4.1 List all possible simple random samples of size $n = 2$ that can be selected from the population {0, 1, 2, 3, 4}. Calculate σ^2 for the population and $V(\bar{y})$ for the sample.

Sample	{0, 1}	{0, 2}	{0, 3}	{0, 4}	{1, 2}	{1, 3}	{1, 4}	{2, 3}	{2, 4}	{3, 4}
$\bar{y}$	0.5	1.0	1.5	2.0	1.5	2.0	2.5	2.5	3.0	3.5

The probability distribution of sample mean $\bar{y}$ is

$\bar{y}$	0.5	1.0	1.5	2.0	2.5	3.0	3.5
$p(\bar{y})$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

$$E(\bar{y}) = 0.5(0.1) + 1.0(0.1) + 1.5(0.2) + 2.0(0.2) + 2.5(0.2) + 3.0(0.1) + 3.5(0.1) = 2$$

$$E(\bar{y}^2) = (0.5)^2(0.1) + (1.0)^2(0.1) + (1.5)^2(0.2) + (2.0)^2(0.2) + (2.5)^2(0.2) + (3.0)^2(0.1) + (3.5)^2(0.1) = 4.75$$

$$V(\bar{y}) = E(\bar{y}^2) - E(\bar{y})^2 = 4.75 - 4 = 0.75$$

The probability distribution of population y is

y	0	1	2	3	4
$p(y)$	0.2	0.2	0.2	0.2	0.2

$$E(y) = 0(0.2) + 1(0.2) + 2(0.2) + 3(0.2) + 4(0.2) = 2$$

$$E(y^2) = 0^2(0.2) + 1^2(0.2) + 2^2(0.2) + 3^2(0.2) + 4^2(0.2) = 6$$

$$\sigma^2 = V(y) = E(y^2) - E(y) = 6 - 4 = 2$$

$$V(\bar{y}) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} = \frac{5-2}{5-1} \cdot \frac{2}{2} = 0.75$$

Ans: $\boldsymbol{\sigma^2 = 2, V(\bar{y}) = 0.75}$

4.2 For the simple random samples generated in Exercise 4.1, calculate s^2 for each sample. Show numerically that

$$E(s^2) = \frac{N}{N-1}\sigma^2$$

```
> y1 <- c(0,1)
> var(y1)
[1] 0.5
> y2 <- c(0,2)
> var(y2)
[1] 2
> y3 <- c(0,3)
> var(y3)
[1] 4.5
> y4 <- c(0,4)
> var(y4)
[1] 8
> y5 <- c(1,2)
> var(y5)
[1] 0.5
```

```
> y6 <- c(1,3)
> var(y6)
[1] 2
> y7 <- c(1,4)
> var(y7)
[1] 4.5
> y8 <- c(2,3)
> var(y8)
[1] 0.5
> y9 <- c(2,4)
> var(y9)
[1] 2
> y10 <- c(3,4)
> var(y10)
[1] 0.5
```

$$E(s^2) = \frac{1}{10}(0.5 + 2 + 4.5 + 8 + 0.5 + 2 + 4.5 + 0.5 + 2 + 0.5) = 2.5$$

$$\frac{N}{N-1}\sigma^2 = \frac{5}{5-1} \cdot 2 = 2.5 = E(s^2)$$

Ans: 0.5; 2; 4.5; 8; 0.5; 2; 4.5; 0.5; 2; 0.5

4.14 State park officials were interested in the proportion of campers who consider the campsite spacing adequate in a particular campground. They decided to take a simple random sample of $n = 30$ from the first $N = 300$ camping parties that visited the campground. Let $y_i = 0$ if the head of the ith party sampled does not think the campsite spacing is adequate and $y_i = 1$ if he does ($i = 1, 2, \ldots, 30$). Use the data in the accompanying table to estimate p, the proportion of campers who consider the campsite spacing adequate.

Camper sampled	Response, y_i
1	1
2	0
3	1
⋮	⋮
29	1
30	1
	$\sum_{i=1}^{30} y_i = 25$

$$\hat{p} = \bar{y} = \frac{25}{30} = 0.833$$

Ans: 0.833

4.15 Use the data in Exercise 4.14 to determine the sample size required to estimate p with a bound on the error of estimation of magnitude $B = 0.05$.

$$B = 0.5, D = \frac{B^2}{4} = \frac{(0.5)^2}{4} = 0.000625$$

$$n = \frac{Np(1-p)}{(N-1)D + p(1-p)} = \frac{300\left(\frac{5}{6}\right)\left(1-\frac{5}{6}\right)}{299(0.000625) + \left(\frac{5}{6}\right)\left(1-\frac{5}{6}\right)} = 127.90 \approx 128$$

Ans: 128

4.17 Use the data in Exercise 4.16, estimate the total number of gallons of water, τ, used daily during the dry spell. Place a bound on the error of estimation.

$n = 100, N = 10{,}000, \bar{y} = 12.5, s^2 = 1252$

$\hat{\tau} = N\bar{y} = 10000(12.5) = 125{,}000$

$$B = 2\sqrt{N^2\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)} = 2\sqrt{10000^2 \cdot \frac{1252}{100} \cdot \frac{10000-100}{10000}} = 70{,}412.50$$

Ans: *B* = 70,412.50

4.19 A dentist was interested in the effectiveness of a new toothpaste. A group of $N = 1000$ schoolchildren participated in a study. Prestudy records showed there was an average of 2.2 cavities every six months for the group. After three months of the study, the dentist sampled $n = 10$ children to determine how they were progressing on the new toothpaste. Using the data in the accompanying table, estimate the mean number of cavities for the entire group and place a bound on the error of estimation.

Child	Number of cavities in the three-month period
1	0
2	4
3	2
4	3
5	2
6	0
7	3
8	4
9	1
10	1

$$\bar{y} = \frac{\sum y_i}{n} = \frac{4+2+3+2+3+4+1+1}{10} = \frac{20}{10} = 2$$

$$s^2 = \frac{\sum y_i^2 - n\bar{y}^2}{n-1} = \frac{60 - 10(4)}{9} = \frac{20}{9} = 2.22, \hat{\mu} = \bar{y} = 2$$

$$B = 2\sqrt{\left(\frac{s^2}{n}\right)\left(\frac{N-n}{N}\right)} = 2\sqrt{\frac{2.22}{10} \cdot \frac{1000-10}{1000}} = 938$$

Ans: $\hat{\mu} = 2, B = 938$

4.30 (Multiple choice) A survey was conducted to determine what adults prefer in cell phone services. The results of the survey showed that 73% of the people wanted email service, with a margin of error of plus or minus 4%. What is meant by the phrase "plus or minus 4%"?

a. They estimate that 4% of the population that was surveyed may change their minds between the time the poll is conducted and the time the survey is published.

b. There is a 4% chance that the true percentage of adults who want email service will not be in the confidence interval of 69–77%.

c. Only 4% of the population was surveyed.

d. It would be unlikely to get the observed sample proportion of 73% unless the actual percentage of all adults who want email service is between 62% and 68%.

e. They probability that the sample proportion is in the confidence interval is .04.

Ans: B

4.44 The Major League Baseball season came to an abrupt end in the middle of 1994 due to a strike. In a poll of 600 adult Americans (*Time*, August 22, 1994), 29% blamed the players for this strike, 34% blamed the owners, and the rest held various other opinions. Does evidence suggest that the true proportions who blame players and owners, respectively, are really different?

$$(0.34 - 0.29) \pm 2\sqrt{\frac{(0.34)(0.66)}{600} + \frac{(0.29)(0.71)}{600} + 2\frac{(0.34)(0.29)}{600}} = 0.05 \pm 0.0647$$

Ans: There is no strong evidence that the two true proportions would differ.

A Sampling Activity—Random Rectangles

The goal is to choose a sample of five rectangles from which to estimate the average area of the 100 rectangles in the display.

1. Without studying the display of rectangles too carefully, quickly choose five that you think represent the population of rectangles on the page. This is your judgment sample.
Judgment sample: 12, 20, 53, 67, 80

2. Find the area of each rectangle (in terms of number of grid cells) in your sample of five and compute the sample mean, that is, the average area of the rectangles in your sample.
R12 <- 1, R29 <- 3, R53 <- 6, R67 <- 12, R80 <- 12
Sample mean = (1 + 3 + 6 + 12 + 12)/5 = 6.8

5. Now, generate five distinct random numbers between 00 and 99. (The rectangle numbered 100 can be called 00.) Find the rectangles that correspond to your random numbers. This is your random sample of five rectangles.

```
> N = 100
> n = 5
> pop = 1:N
> sample1 = sample(pop,n)
> sample1
[1] 70 31 45 79 18
> area1 <- c(10,4,5,12,2)
> ybar1<-mean(area1)
> ybar1
[1] 6.6
> sigmasq1<-1(n-1)*sum((area1-ybar1)^2)
Error: attempt to apply non-function
> sigmasq1<-1/(n-1)*sum((area1-ybar1)^2)
> sigmasq1
[1] 17.8
> vhat_ybar1=(N-n)/N*(sigmasq1/n)
> vhat_ybar1
[1] 3.382
> bound1=2*sqrt((N-n)/N*(vhat_ybar1/n))
> bound1
[1] 1.603222
> tHat1<-N*ybar1
> lower1=tHat1 - N*bound1
> upper1=tHat1 + N*bound1
> lower1
[1] 499.6778
> upper1
[1] 820.3222
```

Additional question: repeat Question 5 for 5 times, and report your 5 sample means and 5 confidence intervals. What proportions of your confidence intervals cover the truth (computed from the table in Question 8)?

```
> sample2=sample(pop,n)
> sample3=sample(pop,n)
> sample4=sample(pop,n)
> sample5=sample(pop,n)
> sample6=sample(pop,n)
> sample2
[1] 82 76 77  2 67
> sample3
[1]  7 39 55 88 38
> sample4
[1] 90 51  2  8 12
> sample5
[1]  7 48 37 92  8
> sample6
[1] 89 32 57 85 59
```

```
> area2<-c(12,12,12,1,9)
> area3<-c(1,5,9,16,5)
> area4<-c(16,8,1,1,1)
> area5<-c(1,8,5,16,1)
> area6<-c(16,4,9,15,9)
> ybar2<-mean(area2)
> ybar3<-mean(area3)
> ybar4<-mean(area4)
> ybar5<-mean(area5)
> ybar6<-mean(area6)
> ybar2
[1] 9.2
> ybar3
[1] 7.2
> ybar4
[1] 5.4
> ybar5
[1] 6.2
> ybar6
[1] 10.6
```

```
> sigmasq2<-1/(n-1)*sum((area2 - ybar2)^2)
> sigmasq3<-1/(n-1)*sum((area3 - ybar3)^2)
> sigmasq4<-1/(n-1)*sum((area4 - ybar4)^2)
> sigmasq5<-1/(n-1)*sum((area5 - ybar5)^2)
> sigmasq6<-1/(n-1)*sum((area6 - ybar6)^2)
> vhat_ybar2<-(N-n)/N*(sigmasq2/n)
> vhat_ybar3<-(N-n)/N*(sigmasq3/n)
> vhat_ybar4<-(N-n)/N*(sigmasq4/n)
> vhat_ybar5<-(N-n)/N*(sigmasq5/n)
> vhat_ybar6<-(N-n)/N*(sigmasq6/n)
> bound2<-2*sqrt((N-n)/N*(vhat_ybar2/n))
> bound3<-2*sqrt((N-n)/N*(vhat_ybar3/n))
> bound4<-2*sqrt((N-n)/N*(vhat_ybar4/n))
> bound5<-2*sqrt((N-n)/N*(vhat_ybar5/n))
> bound6<-2*sqrt((N-n)/N*(vhat_ybar6/n))
```

```
> tHat2<-N*ybar2
> tHat3<-N*ybar3
> tHat4<-N*ybar4
> tHat5<-N*ybar5
> tHat6<-N*ybar6
> lower2<-tHat2 - N*bound2
> lower3<-tHat3 - N*bound3
> lower4<-tHat4 - N*bound4
> lower5<-tHat5 - N*bound5
> lower6<-tHat6 - N*bound6
> upper2<-tHat2 + N*bound2
> upper3<-tHat3 + N*bound3
> upper4<-tHat4 + N*bound4
> upper5<-tHat5 + N*bound5
> upper6<-tHat6 + N*bound6
```

```
> CI2<-c(lower2,upper2)
> CI3<-c(lower3,upper3)
> CI4<-c(lower4,upper4)
> CI5<-c(lower5,upper5)
> CI6<-c(lower6,upper6)
> CI2
[1]  738.9508 1101.0492
> CI3
[1] 504.3688 935.6312
> CI4
[1] 287.0787 792.9213
> CI5
[1] 383.6046 856.3954
> CI6
[1]  872.6789 1247.3211
```

```
> totalArea<-c(rep(1,times=16),rep(2,times=2),rep(3,times=6),rep(4,times=16),rep(5,times=8),
rep(6,times=6),rep(8,times=8),rep(9,times=5),rep(10,times=7),rep(12,times=10),15,rep(16,time
s=10),rep(18,times=5))
> sum(totalArea)
[1] 742
```

Conclusion: only CI2, CI3, CI4, and CI5 cover the truth from Question 8.

Addition problem I: For simple random sampling, prove the following results.

1) Show that each pair of units, i and j ($i \neq j$), has probability $\frac{n(n-1)}{N(N-1)}$ of being selected.

PF: $P(i \text{ being selected}) = \frac{n}{N}, P(j \text{ being selected}) = \frac{n-1}{N-1}$

$$P(i \text{ and } j \text{ being selected}) = P(j \text{ being selected} \mid i \text{ being slected}) \cdot P(i \text{ being selected})$$

$$= \left(\frac{n-1}{N-1} \Big/ \frac{n}{N}\right) \cdot \frac{n}{N} = \frac{n-1}{N-1}$$

2) Show that

$$E(s^2) = \frac{N}{N-1}\sigma^2$$

where s^2 is the sample variance and σ^2 the population variance.

PF: $s^2 = \frac{1}{n-1}\sum_{i\in S}^{n}(y_i - \bar{y})^2$, $E(s^2) = E\left(\frac{1}{n-1}\right)\cdot E(\sum_{i\in S}^{n}(y_i - \bar{y})^2)$

$$E\left(\sum_{i\in S}^{n}(y_i - \bar{y})^2\right) = E\left(\sum_{i\in S}^{n}(y_i^2 - 2y_i\bar{y} + \bar{y}^2)\right) = E\left(\sum_{i\in S}^{n}(y_i^2) - 2\bar{y}\cdot\sum_{i\in S}^{n}y_i + n\bar{y}^2\right)$$

$$= E\left(\sum_{i\in S}^{n}(y_i^2) - n\left(\frac{2\bar{y}}{n}\cdot\sum_{i\in S}^{n}y_i - \bar{y}^2\right)\right) = E\left(\sum_{i\in S}^{n}(y_i^2) - n(2\bar{y}^2 - \bar{y}^2)\right)$$

$$= E\left(\sum_{i\in S}^{n}(y_i^2) - n\bar{y}^2\right)$$

$$\therefore E(s^2) = E\left(\frac{1}{n-1}\right)\cdot E\left(\sum_{i\in S}^{n}(y_i^2) - n\bar{y}^2\right) = \frac{1}{n-1}\cdot\left(E\left(\sum_{i\in S}^{n}(y_i^2)\right) - nE(\bar{y}^2)\right)$$

$$= \frac{1}{n-1}\cdot\left(\sum_{i\in S}^{n}(\mu^2 + \sigma^2) - n\left(\mu^2 + \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}\right)\right)$$

$$= \frac{1}{n-1}\cdot\left(n\mu^2 + n\sigma^2 - n\mu^2 - \frac{N-n}{N-1}\cdot\sigma^2\right) = \frac{1}{n-1}\cdot\left(n\sigma^2 - \frac{N-n}{N-1}\cdot\sigma^2\right)$$

$$= \frac{1}{n-1}\cdot\left(\frac{Nn\sigma^2 - n\sigma^2}{N-1} - \frac{N\sigma^2 - n\sigma^2}{N-1}\right) = \frac{1}{n-1}\cdot\left(\frac{Nn\sigma^2 - N\sigma^2}{N-1}\right)$$

$$= \frac{1}{n-1}\cdot\left(\frac{N\sigma^2(n-1)}{N-1}\right) = \frac{N}{N-1}\sigma^2$$

Additional problem II: The following table gives the distribution of the number of Apple products used by a class of 60 students:

# of Apple products	0	1	2	3	4
# of students	15	24	10	6	5

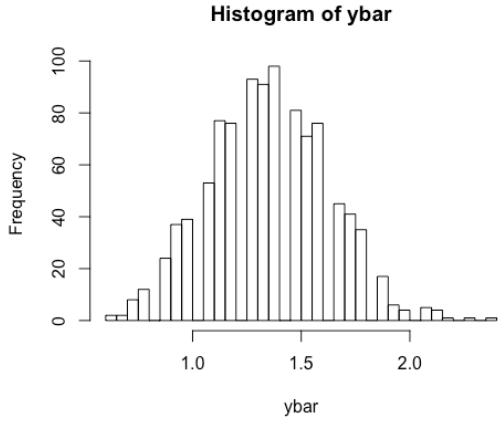
Use R to complete the following exercises. Please submit your R codes and R output showing your answers.

1) Draw a simple random sample of size 15, and compute the sample mean and confidence interval for the class mean number of Apple products used by the students.

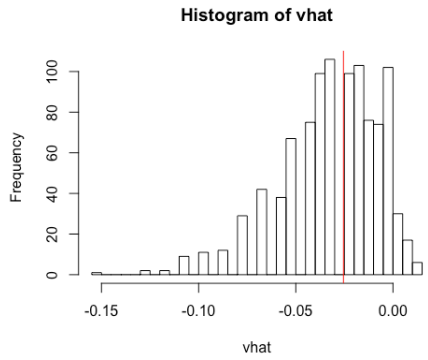
```
> N=60
> n=15
> X<-c(rep(0,times=15),rep(1,times=24),rep(2,times=10),rep(3,times=6),rep(4,times=5))
> samp<-sample(1:N,n)
> applesamp<-X[samp]
> applesamp
 [1] 1 4 3 1 0 1 1 0 1 1 1 2 0 0 3
> ybar<-mean(applesamp)
> ybar
[1] 1.266667
> ssquared<-var(applesamp)
> ssquared
[1] 1.495238
> bound<-2*(sqrt((1-(n/N))*ssquared/n))
> bound
[1] 0.5468525
> left = ybar - bound
> right = ybar + bound
> left
[1] 0.7198142
> right
[1] 1.813519
```

2) Repeat Exercise 1) for 1000 times, and plot a histogram of your 1000 sample means. Find the mean and variance of the 1000 sample means, and compare them to the theoretical mean and variance of the sample mean.

```
> ybar=rep(0,times=R)
> vhat=rep(0,times=R)
> se=rep(0,times=R)
> lower=rep(0,times=R)
> upper=rep(0,times=R)
> for (j in 1:R) {
+     samp=sample(1:N,n)
+     sampled<-X[samp]
+     ybar[j]=mean(sampled)
+     vhat[j]=(1-n/N)*ybar[j]*(1-ybar[j])/(n-1)
+     sigmasq=1/(15-1)*sum((sampled-ybar[j])^2)
+     se[j]=(N-15)/N*(sigmasq/15)
+     bound=Z*sqrt((N-15)/N*(se[j]/15))
+     lower[j]=ybar[j] - bound
+     upper[j]=ybar[j] - bound
+ }
> hist(ybar,breaks=100)
> hist(ybar,breaks=50)
```

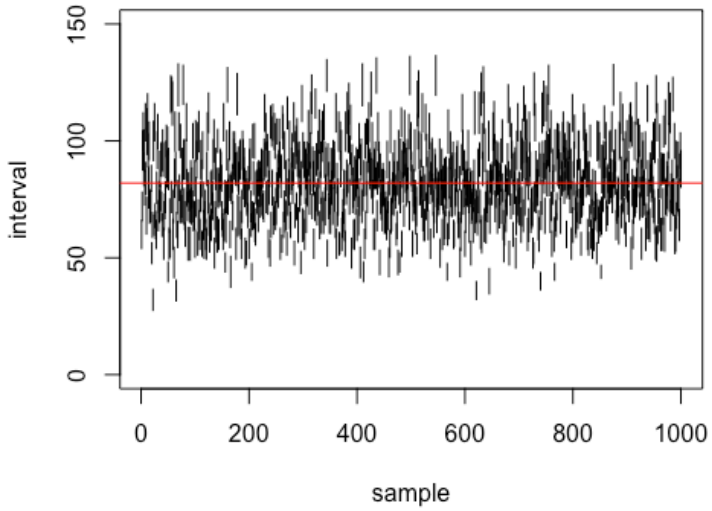



```
> mean(ybar)
[1] 1.366133
> var(ybar)
[1] 0.06618145
> mean(X)
[1] 1.366667
> mean(X)*(1-mean(X))*(N-n)/(N-1)/n
[1] -0.02548023
> hist(vhat,breaks=50)
> abline(v=mean(X)*(1-mean(X))*(N-n)/(N-1)/n,col="red")
> mean(vhat)
[1] -0.03033762
```



3) From Exercise 2), also compute the confidence interval for each of your 1000 samples. What proportions of your 1000 confidence intervals cover the truth?

```
> plot(1,0,type="n",ylim=c(0,150),xlim=c(1,1000),xlab="sample",ylab="interval")
> segments(1:R,lower,1:R,upper)
> abline(h=N*mean(X),col="red")
> mean(lower<=N*mean(X) & N*mean(X)<=upper)
[1] 0.366
```



Ans: 36.6%