Lab10

Insert your name here

10/20/2020

Learning Objective

- Conduct a Chi squared analysis to perform an inferential statistical test (i.e., a hypothesis test) about a categorical population.
- Conduct a t-test to perform a hypothesis test concerning a difference between two means.
- Given a confidence interval, calculate *alpha* or the significance level.

Question 1: M&M Demo Open up the pdf file posted on Moodle titled "Lab10M&Mactivity". One person in your breakout room should open the Google Sheet link (provided on Zoom) and on Moodle, create a copy for your breakout room, share the sheet with the other members of the breakout room and Prof. Sills. The Google Sheet is set up so you can each input the data for your bag of M&Ms (one tab per person) and then pool your data in the "pooled data tab."

Write your names in the PooledData tab of the sheet, and label each tab with your names.

ANSWER THE QUESTIONS FROM THE M&M HANDOUT HERE:

a. Based on the results displayed in your table, which distribution in the table does your sample seem to match best (the 2006 distribution, a uniform distribution, the NJ 2017 distribution, or the TN 2017 distribution)? Why?

My sample most accurately matches the 2017 Distribution, TN distribution because my distribution is not uniform and the proportions are closer to the TN then the NJ distribution.

---

b. In your M&M experiment, what is the number of degrees of freedom?

6-1 = 5

---

c. Based on the table on p.4 of the M&M document, the approximate probability of drawing a sample with your χ^2 value lies between what two probabilities?

0.01 and 0.9

---

d. Using the "pchisq" function in R, what is the probability of drawing a sample with your χ^2 value?

```r
pchisq(6.1,5,lower.tail = FALSE)
```

```
## [1] 0.2966098
```

---

e. Based on the table on p. 4 of the M&M document, what is the critical value of χ^2 for a significance level of 0.05?

11.07

---

f. Based on your analysis, should you reject (or fail to reject) the null hypothesis that the M&M distributions match the published values? Why?

p = 0.297 >0.05 so the null hypothesis is not rejected

---

g. State your conclusion in words: do your results suggest that the current distribution of M&M colors is the same as one of the published distributions, or do they suggest that the distribution has changed?

My conclusion suggests that my distribution is the same as that of one of the published distribuitons. ****

h. If you rejected your null hypothesis, what might be some alternate explanations for your outcome (besides a change in the color distribution of M&Ms)? That is, what are other potential sources of error or bias in our experiments?

Some sources of error in this experiment could be as simple as inaccurately counting the color, and could also be that the sample size was too small.

---

i. Repeat this analysis (in the Pooled Data tab in the Google Sheet file) using the pooled data for your breakout room and the most relevant distribution. Do you reject or fail to reject the null hypothesis based on the your breakout room's combined data? Why?

p = 0.0297 < 0.05 so the null hypothesis is rejected.

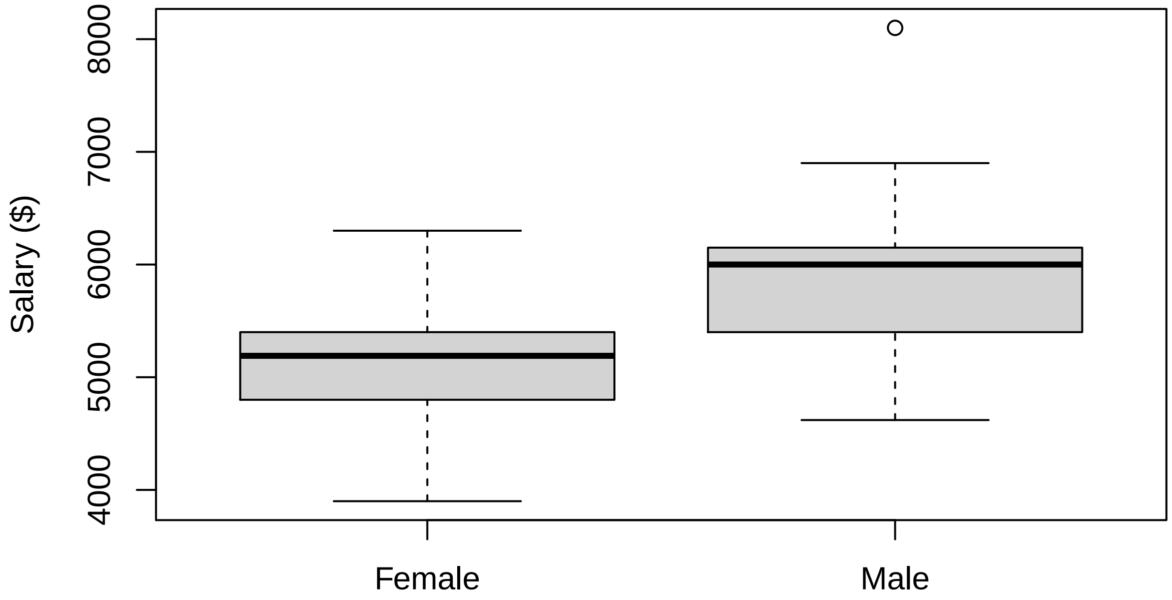
```r
pchisq(12.4,5,lower.tail = FALSE)
```

```
## [1] 0.02969946
```

Question 2: Comparison of two means In Lab 2, you examined a case study on monthly starting salaries for 32 male and 61 female skilled workers at a bank hired between 1967 and 1977 at the "Harris Trust and Savings Bank". You drew box plots of male and female salaries for and used descriptive statistics to analyze the data. This week you can use your new knowledge and apply *inferential statistics* to answer questions about differences between mean male and female salaries.

a. First, recreate the comparative box plot you created in Lab 2 that shows male and female salaries in one plot.

```r
gensal <- read.csv("Salaries.csv")
Male <- gensal$Male
Female <- gensal$Female
boxplot(gensal[c("Female","Male")], ylab = "Salary ($)")
```



b. At a significance level of 5 percent,conduct a hypothesis test to decide whether the true mean salary for males is more than 500 dollars per month higher than the true mean of female salaries. You can used the `t.test()` function to analyze the difference between two mean values as shown before lab.

Comment your code:

```
t.test(Male,Female, 500, alternative = "greater") # performs t test comparing delta of Male and Female
```

```
##
##  Welch Two Sample t-test
##
## data:  Male and Female
## t = 2.354, df = 51.353, p-value = 0.01122
## alternative hypothesis: true difference in means is greater than 500
## 95 percent confidence interval:
##  595.2532      Inf
## sample estimates:
## mean of x mean of y
##  5956.875  5126.613
```

---

c. Using the following equations for the t statistic (or t critical), degrees of freedom, and confidence interval (CI), construct and interpret a two-sided 95% confidence interval for the difference between the true (or population) mean values of male and female banking salaries.

(refer to attachment)

```
sd(Male)
```

```
## [1] NA
```

```
sd(Female)
```

```
## [1] 544.0315
```

```
length(na.omit(Male))
```

```
## [1] 32
```

```r
length(Female)
```

```
## [1] 62
```

t-statistic (or t critical value)

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

---

degrees of freedom:

$$df = \frac{[(se_1)^2 + (se_2)^2]^2}{\frac{(se_1)^4}{n_1 - 1} + \frac{(se_2)^4}{n_2 - 1}}$$

The CI for the difference between two means is covered in Section 7.5 of the textbook (pp.327-329), but here is the equation for this CI:

$$\bar{x}_1 - \bar{x}_2 \, \acute{s} \, (t\, critical\, value)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

---

d. Based on the CI, can you reject Ho? Explain your answer

Muo is outside of 95% then Ho can be rejected (refer to attachment)

---

e. If the CI of the difference between the true means of male and female banking salaries is (683.26, 977.26) dollars, what is α. Comment each line of your code.

Note: to calculate the number of rows for a data column with missing values (referred to as NAs), you can use the following line of code: `length(na.omit(x))`, where x is the data column. The male salary data has NAs.

(refer to attachment)

---

f. After doing some inferential statistics, you can revisit your answer from Assignment 2 and comment on differences between male and female salaries, and state whether you think there was discrimination among male and female salaries. In your explanation refer to your answers from parts b, c, and maybe d.

The data suggests that there is discrimination among male and female salaries. The confidence interval rejected the null hypothesis that the difference in mean salaries between men and women is $500 per month. The data including the calculations from part c and d suggests that the wage gap is greater than $500 per month.
