BACS - HW5

109006241 - helped by 109060082, 109006217

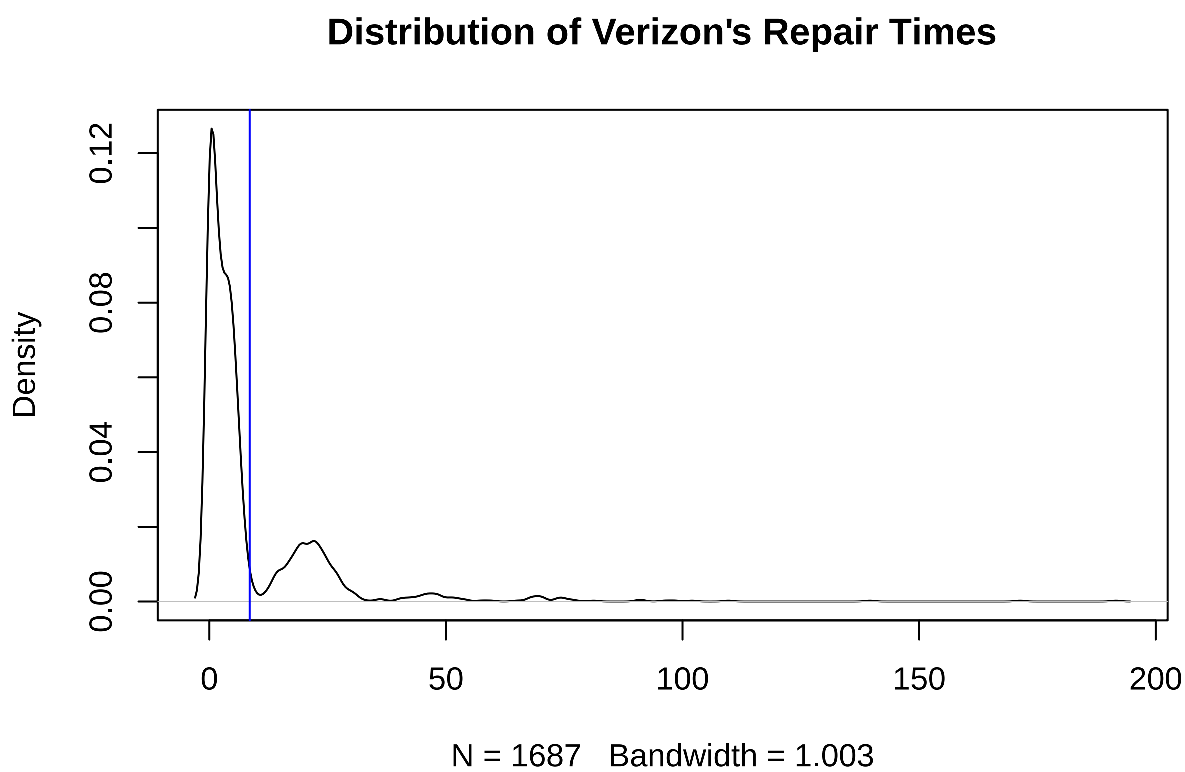
Question 1) The following problem's data is real but the scenario is strictly imaginary The large American phone company Verizon had a monopoly on phone services in many areas of the US. The New York Public Utilities Commission (PUC) regularly monitors repair times with customers in New York to verify the quality of Verizon's services. The file verizon.csv has a recent sample of repair times collected by the PUC.

```
verizon = read.csv('D:/Users/User/Documents/R/BACS/Homeworks/HW5/verizon.csv')
```

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 alpha = 1%) using traditional statistical methods.

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

```
plot(density(verizon$Time), main="Distribution of Verizon's Repair Times")
abline(v=mean(verizon$Time), col="blue")
```



ii. Given what the PUC wishes to test, how would you write the hypothesis? (not graded)

H0: X <= 7.6

H1: X > 7.6

Where X denotes the mean repair time, in minutes.

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

```
n = length(verizon$Time)
verizon_mean = sum(verizon$Time) / n
verizon_sd = sqrt(sum((verizon$Time - mean(verizon$Time))^2) / (n - 1))
verizon_stderror = verizon_sd / sqrt(n)
t = qt(p=0.01, df=n-1, lower.tail=F)
verizon_99CI = verizon_mean - t * verizon_stderror

cat(" Population Mean =", verizon_mean, "\n",
     "99% Confidence Interval =", verizon_99CI, "-", "Inf")
```

```
##  Population Mean = 8.522009
##  99% Confidence Interval = 7.683604 - Inf
```

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

```r
t_statistic = (verizon_mean - 7.6) / (verizon_sd / sqrt(n))
p_value = 1 - pt(t_statistic, df=n-1)
cat(" T-statistic =", t_statistic, "\n",
     "P-value =", p_value)
```

```
##  T-statistic = 2.560762
##  P-value = 0.005265342
```

v. Briefly describe how these values relate to the Null distribution of t (not graded)

> The t-statistic is the standardized version of the distribution with respect to the null distribution. In other words, it is a ratio of of the departure of the estimated value of a parameter from its hypothesized value to its standard error.
>
> The p-value is the probability of the occurrence of a sample with at least the same extremes as the actual sample, assuming the null hypothesis is true in the hypothesis test.

vi. What is your conclusion about the company's claim from this t-statistic, and why?

> The t-statistic is 2.560762, and the p-value, which is 0.005, is less than alpha = 0.01. So, we can reject the null hypothesis, meaning that we can say there is a strong evidence that the mean repair time is more than 7.6 minutes at 99% confidence level.

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
sample_statistic = function(stat_function, sample0) {
  resample = sample(sample0, length(sample0), replace=TRUE)
  return (stat_function(resample))
}

sample_means = replicate(500, sample_statistic(mean, verizon$Time))
```

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

```r
bootstrap_mean_99CI = quantile(sample_means, probs=0.01)
cat("99% CI of the bootstrapped means =", bootstrap_mean_99CI, "-", "Inf")
```

```
## 99% CI of the bootstrapped means = 7.663724 - Inf
```

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

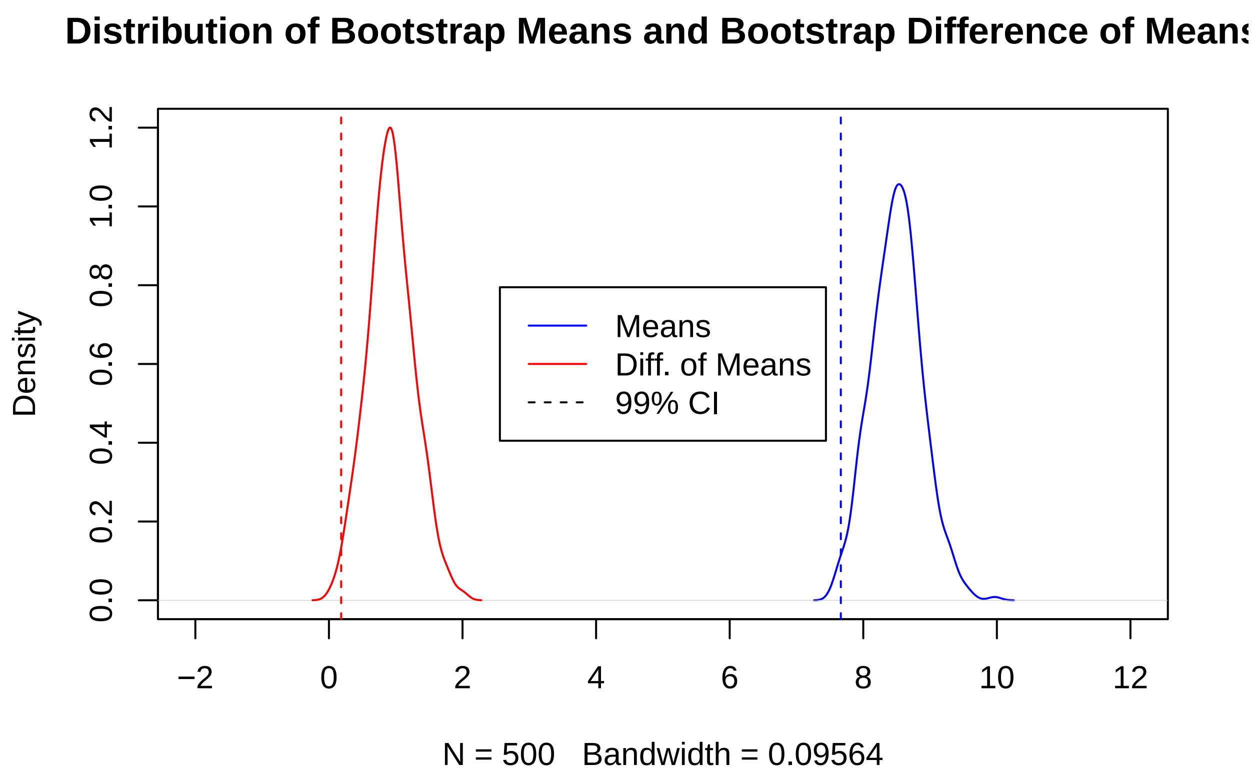
```r
boot_mean_diffs = function(sample0, mean_hyp) {
  resample = sample(sample0, length(sample0), replace=TRUE)
  return (mean(resample) - mean_hyp)
}

mean_diffs = replicate(
```

```
  500,
  boot_mean_diffs(verizon$Time, 7.6)
)

mean_diff_99CI = quantile(mean_diffs, probs=0.01)
cat("99% CI of the bootstrapped difference of means =", mean_diff_99CI, "-", "Inf")
```

```
## 99% CI of the bootstrapped difference of means = 0.1833354 - Inf
```

iii. Plot distribution the two bootstraps above

```
plot(density(sample_means), col="blue", xlim=c(-2,12), ylim=c(0,1.2), main="Distribution of Bootstrap Me
abline(v=bootstrap_mean_99CI, col="blue", lty="dashed")
lines(density(mean_diffs), col="red")
abline(v=mean_diff_99CI, col="red", lty="dashed")
legend("center", legend=c("Means", "Diff. of Means", "99% CI"),
       col=c("blue", "red", "black"),
       lty=c(1, 1, 2))
```



iv. Does the bootstrapped approach agree with the traditional t-test in part [a]?

Yes. The hypothesized mean, which is 7.6, is not included in the bootstrapped 99% CI of the population mean, and also, the 99% CI of the bootstrapped difference between the sample mean and the hypothesized mean does not contain zero. So, we can say that there is a strong evidence that the mean repair time is more than 7.6 minutes at 99% confidence level.

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 alpha = 1%).

```
sample_statistic = function(stat_function, sample0) {
  resample = sample(sample0, length(sample0), replace=TRUE)
  return (stat_function(resample))
}

sample_medians = replicate(500, sample_statistic(median, verizon$Time))
```

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

```
bootstrap_median_99CI = quantile(sample_medians, probs=0.01)
cat("99% CI of the bootstrapped medians =", bootstrap_median_99CI, "-", "Inf")
```

```
## 99% CI of the bootstrapped medians = 3.2196 - Inf
```

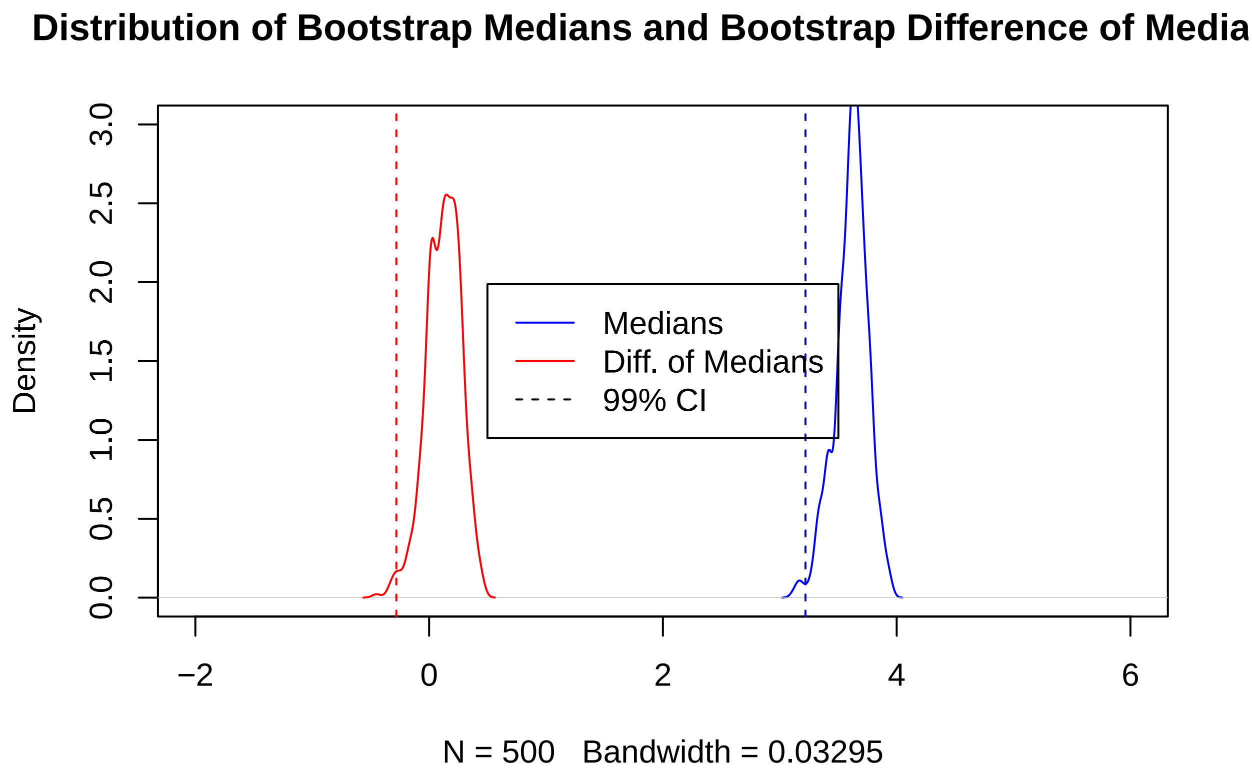
ii. Bootstrapped Difference of Medians: What is the 99% CI of the bootstrapped difference between the sample median and the hypothesized median?

```
boot_median_diffs = function(sample0, median_hyp) {
  resample = sample(sample0, length(sample0), replace=TRUE)
  return (median(resample) - median_hyp)
}

median_diffs = replicate(
  500,
  boot_median_diffs(verizon$Time, 3.5)
)

median_diff_99CI = quantile(median_diffs, probs=0.01)
cat("99% CI of the bootstrapped difference of medians =", median_diff_99CI, "-", "Inf")
```

```
## 99% CI of the bootstrapped difference of medians = -0.28 - Inf
```

iii. Plot distribution the two bootstraps above

```
plot(density(sample_medians), col="blue", xlim=c(-2,6), ylim=c(0,3), main="Distribution of Bootstrap Me
abline(v=bootstrap_median_99CI, col="blue", lty="dashed")
lines(density(median_diffs), col="red")
abline(v=median_diff_99CI, col="red", lty="dashed")
legend("center", legend=c("Medians", "Diff. of Medians", "99% CI"),
       col=c("blue", "red", "black"),
       lty=c(1, 1, 2))
```



iv. What is your conclusion about Verizon's claim about the median, and why?

The hypothesized median, which is 3.5, is included in the bootstrapped 99% CI of the population median. Also, in the 99% CI of the bootstrapped difference between the sample median and the hypothesized median contains zero. So, we can say that there is a strong evidence that the median repair time is no more than 3.5 minutes at 99% confidence level. This contradicts our result in (a) and (b), meaning that using the median instead of mean, we can get a more fair test, yielding a different result from our original test using the mean.

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.

Let's see how hypothesis tests are are affected by various factors:

- diff: the difference we wish to test
- 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?

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.

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

> Systematic error. Because the sample data only includes data from young customers, while missing data from many older customers.

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

> Diff and sd would be affected.

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

> It may increase or decrease, depending on if our diff and sd increases / decreases. Power may increase when diff is increased or sd is decreased.

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

> Type II error.

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.

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

Random error. Because the occurrence of the 20 respondents which reports the data from a wrong wearable device is a random and unpredictable event.

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

Diff, sd, and n would be affected.

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

It will decrease the power.

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

Type II error.

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

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

Neither.

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

Only the alpha would be affected.

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

It will increase the power.

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

Type II error.

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.

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

Systematic error. Because by missing the weekends, the data won't be accurately represented.

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

Diff and sd would be affected.

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

It may increase or decrease, depending on if our diff and sd increases / decreases. Power may increase when diff is increased or sd is decreased.

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

Type II error.