Unit 9 Pre-Class Warm-Up

Adam Yang

The file united_states_senate_2014_v2.csv contains data on the 100 members of the US senate that served in 2014. We will consider this group to be a sample (for example, from some generative process that creates senators).

```
S = read.csv("united_states_senate_2014_v2.csv")
summary(S)
```

```
##           Senator.Names    Gender          State            Party
##  Alan Franken    : 1    Female:20   Alabama   : 2   Democrat   :53
##  Amy Klobuchar   : 1    Male  :80   Alaska    : 2   Independent: 2
##  Angus King      : 1                Arizona   : 2   Republican :45
##  Barbara Boxer   : 1                Arkansas  : 2
##  Barbara Mikulski: 1                California: 2
##  Benjamin Cardin : 1                Colorado  : 2
##  (Other)         :94                (Other)   :88
##             Religion  Campaign.Money.Raised..millions.of...
##  Protestant     :49   Min.   : 0.100
##  Catholic       :27   1st Qu.: 4.575
##  Jewish         :10   Median : 7.550
##  Other Christian: 7   Mean   : 9.645
##  Mormon         : 2   3rd Qu.:13.800
##  Unaffiliated   : 2   Max.   :44.200
##  (Other)        : 3
##  Campaign.Money.Spent..millions.of...   NRA.Rating
##  Min.   : 0.200                       A      :34
##  1st Qu.: 2.975                       F      :34
##  Median : 6.000                       A+     : 9
##  Mean   : 8.227                              : 5
##  3rd Qu.:12.225                       AQ     : 5
##  Max.   :43.400                       C      : 3
##                                       (Other):10
```

You have three questions that you would like to answer with a statistical test.

Question 1: Is there a difference between the amount of money a senator raises and the amount spent?

Question 2: Do female Democratic senators raise more or less money than female Republican senators?

Question 3: Does the NRA prefer male senators or female senators?

For each question, answer the following using the dataset and your background knowledge:

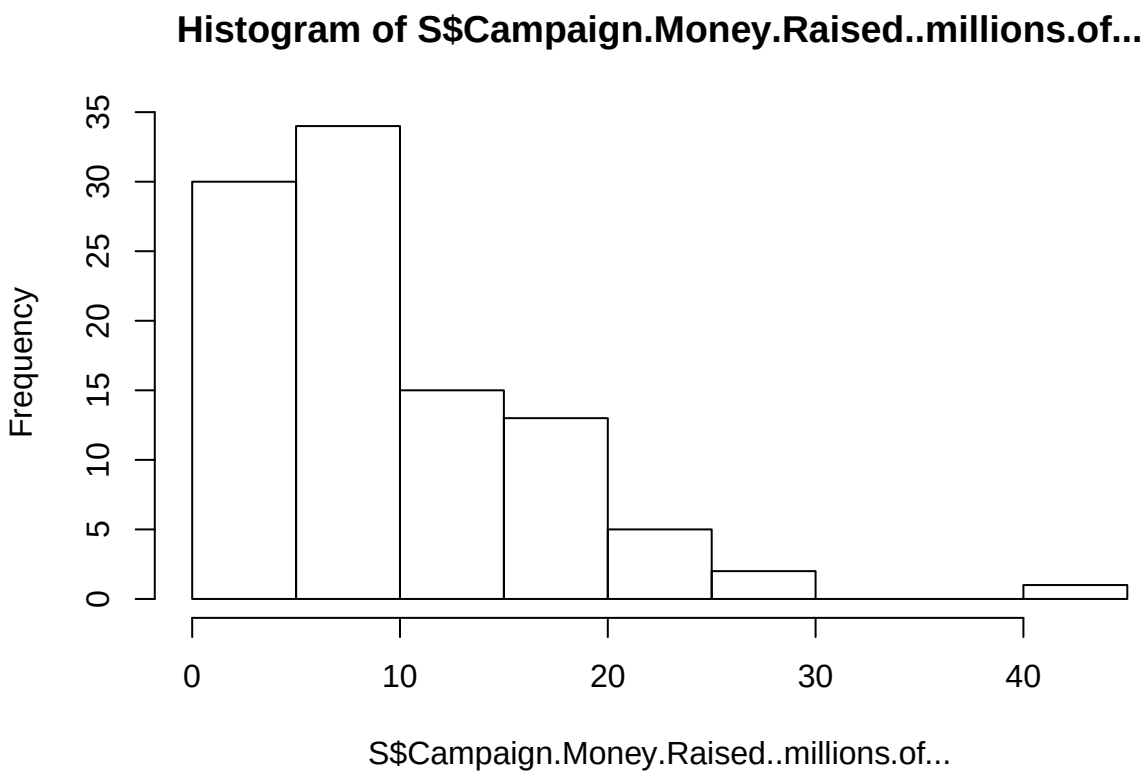
1. Are the assumptions for a t-test met? (you may want to review unit 9.5)
2. Is a paired test or an unpaired test more appropriate?
3. (Unless you argue that a t-test is clearly invalid), conduct a t-test in R and interpret the results.

Answer 1:

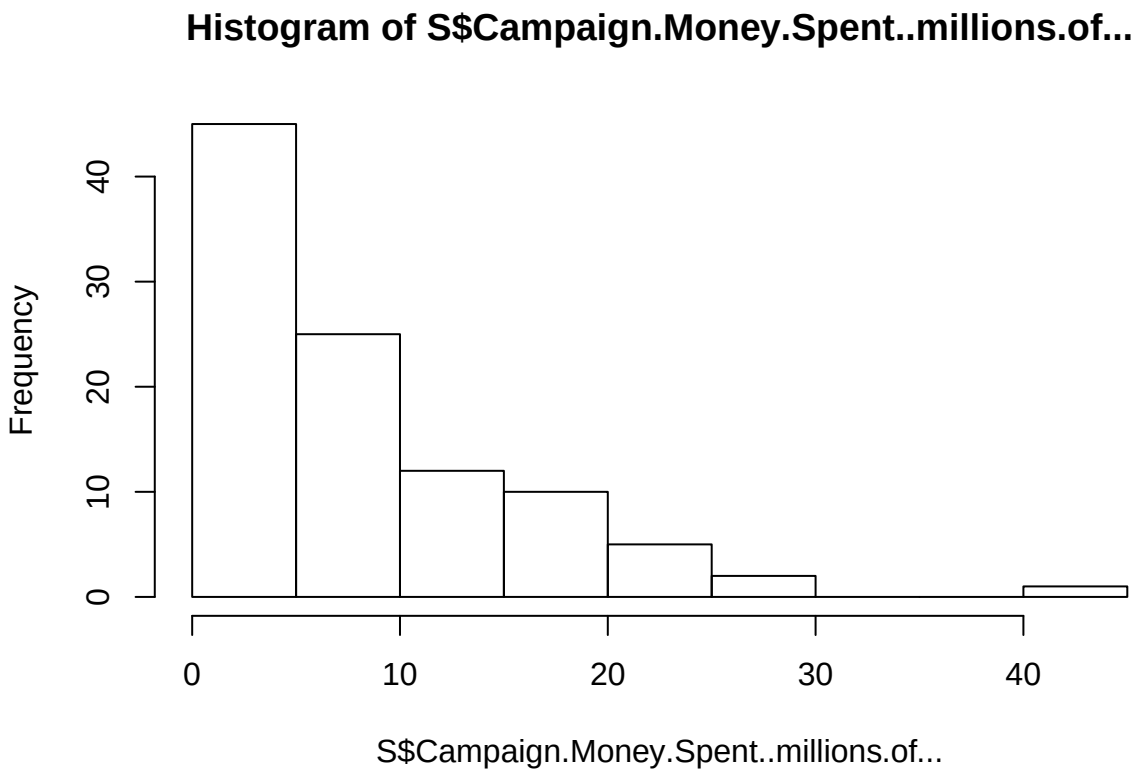
1. The first condition is random sampling. All of the campaign money raised data need to be identically and independently distributed and the campagin money spent data need to be identically and independently distributed. There is an argument that clustering can occur as candidates often fall into groups which can affect how much campagin money is raised. The second condition is normality as the two random

variables need to be drawn from normal distributions. In both cases, the histogram looks to be positively skewed. However, the sample size is 100 which may be big enough for the Central Limit Theorem to help us assume normality. Overall I think the assumptions for a t-test is not met, but it might be fun to see what the results are.

```
hist(S$Campaign.Money.Raised..millions.of...)
```




```
hist(S$Campaign.Money.Spent..millions.of...)
```



2. I believe a paired t-test would be more appropriate because the campaign money raised and spent per senator is linked and not independent. The amount of money spent would depend on how much money was raised because you can't spend money you don't have.

3.

```
t.test(S$Campaign.Money.Raised..millions.of..., S$Campaign.Money.Spent..millions.of..., paired = T)
```

```
##
##  Paired t-test
##
## data:  S$Campaign.Money.Raised..millions.of... and S$Campaign.Money.Spent..millions.of...
## t = 5.9944, df = 99, p-value = 3.329e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9486232 1.8873768
## sample estimates:
## mean of the differences
##                   1.418
```

Accroding to our t-test results, it seems like our p-value is highly significant so we can reject our null hypothesis that there is no difference between the money a senator raises and how much they spend. The result is not surprising to me, however, we must consider the fact that this t-test may not be valid because it is possible that the normality condition is not met. Furthermore, there might be clustering that affects our random sampling conditions.

Answer 2:

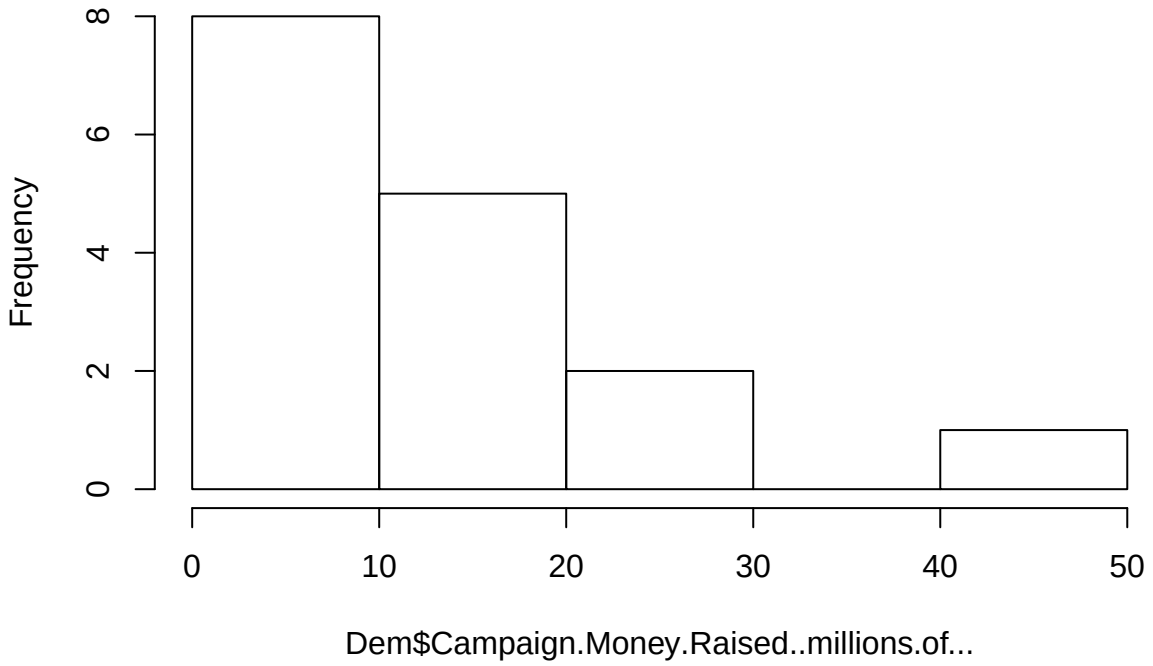
1.

```
# Make DataFrame for Females
Females <- S[S$Gender == "Female",]
# Make DataFrame for female Democrats
```

```r
Dem <- Females[Females$Party == "Democrat",]
# Make DataFrame for female republicans
Rep <- Females[Females$Party == "Republican",]

paste("The number of female Democrats are:", length(Dem))
```

```
## [1] "The number of female Democrats are: 8"
```

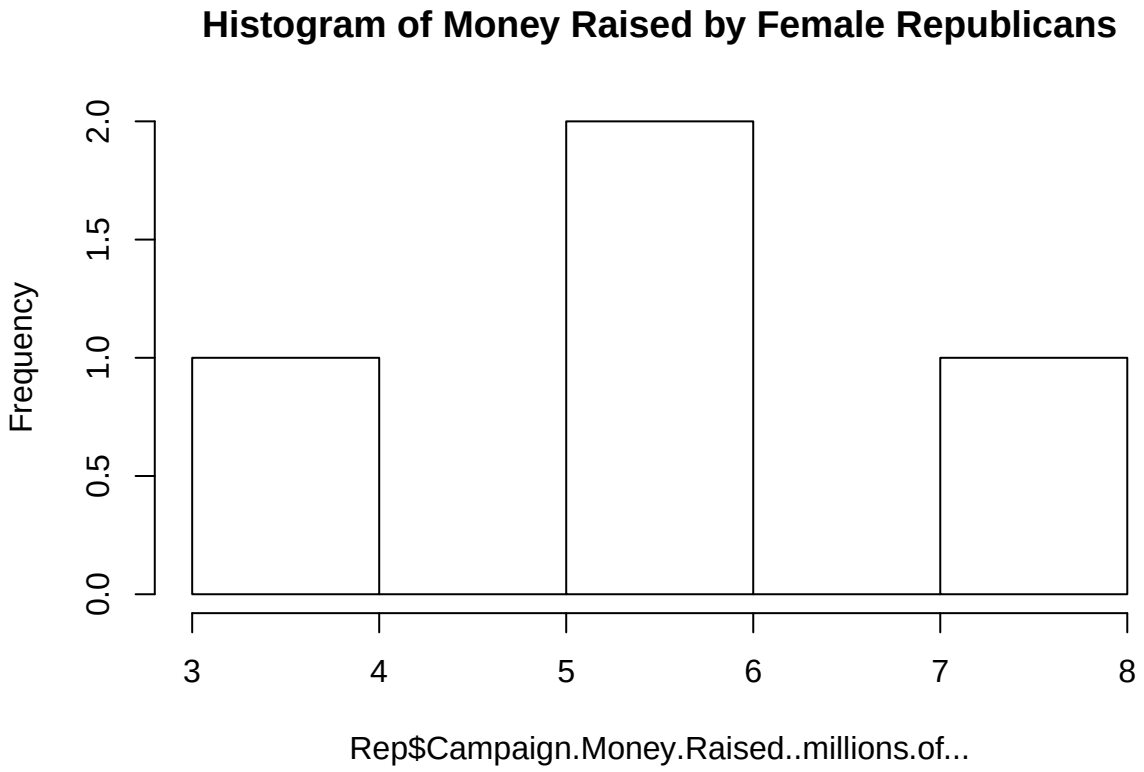
```r
hist(Dem$Campaign.Money.Raised..millions.of..., main = "Histogram of Money Raised by Female Democrats")
```




```r
paste("The number of female Democrats are:", length(Rep))
```

```
## [1] "The number of female Democrats are: 8"
```

```r
hist(Rep$Campaign.Money.Raised..millions.of..., main = "Histogram of Money Raised by Female Republicans
```



According to what is shown above, it looks like our female Democrats distribution has a positive skew. Furthermore for both parties, the sample size is only 8 which is not close to enough for the Central Limit Theorem to take effect. Therefore, the normality conditions are not met. I do not think the assumptions for a t-test are met in this case.

2. In this case I would not use a paired t-test because it does not seem like the two data samples are linked to each other.

3. I think the t-test is clearly invalid based on how small the sample size is for both groups.

Answer 3:

1.

```
# Make DataFrame for Females
Females <- S[S$Gender == "Female",]
# Make DataFrame for Males
Males <- S[S$Gender == "Male",]

# Function that turns a letter grade into a number
Numerify <- function(grade) {
  if(grade == "A+") {
    return(4.3)
  }
  else if(grade == "A"){
    return(4.0)
  }
  else if(grade == "AQ") {
    return(4.0)
  }
  else if(grade == "A-") {
    return(3.7)
```

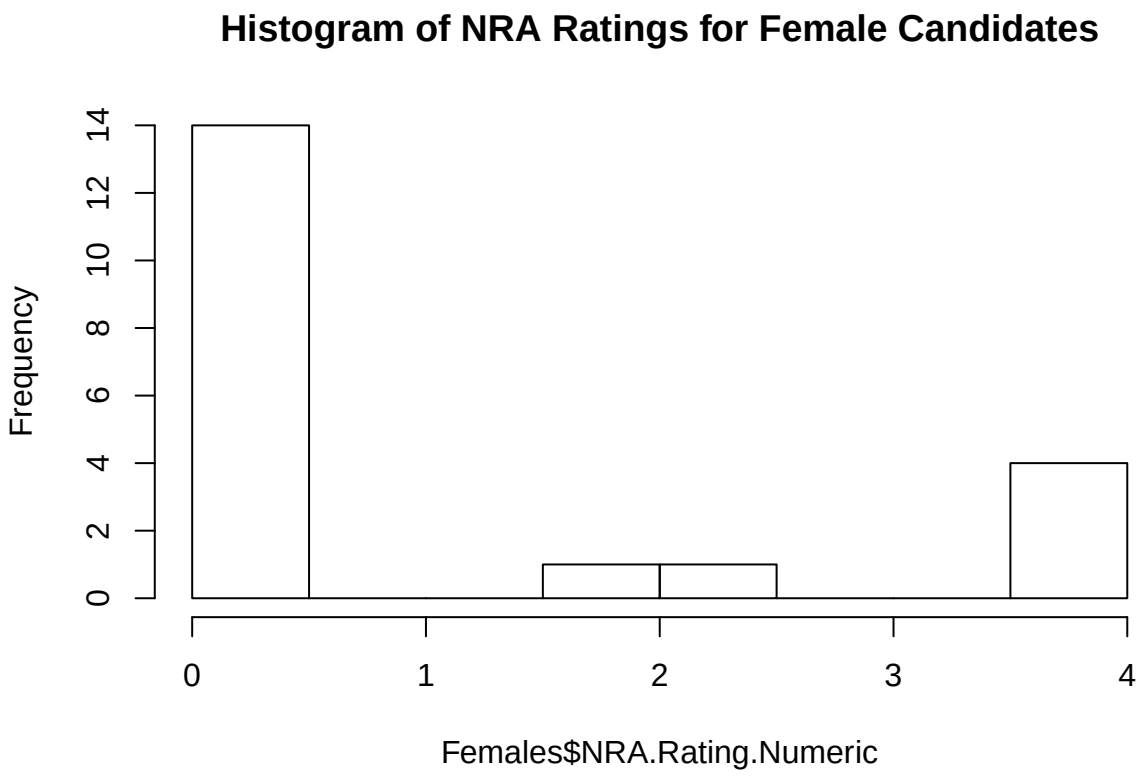
```
  }
  else if(grade == "B+") {
    return(3.3)
  }
  else if(grade == "B") {
    return(3.0)
  }
  else if(grade == "B-") {
    return(2.7)
  }
  else if(grade == "C+") {
    return(2.3)
  }
  else if(grade == "C") {
    return(2.0)
  }
  else if(grade == "C-") {
    return(1.7)
  }
  else if(grade == "D+") {
    return(1.3)
  }
  else if(grade == "D") {
    return(1)
  }
  else if(grade == "D-") {
    return(0.7)
  }
  else if(grade == "F") {
    return(0)
  }
  else {
    return(NA)
  }
}

# Add a new column to Males and Females df, that holds the numeric version of the NRA Grade.
Males$NRA.Rating.Numeric <- mapply(Numerify, Males$NRA.Rating)
Females$NRA.Rating.Numeric <- mapply(Numerify, Females$NRA.Rating)

# Show the sample size of the NRA Grades (get rid of the NA values)
paste("The sample size for female candidates is:", length(na.omit(Females$NRA.Rating.Numeric)))
```

```
## [1] "The sample size for female candidates is: 20"
```

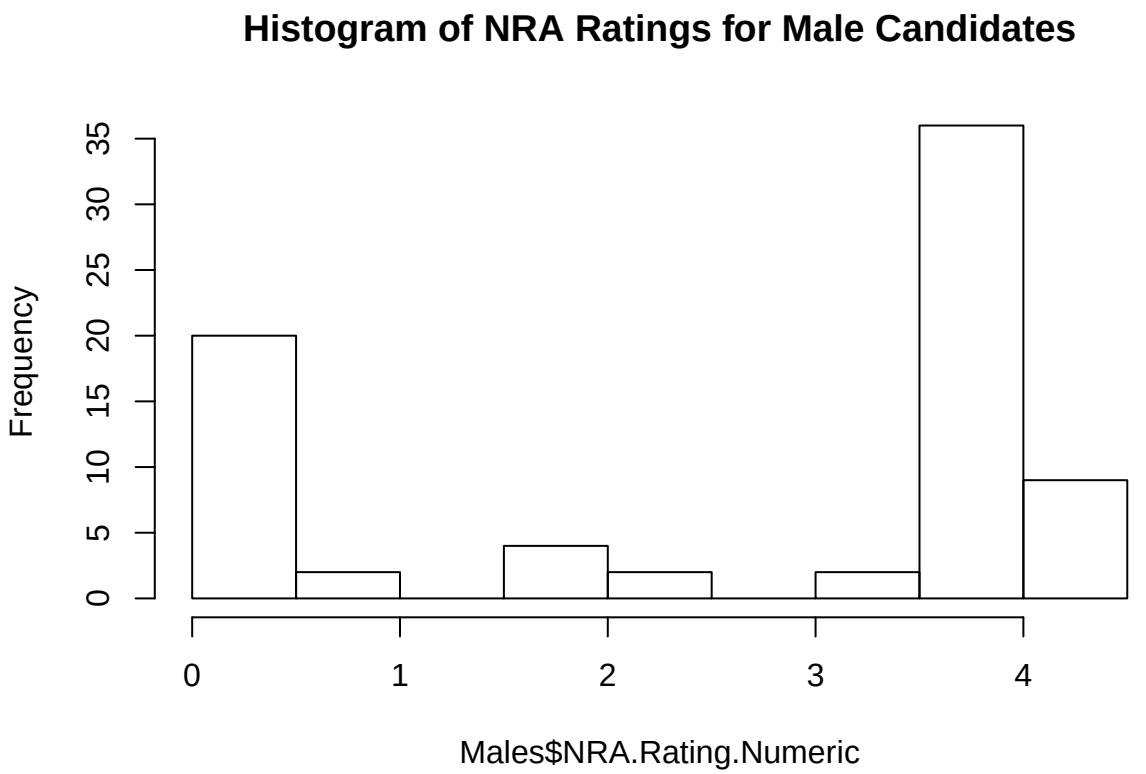
```
paste("The sample size for male candidates is:", length(na.omit(Males$NRA.Rating.Numeric)))
```

```
## [1] "The sample size for male candidates is: 75"
```

```
hist(Females$NRA.Rating.Numeric, main = "Histogram of NRA Ratings for Female Candidates")
```




```
hist(Males$NRA.Rating.Numeric, main = "Histogram of NRA Ratings for Male Candidates")
```



From What is shown above, the 2 distributions are clearly not normal. It seems like the NRA mostly gives As or Fs to candidates and very rarely assign other ratings. The random sampling condition does not seem to be met because there are clusters in both groups for men and women. Usually, if the candidate is a Democrat, he or she will have a poorer NRA rating and if the candidate is a Republican, he or she will have a better

NRA rating. Furthermore, the normality condition doesn't seem to have been met either because the sample size for female candidates is only 20 while the male candidate's sample size is 75. Since both distributions are far from normal, we would need a sample size of at least 30 for the Central Limit Theorem to be valid. Therefore, I don't think a t-test is completely vaid in this scenario.

2. I definitely think an unpaired t-test would be better suited because the two groups are not tied together. The sample size of the two groups aren't even the same.

3. I don't think a t-test is valid in this scenario, but I want to do it just to see what it looks like.

```
t.test(Males$NRA.Rating.Numeric, Females$NRA.Rating.Numeric)
```

```
##
##  Welch Two Sample t-test
##
## data:  Males$NRA.Rating.Numeric and Females$NRA.Rating.Numeric
## t = 3.9574, df = 31.898, p-value = 0.0003964
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8188796 2.5564537
## sample estimates:
## mean of x mean of y
##  2.702667  1.015000
```

According to the t-test, we have a p-value smaller than 0.05 so we can reject the null hypothesis that the two means are the same. However, I still don't think a t-test is valid in this scenario.