{N_d} u_j = \sum_{j=1}^{N_d} u_j = \tau_d \tag{15}$$

Observamos que este resultado representa una generalización para estimar el total de una subpoblación de $\mathcal{P}$ o bien de $\mathcal{P}$. Lo primero es lo demostrado, lo segundo es fácil de ver usando como dominio d el de la población $\mathcal{P}$, teniéndose $N_d = N$, Por lo que $\delta_{di} = 1$ para todo $i = 1, \ldots, N$, es decir:

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} Y_i(1) = N\hat{Y}.$$

siendo este el resultado obtenido en a). □

c) Muestre que la varianza de y en toda la población, σ^2, y la de y en el dominio, σ_d^2, satisfacen aproximadamente la relación

$$\sigma^2 = p_d(\sigma_d^2 + q_d \mu_d^2) \tag{16}$$

siendo μ_d la media de y en el dominio d, $p_d = \dfrac{N_d}{N}$ la proporción de elementos del dominio d en la población y $q_d = 1 - p_d$.

Demostración. Analizemos el siguiente resultado:

$$\begin{aligned} p_d(\sigma_d^2 + q_d\mu_d^2) &= p_d(\sigma_d^2 + (1-p_d)\mu_d^2) \\ &= p_d(\sigma_d^2 + \mu_d^2) - p_d\mu_d^2 \\ &= p_d\left(\frac{1}{N_d}\sum_{j=1}^{N_d} u_k^2\right) - p_d\left(\frac{1}{N_d}\sum_{j=1}^{N_d} u_j\right)^2 \\ &= \frac{N_d}{N}\left(\frac{1}{N_d}\sum_{j=1}^{N_d} u_k^2\right) - \left(\frac{N_d}{N}\right)^2\left(\frac{1}{N_d}\sum_{j=1}^{N_d} u_j\right)^2 \\ &= \frac{1}{N}\sum_{j=1}^{N_d} u_k^2 - \left(\frac{1}{N}\sum_{j=1}^{N_d} u_j\right)^2 \end{aligned} \tag{17}$$

De lo anterior, observamos lo siguiente:

$$\text{Si } N_d \longrightarrow N \quad \Rightarrow \quad p_d(\sigma_d^2 + q_d\mu_d^2) \longrightarrow \underbrace{\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2}_{\sigma^2} \tag{18}$$

Por tanto, si N_d es cercano a N, entonces se cumple que aproximadamente

$$\sigma^2 = p_d(\sigma_d^2 + q_d\mu_d^2) \tag{19}$$

concluyendo la prueba. □

d) Muestre, usando c), que si se desea estimar τ_d con un máximo error de estimación e y una confianza del $100(1-\alpha)\,\%$, el tamaño de muestra apropiado viene dado por

$$n = \frac{\left(\sigma_d^2 + q_d\mu_d^2\right) N z_{1-\frac{\alpha}{2}}^2}{Ne^2 p_d + \left(\sigma_d + q_d\mu_d\right)\mu_d z_{1-\frac{\alpha}{2}}^2}$$

Demostración. En teoría tenemos lo siguiente:

$$P(|\hat{\tau}_d - \tau_d| \leq e) = 1 - \alpha \tag{20}$$

donde e es el error máximo de estimación, siendo a su vez expresado como sigue

$$e \;=\; z_{1-\frac{\alpha}{2}}^2 SE \tag{21}$$

El primer objetivo será determinar SE (error estándar de estimación de $\hat{\tau}_d$), sabiendo que

$SE^2 \approx Var(\hat{\tau}_d)$. Veamos:

$$\begin{aligned}
Var(\hat{\tau}_d) &= E\left[\left(\frac{N}{n}\sum_{i=1}^{n} Y_i\delta_{di}\right)\right]^2 - E[\hat{\tau}_d]^2 \\
&= \frac{N^2}{N_d^2}\underbrace{Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\delta_{di}\right)}_{\approx Var(\bar{Y})} + \frac{\sigma_d(\mu_d-\sigma_d)}{Np_d} \\
&\approx p_d^{-1}\frac{\sigma^2_{N-1}}{n}\left(1-\frac{n}{N}\right) + \underbrace{\left(\frac{\sigma^2}{p_d(\sigma_d^2+q_d\mu_d^2)}\right)}_{\text{por (19)}}\frac{\sigma_d(\mu_d-\sigma_d)}{Np_d} \\
&\approx p_d^{-2}\frac{\sigma^2}{n}\left(1-\frac{n}{N}\right) + p_d^{-2}\frac{\sigma^2\sigma_d(\mu_d-\sigma_d)}{N(\sigma_d^2+q_d\mu_d^2)} \\
&= p_d^{-1}\frac{\sigma^2}{n}\left(1-\frac{n}{N}\left(1+\frac{\sigma_d(\mu_d-\sigma_d)}{(\sigma_d^2+q_d\mu_d^2)}\right)\right) = SE^2
\end{aligned} \tag{22}$$

Por tanto, el error máximo de estimación sería:

$$e = z^2_{1-\frac{\alpha}{2}}\underbrace{p_d^{-1}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}\left(1+\frac{\sigma_d(u_d-\sigma_d)}{\sigma_d^2+q_d\mu_d^2}\right)}}_{SE} \tag{23}$$

Determinemos el tamaño n de la muestra que necesitamos. Primero tomemos

$$\gamma^2 = p_d(\sigma_d+q_d\mu_d)\mu_d \quad \Rightarrow \quad \frac{\gamma^2}{\sigma^2} = \frac{p_d(\sigma_d+q_d\mu_d)\mu_d}{p_d(\sigma_d^2+q_d\mu_d^2)} = 1+\frac{\sigma_d(u_d-\sigma_d)}{\sigma_d^2+q_d\mu_d^2} \tag{24}$$

Reemplazemos (24) en (23), y reformulemos

$$p_de = z^2_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n\left(\frac{\gamma^2}{\sigma^2}\right)}{N}} = z^2_{1-\frac{\alpha}{2}}\frac{\gamma}{\sqrt{n\left(\frac{\gamma^2}{\sigma^2}\right)}}\sqrt{1-\frac{n\left(\frac{\gamma^2}{\sigma^2}\right)}{N}} \tag{25}$$

Lo anterior nos lleva a obtener la siguiente relación:

$$\begin{aligned}
\Rightarrow n\left(\frac{\gamma^2}{\sigma^2}\right) &= \frac{\gamma^2 N z^2_{1-\frac{\alpha}{2}}}{N(p_de)^2+\gamma^2 z^2_{1-\frac{\alpha}{2}}} \\
n &= \frac{\sigma^2 N z^2_{1-\frac{\alpha}{2}}}{N(p_de)^2+p_d(\sigma_d+q_d\mu_d)\mu_d z^2_{1-\frac{\alpha}{2}}} \\
n &= \frac{\left(\frac{\sigma^2}{p_d}\right)N z^2_{1-\frac{\alpha}{2}}}{Ne^2p_d+(\sigma_d+q_d\mu_d)\mu_d z^2_{1-\frac{\alpha}{2}}} \\
n &= \frac{\left(\sigma_d^2+q_d\mu_d^2\right)N z^2_{1-\frac{\alpha}{2}}}{Ne^2p_d+(\sigma_d+q_d\mu_d)\mu_d z^2_{1-\frac{\alpha}{2}}}
\end{aligned} \tag{26}$$

□

e) Obtenga haciendo las estimaciones necesarias, el tamaño de muestra que se necesitaría en una encuesta futura para a), si es que se deseara estimar τ_d con un margen de error no mayor a los 200 litros con una confianza del 95 %. Para ello suponga que para la encuesta tomada en a) se encontró que 60 hogares contaban con servicios de agua y desague y en promedio ellos consumieron en el mes 5,100 litros con una desviación estándar de 380 litros.

Demostración. Se desea hallar n para estimar τ_d, usando los siguientes datos:

$$\begin{aligned} N &= 3000 && (27)\\ e &= 200 && (28)\\ \alpha &= 5\,\% && (29)\\ M &= 60 \text{ :(Número de viviendas que cuentan con servicio de agua y desague)} && (30)\\ \mu_M &= 5100 \text{ :(Promedio de consumo respecto a las viviendas que cuentan con servicio de agua y desague)} && (31)\\ \sigma_M &= 380 \text{ :(Desviación estándar respecto a las viviendas que cuentan con servicio de agua y desague)} && (32) \end{aligned}$$

Analíticamente: Sean w_i para $i = 1, \ldots, M$ los M valores de las viviendas que cuentan con servicio. Sin pérdida general, tomemos $y_i = w_i$ para todo $i = 1, \ldots, M$, e $y_i = 0$ para $i = M+1, \ldots, N$. Veamos lo siguiente

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_{i=1}^{N} y_i^2 - \mu^2 = \frac{1}{N}\sum_{i=1}^{M} w_m^2 - \mu^2 = \frac{1}{N}\left(M(\sigma_M^2 + \mu_M^2)\right) - \left(\frac{M\mu_M}{N}\right)^2 \\ &= \frac{1}{3000}\left(60(380^2 + 5100^2)\right) - \left(\frac{60 \times 5100}{3000}\right)^2 = 512.684 \end{aligned} \tag{33}$$

Usando la regla conservadora, tomemos $p_d = 0.5$, y por conveniencia tomemos $0 \leq \mu_d$. Tenemos la siguiente relación en (26):

$$\underbrace{\frac{\left(\sigma_d^2 + q_d\mu_d^2\right) N z_{1-\frac{\alpha}{2}}^2}{Ne^2p_d + (\sigma_d + q_d\mu_d)\mu_d z_{1-\frac{\alpha}{2}}^2}}_{\text{Falta conocer } \sigma_d, \mu_d} \leq \frac{\left(\sigma_d^2 + q_d\mu_d^2\right) N z_{1-\frac{\alpha}{2}}^2}{Ne^2p_d} = \underbrace{\frac{\sigma^2 z_{1-\frac{\alpha}{2}}^2}{e^2p_d^2}}_{n \text{ (óptimo)}} \tag{34}$$

Calculemos:

$$n = \frac{512.684(1.96)^2}{(200)^2(0.5)^2} = 196.95 \approx 197 \tag{35}$$

□

Ejercicio 17

Mediante un MASs piloto de tamaño n_1 se ha calculado que el tamaño de muestra a tomarse para estimar la media de una variable y con un máximo error de estimación de e y una confianza de $100(1-\alpha)\,\%$ es n. Un colega sugiere que en vez de considerarse estas n observaciones bastaría tan solo tomarse un MASs de tamaño $n–n_1$ de la población no muestreada, pues argumenta que la muestra piloto ya recabo información de y ¿estaría usted de acuerdo con su colega?. Justifique.

Demostración. Sean N elementos en la población y sean

$$P(X_i = x_i \in \mathcal{P}_{n_1}) = \frac{n_1}{N} \tag{36}$$

$$P(X_i = x_i \in \mathcal{P}_{n-n_1}) = \frac{n-n_1}{N} \tag{37}$$

Dado que son eventos mutuamente excluyentes, entonces podemos escribirlo de la siguiente manera:

$$P(X_i \in \mathcal{P}_{n_1} + X_i \in \mathcal{P}_{n-n_1}) = \frac{n_1}{N} + \frac{n-n_1}{N} = \frac{n}{N} \tag{38}$$

Por tanto, si estaría de acuerdo, dado que la probabilidad de tomar las muestras por separado n_1 y $n-n_1$ es la misma que tomar una muestra de tamaño n de la población N. □

Ejercicio 21

En la subsección 2.4.3 obtuvimos el error estándar de estimación para la diferencia de medias del índice de rendimiento api para los años 1999 y 2000.

a) Tome en esta base de datos un MASs de tamaño n = 100 y estime con la libreria survey la diferencia de medias del índice api para estos años.

Demostración. Sean

X : índice de rendimiento API del año 2000; Y : índice de rendimiento API del año 1999.

Listing 13: Diferencia de medias usando librería survey.

```
library(survey)
data(api)

set.seed(100)
N = dim(apipop)[1]
n = 100
index = sample(N,n)
sample1 = apipop[index,]

aux = data.frame(fpc=rep(N,n), pw=rep(N/n,n))
sample1 = cbind(sample1,aux)
disenoMASs = svydesign(id=~1, fpc=~fpc, data=sample1)

means = svymean(~api00+api99, disenoMASs)
mean_dif1 = svycontrast(means, c(api00=1,api99=-1))
mean_dif1
```

Resultado: $\bar{D} = \bar{X} - \bar{Y} = 29.32$. □

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

Demostración.

Listing 14: Intervalo de confianza para diferencia de medias, usando la librería survey.

```
confint(mean_dif1, level=0.95)
```

Resultado: $\bar{D} = \bar{X} - \bar{Y} \in \langle 23.40439; 35.23561 \rangle$. □

c) Con la misma muestra tomada en a) obtenga el IC en b) pero ahora sin usar el paquete survey.

Demostración.

Listing 15: Intervalo de confianza para diferencia de medias sin usar la librería survey.

```
dif2 = sample1[,12]-sample1[,13]
mean_dif2 = mean(dif2)
sd_dif2 = sqrt((1-n/N)*(1/n)*var(dif2))

z = qnorm(0.975,0,1)
lim_inf1 = mean_dif2-z*sd_dif2
lim_sup1 = mean_dif2+z*sd_dif2
lim_inf1
lim_sup1
```

Resultado: $\bar{D} = \bar{X} - \bar{Y} \in \langle 23.40439; 35.23561 \rangle$. □

Ejercicio 22

Suponga que para un MASs de tamaño n sobre una población de tamaño N se tiene interés en estudiar dos variables estadísticas x e y.

a) Muestre que la covarianza entre las medias muestrales de estas variables viene dada por:

$$\mathrm{Cov}(\bar{X}, \bar{Y}) = \left(1 - \frac{n}{N}\right) \frac{\sigma_{xy}}{n} \quad \text{donde } \sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)$$

es la covarianza poblacional entre x e y y μ_x y μ_y las medias poblacionales de x e y, respectivamente.

Demostración. Denotemos $S = \sum_{i=1}^{N} x_i y_i$. Comenzamos reescribiendo las expresiones de interés:

$$\mathrm{Cov}(\bar{X}, \bar{Y}) = \mathrm{E}\left[\left(\bar{X} - \mathrm{E}[\bar{X}]\right)\left(\bar{Y} - \mathrm{E}[\bar{Y}]\right)\right] = \mathrm{E}\left[\bar{X}\bar{Y}\right] - \mu_x \mu_y, \tag{39}$$

$$\sigma_{xy} = \frac{1}{N-1}\left(S - N\mu_x\mu_y\right) \tag{40}$$

Expandimos el producto de las medias muestrales:

$$\begin{aligned}
\mathrm{E}\left[\bar{X}\bar{Y}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{N} x_i\delta_i \frac{1}{n}\sum_{i=1}^{N} y_i\delta_i\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\left(x_1\delta_1 + x_2\delta_2 + \ldots + x_N\delta_N\right)\left(y_1\delta_1 + y_2\delta_2 + \ldots + y_N\delta_N\right)\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{N} x_i y_i \delta_i^2 + \sum_{j\neq 1}^{N} x_1\delta_1 y_j\delta_j + \sum_{j\neq 2}^{N} x_2\delta_2 y_j\delta_j + \ldots + \sum_{j\neq N}^{N} x_N\delta_N y_j\delta_j\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{N} x_i y_i \delta_i^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i\delta_i y_j\delta_j\right] \\
&= \frac{1}{n^2}\mathrm{E}\left[\sum_{i=1}^{N} X_i Y_i + \sum_{i=1}^{N}\sum_{j\neq i}^{N} x_i\delta_i y_j\delta_j\right] \qquad (41)
\end{aligned}$$

Calculamos la esperanza para los dos tipos de términos en (41):

$$\mathrm{E}\left[\sum_{i=1}^{N} X_i Y_i\right] = \mathrm{E}\left[\sum_{i=1}^{N} x_i y_i \delta_i^2\right] = \sum_{i=1}^{N} x_i y_i \mathrm{E}\left[\delta_i^2\right] = \frac{n}{N}\sum_{i=1}^{N} x_i y_i = \frac{n}{N}S \qquad (42)$$

$$\mathrm{E}\left[\sum_{j\neq i}^{N} x_i y_j \delta_i\delta_j\right] = \sum_{j\neq i}^{N} x_i y_j \mathrm{E}\left[\delta_i\delta_j\right] = \frac{n}{N}\left(\frac{n-1}{N-1}\right) x_i \sum_{j\neq i}^{N} y_j = \frac{n}{N}\left(\frac{n-1}{N-1}\right) x_i\left(N\mu_y - y_i\right) \qquad (43)$$

Reemplazamos (42) y (43) en (41):

$$\begin{aligned}
\mathrm{E}\left[\bar{X}\bar{Y}\right] &= \frac{1}{n^2}\left(\frac{n}{N}S + \frac{n}{N}\left(\frac{n-1}{N-1}\right)\sum_{i=1}^{N} x_i\left(N\mu_y - y_i\right)\right) \\
&= \frac{1}{nN}\left(S + \left(\frac{n-1}{N-1}\right)\left(N^2\mu_x\mu_y - S\right)\right) \\
&= \left(\frac{N-n}{nN(N-1)}\right)S + \left(\frac{N(n-1)}{n(N-1)}\right)\mu_x\mu_y \qquad (44)
\end{aligned}$$

Reemplazamos (44) en (39):

$$\begin{aligned}
\mathrm{Cov}(\bar{X},\bar{Y}) &= \left(\frac{N-n}{nN(N-1)}\right)S + \left(\frac{N(n-1)}{n(N-1)}\right)\mu_x\mu_y - \mu_x\mu_y \\
&= \left(\frac{N-n}{nN(N-1)}\right)S - \left(\frac{N-n}{n(N-1)}\right)\mu_x\mu_y \\
&= \left(\frac{N-n}{N}\right)\left(\frac{1}{n}\right)\underbrace{\left(\frac{1}{N-1}\right)\left(S - N\mu_x\mu_y\right)}_{\sigma_{xy}} \\
&= \left(1-\frac{n}{N}\right)\frac{\sigma_{xy}}{n} \qquad (45)
\end{aligned}$$

□

b) Proponga algún estimador insesgado para esta covarianza.

Demostración. Nos situamos en 3 casos:

- **Medias μ_x, μ_y conocidas:**
  Usando (42), podemos efectuar un reemplazo directo en la fórmula (40) de la covarianza poblacional:

$$\begin{aligned}\sigma_{xy} = \frac{1}{N-1}\left(S - N\mu_x\mu_y\right) &= \frac{1}{N-1}\left(\frac{N}{n}E\left[\sum_{i=1}^{n} X_iY_i\right] - N\mu_x\mu_y\right) \\ &= \mathrm{E}\left[\underbrace{\frac{1}{N-1}\left(\frac{N}{n}\sum_{i=1}^{n} X_iY_i - N\mu_x\mu_y\right)}_{\text{estimador}}\right] \qquad (46)\end{aligned}$$

- **Solo una media conocida:**
  Sin pérdida de generalidad, supongamos μ_y conocida y μ_x desconocida. Dado que la media muestral $\bar{X}$ es un estimador de la media poblacional μ_x, es decir $E[\bar{X}] = \mu_x$, además de ser insesgado, entonces podemos sustituir la media desconocida en (46):

$$\begin{aligned}\sigma_{xy} = \frac{1}{N-1}\left(S - N\mu_x\mu_y\right) &= \frac{1}{N-1}\left(\frac{N}{n}E\left[\sum_{i=1}^{n} X_iY_i\right] - NE[\bar{X}]\mu_y\right) \\ &= \mathrm{E}\left[\underbrace{\frac{1}{N-1}\left(\frac{N}{n}\sum_{i=1}^{n} X_iY_i - N\bar{X}\mu_y\right)}_{\text{estimador}}\right] \qquad (47)\end{aligned}$$

- **Medias μ_x, μ_y desconocidas:**
  Primero reformulamos la expresión (44) y adicionamos un factor:

$$\left[\ \frac{n-N}{N(n-1)}S - N\mu_x\mu_y = \frac{(1-N)n}{n-1}\mathrm{E}\left[\bar{X}\bar{Y}\right]\ \right] + \underbrace{\frac{(N-1)n}{(n-1)N}S}_{\text{factor}}$$

$$\Rightarrow \quad S - N\mu_x\mu_y = \frac{(1-N)n}{n-1}\mathrm{E}\left[\bar{X}\bar{Y}\right] + \frac{(N-1)}{(n-1)}\frac{n}{N}S \qquad (48)$$

  Usando (48), expresamos la covarianza poblacional (40) como sigue:

$$\begin{aligned}\sigma_{xy} &= \frac{1}{N-1}\left(\frac{(1-N)n}{n-1}\mathrm{E}\left[\bar{X}\bar{Y}\right] + \frac{(N-1)}{(n-1)}\frac{n}{N}S\right) \\ &= \frac{1}{n-1}\left(\frac{n}{N}S - n\mathrm{E}\left[\bar{X}\bar{Y}\right]\right) \\ &= \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n} X_iY_i\right] - n\mathrm{E}\left[\bar{X}\bar{Y}\right]\right) \\ &= \mathrm{E}\left[\underbrace{\frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}\right)}_{\text{estimador}}\right] \qquad (49)\end{aligned}$$

□

Ejercicio 27

En una investigación para estudiar la relación entre la propensión al consumo de alcohol por parte de adolescentes varones y variables como el control parental, regulación emocional y madurez social, se desea tomar un MASs para sólo el distrito de San Miguel. Puesto que la propensión se medirá mediante una proporción, es de interés estimar esta proporción con un margen de error no mayor a 0.07 y un nivel de confianza del 95 %. Usando en lo posible el paquete survey de R.

a) Halle el tamaño de muestra requerido para este estudio. Para esto y para crear su marco muestral puede hacer uso de la página web del Ministerio de Educación

http://escale.minedu.gob.pe/web/inicio/padron-de-iiee

la cual contiene información de todos los colegios del país en base al censo nacional escolar del 2016.

Demostración. Usamos la fórmula de Hájek para estimar intervalos de confianza en una población finita, aplicada a una proporción:

$$\text{IC} = \left[\bar{p} - z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n-1}\left(1-\frac{n}{N}\right)}, \bar{p} + z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n-1}\left(1-\frac{n}{N}\right)}\right]$$

Definiendo el error de estimación como $e = |p - \bar{p}|$, el intervalo anterior implica:

$$e = z_{1-\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}\left(1-\frac{n}{N}\right)} \to n = \frac{z^2_{1-\alpha/2}\bar{p}(1-\bar{p})N}{z^2_{1-\alpha/2}\bar{p}(1-\bar{p}) + e^2 N}$$

Dados los datos del problema: $z_{.975} \approx 1.96$, $e = 0.07$. Desconocemos la proporción poblacional de propensión al consumo de alcohol, así que asumimos $\bar{p} = 0.5$. Para determinar N debemos examinar nuestro marco muestral, dado por la cantidad de alumnos varones cursando secundaria en IIEE del distrito de San Miguel. Dado que no contamos con el detalle del número de alumnos por género en colegios mixtos, asumimos una proporción de 0.5 para estos casos.

Listing 16: Tamaño de la muestra.

```
z <- qnorm(0.975); e <- 0.07; p <- 0.5

library(data.table)
iiee <- fread(paste0(workdir,"minedu.csv"))[,`Alumnos (2016)`:=as.double(`Alumnos
    (2016)`)]
iiee[Genero=="Mixto",`Alumnos (2016)`:=`Alumnos (2016)`*0.5]
N <- sum(iiee$`Alumnos (2016)`)

### Finalmente, aplicamos la formula a los datos
n <- ceiling((z^2)*p*(1-p)*N/((z^2)*p*(1-p)+(e^2)*N)) #Redondeamos al entero
   superior
```

Resultado: $n = 190$. □

b) Tome la muestra anterior y estime en base a ella el total de alumnos matriculados el 2016 en los colegios de varones de San Miguel, así como la proporción de estudiantes que pertenecen a un colegio de gestión privada. En ambos casos obtenga el error de estimación estimado de los estimadores correspondientes.

Demostración. Bajo los supuestos del ejercicio anterior, generamos una muestra de la población considerando las variables requeridas.

Listing 17: Muestra.

```
iiee[,`Gestion / Dependencia`:=gsub(pattern = " - .*","",iiee$`Gestion /
    Dependencia`)]
poblacion <- iiee[,.(alumnos=sum(`Alumnos (2016)`)),.(Genero,`Gestion /
    Dependencia`)]

set.seed(2017)
muestra <- poblacion[rep(1:4,times=poblacion$alumnos),1:2,with=FALSE][sample(N, n
    )]
```

	Privada	Pública
Mixto	125	43
Varones	3	19

Con las observaciones obtenidas, calculamos los estimadores puntuales y de intervalo:

Listing 18: Estimadores puntuales y de intervalo.

```
#Proporciones:
p_varon <- nrow(muestra[Genero=="Varones"])/n
p_priva <- nrow(muestra[`Gestion / Dependencia`=="Privada"])/n

#ICs:
e_varon <- z*sqrt((p_varon*(1-p_varon)/(n-1))*(1-(n/N)))
e_priva <- z*sqrt((p_priva*(1-p_priva)/(n-1))*(1-(n/N)))

#Estimadores para el total poblacional en colegios de varones:
N_varon <- N*p_varon
Nivaron <- N*(p_varon - e_varon)
Nsvaron <- N*(p_varon + e_varon)
```

Total de alumnos en colegios de varones: 718.71 IC95: [439.93 - 997.48]
Proporción de alumnos en colegios privados: 0.67 IC95: [0.61 - 0.74]

Verificamos que nuestros resultados, salvo por diferencias de redondeo, concuerdan con la salida del paquete survey.

Listing 19: Estimadores puntuales y de intervalo, usando la librería survey.

```
library(survey)
muestra_svy <- svydesign(id=~1,fpc=rep(N,n),data=muestra)

svytotal(~I(Genero=="Varones"), muestra_svy)
svyciprop(~I(`Gestion / Dependencia`=="Privada"), muestra_svy)
```

	total	SE
I(Género == "Varones")FALSE	5488.29	142.24
I(Género == "Varones")TRUE	718.71	142.24

		2.5 %	97.5 %
I(`Gestion / Dependencia`=="Privada")	0.674	0.604	0.74

□

c) ¿Cree usted que el diseño MASs empleado sea apropiado para los fines de este estudio? Indique si no fuera el caso, qué dificultades acarrea este diseño.

Demostración. No. Dado que el estudio tiene como meta comparar la proporción de alumnos propensos a consumir alcohol en relación a otras características, es deseable minimizar el error de estimación para la proporción de cada subgrupo generado por los valores de las variables de interés. Ya que estos valores probablemente no tengan una distribución uniforme en la población total, es posible que un MAS dé como resultado un número de observaciones reducido para algunos subgrupos, con un error de estimación elevado como consecuencia. □