Testing whether average number of friends differ by sex

Daniel Oyeleye

2023-10-12

This markdown file is a report of some sorts on the results of simple data analysis. The Friends.csv file has 2 variables, gender and number of friends of each experimental unit.

I will begin by importing the necessary libraries.

```
library(ggplot2)
library(dplyr)
library(mosaic)
library(readr)
```

I calculated mean, median and so on for number of friends, ran a t-test and ANOVA to test whether the difference in number of friends between genders is statistically significant.

```
Friends = read.csv("C:\\Users\\danie\\Documents\\data projects\\Friends survey analysis\\Friends.csv")
summary(Friends)
```

```
##     gender             friends
##  Length:27112       Min.   :  0.000
##  Class :character   1st Qu.:  2.000
##  Mode  :character   Median :  5.000
##                     Mean   :  6.476
##                     3rd Qu.:  8.000
##                     Max.   :200.000
```

```
t.test(friends ~ gender, data=Friends)
```

```
##
##  Welch Two Sample t-test
##
## data:  friends by gender
## t = -7.0094, df = 22408, p-value = 2.462e-12
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -1.0700567 -0.6023839
## sample estimates:
## mean in group female   mean in group male
##              6.09306              6.92928
```

```r
summary(aov(friends ~ gender, data = Friends))
```

```
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## gender          1    4705    4705   51.67 6.74e-13 ***
## Residuals   27110 2468807      91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

t test at 95% confidence level (0.05 significance level) testing if population mean is 2

```r
t.test(Friends$friends, mu=2)
```

```
##
##  One Sample t-test
##
## data:  Friends$friends
## t = 77.151, df = 27111, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 2
## 95 percent confidence interval:
##  6.361843 6.589249
## sample estimates:
## mean of x
##  6.475546
```

Results show that the difference between the number of friends by gender on average was statistically significant. Now which gender has more friends on average? Lets run a t-test to find out

```r
t.test(friends ~ gender, data = Friends, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  friends by gender
## t = -7.0094, df = 22408, p-value = 1
## alternative hypothesis: true difference in means between group female and group male is greater than
## 95 percent confidence interval:
##  -1.032459       Inf
## sample estimates:
## mean in group female   mean in group male
##              6.09306              6.92928
```

The results show that number of friends of male subjects is greater than the number of friends of female units.

The prior analysis was done using our entire sample would the results change if we used a portion of our sample? Let's find out by selecting 460 and 540 female and male units and find out how many friends they have.

```r
n_male = 460; n_fmale = 540;
male_friends <- Friends %>% filter(gender == "male");
female_friends <- Friends %>% anti_join(male_friends, by = "gender");
```

```
male_friends <- male_friends %>% select(friends);
female_friends <- female_friends %>% select(friends);
female_friends <- female_friends %>% select(friends);
male_sample <- male_friends %>% sample_n(n_male);
female_sample <- female_friends %>% sample_n(n_fmale);
```

We then get a summary of the statistics from the samples

```
favstats(male_sample$friends); # male sample
```

```
##  min   Q1 median Q3 max     mean       sd   n missing
##    0 2.75      5  8 100 6.608696 8.086348 460       0
```

```
favstats(female_sample$friends); # female sample
```

```
##  min Q1 median Q3 max mean       sd   n missing
##    0  2      5  7 100    6 7.601386 540       0
```

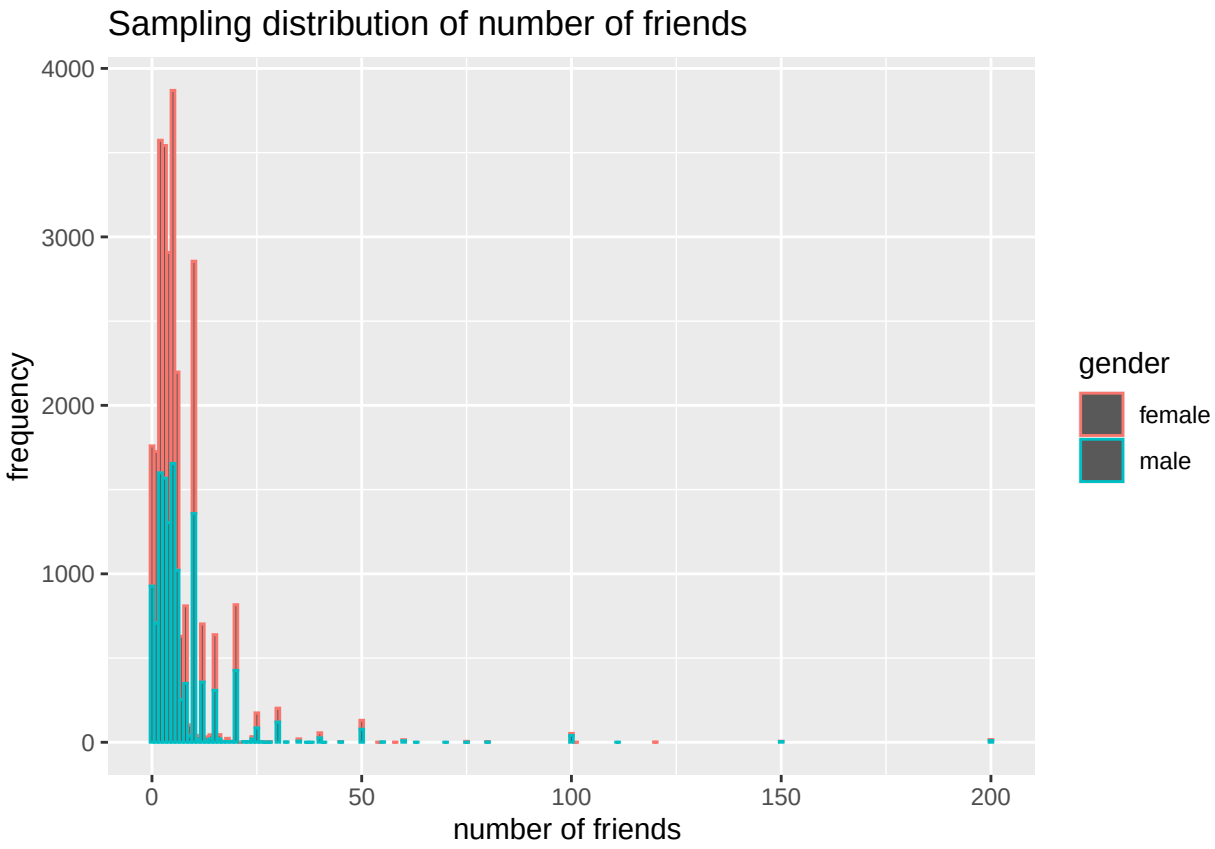
We then use a two sided t-test to find if there is a difference in the average number of friends between genders with a .05 significance level

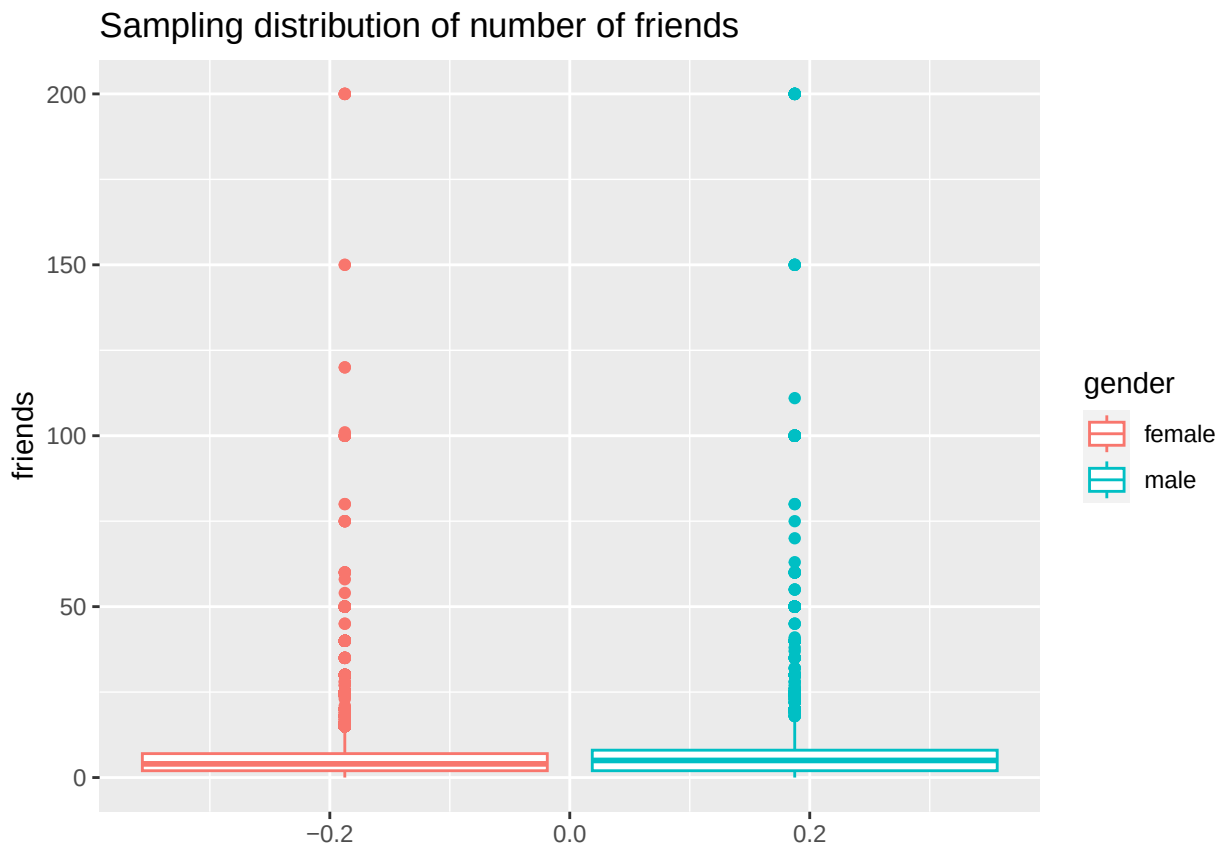
```
t.test(female_sample, male_sample)
```

```
##
##  Welch Two Sample t-test
##
## data:  female_sample and male_sample
## t = -1.2195, df = 951.15, p-value = 0.223
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.5882608  0.3708695
## sample estimates:
## mean of x mean of y
##  6.000000  6.608696
```

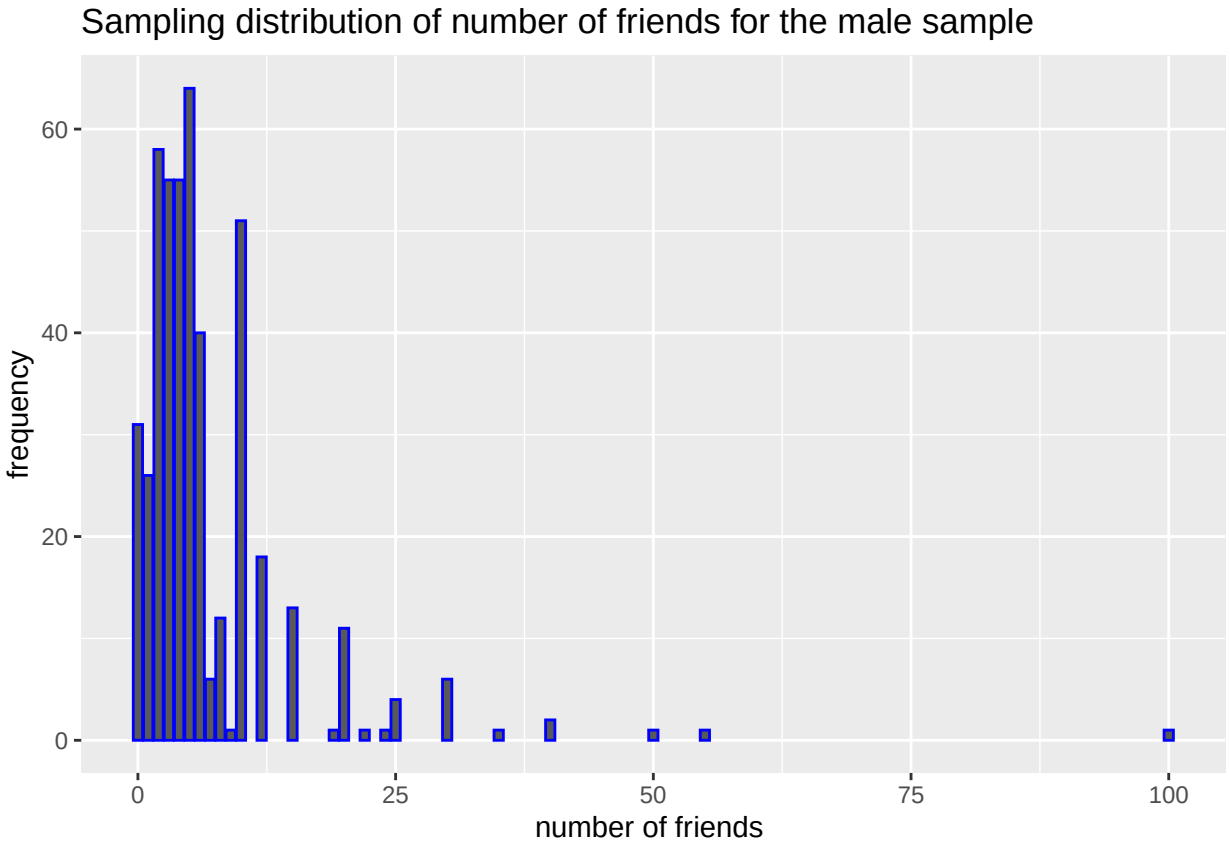
The result show that the difference in number of friends on average between genders in the sample is not significant. ## Plots

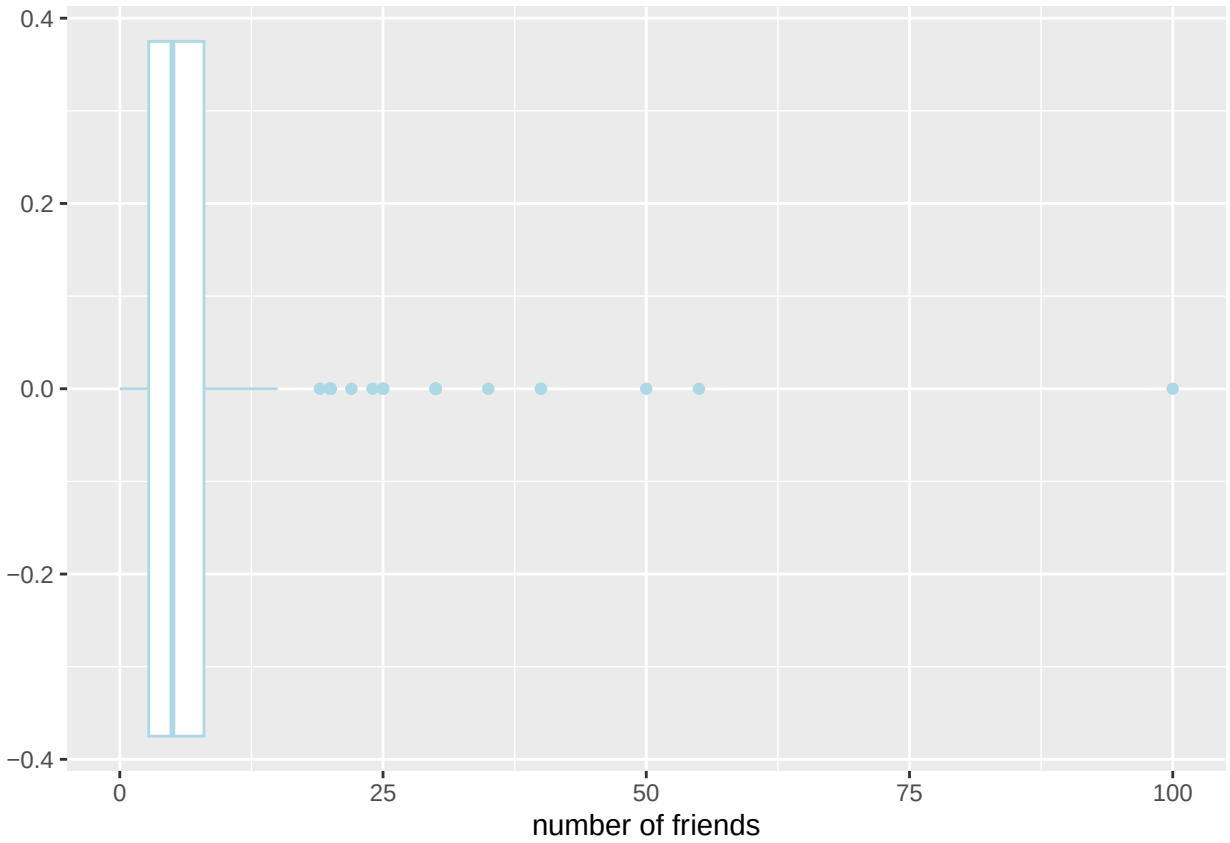
You can also embed plots, for example: The histogram and boxplot of the number of friends by gender

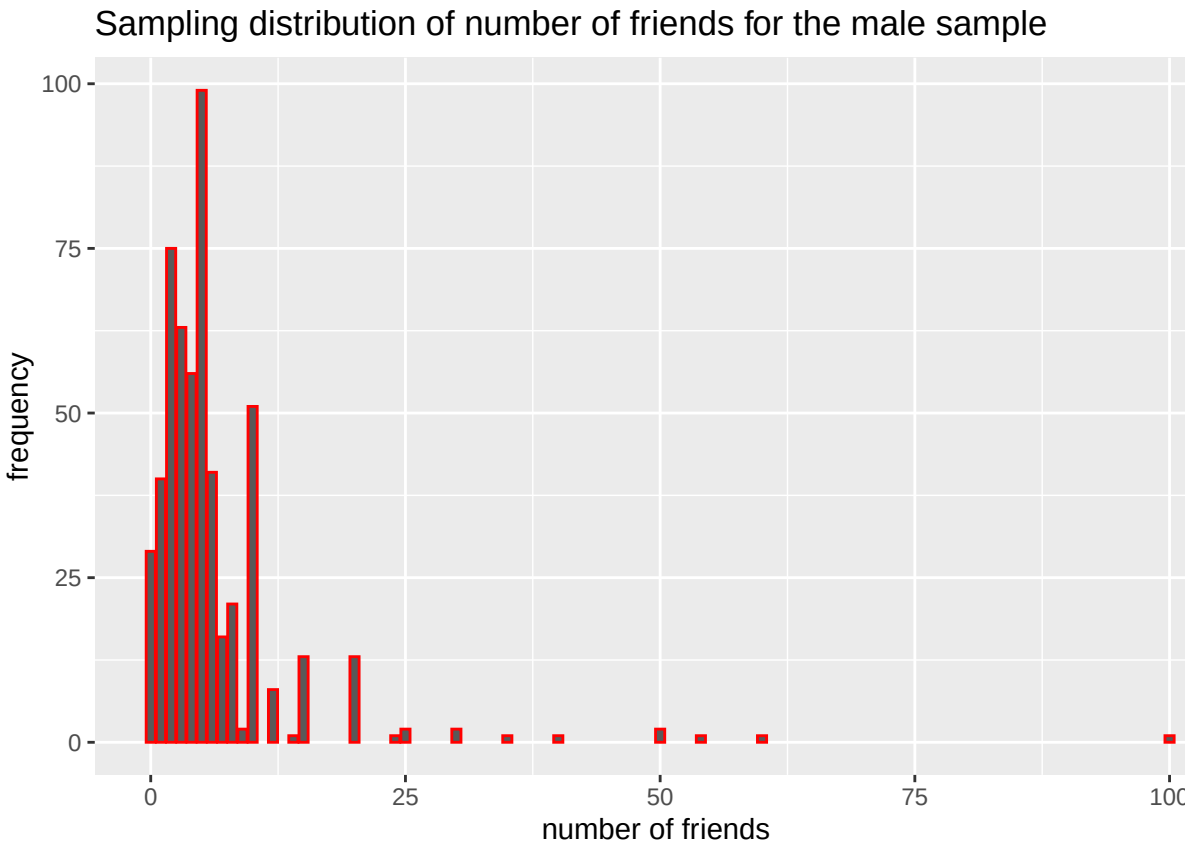

Sampling distribution of number of friends
frequency
4000
3000
2000
1000
0
0
50
100
150
200
number of friends
gender
female
male



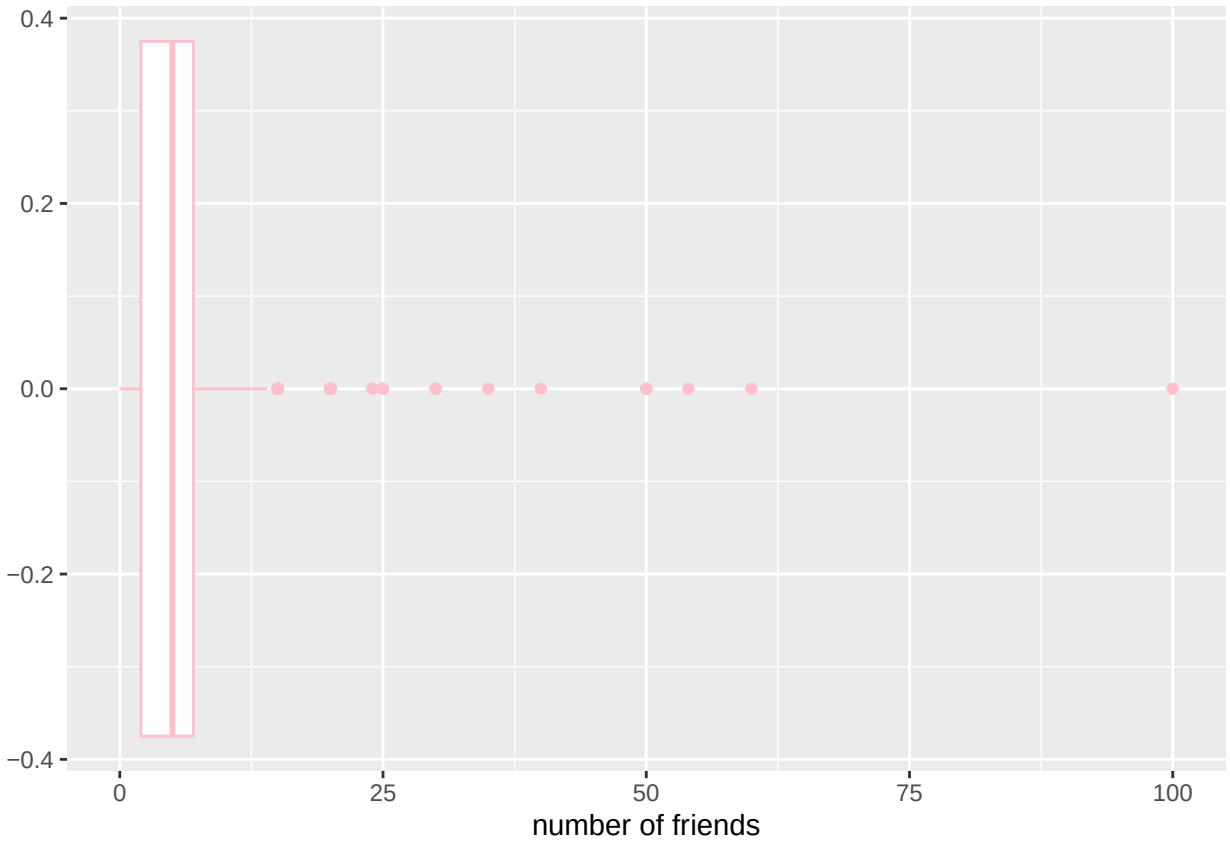
I also plotted the histogram of the distribution of number of friends for each sample Male sample plots:


Sampling distribution of number of friends for the male sample
frequency
60
40
20
0
0
25
50
75
100
number of friends


0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100
number of friends



Female sample plots:


0.4
0.2
0.0
−0.2
−0.4
0
25
50
75
100
number of friends