HW_Week5_108020033

Che-Wei, Chang

2023-03-17

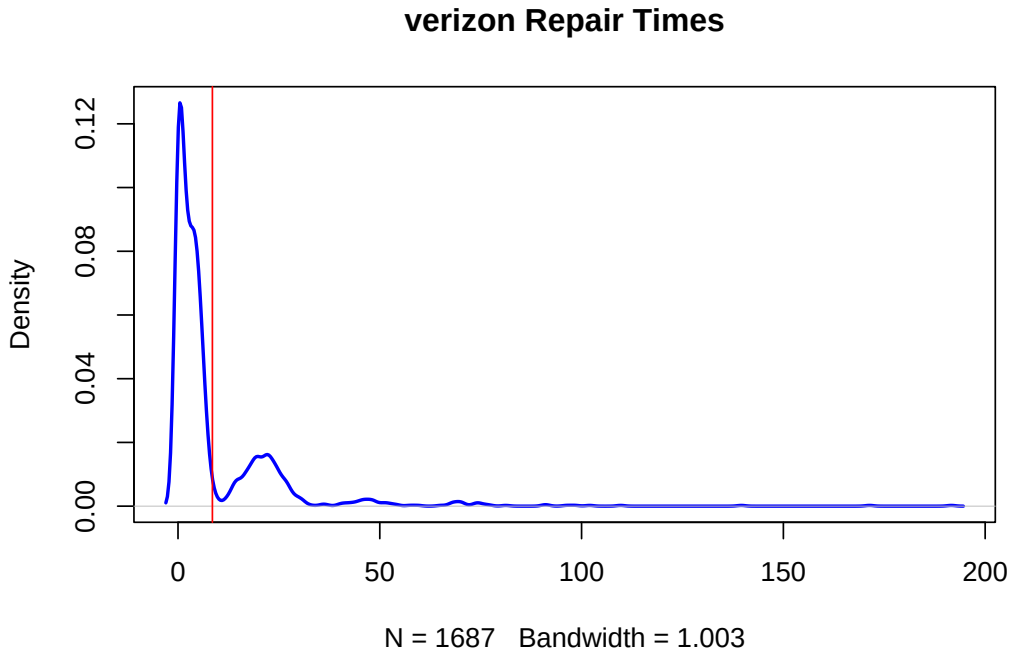
Question 1)

The following problem's data is real but the scenario is strictly imaginary The large American phone company Verizon had a monopoly on phone services in many areas of the US. The New York Public Utilities Commission (PUC) regularly monitors repair times with customers in New York to verify the quality of Verizon's services. The file verizon.csv has a recent sample of repair times collected by the PUC.

a) Imagine that Verizon claims that they take 7.6 minutes to repair phone services for its customers on average. The PUC seeks to verify this claim at 99% confidence (i.e., significance = 1%) using traditional statistical methods.

```
# import the Verizon.csv into df
df <- read.csv("verizon.csv", header = TRUE)

# show the result in density plot
plot(density(df$Time), col = "blue", lwd = 2, main = "verizon Repair Times")

# mark the mean on the plot
abline(v = mean(df$Time), col = "red")
```

i) Visualize the distribution of Verizon's repair times, marking the mean with a vertical line



ii) Given what the PUC wishes to test, how would you write the hypothesis? (not graded)
null hypothesis: the mean repair time of Verizon is equal to 7.6 minutes

alternative hypothesis: the mean repair time of Verizon is different from 7.6 minutes

```
# use variables to store population mean, standard deviation and n
x_bar <- mean(df$Time)
standard_deviaton <- sd(df$Time)
len <- length(df$Time)

#Print the mean
cat("population mean: ", x_bar, "\n")
```

iii) Estimate the population mean, and the 99% confidence interval (CI) of this estimate.

```
## population mean:  8.522009
```

```
# Create CI
part_a_ci <- c(x_bar - 2.58 * (standard_deviaton / sqrt(len)), x_bar + 2.58 * (standard_deviaton / sqrt
cat("99% CI: [", part_a_ci,"]\n")
```

```
## 99% CI: [ 7.593073 9.450946 ]
```

```
# use the function t.test to show the result
t.test(df$Time, mu = 7.6, conf.level = 0.99)
```

iv) Find the t-statistic and p-value of the test

```
##
##  One Sample t-test
##
## data:  df$Time
## t = 2.5608, df = 1686, p-value = 0.01053
## alternative hypothesis: true mean is not equal to 7.6
## 99 percent confidence interval:
##  7.593524 9.450495
## sample estimates:
## mean of x
##  8.522009
```

By the table, we can see that t-statistic is 2.5608 and the p-value is 0.01053

v) Briefly describe how these values relate to the Null distribution of t (not graded) The t-statistic measures the difference between the sample mean and the hypothesized mean in units of the standard error of the mean. The p-value is the probability of observing a t-statistic as extreme or more extreme than the one obtained under the null hypothesis.

vi) What is your conclusion about the company's claim from this t-statistic, and why? Since the p-value $= 0.01053 > 0.01$, we can't reject the null hypothesis: the mean repair time of Verizon is equal to 7.6 minutes.

Therefore, we can't conclude that Verizon's claim is false based on this sample of repair times.

b) Let's re-examine Verizon's claim that they take no more than 7.6 minutes on average, but this time using bootstrapped testing:

```r
# Create the bootstrapped function
boot_mean <- function(sample0) {
  resample <- sample(sample0, length(sample0), replace = TRUE)
  return(mean(resample))
}

# Set the seed and decide the number of bootstrap
set.seed(42379878)
num_boots <- 2000

# Calculate the 99% CI
popu_mean <- replicate(num_boots, boot_mean(df$Time))
CI <- quantile(popu_mean, probs = c(0.005, 0.995))

# Print the result
cat("99% CI: [", CI, "]\n")
```

i. Bootstrapped Percentile: Estimate the bootstrapped 99% CI of the population mean

```
## 99% CI: [ 7.585826 9.499418 ]
```

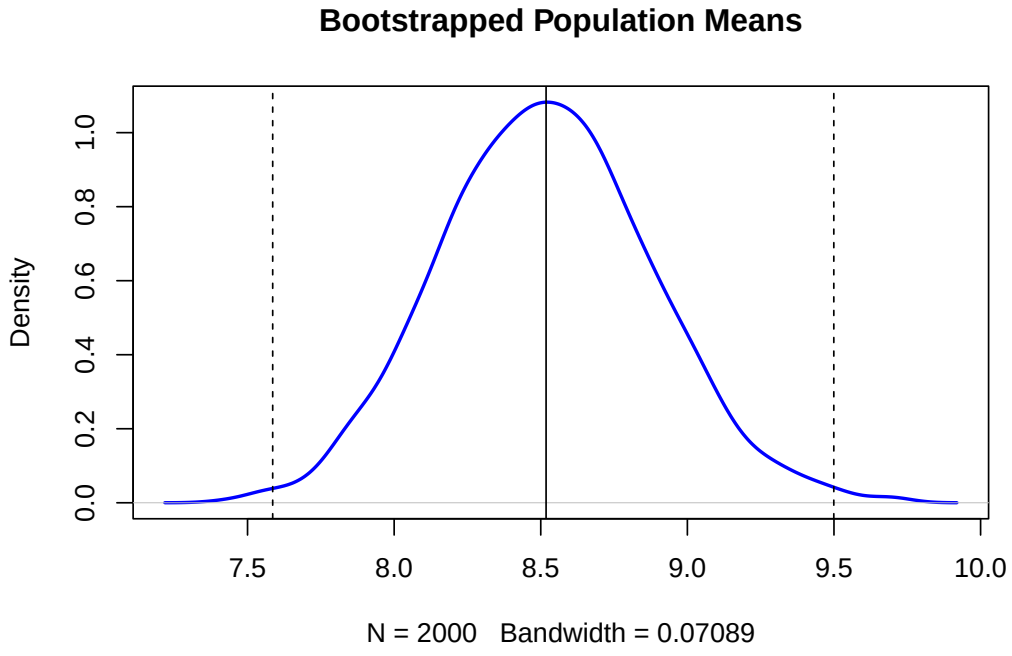
```r
repair_time_hypo <- 7.6

# Create the bootstrapped function
boot_mean_diffs <- function(sample0, mean_hypo) {
  resample <- sample(sample0, length(sample0), replace = TRUE)
  return(mean(resample) - mean_hypo)
}

# Set the seed and decide the number of bootstrap
set.seed(42379878)
num_boots <- 2000

# Calculate the 99% CI
mean_diffs <- replicate(num_boots, boot_mean_diffs(df$Time, repair_time_hypo))
diff_ci_99 <- quantile(mean_diffs, probs = c(0.005, 0.995))

# Print the result
cat("99% CI: [", diff_ci_99, "]\n")
```

ii. Bootstrapped Difference of Means: What is the 99% CI of the bootstrapped difference between the sample mean and the hypothesized mean?

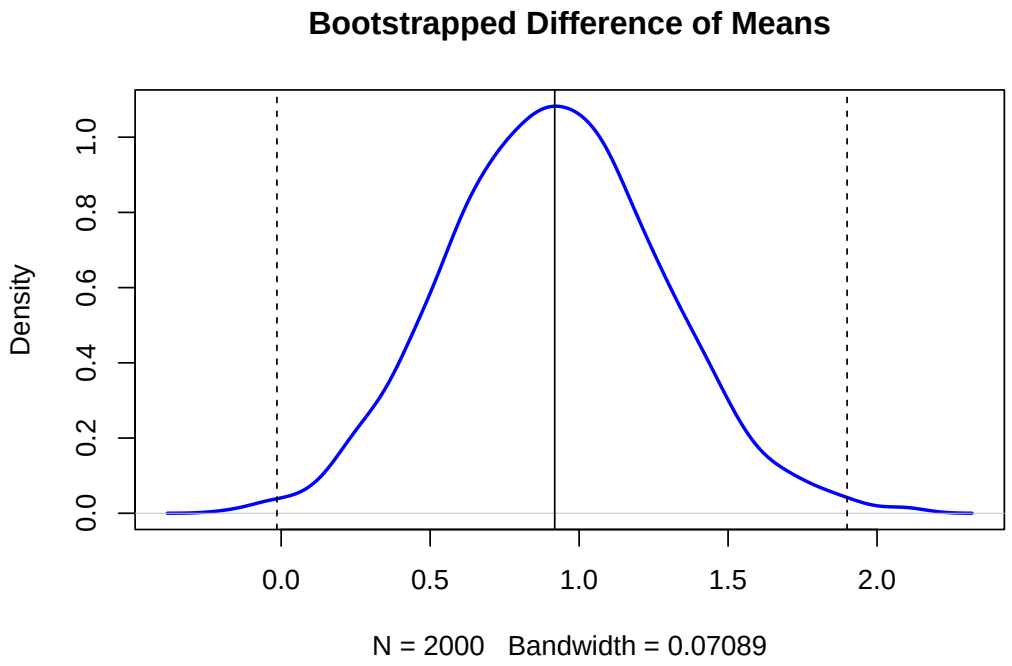
```
## 99% CI: [ -0.01417365 1.899418 ]
```

```r
# Show the result of plot i
plot(density(popu_mean), main = "Bootstrapped Population Means", col = "blue", lwd = 2)
abline(v = mean(popu_mean))
abline(v = CI, lty = "dashed")
```

iii. Plot distribution the two bootstraps above




```r
# Show the result of plot ii
plot(density(mean_diffs), main = "Bootstrapped Difference of Means", col = "blue", lwd = 2)
abline(v = mean(mean_diffs))
abline(v = diff_ci_99, lty = "dashed")
```



iv. Does the bootstrapped approach agree with the traditional t-test in part [a]? Yes, since 7.6 is in the 99% CI in i. and 0 is in the 99% CI in ii., we can't reject the null hypothesis in part [a].

Therefore, the bootstrapped approach agree with the traditional t-test in part [a].

c) Finally, imagine that Verizon notes that the distribution of repair times is highly skewed by outliers, and feel that testing the mean in not fair because the mean is sensitive to outliers. They claim that the median is a more fair test, and claim that the median repair time is no more than 3.5 minutes at 99% confidence (i.e., significance = 1%).

```
# Create the bootstrapped function
boot_median <- function(sample0) {
  resample <- sample(sample0, length(sample0), replace = TRUE)
  return(median(resample))
}

# Set the seed and decide the number of bootstrap
set.seed(42379878)
num_boots <- 2000

# Calculate the 99% CI
median_boots <- replicate(num_boots, boot_median(df$Time))
median_boots_99ci <- quantile(median_boots, probs = c(0.005, 0.995))

# Print the result
cat("99% CI: [", median_boots_99ci,"]\n")
```

i. Bootstrapped Percentile: Estimate the bootstrapped 99% CI of the population median

```
## 99% CI: [ 3.22 3.93 ]
```

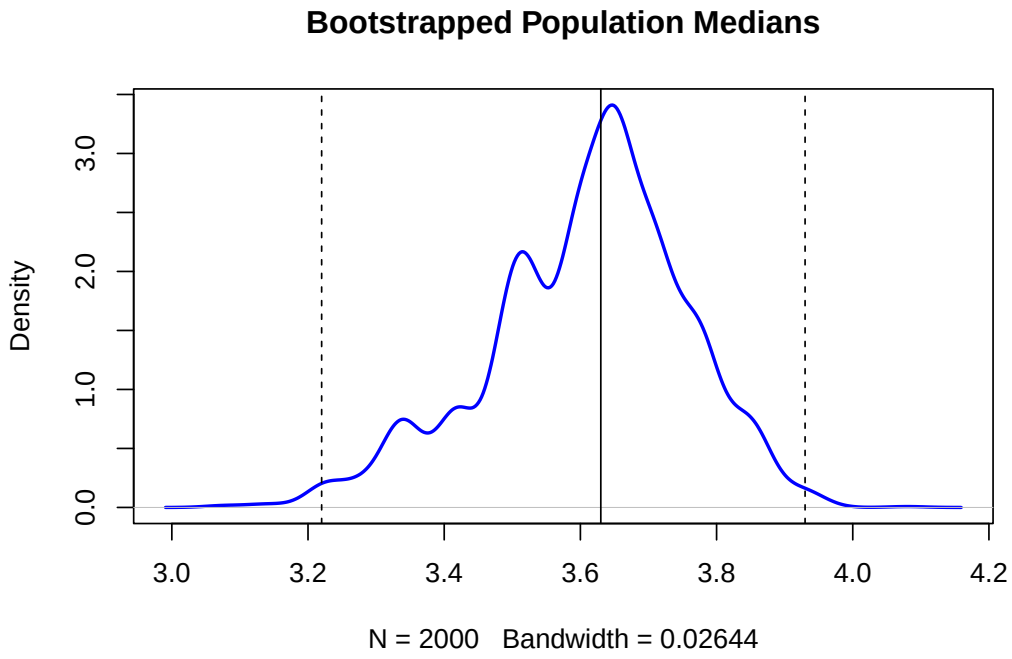
```
median_repair_time_hypo <- 3.5

# Create the bootstrapped function
boot_median_diffs <- function(sample0, median_hypo) {
  resample <- sample(sample0, length(sample0), replace = TRUE)
  return(median(resample) - median_hypo)
}

# Set the seed and decide the number of bootstrap
set.seed(42379878)
num_boots <- 2000

# Calculate the 99% CI
median_diffs <- replicate(num_boots, boot_median_diffs(df$Time, median_repair_time_hypo))
median_diffs_99ci <- quantile(median_diffs, probs = c(0.005, 0.995))

# Print the result
cat("99% CI: [", median_diffs_99ci, "]\n")
```

ii. Bootstrapped Difference of Medians: What is the 99% CI of the bootstrapped difference between the sample median and the hypothesized median?

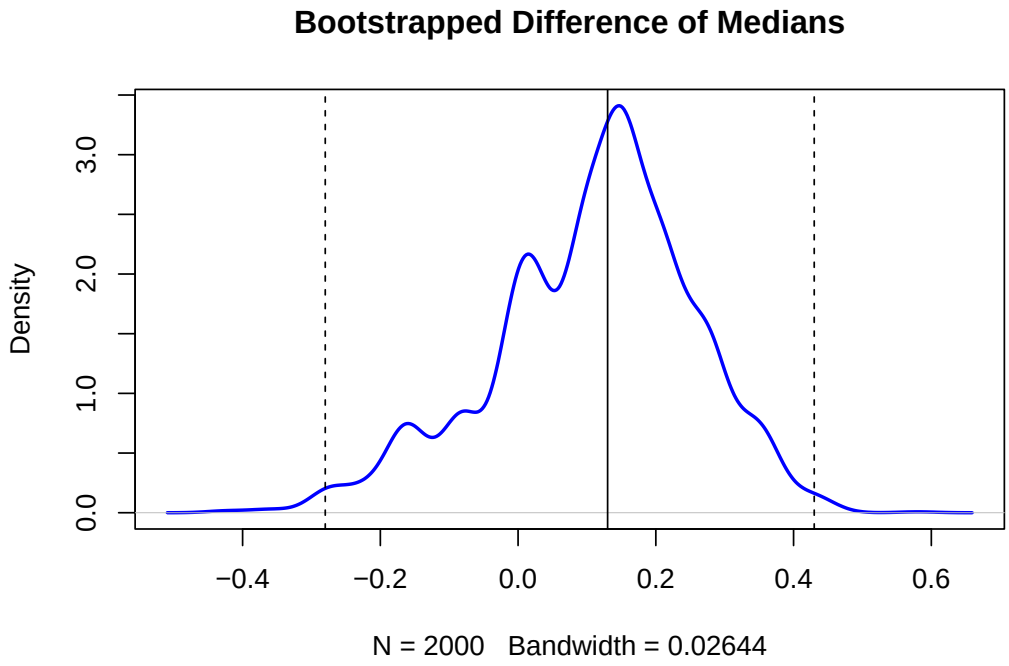
```
## 99% CI: [ -0.28 0.43 ]
```

```r
# Show the result of plot i
plot(density(median_boots), main = "Bootstrapped Population Medians", col = "blue", lwd = 2)
abline(v = median(median_boots))
abline(v = median_boots_99ci, lty = "dashed")
```

iii. Plot distribution the two bootstraps above




```r
# Show the result of plot ii
plot(density(median_diffs), main = "Bootstrapped Difference of Medians", col = "blue", lwd = 2)
abline(v = median(median_diffs))
abline(v = median_diffs_99ci, lty = "dashed")
```



iv. What is your conclusion about Verizon's claim about the median, and why? Yes, since 3.5 is in the 99% CI of population median, 0 is in the 99% CI of difference between the sample median and the hypothesized median, we can't reject the null hypothesis.

Therefore, we can't conclude that Verizon's claim is false.

Question 2)

Load the compstatslib package and run interactive_t_test(). You will see a simulation of null and alternative distributions of the t-statistic, along with significance and power. If you do not see interactive controls (slider bars), press the gears icon ( ) on the top-left of the visualization.

```
# import the library
# library(compstatslib)

# Show the plot
# interactive_t_test()
```

Recall the form of t-tests, where t = (x-0)s/n Let's see how hypothesis tests are are affected by various factors: diff: the difference we wish to test (x-0) sd: the standard deviation of our sample data (s) n: the number of cases in our sample data alpha: the significance level of our test (e.g., alpha is 5% for a 95% confidence level) Your colleague, a data analyst in your organization, is working on a hypothesis test where he has sampled product usage information from customers who are using a new smartwatch. He wishes to test whether the mean (xi) usage time is higher than the usage time of the company's previous smartwatch released two years ago ( ): Hnull: The mean usage time of the new smartwatch is the same or less than for the previous smartwatch. Halt: The mean usage time is greater than that of our previous smartwatch. After collecting data from just n=50 customers, he informs you that he has found diff=0.3 and sd=2.9. Your colleague believes that we cannot reject the null hypothesis at alpha of 5%. Use the slider bars of the simulation to the values your colleague found and confirm from the visualization that we cannot reject the null hypothesis. Consider the scenarios (a – d) independently using the simulation tool. For each scenario, start with the initial parameters above, then adjust them to answer the following questions:

i. Would this scenario create systematic or random error (or both or neither)?

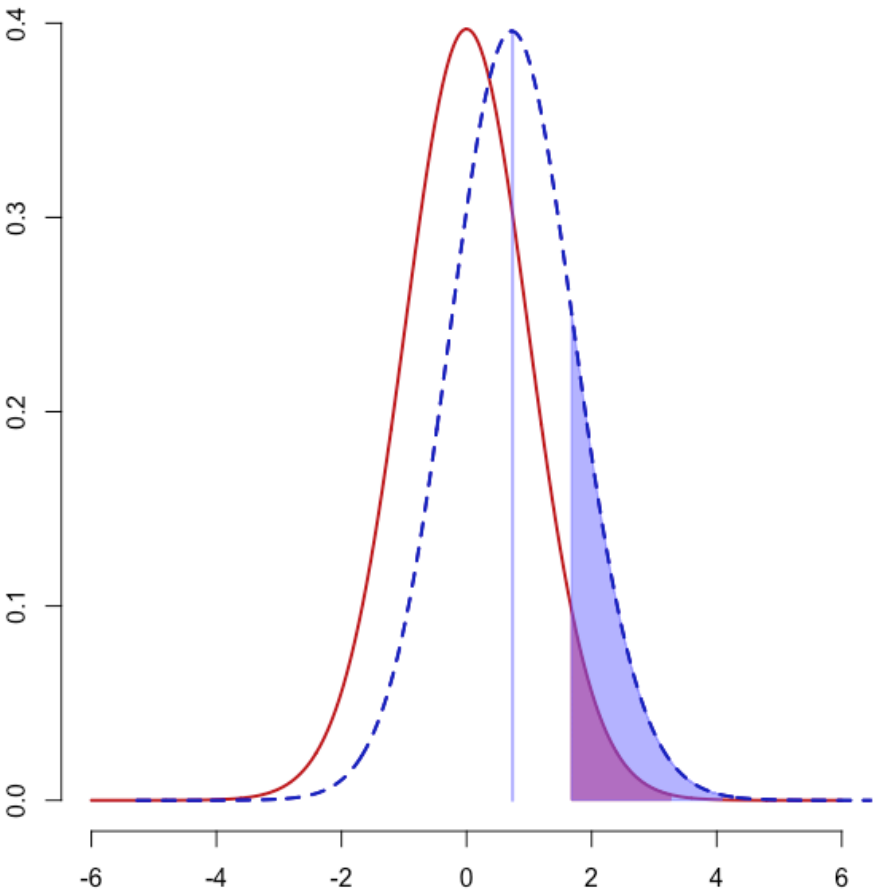
ii. Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?

iii. Will it increase or decrease our power to reject the null hypothesis?

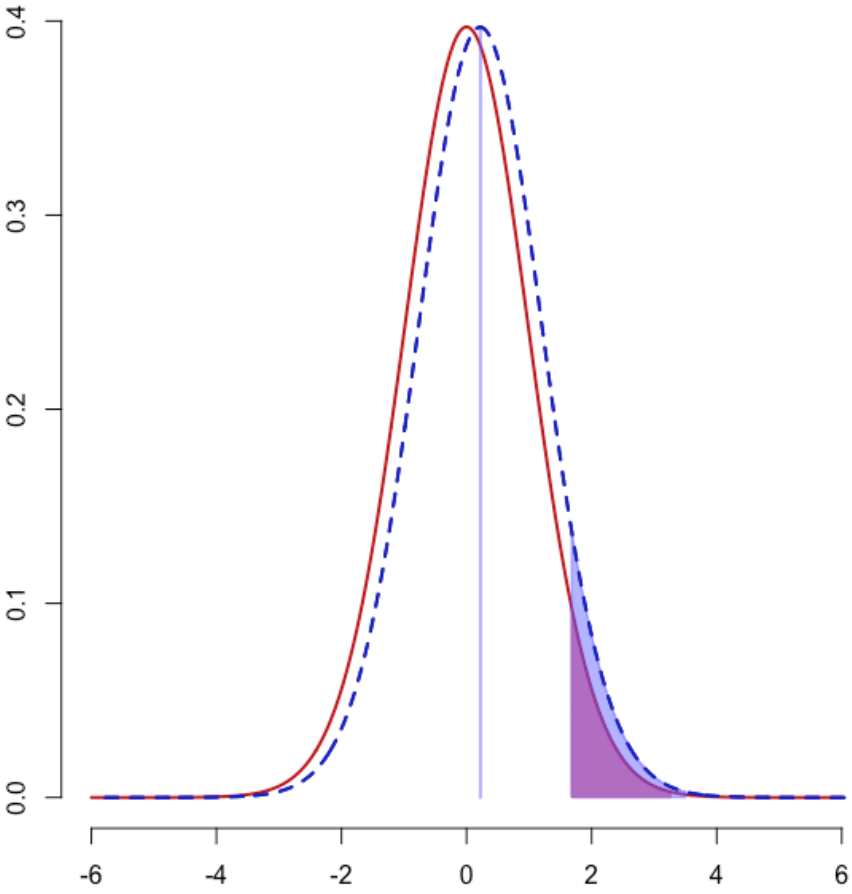
iv. Which kind of error (Type I or Type II) becomes more likely because of this scenario?

```
# This is the plot that we set df = 0.3, sd = 2.9, n = 50, alpha = 5%
knitr::include_graphics("original.png")
```

a) You discover that your colleague wanted to target the general population of Taiwanese users of the product. However, he only collected data from a pool of young consumers, and missed many older customers who you suspect might use the product much less every day.



```
# This is the plot that we decrease df and increase sd
knitr::include_graphics("senario_a.png")
```



Ans:

i. This scenario would create systematic error because the sample is not representative of the population.

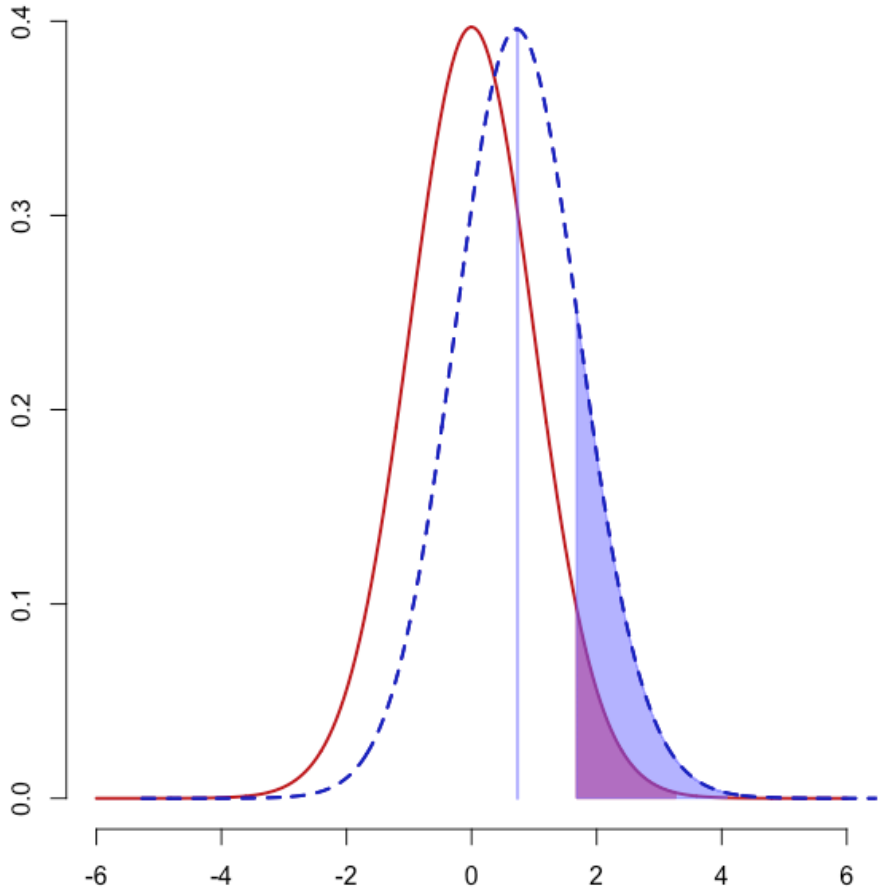
ii. diff and sd would be affected because when we consider older customers into the situation, diff will decrease and sd will increase.

iii. It would decrease our power to reject the null hypothesis because the diff decreases and sd increases.

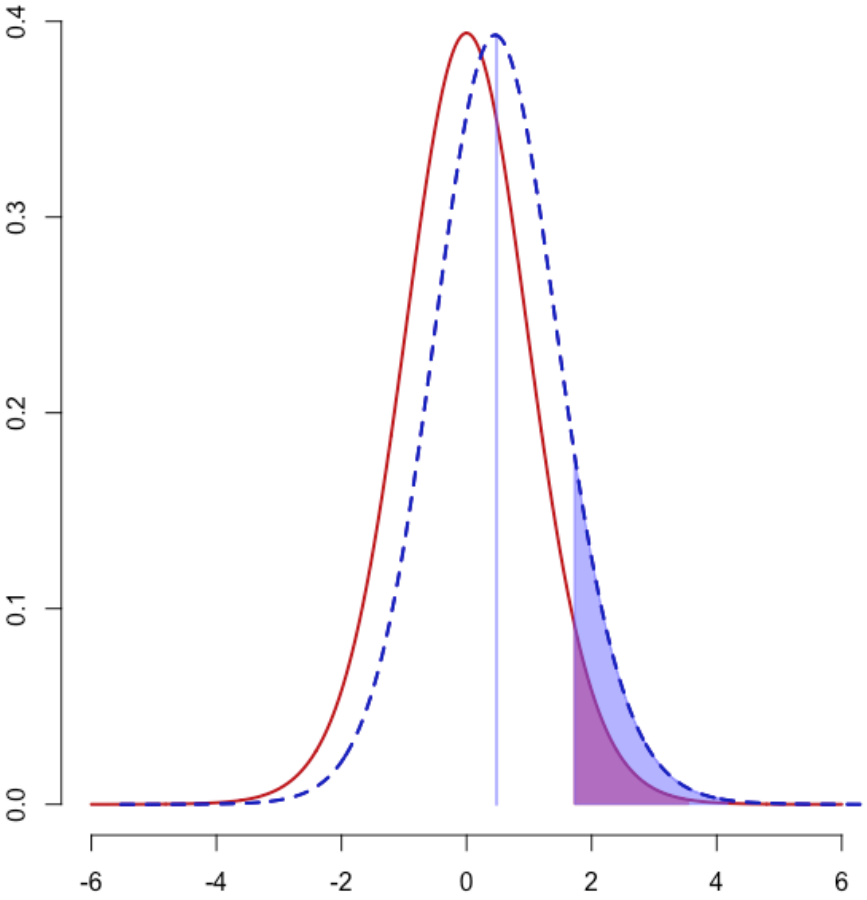
iv. Type II error becomes more likely because of this scenario.

```
# This is the plot that we set df = 0.3, sd = 2.9, n = 50, alpha = 5%
knitr::include_graphics("original.png")
```

b) You find that 20 of the respondents are reporting data from the wrong wearable device, so they should be removed from the data. These 20 people are just like the others in every other respect.




```
# This is the plot that we decrease n
knitr::include_graphics("senario_b.png")
```



Ans:

i. This scenario would create random error because the 20 respondents reporting data from the wrong wearable device are randomly selected.

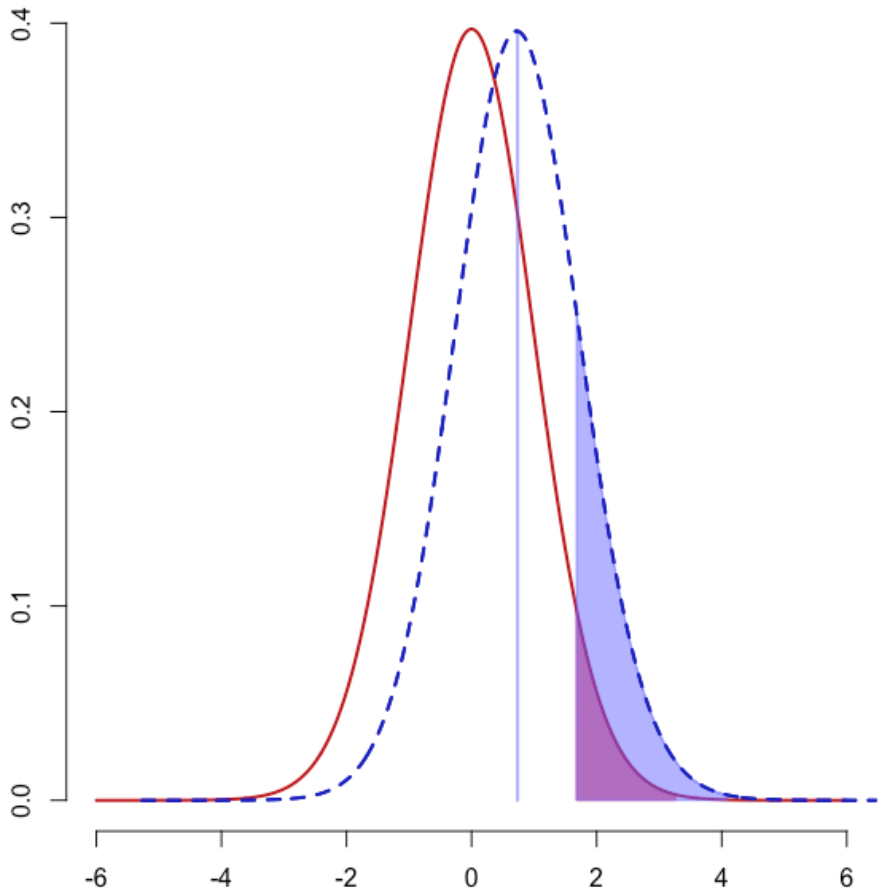
ii. n would be affected because we need to remove the wrong data.

iii. It would decrease our power to reject the null hypothesis because the sample size is smaller.

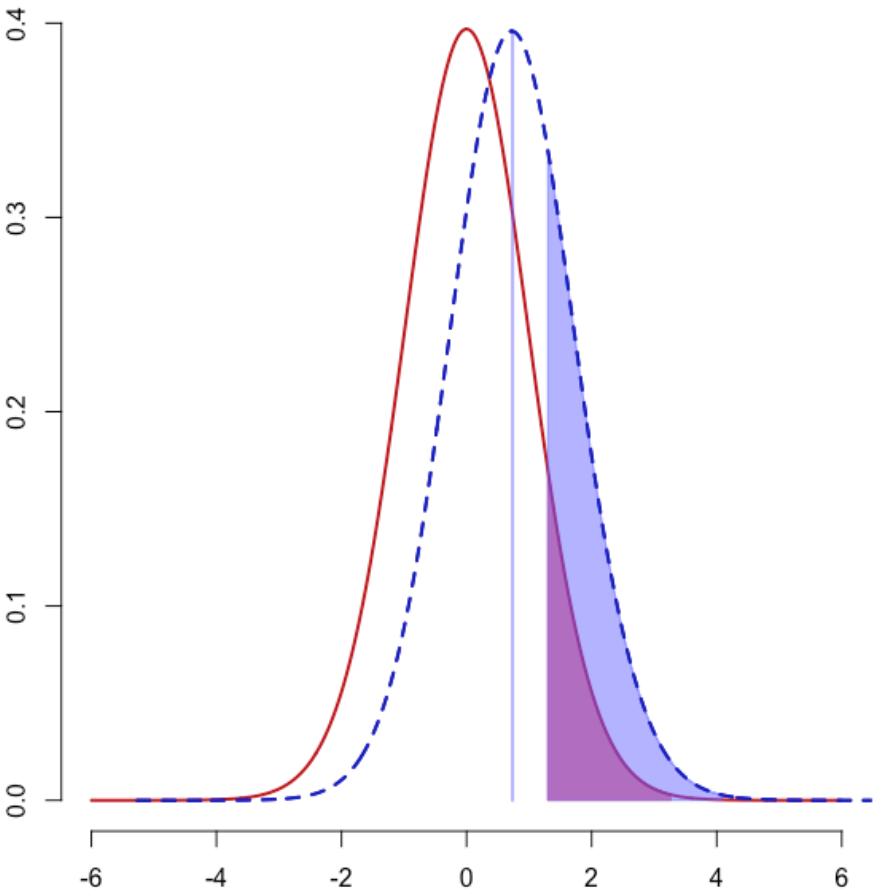
iv. Type II error becomes more likely because of this scenario.

```
# This is the plot that we set df = 0.3, sd = 2.9, n = 50, alpha = 5%
knitr::include_graphics("original.png")
```

c) A very annoying professor visiting your company has criticized your colleague's "95% confidence" criteria, and has suggested relaxing it to just 90%.




```
# This is the plot that we increase alpha
knitr::include_graphics("senario_c.png")
```



Ans:

i. This scenario doesn't have systematic error or random error because we just change the level of significance.

ii. alpha would be affected because the level of significance has been changed.

iii. It would increase our power to reject the null hypothesis because the level of significance is higher.

iv. Type I error becomes more likely because of this scenario.

d) Your colleague has measured usage times on five weekdays and taken a daily average. But you feel this will underreport usage for younger people who are very active on weekends, whereas it over-reports usage of older users. Ans:

i. This scenario would create systematic error because the sample is biased towards weekdays and not representative of the population.

ii. diff and sd would be affected but we can't know diff will increase or decrease since we don't the proportion of the young and older user in the data.

iii. We can't conclude whether the power will increase or decrease since we don't know diff will increase or decrease.

iv. We can't conclude type I or type II error will become more likely since the variation of diff can't be concluded.