Discussion 6

1. The length of time a patient stays in a hospital is a variable of great interest for insurance and resource allocation purposes. In a given hospital, a simple random sample of lengths of stay in the intensive care unit was taken. The data are (in hours):

```
library(tidyverse)
hospital_stay <- tibble(hours = c(10, 20, 40, 60, 120, 150, 200, 300, 400))
```

   a. Create a normal Q-Q plot of the data. Is it reasonable to assume the distribution of length of stay is normal? Explain your answer.

   b. Construct a 95% confidence interval for the mean length of stay if we are willing to assume that the distribution is normal.

   c. Your collaborator is not happy with assuming that the data are normal. To avoid that assumption, you decide to find use a bootstrap approach to find a 95% confidence interval. We will go through the motions step by step for the first bootstrap sample, then repeat 5000 times in a more automated way.

      i. Find the average, standard deviation, and sample size of the sample. Create objects called `xbar_orig`, `std_dev`, and `sample_size`:

```
set.seed(154812)
xbar_orig <- mean(hospital_stay$hours)
sample_size <- nrow(hospital_stay)
```

      ii. Create a bootstrap sample of same size as the data by sampling with replacement from the data. Use the code below, but fill in the blanks. Take a look at the resulting sample. Comment on what you see.

```
bootstrap_sample <- sample_n(hospital_stay, size = ..., replace = ...)
```

      iii. Calculate $T_{\text{boot}} = \frac{\bar{x}_{\text{boot}} - \bar{x}_{\text{orig}}}{s/\sqrt{n}}$.

```
T_boot <- (mean(bootstrap_sample$hours) - xbar_orig)/
  (sd(bootstrap_sample$hours)/sqrt(sample_size))
```

      iv. We now have one value for T. We need a whole lot more, so that we can get a histogram that estimates the distribution. The code below will help you repeat this process 5000 times. Take a look at the object after you run the code to see what it actually looks like. (I.e., run `bootstrap_samples` in the console.)

```
bootstrap_samples <- tibble(i = 1:5000) %>%
  mutate(bootstrap_sample = map(i, ~sample_n(hospital_stay, size = 9, replace = TRUE)$hours),
         bootstrap_mean = map_dbl(bootstrap_sample, mean),
         bootstrap_sd = map_dbl(bootstrap_sample, sd),
         bootstrap_T = (bootstrap_mean - xbar_orig)/(bootstrap_sd/sqrt(9)))
```

v. Now that we have 5000 values of T, we want to take a look at the distribution of it. Create a histogram of the `bootstrap_T` values. You can use the code below, but don't forget to fill in the blanks!

```
ggplot(data = bootstrap_samples,
       aes(... = ...)) +
  geom_...(...)
```

vi. We now have a good idea of what the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ looks like, i.e. very similar to the histogram above. We now want to find our *critical values*, i.e. values such that we have $\alpha/2$ to the left of one of them, and $\alpha/2$ to the right of the other. I.e. we want to find the $\alpha/2$ and $1-\alpha/2$ quantiles of the 5000 T values. (Again, fill in the blanks below.) Compare the values you get to the histogram created above.

```
bootstrap_samples %>%
  summarize(t_crit1 = quantile(..., ...),
            t_crit2 = quantile(..., ...))
```

vii. Finally, we can construct our confidence interval: a 95% CI for the true mean μ is $[\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} - t_{1-\alpha/2}\frac{s}{\sqrt{n}}]$.

d. Write one sentence to interpret your CIs from b and c.

e. Compare the two CIs in (b) and (c). Which one do you think makes more sense?

2. Specifications for a water pipe call for a mean breaking strength μ of more than 2000 lbs per linear foot. To verify a particular batch of pipe, engineers will randomly select n sections of pipe from the batch that are 1ft long, measure their breaking strengths, and perform a hypothesis test. The batch of pipe will not be used unless the engineers can conclude that the mean breaking strength for the whole batch is greater than 2000.

    a. Specify appropriate null and alternative hypotheses for this situation.

    b. What kind of evidence from the sample do you need to reject the null hypothesis?

    c. Explain in non-statistical language what a Type I error would be in this context.

    d. Explain in non-statistical language what a Type II error would be in this context.

    e. Which type of Error, Type I or Type II, is worse in this situation? Justify your choice.