9.2 A nurseryman wants to estimate the average height of seedlings in a large field that is divided into 50 plots that vary slightly in size. He believe the heights are fairly constant throughout each plot but may vary considerably from plot to plot. Therefore, he decides to sample 10% of the trees within each of 10 plots using a two-stage cluster sample. The data are as given in the accompanying table. Estimate the average height of seedlings in the field and place a bound on the error of estimation.

Plot	Number of seedlings (M_i)	Number of seedlings sampled (m_i)	Heights of seedlings (in inches)	$\bar{y}_i$	s_i^2
1	52	5	12, 11, 12, 10, 13	11.60	1.30
2	56	6	10, 9, 7, 9, 8, 10	8.83	1.37
3	60	6	6, 5, 7, 5, 6, 4	5.50	1.10
4	46	5	7, 8, 7, 7, 6	7.00	0.50
5	49	5	10, 11, 13, 12, 12	11.60	1.30
6	51	5	14, 15, 13, 12, 13	13.40	1.30
7	50	5	6, 7, 6, 8, 7	6.80	0.70
8	61	6	9, 10, 8, 9, 9, 10	9.17	0.57
9	60	6	7, 10, 8, 9, 9, 10	8.83	1.37
10	45	6	12, 11, 12, 13, 12, 12	12.00	0.40

```
> rm(list=ls())
> exercise9.2 <- read.delim("~/Documents/Rutgers/Spring 2020/Stat 476/Homework/Homework 06/exercise9.2.txt")
>   View(exercise9.2)
> data=exercise9.2
> Mi=c(52,56,60,46,49,51,50,61,60,45)
> yi=c(11.6,8.83,5.5,7,11.6,13.4,6.8,9.17,8.83,12)
> sum(Mi*yi)
[1] 4970.65
> sum(Mi*yi)/sum(Mi)
[1] 9.378585
> Mbar=mean(Mi)
> Mbar
[1] 53
> n=10
> mu=sum(Mi*yi)/sum(Mi)
> sb2=sum((Mi*yi-Mbar*mu)^2)/(n-1)
> sb2
[1] 15678.39
> mi=c(5,6,6,5,5,5,5,6,6,6)
> N=50
> si2=c(1.3,1.37,1.1,0.5,1.3,1.3,0.7,0.57,1.37,0.4)
> Vhat=(1-n/N)/(Mbar^2*n)*sb2+sum(Mi^2*(1-mi/Mi)*si2/mi)/(n*N*Mbar^2)
> Vhat
[1] 0.4498783
> B=2*sqrt(Vhat)
> B
[1] 1.341459
```

Ans: $\hat{\mu} = 9.378585$; $B = 1.341459$

9.3 In Exercise 9.2, assume that the nurseryman knows there are approximately 2600 seedlings in the field. Use this additional information to estimate average height and place a bound on the error of estimation.

```
> M=2600
> mu=N/M*sum(Mi*yi)/n
> mu
[1] 9.558942
> Mbar=M/N
> Mbar
[1] 52
> sb2=sum((Mi*yi-Mbar*mu)^2)/(n-1)
> sb2
[1] 15678.39
> Vhat=(N-n)/(N*n*Mbar^2)*sb2+sum(Mi^2*(Mi-mi)/Mi*si2/mi)/(N*n*Mbar^2)
> Vhat
[1] 0.4673477
> B=2*sqrt(Vhat)
> B
[1] 1.367257
```

Ans: $\hat{\mu} = 9.558942; B = 1.367257$

9.4 A supermarket chain has stores in 32 cities. A company official wants to estimate the proportion of stores in the chain that do not meet a specified cleanliness criterion. Stores within each city appear to possess similar characteristics; therefore, she decides to select a two-stage cluster sample containing one-half of the stores within each of four cities. Cluster sampling is desirable in this situation because of travel costs. The data collected are given in the accompanying table. Estimate the proportion of stores not meeting the cleanliness criterion and place a bound on the error of estimation.

City	Numbers of stores in city	Numbers of stores sampled	Number of stores not meeting criterion
1	25	13	3
2	10	5	1
3	18	9	4
4	16	8	2

```
> rm(list=ls())
> N=32
> n=4
> Mi=c(25,10,18,16)
> mi=c(13,5,9,8)
> pi=c(3/13,1/5,4/9,2/8)
> phat=sum(Mi*pi)/sum(Mi)
> phat
[1] 0.2865106
> sr2=sum(Mi^2*(pi-phat)^2)/(n-1)
> sr2
[1] 3.704388
> qi=1-pi
> Mbar=mean(Mi)
> Mbar
[1] 17.25
> Vhat=(N-n)*sr2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)/Mi*pi*qi/(mi-1))/(n*N*Mbar^2)
> Vhat
```

```
[1] 0.003113561
> B=2*sqrt(Vhat)
> B
[1] 0.1115986
```

Ans: $\hat{p} = 0.2865106; B = 0.1115986$

9.5 Repeat Exercise 9.4, given that the chain contains 450 stores. [*Hint*: Use the unbiased estimator of Eq. (9.1) and adapt it to proportions.]

```
> M=450
> Mbar=M/N
> Mbar
[1] 14.0625
> phat=N*sum(Mi*pi)/(M*n)
> phat
[1] 0.351453
> sb2=sum((Mi*pi-Mbar*phat)^2)/(n-1)
> sb2
[1] 6.526134
> Vhat=(N-n)*sb2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)*pi*qi/(Mi*(mi-1)))/(N*n*Mbar^2)
> Vhat
[1] 0.007806349
> B=2*sqrt(Vhat)
> B
[1] 0.1767071
```

Ans: $\hat{p} = 0.351453; B = 0.1767071$

9.11 A market research firm constructed a sampling plan to estimate the weekly sales of brand A cereal in a certain geographic area. The firm decided to sample cities within the area and then to sample supermarkets within cities. The number of boxes of brand A cereal sold in a specified week is the measurement of interest. Five cities are sampled from the 20 in the area. Using the data given in the accompanying table, estimate the average sales for the week for all supermarkets in the area. Place a bound on the error of the estimation. Is the estimator you used unbiased?

City	Number of supermarkets (M_i)	Number of supermarkets sampled (m_i)	$\bar{y}_i$	s_i^2
1	45	9	102	20
2	36	7	90	16
3	20	4	76	22
4	18	4	94	26
5	28	6	120	12

```
> rm(list=ls())
> Mi=c(45,36,20,18,28)
> mi=c(9,7,4,4,6)
> yi=c(102,90,76,94,120)
> si2=c(20,16,22,26,12)
> n=5
> mu=sum(Mi*yi)/sum(Mi)
> mu
[1] 97.97279
> sr2=sum(Mi^2*(yi-mu)^2)/(n-1)
> sr2
[1] 173463.3
```

```
> N=20
> Mbar=mean(Mi)
> Vhat=(N-n)*sr2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)*si2/(Mi*mi))/(n*N*Mbar^2)
> Vhat
[1] 30.22544
> B=2*sqrt(Vhat)
> B
[1] 10.99553
```

Ans: $\hat{\mu}_r = 97.97279; B = 10.99553$

9.12 In Exercise 9.11, do you have enough information to estimate the total number of boxes of cereal sold by all supermarkets in the area during the week? If so, explain how you would estimate this total, and place a bound on the error of estimation.

$$\hat{V}(\hat{\tau}) = M^2\hat{V}(\hat{\mu}) = \left(1 - \frac{n}{N}\right)\frac{N^2 s_b^2}{n} + \frac{N}{n}\sum_{i=1}^{n} M_i^2\left(1 - \frac{m_i}{M_i}\right)\left(\frac{s_i^2}{m_i}\right)$$

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(M_i\bar{y}_i - \bar{M}\hat{\mu})^2$$

Solution: Yes, there is enough information. We can estimate M by multiplying our estimated Mbar by the number of clusters in the population, N.

```
> t=N*(sum(Mi)/n)*mu
> t
[1] 57608
> Vhat=Vhat*((N*(sum(Mi)/n))^2)
> Vhat
[1] 10450266
> B=2*sqrt(Vhat)
> B
[1] 6465.374
```

Ans: $\hat{t} = 57608; B = 6465.374$

9.14 Suppose a sociologist wants to estimate the total number of retired people residing in a certain city. She decides to sample blocks and then sample households within blocks. (Block statistics from the Census Bureau aid in determining the number of households in each block.) Four blocks are randomly selected from the 300 of the city. From the data in the accompanying table, estimate the total number of retired residents in the city and place a bound on the error of estimation.

Block	Number of households (M_i)	Number of households sampled	Number of retired residents per household
1	18	3	1, 0, 2
2	14	3	0, 3, 0
3	9	3	1, 1, 2
4	12	3	0, 1, 1

```
> rm(list=ls())
> N=300
> n=4
> Mi=c(18,14,9,12)
> mi=c(3,3,3,3)
```

```
> yi=c(3/3,3/3,4/3,2/3)
> s2i=c(var(c(1,0,2)),var(c(0,3,0)),var(c(1,1,2)),var(c(0,1,1)))
> sb2=sum((Mi*yi-sum(Mi*yi)/n)^2)/(n-1)
> sb2
[1] 35.66667
> mu=sum(Mi*yi)/sum(Mi)
> mu
[1] 0.9811321
> t=N*(sum(Mi)/n)*mu
> t
[1] 3900
> Vhat=N^2*(1-n/N)/n*sb2+N/n*sum(Mi^2*(1-mi/Mi)*s2i/mi)
> Vhat
[1] 404450.2
> B=2*sqrt(Vhat)
> B
[1] 1271.928
```

Ans: $\hat{t} = 3900; B = 1271.928$

9.15 Using the data in Exercise 9.14, estimate the average number of retired residents per household and place a bound on the error of estimation.

```
> mu
[1] 0.9811321
> Vhat=(N-n)*sr2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)*si2/(Mi*mi))/(n*N*Mbar^2)
> Vhat
[1] 0.0127051
> B=2*sqrt(Vhat)
> B
[1] 0.2254338
```

Ans: $\hat{\mu}_r = 0.9811321; B = 0.2254338$

9.16 From the data in Exercise 9.14, can you estimate the average number of retired residents per block? How can you construct this estimate and place a bound on the error of estimation?

```
> mean(Mi)
[1] 13.25
> 13.25*mu
[1] 13
> Vhat=Vhat*mean(Mi)
> B=2*sqrt(Vhat)
> B
[1] 0.8205913
```

Ans: $\hat{\mu}_{\text{block}} = 13; B = 0.8205913$

Sampling from Real Populations:

Return to the Problem 8.4. You are asked to sample a total of 30 trees to estimate the proportion of diseased trees. Use two-stage cluster sampling, with either rows or columns as clusters, as follows.

```
> rm(list=ls())
> set.seed(0)
> TreeGrid <- c(0,0,0,1,1,
+               0,0,1,1,1,
+               0,0,0,0,0,
+               0,0,1,0,1,
```

```
+             0,0,0,0,1,
+             1,0,1,0,0,
+             0,0,1,0,1,
+             0,1,1,0,1,
+             0,0,0,1,1,
+             0,0,0,1,1,
+             0,0,0,1,0,
+             0,1,1,1,1,
+             0,1,0,1,1,
+             0,0,0,1,1,
+             0,1,0,1,1,
+             0,0,1,1,1,
+             0,0,1,1,1,
+             0,0,0,1,1,
+             0,0,1,1,1,
+             0,0,0,0,0,
+             1,0,0,1,0,
+             0,1,0,0,1,
+             0,0,1,1,0,
+             0,0,0,1,1,
+             0,0,0,1,1,
+             0,1,0,0,1,
+             0,0,1,0,1,
+             1,0,0,0,1,
+             0,0,0,0,1,
+             0,0,1,1,1)
> TreeGrid <- matrix(TreeGrid, nrow=30, ncol=5, byrow=T)
```

1) Draw 10 rows, and then draw 3 cells from each selected row.

```
> random_num=sample(30,10)
> random_num
 [1] 27  8 11 16 24  6 22 26 15 14
> Q1=rbind(TreeGrid[random_num,])
> sam1=sample(1:5,3)
> sam2=sample(1:5,3)
> sam3=sample(1:5,3)
> sam4=sample(1:5,3)
> sam5=sample(1:5,3)
> sam6=sample(1:5,3)
> sam7=sample(1:5,3)
> sam8=sample(1:5,3)
> sam9=sample(1:5,3)
> sam10=sample(1:5,3)
> y1=Q1[1,sam1]
> y2=Q1[2,sam2]
> y3=Q1[3,sam3]
> y4=Q1[4,sam4]
> y5=Q1[5,sam5]
> y6=Q1[6,sam6]
> y7=Q1[7,sam7]
> y8=Q1[8,sam8]
> y9=Q1[9,sam9]
> y10=Q1[10,sam10]
> pi=c(mean(y1),mean(y2),mean(y3),mean(y4),mean(y5),mean(y6),mean(y7),mean(y8),mean(y9),mean(y10))
> phat=sum(Mi*pi)/sum(Mi)
> phat
[1] 0.4333333
```

```
> sr2=sum(Mi^2*(pi-phat)^2)/(n-1)
> sr2
[1] 5.694444
> N=50
> M=150
> n=3
> qi=1-pi
> Mbar=mean(Mi)
> Mbar
[1] 5
> Vhat=(N-n)*sr2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)/Mi*pi*qi/(mi-1))/(n*N*Mbar^2)
> Vhat
[1] 0.07403704
> B=2*sqrt(Vhat)
> B
[1] 0.544195
```

Ans: $\hat{p} = 0.4333333$; $B = 0.544195$

2) Use the sample data from 1) to estimate the variance that would be obtained if drawing 6 rows and then 5 cells from each selected row.

```
> TreeGrid2 <- matrix(TreeGrid, nrow=N, ncol=M, byrow=T)
> n=6
> m=5
> y=matrix(0,n,m)
> sam1=sample(N,n)
> for (i in 1:n) {
+
+    sam2 <- sample(M,m)
+    y[i,] <- TreeGrid2[ sam1[i], sam2 ]
+ }
> y=c(t(y))
> cluster.y=rep(1:n,each=m)
> cluster.y=factor(cluster.y)
> anova.sam=anova(lm(y~cluster.y))
> anova.sam
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
cluster.y  5 1.7667 0.35333  1.6308 0.1901
Residuals 24 5.2000 0.21667
> MSb.y=anova.sam[1,3]
> MSw.y=anova.sam[2,3]
> Vhat <- (1-n/N)/n *(MSb.y/m) + (1-m/M)/(N*m) *MSw.y
> Vhat
[1] 0.009422222
```

Ans: $V = 0.009422222$

3) Draw 3 columns, and then draw 10 cells from each selected column.

```
> random_num=sample(5,3)
> random_num
[1] 4 1 2
> Q3=cbind(TreeGrid[,random_num])
> y1=Q3[sam1,1]
> y2=Q3[sam2,2]
```

```
> y3=Q3[sam3,3]
> pi=c(mean(y1),mean(y2),mean(y3))
> Mi=rep(30,3)
> phat=sum(Mi*pi)/sum(Mi)
> phat
[1] 0.3333333
> n=10
> sr2=sum(Mi^2*(pi-phat)^2)/(n-1)
> sr2
[1] 20.66667
> Vhat=(N-n)*sr2/(N*n*Mbar^2)+sum(Mi^2*(Mi-mi)/Mi*pi*qi/(mi-1))/(n*N*Mbar^2)
> Vhat
[1] 0.01714667
> B=2*sqrt(Vhat)
> B
[1] 0.2618906
```

Ans: $\hat{p} = 0.3333333; B = 0.2618906$

4) Use the sample data from 3) to estimate the variance that would be obtained if using stratified sampling: drawing 6 cells from each of the 5 columns.

```
> MSb.est=M*(MSb.y/m-(1-m/M)*MSw.y/m)
> MSw.est=MSw.y
> n=6
> m=5
> (1-n/N)/n *(MSb.est/M) + (1-m/M)/(m*n) *MSw.est
[1] 0.009422222
```

Ans: $V = 0.009422222$