ST 502 HW 3 Chapter 7 Problem 65

Bruce Campbell

January 29, 2017

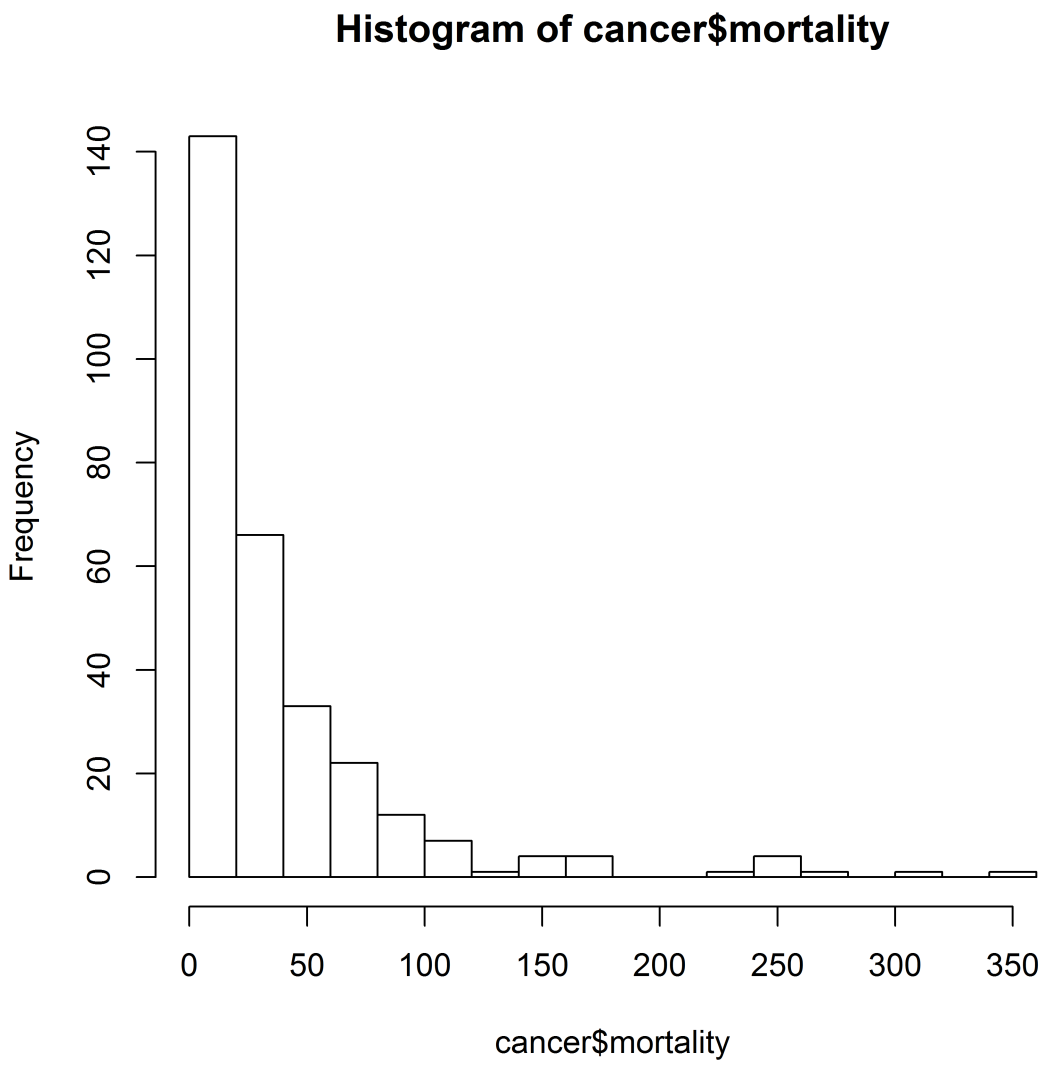
Rice, John A.. Mathematical Statistics and Data Analysis Chapter 7 Problem 65

The disk file cancer contains values for breast cancer mortality from 1950 to 1960 (y) and the adult white female population in 1960 (x) for 301 counties in North Carolina, South Carolina, and Georgia.

- a. Make a histogram of the population values for cancer mortality. -
- b. What are the population mean and total cancer mortality? What are the population variance and standard deviation?
- c. Simulate the sampling distribution of the mean of a sample of 25 observations of cancer mortality.
- d. Draw a simple random sample of size 25 and use it to estimate the mean and total cancer mortality.
- e. Estimate the population variance and standard deviation from the sample of part (d).
- f. Form 95% confidence intervals for the population mean and total from the sample of part (d). Do the intervals cover the population values?
- g. Repeat parts (d) through (f) for a sample of size 100.

Histogram and Summary Statistics

```
library(pander)
library(plyr)
library(dplyr)
library(readr)

# cancer <- read_csv('C:/E/GD/NCSU/st-502/Rice-DataSets/Chapter
# 7/cancer.csv')
cancer <- read_csv("D:/GD/NCSU/st-502/Rice-DataSets/Chapter 7/cancer.csv")
colnames(cancer) <- c("mortality", "population")

hist(cancer$mortality, 20)
```




```
summary(cancer)
```

```
##    mortality        population
##  Min.   :  0.00   Min.   :  559
##  1st Qu.: 11.00   1st Qu.: 2955
##  Median : 22.00   Median : 6534
##  Mean   : 39.99   Mean   :11324
##  3rd Qu.: 48.00   3rd Qu.:14014
##  Max.   :360.00   Max.   :88456
```

```
pander(data.frame(sum(cancer$mortality)), caption = "Toal Mortality")
```

Table 1: Toal Mortality

sum.cancer.mortality.
11996

```
pander(data.frame(var(cancer$mortality)), caption = "Variance Mortality")
```

Table 2: Variance Mortality

var.cancer.mortality.
2602

```
pander(data.frame(sd(cancer$mortality)), caption = "Standard Deviation Mortality")
```

Table 3: Standard Deviation Mortality

sd.cancer.mortality.
51.01

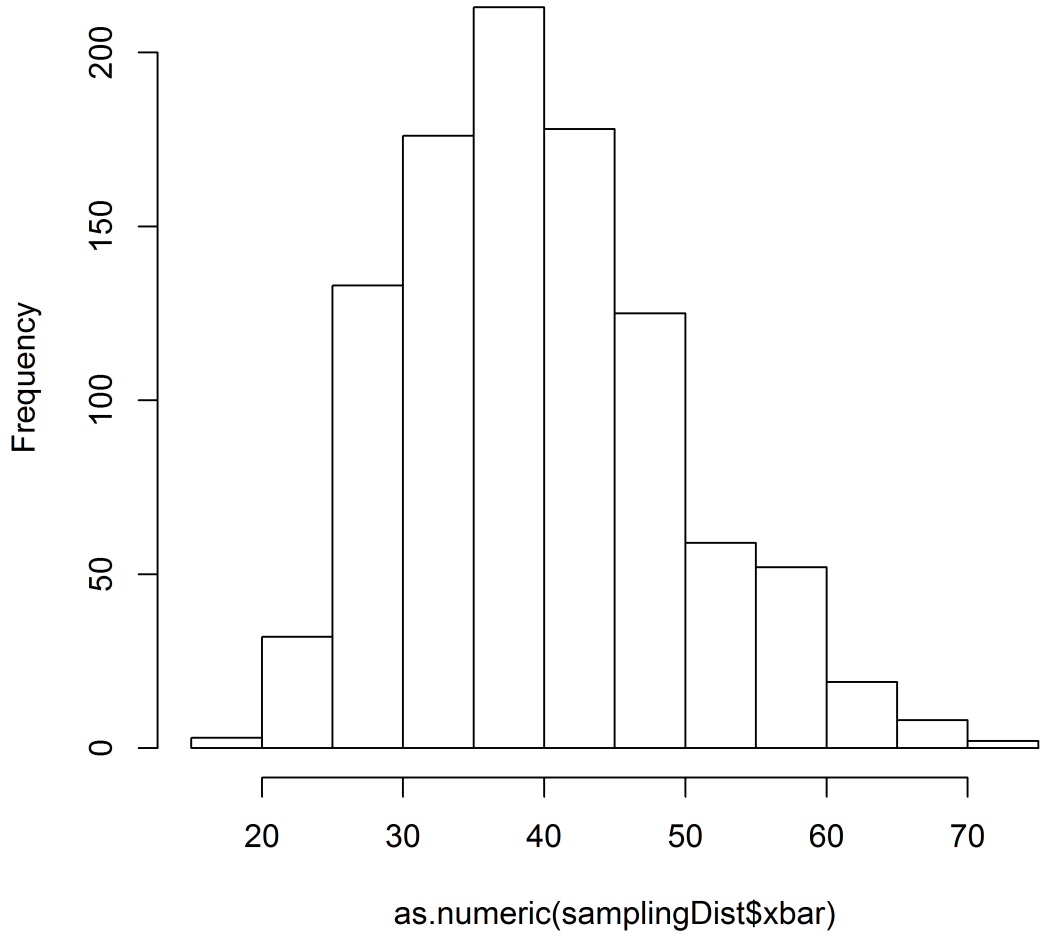
Simulate the sampling distribution of the mean of a sample of 25 observations of cancer mortality.

```
set.seed(314)
sampleSize <- 25
numSamples <- 1000

totalSamples <- choose(nrow(cancer), sampleSize)
pander(data.frame(totalSamples), "Total Possible Samples under SRS")
```

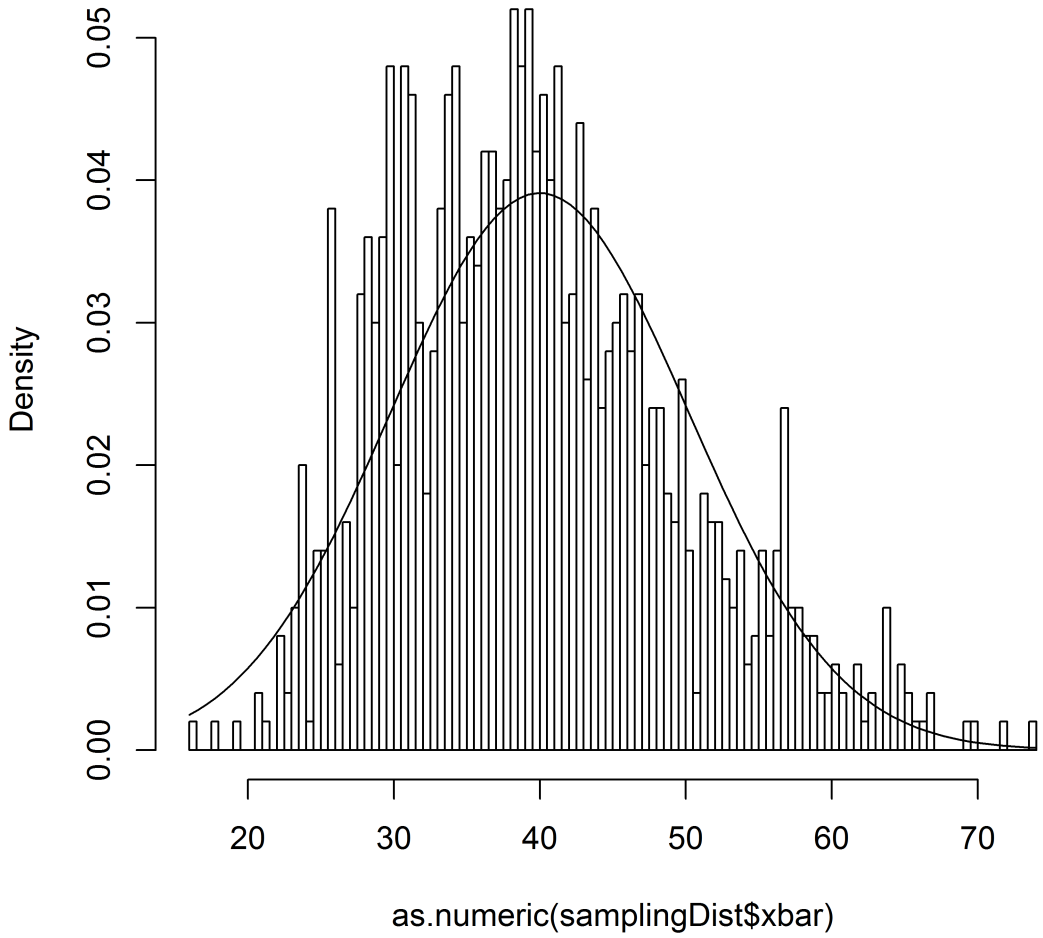
Table 4: Total Possible Samples under SRS

totalSamples
1.953e+36

```
samplingDist <- data.frame(matrix(ncol = 1, nrow = numSamples))
colnames(samplingDist) <- "xbar"
for (i in 1:numSamples) {
    srsIndex <- sample(nrow(cancer), sampleSize, replace = FALSE)
    srs <- cancer[srsIndex, ]
    Xbar <- mean(srs$mortality)
    samplingDist[i, 1] = Xbar
}

hist(as.numeric(samplingDist$xbar), 20, main = c("Histogram of Simulation of Sampling distribution for S
    "Sample size =25", "Number Of Samples = 1000"))
```




```
hist(as.numeric(samplingDist$xbar), freq = FALSE, 100, main = c("Normalized Histogram with actual sampli
curve(dnorm(x, mean = mean(cancer$mortality), sd = sd(cancer$mortality)/sqrt(sampleSize)),
    add = TRUE)
```



Draw a simple random sample of size 25 and use it to estimate the mean and total cancer mortality.

```
set.seed(314)
srsIndex <- sample(nrow(cancer), sampleSize, replace = FALSE)
srs <- cancer[srsIndex, ]
Xbar <- mean(srs$mortality)

totalMortality <- sum(srs$mortality)

pander(data.frame(Xbar), caption = "Estimated mean from single SRS")
```

Table 5: Estimated mean from single SRS

Xbar
16.36

```
pander(data.frame(totalMortality), caption = "Estimated total mortality from single SRS")
```

Table 6: Estimated total mortality from single SRS

totalMortality
409

Estimate the population variance and standard deviation from the sample of part (d).

If $E[X] = \mu$ and $Var(X) = \sigma^2$ then we know that $E[\bar{X}] = \mu$ and

$$Var(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

From this we can estimate $E[\bar{X}]$ and $Var(X)$ from our simulated sampling distribution.

```
muHat <- mean(samplingDist$xbar)
finitePopulationCorrection <- (nrow(cancer) - sampleSize)/(nrow(cancer) - 1)
sigmaHat <- sd(samplingDist$xbar) * sqrt(sampleSize) * 1/finitePopulationCorrection


pander(data.frame(estimate = c(muHat, sigmaHat), actual = c(mean(cancer$mortality),
    sd(cancer$mortality)), parameter = c("mean", "standard deviation")), caption = "Avtual Versus Estima
```

Table 7: Avtual Versus Estimate from Sampling Distribution Sample Size = 25

estimate	actual	parameter
39.65	39.99	mean
52.03	51.01	standard deviation

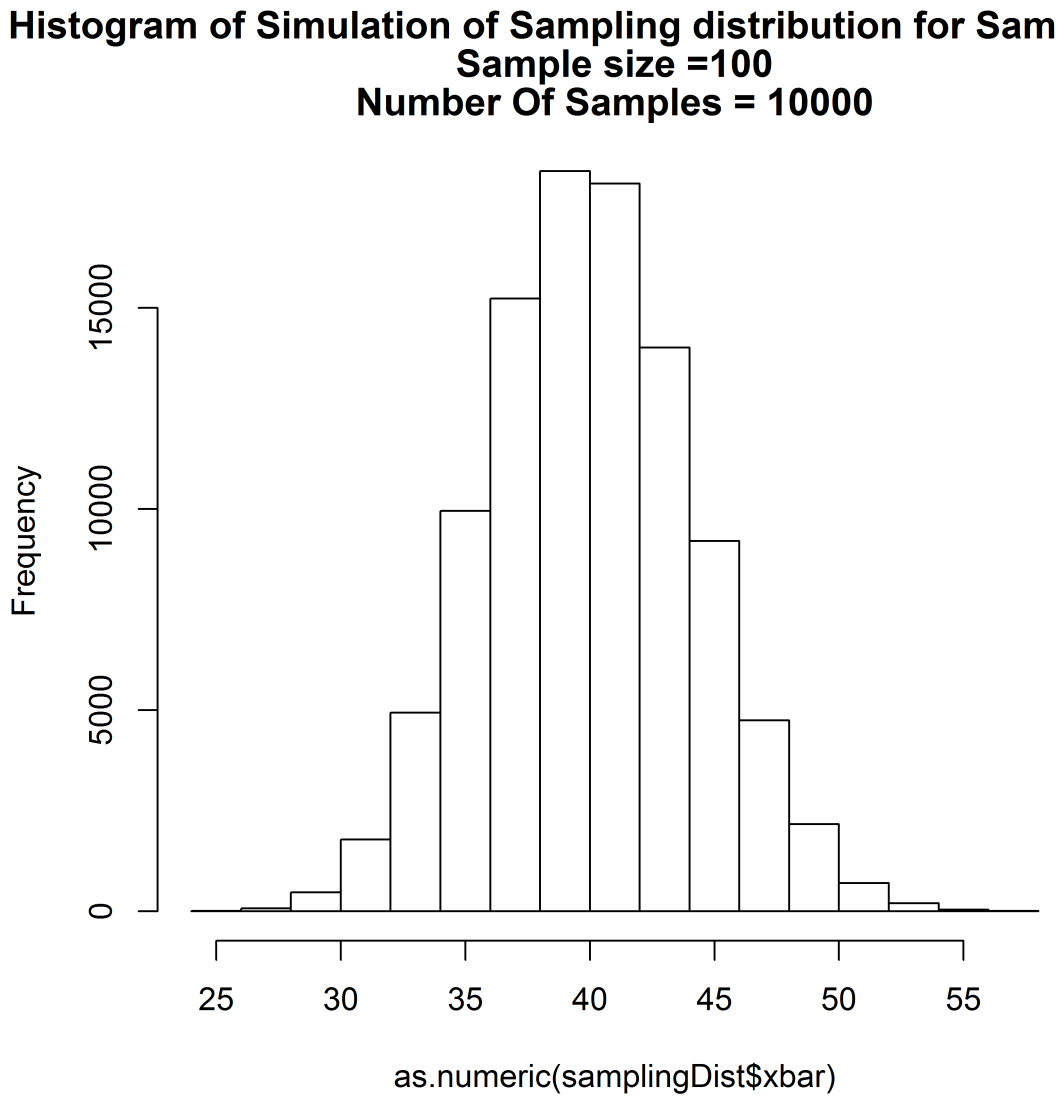
Repeat parts (d) through (f) for a sample of size 100.

```
set.seed(314)
sampleSize <- 100
numSamples <- 1e+05

totalSamples <- choose(nrow(cancer), sampleSize)
pander(data.frame(totalSamples), "Total Possible Samples under SRS")
```

Table 8: Total Possible Samples under SRS

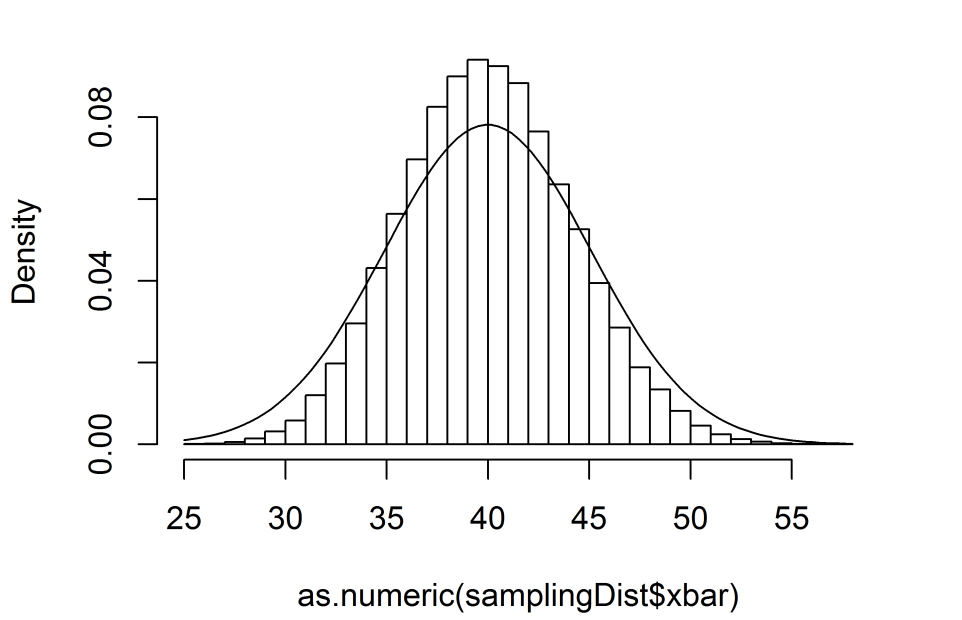
totalSamples
4.158e+81

```
samplingDist <- data.frame(matrix(ncol = 1, nrow = numSamples))
colnames(samplingDist) <- "xbar"
for (i in 1:numSamples) {
    srsIndex <- sample(nrow(cancer), sampleSize, replace = FALSE)
    srs <- cancer[srsIndex, ]
    Xbar <- mean(srs$mortality)
    samplingDist[i, 1] = Xbar
}

hist(as.numeric(samplingDist$xbar), 20, main = c("Histogram of Simulation of Sampling distribution for S
    "Sample size =100", "Number Of Samples = 10000"))
```



Here we plot the histogram overlayed with the dirstribution function for $N(\mu, \sigma_{\overline{X}})$

```
hist(as.numeric(samplingDist$xbar), freq = FALSE, 25, main = c("Normalized Histogram with actual samplin
curve(dnorm(x, mean = mean(cancer$mortality), sd = sd(cancer$mortality)/sqrt(sampleSize)) *
    sqrt((numSamples - sampleSize)/(numSamples - 1)), add = TRUE)
```



Draw a simple random sample of size 100 and use it to estimate the mean and total cancer mortality.

```
srsIndex <- sample(nrow(cancer), sampleSize, replace = FALSE)
srs <- cancer[srsIndex, ]
Xbar <- mean(srs$mortality)

totalMortality <- sum(srs$mortality)

pander(data.frame(Xbar), caption = "Estimated mean from single SRS")
```

Table 9: Estimated mean from single SRS

Xbar
44.65

```
pander(data.frame(totalMortality), caption = "Estimated total mortality from single SRS")
```

Table 10: Estimated total mortality from single SRS

totalMortality
4465

Estimate the population variance and standard deviation from the sample of part (d).

```
muHat <- mean(samplingDist$xbar)

sigmaHat <- sd(samplingDist$xbar) * sqrt(sampleSize)


pander(data.frame(estimate = c(muHat, sigmaHat), actual = c(mean(cancer$mortality),
    sd(cancer$mortality)), parameter = c("mean", "standard deviation")), caption = "Avtual Versus Estima
```

Table 11: Avtual Versus Estimate from Sampling Distribution
Sample Size = 100

estimate	actual	parameter
40	39.99	mean
41.69	51.01	standard deviation