Diseños complejos

Tarea Examen 3. Tipo A

Profesor. Gonzalo Pérez.
Ayudante. Juan Andrés Cervantes.
Ayudante. José Angel Román.

Integrantes:	Franco Zarraga Daniel
	Gomez Jimenez Aaron Mauricio
	Sandoval Mendoza Jorge

1. **Comparación de dieseños de muestreo**

   Suponga que se realizará una nueva elección de diputaciones a nivel federal y le han encargado realizar el diseño de muestreo. Para esto cuenta con la información de los resultados a nivel acta (casilla) de los cómputos distritales de 2021 (https://computos2021.ine.mx/base-de-datos).

   Por simplicidad suponga que se centrará en estudiar los diseños de manera que elegirá el que sirva para realizar de la mejor forma la estimación del porcentaje de votos a favor de una coalición integrada por Morena, PT y P. Verde a nivel nacional.

   En el denominador considerará sólo el total de votos válidos, es decir, que no se considerarán los votos nulos para efectos del cálculo del porcentaje.

   Consideré los siguientes tres diseños a comparar:

   I. Se seleccionan 900 casillas de las 163,666 usando un muestreo aleatorio simple sin reemplazo.

   II. Se considera una estratificación a partir de los 300 diferentes distritos electorales del país. El diseño de muestreo considerado en cada estrato corresponde a un muestreo aleatorio simple sin reemplazo de 3 casillas.

   III. Se considera una estratificación considerando las cinco circunscripciones electorales de México. En cada estrato se usa un muestreo aleatorio simple sin reemplazo para seleccionar 6 distritos electorales y en cada distrito electoral seleccionado se usa un muestreo aleatorio simple para seleccionar 30 casillas.

   Realice la comparación de los tres diseños a partir de simulaciones. Es decir, repita 5000 veces lo siguiente. Seleccione una muestra con cada diseño y realice la estimación, en este caso deberá usar un estimador de razón.

A partir de las 5000 estimaciones estime el ECM del estimador que se obtendría para cada diseño y con estos resultados indique cuál diseño parece ser el mejor.

Nota. **Se deben sumar varias columnas de la base de datos**, aquellas que contengan información de al menos un partido que pertenezca a la coalición.

Punto extra opcional. Considerando los diseños en I) y II) calcular las varianzas poblacionales y comparar éstas con los valores obtenidos a partir de simulaciones. Comente.

Solución.

Tenemos que según la base crearemos dos variabled e interés. Total de votos para la coalición Morena, PT y P. Verde Morena, PT y P. Verde *TOTAL_MORENA_PT_PVME* (variable y) y total de votos válidos *TOTAL_VOTOS_VALIDOS* (variable z).

```
library(tidyverse)
library(psych)
library(latex2exp)
Diputaciones <- read.csv(file = 'DiputacionesExamen.csv', sep=',',
                         fileEncoding="latin1")
TOTAL_VOTOS_VALIDOS<-c()
for (i in 1:length(Diputaciones$ID_ESTADO)){
  TOTAL_VOTOS_VALIDOS[i]<-Diputaciones$TOTAL_VOTOS_CALCULADOS[i]
  -Diputaciones$VOTOS.NULOS[i]
}

TOTAL_MORENA_PT_PVME<-(Diputaciones$PT+ Diputaciones$MORENA+ Diputaciones$PVEM
                       +Diputaciones$PT.MORENA+ Diputaciones$PVEM.PT
                       + Diputaciones$PVEM.MORENA +Diputaciones$PVEM.PT.MORENA)
Datos<-as.data.frame(Diputaciones[,c(1:8)])
Datos<-cbind(Datos,TOTAL_VOTOS_VALIDOS,TOTAL_MORENA_PT_PVME)
```

Nuestro parámetro de interés será:

$$\widehat{R} = \frac{\widehat{t}_{\pi y}}{\widehat{t}_{\pi z}}$$

Nuestro parámetro de interés tiene un valor muestral real de:

```
(Porcentaje=sum(Datos$TOTAL_MORENA_PT_PVME)/sum(Datos$TOTAL_VOTOS_VALIDOS)*100)

## [1] 42.77757
```

I.- **Muestreo Aleatorio Simple:**

Definimos una semilla, número de datos totales (N) y número de datos a seleccionar (n). Se

seleccionan 900 casillas de las 163,666. Además tomamos las probabilidades de inclusión.

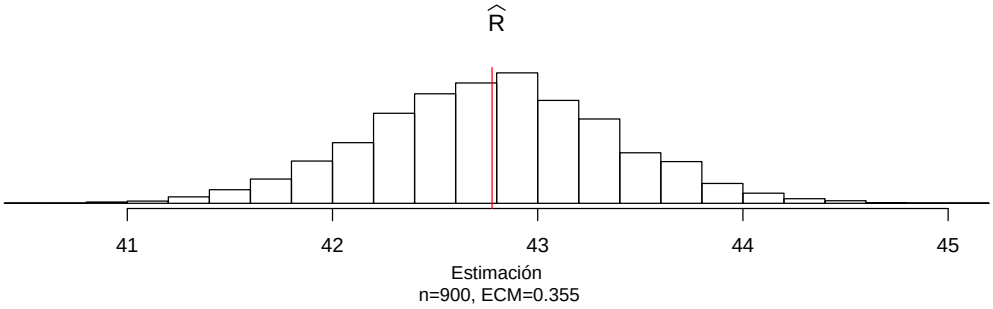
```
set.seed(1971)
(N <- nrow(Datos))

## [1] 163666

(n=900)

## [1] 900

pi_k=n/N
pi_kl=(n*(n-1))/(N*(N-1))
```

Usaremos una función básica en R para seleccionar con un m.a.s., que nos da como resultado los índices de las observaciones seleccionadas.

```
sindex=sample(1:N, n)
head(sindex,10)

##  [1] 156411   3561 104781  32952  10013  71189  90515 139251  48940 150060
```

Repetiremos la selección 5000 veces para analizar cómo se ve la distribución del estimador HT.

```
B=5000
ParMuestra = replicate(B, sample(1:N, n))
Porcentaje.HT= sapply(1:B,function(x, Indices, Pob)
{sum(Pob[Indices[,x],"TOTAL_MORENA_PT_PVME"]/pi_k)*100/
      sum(Pob[Indices[,x],"TOTAL_VOTOS_VALIDOS"]/pi_k)}
, Indices=ParMuestra, Pob=Datos)
h1=hist(Porcentaje.HT, plot=FALSE, breaks=25)
plot(h1, freq=FALSE, main=TeX("$\\widehat{R}$"),
     xlab=paste("\n Estimación \n n=",n, ", ECM=",
                round(mean((Porcentaje.HT-Porcentaje)^2), 3) ,  sep=""),
     ylab="", cex.lab=.9, yaxt='n')
abline(v=Porcentaje, col="red")
```



Obtenemos un error cuadrático medio de 0.355 %, teneindo así una estimación muy buena.

```
mean(Porcentaje.HT)

## [1] 42.7749
```

Por otro lado el promedio de las simulaciónes es de 44.7749 %. Un valor muy parecido al muestral.

Varianza

Por lo visto en clase en la expresión (69), podemos ver la varianza de nuestro estimador de razón como:

$$Var(\widehat{R}) = \sum_{k\in S}\sum_{l\in S}\Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l}$$

Además también se vieron las variables de linealización:

$$v_k = \frac{1}{t_z}\left(y_k - \frac{t_y}{t_z}z_k\right)$$

Usando R tenemos la siguiente función:

```
#Varianza poblacional del estimador HT para m.a.s
VarPob=0
# Probabilidades de inclusión
pi_k=n/N
pi_kl=(n*(n-1))/(N*(N-1))
pi_kl-pi_k**2

t_y=sum(Datos$TOTAL_MORENA_PT_PVME)
t_z=sum(Datos$TOTAL_VOTOS_VALIDOS)

Varianza=function(){
  for(k in 1:N) {
    v_k=(1/t_z*(TOTAL_MORENA_PT_PVME[k]-Porcentaje*TOTAL_VOTOS_VALIDOS[i]))
```

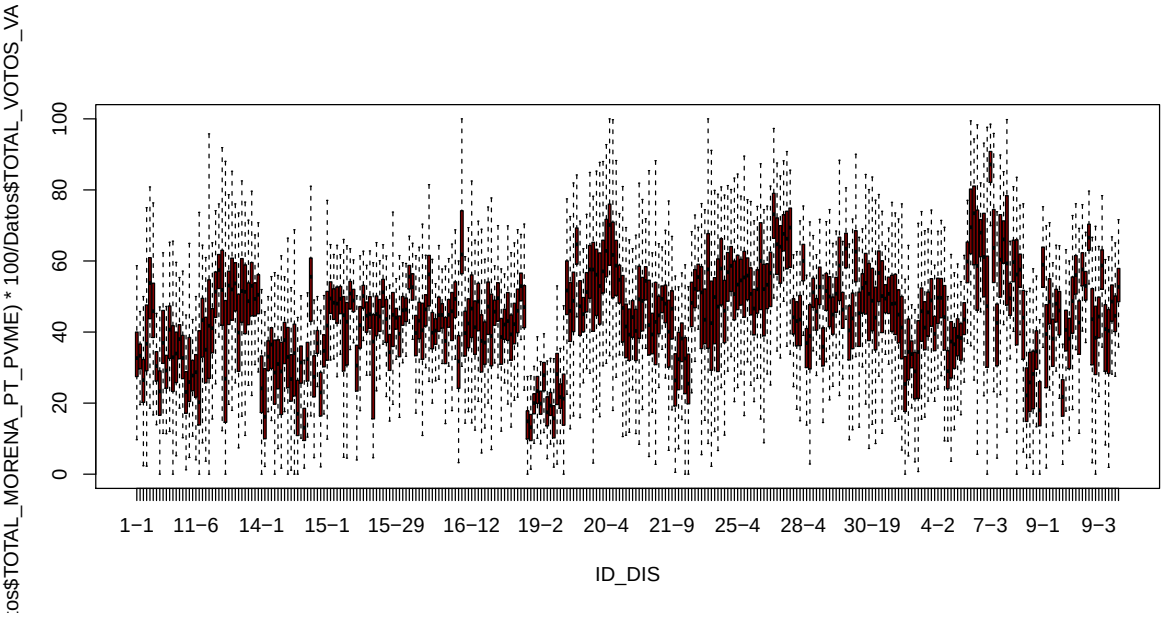
```
    for(l in 1:N){
      v_l=(1/t_z*(TOTAL_MORENA_PT_PVME[l]-Porcentaje*TOTAL_VOTOS_VALIDOS[l]))
      if(k==l){
        VarPob=VarPob + (pi_k - pi_k**2)*(v_k**2/pi_k**2)
      }else{
        VarPob=VarPob + (pi_kl-pi_k**2)*(v_k/pi_k)*(v_l/pi_k)
      }
    }
  }
}

VarPob=Varianza()
VarPob
```

II.- **Estartificación por distritos:**

Creamos un identificador para cada distrito, de tal manera que haya 300 ID´s diferentes para los 300 distritos electorales.

```
Datos<-mutate(Datos, ID_DIS=(paste(ID_ESTADO,"-",ID_DISTRITO, sep="")))
length(unique(Datos[,11]))

## [1] 300
```

```
summary(count(Datos,ID_DIS))

##     ID_DIS                n
##  Length:300         Min.   :397.0
##  Class :character   1st Qu.:499.0
##  Mode  :character   Median :541.5
##                     Mean   :545.6
##                     3rd Qu.:584.0
##                     Max.   :891.0

boxplot((Datos$TOTAL_MORENA_PT_PVME)*100/Datos$TOTAL_VOTOS_VALIDOS
     ~ ID_DIS, data = Datos, col = "darkred", outline=FALSE)
```



Creamos un identificador para cada casilla.

```
Datos=Datos[order(Datos$NOMBRE_ESTADO),]
ID<-c()
for(i in 1:length(Datos$CLAVE_CASILLA)){
  ID[i]<-i
}
Datos<-cbind(Datos, ID)
```

En total escogeremos 900 casillas, y de cada distrito escogeremos 3. Asignando del tamaño de muestra en cada ID_DIS = 3.

```
n=900
(nhCalculo=Datos %>% group_by(ID_DIS) %>% summarise(Nh2=n(), nh2=3,
           pik2=nh2/Nh2 ) )
```

```
## # A tibble: 300 x 4
##    ID_DIS   Nh2   nh2    pik2
##    <chr>  <int> <dbl>   <dbl>
##  1 1-1      537     3 0.00559
##  2 1-2      566     3 0.00530
##  3 1-3      620     3 0.00484
##  4 10-1     674     3 0.00445
##  5 10-2     655     3 0.00458
##  6 10-3     655     3 0.00458
##  7 10-4     573     3 0.00524
##  8 11-1     480     3 0.00625
```

```
##  9 11-10    584     3 0.00514
## 10 11-11    441     3 0.00680
## # ... with 290 more rows

sum(nhCalculo$nh2)

## [1] 900
```

Guardamos esta información en la base de datos (marco muestral)

```
Datos=Datos %>% group_by(ID_DIS) %>% mutate(Nh2=n(), nh2=3,
      pik2=nh2/Nh2) %>% ungroup()
```

Realizamos la selección de índices mediante una función.

```
Datosdiv <- split(Datos, list(Datos$ID_DIS))
indexstrata1=function(){
  samples <- lapply(Datosdiv, function(x) sample(x$ID, unique(x$nh2), FALSE))
  IndexS <- unlist(samples)
  return(IndexS)
}
length(indexstrata1())

## [1] 900
```

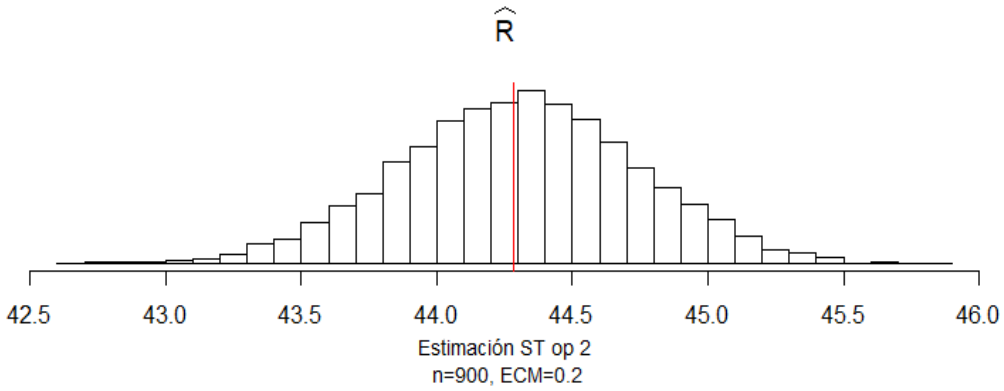
Haremos las simulaciones y ordenaremos por ID_DIS, pues la salida de las funciones son directamente ese valor.

```
set.seed(123)
ParMuestra.2 = replicate(B, indexstrata1())
Datos=Datos[order(Datos$ID),]
```

Finalmente hacemos las simulaciones.

```
Porcentaje.HT.2= sapply(1:B,function(x, Indices, Pob)
  {sum(Pob[Indices[,x],"TOTAL_MORENA_PT_PVME"]/Pob[Indices[,x],"pik2"])*100
     /sum((Pob[Indices[,x],"TOTAL_VOTOS_VALIDOS"]/Pob[Indices[,x],"pik2"]))},
  Indices=ParMuestra.2, Pob=Datos)
h2=hist(Porcentaje.HT.2, plot=FALSE, breaks=25)
plot(h2, freq=FALSE, main=TeX("$\\widehat{R}$"),
     xlab=paste("\n Estimación ST op 2 \n n=",dim(ParMuestra.2)[1], ", ECM=",
     round(mean((Porcentaje.HT.2-Porcentaje)^2), 2) ,sep=""),
```

```
     ylab="", cex.lab=.9, yaxt='n')
abline(v=Porcentaje, col="red")
```



Obtenemos un error cuadrático medio de 0.2 %, teneindo así una estimación muy buena.

```
mean(Porcentaje.HT.2)
```

Por otro lado el promedio de las simulaciónes es de 44.3 %. Un valor muy parecido al muestral pero mas alejado que con un m.a.s.

Varianza

Nuevamente usamos la expresión (69) y usamos R como calculadora, solo que ahora usamos las probabilidades de inclusión para este diseño estratificado.

```
VarPob1=0
Varianza1=function(){
  for(k in 1:N) {
    v_k=(1/t_z*(TOTAL_MORENA_PT_PVME[k]-Porcentaje*TOTAL_VOTOS_VALIDOS[i]))
    for(l in 1:N){
      v_l=(1/t_z*(TOTAL_MORENA_PT_PVME[l]-Porcentaje*TOTAL_VOTOS_VALIDOS[l]))
      if(k==l){
        VarPob1=VarPob1 + (Datos$pik2[k]
                - Datos$pik2[k]**2)*(v_k**2/Datos$pik2[k]**2)
      }else{
        #Pues Delta_kl =0
        VarPob1=VarPob1 + 0
      }
    }
  }
}
```

III.- **Estratificación por circunscripciones:**

Para esta estrategia, primero clasificaremos a los estados por circrunscripciones, con la ayuda del ordenamiento por nombre de estado de nuestra base y la variable *NOM_ENT* que contiene el nombre de las entidades federativas.

```
Datos=Datos[order(Datos$NOMBRE_ESTADO),]
NOM_ENT<-unique(Datos$NOMBRE_ESTADO)
```

Creamos un ciclo que asigna la circunscripción, basandose en el nombre del estado.

```
CIRC<-c()
a=0
for(i in 1:length(Datos$NOMBRE_ESTADO)){
  a=Datos$NOMBRE_ESTADO[i]
  if(a==NOM_ENT[2]| a==NOM_ENT[3]|
     a==NOM_ENT[6]| a==NOM_ENT[10]|
     a==NOM_ENT[14]| a==NOM_ENT[18]|
     a==NOM_ENT[25]| a==NOM_ENT[26]){
    CIRC[i]=1
  }
  if(a==NOM_ENT[1]| a==NOM_ENT[8]|
     a==NOM_ENT[11]| a==NOM_ENT[19]|
     a==NOM_ENT[22]| a==NOM_ENT[24]|
     a==NOM_ENT[28]| a==NOM_ENT[32]){
    CIRC[i]=2
  }
  if(a==NOM_ENT[4]| a==NOM_ENT[5]|
     a==NOM_ENT[20]| a==NOM_ENT[23]|
     a==NOM_ENT[27]| a==NOM_ENT[30]|
     a==NOM_ENT[31]){
    CIRC[i]=3
  }
  if(a==NOM_ENT[7]| a==NOM_ENT[12]|
     a==NOM_ENT[17]| a==NOM_ENT[21]|
     a==NOM_ENT[29]){
    CIRC[i]=4
  }
  if(a==NOM_ENT[9]| a==NOM_ENT[13]|
     a==NOM_ENT[15]| a==NOM_ENT[16]){
```

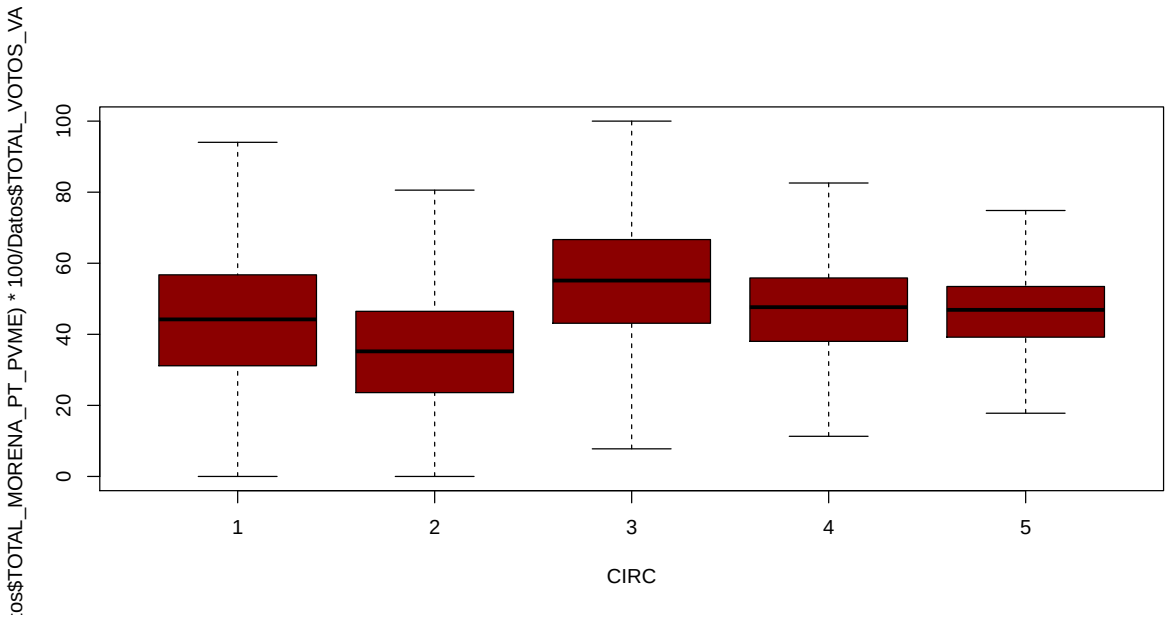
```
    CIRC[i]=5
  }
}
```

Veamos cuantas circunscripciones hay.

```
CIRC<-as.factor(CIRC)
levels(CIRC)

## [1] "1" "2" "3" "4" "5"
```

Vemos como se distribuyen los porcentajes en cada circunscripción.

```
count(Datos,CIRC)

##   CIRC     n
## 1    1 34970
## 2    2 34544
## 3    3 32521
## 4    4 30312
## 5    5 31319

boxplot((Datos$TOTAL_MORENA_PT_PVME)*100/Datos$TOTAL_VOTOS_VALIDOS
        ~ CIRC, data = Datos, col = "darkred", outline=FALSE)
```



Crearemos una base auxiliar que solamente tenga la información de una casilla por distrito electoral, con la finalidad de poder escojer los 6 distritos por circunscripción.

```
BaseAux<-rbind(Datos[1,])
for (i in 1:length(Datos[,1])){
  if(Datos$ID_DIS[i]!=BaseAux$ID_DIS[nrow(BaseAux)]){
    BaseAux<-rbind(BaseAux,Datos[i,])
  }
}
```


```
BaseAux<-rbind(Datos[1,])
for (i in 1:length(Datos[,1])){
  if(Datos$ID_DIS[i]!=BaseAux$ID_DIS[nrow(BaseAux)]){
    BaseAux<-rbind(BaseAux,Datos[i,])
  }
}
```


Obteneindo una base con sólo 300 renglones corresponidentes a los 300 dsitritos electorales.

```
count(BaseAux,CIRC)
```

```
##   CIRC  n
## 1    1 60
## 2    2 62
## 3    3 60
## 4    4 56
## 5    5 62
```

Con esta base ya podemos hacer la primera estratificación, en donde seleccionaremos 30 distritos, 6 por cada circunscripción.

```
n1=6*5
(nhCalculo1=BaseAux %>% group_by(CIRC) %>%
           summarise(Nh21=n(), nh21=6, pik21=nh21/Nh21 ) )
```

```
## # A tibble: 5 x 4
##    CIRC  Nh21  nh21  pik21
##   <int> <int> <dbl>  <dbl>
## 1     1    60     6 0.1
## 2     2    62     6 0.0968
## 3     3    60     6 0.1
## 4     4    56     6 0.107
```

```
## 5     5    62     6 0.0968
```

```
sum(nhCalculo1$nh21)
```

```
## [1] 30
```

Guardamos esta información en la base de datos (marco muestral).

```
nhCalculo1[2]
```

```
## # A tibble: 5 x 1
##    Nh21
##   <int>
## 1    60
## 2    62
## 3    60
## 4    56
## 5    62
```

```
BaseAux=BaseAux %>% group_by(CIRC) %>%
        mutate(Nh21=n(), nh21=6, pik21=nh21/Nh21 )  %>%
        ungroup()
```

De igual manera lo almacenamos en la base original.

```
Datos=merge(Datos,nhCalculo1)
```

Empezamos con la selección mediante una función que noa arrojará los distritos selccionados.

```
Datosdiv1 <- split(BaseAux, list(BaseAux$CIRC))
indexstrata11=function(){
  samples <- lapply(Datosdiv1, function(x) sample(x$ID_DIS,
                       unique(x$nh21), FALSE))
  IndexS1 <- unlist(samples)
  return(IndexS1)
}
(a=indexstrata11())
```

```
##    11    12    13    14    15    16    21    22    23    24    25    26    31
##  25-1  18-2   8-9 14-11   2-8  14-9  24-5 11-15 11-13 11-11  11-2   1-2  7-12
##    32    33    34    35    36    41    42    43    44    45    46    51    52
```

```
## 30-17  20-5 30-16  20-8   4-1 21-13  9-12  9-14  17-2  9-10  9-19  15-8 15-39
##    53    54    55    56
##  16-4 15-18  15-2  15-1
## 300 Levels: 1-1 1-2 1-3 10-1 10-2 10-3 10-4 11-1 11-10 11-11 11-12 ... 9-9

length(a)

## [1] 30
```

Con esta selección podemos obtener nuestra segunda base, quedannos sólo con estos distritos, y eliminando los demas de la base original (Datos).

```
ID_DIS=as.vector(a)
P<-as.data.frame(ID_DIS)
BaseAux1=merge(Datos,P)
length(unique(BaseAux1$ID_DIS))

## [1] 30
```

Con esta base ya podemos hacer la segunda estratificación, para obtener las 30 casillas por cada distrito.

```
n12=30*30
(nhCalculo12=BaseAux1 %>% group_by(ID_DIS)
             %>% summarise(Nh212=n(), nh212=30, pik212=nh212/Nh212 ))

## # A tibble: 30 x 4
##    ID_DIS Nh212 nh212 pik212
##    <fct>  <int> <dbl>  <dbl>
##  1 1-2      566    30 0.0530
##  2 11-11    441    30 0.0680
##  3 11-13    563    30 0.0533
##  4 11-15    465    30 0.0645
##  5 11-2     542    30 0.0554
##  6 14-11    538    30 0.0558
##  7 14-9     507    30 0.0592
##  8 15-1     582    30 0.0515
##  9 15-18    504    30 0.0595
## 10 15-2     446    30 0.0673
## # ... with 20 more rows

sum(nhCalculo12$nh212)
```

```
## [1] 900
```

Análogamente guardamos esta información en la base de datos (marco muestral).

```
BaseAux1=BaseAux1 %>% group_by(ID_DIS) %>%
         mutate(Nh212=n(), nh212=30, pik212=nh212/Nh212) %>% ungroup()
```

También las agregamos a nuestra base original. pero en lugar de sólo sacar las probabiliades únicamente de los 30 distritos selecinados, lo hacemos para los 300 distritos.

```
(nhCalculo13=Datos %>% group_by(ID_DIS) %>%
             summarise(Nh212=n(), nh212=30, pik212=nh212/Nh212 ) )

## # A tibble: 300 x 4
##    ID_DIS Nh212 nh212 pik212
##    <fct>  <int> <dbl>  <dbl>
##  1 1-1      537    30 0.0559
##  2 1-2      566    30 0.0530
##  3 1-3      620    30 0.0484
##  4 10-1     674    30 0.0445
##  5 10-2     655    30 0.0458
##  6 10-3     655    30 0.0458
##  7 10-4     573    30 0.0524
##  8 11-1     480    30 0.0625
##  9 11-10    584    30 0.0514
## 10 11-11    441    30 0.0680
## # ... with 290 more rows

Datos=merge(Datos,nhCalculo13)
```

Ponemos las probabiliades finales, que son la multiplicación de que se selccione el distrito y luego la casiilla.

```
Datos$PROBABILIDAD=Datos$pik21*Datos$pik212
```

Empezamos con la selección mediante una función que nos arrojará el ID de los distritos selccionados. Finalizando con el diseño de muestreo.

```
Datosdiv12 <- split(BaseAux1, list(BaseAux1$ID_DIS))
indexstrata112=function(){
  samples <- lapply(Datosdiv12, function(x) sample(x$ID, unique(x$nh212), FALSE))
  IndexS12 <- unlist(samples)
```

```
  return(IndexS12)
}
```

Ahora haremos una función que haga las dos estratificaciones.

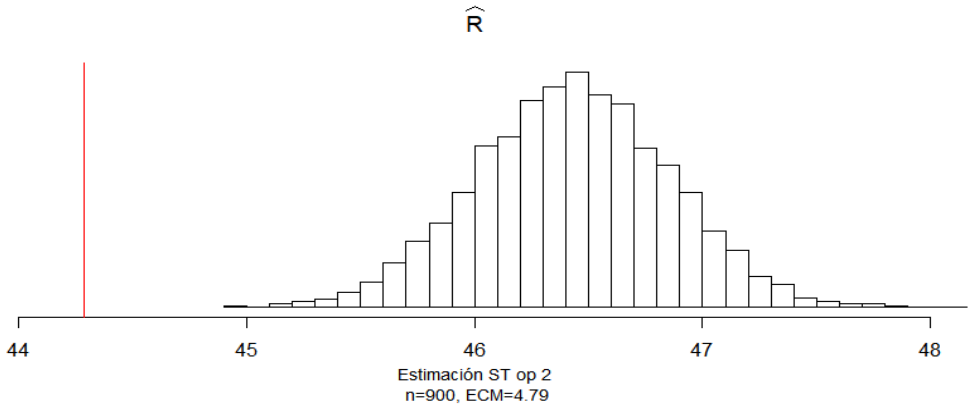
```
Estrato<-function(){
  #Hacemos la selección de los 6 distritos por cada circunscripción
  a=(indexstrata11())
  #Ceramos la base para seleccionar casillas
  #Usamos un for para llenarla
  ID_DIS=as.vector(a)
  P<-as.data.frame(ID_DIS)
  BaseAux1=merge(Datos,P)
  #Selección de casillas
  b=(indexstrata112())
  return(b)
}
```

Con la función anterior podemos estudiar la distribución vía simulación.

```
ParMuestra.3 = replicate(B, Estrato())
Datos=Datos[order(Datos$ID),]
```

Finanlmente graficamos.

```
Porcentaje.HT.3= sapply(1:B,function(x, Indices, Pob)
{sum(Pob[Indices[,x],"TOTAL_MORENA_PT_PVME"]/Pob[Indices[,x],"PROBABILIDAD"])*100
    sum((Pob[Indices[,x],"TOTAL_VOTOS_VALIDOS"]/Pob[Indices[,x],"PROBABILIDAD"]))
Indices=ParMuestra.3, Pob=Datos)
h3=hist(Porcentaje.HT.3, plot=FALSE, breaks=25)
plot(h3, freq=FALSE, main=TeX("$\\widehat{R}$"),
     xlab=paste("\n Estimación ST op 2 \n n=",dim(ParMuestra.3)[1],", ECM=",
     round(mean((Porcentaje.HT.3-Porcentaje)^2), 2) ,  sep=""),
     ylab="", cex.lab=.9, yaxt='n', xlim = c(44,48))
abline(v=Porcentaje, col="red")
```



Teneos un error cuadrático medio del 4.79 %, lo que es un error grande, pues hablamos de porcentajes.

```
mean(Porcentaje.HT.3)
```

Por otro lado el promedio de las simulaciones es de 46.4 %, siendo un valor muy distinto a nuestro valor muestral.

Conclusiones: El diseño de muestreo m.a.s. es el que en promedio se parece más al valor verdadero, pero la estratificación por medio de distritos tiene un error cuadrático medio menor, por lo tanto, este diseño de muestreo es el mejor. En cunato a la estratificación por circunscripciones, tenemos una estimación muy mala, la que atribuimos a que se está realizndo practicamente un muestreo por conglomerado, pues se selecionan pocos distritos electorales y muchas casillas de estos distritos, sesgando considerablemente la muestra. Una propuesta para mejorar sería selcionar más distritos y menos casillas por distritos, por ejemplo slecionar 30 y 6 respectivamente.

2. **Muestreo bietápico** Un estudiante de posgrado tiene una colección de 300 libros en su librero. Para estimar el número total de palabras en su colección de libros, selecciona una muestra s_I de dos libros usando un muestreo aleatorio simple sin reemplazo. Para los dos libros seleccionados tiene el número de páginas de cada uno. Posteriormente, en cada libro seleccionado selecciona dos páginas usando un muestreo aleatorio simple sin reemplazo. Una vez seleccionada la página registra de cada una el número de palabras que contiene. La información muestral es como sigue.

Muestras seleccionadas en cada etapa y valores observados

Libro seleccionado (s_I)	Número de páginas (N_i)	Páginas seleccionadas (s_i)	Número de Palabras en la página (y_k)
195	200	61	23
		112	25
288	100	99	20
		11	20

Aquí las páginas conforman la población de interés y los libros son las upm.

i) Calcule las probabilidades de inclusión de primer y segundo orden correspondientes a las páginas seleccionadas.

ii) Calcule el valor estimado del número total de palabras en la colección de libros.

iii) Dé una estimación insesgada de la varianza del estimador usado en ii).

iv) Dé una aproximación de la varianza del estimador usado en ii) que sólo use las probabilidades de inclusión de primer orden o los factores de expansión.

Solución.

Analizando el problema vemos que se nos dan la primera etapa del muestreo bietapico las ump, es decir los libros, de los 300 libros se seleccionan 2 de ellos con un muestreo simple sin reemplazo, suguiendo el resultado 5.4 de las notas(p. 76) tenemos que la probabilidad de inclusión de primer orden para un muestreo bietapico esta dada por

$$\pi_k = \pi_{k|i}\pi_{Ii}$$

Y las probabilidades de inclusión de segundo orden estan dadas por

$$\pi_{kl} = \pi_{Ii}\pi_{kl|i}$$

si i=j

$$\pi_{kl} = \pi_{Iij}\pi_{k|i}\pi_{l\ midj}$$

$sii \neq j$

Así tenemos que

$$\pi_k = \frac{2}{300} * \frac{2}{N_i}$$

Para las probabilidades de segundo orden tenemos que

$$\pi_{kl} = \frac{2}{300} * \frac{2}{N_i(N_i - 1}$$

$sii = j$

$$\pi_{kl} = \frac{2}{(300)(299)} * \frac{2}{N_i} * \frac{2}{N_k}$$

$sii \neq j$

Utilizando la muestra dada tenemos que al hacer la selección se obtiene $S_I = 195, 288$ con $N_1 95 = 200$ y $N_2 88 = 100$, sustituyendo obtenemos para $k = 195$

$$\pi_k = \frac{2}{300} * \frac{2}{200} = \frac{4}{60000}$$

Para la probabilidad de segundo orden tenemos que

$$\pi_{kl} = \frac{2}{300} * \frac{2}{N_i(N_i - 1)} = \frac{4}{60000 * 199} = \frac{4}{11,940,000}$$

Para $k = 288$

$$\pi_k = \frac{2}{300} * \frac{2}{100} = \frac{4}{30000}$$

Para la probabilidad de segundo orden si $i = j$

$$\pi_{kl} = \frac{2}{300} * \frac{2}{N_i(N_i - 1)} = \frac{4}{30000 * 99} = \frac{4}{2,970,000}$$

Si $sii \neq j$

$$\pi_{kl} = \frac{2}{(300)(299)} * \frac{2}{200} * \frac{2}{100} = \frac{8}{1,794,000,000}$$

Basandonos en el resultado 5.5 de las notas el estimador de HT para totales para un diseño bietapico, con $\theta = \mathrm{t}_y$ es

$$\widehat{t}_{y\pi} = \sum_{i \in S_I} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} = \frac{\widehat{t}_{1\pi}}{\pi_{I1}} + \frac{\widehat{t}_{2\pi}}{\pi_{I2}}$$

Sustituyendo k=195, $\pi_{k|i} = \frac{2}{200}$ y los datos del muestreo obtenemos que

$$\widehat{t}_{1\pi}\pi_{I1} = \sum_{K\in s_1} \frac{y_k}{\pi_{k|i}} = \frac{23}{\frac{2}{200}} + \frac{25}{\frac{2}{200}} = \frac{48*200}{2} = 48(100) = 4800$$

Con k=288 y $\pi_{k|i} = \frac{2}{100}$

$$\widehat{t}_{2\pi}\pi_{I1} = \sum_{K\in s_2} \frac{y_k}{\pi_{k|i}} = \frac{20}{\frac{2}{100}} + \frac{20}{\frac{2}{00}} = \frac{40*100}{2} = 20(100) = 2000$$

Asi

$$\widehat{t}_{y\pi} = \sum_{i\in S_I} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} = \frac{\widehat{t}_{1\pi}}{\pi_{I1}} + \frac{\widehat{t}_{2\pi}}{\pi_{I2}} = \frac{4800}{\frac{2}{300}} + \frac{2000}{\frac{2}{300}} = \frac{6800*300}{2} = 1,020,000$$

Así la estimación total de palabras es 1,020,000

ii) Calcule el valor estimado del número total de palabras en la colección de libros.

Basandonos en el resultado 5.5 de las notas el estimador de HT para totales para un diseño bietapico, con θ =t_y es

$$\widehat{t}_{y\pi} = \sum_{i\in S_I} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} = \frac{\widehat{t}_{1\pi}}{\pi_{I1}} + \frac{\widehat{t}_{2\pi}}{\pi_{I2}}$$

Sustituyendo k=195, $\pi_{k|i} = \frac{2}{200}$ y los datos del muestreo obtenemos que

$$\widehat{t}_{1\pi}\pi_{I1} = \sum_{K\in s_1} \frac{y_k}{\pi_{k|i}} = \frac{23}{\frac{2}{200}} + \frac{25}{\frac{2}{200}} = \frac{48*200}{2} = 48(100) = 4800$$

Con k=288 y $\pi_{k|i} = \frac{2}{100}$

$$\widehat{t}_{2\pi}\pi_{I1} = \sum_{K\in s_2} \frac{y_k}{\pi_{k|i}} = \frac{20}{\frac{2}{100}} + \frac{20}{\frac{2}{00}} = \frac{40*100}{2} = 20(100) = 2000$$

Asi

$$\widehat{t}_{y\pi} = \sum_{i\in S_I} \frac{\widehat{t}_{i\pi}}{\pi_{Ii}} = \frac{\widehat{t}_{1\pi}}{\pi_{I1}} + \frac{\widehat{t}_{2\pi}}{\pi_{I2}} = \frac{4800}{\frac{2}{300}} + \frac{2000}{\frac{2}{300}} = \frac{6800*300}{2} = 1,020,000$$

Así la estimación total de palabras es 1,020,000

iii) Dé una estimación insesgada de la varianza del estimador usado en ii).

Basandonos en el resultado 5.6 (pag.78) ecuación 160 tenemos que

$$\widehat{V}(\widehat{t}_{y\pi}) = \sum_{i\in s_I}\sum_{j\in s_I} \widehat{\Delta}_{Iij} \frac{\widehat{t}_{i\pi}}{\pi Ii} \frac{\widehat{t}_{j\pi}}{\pi Ij} + \sum_{i\in S_I} \frac{\widehat{V}_i}{\pi Ii}$$

Primero calculemos si $i = j$

$$\widehat{\Delta}_{Iij} = \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}}$$

Como ya sabemos

$$\pi_{Iij} = \frac{2}{300}, \pi_{Iij} = \frac{2}{300}$$

Ademas ya vimos que si i$\neq j$

$$\pi_{Iij} = \frac{2}{(300)(299)}$$

si $i = j$

$$\pi_{Iij} = \frac{2}{300}$$

Así si $i \neq j$ tenemos que

$$\widehat{\Delta}_{Iij} = \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} = \frac{\frac{2}{(300)(299)} - \frac{2}{300} * \frac{2}{300}}{\frac{2}{(300)(299)}} = 1 - \frac{598}{300} = -\frac{149}{150}$$

Si i=j tenemos que

$$\widehat{\Delta}_{Iij} = \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Iij}} = \frac{\frac{2}{(300)} - \frac{2}{300} * \frac{2}{300}}{\frac{2}{(300)}} = 1 - \frac{2}{300} = \frac{149}{150}$$

Sustituyendo todos los valores obtenemos que

$$\sum_{i\in s_I}\sum_{j\in s_I} \widehat{\Delta}_{Iij}\frac{\widehat{t}_{i\pi}}{\pi Ii}\frac{\widehat{t}_{j\pi}}{\pi Ij} = \frac{149}{150}\frac{(4800)^2}{(\frac{2}{300})^2} - \frac{149}{150}\frac{(4800)(2000)}{(\frac{2}{300})^2} + (\frac{-149}{150}\frac{(4800)(2000)}{(\frac{2}{300})^2} + \frac{149}{150}\frac{(2000)^2}{(\frac{2}{300})^2}) = 175224000000$$

Ahora para calcular

$$\sum_{i\in s_I}\frac{\widehat{V}_i}{\pi_{Ii}} = \frac{\widehat{V}_1}{\frac{2}{300}} + \frac{\widehat{V}_2}{\frac{2}{300}}$$

si $i = j$

$$\widehat{V}_1\widehat{\Delta}_{kl|i} = \frac{198}{200}$$

si $i \neq j$

$$\widehat{\Delta}_{kl|i} = -\frac{198}{200}$$

Analogamente para $\widehat{V_2}$ si i=j

$$\widehat{\Delta}_{kl|i} = \frac{98}{100}$$

si $i \neq j$

$$\widehat{\Delta}_{kl|i} = \frac{-98}{100}$$

Así calculando $\widehat{V_1}$ tenemos que

$$\widehat{V_1} = (\frac{198}{200}\frac{23*25}{(\frac{2}{200})^2} - \frac{198}{200}\frac{23*25}{(\frac{2}{200})^2} + (\frac{-198}{200}\frac{23*25}{(\frac{2}{200})^2} + \frac{198}{200}\frac{(25)^2}{(\frac{2}{200}{}^2)} = 22809600$$

Para $\widehat{V_2}$

Ahora

$$\widehat{V_1} = \frac{22809600}{\frac{1}{150}} = 3421440000$$

Por lo tanto

$$\widehat{V}(\widehat{t}_{y\pi}) = \sum_{i\in s_I}\sum_{j\in s_I}\widehat{\Delta}_{Iij}\frac{\widehat{t}_{i\pi}}{\pi Ii}\frac{\widehat{t}_{j\pi}}{\pi Ij} + \sum_{i\in S_I}\frac{\widehat{V}_i}{\pi Ii} = 3421440000 + 0 + 175224000000 = 178645440000$$

iv) Dé una aproximación de la varianza del estimador usado en ii) que sólo use las probabilidades de inclusión de primer orden o los factores de expansión.

Sabemos que este estimador que cumple las condiciones es el siguiente

$$\widehat{V}(\widehat{t}_{y\pi} = 2\sum_{i\in I}^{n_1}(\sum_{i\in I}^{m}\frac{y_k}{\pi_k} - \frac{\widehat{t}_{y\pi}}{2})^2 = 2(\frac{1440000}{2} - \frac{10200000}{2})^2 + 2(\frac{600000}{2} - \frac{1020000}{2}{}^2 = 4(210000)^2 = 176400000000$$

3. **Estimación básica de una encuesta con diseño multietápico**
   Considere la Encuesta Nacional de Vivienda (ENVI) 2020 https://www.inegi.org.mx/programas/envi/2020/

   Suponga que será el encargado de generar los resultados básicos presentados en el tabulado llamado Cuadro 5.1, ver Figura **??** (https://www.inegi.org.mx/contenidos/programas/envi/2020/tabulados/envi_2020_tema_05_xlsx.zip)

   En particular realice lo siguiente

   I. Describa brevemente el diseño de muestreo usado en la encuesta. Es decir, si es muestreo aleatorio simple, tiene estratificación, es por conglomerados, etc.

   II. Identifique las variables asociadas al diseño de muestreo que están presentes en la base de datos a usar (THOGAR en https://www.inegi.org.mx/programas/envi/2020/#Microdatos)

   III. Identifique la pregunta y variable asociada a la identificación de los Hogares con necesidad de rentar, comprar o construir una vivienda independiente de la que habitan.

   IV. Con esta información, estime el número total de hogares y el porcentaje de hogares que tienen una necesidad de vivienda a nivel nacional y por entidad federativa.

   V. Calcule intervalos de confianza para los parámetros estimados en el inciso anterior. Comente sobre los resultados obtenidos.

   VI. **Punto extra opcional**. Considerando el porcentaje de hogares que tienen una necesidad de vivienda por entidad federativa, realice un mapa de calor (Geographic Heat Map) y comente los resultados.

INEGI. Encuesta Nacional de Vivienda. ENVI 2020. Tabulados básicos. 2021 ÍNDICE

Estimaciones puntuales.

Hogares por entidad federativa, según condición de tener necesidad de rentar, comprar o construir una vivienda independiente de la que habitan

Cuadro 5.1

Entidad federativa	Total de hogares	Condición de necesidad de vivienda					
		Sí		No		No especificado	
		Absolutos	Relativos	Absolutos	Relativos	Absolutos	Relativos
Estados Unidos Mexicanos	**36 210 467**	**7 628 562**	**21.1**	**28 529 481**	**78.8**	**52 424**	**0.1**
Aguascalientes	397 770	61 505	15.5	335 194	84.3	1 071	0.3
Baja California	1 156 528	281 959	24.4	872 108	75.4	2 461	0.2
Baja California Sur	246 920	64 373	26.1	182 547	73.9	0	0.0
Campeche	262 489	67 321	25.6	195 002	74.3	166	0.1
Coahuila de Zaragoza	913 569	163 939	17.9	747 788	81.9	1 842	0.2
Colima	234 272	40 317	17.2	193 611	82.6	344	0.1
Chiapas	1 460 368	400 920	27.5	1 057 487	72.4	1 961	0.1
Chihuahua	1 147 667	180 625	15.7	965 828	84.2	1 214	0.1
Ciudad de México	2 808 652	759 121	27.0	2 046 170	72.9	3 361	0.1

INEGI. Encuesta Nacional de Vivienda. ENVI 2020. Tabulados básicos. 2021 ÍNDICE

Error estándar.

Hogares por entidad federativa, según condición de tener necesidad de rentar, comprar o construir una vivienda independiente de la que habitan

Cuadro 5.1

Entidad federativa	Total de hogares	Condición de necesidad de vivienda					
		Sí		No		No especificado	
		Absolutos	Relativos	Absolutos	Relativos	Absolutos	Relativos
Estados Unidos Mexicanos	**250 260.7**	**105 958.2**	**0.3**	**220 719.5**	**0.3**	**7 310.7**	**0.0**
Aguascalientes	9 362.6	3 974.8	0.9	8 564.8	0.9	475.7	0.1
Baja California	30 667.9	19 337.4	1.5	27 919.5	1.5	1 244.3	0.1
Baja California Sur	9 038.1	5 173.9	1.6	6 449.4	1.6	0.0	0.0
Campeche	10 080.2	4 445.6	1.3	8 161.9	1.3	129.1	0.0
Coahuila de Zaragoza	20 581.8	10 105.6	1.0	19 628.4	1.1	934.0	0.1
Colima	8 612.2	3 021.6	1.1	7 590.5	1.1	201.2	0.1
Chiapas	65 259.8	24 523.2	1.2	51 431.1	1.2	1 401.0	0.1
Chihuahua	21 729.7	11 366.6	1.0	22 561.5	1.0	882.9	0.1
Ciudad de México	53 774.2	39 095.9	1.2	48 555.5	1.2	2 378.1	0.1

Solución.

I. Diseño de muestreo usado en la encuesta.

El diseño de muestreo para la ENVI 2020 es probabilístico, a su vez es bietápico, **estratificado** y **por conglomerados**, donde la unidad última de selección es la vivienda.

Conglomerados:

Las unidades primarias de muestreo *upm* están constituidas por agrupaciones de viviendas con características diferenciadas dependiendo del ámbito al que pertenecen:

- En urbano alto.

- En complemento urbano.
- En rural.

El total de $upm's$ formadas fue de $245,279$.

Estratos:

La división política del país y la conformación de localidades diferenciadas por su tamaño, forman una estratificación geográfica. En cada entidad federativa se distinguen tres ámbitos, divididos a su vez en zonas, como se indica en el siguiente cuadro:

Ámbito	Zona	Descripción
Urbano alto	01 a 09	Ciudades con 100 000 o más habitantes
Complemento urbano	25	De 50 000 a 99 999 habitantes
	35	De 15 000 a 49 999 habitantes
	45	De 5 000 a 14 999 habitantes
	55	De 2 500 a 4 999 habitantes
Rural	60	Localidades menores de 2 500 habitantes

Se formaron también cuatro estratos **sociodemográficos** en los que se agruparon todas las UPM del país, esta estratificación considera las características sociodemográficas de los habitantes de las viviendas, las características físicas y el equipamiento de las mismas.

De esta forma, cada UPM fue clasificada en un único estrato geográfico y uno sociodemográfico. Como resultado, se tienen un total de 683 estratos en todo el ámbito nacional.

II. Variables asociadas al diseño de muestreo que están presentes en la base de datos a usar (THOGAR).

La base de datos consta de 14 variables, de las cuales, las de interés para el diseño de muestreo son:

- P3A1 1: Tener la necesidad o planear rentar, comprar o construir una vivienda. (Sí=1, No=2, No sabe=9).
- EST DIS: Estrato de diseño.
- UPM DIS: Unidad primaria de muestreo diseño.
- FACTOR: Factor de expansión.
- ENT: Entidad Federativa.

III. Identifique la pregunta y variable asociada a la identificación de los Hogares con necesidad de rentar, comprar o construir una vivienda independiente de la que habitan.

- **Pregunta:** 3a.1 ¿Alguna de las personas de este hogar tiene necesidad o está planeando rentar, comprar o construir una vivienda, independientemente de la que habitan actualmente?

- **Variable:** P3A1 1

IV. Estimación del número total de hogares y el porcentaje de hogares que tienen una necesidad de vivienda a nivel nacional y por entidad federativa.

V. Calcule intervalos de confianza para los parámetros estimados en el inciso anterior. Comente sobre los resultados obtenidos.

Las tablas se encuentran en las siguientes páginas, su codigo se anexa en un archivo .R

- Tabla 1: Totales
- Tabla 2: Limites inferiores para intervalos del 90 porciento de confianza.
- Tabla 3: Limites superiores para intervalos del 90 porciento de confianza