HW#7 87 F21

Lucía Carrera

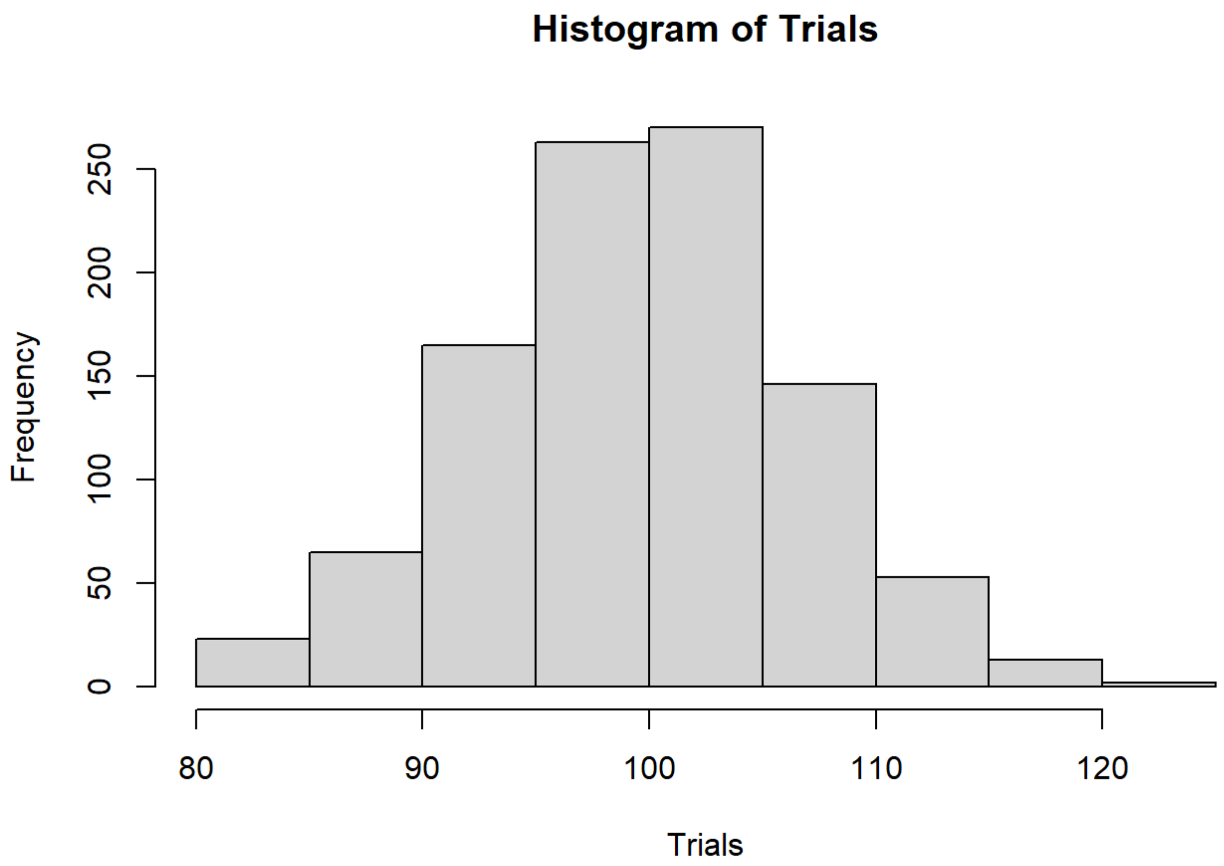
12/3/2021

```
knitr::opts_chunk$set(echo = TRUE)
```

Question 1. A hypothesis test

In my case p-value was 0.02 which is less than 0.05, so I reject the null hypothesis that 50% students would say they are better drivers than average, in favor of alternative hypothesis that more than 50% students would say they are better drivers.

```
# Suppose you surveyed a random sample of 200 UVM students, and 115 of them said th
ey are 'Better than Average' or "Way Better than Average" drivers.    Do a hypothes
is test to see if the true proportion of UVM students who would say they are better
or way better drivers is a majority.   That is, use the null hypothesis "50% of stu
dents would say they are better or way better than average".   Generate 1000 trial
s, compute the 'p-value' and state your decision (above), in terms of the problem.

# Observer Result (OR)
OR <- 115/200

# Null Hypothesis: 50% of students would say they are better or way better than ave
rage
Trials <- rbinom(n = 1000, size = 200, prob = .5)
hist(Trials)
```



```r
table(Trials)
```

```
## Trials
##  80  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101 
##   1   5   8   9   6   8  13  18  20  27  23  35  36  44  48  47  47  54  67  66 
## 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 121 124 
##  73  54  39  38  35  35  24  29  23  16   9  11  10   7   4   4   1   4   1   1
```

```r
# pvalue
pValue <- (8 + 6 + 1 + 2 + 1 + 1 + 1) /1000
pValue
```

```
## [1] 0.02
```

Question 2. Confidence Intervals for Percentages

part a – Driving

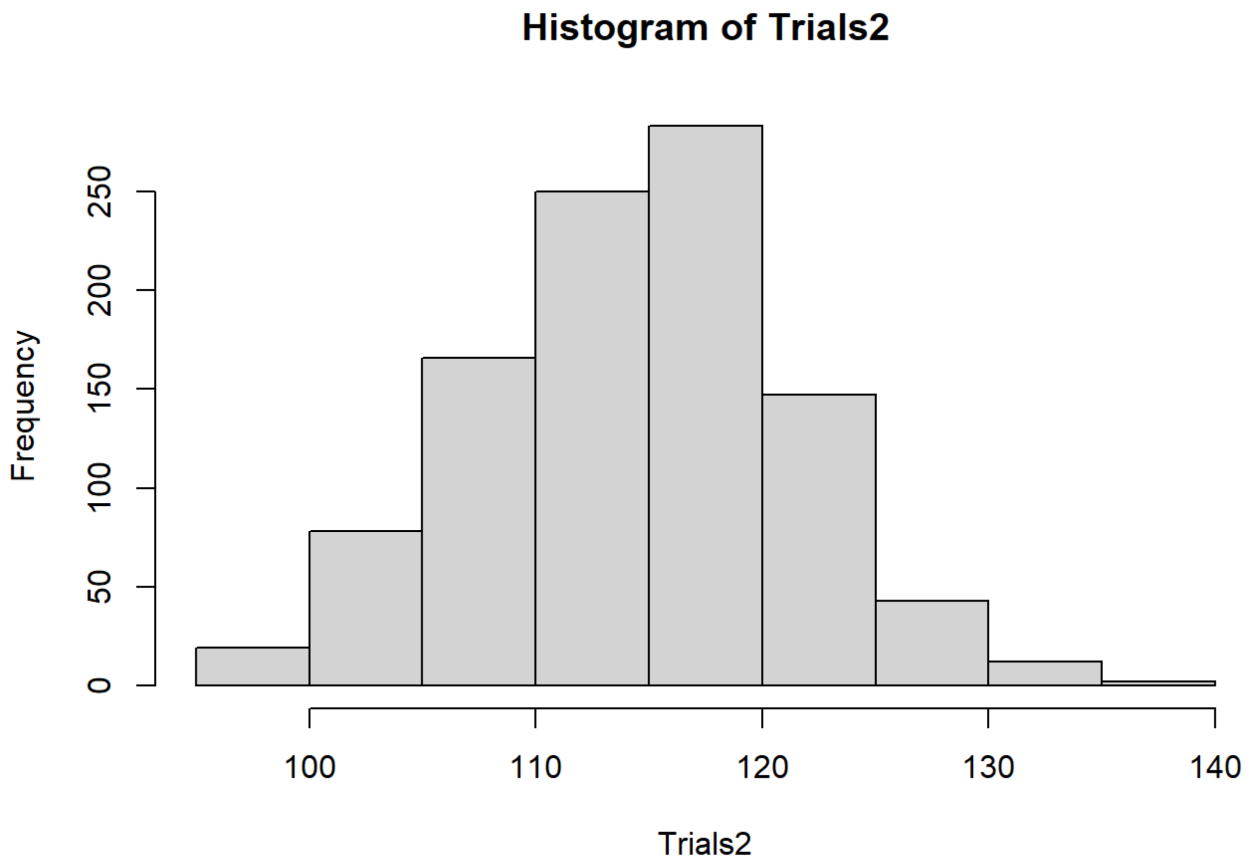
The point estimate is 57.5%.

I’m 95% confident that the percentage of UVM students who say they are better driver than average is captured in the interval from 50.83% to 64.17%.

I believe that majority of UVM students will say they are better drivers than average, since the whole 95% confidence interval is above 50%.

I’m 95% sure about this conclusion.

```
# Use the same data as in question 1 (a random sample of 200 UVM students, and 115
of them said they are 'Better than Average' or "Way Better than Average" drivers.)
Using methods from class, find a 95% bootstrap confidence interval for the true per
centage of students.  Have R calculate and print the lower and upper limit, roundin
g each value to 2 decimal places. (xx.xx  to xx.xx).  Also have R do a histogram of
the bootstrap percentages.

#  Express the trials as percentages;

#  Generate trials using the PE for the probability
Trials2 <- rbinom(n = 1000, size = 200, prob = 115/200)

#  histogram and table
hist(Trials2)
```



```
table(Trials2)
```

```
## Trials2
##  95  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
##   1   2   2   4   7   3   8   9  13  17  31  17  23  42  45  39  37  55  51  55
## 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 134 135
##  52  59  60  61  47  56  49  22  24  32  20  16   8   4  10   5   7   2   1   2
## 136
##   2
```

```
# Standard Error
SE <- sd(Trials2)

# Margin of error
Perc <- 100 * Trials / 200

# Point Estimate
PE <- 100 * 115/200
PE
```

```
## [1] 57.5
```

```
SE <- sd(Perc)
ME <- 2 * SE

# upper and lower limit
lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,2)
upperlimit <- round(upperlimit,2)



# In the text before this chunk, state the point estimate, and write a sentence of
interpretation above (I’m 95% sure…..)  Based on your interval, do you believe that
a majority of all UVM students would say they are better or way better than average
drivers?  How do you know?  How sure are you of the conclusion about a majority?
(write answers above)

print(paste("I am 95% confident the true mean difference of students saying they ar
e better drivers is captured in the interval from",lowerlimit, upperlimit))
```

```
## [1] "I am 95% confident the true mean difference of students saying they are bet
ter drivers is captured in the interval from 50.51 64.49"
```

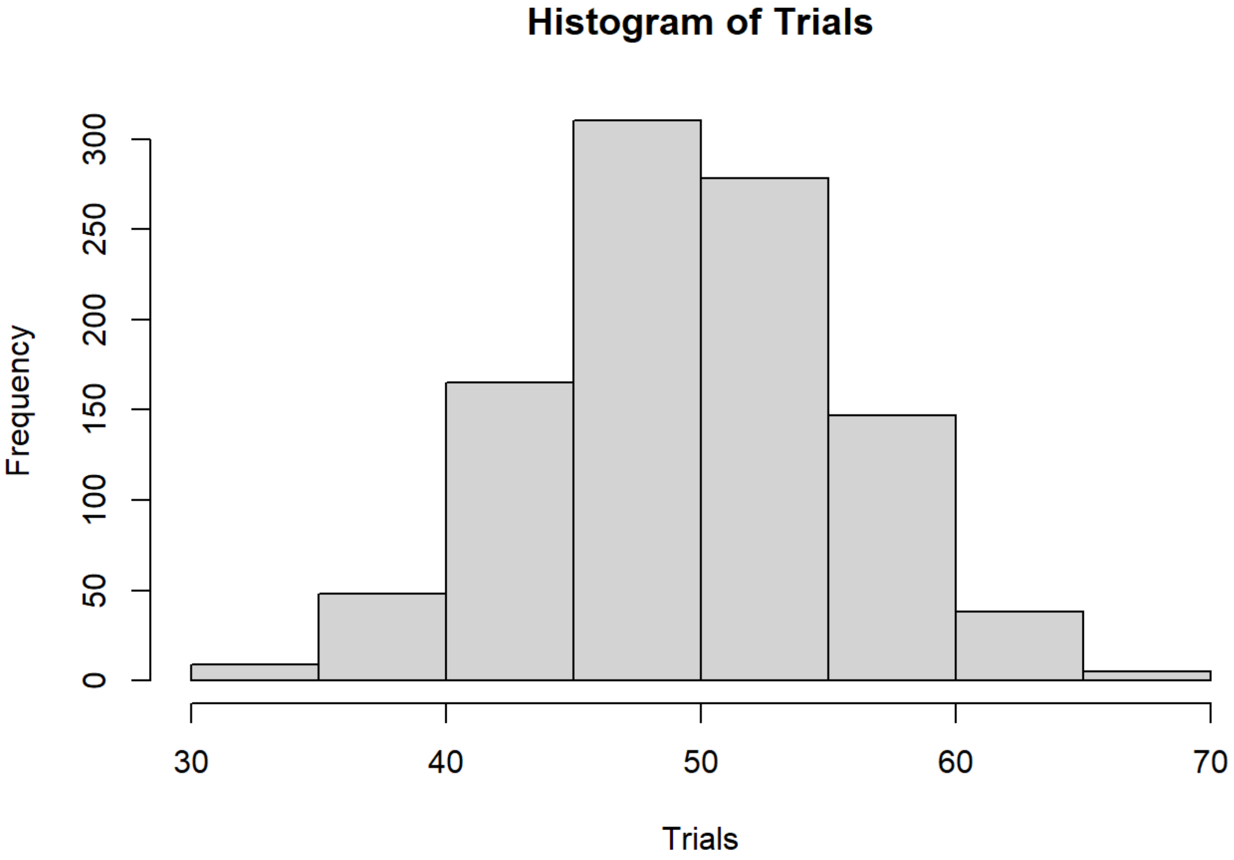
part b – Favorite app

Point estimate was 25%.

I’m 95% confident that the percentage of UVM students who say that their favorite app is Instagram is in interval from 19.01% to 30.99%. I believe that minority of UVM students will say that their favorite app is Instagram, since the whole 95% confidence interval is under 50%.

I’m 95% sure about this conclusion.

```
# Suppose in the same random sample of 200 UVM students, 50 of them said their favorite app is Instagram (the most popular app in this sample).    Using methods from class, find a 95% bootstrap confidence interval for the true percentage of students who prefer Instagram.  Have R calculate and print the lower and upper limit, rounding each value to 2 decimal places. (xx.xx  to xx.xx). Also have R do a histogram of the bootstrap percentages.

# probability
p = 50/200

# Trials
Trials <- rbinom(n = 1000, size = 200, prob = p)
hist(Trials)
```



```
# Percentage
Perc <- 100 * Trials / 200

# Point of Estimate
PE <- 100 * p
PE
```

```
## [1] 25
```

```
SE <- sd(Perc)

# Margin of Error
ME <- 2 * SE

# upper and lower limit
lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,2)
upperlimit <- round(upperlimit,2)
upperlimit
```

```
## [1] 31.09
```

```
# paste
print(paste(lowerlimit, upperlimit))
```

```
## [1] "18.91 31.09"
```

```
# In the text before this chunk, state the point estimate, and write a sentence of
interpretation above (I’m 95% sure…..)   Based on your interval, do you believe tha
t a minority of all UVM students (less than 50%) like Instagram the best?   How do
you know?   How sure are you of the conclusion about a minority? (write answers abo
ve)
```

Question 3 – Confidence Intervals for Means

part a - CI for mean earnings

The point estimate is 7458.578. I’m 95% confident that the true mean earnings in dollars for UVM students is in an interval from 5825.54 to 9091.61.

```
# Use the attached data file surveyC_S21.csv, download it, and read in as a data fr
ame called, s.  Assume the data is a random sample of 200 UVM students.  Using the
bootstrap method, and code similar to the code from class, find a 95% confidence in
terval for the true mean earnings for UVM students.  Have R do a histogram of the b
ootstrap means, calculate the standard deviation of the bootstrap means (the SE), c
alculate the margin of error, and Upper and Lower limits of the confidence interva
l.  Have R calculate and print the lower and upper limit, rounding each value to 2
decimal places. (xx.xx  to xx.xx).

# data loading
s <- read.csv("surveyC_F21.csv")

library(tidyverse)
```

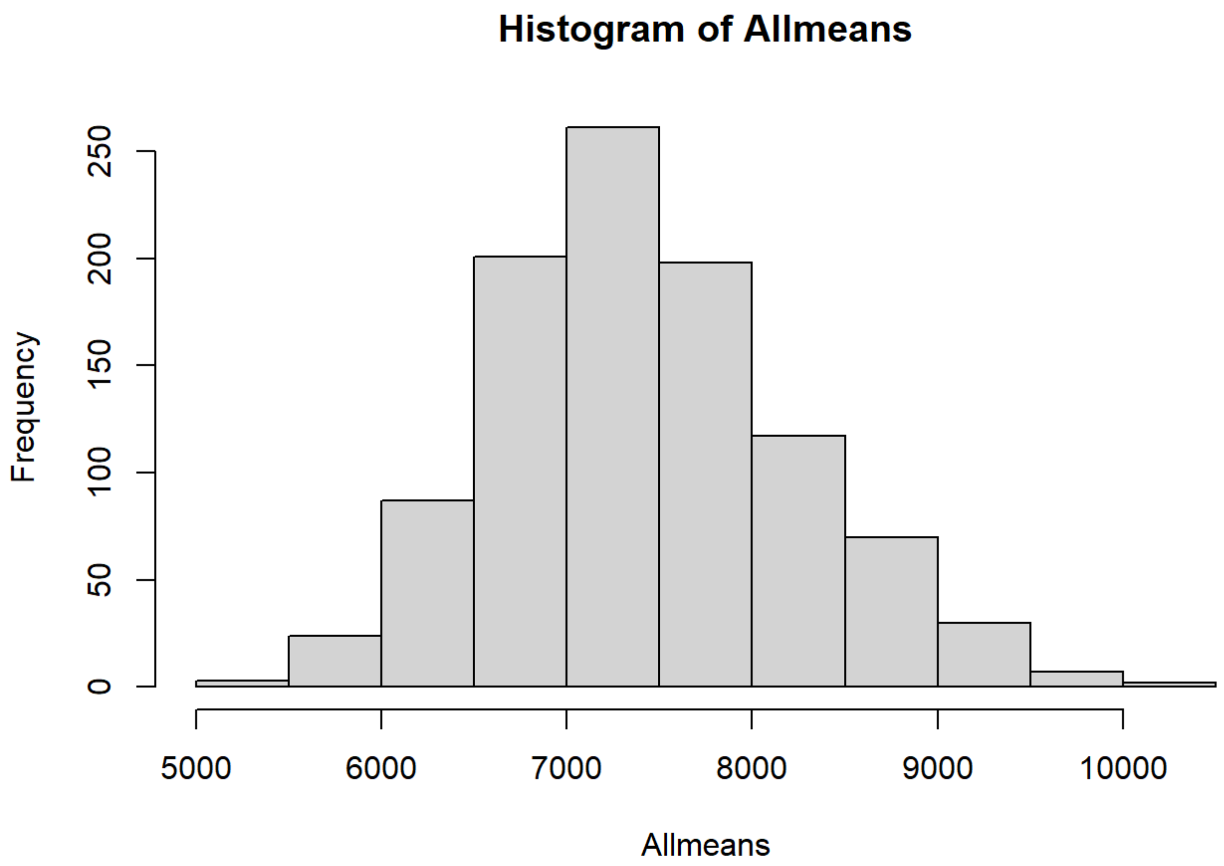
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
os <- na.omit(s$Earnings)

# point estimate
PE <- mean(os)
PE
```

```
## [1] 7458.578
```

```
# for loop for boostrap confidence
Allmeans <- c()
for (i in 1:1000){

  mySample <- mean(sample(os, size = length(os), replace = TRUE))
  Allmeans[i] <- mean(mySample)
}

# histogram
hist(Allmeans)
```



```
# Standard Error
SE <- sd(Allmeans)

# Margin of error
ME <- 2*SE

# upper and lower limit
lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,2)
upperlimit <- round(upperlimit,2)

# printing of limits
lowerlimit
```

```
## [1] 5828.18
```

```
upperlimit
```

```
## [1] 9088.98
```

```
#  Above this code chunk, state the point estimate, then state the interval in a co
mplete sentence in terms of the problem, as we’ve done in class.  (I’m 95% sure…..)
(Hint:   Remember that you need to specify that Earnings is in the data frame, s:
os <- s$Earnings and you may need to remove missing values)
```

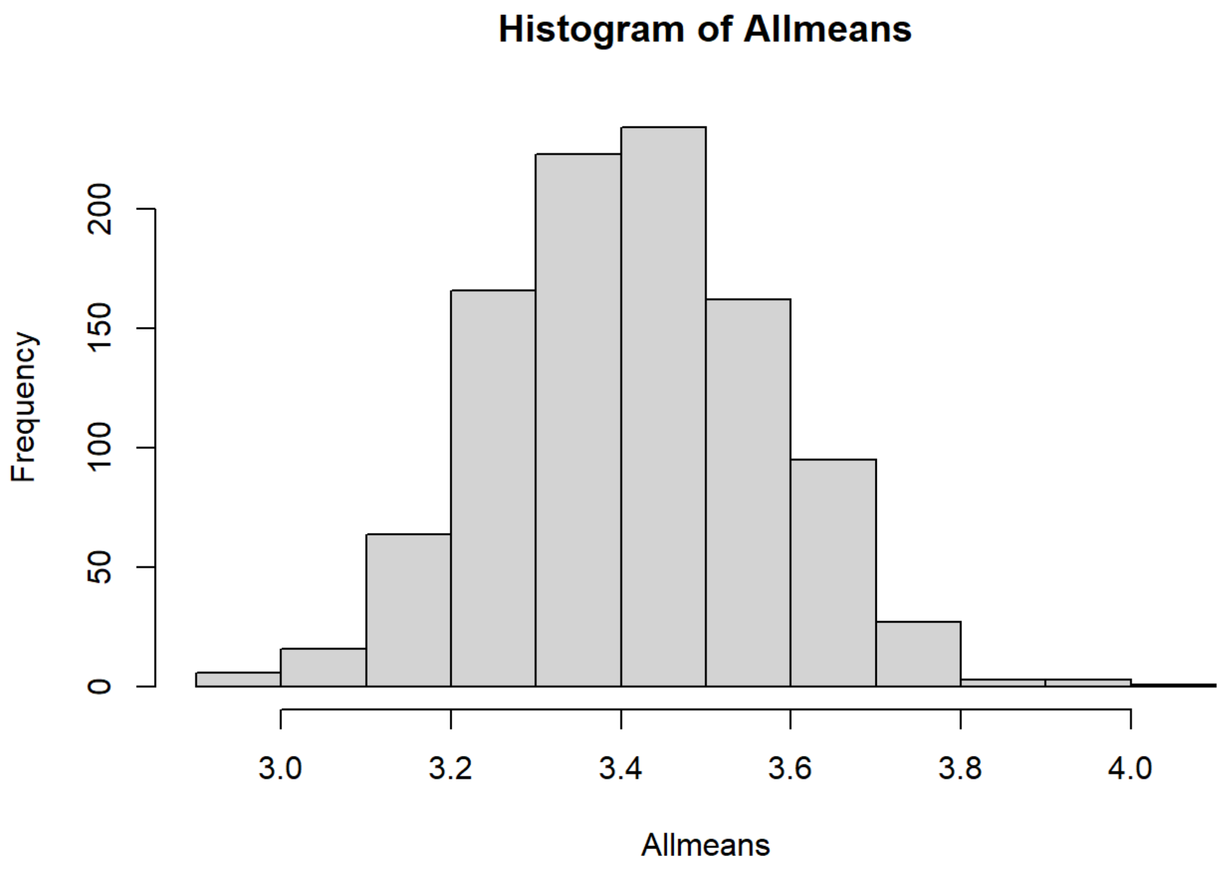
part b - CI for mean political leaning

The point estimate is 3.40201. I’m 95% confident that the true mean of political standing for UVM students is in an interval from 3.062 to 3.742, meaning that they tend to be more liberal as they are in the lower half of the scale.

```
# Once you have your code working for problem 3a, change it so that you can find the bootstrap confidence interval for the true mean response to the question ‘Circle where you political leaning falls, where 0 is most LIBERAL and 10 is most CONSERVATIVE.’ Produce the same output, described above. (Include the same output, including the histogram of the bootstrap means.)



os <- na.omit(s$Political)

PE <- mean(os)
PE
```

```
## [1] 3.40201
```

```
# for loop for bootstrap
Allmeans <- c()
for (i in 1:1000){

  samp <- mean(sample(os, size = length(os), replace = TRUE))
  Allmeans[i] <- mean(samp)
}

# histogram
hist(Allmeans)
```



```
# Standard Error
SE <- sd(Allmeans)

# Margin of error
ME <- 2*SE

lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,3)
upperlimit <- round(upperlimit,3)

#upper and lowerlimit
lowerlimit
```

```
## [1] 3.078
```

```
upperlimit
```

```
## [1] 3.726
```

```
#  Above this code chunk, state the point estimate, and state the interval in a com
plete sentence in terms of the problem, as we've done in class.  (I'm 95% sure…..)
If a value of 5 is 'average' or 'centrist,” does your confidence interval allow you
to say that UVM students consider themselves to be more liberal than average?  Expl
ain..
```

Question 4 – Confidence Intervals for Medians

part a - CI for Earnings median

The point estimate is 5000. I'm 95% confident that the true median earnings in dollars for UVM students is in an interval from 4390.76 to 5609.24.

```
# a.  Once you have your code working for problem 3, copy and change it so that you
can find the bootstrap confidence interval for the true median earnings here.  Abov
e, state the point estimate and write the Earnings confidence interval in words, in
terms of the problem here, as we've done in class.  (I'm 95% sure...)  You don't ne
ed to print bootstrap histograms here.

#
os <- na.omit(s$Earnings)

PE <- median(os)
PE
```

```
## [1] 5000
```

```
Allmeds <- c()
for (i in 1:1000){

  samp <- median(sample(os, size = length(os), replace = TRUE))
  Allmeds[i] <- median(samp)
}

# Standard Error
SE <- sd(Allmeds)

# Margin of error
ME <- 2*SE

lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,2)
upperlimit <- round(upperlimit,2)

lowerlimit
```

```
## [1] 4378.58
```

```
upperlimit
```

```
## [1] 5621.42
```

part b - CI for Political median

The point estimate is 3. I'm 95% confident that the true median of political standing for UVM students is in an interval from 2.7 to 3.3, meaning that they tend to be more liberal as they are in the lower half of the scale.

```
# b. Once you have your code working for problem 3, copy and change it so that you
can find the bootstrap confidence interval for the true median political leaning he
re.  Above, state each point estimate, and state the interval in a complete sentenc
e in terms of the problem, as we’ve done in class.  (I’m 95% sure…..)  you don't ne
ed to print bootstrap histograms here.
os <- na.omit(s$Political)

PE <- median(os)
PE
```

```
## [1] 3
```

```
Allmeds <- c()
for (i in 1:1000){

  samp <- median(sample(os, size = length(os), replace = TRUE))
  Allmeds[i] <- median(samp)
}

# Standard Error
SE <- sd(Allmeds)

# Margin of error
ME <- 2*SE

lowerlimit <- PE - ME
upperlimit <- PE + ME
lowerlimit <- round(lowerlimit,2)
upperlimit <- round(upperlimit,2)

lowerlimit
```

```
## [1] 2.69
```

```
upperlimit
```

```
## [1] 3.31
```

```
print(paste("I am 95% confident that the true median earnings of UVM students is wi
thin the range",lowerlimit, "to", upperlimit))
```

```
## [1] "I am 95% confident that the true median earnings of UVM students is within
the range 2.69 to 3.31"
```

part c - CIs for medians versus means

Answer the following questions in a one short paragraph:

Were your median point estimate and mean point estimate for earnings almost the same?

No, mean PE (7458.578) of earnings is about 1.4 times higher than the median PE (5 000). This shows that the mean is much more influenced by outliers than the median.

Were your median CI and mean CI for earnings almost the same?

Not really, my mean CI of earnings was broader than the median confidence interval. It was 1.5 times broader than the median confidence interval.

Were your median point estimate and mean point estimate for political almost the same?

They were similar, the median point estimate for political standing is 3 and the mean is roughly 3.4, having a difference of 1.13.

Were your median CI and mean CI for political almost the same?

The median confidence interval for political standing is just a bit broader than the mean confidence interval, 1.09 times broader to be exact.

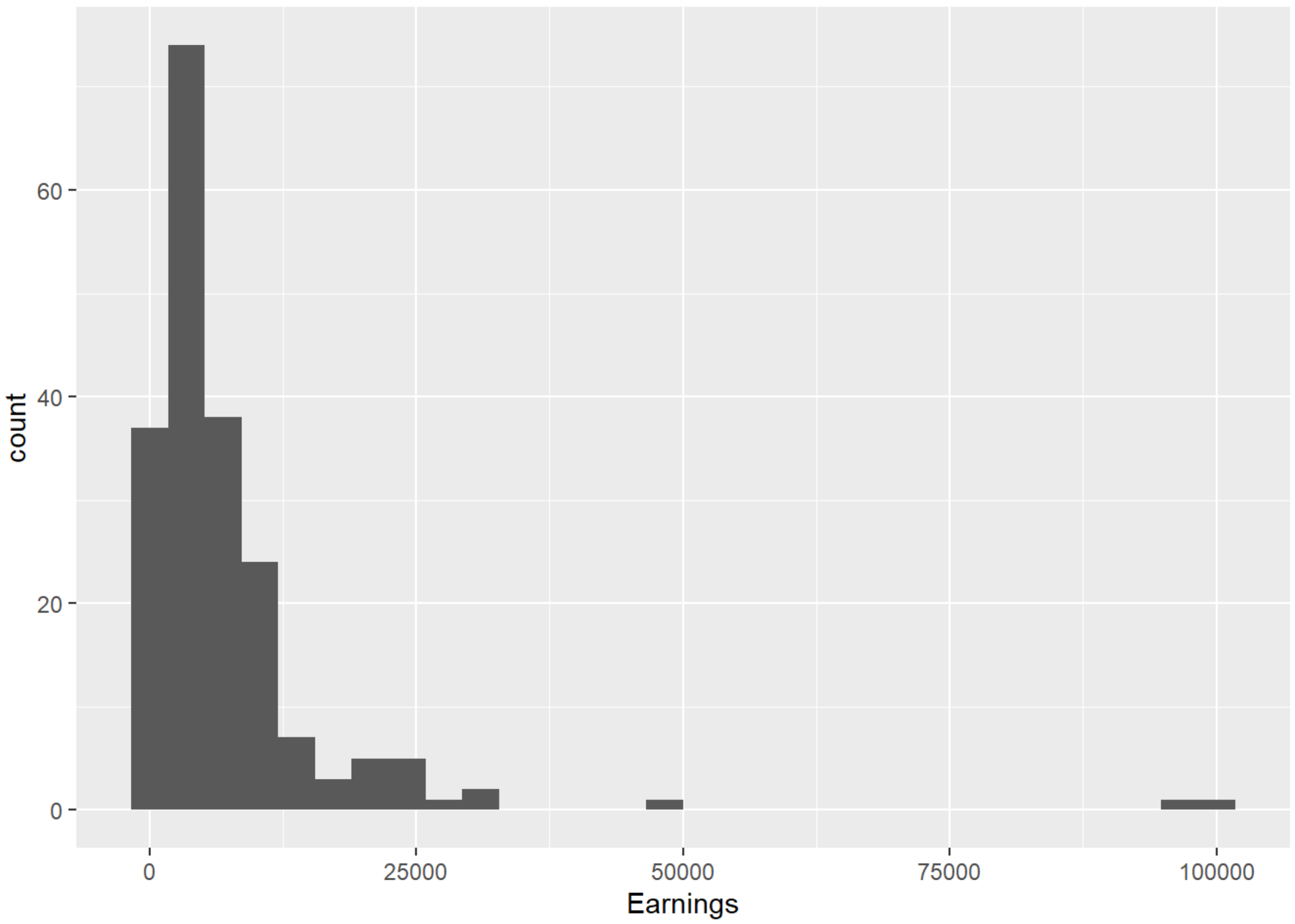
Explain above why they are very different in one case, and not a lot different in the other. (Hint: make a histogram of the original data for earnings and for political leaning.)

The political leaning values for mean and median is much more similar to the one compared to the income. I think one of the main reasons is because political has such a small range, from 1 to 10, while earnings varies has a bigger scope. The larger the scope the more the mean can vary and have outliers.

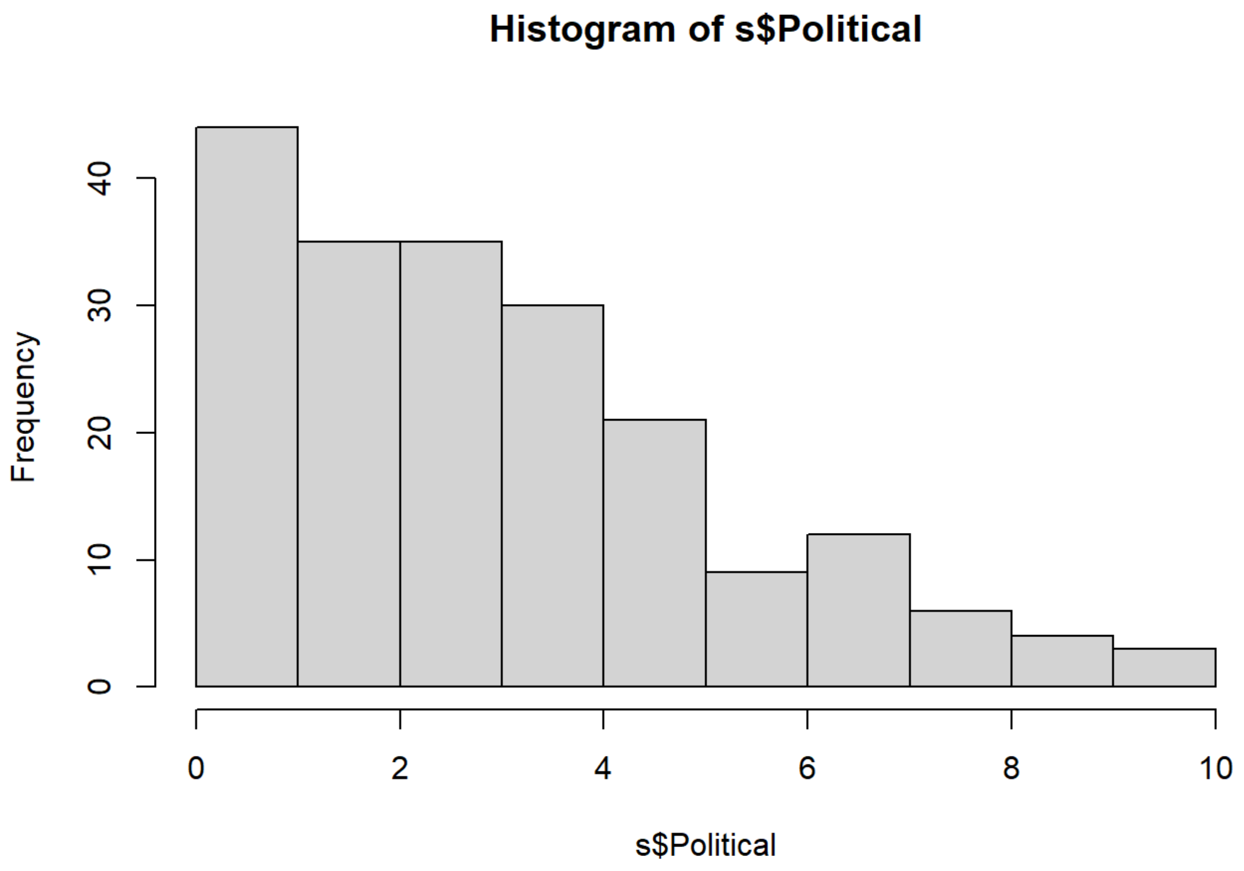
```
# histogram earnings
ggplot(data = s,
       mapping = aes(x = Earnings, color = Earnings)) +
geom_histogram()
```

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

```
## Warning: Removed 1 rows containing non-finite values (stat_bin).
```



```
# histogram political
hist(s$Political)
```


Histogram of s$Political
Frequency
0
10
20
30
40
0
2
4
6
8
10
s$Political