STA2050 - Assignment 1

Chesia Anyika

2024-02-18

Name: Chesia Anyika

ID: 665567

Libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.0     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors
```

```
library(SDAResources)
```

```
##
## Attaching package: 'SDAResources'
##
## The following object is masked from 'package:ggplot2':
##
##     seals
```

```
library(survey)
```

```
## Loading required package: grid
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
```

```
## Loading required package: survival
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
```

Question 1

A university has 807 faculty members. For each faculty member, the number of refereed publications was recorded. This number is not directly available on the database, so requires the investigator to examine each record separately. A frequency table for number of refereed publications is given below for an SRS of 50 faculty members.

Refereed Publications	0	1	2	3	4	5	6	7	8	9	10
Faculty Members	28	4	3	4	4	2	1	0	2	1	1

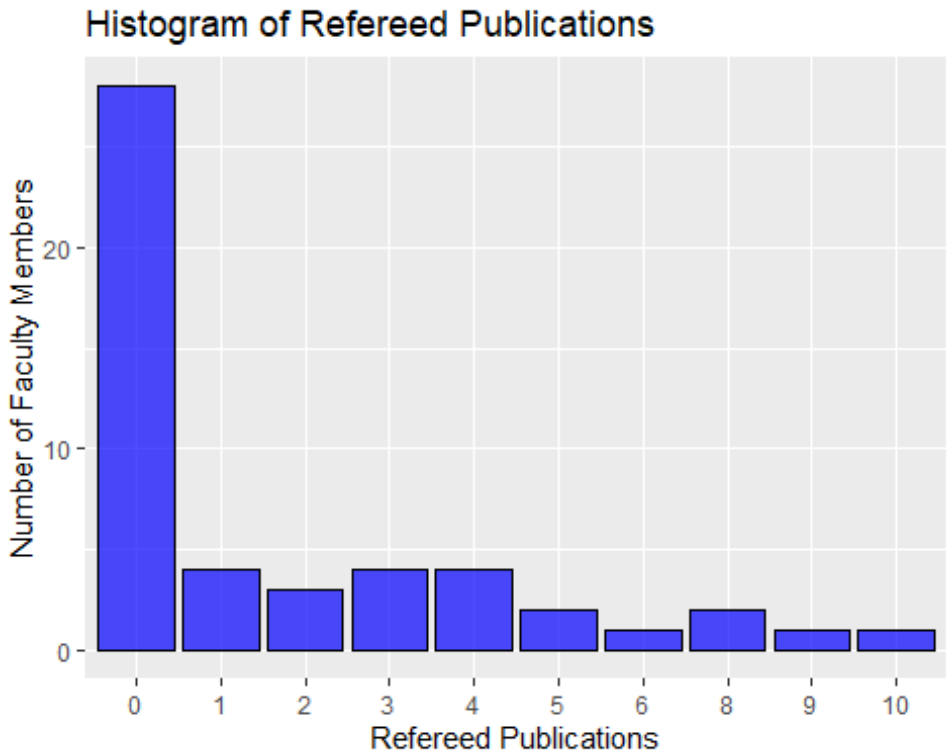
a) Plot the data using a histogram. Describe the shape of the data. (2 marks)

First I created a data-frame with the above data:

```
#create dataframe
df1 <- data.frame(
  Publications = rep(0:10, times = c(28, 4, 3, 4, 4, 2, 1, 0, 2, 1, 1))
)

#view head and tail of dataframe
head(df1)

##   Publications
## 1            0
## 2            0
## 3            0
## 4            0
## 5            0
## 6            0

tail(df1)

##    Publications
## 45            5
## 46            6
## 47            8
## 48            8
## 49            9
## 50           10
```

I then plotted a histogram using `geom_bar()` function in the `ggplot2` package

```
# Create the histogram #library(tidyverse)
ggplot(df1, aes(x = factor(Publications))) +
  geom_bar(fill = "blue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of Refereed Publications",
       x = "Refereed Publications",
       y = "Number of Faculty Members")
```



The histogram is **right skewed**, such that the mode of the distribution is less than the mean. Majority of faculty members have 0 refereed publications, followed by between $1-4$ refereed publications, then $5-10$. Thus more faculty members tend to have 4 or less refereed publications.

b)Estimate the mean number of publications per faculty member, and give the SE for your estimate. (3 marks)

To estimate the population mean $\bar{Y}_U$ from the SRS, I used the formula:

$$\bar{y}_S = \frac{1}{n}\sum_{i=1}^{k} x_i f_i$$

Where

- n is the total number of faculty members in the sample (50 in this case)

- k is the number of categories or levels of publications (in this case, 11 levels from 0 to 10)
- x_i is the midpoint of the i^{th} category.
- f_i is the frequency of faculty members in the i^{th} category.

I computed this as follows:

```
#define the sample values
n = 50
xi <- 0:10
fi <- c(28, 4, 3, 4, 4, 2, 1, 0, 2, 1, 1)

#calculate the estimated mean
ys <- sum(xi*fi)/n

cat('Estimated Mean: ', ys)

## Estimated Mean:  1.78
```

The estimated number of publications per faculty memeber is 1.78.

To calculate the standard error I used the formula

$$SE = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{y}_s)^2 f_i}{n(n-1)}}$$

Where:

x_i is the midpoint of the i^th category

$\bar{y}_s$ is the estimated mean

f_i is the frequency of faculty members in the i^th category

n is the total number of faculty members in the sample

I computed this as follows:

```
SE <- sqrt(sum((xi - ys)^2 * fi) / (n * (n - 1)))

cat('Standard Error: ', SE)

## Standard Error:  0.379355
```

The Standard Error is 0.3794

c)Estimate the proportion of faculty members with no publications and give a 95% CI (2marks)

To estimate the population proportion we use the sample proportion, calculated as follows:

$$\hat{p} = \frac{x}{n}$$

Where

- x is the number of individuals with a particular characteristic
- n is the total number of individuals in the sample

I computed this as follows

```
#define parameters
x = 28
n = 50

#compute proportion
phat <- x/n

#print result
cat('Estimated Population Proportion with 0 Publications: ', phat )

## Estimated Population Proportion with 0 Publications:  0.56
```

The Estimated Population Proportion with 0 publications is 0.56

To compute the Confidence Interval, we use the following formula:

$$CI = \hat{p} \pm Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

OR:

$$CI = \hat{p} \pm MarginOfError$$

$\hat{p}$ is the estimated proportion

Z is the z-score that corresponds with the desired confidence level

n is the total number of observations in the sample

I computed this as follows:

```
#define confidence level
alpha <- 0.95

#define standard error for proportion
SEp <- sqrt((phat * (1 - phat)) / n)

#calculate the z_score
z.score <- qnorm((1 + alpha) / 2)

#calculate the margin of error
ME <- z.score * SEp
```

```r
#Confidence Interval
lower <- phat - ME
upper <- phat + ME

cat('Confidence Interval: [', lower, ',', upper, ']')
```

```
## Confidence Interval: [ 0.4224111 , 0.6975889 ]
```

The confidence Interval for the Estimated Proportion of Faculty with 0 publications is:

$$CI = [0.4224111, 0.6975889]$$

Question 2

Suppose that a city has 90,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 10,000 are condominiums. You believe that the mean electricity usage is about twice as much for houses as for apartments or condominiums, and that the standard deviation is proportional to the mean so that S1 = 2S2 = 2S3. How would you allocate a stratified sample of 900 observations if you wanted to estimate the mean electricity consumption for all households in the city? (3marks)

To estimate mean electricity consumption for all houses in the city, I would allocate a stratified sample of 900 as follows:

Step 1: Compute Proportions per stratum

$$p_i = \frac{xi}{N}$$

Where

p_i is the proportion per stratum

x_i is the number of units per stratum

N is the population total

I computed this as follows:

```r
#define population total
N = 90000

#define number of units per stratum
x1 = 35000
x2 = 45000
x3 = 10000

#compute proportions
p1 <- x1/N
p2 <- x2/N
```

```
p3 <- x3/N

#print
cat('Proportion of Houses (p1): ', p1, '\n Proportion of Apartments (p2): ',
p2, '\n Proportion of Condominiums (p3): ', p3)

## Proportion of Houses (p1):  0.3888889
##  Proportion of Apartments (p2):  0.5
##  Proportion of Condominiums (p3):  0.1111111
```

Step 2: Determine the sample sizes

I multiplied the total sample size by the proportions calculated in order to determine the number of observations per stratum, as per the formula

$$n_i = p_i \times n$$

Where

n_i is the sample size per stratum

p_i is the proportion per stratum

n is the total sample size

I computed this as follows

```
#define the total sample size
n = 900

#compute sample sizes per stratum
n1 <- p1 * n
n2 <- p2 * n
n3 <- p3 * n

#print
cat('Houses Sample Size:', n1, '\n Apartments Sample Size: ', n2, '\n
Condominiums Sample Size: ', n3)

## Houses Sample Size: 350
##  Apartments Sample Size:  450
##  Condominiums Sample Size:  100
```

Question 3

The data file agstrat.dat contains information on other variables. For each of the following quantities, plot the data, and estimate the population mean for that variable along with its standard error and a 95% CI.

a) Number of acres devoted to farms, 1987 (5 marks)

First I imported the data set, and viewed the column names.

```
#impot the data #library(SDAResources)
data(agstrat)

#view column names
names(agstrat)
```

```
##  [1] "county"   "state"    "acres92"  "acres87"  "acres82"  "farms92"
##  [7] "farms87"  "farms82"  "largef92" "largef87" "largef82" "smallf92"
## [13] "smallf87" "smallf82" "region"   "rn"       "strwt"
```

The number of acres devoted to farms, 1987 corresponds to the variable `acres87` .

Part 1: Plot the Data

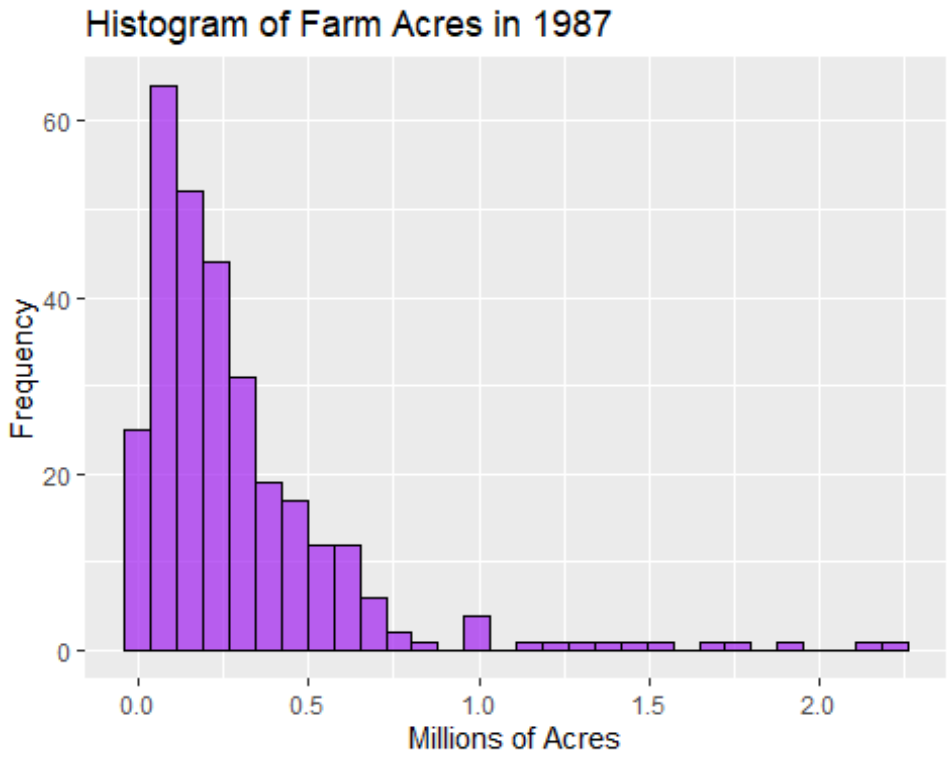
I plotted a **histogram** and a **Boxplot** of the data to explore different aspects of it.

Histogram

I used the `geom_histogram()` function from the `ggplot2` library to plot the histogram below:

```
#plot histogram #library(tidyverse)
ggplot(agstrat, aes(x = acres87 / 10^6)) +
  geom_histogram(fill = "purple", color = "black", alpha = 0.7) +
  labs(x = "Millions of Acres", y = "Frequency") +
  ggtitle("Histogram of Farm Acres in 1987")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

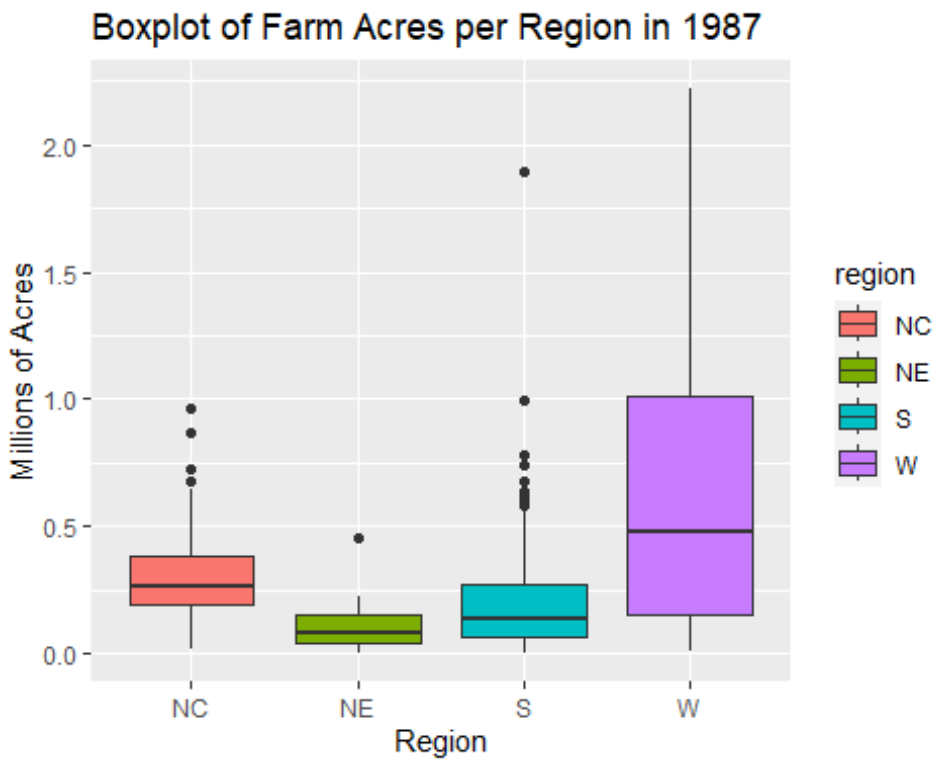


The histogram is **right skewed**, such that the mode of the distribution is less than the mean. This implies that majority of farmers planted less acres than the mean amount of acres planted in 1987.

Boxplot

I used the `geom_boxplot()` function from the `ggplot2` package to plot the following boxplot, in order to explore the acreage planted per region:

```
ggplot(agstrat, aes(x = region, y = acres87 / 10^6, fill = region)) +
  geom_boxplot() +
  labs(x = "Region", y = "Millions of Acres") +
  ggtitle("Boxplot of Farm Acres per Region in 1987")
```



The Boxplot shows that the **State W** had the **highest mean** of land dedicated to farms in 1987, as well as the **largest interquartile range**.

Part 2: Estimation of Population Mean, Standard Error and Confidence Interval

To calculate the stated statistics, I created a Survey Design.

```
# create a variable containing population stratum sizes #library(survey)
popsize_recode <- c('NC' = 1054, 'NE' = 220, 'S' = 1382, 'W' = 422)

#substitute population sizes for strata names
agstrat$popsize <- popsize_recode[agstrat$region]
table(agstrat$popsize) #check the new variable
```

```
##
##  220  422 1054 1382
##   21   41  103  135
```

I then input the design information.

```
dstr <- svydesign(id = ~1, strata = ~region, weights = ~strwt, fpc =
~popsize, data = agstrat)

dstr

## Stratified Independent Sampling design
## svydesign(id = ~1, strata = ~region, weights = ~strwt, fpc = ~popsize,
##     data = agstrat)
```

Mean and Standard Error (Using survey information)

```
#calculate mean, SE and confidence interval
smean <- svymean(~acres87, dstr)
smean

##           mean    SE
## acres87 298547 16293
```

Confidence Interval

```
confint(smean, level=.95, df=degf(dstr))

##            2.5 %   97.5 %
## acres87 266482.4 330611.8
```

The computed statistics are

$$Mean = 298547$$

$$SE = 16293$$

$$CI = [266482.4, 330611.8]$$

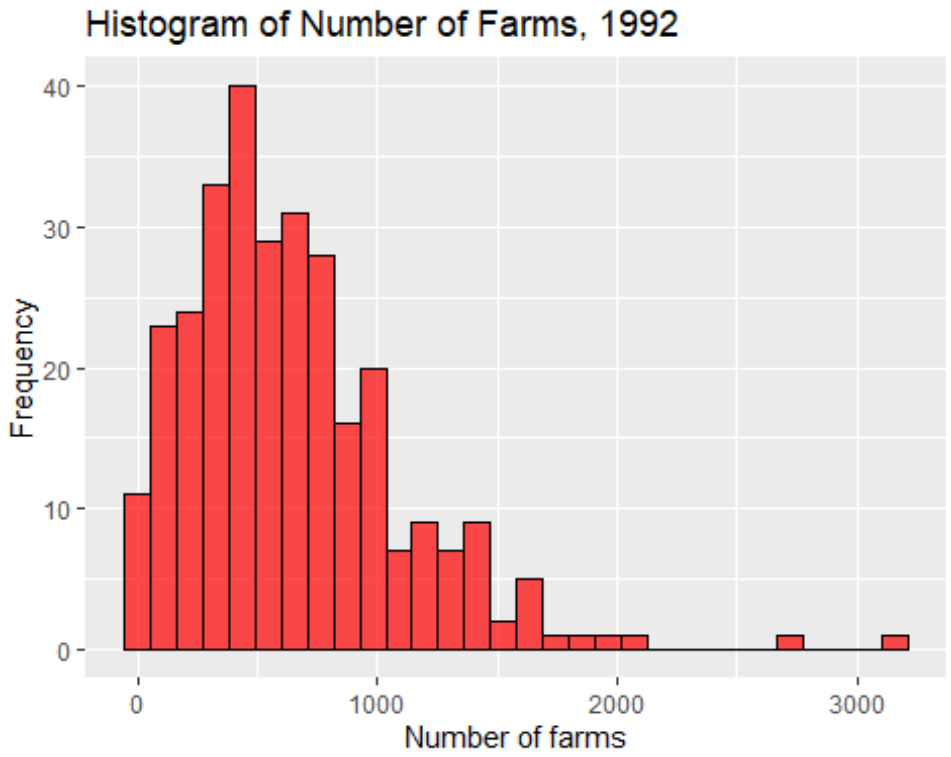
b) Number of farms, 1992(5marks)

The number of farms, 1992 corresponds to the variable `farms92` .

Part 1: Plot the data

Histogram

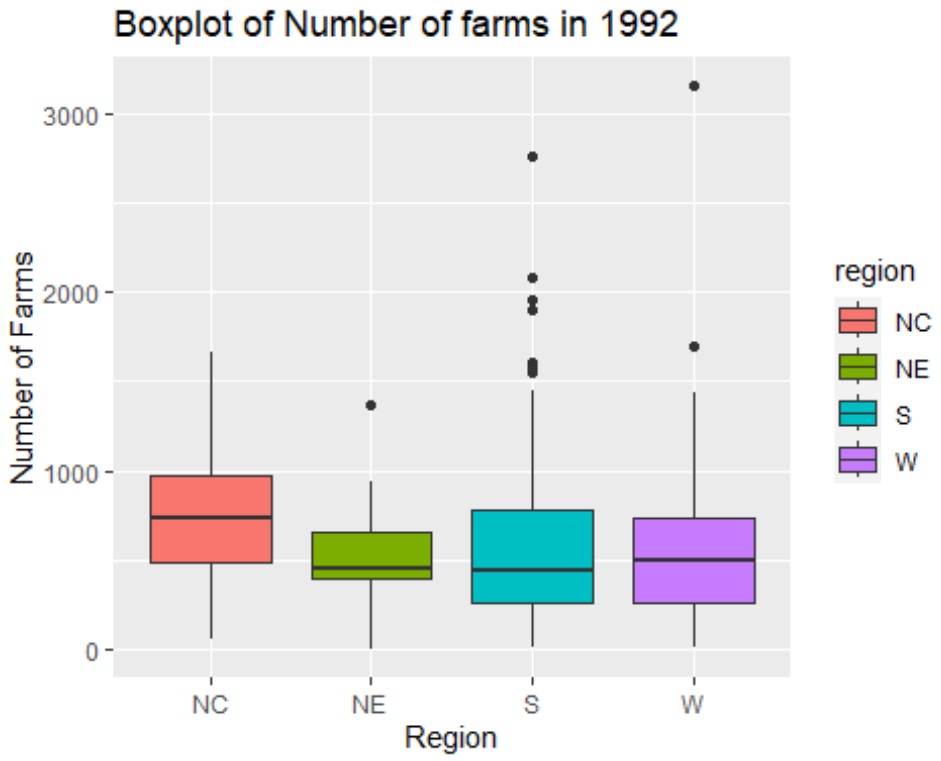
```
#plot histogram #library(tidyverse)
ggplot(agstrat, aes(x = farms92 )) +
  geom_histogram(fill = "red", color = "black", alpha = 0.7) +
  labs(x = "Number of farms", y = "Frequency") +
  ggtitle("Histogram of Number of Farms, 1992")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The above histogram is **right skewed** suggesting that the mode of the distribution is less than the mean and the median. This implies that the number of farms in 1992 tended to be below the mean of the distribution, on the lower end.

Boxplot

```
ggplot(agstrat, aes(x = region, y = farms92, fill = region)) +
  geom_boxplot() +
  labs(x = "Region", y = "Number of Farms") +
  ggtitle("Boxplot of Number of farms in 1992")
```



The **NC** region has the highest mean of number of farms in 1992, while the **S** region exhibits the most variance.

Part 2: Estimation of Population Mean, Standard Error and Confidence Interval

I computed this using the survey design I created in part (a)

Mean and Standard Error

```
#calculate mean, SE and confidence interval
smean1 <- svymean(~farms92, dstr)
smean1

##           mean     SE
## farms92 637.16 24.277
```

Confidence Interval

```
confint(smean1, level=.95, df=degf(dstr))

##             2.5 %   97.5 %
## farms92 589.3853 684.9422
```

The computed statistics are as follows:

$$ Mean = 637.16 \\ SE = 24.277 \\ CI = [589.3853, 684.9422] $$

c) Number of farms with 1000 acres or more, 1992 (5 marks)

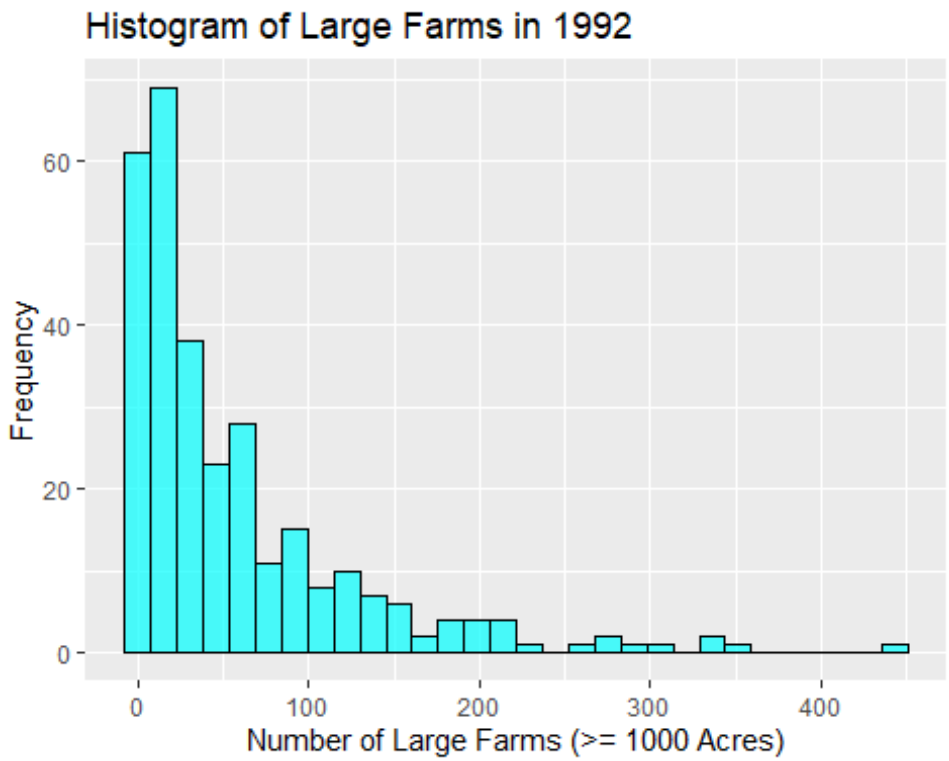
Part 1: Plot the data

The number of farms with 1000 acres or more, 1992 corresponds with the `largef92` variable.

Histogram

```
#plot histogram #library(tidyverse)
ggplot(agstrat, aes(x = largef92)) +
  geom_histogram(fill = "cyan", color = "black", alpha = 0.7) +
  labs(x = "Number of Large Farms (>= 1000 Acres)", y = "Frequency") +
  ggtitle("Histogram of Large Farms in 1992")
```

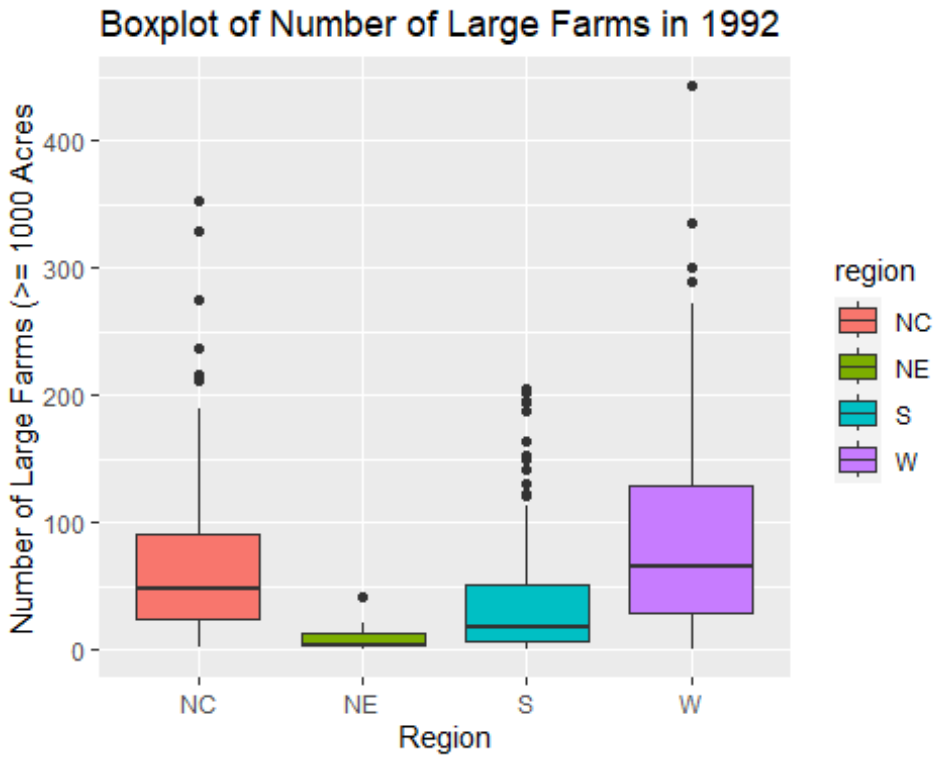
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram is **right skewed**, thus the mod of the distribution is less than the median and mean. Thus, in 1992 there tended to be less large farms as compared to the mean number that existed.

Boxplot

```
ggplot(agstrat, aes(x = region, y = largef92, fill = region)) +
  geom_boxplot() +
  labs(x = "Region", y = "Number of Large Farms (>= 1000 Acres") +
  ggtitle("Boxplot of Number of Large Farms in 1992")
```



The **W** region has the highest mean of number of large farms in 1992, as well as shows the highest amount of variation, and the largest interquartile range.

Part 2: Estimation of Population Mean, Standard Error and Confidence Interval

I computed this using the survey design I created in part (a)

Mean and Standard Error

```
#calculate mean, SE and confidence interval
smean3 <- svymean(~largef92, dstr)
smean3

##              mean     SE
## largef92 56.698 3.5577
```

Confidence Interval

```
confint(smean3, level=.95, df=degf(dstr))

##                2.5 %   97.5 %
## largef92 49.69636 63.69954
```

The computed statistics are:

$$ Mean = 56.698\\ SE = 3.5577\\ CI = [49.696, 63.700] $$

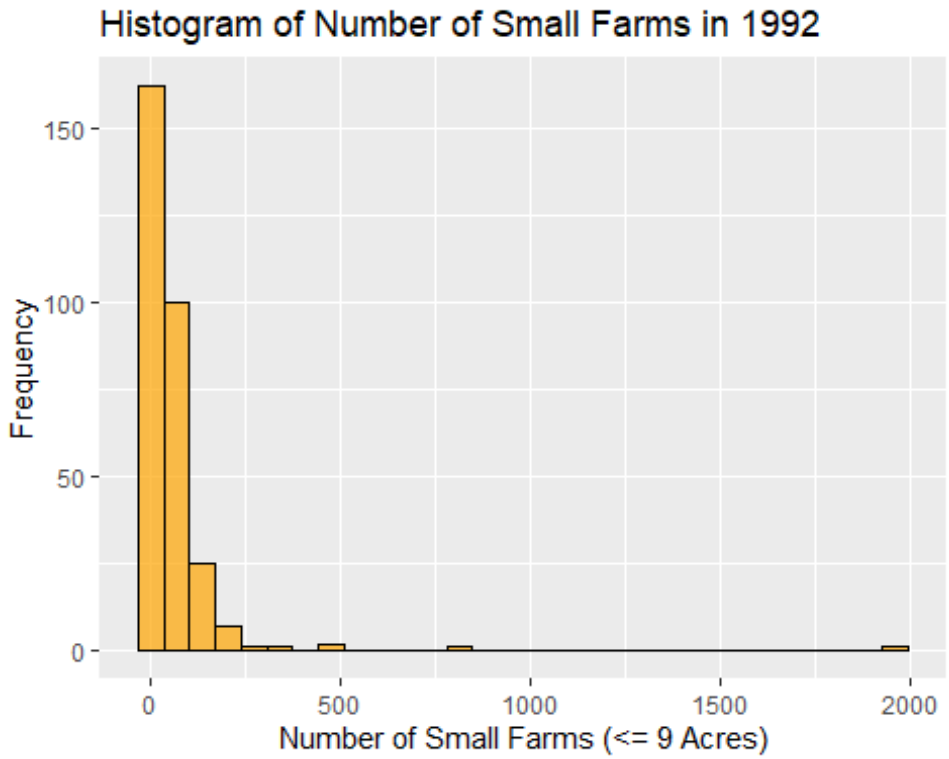
d) Number of farms with 9 acres or fewer, 1992 (5 marks)

The number of farms with 9 acres or fewer, 1992, corresponds with the variable `smallf92` .

Part 1: Plot the Data

Histogram

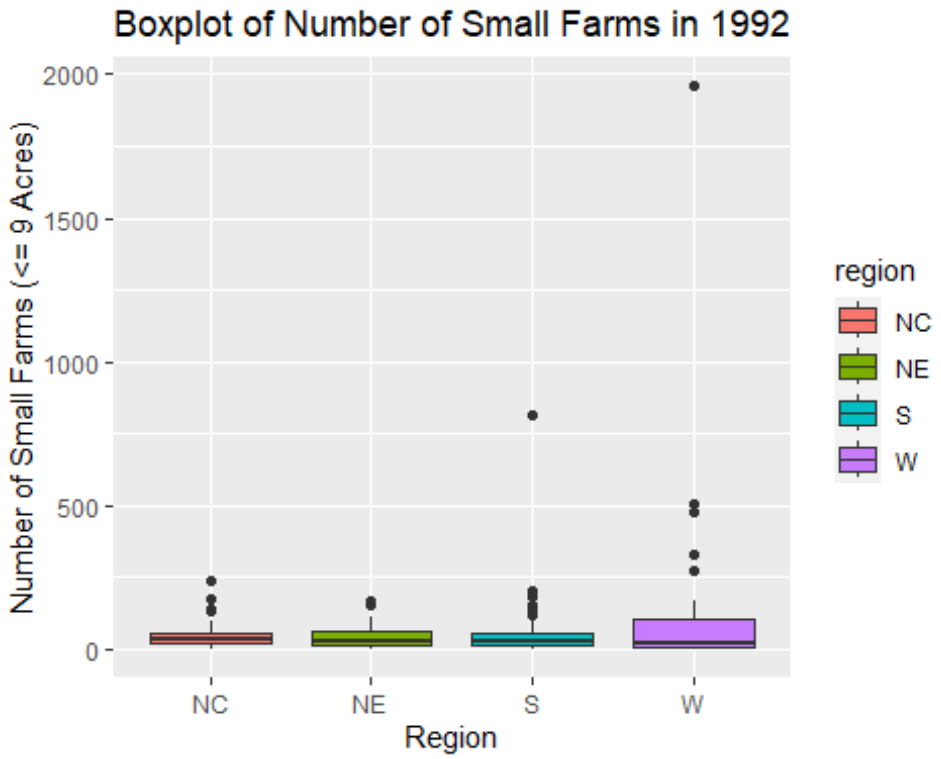
```
#plot histogram #library(tidyverse)
ggplot(agstrat, aes(x = smallf92)) +
  geom_histogram(fill = "orange", color = "black", alpha = 0.7) +
  labs(x = "Number of Small Farms (<= 9 Acres)", y = "Frequency") +
  ggtitle("Histogram of Number of Small Farms in 1992")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram is very **right skewed**, suggesting the mode of the distribution is smaller than the mean and the median. This suggests that the number of small farms in 1992 tended to be less than the mean value for the year.

Boxplot

```
ggplot(agstrat, aes(x = region, y = smallf92, fill = region)) +
  geom_boxplot() +
  labs(x = "Region", y = "Number of Small Farms (<= 9 Acres)") +
  ggtitle("Boxplot of Number of Small Farms in 1992")
```



There is little difference in the mean of number of small farms in 1992, but the **W** region has the highest interquartile range, and the largest variation among the regions.

Part 2: Estimation of Population Mean, Standard Error and Confidence Interval

I computed this using the survey design I created in part (a)

Mean and Standard Error

```
#calculate mean, SE and confidence interval
smean4 <- svymean(~smallf92, dstr)
smean4

##             mean     SE
## smallf92 56.863 7.2014
```

Confidence Interval

```
confint(smean4, level=.95, df=degf(dstr))

##              2.5 %   97.5 %
## smallf92 42.69033 71.03526
```

The computed statistics are as follows:

$$ Mean = 56.863\\ SE = 7.201\\ CI = [42.69, 71.04] $$