BACS HW5 - 109006234

Credit: 109006278

March 19th 2023

Supporting Functions

```
plotFunct <- function(data, title){
  hist(data, prob=TRUE, main = title)
  lines(density(data), lwd = 2, col = "#fb8e8e", main = title)
  abline(v = mean(data), col = "blue")
}
```

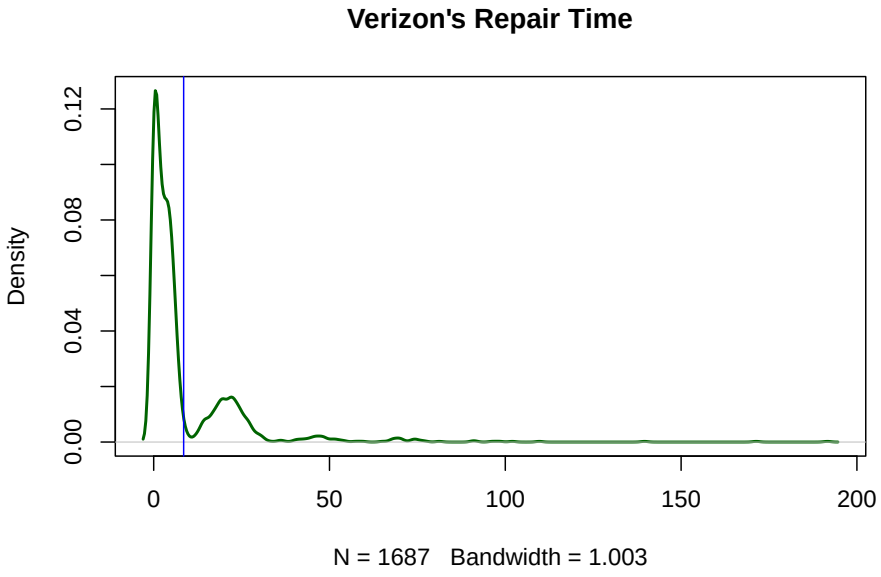
Problem 1.

(a) Claim: Verizon takes 7.6 minutes to repair a phone. PUC seeks to veritfy this claim at 99% confindence. (significance = 1%)

```
datas <- read.csv("G:/My Drive/111_2_BACS/HW5/verizon.csv")
ver_time <- datas$Time
ver_mean <- mean(ver_time)
ver_sd <- sd(ver_time)
ver_size <- length(ver_time)
```

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

```
plot(density(ver_time), col="darkgreen", lwd=2,
    main = "Verizon's Repair Time")
abline(v=mean(ver_time), col="blue")
```



(ii) Given what the PUC wishes to test, how would you write the hypothesis?
Given the data above, the null hypothesis will be that Verizon's average repair time is 7.6 minutes. The alternative hypothesis will be that Verizon's average repair time is not equal to 7.6 minutes.

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

```
## [1] "The estimated mean is : 8.52200948429164"
```

```
## [1] "The 99% Confindence Interval is between 7.59307342589326 and 9.45094554269003"
```

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

```
## [1] "T-Statistics:  2.56076233446444"
```

```
## [1] "P-Value:  0.010530684588578"
```

(v) Briefly describe how these values relate to the Null distribution of t
The t-statistic measures how many standard deviations the sample mean is away from the null hypothesis mean, while the p-value measures the probability of observing a t-statistic as extreme or more extreme than the one observed, given the null distribution of t.

(vi) What is your conclusion about the company's claim from this t-statistic, and why?
As we have observed, we have 1% (0.005 as it is two-sided) significance level. Then, we obtained the P-value of the dataset to be smaller than 0.01. This means that we fail to reject the null hypothesis, meaning that Verizon's repairing time is 7.6 minutes.

(b) Bootstrapped Testing - Mean repair time: 7.6 minutes

```
boot_mean <- function(sample0) {
  resample <- sample(sample0, length(ver_time), replace=TRUE)
  return(mean(resample))
}
set.seed(10)
sample_means <- replicate(2000, boot_mean(ver_time))
```

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

```
mean(sample_means)
```

```
## [1] 8.512661
```

```
sd_error <- sd(sample_means) / length(sample_means)^0.5
sd_error
```

```
## [1] 0.008168943
```

```
ci99_boot <- mean(sample_means) + c(-2.33, 2.33)*sd_error
print(paste("The 99% Confindence Interval is between", ci99_boot[1], "and", ci99_boot[2]))
```

```
## [1] "The 99% Confindence Interval is between 8.49362745397299 and 8.53169472741409"
```

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

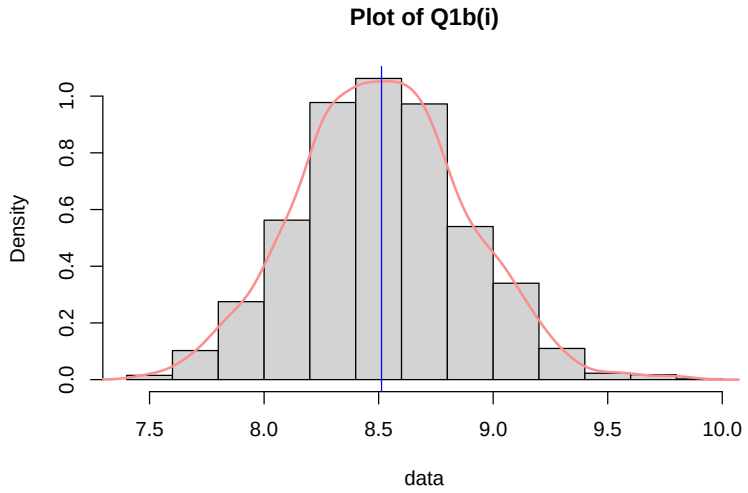
```
diff <- sample_means - 7.6
mean(abs(diff))
```

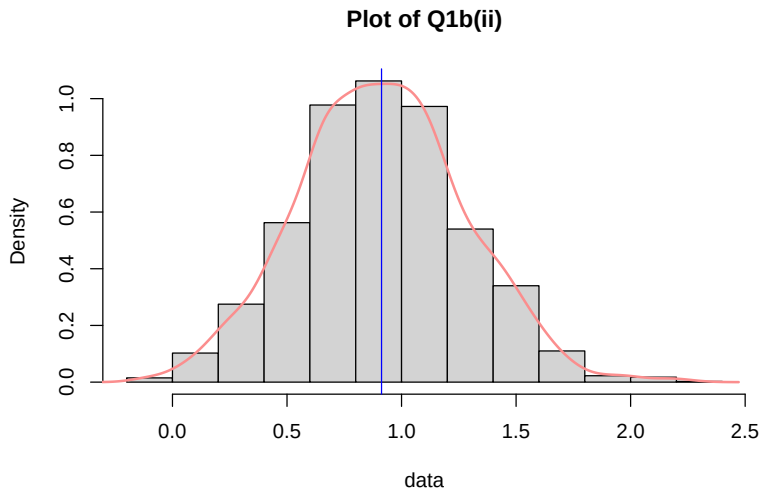
```
## [1] 0.9132078
```

```
diff_sderror <- sd(diff) / length(diff)^0.5
diff_ci99 <- mean(abs(diff)) + c(-2.33, 2.33) * diff_sderror
print(paste("The 99% Confindence Interval is between", diff_ci99[1], "and", diff_ci99[2]))
```

```
## [1] "The 99% Confindence Interval is between 0.894174193747735 and 0.93224146718884"
```

(iii) Plot distribution the two bootstraps above





(iv) Does the bootstrapped approach agree with the traditional t-test in part (a)?
From both plots above, we can observe that both plots have a similar shape but with a different x-axis. We can see that the mean of the traditional t-test falls in between its 99% confidence interval.
In the bootstrapped approach, we also observe that the mean of the bootstrapped test falls between its 99% confidence interval. As both falls within its 99% confidence interval, both hypothesis do not reject the null hypothesis.

(c) Bootstraping Testing - Median repair time: 3.5 minutes

```
boot_median <- function(sample0) {
  resample <- sample(sample0, length(ver_time), replace=TRUE)
  return(median(resample))
}
set.seed(10)
sample_median <- replicate(2000, boot_median(ver_time))
```

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

```
mean(sample_median)
```

```
## [1] 3.614585
```

```
sd_error <- sd(sample_median) / length(sample_median)^0.5
ci99_boot_median <- mean(sample_median) + c(-2.33, 2.33)*sd_error
print(paste("The 99% Confindence Interval is between", ci99_boot_median[1],
    "and", ci99_boot_median[2]))
```

```
## [1] "The 99% Confindence Interval is between 3.60713392479774 and 3.62203607520226"
```

(ii) Bootstrapped Difference of Means: What is the 99% CI of the bootstrapped difference between the sample median and the hypothesized median?

```
diff_median <- sample_median - 3.5
mean(abs(diff_median))
```

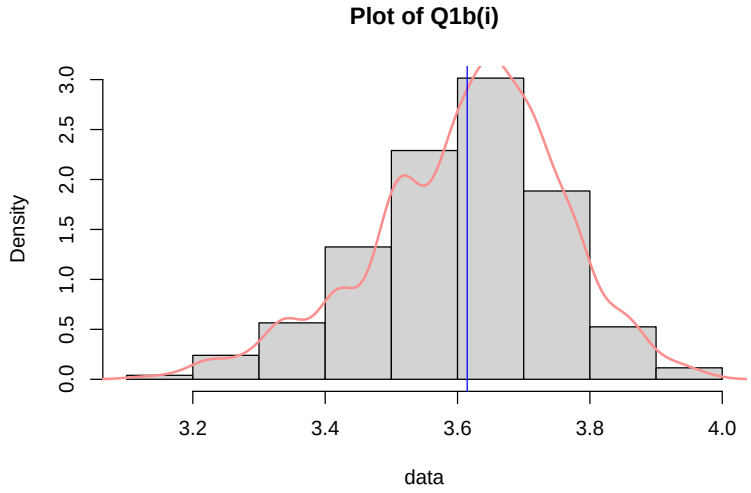
```
## [1] 0.154665
```

```
diff_sderror_median <- sd(diff_median) / length(diff_median)^0.5
diff_ci99_median <- mean(abs(diff_median)) + c(-2.33, 2.33) * diff_sderror_median
print(paste("The 99% Confindence Interval is between", diff_ci99_median[1],
    "and", diff_ci99_median[2]))
```

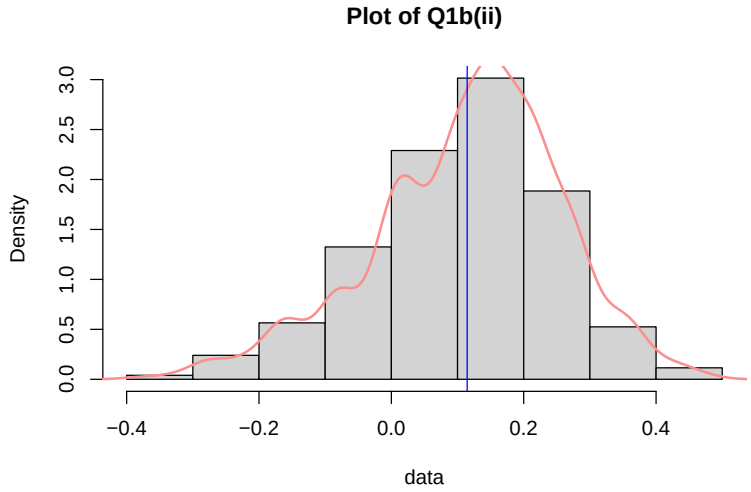
```
## [1] "The 99% Confindence Interval is between 0.147213924797743 and 0.162116075202257"
```

(iii) Plot distribution the two bootstraps above

```
plotFunct(sample_median, "Plot of Q1b(i)")
```




```
plotFunct(diff_median, "Plot of Q1b(ii)")
```



(iv) What is your conclusion about Verizon's claim about the median, and why?
As we have observed the values above, we obtained that the mean falls between the 99% confidence interval, which means we fail to reject the hypothesis that the median of Verizon's repair time is 3.5 minutes.

Problem 2. - T-tests 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)?
This scenario could create systematic error in the statistical analysis because the sample used to collect the data is not representative of the entire population of Taiwanese users of the product, as the target was.

(ii) Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?
As the older users use the product much less, the standard deviation would be affected because if the data is included/excluded, the result would be a biased estimate of the entire population.

(iii) Will it increase or decrease our power to reject the null hypothesis?
In the scenario described, collecting data only from a pool of young consumers instead of the general population of Taiwanese users may decrease the power to reject the null hypothesis. This is because the sample may not accurately reflect the characteristics and behaviors of the entire population.

(iv) Which kind of error (Type I or Type II) becomes more likely because of this scenario?
In the scenario described, collecting data only from a pool of young consumers instead of the general population of Taiwanese users may increase the risk of 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)?
This scenario could create random error in the statistical analysis because the 20 respondents all have the same characteristics.

(ii) Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?
As the sample size decreases, the standard error of the mean increases and the t-value decreases, hence sample size would be affected.

(iii) Will it increase or decrease our power to reject the null hypothesis?
Removing 20 respondents from the data will decrease the sample size, which can reduce the power to reject the null hypothesis.

(iv) Which kind of error (Type I or Type II) becomes more likely because of this scenario?
As from the explaination above, if sample size decreases, then the power to reject the null hypothesis decreases, which means that Type II error becomes more likely.

(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)?
Relaxing the confidence criteria from 95% to 90% would not necessarily create systematic or random errors in and of itself.

(ii) Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?
The significance level would be affected because it would be increased from 0.05 to 0.1.

(iii) Will it increase or decrease our power to reject the null hypothesis?
The power to reject the null hypothesis would increase because we are allowing for more uncertainty in our results.

(iv) Which kind of error (Type I or Type II) becomes more likely because of this scenario?
As the scenario would simply decrease the level of confidence required to reject the null hypothesis, meaning that the alternative hypothesis would be more willing to accepted even if the evidence is not as strong. This would increase the likelihood of making a Type I error but decrease the likelihood of making a 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)?
This scenario could create systematic error in the statistical analysis because the sample used to collect the data may not be representative of the entire population of users, and therefore, may not accurately reflect the characteristics of the population parameter being estimated.

(ii) Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?
The difference between the sample mean and the hypothesized population mean would be affected, as the sample mean may not be a good estimate of the true population mean. Additionally, the standard deviation of the sample may also be affected if the measurement process has different levels of variability for younger and older users.

(iii) Will it increase or decrease our power to reject the null hypothesis?
The larger the standard deviation, our power to reject the null hypothesis test is more likely. In additional to that, a larger difference between the sample mean and the hypothesized value results in a larger test statistic, which increases the power of the test.

(iv) Which kind of error (Type I or Type II) becomes more likely because of this scenario?
The scenario described here could increase the likelihood of both Type I and Type II errors, as it could result in a biased estimate of the population mean and affect the accuracy and power of the hypothesis test.