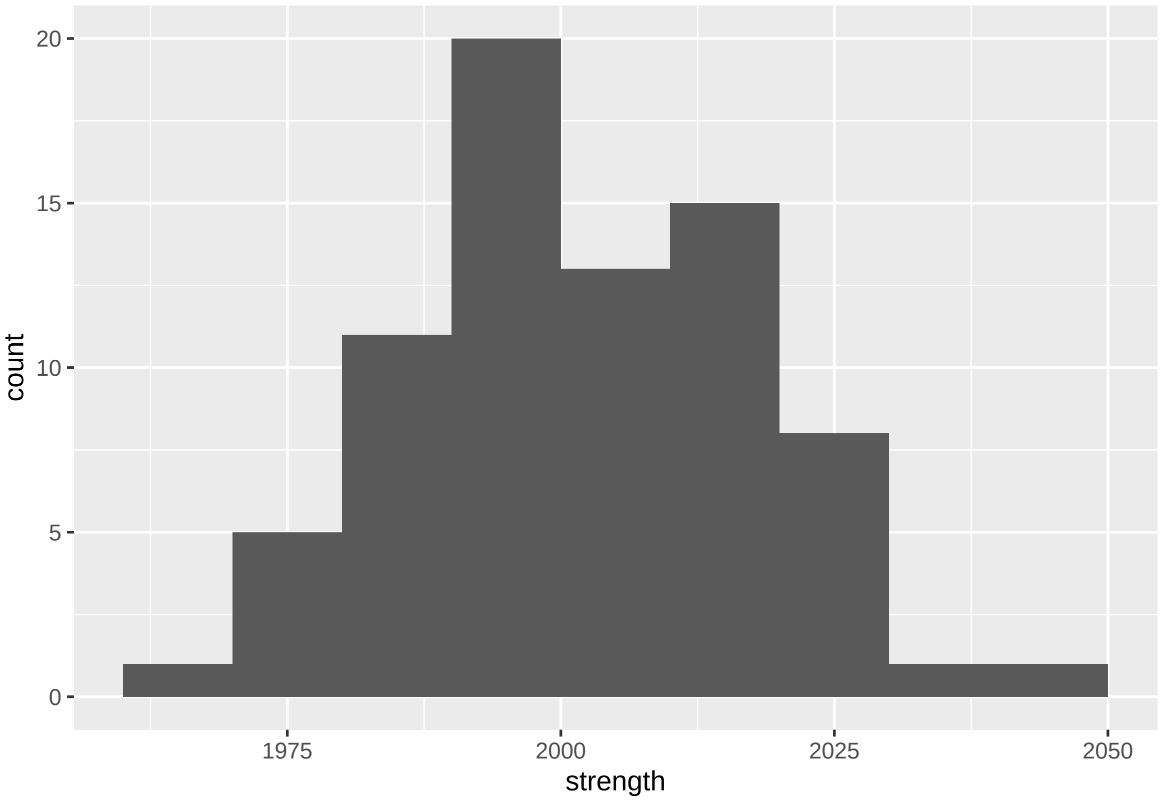
Discussion 7

1. Specifications for a water pipe call for a mean breaking strength μ of more than 2000 lb per linear foot. Engineers will perform a test to decide whether or not to use a certain kind of pipe. A random sample of 1 ft sections of pipe is selected and their breaking strengths are measured. The pipe will not be used unless the engineers can conclude (statistically, not with certainty) that the mean breaking strength is greater than 2000.

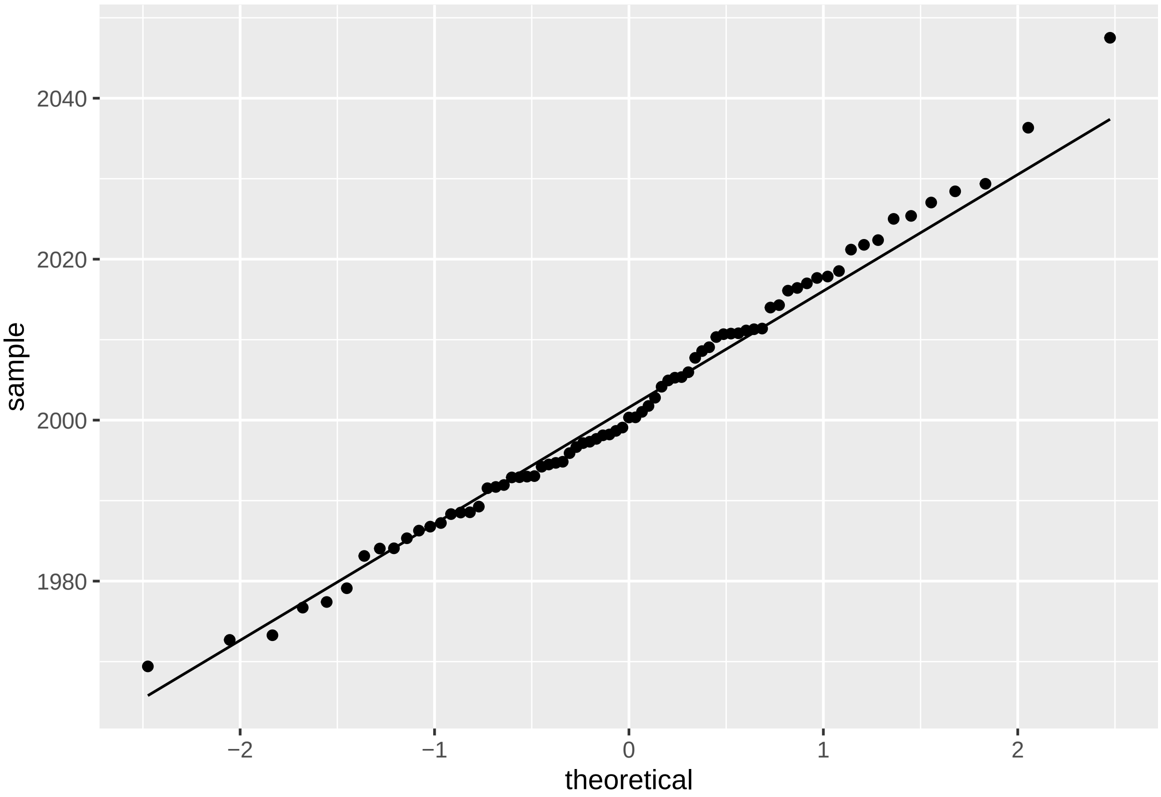
   a. Specify appropriate null and alternative hypotheses for this situation.
   Solution: $H_0 : \mu = 2000,\ H_a : \mu > 2000$

   b. Based on last week's analsysis, the engineers chose to obtain a sample of 75 random 1 foot pipe sections. The data is provided in a .csv file. Perform a one sample t-test at the 5% level after checking that the assumptions for testing are well met and interpret the results in context.

```r
library(tidyverse)
pipes <- read_csv(here::here("Discussions/disc07/pipes.csv"))

ggplot(pipes, aes(x = strength)) +
  geom_histogram(binwidth = 10, boundary = 1970)
```



```
ggplot(pipes, aes(sample = strength)) +
  geom_qq() +
  geom_qq_line()
```



```
sum_stats <- pipes %>%
  summarize(n = n(),
            Mean = mean(strength),
            SD = sd(strength),
            T_obs = (Mean - 2000)/(SD/sqrt(n)))
sum_stats
```

```
## # A tibble: 1 x 4
##       n  Mean    SD T_obs
##   <int> <dbl> <dbl> <dbl>
## 1    75 2002.  15.7  1.13
```

Solution: We are assuming these 75 samples are randomly and independently chosen (ie, we aren't only testing from 1 long pipe 75 times). From the qqplot and histogram of the sample data, we do not have strong evidence that the sample came from a non-normal population. Addiitionally, with such a large sample size, we are confident the CLT has kicked in and the distribution of sample means would be normal, even if the population data is not. T-test T-statistic: $t_{obs} = \frac{2002.053799-2000}{15.7074943/\sqrt{75}} = 1.1324$. Then, we can compare that observed test statistic to the critical values in the T_{74} distribution, or calculate the p-value. Let's do both. Since $1.1324 = T_{\text{obs}} < t_{74,0.05} = 1.6657$, our observed value is NOT far from the null (i.e. 0), hence we do NOT reject. Since $P(T_{74} > T_{\text{obs}}) = P(T_{74} > 1.1324) = 0.8694 > 0.05$, we do not reject – the probability of observing something more extreme is low.

```r
library(distributions3)
T_74 <- StudentsT(df = 74)

## Critical Value
quantile(T_74, 1-0.05)
```

```
## [1] 1.665707
```

```r
## P-value
1-cdf(T_74, sum_stats$T_obs)
```

```
## [1] 0.130571
```

c. Another scientist in the lab suggests instead of a t test, a z test could be performed. Explain why either a t or z test will give very similar conclusions in this case.
Solution: Since our sample size is so large (75), we are confident that the sample standard deviation is a good approximation for the population standard deviation. Additionally, with 74 degrees of freedom, the t-distribution will be very similar to the standard Normal, so if we calculate the critical value and p-value based on the normal rather than the t, we get very similar results:

```r
Z <- Normal()

## Critical Value
quantile(Z, 1 - 0.05)
```

```
## [1] 1.644854
```

```r
## P-value
1-cdf(Z, sum_stats$T_obs)
```

```
## [1] 0.1287431
```

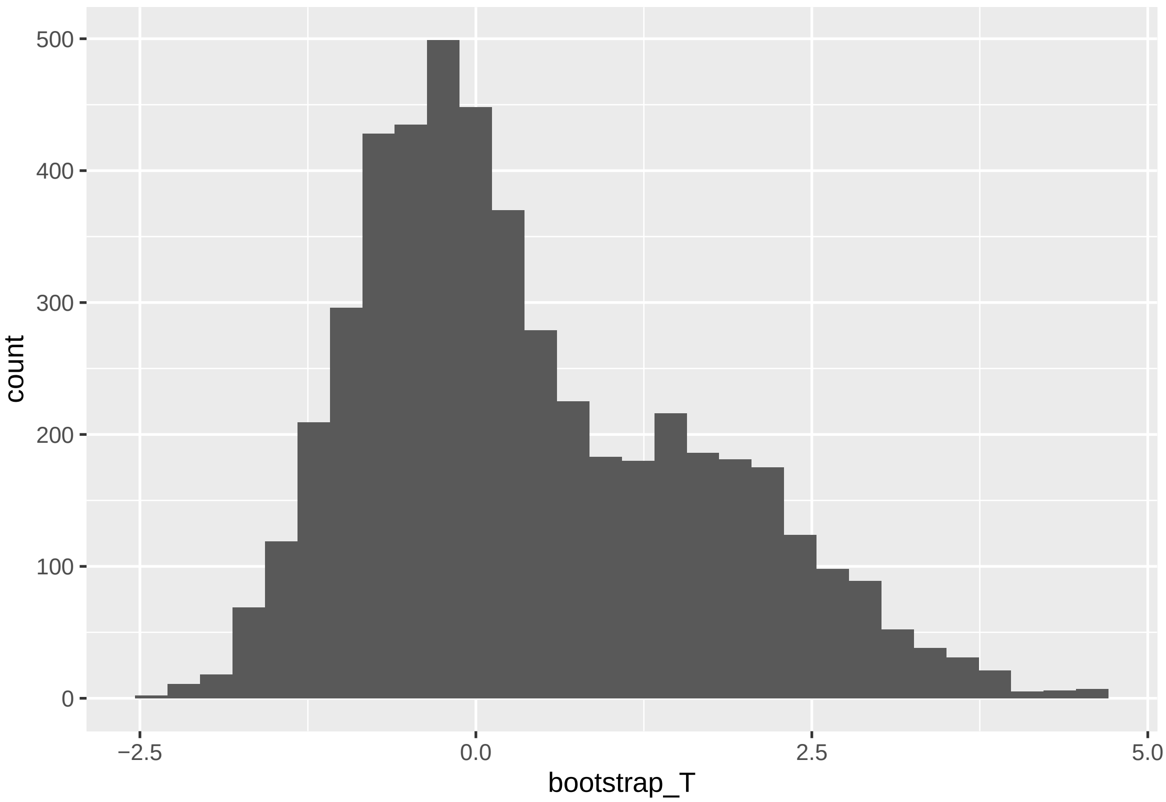
2. A crop scientist evaluating lettuce yields plants 20 plots, treats them with a new fertilizer, lets the lettuce grow, and then measures yield in numbers of heads per plot. The results are provided in a .csv file.
The old fertilizer led to an average yield of 145 heads per plot. Test whether the new fertilizer leads to an improved yield via the following steps.

a. Perform a bootstrap test at a 0.1 significance level. State all assumptions needed.

```r
lettuce <- read_csv(here::here("Discussions/disc07/lettuce.csv"))

bootstrap_samples <- tibble(i = 1:5000) %>%
  mutate(bootstrap_sample = map(i, ~sample_n(lettuce,
                                           size = nrow(lettuce), replace = T)$yields),
         bootstrap_mean = map_dbl(bootstrap_sample, mean),
         bootstrap_sd = map_dbl(bootstrap_sample, sd),
         bootstrap_T = (bootstrap_mean - mean(lettuce$yields))/(bootstrap_sd/sqrt(nrow(lettuce))))

(critical_values <- bootstrap_samples %>%
  summarize(t_left = quantile(bootstrap_T, 0.05),
            t_right = quantile(bootstrap_T, 0.95)))
```

```
## # A tibble: 1 x 2
##   t_left t_right
##    <dbl>   <dbl>
## 1  -1.28    2.78
```

```r
T_obs <- (mean(lettuce$yields) - 145)/(sd(lettuce$yields)/sqrt(nrow(lettuce)))

p_value <- sum(bootstrap_samples$bootstrap_T > T_obs)/5000

ggplot(bootstrap_samples, aes(x = bootstrap_T)) +
  geom_histogram(bins = 30)
```



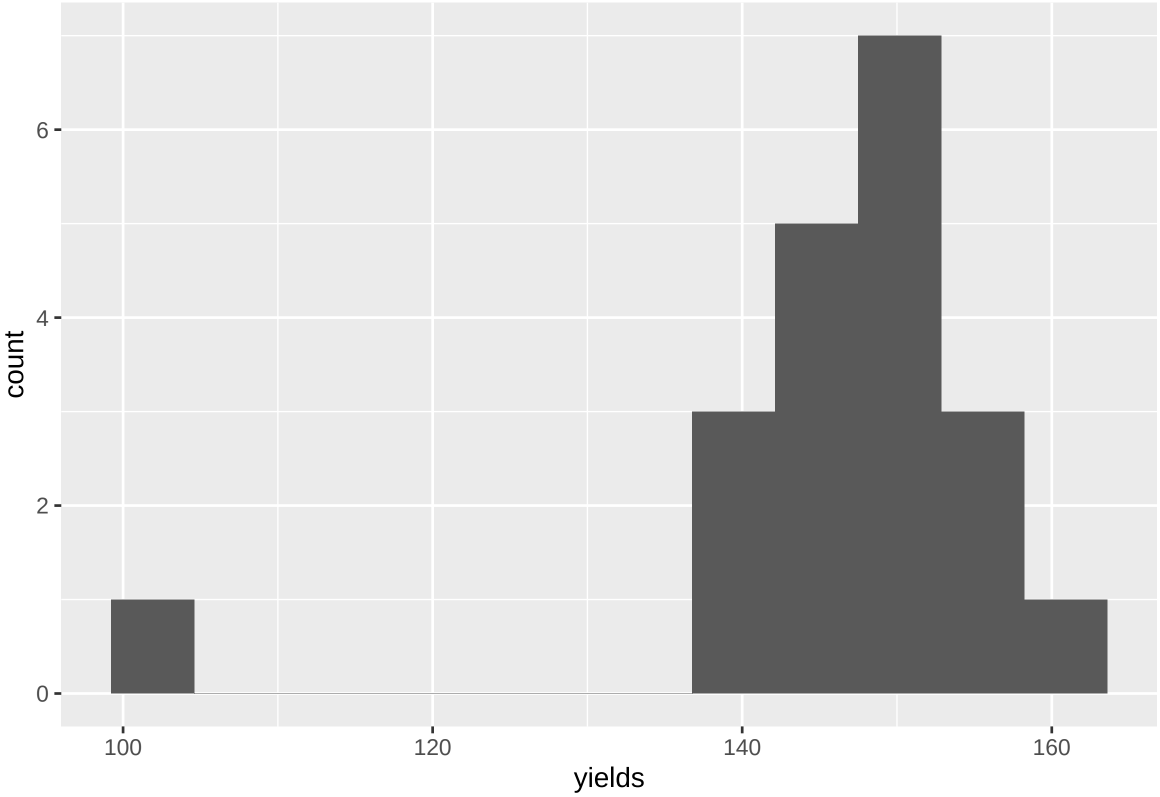
Solution: We are testing $H_0 : \mu = 145$ against $H_A : \mu > 145$ using a bootstrap test. For this to work, we need to assume that the samples are independent of each other. We need to choose a significance level. Let's use $\alpha = 0.1$. Our observed test statistics is $T_{\text{obs}} = 0.567$. The p-value is found using the bootstrap samples as the proportion of bootstrap T's that are "more extreme" than what we observe. Since $H_A : \mu > 145$, "more extreme" = "greater". So, the p-value is $\frac{\text{\# bootstrap T's} > T_{\text{obs}}}{\text{\# bootstrap samples}} = \frac{1854}{5000} = 0.371$. Since this is greater than our chosen significance level ($\alpha = 0.1$), we do NOT reject the null in favor of the alternative.

b. Find a 90% bootstrap confidence interval.
   Solution: A 90% CI for mu is given as $[\bar{x}_{\text{obs}} - \hat{t}_{0.05}\frac{s}{\sqrt{n}}, \bar{x}_{\text{obs}} - \hat{t}_{0.95}\frac{s}{\sqrt{n}}] = [146.5 - 2.7757306 \cdot \frac{11.8343834}{\sqrt{20}}] = [139.155, 149.875]$.

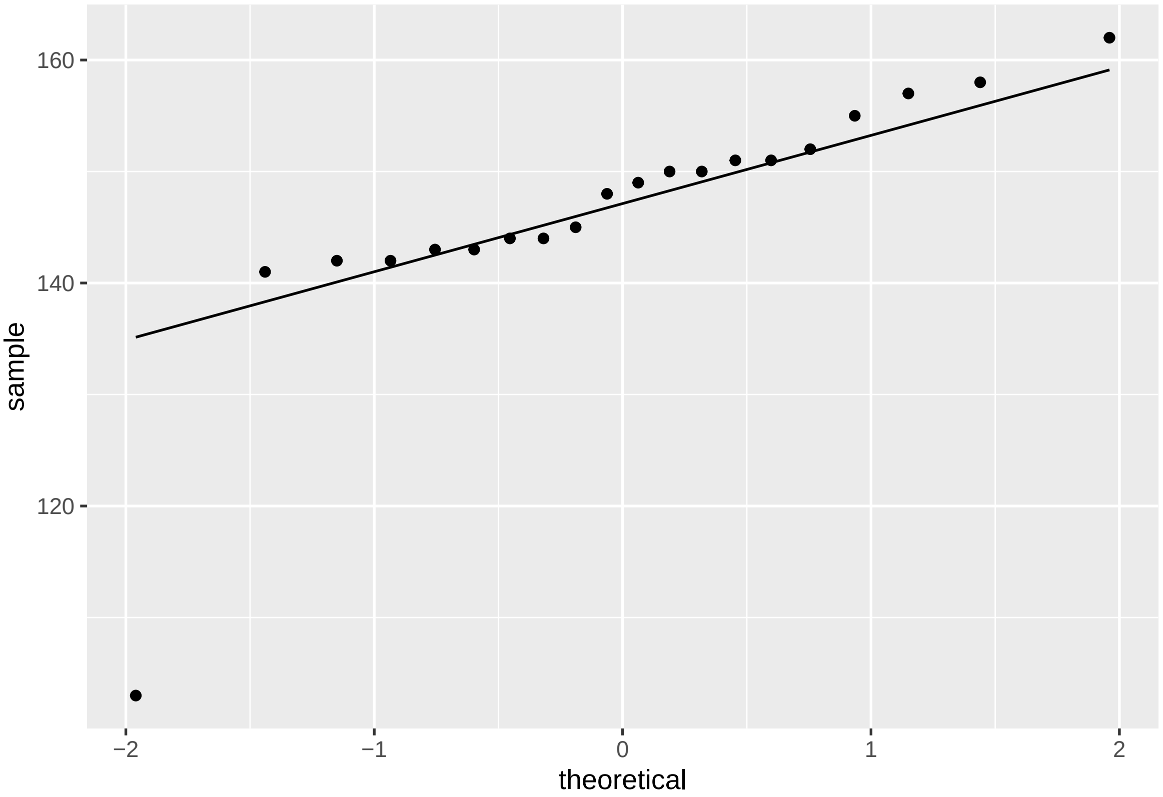
c. Perform a t-test and find a corresponding confidence interval using $\alpha = 0.1$. State all assumptions needed, and whether or not you find them reasonable. Compare to results from b and c.
   Solution: Calculations below. For us to do this, we need to add the assumption that $\bar{X} \sim N$, i.e. either the data are normal, or CLT gives us normality. From the histogram and QQ-plot

below, the data don't seem to be normally distributed, so we would have to rely on CLT. With 20 observations, this assumption seems hard to defend. Comparison: p-value smaller, confidence interval shifted a bit.

```
ggplot(lettuce,
       aes(x = yields)) +
  geom_histogram(bins = 12)
```



```
ggplot(lettuce,
       aes(sample = yields)) +
  geom_qq() +
  geom_qq_line()
```



```
lettuce %>%
  summarize(Mean = mean(yields),
            SD = sd(yields),
            n = n(),
            T_obs = (Mean - 145)/(SD/sqrt(n)),
            crit_val = quantile(StudentsT(df = n-1), 1-0.1),
            p_val = 1 - cdf(StudentsT(df = n-1), T_obs),
            LL = Mean - quantile(StudentsT(df = n-1), 1-0.05)*SD/sqrt(n),
            UL = Mean + quantile(StudentsT(df = n-1), 1-0.05)*SD/sqrt(n))
```

```
## # A tibble: 1 x 8
##    Mean    SD     n T_obs crit_val p_val    LL    UL
##   <dbl> <dbl> <int> <dbl>    <dbl> <dbl> <dbl> <dbl>
## 1  146.  11.8    20 0.567     1.33 0.289  142.  151.
```

```
## Perform test using build-in t.test function:
t.test(x = lettuce$yields, conf.level = 0.9, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  lettuce$yields
## t = 55.361, df = 19, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 0
## 90 percent confidence interval:
##  142.9865      Inf
## sample estimates:
## mean of x
##     146.5
```

```
## Get confidence interval from build-in t.test function:
t.test(x = lettuce$yields, conf.level = 0.9)$conf.int
```

```
## [1] 141.9243 151.0757
## attr(,"conf.level")
## [1] 0.9
```