5.2 A corporation desires to estimate the total number of worker-hours lost, for a given month, because of accidents among all employees. Because laborers, technicians, and administrators have different accident rates, the researcher decides to use stratified random sampling, with each group forming a separate stratum. Data from previous years suggest the variances shown in the accompanying table for the number of worker-hours lost per employee in the three groups, and current data give the stratum sizes. Determine the Neyman allocation for a sample of *n* = 30 employees.

I (laborers)	II (technicians)	III (administrators)
$\sigma_1^2 = 36$	$\sigma_2^2 = 25$	$\sigma_3^2 = 9$
$N_1 = 132$	$N_2 = 92$	$N_3 = 27$

$$\sum_{i=1}^{L} N_i\sigma_i = 132(6) + 92(5) + 27(3) = 1333; n_1 = 30\left(\frac{132\sqrt{36}}{1333}\right) = 17.82 \approx 18;$$

$$n_2 = 30\left(\frac{92\sqrt{25}}{1333}\right) = 10.35 \approx 10; n_3 = 30\left(\frac{27\sqrt{9}}{1333}\right) = 1.82 \approx 2; 18 + 10 + 2 = 30$$

Ans: n_1 = 18, n_2 = 10, n_3 = 2

5.5 A corporation wishes to obtain information on the effectiveness of a business machine. A number of division heads will be interviewed by telephone and asked to rate the equipment on a numerical scale. The divisions are located in North America, Europe, and Asia. Hence, stratified sampling is used. The costs are larger for interviewing division heads located outside North America. The accompanying table gives the costs per interview, approximate variances of the ratings, and *N* that have been established. The corporation wants to estimate the average rating with $V(\bar{y}_{st}) = 0.1$. Choose the sample size *n* that achieves this bound, and find the appropriate allocation.

Stratum I (North America)	Stratum II (Europe)	Stratum III (Asia)
$c_1 = \$9$	$c_2 = \$25$	$c_3 = \$36$
$\sigma_1^2 = 2.25$	$\sigma_2^2 = 3.24$	$\sigma_3^2 = 3.24$
$N_1 = 112$	$N_2 = 68$	$N_3 = 39$

Solution: *N* = 112 + 68 + 39 = 219

$$\sum_i \left(\frac{N_i\tilde{\sigma}_i}{\sqrt{c_i}}\right) = \frac{112\sqrt{2.25}}{\sqrt{9}} + \frac{68\sqrt{3.24}}{\sqrt{25}} + \frac{39\sqrt{3.24}}{\sqrt{36}} = 92.18$$

$$\sum_i N_i\tilde{\sigma}_i\sqrt{c_i} = 112\sqrt{2.25}\sqrt{9} + 68\sqrt{3.24}\sqrt{25} + 39\sqrt{3.24}\sqrt{36} = 1537.2$$

$$\sum_i N_i\tilde{\sigma}_i^2 = 112(2.25) + 68(3.24) + 39(3.24) = 598.68$$

$$n = \frac{92.18(1537.2)}{219^2(0.1) + 598.68} = 26.3 \approx 27; n_1 = 27\frac{112\sqrt{2.25}/\sqrt{9}}{92.18} = 16.40 \approx 16;$$

$$n_2 = 27\frac{68\sqrt{3.24}/\sqrt{25}}{92.18} = 7.17 \approx 7;\ n_3 = 27\frac{39\sqrt{3.24}/\sqrt{36}}{92.18} = 3.43 \approx 3$$

$16 + 7 + 3 = 26$; round up n_3 because 3.43 is closer to the next higher integer than n_1 and n_2.

Ans: *n* = 27, n_1 = 16, n_2 = 7, n_3 = 4

5.13 A county government is interested in expanding the facilities of a day-care center for mentally retarded children. The expansion will increase the cost of enrolling a child in the center. A sample survey will be conducted to estimate the proportion of families with retarded children that will make use of the expanded facilities. The families are divided into those who use the existing facilities and those who do not. Some families live in the city in which the center is located, and some live in the surrounding suburban and rural areas. Thus, stratified random sampling is used, with users in the city, users in the surrounding county, nonusers in the city, and nonusers in the country forming strata 1, 2, 3, and 4, respectively. Approximately 90% of the present users and 50% of the present nonusers will user the expanded facilities. The cost of obtaining an observation from a user is \$4 and from a nonuser is \$8. The difference in cost results because nonusers are difficult to locate.

Existing records give $N_1 = 97$, $N_2 = 43$, $N_3 = 145$, and $N_4 = 68$. Find the appropriate sample size and allocation necessary to estimate the population proportion with a bound of 0.05 on the error of estimation.

Solution: $N = 97 + 43 + 145 + 68 = 353$; $p_1 = p_2 = 0.9$, $p_3 = p_4 = 0.5$; $c_1 = c_2 = 4$, $c_3 = c_4 = 8$

$$\frac{N_1\tilde{\sigma}_1}{\sqrt{c_1}} = \frac{97\sqrt{0.9(0.1)}}{\sqrt{4}} = 14.55; \frac{N_2\tilde{\sigma}_2}{\sqrt{c_2}} = \frac{43\sqrt{0.9(0.1)}}{\sqrt{4}} = 6.45;$$

$$\frac{N_3\tilde{\sigma}_3}{\sqrt{c_3}} = \frac{145\sqrt{0.5(0.5)}}{\sqrt{8}} = 25.63; \frac{N_4\tilde{\sigma}_4}{\sqrt{c_4}} = \frac{68\sqrt{0.5(0.5)}}{\sqrt{8}} = 12.02$$

$$\sum_i \left(\frac{N_i\tilde{\sigma}_i}{\sqrt{c_i}}\right) = \sum_i N_i\sqrt{\frac{p_i(1-p_i)}{c_i}} = 14.55 + 6.45 + 25.63 + 12.02 = 58.65$$

$$w_1 = \frac{14.55}{58.65} = .248; w_2 = \frac{6.45}{58.65} = .110; w_3 = \frac{25.63}{58.65} = .437; w_4 = \frac{12.02}{58.65} = .205$$

$$\frac{N_1^2\tilde{\sigma}_1^2}{w_1} = \frac{97^2(0.9)(0.1)}{0.248} = 3414.56; \frac{N_2^2\tilde{\sigma}_2^2}{w_2} = \frac{43^2(0.9)(0.1)}{0.110} = 1512.82$$

$$\frac{N_3^2\tilde{\sigma}_3^2}{w_3} = \frac{145^2(0.5)(0.5)}{0.437} = 12028.03; \frac{N_4^2\tilde{\sigma}_4^2}{w_4} = \frac{68^2(0.5)(0.5)}{0.205} = 5639.02$$

$$N_1\tilde{\sigma}_1^2 = 97(0.9)(0.1) = 8.73; N_2\tilde{\sigma}_2^2 = 43(0.9)(0.1) = 3.87$$

$$N_3\tilde{\sigma}_3^2 = 145(0.5)(0.5) = 36.25; N_4\tilde{\sigma}_4^2 = 68(0.5)(0.5) = 17$$

$$D = \frac{0.05^2}{4} = 0.000625; n = \frac{3414.56 + 1512.82 + 12028.03 + 5639.02}{353^2(0.000625) + 8.73 + 3.87 + 36.25 + 17} = 157.20 \approx 158$$

$$n_1 = 158(.248) = 39.18 \approx 39; n_2 = 158(.110) = 17.38 \approx 17$$

$$n_3 = 158(.437) = 69.05 \approx 69; n_4 = 158(.205) = 32.39 \approx 32$$

$39 + 17 + 69 + 32 = 157$; round up n_4 to 33 because n_4 is the closest to the next higher integer.

Ans: $n = 158$, $n_1 = 39$, $n_2 = 17$, $n_3 = 69$, $n_4 = 33$

5.14 The survey in Exercise 5.13 is conducted and. yields and following proportion of families who will use the new facilities:

$$\hat{p}_1 = 0.87 \quad \hat{p}_2 = 0.93 \quad \hat{p}_3 = 0.60 \quad \hat{p}_4 = 0.53$$

Estimate the population proportion p, and place a bound on the error of estimation. Was the desired bound achieved?

$$\hat{p}_{st} = \frac{1}{N}\sum_{i=1}^{L} N_i\hat{p}_i = \frac{97(0.87) + 43(0.93) + 145(0.60) + 68(0.53)}{353} = 0.70$$

$$\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2\left(1 - \frac{n_i}{N_i}\right)\left(\frac{\hat{p}_i(1-\hat{p}_i)}{n_i - 1}\right)$$

$$= \frac{1}{355^2}\left[97^2\left(1 - \frac{39}{97}\right)\left(\frac{.87(.13)}{39-1}\right) + 43^2\left(1 - \frac{17}{43}\right)\left(\frac{.93(.07)}{17-1}\right) + 145^2\left(1 - \frac{69}{145}\right)\left(\frac{.6(.4)}{69-1}\right) + 68^2\left(1 - \frac{33}{68}\right)\left(\frac{.53(.47)}{33-1}\right)\right] = 0.000625$$

$$\hat{p}_{st} \pm 2\sqrt{\hat{V}(\hat{p}_{st})} = 0.70 \pm 2\sqrt{0.000625} = 0.70 \pm 0.05$$

Ans: $\hat{p}_{st} = 0.70$; $\hat{p}_{st} \pm 2\sqrt{\hat{V}(\hat{p}_{st})} = 0.70 \pm 0.05$; Yes, the desired bound is achieved

5.15 Suppose in Exercise 5.13 that the total cost of sampling is fixed at \$400. Choose the sample size and allocation that minimizes the variance of the estimator $\hat{p}_{st}$ for this fixed cost.

$$\sum_i n_i c_i = 400; n\big(4(.248) + 4(.110) + 8(.437) + 8(.205)\big) = 400, n = 60.90 \approx 61$$

$n_1 = 61(.248) = 15.13 \approx 15; n_2 = 61(.110) = 6.71 \approx 7;$

$n_3 = 61(.437) = 26.67 \approx 27; n_4 = 61(.205) = 12.51 \approx 13;$

$15 + 7 + 27 + 13 = 62 > 61$; round down n_4 to 12.

$15(4) + 4(4) + 27(8) + 12(8) = 388 < 400$

Ans: *n* = 61, *n*₁ = 15, *n*₂ = 7, *n*₃ = 27, *n*₄ = 12

5.21 A quality control inspector must estimate the proportion of defective microcomputer chips coming from two different assembly operations. She knows that, among the chips in the lot to be inspected, 60% are from assembly operation A and 40% are from assembly operation B. In a random sample of 100 chips, 38 turn out to be from operation A and 62 from operation B. Among the sampled chips from operation A, six are defective. Among the sampled chips from operation B, ten are defective.

a. Considering only the sample random sample of 100 chips, estimate the proportion of defectives in the lot, and place a bound on the error of estimation.

$$\hat{p} = \frac{\sum y_i}{n} = \frac{6 + 10}{100} = 0.16; B = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = 2\sqrt{\frac{.16(.84)}{99}} = 0.074$$

Ans: $\hat{p} = 0.16; B = 0.074$

b. Stratifying the sample, after selection, into chips from operation A and B, estimate the proportion of defectives in the population, and place a bound on the error of estimation. Ignore the fpc in both cases. Which answer do you find more acceptable?

$$\hat{p}_{post} = \sum_{i} \frac{N_i}{N}\bar{p}_i = 0.6\frac{6}{38} + 0.4\frac{10}{62} = 0.16;$$

$$\hat{V}(\bar{y}_{post}) = \frac{1}{n}\left(1-\frac{n}{N}\right)\sum_{i=1}^{L}\frac{N_i}{N}s_i^2 + \frac{1}{n^2}\sum_{i=1}^{L}\left(1-\frac{n_i}{N_i}\right)s_i^2 \approx \frac{1}{n}\sum_{i}\frac{N_i}{N}s_i^2 + \frac{1}{n^2}\sum_{i}s_i^2$$

$$= \frac{1}{100}\left[0.6\left(\frac{6}{38}\right)\left(\frac{32}{38}\right) + 0.4\left(\frac{10}{62}\right)\left(\frac{52}{62}\right)\right] + \frac{1}{100^2}\left[\left(\frac{6}{38}\right)\left(\frac{32}{38}\right) + \left(\frac{10}{62}\right)\left(\frac{52}{62}\right)\right]$$

$$= 0.00137$$

$$B_{post} = 2\sqrt{\hat{V}(\bar{y}_{post})} = 2\sqrt{0.00137} = 0.074$$

Ans: $\hat{\boldsymbol{p}}_{\boldsymbol{post}} = \mathbf{0.16}; \boldsymbol{B}_{\boldsymbol{post}} = \mathbf{0.074}$; no good reason to poststratify

5.30 Wage earners in a large firm are stratified into management and clerical classes, the first having 300 and the second having 500 employees. To assess attitude on sick-leave policy, independent random samples of 100 workers each were selected, one sample from each of the classes. After the sample data were collected, the responses were divided according to gender. In the table of results, *a* = Number who like the policy; *b* = Number who dislike the policy; and *c* = Number who have no opinion on the policy.

	Management, $N_1 = 300$	Clerical, $N_2 = 500$	Total, $N = 800$
Male	$a = 60$ $b = 15$ $c = 5$	$a = 24$ $b = 4$ $c = 2$	110
Female	$a = 10$ $b = 7$ $c = 3$	$a = 42$ $b = 20$ $c = 8$	90
Total	$n_1 = 100$	$n_2 = 100$	$n = 200$

Find an estimate and an estimated variance of that estimate for each parameter listed:

a. Proportion of managers who like the policy

$$\bar{y}_{st} = \sum_{i}\frac{N_i}{N}\bar{y}_i = \frac{60+10}{100} = 0.7$$

b. Proportion of wage earners who like the policy

$$\bar{y}_{st} = \sum_{i}\frac{N_i}{N}\bar{y}_i = \frac{24+42}{100} = 0.66$$

c. Total number of female wage earners who dislike the policy

$$\bar{y}_{st} = \sum_{i}\frac{N_i}{N}\bar{y}_i = \frac{7+20}{90} = 0.3$$

d. Difference between the proportion of male managers who like the policy and the proportion of female managers who like the policy.

$$\frac{60}{60+15+5}-\frac{10}{10+7+3}=0.75-0.5=0.25$$

e. Difference between the proportion of managers who like the policy and the proportion of managers who dislike the policy

$$\frac{60+10}{100}-\frac{15+7}{100}=0.7-0.22=0.48$$

Ans: a) 0.7, b) 0.66, c) 0.3, d) 0.25, e) 0.48

5.43 Reed and Chagnon (1987) wanted to estimate the number of greater snow geese on Bylot Island, in Canada's Northwest Territories (the island is an important breeding ground for the bird). They gridded the island into 400 2 × 2 km plots (excluding areas that were patently not usable by geese). They divided the 400 plots into three strata (high, medium, and low quality), based on ecological factors that are known to be associated with goose abundance. Using their estimates of stratum SDs and sizes, determine the optimal allocation for a sample of 83 (their actual sample size).

Stratum	Stratum size	Sample size	Mean	SD
High quality	65	34	412.5	316.9
Medium quality	127	28	136.8	127.7
Low quality	208	21	16.2	30.5

$$\sum_{i=1}^{L} N_i\sigma_i = 65(316.9)+127(127.7)+208(30.5)=43160.4;$$

$$n_1=83\left(\frac{65(316.9)}{43160.4}\right)=39.61\approx 40; n_2=83\left(\frac{127(127.7)}{43160.4}\right)=31.19\approx 31;$$

$$n_3=83\left(\frac{208(30.5)}{43160.4}\right)=12.20\approx 12; 40+31+12=83$$

Ans: n_1 = 40, n_2 = 31, n_3 = 12

5.44 Using the data from Exercise 5.43, estimate the total number of geese on the island with a 95% CI.

$$\hat{y}=40(412.5)+31(136.8)+12(16.2)=20935.2$$

$$\hat{V}(\hat{p}_{st})=\sum_{i=1}^{L}\left(\frac{N_i}{N}\right)^2\left(1-\frac{n_i}{N_i}\right)\frac{\tilde{\sigma}_i^2}{n_i-1}$$

$$=\left(\frac{65}{400}\right)^2\left(1-\frac{40}{65}\right)\frac{316.9^2}{40-1}+\left(\frac{127}{400}\right)^2\left(1-\frac{31}{127}\right)\frac{127.7^2}{31-1}$$

$$+\left(\frac{208}{400}\right)^2\left(1-\frac{12}{208}\right)\frac{30.5^2}{12-1}=89.12$$

95% CI: $20935.2 \pm 2\sqrt{89.12}=20935.2\pm 18.88$

Ans: $\hat{y}=\mathbf{20935.2}$; 95% CI: $\mathbf{20935.2 \pm 18.88}$

5.45 Assuming optimal allocation, what sample size would be required to estimate goose abundance with a 10% relative margin of error?

$$w_1 = \frac{65(316.9)}{43160.4} = .477; w_2 = \frac{127(127.7)}{43160.4} = .378; w_3 = \frac{208(30.5)}{43160.4} = .147;$$

$$\frac{N_1^2\tilde{\sigma}_1^2}{w_1} = \frac{[65(316.9)]^2}{0.477} = 889514050.84; \frac{N_2^2\tilde{\sigma}_2^2}{w_2} = \frac{[127(127.7)]^2}{0.378} = 695820847.65$$

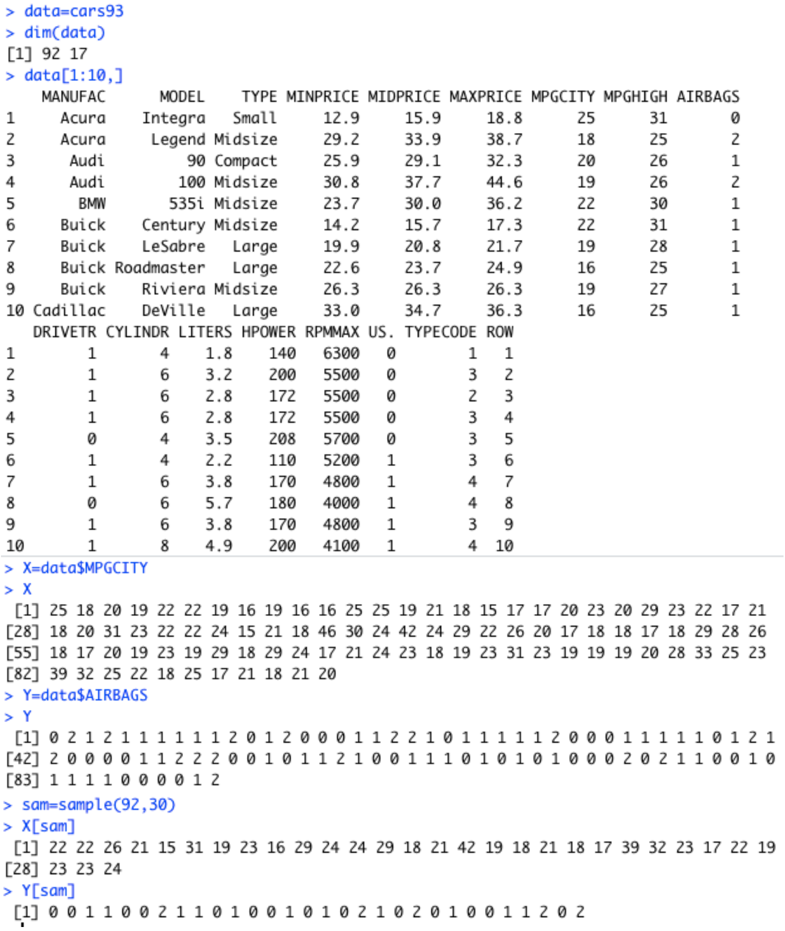
$$\frac{N_3^2\tilde{\sigma}_3^2}{w_3} = \frac{[208(30.5)]^2}{0.147} = 273784598.64$$

$$D = \frac{0.1^2}{4} = 0.0025; n = \frac{889514050.84 + 695820847.65 + 273784598.64}{400^2(0.0025) + 65(316.9)^2 + 127(127.7)^2 + 208(30.5)^2}$$

$$= \frac{1859119497.13}{8792582.48} = 211.44 \approx 212$$

Ans: *n* = 212

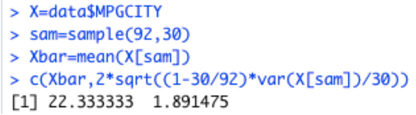
Sampling from Real Populations



```
> data=cars93
> dim(data)
[1] 92 17
> data[1:10,]
    MANUFAC      MODEL    TYPE MINPRICE MIDPRICE MAXPRICE MPGCITY MPGHIGH AIRBAGS
1     Acura    Integra   Small     12.9     15.9     18.8      25      31       0
2     Acura     Legend Midsize     29.2     33.9     38.7      18      25       2
3      Audi         90 Compact     25.9     29.1     32.3      20      26       1
4      Audi        100 Midsize     30.8     37.7     44.6      19      26       2
5       BMW       535i Midsize     23.7     30.0     36.2      22      30       1
6     Buick    Century Midsize     14.2     15.7     17.3      22      31       1
7     Buick    LeSabre   Large     19.9     20.8     21.7      19      28       1
8     Buick Roadmaster   Large     22.6     23.7     24.9      16      25       1
9     Buick    Riviera Midsize     26.3     26.3     26.3      19      27       1
10 Cadillac    DeVille   Large     33.0     34.7     36.3      16      25       1
   DRIVETR CYLINDR LITERS HPOWER RPMMAX US. TYPECODE ROW
1        1       4    1.8    140   6300   0        1   1
2        1       6    3.2    200   5500   0        3   2
3        1       6    2.8    172   5500   0        2   3
4        1       6    2.8    172   5500   0        3   4
5        0       4    3.5    208   5700   0        3   5
6        1       4    2.2    110   5200   1        3   6
7        1       6    3.8    170   4800   1        4   7
8        0       6    5.7    180   4000   1        4   8
9        1       6    3.8    170   4800   1        3   9
10       1       8    4.9    200   4100   1        4  10
> X=data$MPGCITY
> X
 [1] 25 18 20 19 22 22 19 16 19 16 16 25 25 19 21 18 15 17 17 20 23 20 29 23 22 17 21
[28] 18 20 31 23 22 22 24 15 21 18 46 30 24 42 24 29 22 26 20 17 18 18 17 18 29 28 26
[55] 18 17 20 19 23 19 29 18 29 24 17 21 24 23 18 19 23 31 23 19 19 19 20 28 33 25 23
[82] 39 32 25 22 18 25 17 21 18 21 20
> Y=data$AIRBAGS
> Y
 [1] 0 2 1 2 1 1 1 1 1 1 2 0 1 2 0 0 0 1 1 2 2 1 0 1 1 1 1 1 2 0 0 0 1 1 1 1 1 0 1 2 1
[42] 2 0 0 0 0 1 1 2 2 2 0 0 1 0 1 1 2 1 0 0 1 1 1 0 1 0 1 0 1 0 0 0 2 0 2 1 1 0 0 1 0
[83] 1 1 1 1 0 0 0 0 1 2
> sam=sample(92,30)
> X[sam]
 [1] 22 22 26 21 15 31 19 23 16 29 24 24 29 18 21 42 19 18 21 18 17 39 32 23 17 22 19
[28] 23 23 24
> Y[sam]
 [1] 0 0 1 1 0 0 2 1 1 0 1 0 0 1 0 1 0 2 1 0 2 0 1 0 0 1 1 2 0 2
```

5.3 The CARS93 data, in Appendix C, has cars classified as to being one of six different types, small, compact, midsize, large, sporty, or van. A numerical type code is given in the data set, in addition to the actual name of the type. The goal of this activity is to see if poststratification on car type pays any dividends when estimating average city gasoline mileage or proportion of cars with air bags for the cars in this population.

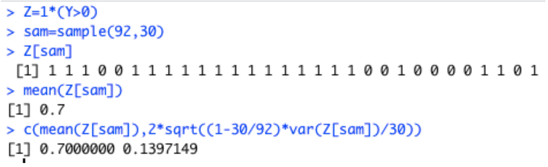
a. Select a random sample of cars from this population. Estimate the average city miles per gallon (mpg) for these cars, with a bound on the error of estimation.



```
> X=data$MPGCITY
> sam=sample(92,30)
> Xbar=mean(X[sam])
> c(Xbar,2*sqrt((1-30/92)*var(X[sam])/30))
[1] 22.333333  1.891475
```

Ans: 22.333333± 1.891475 mpg

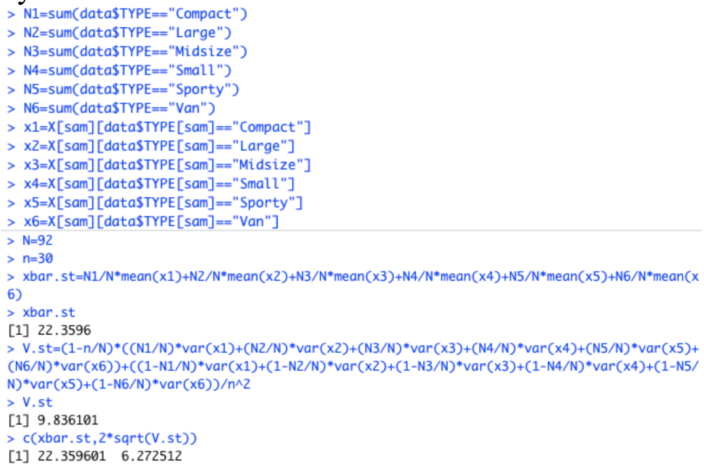
b. Estimate the proportion of these cars that have at least one air bag, with a bound on the error of estimation.



```
> Z=1*(Y>0)
> sam=sample(92,30)
> Z[sam]
 [1] 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 1 1 0 1
> mean(Z[sam])
[1] 0.7
> c(mean(Z[sam]),2*sqrt((1-30/92)*var(Z[sam])/30))
[1] 0.7000000 0.1397149
```

Ans: 0.70 ± 0.1397149

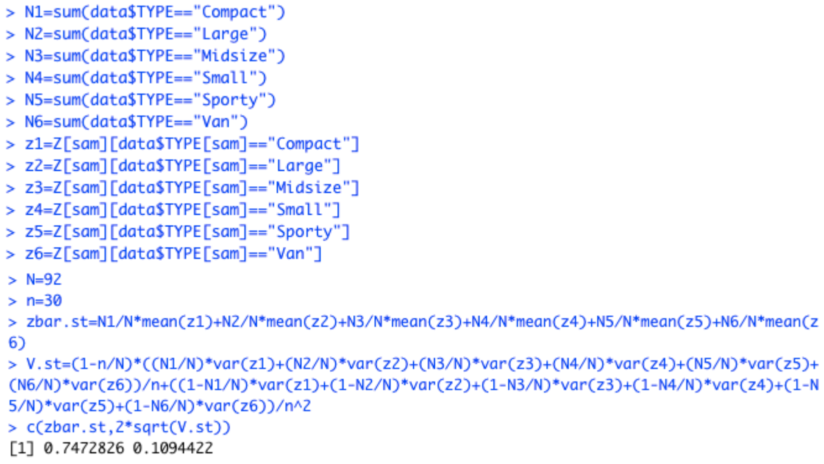
c. Using the data from part (a), poststratify on the car type and then estimate the average city mpg by this method.



```
> N1=sum(data$TYPE=="Compact")
> N2=sum(data$TYPE=="Large")
> N3=sum(data$TYPE=="Midsize")
> N4=sum(data$TYPE=="Small")
> N5=sum(data$TYPE=="Sporty")
> N6=sum(data$TYPE=="Van")
> x1=X[sam][data$TYPE[sam]=="Compact"]
> x2=X[sam][data$TYPE[sam]=="Large"]
> x3=X[sam][data$TYPE[sam]=="Midsize"]
> x4=X[sam][data$TYPE[sam]=="Small"]
> x5=X[sam][data$TYPE[sam]=="Sporty"]
> x6=X[sam][data$TYPE[sam]=="Van"]
> N=92
> n=30
> xbar.st=N1/N*mean(x1)+N2/N*mean(x2)+N3/N*mean(x3)+N4/N*mean(x4)+N5/N*mean(x5)+N6/N*mean(x
6)
> xbar.st
[1] 22.3596
> V.st=(1-n/N)*((N1/N)*var(x1)+(N2/N)*var(x2)+(N3/N)*var(x3)+(N4/N)*var(x4)+(N5/N)*var(x5)+
(N6/N)*var(x6))+((1-N1/N)*var(x1)+(1-N2/N)*var(x2)+(1-N3/N)*var(x3)+(1-N4/N)*var(x4)+(1-N5/
N)*var(x5)+(1-N6/N)*var(x6))/n^2
> V.st
[1] 9.836101
> c(xbar.st,2*sqrt(V.st))
[1] 22.359601  6.272512
```

Ans: 22.359601 ± 6.272512 mpg

d. Using the data from part (b), poststratify on car type and then estimate the proportion of cars that have at least one air bag by this method.



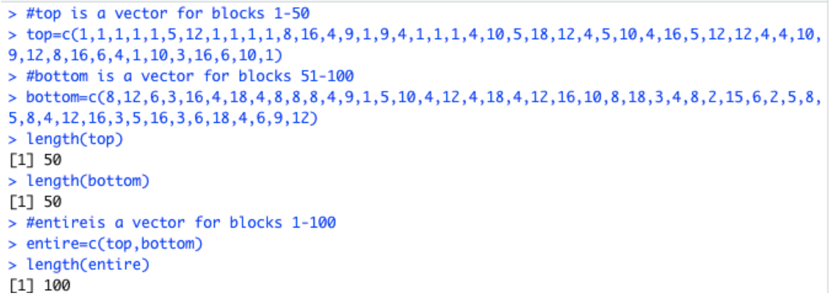
```
> N1=sum(data$TYPE=="Compact")
> N2=sum(data$TYPE=="Large")
> N3=sum(data$TYPE=="Midsize")
> N4=sum(data$TYPE=="Small")
> N5=sum(data$TYPE=="Sporty")
> N6=sum(data$TYPE=="Van")
> z1=Z[sam][data$TYPE[sam]=="Compact"]
> z2=Z[sam][data$TYPE[sam]=="Large"]
> z3=Z[sam][data$TYPE[sam]=="Midsize"]
> z4=Z[sam][data$TYPE[sam]=="Small"]
> z5=Z[sam][data$TYPE[sam]=="Sporty"]
> z6=Z[sam][data$TYPE[sam]=="Van"]
> N=92
> n=30
> zbar.st=N1/N*mean(z1)+N2/N*mean(z2)+N3/N*mean(z3)+N4/N*mean(z4)+N5/N*mean(z5)+N6/N*mean(z
6)
> V.st=(1-n/N)*((N1/N)*var(z1)+(N2/N)*var(z2)+(N3/N)*var(z3)+(N4/N)*var(z4)+(N5/N)*var(z5)+
(N6/N)*var(z6))/n+((1-N1/N)*var(z1)+(1-N2/N)*var(z2)+(1-N3/N)*var(z3)+(1-N4/N)*var(z4)+(1-N
5/N)*var(z5)+(1-N6/N)*var(z6))/n^2
> c(zbar.st,2*sqrt(V.st))
[1] 0.7472826 0.1094422
```

Ans: 0.7472826±0.1094422

e. Comparing the above results, comment on when poststratification might produce big gains in terms of the error of estimation.

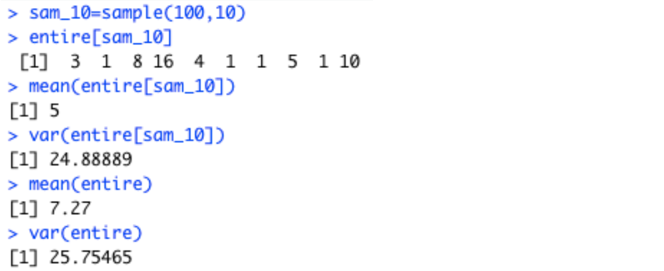
Ans: Poststratifying MPG does not show much benefit (larger bound); poststratifying proportion of cars with at least one air bag actually shows a smaller bound.

5.5. Return to the sheet of rectangles provided for the activity at the end of Chapter 4. Select a random sample of ten rectangles and use them to construct an estimate of the average area of the rectangles on the page. Then, select a stratified random sample choosing five rectangles from those numbered 1–50 and five from among those numbered 51–100. Does the stratification appear to provide any advantage? Compare your results with those of the rest of the class and comment on the general pattern.

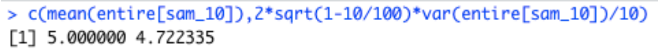


```
> #top is a vector for blocks 1-50
> top=c(1,1,1,1,1,5,12,1,1,1,1,8,16,4,9,1,9,4,1,1,1,4,10,5,18,12,4,5,10,4,16,5,12,12,4,4,10,
9,12,8,16,6,4,1,10,3,16,6,10,1)
> #bottom is a vector for blocks 51-100
> bottom=c(8,12,6,3,16,4,18,4,8,8,8,4,9,1,5,10,4,12,4,18,4,12,16,10,8,18,3,4,8,2,15,6,2,5,8,
5,8,4,12,16,3,5,16,3,6,18,4,6,9,12)
> length(top)
[1] 50
> length(bottom)
[1] 50
> #entireis a vector for blocks 1-100
> entire=c(top,bottom)
> length(entire)
[1] 100
```

Step 1: Select ten random rectangles out of 100

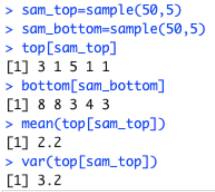


```
> sam_10=sample(100,10)
> entire[sam_10]
 [1]  3  1  8 16  4  1  1  5  1 10
> mean(entire[sam_10])
[1] 5
> var(entire[sam_10])
[1] 24.88889
> mean(entire)
[1] 7.27
> var(entire)
[1] 25.75465
```

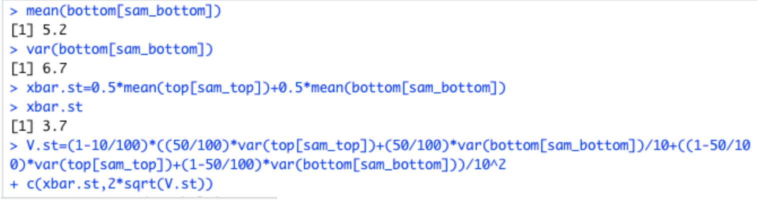


```
> c(mean(entire[sam_10]),2*sqrt(1-10/100)*var(entire[sam_10])/10)
[1] 5.000000 4.722335
```

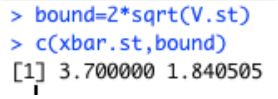
Step 2: Stratified sampling



```
> sam_top=sample(50,5)
> sam_bottom=sample(50,5)
> top[sam_top]
[1] 3 1 5 1 1
> bottom[sam_bottom]
[1] 8 8 3 4 3
> mean(top[sam_top])
[1] 2.2
> var(top[sam_top])
[1] 3.2
```



```
> mean(bottom[sam_bottom])
[1] 5.2
> var(bottom[sam_bottom])
[1] 6.7
> xbar.st=0.5*mean(top[sam_top])+0.5*mean(bottom[sam_bottom])
> xbar.st
[1] 3.7
> V.st=(1-10/100)*((50/100)*var(top[sam_top])+(50/100)*var(bottom[sam_bottom])/10+((1-50/10
0)*var(top[sam_top])+(1-50/100)*var(bottom[sam_bottom]))/10^2
+ c(xbar.st,2*sqrt(V.st))
```



```
> bound=2*sqrt(V.st)
> c(xbar.st,bound)
[1] 3.700000 1.840505
```

Ans: Poststratifying shows a smaller bound but a significantly smaller sample mean from population mean (7.27). This can be attributed to the small sample size.

Additional problem I: For stratified random sampling, let μ be the population mean, σ^2 be the population variance, and $S^2 = \frac{N}{N-1}\sigma^2$. Show that

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{L}\{(N_i - 1)S_i^2 + N_i(\mu_i - \mu)^2\}$$

where μ_i is the stratum population mean, σ_i^2 is the stratum population variance, and $S_i^2 = \frac{N_i}{N_i-1}\sigma_i^2$ for $i = 1, \ldots, L$.

Solution:

$$S_i^2 = \frac{N_i}{N_i - 1}\sigma^2; \mu = \sum_{i=1}^{L}\left(\frac{N_i}{N}\mu_i\right) \to \mu N = \sum_{i=1}^{L}(N_i\mu_i)$$

$$\frac{1}{N-1}\sum_{i=1}^{L}\{(N_i - 1)S_i^2 + N_i(\mu_i - \mu)^2\}$$

$$= \frac{1}{N-1}\sum_{i=1}^{L}\left\{(N_i - 1)\left(\frac{N_i}{N_i - 1}\sigma^2\right) + N_i\mu_i^2 - 2N_i\mu_i\mu + N_i\mu^2\right\}$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{L} N_i(\sigma^2 + \mu^2) + \sum_{i=1}^{L}(N_i\mu_i^2) - 2\mu\sum_{i=1}^{L} N_i\mu_i\right]$$

$$= \frac{1}{N-1}\left[\sum_{i=1}^{L} N_i(\sigma^2 + \mu^2) + N\mu^2 - 2N\mu^2\right] = \frac{1}{N-1}[N\sigma^2 + N\mu^2 - N\mu^2]$$

$$= \frac{N}{N-1}\sigma^2 = S^2$$