Homework 4

Instructions: To receive credit, you must submit your assignment to Canvas before **6pm, Friday, February 21st**. 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. This assignment is worth 50 points.

1. Let F be an RV that represents the operating temperature in Fahrenheit of one instance of a manufacturing process, and assume $F \sim N(100, Var(F) = 5^2)$. Let C be an RV that represents the same process, but measured in Celsius. Fahrenheit can be converted to Celsius using $C = \frac{5}{9}(F - 32)$. Using `R` when needed, solve the following:

    a. Find the probability that one randomly selected instance of the process will have operating temperature greater than 98.6 Fahrenheit.
    (3pts)Solution:

```
library(distributions3)
F <- Normal(100, 5)
1 - cdf(F, 98.6)
```

```
## [1] 0.6102612
```

    b. Find the distribution of C. (Hint: $C \sim ?(?, ?)$)
    (3pts)Solution: C is normal, since it is a linear function of a normal RV. $E(C) = 5/9(E(F) - 32) = 37.78$ and $\text{Var}(C) = 5^2/9^2\text{Var}(F) = \frac{25}{81}25 \approx 7.72$ So, $C \sim N(37.78, 7.72)$.

    c. Find the probability that one randomly selected instance of the process will have operating temperature below 32 Celsius.
    (3pts)Solution:

```
C <- Normal(37.78, sqrt(7.72))
cdf(C, 32)
```

```
## [1] 0.0187505
```

    d. Above what temperature (in Celsius) is the top 10% of operating temperatures?
    (3pts)Solution

```
quantile(C, 1 - 0.1)
```

```
## [1] 41.34078
```

    e. Find the probability in a sample of 6 instances, more than 4 instances have operating temperature above 32 Celsius.(Assuming observations in the sample are independent)
    (3pts)Solution: Let X be the number of instances that have operating temperature above 32 celsius. Then $X \sim \text{Binomial}(n = 6, \pi = P(C \geq 32) = 0.9812495)$. So $P(X > 4) = P(X = 5) + P(X = 6) = 0.9949845$

f. Find the distribution of $\bar{C}$ for n=6, then find the probability that the average operating temperature in a sample of 6 instances is between 36 and 40 Celsius.
(3pts)Solution: $\bar{C} \sim N(37.78, 7.72/6^2)$.

```
Cbar <- Normal(37.78, sqrt(7.72/6))
cdf(Cbar,40) - cdf(Cbar, 36)
```

```
## [1] 0.916537
```

2. Using R, generate 9 samples of size 10, and 9 samples of size 20, and 9 samples of size 100 from a standard normal distribution. For each sample, create a QQ-plot. Do you see any relationship between sample size, and QQ-plots? What does this show you in terms of our ability to determine normality based on a QQ-plot?

   Hint: the code below will generate 9 samples of size 5 from a standard normal distribution, and create the QQ-plots. You can copy/paste this code, and adjust where appropriate to get the plots you need.

```
library(tidyverse)
library(distributions3)

random_samples <- data.frame(i = 1:9) %>%
  mutate(datasets = map(i, ~random(d = Normal(), n = 5))) %>%
  unnest_longer(col = datasets)

ggplot(random_samples,
       aes(sample = datasets)) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~i, nrow = 3) +
  ggtitle(label = "Size 5") ## this just adds a label to the plot
```

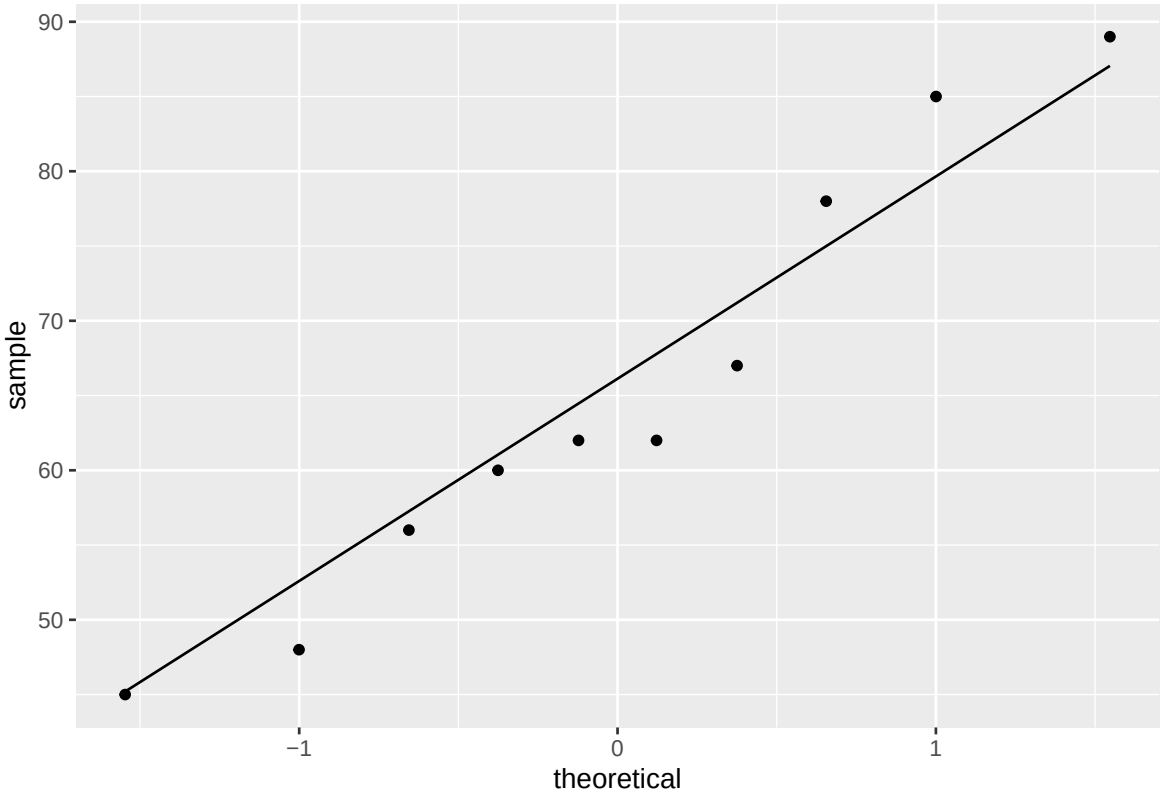
(10pts)Solution: when the sample size is 10, some of the QQ-plots do not exactly look super linear. As sample size increases, the plots look more linear. For small samples, it is hard to use QQ-plots to determine normality.

3. The data below record the number of hours a team of workers takes to assemble a custom-built motorcycle. The data are recorded for 10 different teams each assembling a motorcycle.

```
assembling_time <- data.frame(hours = c(89, 78, 48, 85, 67, 45, 60, 62, 62, 56))
```

   a. Create a QQ-plot in R and comment on the assumption that the population of times to assemble a motorcycle is well-approximated by a normal distribution.
   (3pts)Solution: The QQ-plot below looks fairly linear. The data might be from a normal distribution.

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



b. Let's say we assume the data are well-approximated by a normal distribution. What kind of distribution does $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ follow?
(3pts)Solution: since each of the X_i's follow normal distributions, the average will also follow a normal distribution.

c. Estimate the true mean and the true variance.
(3pts)Solution: We estimate the true mean and variance with the sample equivalents:

```
assembling_time %>%
  summarize(mean = mean(hours), variance = var(hours))
```

```
##   mean variance
## 1 65.2 217.9556
```

d. Pretend that the true mean is 85 minutes, and the true standard deviation is 15. If these were the true values, what would the exact distribution of $\bar{X}$ be?
(3pts)Solution: $\bar{X} \sim N(85, 15^2/10)$.

e. Pretend that the true mean is 85 minutes, and the true standard deviation is 15. What is the probability that, if we were to repeat the experiment (i.e. record the time it takes to assemble a custom-built motorcycle 10 times), the average $\bar{X}$ would be smaller than the average we got the first time?
(3pts)Solution: $P(\bar{X} < 65.2)$

```
Xbar <- Normal(85, 15/sqrt(10))
cdf(Xbar, 65.2)
```

```
## [1] 1.495132e-05
```

f. Pretend the true mean and standard deviation are our estimates. How fast must a team assemble a motorcycle to be in the top 20%?
(3pts)Solution: To be in the top 20% means having one of the 20% fastest assembling times. So, we want to find x such that $P(X < x) = 0.2$:

```
X <- Normal(65.2, sqrt(217.9556))
quantile(X, 0.2)
```

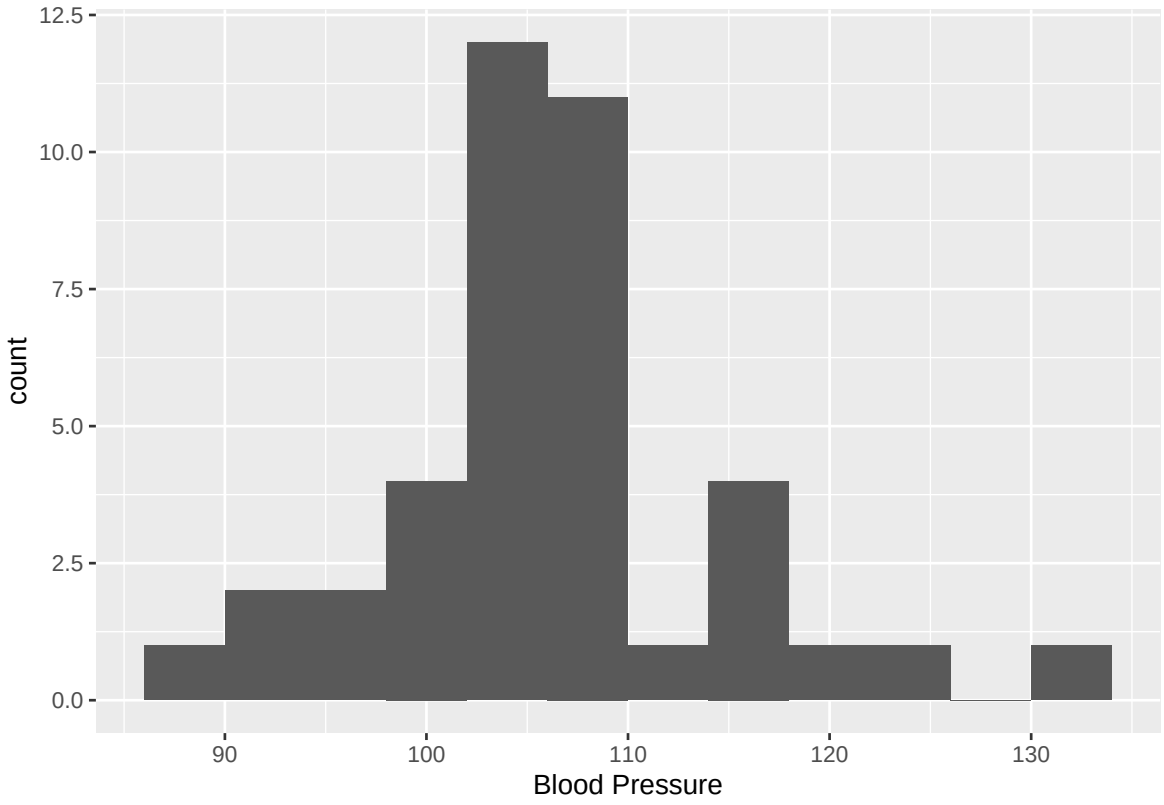
```
## [1] 52.77488
```

4. The file `blood_pressure.csv` contains two columns: `Blood Pressure` with 40 measurements of systolic blood pressure from two groups of individuals, and `Treated` that indicates if the individuals were treated with a new blood pressure medicine (1) or placebo (0). We want to assess if the blood pressure medicine is effective in lowering the blood pressure of the patients.

```
dataset <- read_csv("blood_pressure.csv")
```

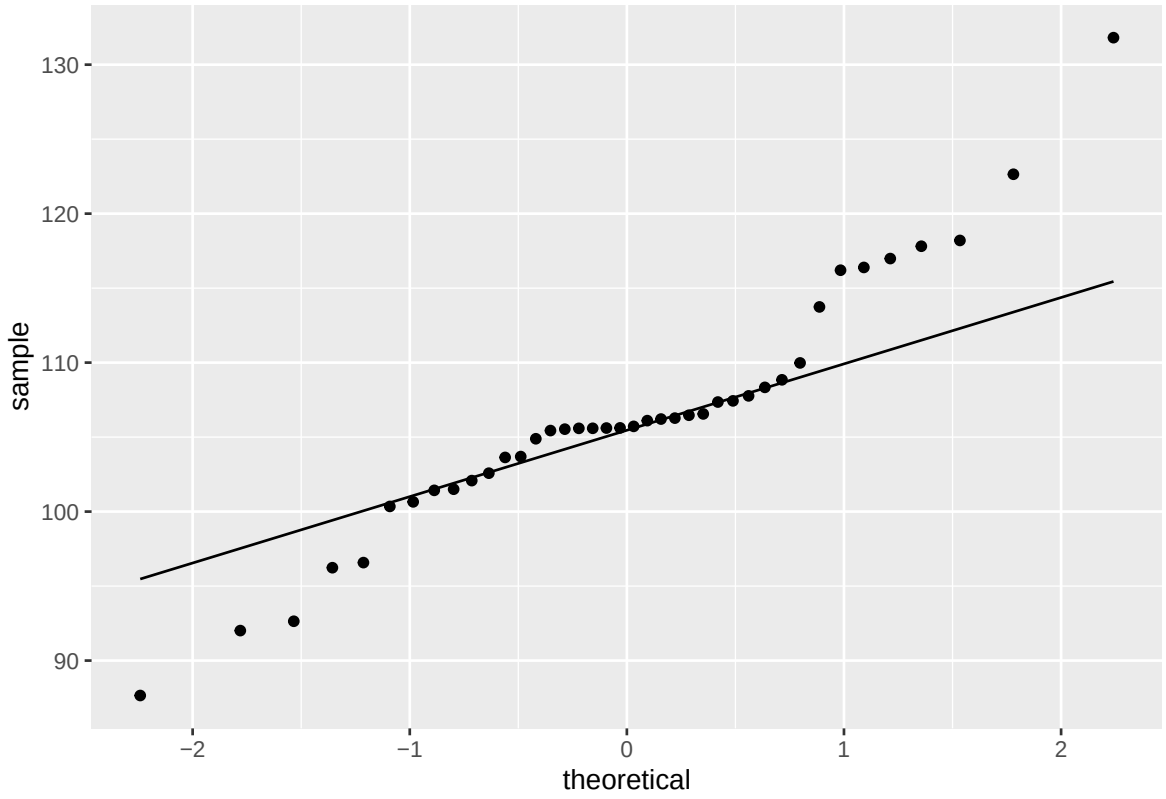
a. Create a histogram of the blood pressure measurements. Comment on the shape of the data. Do you think it is normally distributed?

```
ggplot(dataset,
       aes(x = `Blood Pressure`)) +
  geom_histogram(binwidth = 4)
```



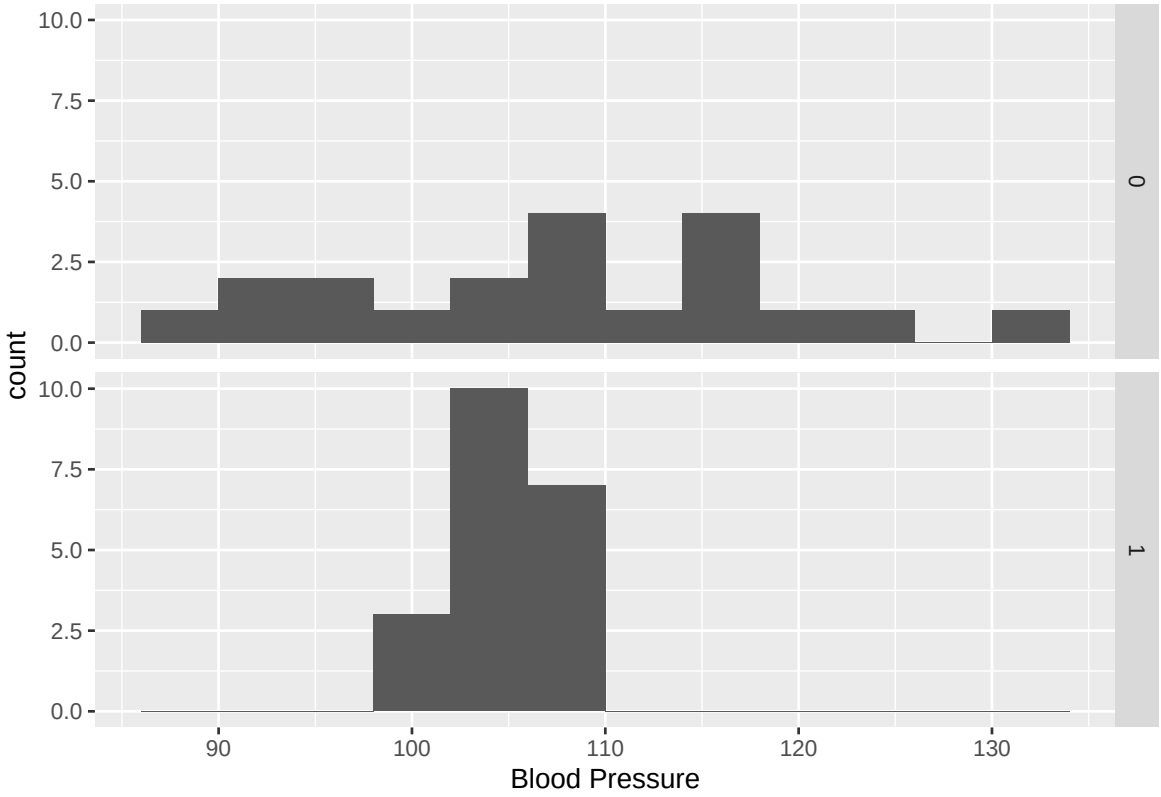
b. Create a QQ-plot. What would you conclude?

```
ggplot(dataset,
       aes(sample = `Blood Pressure`)) +
  geom_qq() +
  geom_qq_line()
```



c. We are really interested in comparing the two groups (treated and untreated). Create a histogram for each group. Comment on the shape of the data. Do you think each group is normally distributed?

```
ggplot(dataset,
       aes(x = `Blood Pressure`)) +
  geom_histogram(binwidth = 4) +
  facet_grid(Treated ~ .)
```



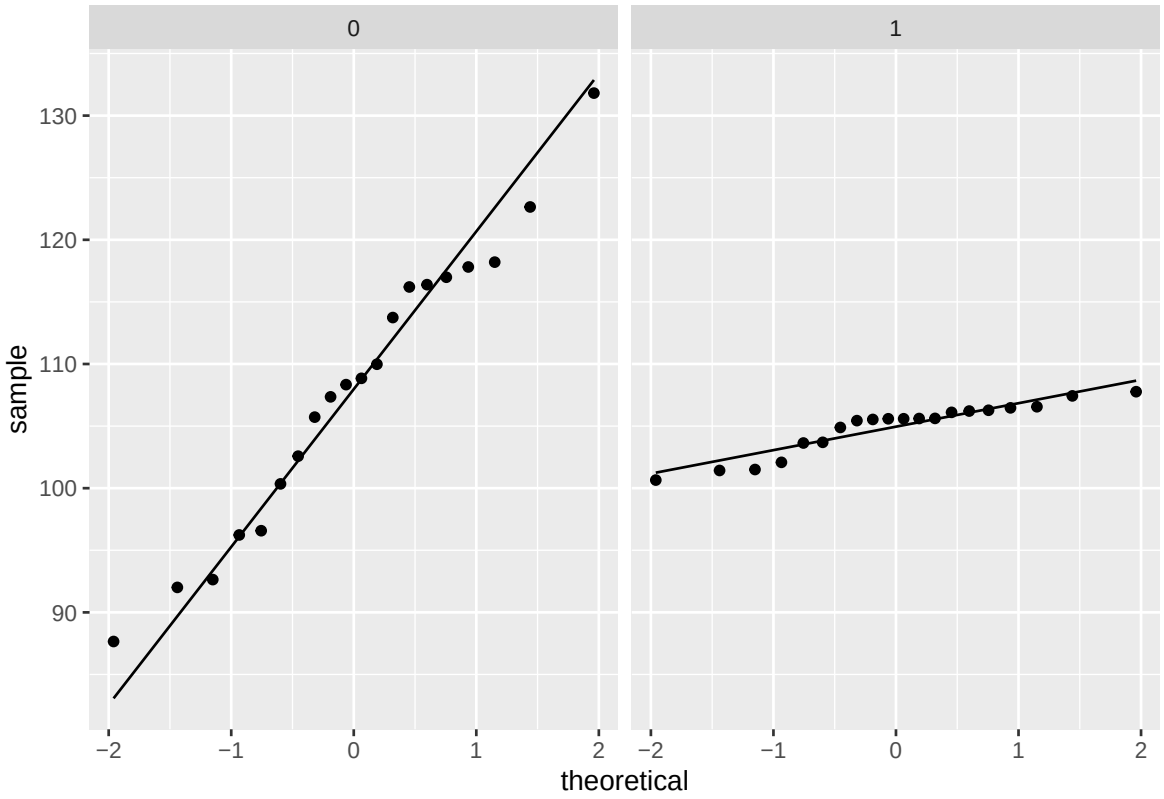
d. Compute the estimated mean and variance for each group, and overall.

```
overall <- dataset %>%
  summarize(Mean = mean(`Blood Pressure`),
            Variance = var(`Blood Pressure`)) %>%
  mutate(Treated = 'Overall')

grouped <- dataset %>%
  group_by(Treated) %>%
  summarize(Mean = mean(`Blood Pressure`),
            Variance = var(`Blood Pressure`)) %>%
  mutate(Treated = as.character(Treated))

bind_rows(
  overall,
  grouped
) %>%
  select(Treated, Mean, Variance)
```

```
## # A tibble: 3 x 3
##   Treated  Mean Variance
##   <chr>   <dbl>    <dbl>
## 1 Overall  107.    68.9
## 2 0        108.   132.
## 3 1        105.     4.22
```

e. Create a QQ-plot for each group. What would you conclude?

```
ggplot(dataset,
       aes(sample = `Blood Pressure`)) +
  facet_grid(~Treated) +
  geom_qq() +
  geom_qq_line()
```



f. Let's pretend that both samples are from normal distributions, and that the true mean for both is 106. The variances are the group variances found in d. Let $X_1, ..., X_{20}$ be random variables giving the blood pressure of the treated patients, and $Y_1, ..., Y_{20}$ the blood pressure of the untreated patients. Find the distributions of $\bar{X}$ and $\bar{Y}$, respectively.
(2pts)Solution: $\bar{X} \sim N(106, 4.221966/20)$ and $\bar{Y} \sim N(106, 131.877605/20)$

```
Xbar <- Normal(mu = 106, sigma = sqrt(4.221966/20))
Ybar <- Normal(mu = 106, sigma = sqrt(131.877605/20))
```

g. What is the probability that $\bar{X}$, if the experiment was repeated, came out to be smaller than the average we observed in this experiment?
(1pt)Solution:

```
cdf(Xbar, mean(filter(dataset, Treated == 1)$`Blood Pressure`))
```

```
## [1] 0.008524488
```

h. What is the probability that $\bar{Y}$, if the experiment was repeated, came out to be greater than the average we observed in this experiment?
(1pt)Solution:

```r
1-cdf(Ybar, mean(filter(dataset, Treated == 0)$`Blood Pressure`))
```

```
## [1] 0.2063875
```

i. The probabilities found in g and h are, in some sense, a measure of how unusual the observed data are, *if the true means in the two groups were indeed 106*. If the probability is small, the observed data are unlikely to be from a population with mean 106, and we would therefore second guess our assumption about the true mean. On the other hand, if the probability is large, the observed data are likely, and there would be no reason to second guess that assumption.
If we use 0.1 as a cut-off for what is unlikely (i.e. if the probability is below 10% the event is unlikely), would you say that the observed data are unlikely if the true means were 106? What conclusion would you draw about the likelihood of the populations from which the samples were taken having same true mean? What conclusion would this lead you to in terms of the efficacy of the treatment? I.e. do you think the treatment actually lowers the blood pressure?
Solution: Since the probability for $\bar{Y}$ is 0.1, the data are likely to be from populations with true mean 106. The probability fo $\bar{X}$ is very small (compared to 0.1), so the treated group seems to have true mean different from 106. Since the observed value of $\bar{x}$ is smaller than 106, we would think the true mean is less than 106. This also means that the two groups (treated and untreated) could very well be from different populations with different true means, so the treatment does seem to work.