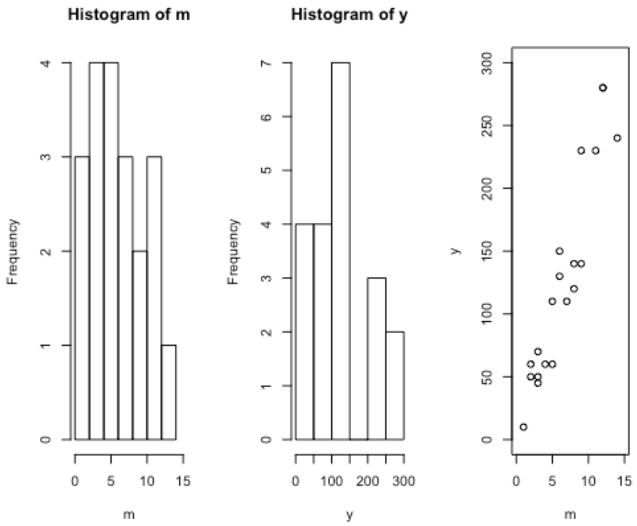
8.2 A manufacturer of band saws wants to estimate the average repair cost per month for the saws he has sold to certain industries. He cannot obtain a repair cost for each saw, but he can obtain the total amount spent for saw repairs and the number of saws owned by each industry. Thus, he decides to use cluster sampling, with each industry as a cluster. The manufacturer selects a simple random sample of $n = 20$ from the $N = 96$ industries he services. The data on total cost of repairs per industry and number of saws per industry are as given in the accompanying table. Estimate the average repair cost per saw for the past month and place a bound on the error of estimation.

Industry	Number of saws	Total repair cost for past month (dollars)
1	3	50
2	7	110
3	11	230
4	9	140
5	2	60
6	12	280
7	14	240
8	3	45
9	5	60
10	9	230
11	8	140
12	6	130
13	3	70
14	2	50
15	1	10
16	4	60
17	12	280
18	6	150
19	5	110
20	8	120

```
> rm(list=ls())
> exercise8.2 <- read.delim("~/Documents/Rutgers/Spring 2020/Stat 476/Homework/Homework 05/e
xercise8.2.txt")
>   View(exercise8.2)
> data=exercise8.2
> N=96
> n=20
> m=data$Saws
> y=data$RepairCo
> par(mfrow=c(1,3))
> hist(m,xlim=c(0,15))
> hist(y,xlim=c(0,300))
> plot(m,y,xlim=c(0,15),ylim=c(0,300))
```




```
> mbar=mean(m)
> ybar=mean(y)
> c(mbar,ybar)
[1]   6.50 128.25
> sm2=var(m)
> sy2=var(y)
> sqrt(c(sm2,sy2))
[1]  3.79057 83.11810
> smy=cov(m,y)
> smy
[1] 295.3947
> r=sum(y)/sum(m)
> r
[1] 19.73077
> sr2=sy2+r^2*sm2-2*r*smy
> sr2
[1] 845.5607
> Vhat=(1-n/N)/n*sr2/mean(m)^2
> Vhat
[1] 0.792192
> B=2*sqrt(Vhat)
> B
[1] 1.780103
```

Ans: $\bar{y}_r = 19.73077, B = 1.780183$

8.3 For the data in Exercise 8.2, estimate the total amount spent by the 96 industries on band saw repairs. Place a bound on the error of estimation.

```
> N*sum(y)/n
[1] 12312
> ybarT=sum(y)/20
> ybarT
[1] 128.25
> Vhat=N^2*(N-n)/(N*n)*(sum(c((y-ybarT)^2))/(n-1))
> Vhat
[1] 2520264
> B=2*sqrt(Vhat)
> B
[1] 3175.068
```

Ans: $\hat{\tau} = 12312; B = 3175.068$

8.4 After checking his sales records, the manufacturer in Exercise 8.2 finds that he sold a total of 710 band saws to these industries. Using this additional information, estimate the total amount spent on saw repairs by these industries and place a bound on the error of estimation.

```
> M=710
> totalPopHat=M*sum(y)/sum(m)
> totalPopHat
[1] 14008.85
> sr2=sum((y-m*r)^2)/(20-1)
> Vhat=N^2(1-n/N)*sr2/n
Error: attempt to apply non-function
> Vhat=N^2*(1-n/N)*sr2/n
> B=2*sqrt(Vhat)
> B
[1] 1110.785
```

Ans: $M\bar{y} = 14008.85; B = 1110.785$

8.5 The same manufacturer (Exercise 8.2) wants to estimate the average repair cost per saw for the next month. How many clusters should he select for his sample if he wants the bound on the error of estimation to be less than $2?

```
> D=2^2*(M/N)^2/4
> D
[1] 54.69835
> nhat=N*sr2/(N*D+sr2)
> nhat
[1] 13.3146
```

Ans: $n = 14$

8.8 An industry is considering revision of its retirement policy and wants to estimate the proportion of employees that favor the new policy. The industry consists of 87 separate plants located throughout the United States. Because results must be obtained quickly and with little cost, the industry decides to use cluster sampling with each plant as a cluster. A simple random sample of 15 plants is selected, and the opinions of the employees in these plants are obtained by questionnaire. The results are as shown in the accompanying table. Estimate the proportion of employees in the industry who favor the new retirement policy and place a bound on the error of estimation.

Plant	Number of employees	Number favoring new policy
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63
6	48	31
7	65	38
8	49	30
9	73	54
10	61	45
11	58	51
12	52	29
13	65	46
14	49	37
15	55	42

```
> rm(list=ls())
> exercise8.8 <- read.delim("~/Documents/Rutgers/Spring 2020/Stat 476/Homework/Homework 05/exercise8.8.txt")
>   View(exercise8.8)
> data=exercise8.8
> N=87
> n=15
> m=data$Employee
> y=data$Favor
> mbar=mean(m)
> ybar=mean(y)
> c(mbar,ybar)
[1] 60.73333 43.06667
> sm2=var(m)
> sy2=var(y)
> sqrt(c(sm2,sy2))
[1] 14.007481  9.572779
> smy=cov(m,y)
> smy
[1] 106.8762
> r=ybar/mbar
> r
[1] 0.7091109
> Vhat=(1-n/N)/n*var(y-r*m)/mbar^2
> Vhat
[1] 0.0005792499
> B=2*sqrt(Vhat)
> B
[1] 0.04813522
```

Ans: $\hat{p} = 0.7091109; B = 0.04813522$

8.9 The industry in Exercise 8.8 modified its retirement policy after obtaining the results of the survey. It now wants to estimate the proportion of employees in favor of the modified policy. How many plants should be sampled to have a bound of 0.08 on the error of estimation? Use the data from Exercise 8.8 to approximate the results of the new survey.

```
> N*(sy2+r^2*sm2-2*r*smy)/(N*(0.08^2*mbar^2/4)+(sy2+r^2*sm2-2*r*smy))
[1] 6.101613
```

Ans: $n = 7$

8.20 Block statistics report the number of housing units, the number of residents, and the total number of rooms within housing units for a random sample of eight blocks selected from a large city. (Assume the number of blocks in the city is very large.) The data are given in the accompanying table.

Block	Number of housing units	Number of residents	Number of rooms
1	12	40	58
2	14	39	72
3	3	12	26
4	20	52	98
5	12	37	74
6	8	33	57
7	10	41	76
8	6	14	48

a) Estimate the average number of residents per housing unit and place a bound on the error of estimation.

```
> rm(list=ls())
> exercise8.20 <- read.delim("~/Documents/Rutgers/Spring 2020/Stat 476/Homework/Homework 05/
exercise8.20.txt")
>   View(exercise8.20)
> data=exercise8.20
> n=8
> m=data$Housing
> y=data$Resident
> r=sum(y)/sum(m)
> r
[1] 3.152941
> sm2=var(m)
> sy2=var(y)
> smy=cov(m,y)
> sr2=sy2+r^2*sm2-2*r*smy
> sr2
[1] 47.84562
> mean(m)
[1] 10.625
> Vhat=sr2/(8*mean(m)^2)
> Vhat
[1] 0.05297784
> B=2*sqrt(Vhat)
> B
[1] 0.4603383
```

Ans: $\widehat{\mathbf{resident/housing}} = \mathbf{3.152941}, \boldsymbol{B} = \mathbf{0.4603383}$

b) Estimate the average number of rooms per resident and place a bound on the error of estimation.

```
> m=data$Resident
> y=data$Rooms
> r=sum(y)/sum(m)
> r
[1] 1.899254
> sm2=var(m)
> sy2=var(y)
> smy=cov(m,y)
> sr2=sy2+r^2*sm2-2*r*smy
> sr2
[1] 120.8712
> mean(m)
[1] 33.5
> Vhat=sr2/(n*mean(m)^2)
> Vhat
[1] 0.01346305
> B=2*sqrt(Vhat)
> B
[1] 0.2320607
```

Ans: $\widehat{\mathbf{room/resident}} = \mathbf{1.899254}; \boldsymbol{B} = \mathbf{0.2320607}$

8.21 A certain type of circuit board manufactured for installation in computers has 12 microchips per board. During the quality control inspection of ten of these boards, the numbers of defective microchips on each of the ten boards were as follows: 2, 0, 1, 3, 2, 0, 0, 1, 3, 4. Estimate the proportion of defective microchips in the population from which this sample was drawn and place a bound on the error of estimation.

```
> n=10
> Mbar=12
> a=c(2,0,1,3,2,0,0,1,3,4)
> phat=sum(a)/(n*Mbar)
> phat
[1] 0.1333333
> sp2=sum(c((a-phat*12)^2))/(n-1)
> sp2
[1] 2.044444
> Vhat=sp2/(n*Mbar^2)
> Vhat
[1] 0.001419753
> B=2*sqrt(Vhat)
> B
[1] 0.07535922
```

Ans: $\hat{p} = 0.1333333, B = 0.07535922$

8.22 Refer to Exercise 8.21. Suppose the sample of ten boards used there came from a shipment of 50 such boards. Estimate the total number of defective microchips in the shipment and place a bound on the error of estimation.

```
> n=10
> N=50
> M=50*12
> a=c(2,0,1,3,2,0,0,1,3,4)
> m=c(12,12,12,12,12,12,12,12,12,12)
> r=sum(a)/sum(m)
> r
[1] 0.1333333
> M*r
[1] 80
> Vhat=N^2*(1-n/N)/n*var(a-r*m)
> Vhat
[1] 408.8889
> B=2*sqrt(Vhat)
> B
[1] 40.442
```

Ans: $\hat{\tau} = 80; B = 40.442$

Sampling from Real Populations:

8.4 Think of the accompanying grid of 0s and 1s as a rough aerial map of a planted forest in which the 1s represent diseased trees. You can see something of the pattern of the trees, but assure that you cannot count them accurately from the air. You need to conduct a ground survey to estimate the proportion of diseased trees. From the 150 trees in the forest you are to sample 30 trees from which to construct your estimate. Design and carry out a sample survey and complete the analysis (estimate the proportion and calculate a margin of error) for each of the following designs:
a. Simple random sample

```
> set.seed(0)
> sam=sample(150,30)
> sam
 [1] 135  40  56  85 133  30 130 136  94  89   9  29  25  95  53 104  67  96 131  50
[21] 102 121  28  83  16  34  48   2  47 106
> select=rep(FALSE,150)
> select[sam]=TRUE
> 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[sam]
 [1] 1 1 0 1 1 0 1 1 1 1 1 0 1 1 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0
> phat=sum(TreeGrid[sam])/30
> phat
[1] 0.5
> Vhat=(1-30/150)*phat*(1-phat)/(30-1)
> Vhat
[1] 0.006896552
> B=2*sqrt(Vhat)
> B
[1] 0.166091
```

Ans: $\hat{p} = 0.5; B = 0.166091$

c. Stratified random sample with either rows of columns as strata, explaining your choice

```
      [,1] [,2] [,3] [,4] [,5]
 [1,]    0    0    0    1    1
 [2,]    0    0    1    1    1
 [3,]    0    0    0    0    0
 [4,]    0    0    1    0    1
 [5,]    0    0    0    0    1
 [6,]    1    0    1    0    0
 [7,]    0    0    1    0    1
 [8,]    0    1    1    0    1
 [9,]    0    0    0    1    1
[10,]    0    0    0    1    1
[11,]    0    0    0    1    0
[12,]    0    1    1    1    1
[13,]    0    1    0    1    1
[14,]    0    0    0    1    1
[15,]    0    1    0    1    1
[16,]    0    0    1    1    1
[17,]    0    0    1    1    1
[18,]    0    0    0    1    1
[19,]    0    0    1    1    1
[20,]    0    0    0    0    0
[21,]    1    0    0    1    0
[22,]    0    1    0    0    1
[23,]    0    0    1    1    0
[24,]    0    0    0    1    1
[25,]    0    0    0    1    1
[26,]    0    1    0    0    1
[27,]    0    0    1    0    1
[28,]    1    0    0    0    1
[29,]    0    0    0    0    1
[30,]    0    0    1    1    1
> sam1=sample(30,6)
> sam2=sample(30,6)
> sam3=sample(30,6)
> sam4=sample(30,6)
> sam5=sample(30,6)
```

```
> C1=TreeGrid[,1][sam1]
> C1
[1] 0 0 0 0 0 0
> C2=TreeGrid[,2][sam2]
> C2
[1] 0 0 0 0 1 0
> C3=TreeGrid[,3][sam3]
> C3
[1] 0 0 0 1 0 0
> C4=TreeGrid[,4][sam4]
> C4
[1] 1 0 1 1 1 0
> C5=TreeGrid[,5][sam5]
> C5
[1] 1 0 1 1 0 0
> phat=mean(c(C1,C2,C3,C4,C5))
> phat
[1] 0.3
> Vhat=(30/150)^2*(1-6/30)*(var(C1)+var(C2)+var(C3)+var(C4)+var(C5))/6
> Vhat
[1] 0.0048
> B=2*sqrt(Vhat)
> B
[1] 0.1385641
```

Ans: $\widehat{p}_{st} = 0.3; B = 0.1385641$

d. Cluster sample with either rows or columns as clusters, explaining your choice

```
> #pick the first 6 rows from all five columns
> C1=TreeGrid[,1][1:6]
> C1
[1] 0 0 0 0 0 1
> C2=TreeGrid[,2][1:6]
> C2
[1] 0 0 0 0 0 0
> C3=TreeGrid[,3][1:6]
> C3
[1] 0 1 0 1 0 1
> C4=TreeGrid[,4][1:6]
> C4
[1] 1 1 0 0 0 0
> C5=TreeGrid[,5][1:6]
> C5
[1] 1 1 0 1 1 0
> y=c(C1,C2,C3,C4,C5)
> r=mean(y)
> r
[1] 0.3333333
> sy2=var(y)
> sy2
[1] 0.2298851
> Vhat=(30/150)^2*sy2/30
> Vhat
[1] 0.0003065134
> B=2*sqrt(Vhat)
> B
[1] 0.03501505
```

Ans: $\hat{p}_c = 0.3333333; B = 0.03501505$

Compare your results with those of other students, and discuss which sampling design you think is best.

```
> p_actual=mean(TreeGrid)
> p_actual
[1] 0.42
> v_actual=(p_actual)*(1-p_actual)
> v_actual
[1] 0.2436
```

```
> Vhat_srs=0.006896552
> Vhat_st=0.0048
> Vhat_cluster=0.0003065134
> # SRS vs stratified
> Vhat_st/Vhat_srs
[1] 0.696
> # SRS vs cluster
> Vhat_cluster/Vhat_srs
[1] 0.04444444
> # stratified vs cluster
> Vhat_cluster/Vhat_st
[1] 0.06385696
```

Conclusion: cluster sampling yields not only an estimate proportion closest to the actual proportion but also the smallest variance.