ST 314 Data Analysis 05

Topics:

- One and Two Sample t Procedures
- Describing Visual Displays
- Hypothesis Testing and Confidence Intervals
- Lessons Covered: 26 - 30
- Textbook Chapter (Optional): 8 and 9

Grading:

- Points are listed next to each question and should total 25 points overall.
- Grading will be based on the content of the data analysis as well as the overall appearance of the document.
- Late assignments will not be graded.

Deadlines:

- Final Submission: **Thursday, November 19th**. All submissions must be PDF files.

Instructions:

- Clearly label and **type answers** to the questions on the proceeding pages in Word, Google Docs, or other word processing software.
- Insert **diagrams or plots as a picture** in an appropriate location.
- Math Formulas need to be typed with Math Type, LaTeX, or clearly using key board symbols such as +, -. *, /, sqrt() and ^
- Submit assignment to the Gradescope link as a PDF. **Indicate the pages to the individual questions and also verify the correct document has been uploaded. Failing to follow this direction may result in point deductions.**

Allowances:

- You may use any resources listed or posted on the Canvas page for the course.
- You are encouraged to discuss the problems with other students, the instructor and TAs, however, all work must be your own words. Duplicate wording will be considered plagiarism.
- Outside resources need to be cited. Websites such as Chegg, CourseHero, Koofers, etc. are discouraged, but if used need to be cited and used within the boundaries of academic honesty.

ST 314 Data Analysis 05

Part 1. (15 points)
The microbeersW19.csv dataset is a representative sample of 1,244 microbrews from around the United States. The variable `abv` represents the percent of alcohol by volume for each craft beer. According to the National Institute of Health, one standard serving of alcohol is 12 ounces of regular beer, which is usually about 5% alcohol by volume (abv).

Does the sample of microbrews provide evidence the average alcohol by volume of all craft beers is different from a standard serving of beer at 5% abv?

Use this dataset and the R script DA5_t_procedures.R to complete the following:

a (1 point) What is the parameter of interest in this scenario? Provide the symbol and context.

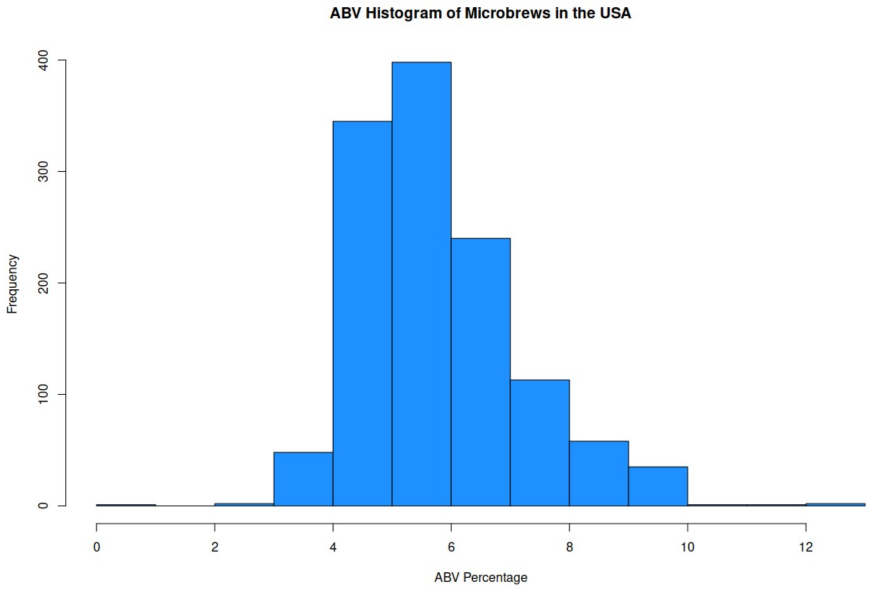
The parameter of interest in this scenario is the average ABV value of all the microbrew beers listed. This would be represented with the character μ. In R, the ABV would be represented by microbeers$abv, the set of all ABV values for the beers in question.

b (1 point) State the null and alternative hypothesis to answer the question of interest.

The null hypothesis and alternative hypothesis are as follows:

$$H_0 : \mu_0 = 5.0\, Percent\ ABV$$
$$H_a : \mu_0 \neq 5.0\, Percent\ ABV$$

c (2 point) Make a histogram or boxplot to visualize the variable abv. Is there visual evidence the average alcohol by volume is different than 5%?



There is slight visual evidence that the ABV might not be exactly 5% – the peak in the graph is a little higher than 5, as well as there appearing to be more samples above 5 than there are below it.

d (1 point) Calculate the sample mean and standard deviation using R. State the values.

$$Mean(\bar{x})=5.904$$
$$Standard\ Deviation(\sigma_{\bar{x}})=1.374$$

e (1 point) Check the conditions for inference. State them and whether they are met.

- Is the data random?
  - Yes, this sample is random, as it is a representative sample of the populations of microbrew.
- Is the sampling distribution normal based on conditions for the CLT?
  - Yes, as the population is sufficiently large.
- Do we know σ?
  - No, we only know $\sigma_{\bar{x}}$ - the standard deviation of the sample.

f (1 point) Calculate the test statistic by hand. Show work.

$$t\ stat=\frac{\bar{x}-\mu_0}{s/\sqrt{n}}=\frac{5.904-5}{1.374/\sqrt{1244}}=23.206$$

g (1 point) State the p-value. Is it one or two sided?

The p-value will be two sided. It has value $p\,value = 2.2\text{e-}16$

h (2 points) Calculate the 95% Confidence Interval by hand. Show work.

$$\bar{x} \pm t_{1243,0.975} * (\frac{s_x}{\sqrt{n}}) = 5.904 \pm 1.962 * (\frac{1.374}{\sqrt{1244}}) = (5.828, 5.980)$$

i (1 point) Use the `t.test()` command in R to verify the results of the t test. How do your answers compare?

```
t.test(microbeers$abv, mu=5,alternative = "two.sided")


      One Sample t-test

data:  microbeers$abv
t = 23.21, df = 1243, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 5.827753 5.980607
sample estimates:
mean of x
 5.90418
```

The results of the T test are very close to what I calculated.

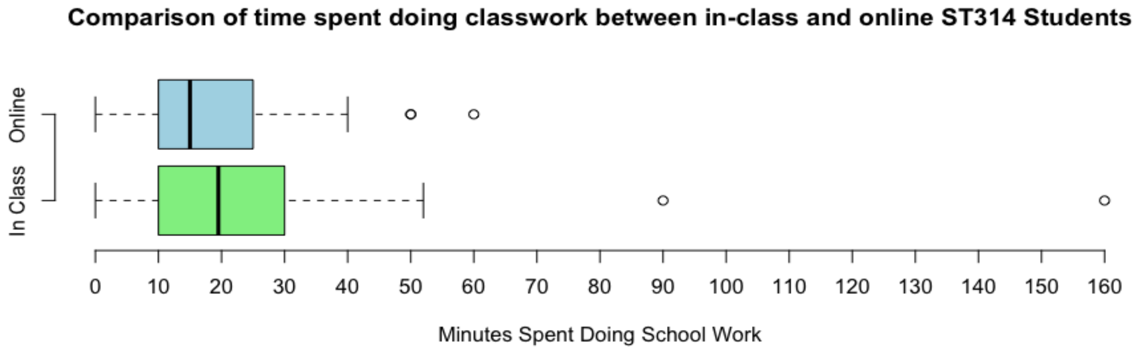
j (4 points) From the R output, write a four-part conclusion describing the results. Use $\alpha = 0.05$. Provide a statement in terms of the alternative hypothesis. State whether (or not) to reject the null. Give in context an interpretation of the point and interval estimate. Include any other information you might feel to relevant.

There is extremely strong evidence the average ABV is different than 5.0 %. We can reject the null hypothesis based on a significance level of 0.05 (t = 23.21, p < 2.2e-16). The 95% CI estimates the population average ABV to be between 5.828% and 5.981% with a best guess of 5.904. Therefore, we can assert that the average ABV of this dataset does not have the same ABV as a standard beer.

ST 314 Data Analysis 05

Part 2. (10 points)

The goal of this analysis is to compare the average time spent doing schoolwork during a week for ST314 students who are either in-class students (attend lectures in person) or online students (completing the course online). This data is from the combined ST314 Winter 2019 student information survey. The following software output is an analysis of this data:



	Mean	**Std. Dev.**	**N**
In Class	20.31	17.22	150
Online	18.31	12.67	67

```
	Welch Two Sample t-test

data:  st314data$SchoolWorkHours by st314data$Course
t = 0.95644, df = 168.89, p-value = 0.3402
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.127949  6.127750
sample estimates:
mean in group In class   mean in group Online 
              20.31333               18.31343 
```

Do these data provide evidence of a difference between the average time spent doing schoolwork in a week among in-class and online students? Use a significance level of 0.05 and answer the following questions using the software output**.**

a (2 points) Describe the side-by-side boxplot. Is there visual evidence time spent doing schoolwork is different among in-class and online students? Explain your answer in 1-2 sentences.

There is indeed visual evidence that indicates that online ST314 participants spend less time doing classwork than ST314 in person / in class students. These appear to be different by about 2-5 minutes per student at most, so not a significant difference.

b (2 point) State the null and alternative hypothesis to answer the question of interest. Is the alternative one or two sided?

ST 314 Data Analysis 05

Hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

c (2 points) Check conditions for the test. State each condition and whether it is met. If not met, state why. Then continue, even if the conditions aren't met.

- Is the data random?
  - No, the data is not random, as this covers all ST314 students – the entire population.
- Is the sampling distribution normal based on conditions for the CLT?
  - No, it's a complete sample of the population.
- Do we know σ?
  - No.

d (4 points) From the R output, write a four-part conclusion describing the results. Provide a statement in terms of the alternative hypothesis. State whether (or not) to reject the null. Give in context an interpretation of the point and interval estimate. Make sure to provide a *direction* to your interval, for example, one group had a smaller (or larger) mean than the other, include this relationship in your point and interval estimate. Include any other information you might feel to relevant.

There is weak evidence the time spent on class assignments by online students is different than that spent by in class students. We can not reject the null hypothesis based on a significance level of 0.05 (t = 0.956, p = 0.340. The 95% CI estimates that In Class students spend between -2.128 and 6.128 minutes longer than online students with a best guess of 2.000 minutes more spent by In Class students. Therefore, we can say that it would be a good guess to represent Online students as spending about 2 minutes less on assignments than in class students.