Homework 6

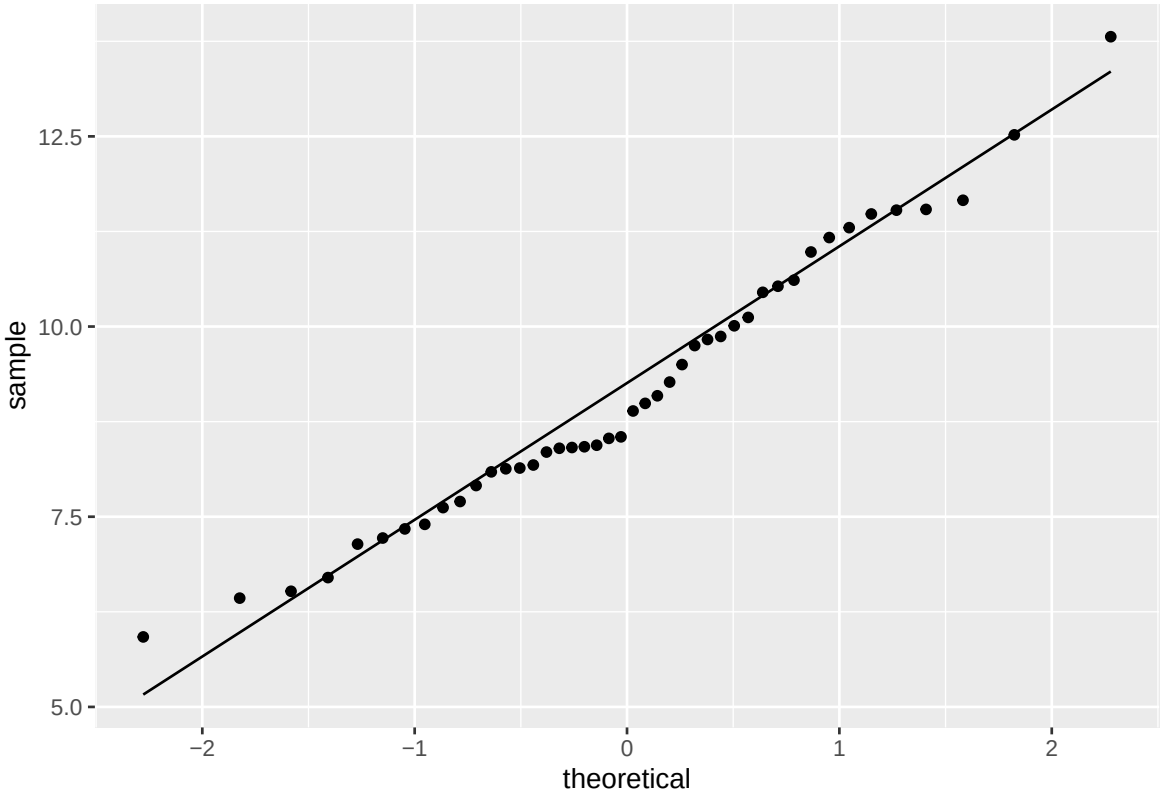
Instructions: To receive credit, you must submit your assignment to Canvas before **6pm, Friday, March 13**. The file submission must be a knitted .html file, made using RMarkdown. The code you used to answer the questions should be included in your file. You do not need to submit your .rmd file.

1. A simple random sample was taken of 44 water bottles from a bottling plant's warehouse. The dissolved oxygen content (in mg/L) was measured for each bottle, with the results below.

```
library(tidyverse)
doc_data <- tibble(oxygen_content = c(11.53, 8.35, 11.66, 11.54, 9.83,
                                      5.92, 7.14, 8.41, 8.99, 13.81,
                                      10.53, 7.4, 6.7, 8.42, 8.4, 8.18,
                                      9.5, 7.22, 9.87, 6.52, 8.55,
                                      9.75, 9.27, 10.61, 8.89, 10.01,
                                      11.17, 7.62, 6.43, 9.09, 8.53,
                                      7.91, 8.13, 7.7, 10.45, 11.3,
                                      10.98, 8.14, 11.48, 8.44, 12.52,
                                      10.12, 8.09, 7.34))
```

   a. Create a histogram and QQ plot of the data. Is it plausible that the data was drawn from a population that follows a normal distribution? Explain your answer.
   Solution: The plots are below. The histogram is fairly bell shaped, and the QQ plot shows a reasonably straight line, so normality seems reasonable.

```
ggplot(doc_data,
       aes(sample = oxygen_content)) +
  geom_qq() +
  geom_qq_line()
```



b. Find a 98% confidence interval for the population mean dissolved oxygen content. Show your work.
Solution: Since the data is plausibly normal, but we don't know sigma, we can use a T-based interval. $\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$. For the given 98% confidence level, we have $(1-\alpha)100\% = 98\% implies \alpha = 0.02$, and we need $t_{n-1,\alpha/2} = t_{43,0.01}$. Using the T-table, the correct critical value is between 2.4 and 2.5. We'll use 2.45. (The exact value is about 2.42). R computes the sample mean and sample standard deviation as $\bar{x} = 9.146$ and $s = 1.776$. Our interval is thus $9.146 \pm 2.45\frac{1.776}{\sqrt{44}} = 9.146 \pm 0.656$ or $(8.490, 9.802)$. Alternatively, you could argue that the sample size, $n = 44$, is large enough that a T is a lot like a Z, and instead use the z multiplier, $z_{0.01} = 2.33$. The interval would then be $(8.522, 9.770)$, which is pretty similar to the T-based interval.

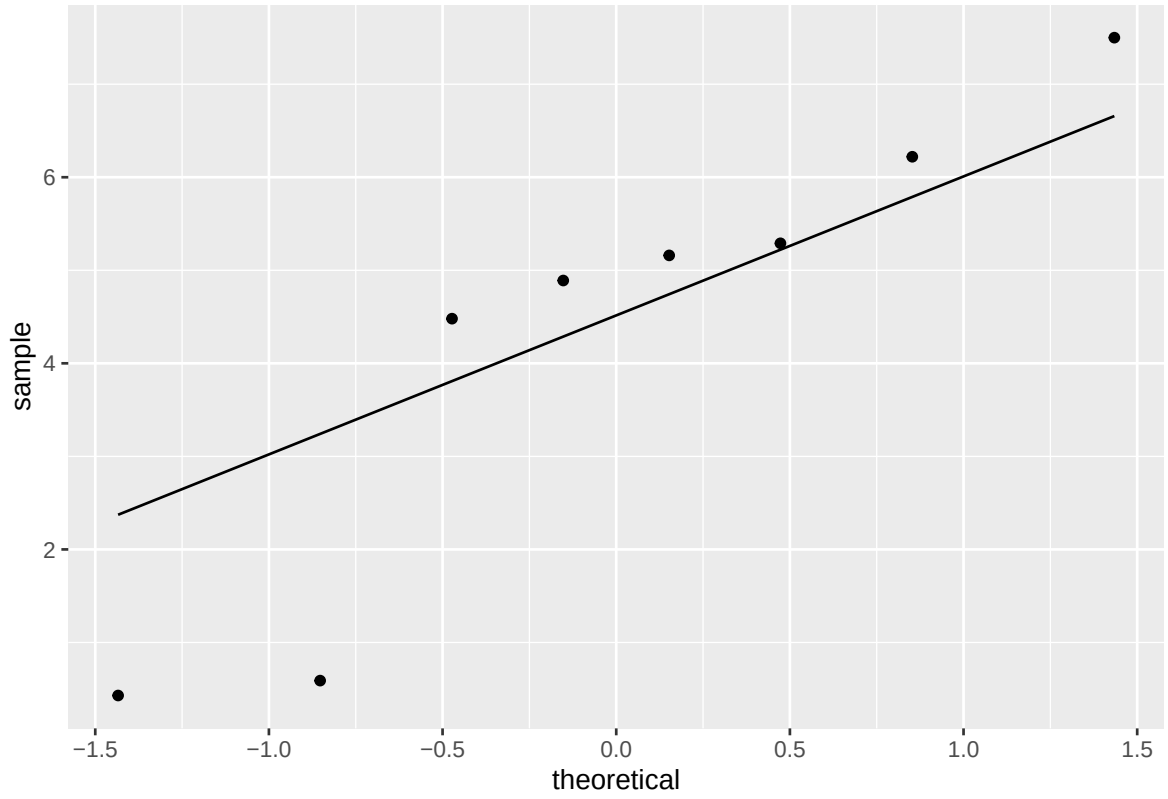
c. Interpret the interval you created in (b).
Solution: You could say that you have 98% confidence that your interval contains the true population mean, μ. A better statement would be that you created this interval using a procedure that has a 98% chance of covering the population mean μ. Or you could say that if you were to collect many simple random samples of size 44 from this population and compute an interval corresponding to each sample, you would expect that about 98% of those intervals would contain the true μ. However, you cannot be sure that the one interval that you actually calculated does or does not contain μ.

2. Consider the following data that represents a random sample out of a very large population:

```
data_sample <- tibble(x = c(0.43, 4.89, 5.29, 5.16, 0.59, 7.50, 4.48, 6.22))
```

a. Create a QQ plot of the data. Do you think it is reasonable to assume that the population distribution is normal? Explain your answer.
Solution: The QQ plot is below. The plot is not very straight, so normality does not look like a good assumption.

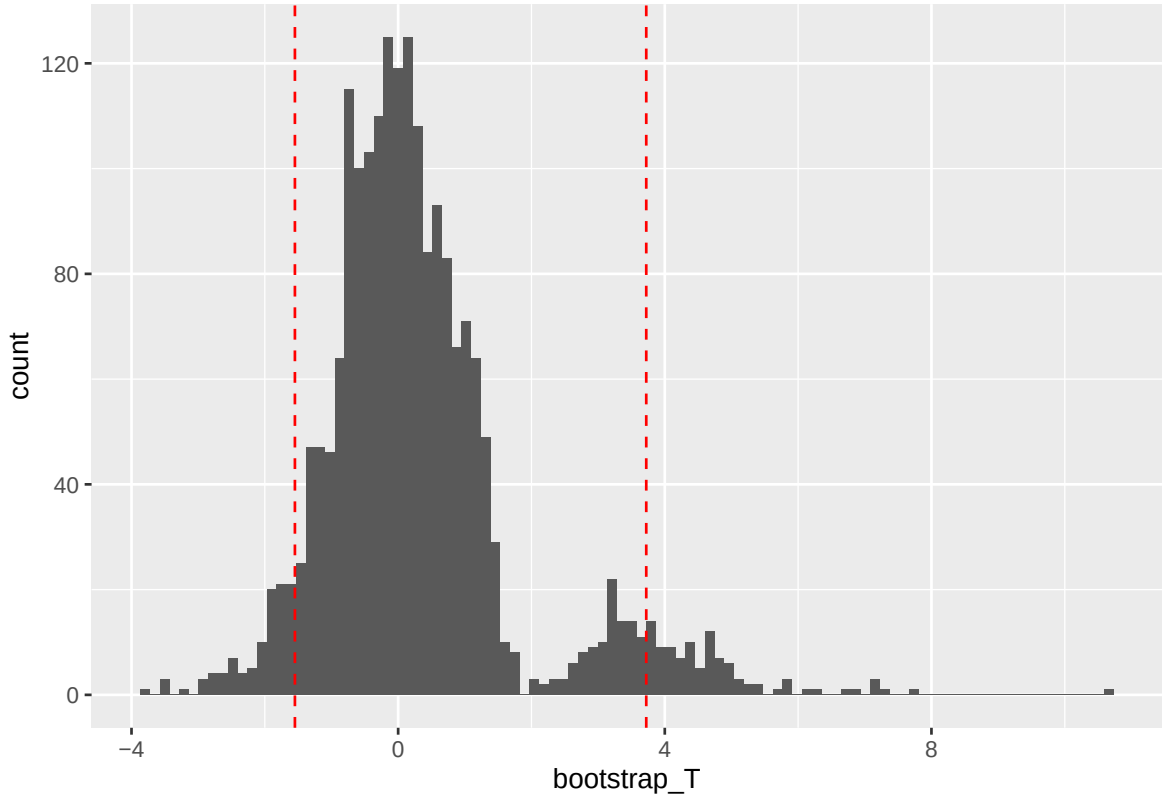
```r
ggplot(data_sample,
       aes(sample = x)) +
  geom_qq() +
  geom_qq_line()
```



b. Denote μ to be the population mean. Regardless of your answer to (a), use R to perform the bootstrap with 2000 resamplings to create a 90% CI for μ. Show the estimated distribution using a histogram. Since answers will differ for this question, it is critical that you show your R code and output to get full credit. Use the code from discussion March 4th/5th.

```r
sample_size <- nrow(data_sample)

xbar <- mean(data_sample$x)
SD <- sd(data_sample$x)

bootstrap_samples <- tibble(i = 1:2000) %>%
  mutate(bootstrap_sample = map(i, ~sample_n(data_sample, size = sample_size, replace = TRUE)$x),
         bootstrap_mean = map_dbl(bootstrap_sample, mean),
         bootstrap_sd = map_dbl(bootstrap_sample, sd),
         bootstrap_T = (bootstrap_mean - xbar)/(bootstrap_sd/sqrt(sample_size)))

t_left <- quantile(bootstrap_samples$bootstrap_T, 0.05)
t_right <- quantile(bootstrap_samples$bootstrap_T, 0.95)

LL <- xbar - t_right*SD/sqrt(sample_size)
UL <- xbar - t_left*SD/sqrt(sample_size)
```

```
ggplot(bootstrap_samples,
       aes(x = bootstrap_T)) +
  geom_histogram(bins = 100) +
  geom_vline(xintercept = c(t_left, t_right),
             linetype = "dashed", color = "red")
```



Solution: The critical values from the estimated distribution are $\hat{t}_{0.05} = 3.721928$ and $\hat{t}_{0.95} = -1.5486757$ – the red dotted line on the histogram above. The 90% confidence interval is then found as $[\bar{x} - \hat{t}_{0.05}, \bar{x} - \hat{t}_{0.95}]$, which is [0.992, 5.705].

c. Regardless of your answer to (a), assume the population distribution is normal and use that fact to create a 90% CI for μ.
Solution: Since we are assuming the data is normal for this problem, and σ is unknown, we should use a t multiplier. From the table, $t_{(7,0.05)} = 1.9$, so the interval is $4.32 \pm 1.9\frac{2.53}{\sqrt{8}}$ or $(2.62, 6.02)$.

d. Compare your answers to parts (b) and (c). Which one do you think is more correct, and why?
Solution: Since the QQ plot indicates that the population is not likely normal, the bootstrap interval is better. You can also see this because the critical values from the bootstrap are not ± 1.9 as they were for the T-based interval.

3. Suppose you are in charge of inventory maintenance at a bicycle shop. **One of your jobs is to ensure that the tire pressure in each of the display bicycles is between 65-85 PSI (pounds per square inch)**. If the pressure is too low, then there is a risk of wheel damage when a customer rides out on one. On the other end, if the pressure is too high, there is a (small) risk of the tire exploding! Of course, you don't know the true pressure of any particular tire. Instead, you have the output from your pressure gauge. This will be similar to the true pressure, but not necessarily the same. Laboratory testing of the particular gauge you use has shown that there is a ± 2 PSI error margin, and so to be careful, you decide that you will adjust the pressure on any tire that has a measured pressure above 81

PSI or below 69 PSI, giving you 2x the error margin on either side. Previous testing shows that this procedure will give the following results:

	Measure inside 69-81PSI	Measure outside 69-81 PSI
PSI within 65-85 PSI	5000	105
PSI outside 65-85 PSI	3	216

a. State (in words) the appropriate null and alternate hypothesis.
Solution: Since your job is to ensure that the tire pressure in each of the display bicycles is between 65-85 PSI, H_0: the true pressure is in the range 65-85 PSI. Hence, H_1: the true pressure is ***not*** in the range 65-85 PSI.

b. Calculate the value of α. What situation does the Type I error rate represent here?
Solution: α= P(Type I error)= P(rejecting H_0| H_0 is true)=P(Rejection Region $|H_0$ is true)
$= \frac{\text{No. of cases of rejecting null when null is actually true}}{\text{No. of cases when null is actually true}} = \frac{105}{105+5000} = 0.0206$. This is the situation where the true pressure is not in the desired range 65-85 PSI, but our gauge suggests that it is.

c. Calculate the value of β. What situation does the Type II error rate represent here?
Solution: β= P(Type II error)= P(Fail to reject H_0| H_0 is false)=1−P(Rejection Region $|H_0$ is false)
$= \frac{\text{No. of cases of failing to reject null when null is actually false}}{\text{No. of cases when null is actually false}} = \frac{3}{3+216} = 0.0137$. This is the situation where the true pressure is inside the desired range 65-85 PSI, but our gauge suggests that it is not.

d. Calculate the *Power* of this test. What situation does test *Power* represent here?
Solution: $Power = 1 - \beta = 0.986$. This is the situation where the true pressure is inside the desired range 65-85 PSI, and our gauge measurement is within 69-81 PSI, i.e. it matches with reality.

e. What is the rejection region for the test you are conducting?
Solution: The rejection region is $[0, 69)$ and $(81, \infty]$ PSI, since any measurement within either of these ranges will cause us to reject the null hypothesis that the true pressure is within the range of 65-85 PSI.

f. Practically speaking, which error seems more problematic: a Type I error or a Type II error?
Solution: In this case, a Type II error is worse. If we commit a Type I error, then we adjust the tire pressure even though we didn't have to. When we commit a Type II error, then the tire pressure remains in a dangerous range (either too low or too high).

g. If we wanted to decrease the Type I error rate, how would we change the rejection region?
Solution: Looking into the definition of α in part (a), it is clear that we should **decrease the rejection region to decrease the Type I error rate**, or equivalently, increase the acceptance region.

h. If we wanted to decrease the Type II error rate, should we increase or decrease the rejection region?
Solution: Looking into the definition of β in part (b), it is clear that we should **increase the rejection region to decrease the Type II error rate**, or similarly, decrease the acceptance region.