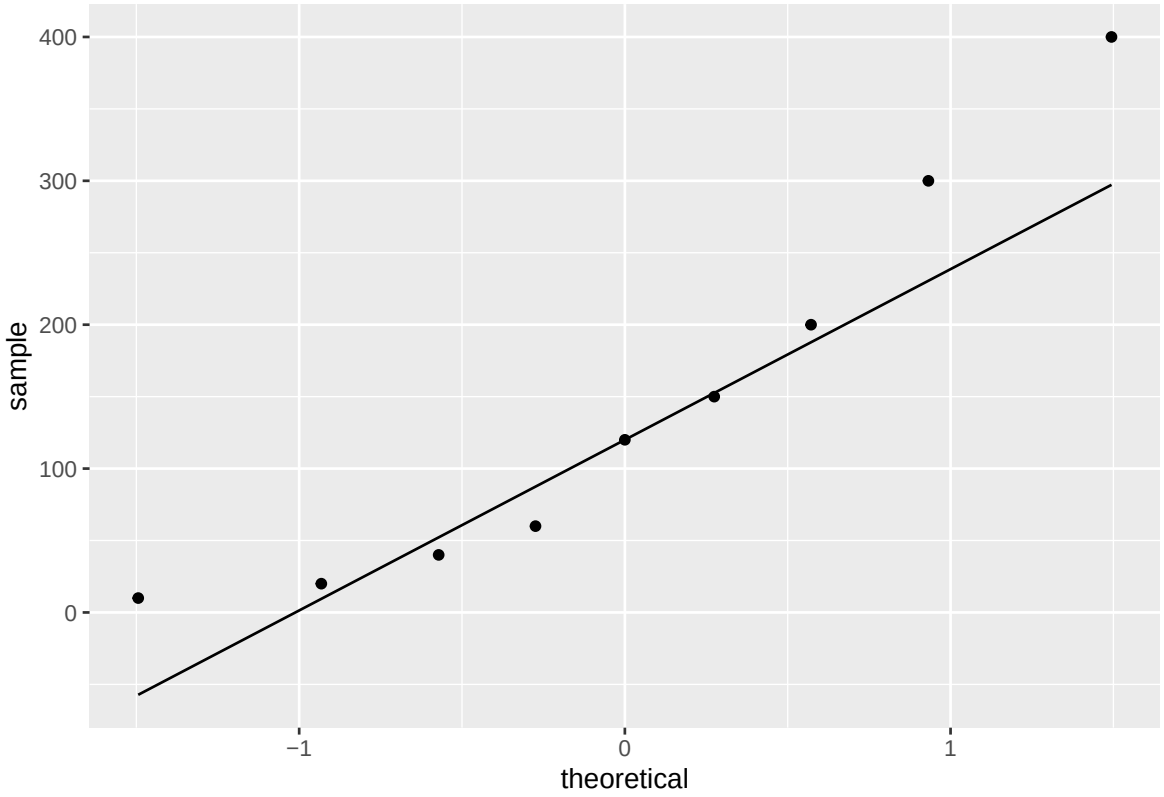
Discussion 6

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.
   Solution: The plot is below. This one is a little bit borderline. There is some curve to the plot, but it isn't too bad. With the sample size being pretty small, we might want to err on the side of caution and conclude that normality is questionable.

```
ggplot(data = hospital_stay,
       aes(sample = hours)) +
  geom_qq() +
  geom_qq_line()
```



b. Construct a 95% confidence interval for the mean length of stay if we are willing to assume that the distribution is normal.
Solution: We calculate the sample mean and sample standard deviation to be 144.44 and 134.55, respectively. We will use a t multiplier since we are assuming the data is normal, but we do not know σ. We then determine the critical value $t_{n-1,\alpha/2} = t_{8,0.025} = 2.31$. The interval is thus $144.44 \pm 2.31(\frac{134.55}{\sqrt{9}}) = 144.44 \pm 103.60 = (40.84, 248.04)$. Some R code to help with this is below.

```
library(distributions3)
```

```
##
## Attaching package: 'distributions3'

## The following objects are masked from 'package:stats':
##
##     Gamma, quantile

## The following object is masked from 'package:grDevices':
##
##     pdf
```

```
T_8 <- StudentsT(df = 8)
quantile(T_8, 0.975)
```

```
## [1] 2.306004
```

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`:
Solution: Note: the first line below simply ensures that we get the same results every time we run. It is not strictly necessary, but if you want to double check your results against the ones shown below, and avoid discrepencies due to randomness in the resampling, include it.

```
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.

Solution: The sample has several repeated observations. This is not surprising, since we sample with replacement, but it is important, because otherwise we would just get the same sample over and over again.

```
bootstrap_sample <- sample_n(hospital_stay, size = 9, replace = TRUE) # size = n
bootstrap_sample
```

```
## # A tibble: 9 x 1
##   hours
##   <dbl>
## 1   300
## 2    60
## 3    40
## 4   400
## 5    20
## 6    40
## 7   150
## 8   400
## 9    10
```

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

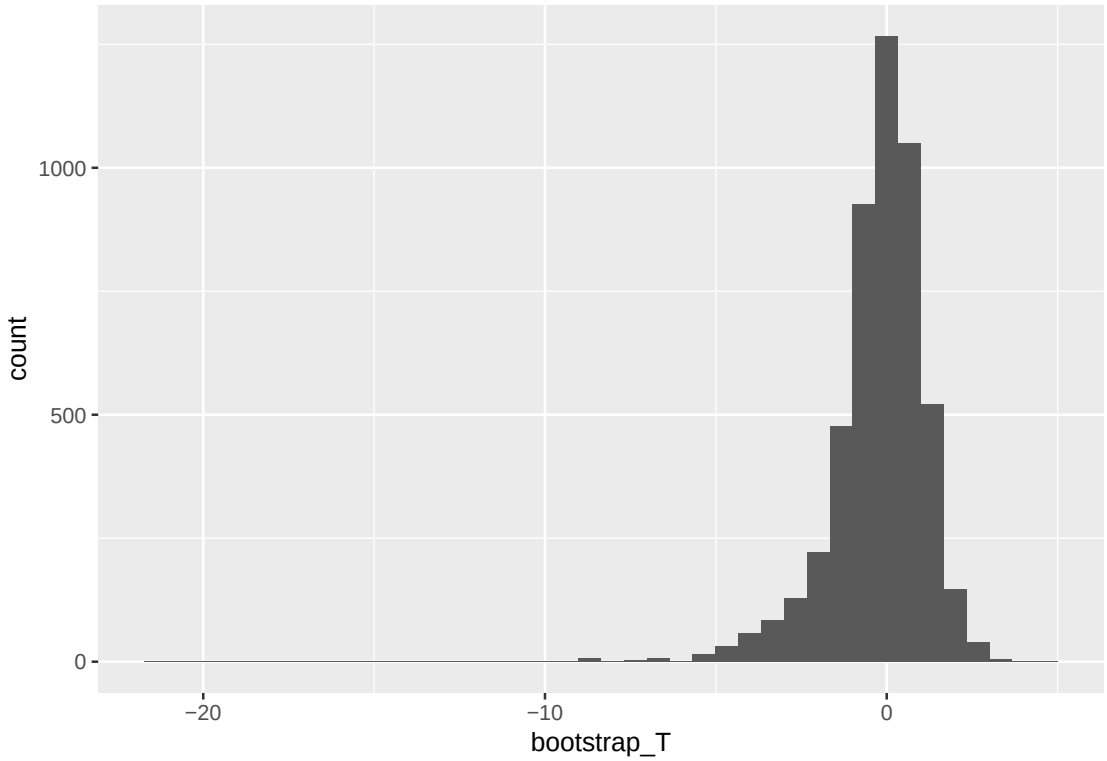
```
T_boot <- (mean(bootstrap_sample$hours) - xbar_orig)/
  (sd(bootstrap_sample$hours)/sqrt(sample_size))
```

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(x = bootstrap_T)) +
  geom_histogram(bins = 40)
```

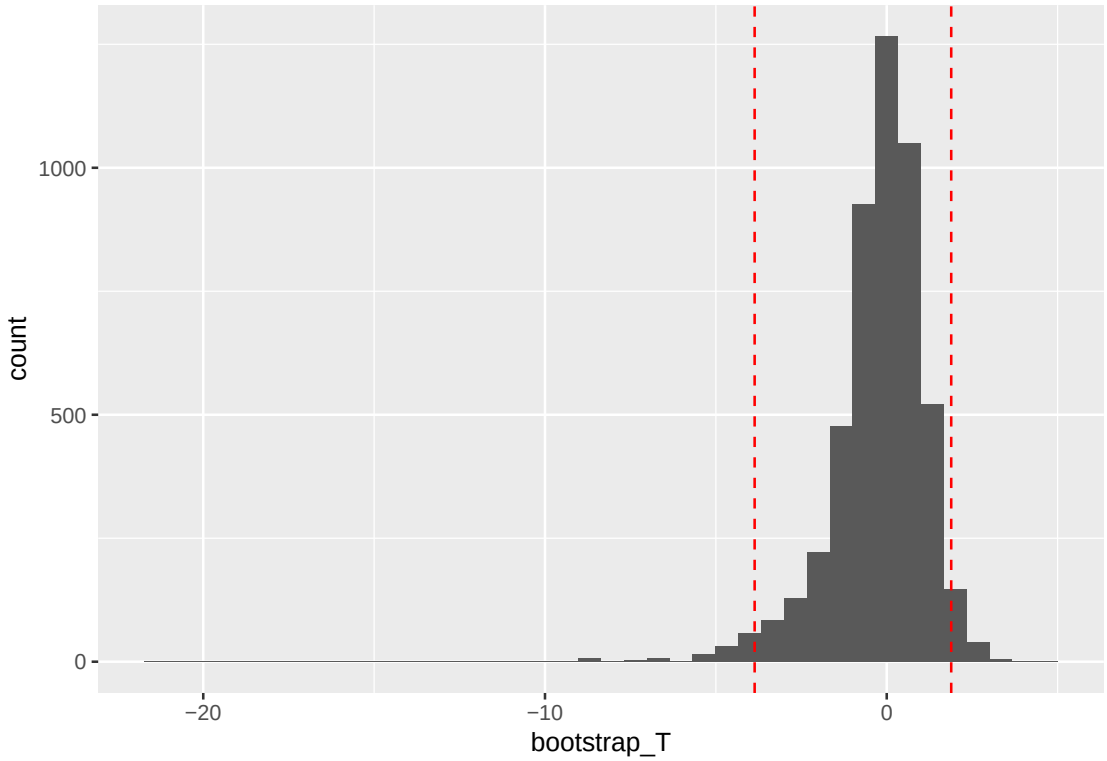


vi. We now have a good idea of what the distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ looks like, i.e. very similar to the histogram above. We now want to find our *critical values*, i.e. values such that we have $\alpha/2$ to the left of one of them, and $\alpha/2$ to the right of the other. I.e. 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. **Solution**: The critical values are overlayed the histogram below. It seems pretty reasonable that 2.5% of the area is to the left and right of the cut-offs, respectively.

```
bootstrap_samples %>%
  summarize(t_crit1 = quantile(bootstrap_T, 0.025),
            t_crit2 = quantile(bootstrap_T, 0.975))
```

```
## # A tibble: 1 x 2
##   t_crit1 t_crit2
##     <dbl>   <dbl>
## 1   -3.86    1.89
```

```
ggplot(data = bootstrap_samples,
       aes(x = bootstrap_T)) +
    geom_histogram(bins = 40) +
    geom_vline(xintercept = quantile(bootstrap_samples$bootstrap_T,
                                     c(0.025, 0.975)),
               color = "red", linetype = "dashed")
```



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

```
s <- sd(hospital_stay$hours)
bootstrap_samples %>%
  summarize(t_crit1 = quantile(bootstrap_T, 0.025),
            t_crit2 = quantile(bootstrap_T, 0.975),
            LL = xbar_orig - t_crit2*s/sqrt(9),
            UL = xbar_orig - t_crit1*s/sqrt(9))
```

```
## # A tibble: 1 x 4
##   t_crit1 t_crit2    LL    UL
##     <dbl>   <dbl> <dbl> <dbl>
## 1   -3.86    1.89  59.8  318.
```

d. Write one sentence to interpret your CIs from b and c.
Solution: If we were to repeatedly sample from the population and use the same procedure to construct a CI, about 95% of our intervals would contain the true value of the population mean.

e. Compare the two CIs in (b) and (c). Which one do you think makes more sense?
Solution: We probably prefer the bootstrap-based CI, since we are concerned about the validity of the normality assumption. You can see that the intervals are not very similar. The bootstrap interval is quite a bit larger.

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.
Solution: A reasonable null would be that $\mu = 2000$. The alternative would be that $\mu > 2000$. Notice that in this case it is bad practice to run the test by setting the alternative to be $\mu < 2000$ and declaring the batch of pipe safe if we do not reject the null. The problem is that the Neyman-Pearson paradigm assumes that the null is true without any evidence. Not rejecting the null does not imply that the null is true, but merely that we did not have sufficient evidence to reject it. For example, this could happen in cases where the test does not have sufficient power. It is better to use the alternative $\mu > 2000$, since we only reject this (and declare the pipe safe) if we have sufficient evidence to do so. We are in essence assuming that the pipe is not safe and only changing that premise if the data tells us to.

b. What kind of evidence from the sample do you need to reject the null hypothesis?
Solution: The larger the observed sample mean is than 2000, the more evidence there is against the null.

c. Explain in non-statistical language what a Type I error would be in this context.
Solution: A Type I error occurs when the data tells you to reject, but in reality, the null is true. In this case, this would mean that we observed a large sample mean breaking strength, but the true mean breaking strength is 2000 (or less).

d. Explain in non-statistical language what a Type II error would be in this context.
Solution: A Type II error occurs when the data tells you to not reject, but in reality, the null is false. In this case, this would mean that we observed a small sample mean breaking strength, but the true mean breaking strength is greater than 2000.

e. Which type of Error, Type I or Type II, is worse in this situation? Justify your choice.
Solution: Both types of errors are bad, but it's probably worse to claim that the pipe is safe when it really isn't. You risk a flood and potential open ended costs! A Type II error would be to fail to certify as safe a batch that is actually safe. Tossing the batch of course costs money, but presumably this cost is fixed. So a Type I error is probably worse here.