two sample t-test

K Iwasaki

September 11, 2017

```
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
```

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

Dependent t-test (paird t-test)

0. Analyse the problem.

There is a claer pairing in this case. Money raised vs. money spent for each senator. Since we have a large sample, we can use CLT to justify parametric test.

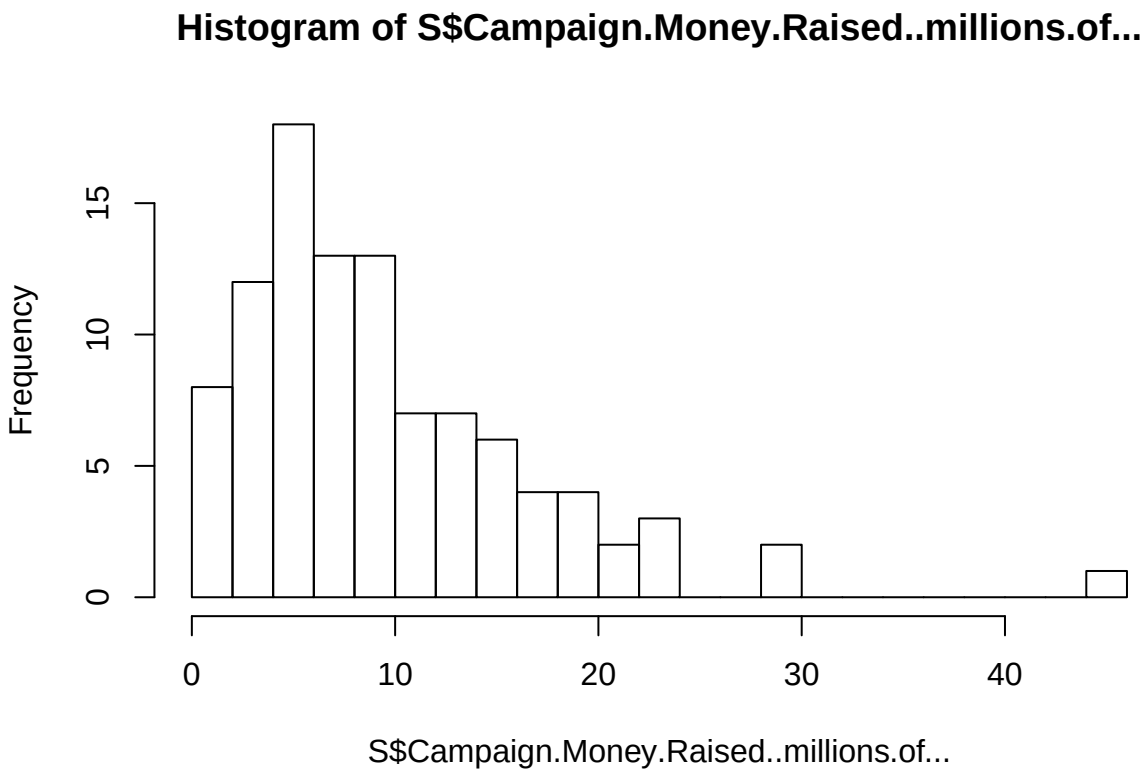
1. Construct hypothesis.

H_0 is there is no difference between them.

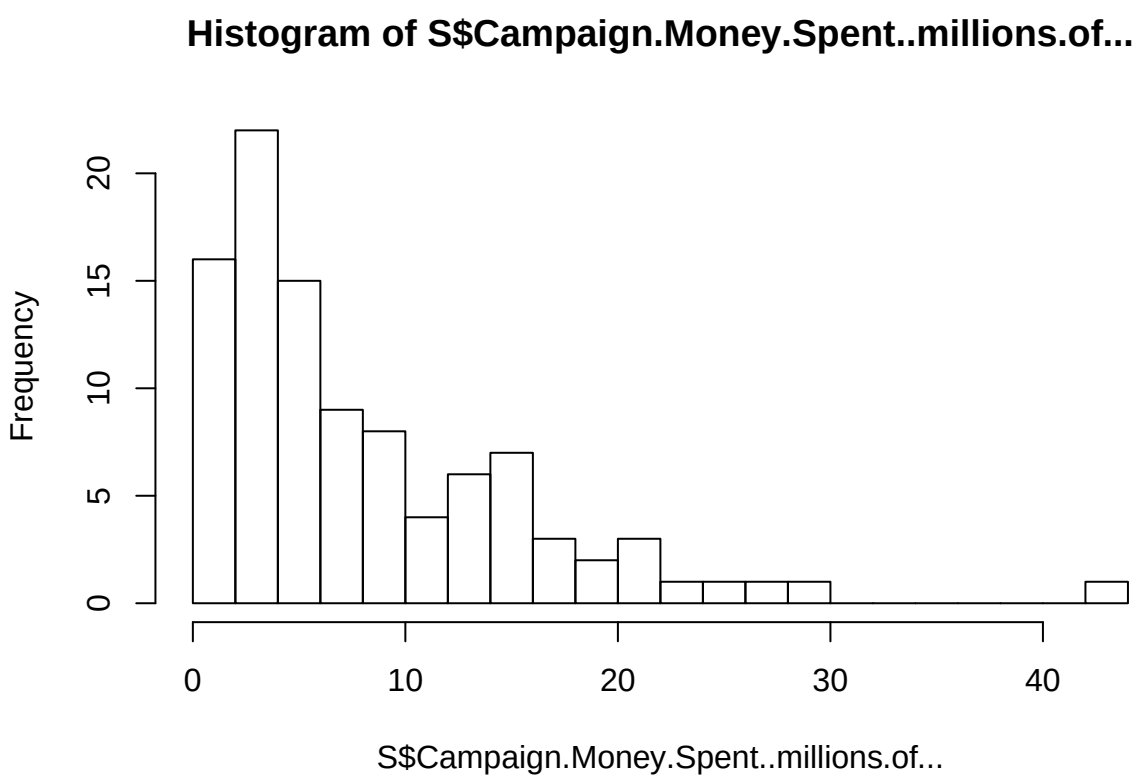
2. Check the assumption of normality

The distributions for the money raised and spent are not normal but should be fine because of CLT.

```
hist(S$Campaign.Money.Raised..millions.of..., breaks = 20)
```

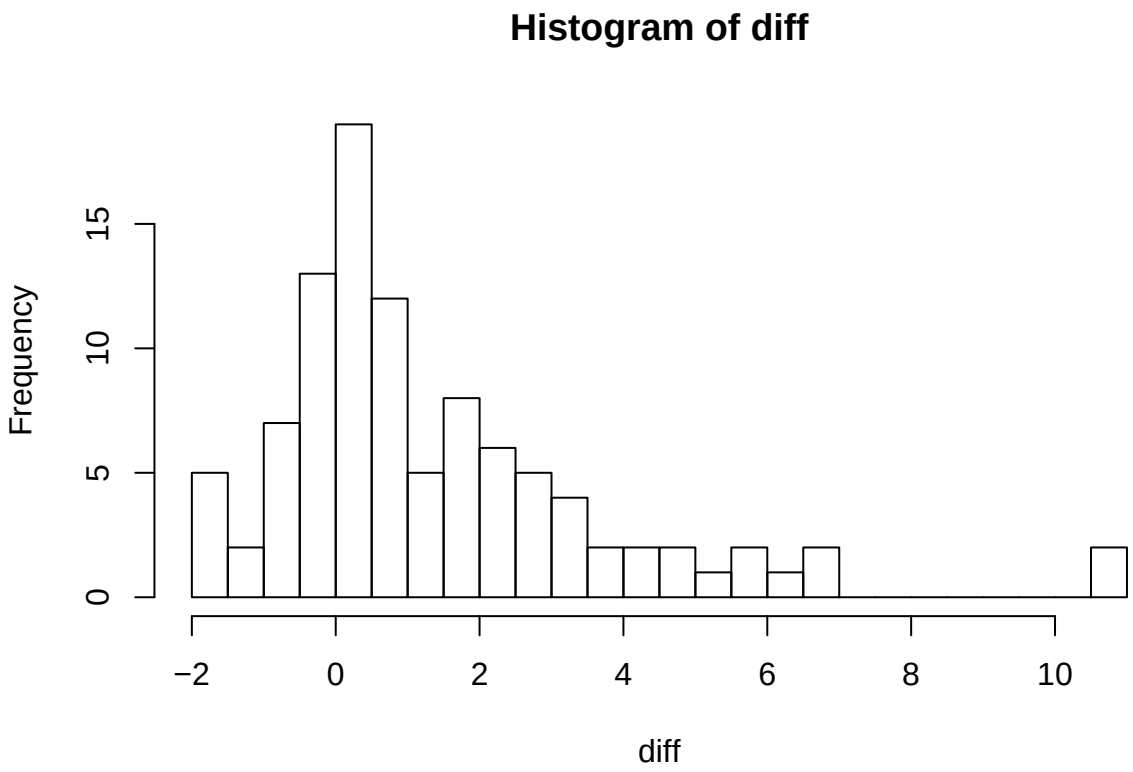



```
hist(S$Campaign.Money.Spent..millions.of..., breaks = 20)
```




```
diff = S$Campaign.Money.Raised..millions.of... - S$Campaign.Money.Spent..millions.of...

hist(diff, breaks = 20)
```



3. Check the assumption of equal variances.

Since they are paird. This is ok. There is onlly one distribution to test.

4. Run the t.test

```
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
```

p-value is very small. Reject null hypothesis that there is no difference in the money raised and money spent for the senatros.

5. Calculate Effect Size

The differences in the mean is 1.418

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

Independent t-test

0. Analyze the problem

The data we are interested in are not paired.

1. Construct hypothesis.

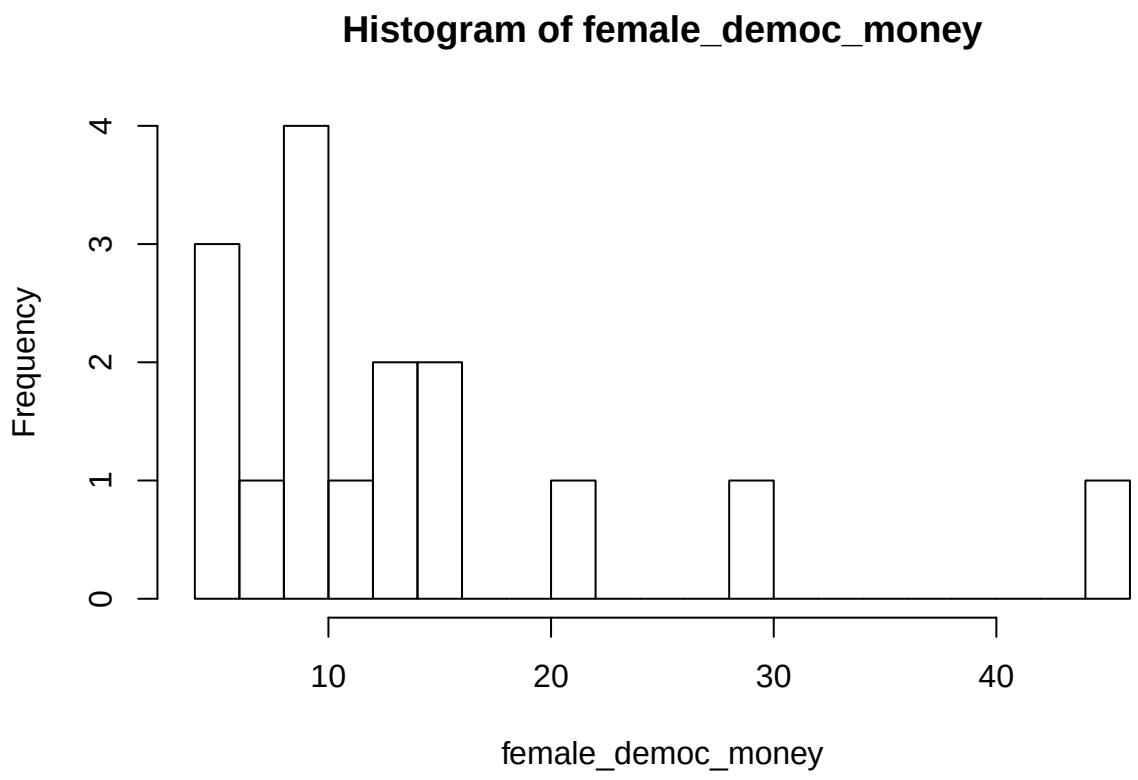
H_0 is there is no difference between average money raised by female democratic senators and average money raised by female republican senators. Two tailed test since we don't know less or more. Also, we doon't have many observation.

2. Check the assumpiton of normality

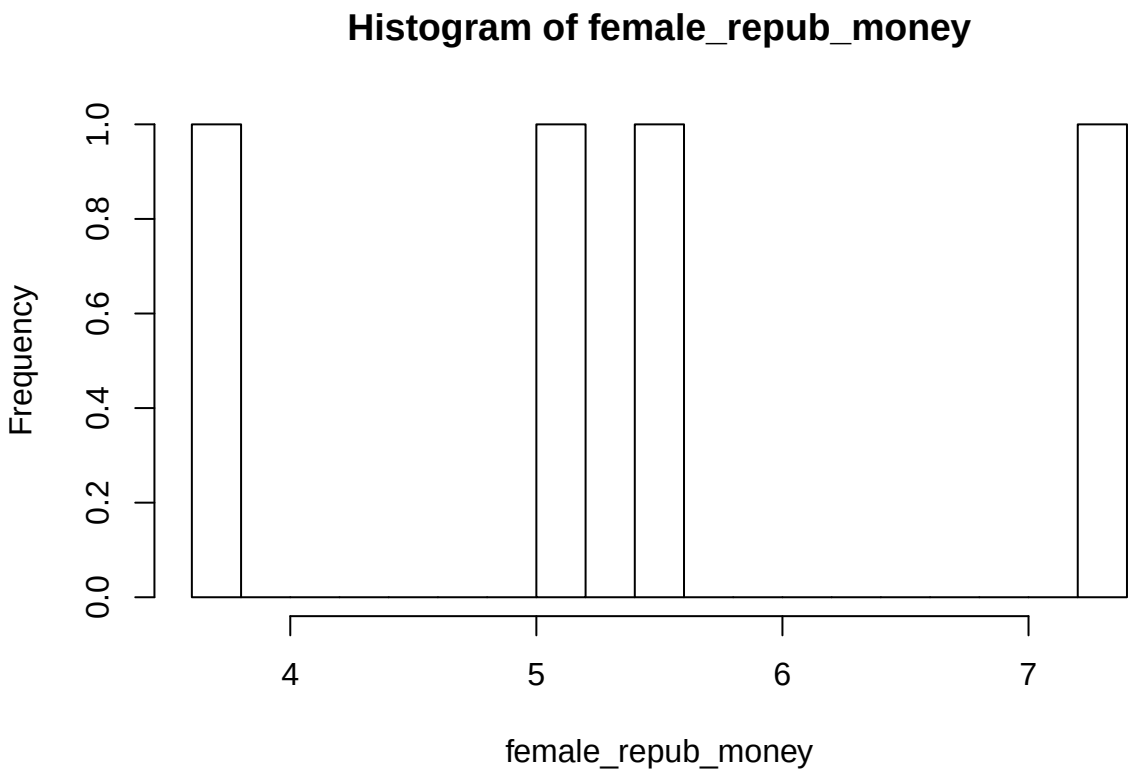
```
female_democrat_mask = (S$Gender == "Female") & (S$Party == "Democrat")
female_republican_mask = (S$Gender == "Female") & (S$Party == "Republican")

female_democ_money = S[female_democrat_mask, ]$Campaign.Money.Raised..millions.of...
female_repub_money = S[female_republican_mask, ]$Campaign.Money.Raised..millions.of...

hist(female_democ_money, breaks = 20)
```




```
hist(female_repub_money, breaks = 20)
```



3. Check the assumption of equal variances

```
library(car)
leveneTest(Campaign.Money.Raised..millions.of... ~ Party, data = S[S$Gender=="Female",])
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.5374 0.2309
##       18
```

```
wilcox.test(Campaign.Money.Raised..millions.of... ~ Party, data = S[S$Gender=="Female",])
```

```
## Warning in wilcox.test.default(x = c(15.3, 13.8, 11.7, 9.7, 29.7, 9.9,
## 6.2, : cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Campaign.Money.Raised..millions.of... by Party
## W = 58, p-value = 0.01593
## alternative hypothesis: true location shift is not equal to 0
```

```
t.test(Campaign.Money.Raised..millions.of... ~ Party, data = S[S$Gender=="Female",])
```

```
##
##  Welch Two Sample t-test
##
## data:  Campaign.Money.Raised..millions.of... by Party
## t = 3.2476, df = 17.102, p-value = 0.004708
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.035148 14.277352
## sample estimates:
##   mean in group Democrat mean in group Republican
##                 14.08125                  5.42500
```