Homework Week 5

Each code piece that evaluates to some value has this value directly after the code in the document, usually displayed as `## value` or as an image.

```
## read the csv file and store in a data frame.
vdata <- read.csv("verizon.csv")
```

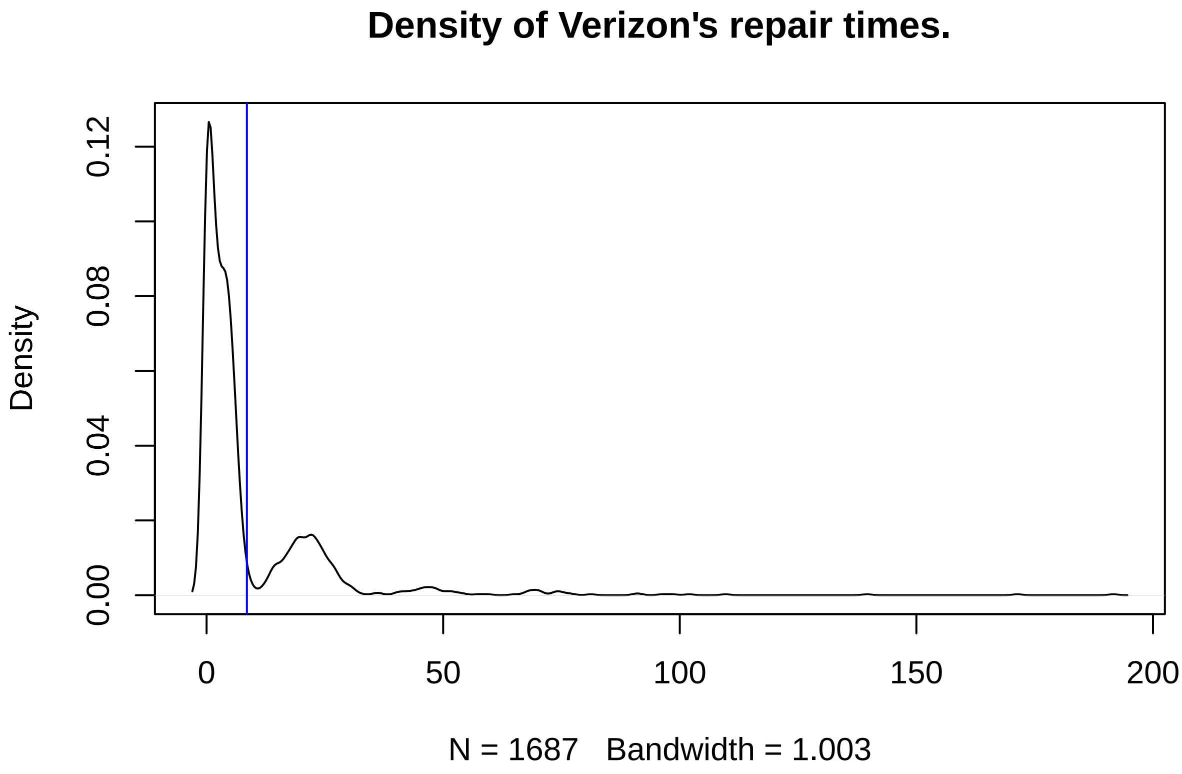
1

(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 a = 1%) using traditional statistical methods.

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

```
## In this case we are only interested in the `Time` column.
vdist <- density(vdata$Time)
vmean <- mean(vdata$Time)

plot(vdist, main="Density of Verizon's repair times.")
abline(v=vmean, col=rgb(0.0, 0.0, 1.0, 1.0))
```



(ii) – Given what the PUC wishes to test, how would you write the hypothesis?

I would probably state the hypothesis as: At Verizon 99% of all customer phone repairs are performed within 7.6 minutes.

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

Keep in mind that the data is not normal so the limits may or may not make sense at all.

```r
vmean <- mean(vdata$Time)
vmean
```

```
## [1] 8.522009
```

```r
z <- qnorm(c(0.005, 0.995)) # z values for 99% CI

vsd <- sd(vdata$Time)
vlen <- length(vdata$Time)

vmargin <- z * (vsd / sqrt(vlen))

vlimits <- c(vmean+vmargin)
names(vlimits) <- c("0.5%", "99.5%")
vlimits
```

```
##     0.5%    99.5%
## 7.594575 9.449444
```

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

```r
## Code from lecture 5.
time_hyp <- 7.6 # hypothesis states that time <= 7.6 m

vserr <- (sd(vdata$Time) / sqrt(length(vdata$Time)))
ot <- (mean(vdata$Time) - time_hyp) / vserr
df <- length(vdata$Time) - 1
op <- 1 - pt(ot, df)

c("t"=ot, "p"=op)
```

```
##           t           p
## 2.560762334 0.005265342
```

(v) – Briefly describe how these values relate to the Null distribution of t.

The null distribution is a standard normal distribution that is used with the t value to estimate the likelihood of a hypothesis being correct. A t value of 0 means that the null hypothesis is exactly correct. The p value is a measure of probability which states the likelihood of our desired t value to appear, as the distribution of t values is normal then a higher (and lower) t value will result in a lower p value. The null hypothesis in this case corresponds to 7.6 minutes.

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

Since we are looking for a 99 % CI we want to have a p value above 1 %. This would correspond to a p value >0.01, but as our current p is 0.0052653 (not above 0.01!) we can reject the company's claim.

(b)

(i) – Estimate the bootstrapped 99% CI of the population mean.

```
## Code from lecture 5.
sample_statistic <- function(stat_function, sample0) {
    resample <- sample(sample0, length(sample0), replace=TRUE)
    stat_function(resample)
}

vsample <- sample(vdata$Time, 500, TRUE) # magic number
n_boot <- 2000

sample_means <- replicate(n_boot, sample_statistic(mean, vsample))

quantile(sample_means, probs = c(0.005, 0.995))
```

```
##      0.5%     99.5%
##  6.825605 10.519957
```

(ii) – What is the 99% CI of the bootstrapped difference between the sample mean and the hypothesized mean?

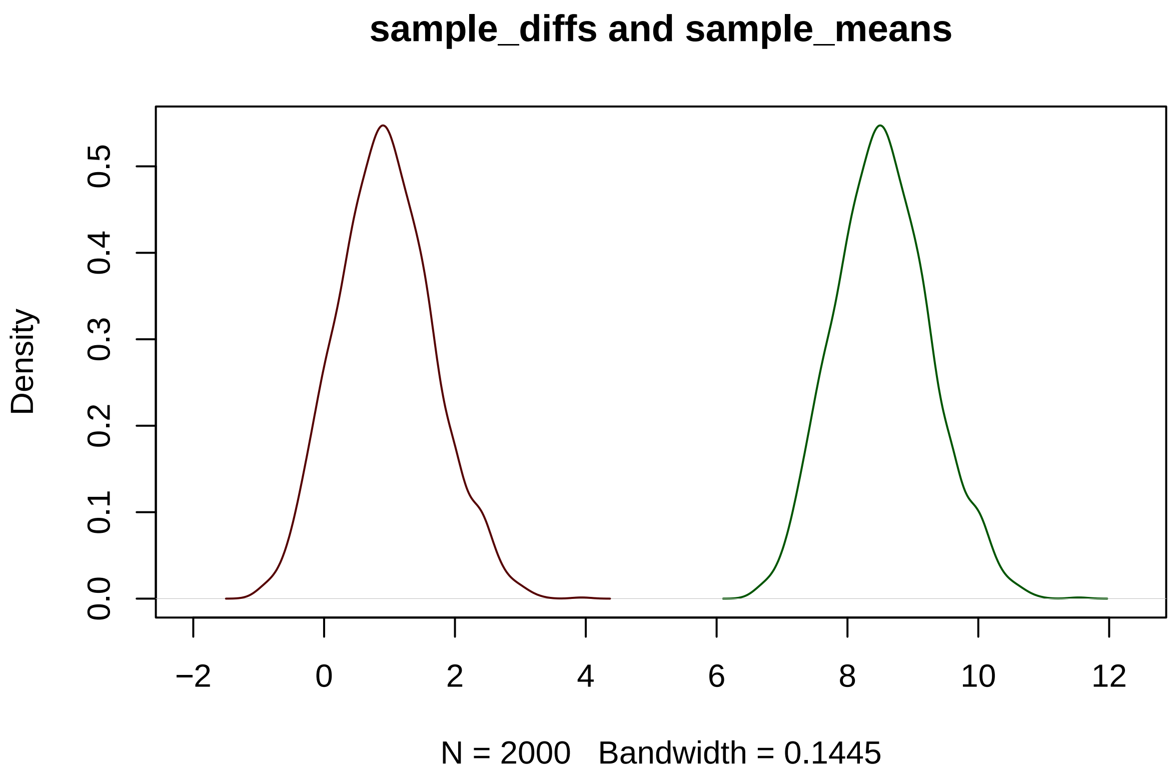
Instead of bootstrapping anew and defining `stat_function` as `function (sample0) { mean(sample0) - time_hyp }` we simply use the sample we have, `sample_means`, and subtract `time_hyp` from every element.

```
## recall that time_hyp is defined previously as 7.6
sample_diffs <- sample_means - time_hyp

quantile(sample_diffs, c(0.005, 0.995))
```

```
##       0.5%      99.5%
## -0.7743949  2.9199566
```

(iii) – Plot distribution the two bootstraps above.

```
plot(
    density(sample_means),
    col="#005500",
    xlim=c(-2, 12.3),
    main="sample_diffs and sample_means"
)

lines(density(sample_diffs), col="#550000")
```



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

Sort of. It agrees that the company's claim cannot be true, but it states that the company's claim is false to a much higher degree. As shown below the `p` value is so small it is displayed as 0, so it appears to be practically impossible for the hypothesised value to appear.

```
## Code from lecture 5.
vserr <- (sd(sample_means) / sqrt(length(sample_means)))
t <- (mean(sample_means) - time_hyp) / vserr
df <- length(sample_means) - 1
p <- 1 - pt(t, df)

c("t"=t, "p"=p)
```

```
##        t        p
## 57.37708  0.00000
```

(c) – New claim that median repair time is 3.5 min at 99% CI.

(i) – Estimate bootstrapped 99% CI of pop. median.

Let's use the same sample as before, `vsample`.

```
sample_medians <- replicate(n_boot, sample_statistic(median, vsample))
quantile(sample_medians, c(0.005, 0.995))
```

```
##  0.5% 99.5%
## 2.620 4.125
```

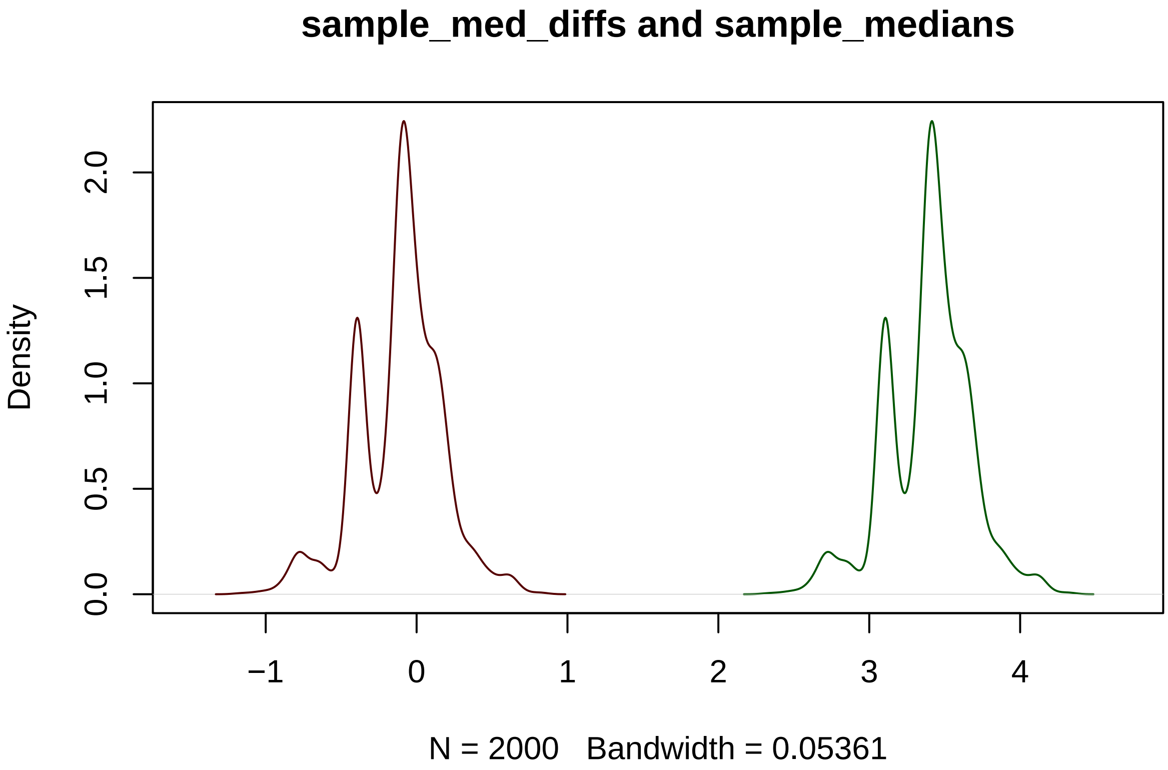
(ii) – Show 99% CI of bootstrapped difference between sample median and hyp. median.

```
median_hyp <- 3.5
sample_med_diffs <- sample_medians - median_hyp
quantile(sample_med_diffs, c(0.005, 0.995))
```

```
##   0.5%  99.5%
## -0.880  0.625
```

(iii) – Plot both bootstrapped distributions.

```
plot(
    density(sample_medians),
    col="#005500",
    xlim=c(-1.5, 4.7),
    main="sample_med_diffs and sample_medians"
)

lines(density(sample_med_diffs), col="#550000")
```



(iv) – Conclusion.

Yes, I think their statement is reasonable. Although, you would have to consider that the median of the total sample may not equal the bootstrapped median, for example:

```
c("vdata"=median(vdata$Time), "vsample"=median(vsample))
```

```
##   vdata vsample
##    3.63    3.42
```

We can see that the bootstrapped median is below Verizon's threshhold, but the full population median is above it.

Let's also try to find the t and p values and see what they say:

```r
vserr <- (sd(sample_medians) / sqrt(length(sample_medians)))
t <- (mean(sample_medians) - median_hyp) / vserr
df <- length(sample_medians) - 1
p <- 1 - pt(t, df)

c("t"=t, "p"=p)
```

```
##         t         p
## -17.28248   1.00000
```

2

For each scenario:

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?*

Initial

```
knitr::include_graphics("./50-0.3-2.9.png")
```

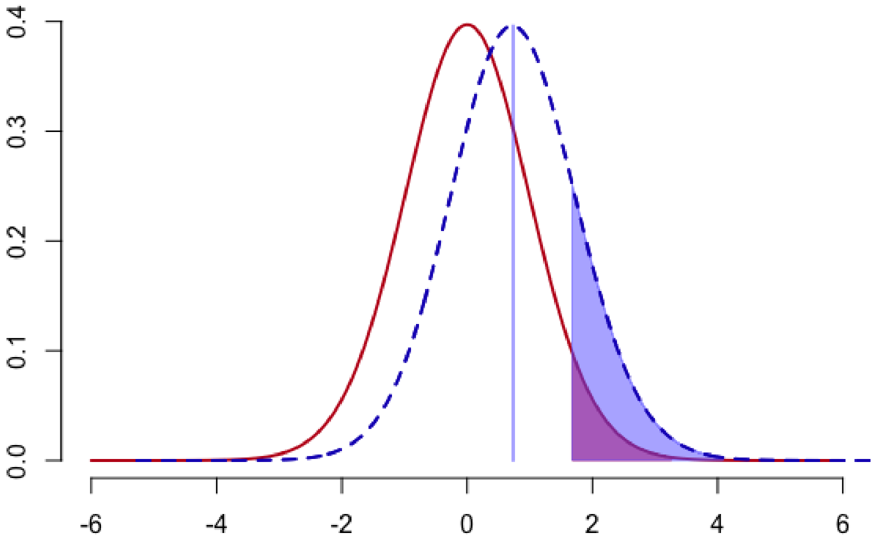


Figure 1: n=50, diff=0.3, sd=2.9

As we can see in the image, the null hypothesis is likely to be correct. So my colleague was right.

(a) – Faulty pool of customers.

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

ii. *Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?* That depends, if the data is extended to suit the general population better (i.e. all the old data remains, and new is added) then n and all values affected by n. Regardless if n changes, if we assume that the new data actually has a decreased usage then mean will change, causing the diff to decrease (as the mean is lower than previously), moving the t value closer to 0.

iii. *Will it increase or decrease our power to reject the null hypothesis?* Decrease as is reasoned in ii.

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

(b) – Faulty no. of samples.

i. *Would this scenario create systematic or random error (or both or neither)?* Random error? I think this is the definition of a systematic error. However, assuming that the error is fixed by simply removing those 20 from the sample and re-calculating all the values then I think random error from that point.

ii. *Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?* n, which in turn causes all other variables with the exception of alpha to potentially change. So n, t, sd, and diff may all be affected.

iii. *Will it increase or decrease our power to reject the null hypothesis?* If we assume that the distribution of data is the same as previously it would decrease our power to reject the null hypothesis. This can be shown using the function from `compstatslib`, where a smaller `n` causes the `t` value to come closer to 0, meaning that the null hypothesis is more likely.

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

(c) – Relaxed confidence level.

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?* Alpha.

iii. *Will it increase or decrease our power to reject the null hypothesis?* It will increase our power. The hypothesis will still be rejected if we consider an alpha value of 0.1 but it is now much closer to the cut off point for rejection.

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

(d) – Data is exclusively from weekdays.

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

ii. *Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?* If the data is changed, then `n`, which in turn may change all other parts.

iii. *Will it increase or decrease our power to reject the null hypothesis?* Depends on the data added, if we now assume that younger people are in fact very active on weekends then this may cause an increase in usage. And an increase in usage would make the mean higher, producing a larger t value and giving us more power to reject the null hypothesis.

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