```
> ## AGYEMANG ERIC
> ## MAT 450
> ## HOMEWORK 6
> library(survey)
> library(SDaA)

> # QUESTION 3
> Store=c("A","B","C","D")
> Size=c(100,200,300,1000)
> phi_i=c(1/16,2/16,3/16,10/16)
> ti=c(75,75,75,75)
> t=sum(ti)
> T_phi_i=ti/phi_i
> T_i=(T_phi_i-t)^2
> dat=cbind.data.frame(Store,Size,phi_i,ti,T_phi_i,T_i)
> dat
  Store Size  phi_i ti T_phi_i    T_i
1     A  100 0.0625 75    1200 810000
2     B  200 0.1250 75     600  90000
3     C  300 0.1875 75     400  10000
4     D 1000 0.6250 75     120  32400
>
> Et_phi =sum(phi_i*T_phi_i)
> Et_phi
[1] 300
> # As the E[tˆψ] required

> Vt_phi = sum(phi_i*T_i)
> Vt_phi
[1] 84000
> # As the V[tˆψ] required
> ##############################################################
>
> # QUESTION 4
> Store=c("A","B","C","D")
> Size=c(100,200,300,1000)
> phi_i=c(7/16,3/16,3/16,3/16)
> ti=c(11,20,24,245)
> T_phi_i=ti/phi_i
> t=sum(ti)
> t
[1] 300
> T_i=(T_phi_i-t)^2
> dat=data.frame(Store,Size,phi_i,ti, T_phi_i,T_i)
> dat
   Store Size phi_i  ti    T_phi_i      T_i
1     A  100 0.4375  11   25.14286   75546.45
2     B  200 0.1875  20  106.66667   37377.78
3     C  300 0.1875  24  128.00000   29584.00
4     D 1000 0.1875 245 1306.66667 1013377.78
>
> Et_phi =sum(phi_i*T_phi_i)
> Et_phi
[1] 300
```

```
> # As the E[tˆψ]= t= 300. Hence unbiased estimator.
> Vt_phi = sum(phi_i*T_i)
> Vt_phi
[1] 235615.2
> # As the V[tˆψ] required
```

This is a poor sampling design. Store A, with the smallest sales, is sampled with the largest probability, while Store D is sampled with a smaller probability. The ψ_i used in this exercise prod uce a higher variance than simple random sampling.

```
> # QUESTION 9a)
> library(pps)
> set.seed(1000)
> View(statepps)
> T=sum(statepps$landarea)
> T
[1] 3536281

> #As the total land area
>
> samp<-ppswr(statepps$landarea,10)
> samp
 [1] 11 38  2 35 26  2 38 28  4  5
> sampp<-statepps[c(samp),c(1,2,4,5)]
> sampp

             state      counties landarea cumland
11         Georgia       159     57919 1165260
38          Oregon        36     96003 2708173
2           Alaska        25    570374  621124
35    North Dakota        53     68994 2502538
26        Missouri       115     68898 1867609
2.1         Alaska        25    570374  621124
38.1        Oregon        36     96003 2708173
28        Nebraska        93     76878 2090043
4         Arkansas        75     52075  786841
5       California        58    155973  942814
>
> phi=sampp$landarea/T
>
> sampl<-cbind(sampp,phi)
> sampl
             state     counties landarea cumland     phi
11         Georgia       159     57919    1165260   0.01637851
38          Oregon        36     96003    2708173   0.02714801
2           Alaska        25    570374    621124    0.16129205
35    North Dakota        53     68994    2502538   0.01951033
26        Missouri       115     68898    1867609   0.01948318
2.1         Alaska        25    570374    621124    0.16129205
38.1        Oregon        36     96003    2708173   0.02714801
28        Nebraska        93     76878    2090043   0.02173979
4         Arkansas        75     52075    786841    0.01472592
5       California        58    155973    942814    0.04410651

> # As the required sample of size 10 with replacement and ψi for
    each state in each sample.
```

```
> # QUESTION 9b)
> set.seed(1000)
> samp2<-ppswr(statepps$popn,10)
> samp2
 [1] 14 38  5 35 26  5 37 31 10 11
>
> T2=sum(statepps$popn)
> T2
[1] 255077117
> #As the total population

> sampp2<-statepps[c(samp),c(1,2,6,7)]
> sampp2
# A tibble: 10 x 3

           state      counties  popn    cumpopn
11         Georgia       159   6773364   70123230
38          Oregon        36   2971567 193875268
2           Alaska        25    587766    4725277
35    North Dakota        53    634031 176677048
26        Missouri       115   5190719 136821145
2.1         Alaska        25    587766    4725277
38.1        Oregon        36   2971567 193875268
28        Nebraska        93   1600524 139244016
4         Arkansas        75   2394253   10951898
5       California        58  30895356   41847254

> Phi=sampp2$popn/T2
> sampl2<-cbind(sampp2,phi)
> sampl2
            state    counties popn     cumpopn       phi
 11         Georgia       159  6773364   70123230 0.026554181
 38          Oregon        36  2971567  193875268 0.011649681
 2           Alaska        25   587766    4725277 0.002304268
 35    North Dakota        53   634031  176677048 0.002485644
 26        Missouri       115  5190719  136821145 0.020349607
 2.1         Alaska        25   587766    4725277 0.002304268
 38.1        Oregon        36  2971567  193875268 0.011649681
 28        Nebraska        93  1600524  139244016 0.006274667
 4         Arkansas        75  2394253   10951898 0.009386389
 5       California        58 30895356   41847254 0.121121629

 >

>
> # As the required sample of size 10 with replacement and ψi for each
  state in each sample

>
```

QUESTION 9C)

The two samples differ to the great extent by reason that the samples are selected using the cumulative size method which generates the random sample. Also, the countries selected in each sample are different.

The states present in each sample are Georgia, Oregon, Alaska, North Dakota, Missouri, California, Nebraska, and Arkansas.

```
>
> #QUESTION 10 a)
> SamplingWeight<-1/sampl2$phi
> dat<-cbind(sampl2,SamplingWeight)
> stat_pps<- svydesign(id=~1, fpc=~phi, weights =~SamplingWeight, data=sampl)

#Estimate of the total and standard Error of the total
> svytotal(~sampl2$counties,stat_pps)
                total      SE
sampl2$counties 84131      19539
```

Hence the estimated total number of counties in the United States is 84131 and its standard error is 19539.

```
> #QUESTION 10 b)
> sampl2$fpc<-51
> stat_pps<- svydesign(id=~1, fpc=~fpc, data=sampl2)
> svytotal(~sampl2$counties,stat_pps)
                 total     SE
sampl2$counties 3442.5 632.75
```

As the values for the estimated total and its standard error are calculated by Tom. These values significantly differ from mine. The total differ by 80688.5 while the SE differ by 18906.3. which is bias.

QUESTION 26)

(26) The probability of inclusion $\pi_i = \frac{2M_i}{\sum_{j=1} M_j}$. calculating $\psi_i = \frac{\pi_i}{2}$ and $a_i = \frac{\psi_i(1-\psi_i)}{(1-\pi_i)}$ for each of the pus in the table below:

Psu i	M_i	π_i	ψ_i	a_i
1	5	0.40	0.20	0.26667
2	4	0.32	0.16	0.19765
3	8	0.64	0.32	0.60444
4	5	0.40	0.20	0.26667
5	3	0.24	0.12	0.13895
TOTAL	25	2.00	1.00	1.47487

By the Brewer's method, P(selecting Psu i on the 1st draw) = $\frac{a_i}{\sum_{j=1} a_j}$ and P(psu j on 2nd draw / Psu i on 1st draw) = $\frac{\psi_i}{(1-\psi_i)}$

Then the $P\{s=(1,2)\} = \frac{0.26667}{1.474437} \times \frac{0.16}{0.8} = 0.036174$

$P\{s=(2,1)\} = \frac{0.19765}{1.474437} \times \frac{0.2}{0.84} = 0.031918$

$\pi_{12} = P\{s=(1,2)\} + P\{s=(2,1)\} = 0.036174 + 0.0319$

$= 0.068092$

We then calculate π_{ij} in the table below:

$i \backslash j$	1	2	3	4	5
1	—	0.068	0.193	0.090	0.049
2	0.068	—	0.148	0.068	0.036
3	0.193	0.148	—	0.193	0.107
4	0.090	0.068	0.193	—	0.049
5	0.049	0.036	0.107	0.049	—
Sum	0.400	0.320	0.640	0.400	0.240

using (6.21) we can calculate the variance of the Horvitz – Thompson estimator in the table ~~[illegible]~~ below:

i	j	π_{ij}	π_i	π_j	t_i	t_j	$(\pi_i\pi_j - \pi_{ij})\left(\frac{t_i}{\pi_i} - \frac{t_j}{\pi_j}\right)^2$
1	2	0.068	0.40	0.32	20	25	47.39
1	3	0.193	0.40	0.64	20	38	5.54
1	4	0.090	0.40	0.40	20	24	6.96
1	5	0.049	0.40	0.24	20	21	66.73
2	3	0.148	0.32	0.64	25	38	20.13
2	4	0.068	0.32	0.40	25	24	19.68
2	5	0.036	0.32	0.24	25	21	3.56
3	4	0.193	0.64	0.40	38	24	0.02
3	5	0.107	0.64	0.24	38	21	37.16
4	5	0.049	0.40	0.24	24	21	35.88
Sum		1					243.07

For the population, $t = 128$. We see that $\sum P(s)\hat{t}_{HTS} = 128$ and $\sum P(s)(\hat{t}_{HTS} - 128)^2 = 243.07$ which confirms that we are correct in our calculations.