Statistical Inference Week 4 Project 1

jedraynes

```
# load packages
library(ggplot2)
library(dplyr)
```

Part 2

EDA

We'll start with some exploratory data analysis to get a sense of the data we have.

```
# load and inspect data
df <- datasets::ToothGrowth
head(df)
```

```
##     len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

Upon first inspection, there are three columns (length, supplement type, and dose) with 60 rows.

```
# inspect data
print(str(df))
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
## NULL
```

```
print(colSums(is.na(df)))
```

```
##  len supp dose
##    0    0    0
```

It appears there are no nulls in the data and there are numeric (length and dose) and factor (supplement type) data types.

Summary and Analysis

```
# summarize the data
summary(df)
```

```
##       len          supp         dose
## Min.   : 4.20   OJ:30   Min.   :0.500
## 1st Qu.:13.07   VC:30   1st Qu.:0.500
## Median :19.25           Median :1.000
## Mean   :18.81           Mean   :1.167
## 3rd Qu.:25.27           3rd Qu.:2.000
## Max.   :33.90           Max.   :2.000
```

The length, our dependent variable has a minimum of 4.20 and a maximum of 33.90. There are two supplements types of vitamin C: orange juice (OJ) or ascorbic acid (VC). Finally, the dosage is between 0.5 and 2 mg/day. The dataset, per the documentation, displays the "effect of vitamin C on tooth growth in guinea pigs". The source can be found here (https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/ToothGrowth).

Let's take a look at the confidence intervals of tooth length using the data.

```
# len confidence interval
quantile(df$len, c(0.025, 0.975))
```

```
##   2.5%  97.5%
## 5.485 31.740
```

So, given a 95% confidence interval on the tooth length, the upper and lower bounds can be seen in the output above.

Next, let's complete a hypothesis test on the data. Let's compare the mean of tooth length of the OJ group with the mean tooth length of the population. We'll set H_0, or our null hypothesis, to mu_OJ >= 18.81. If you recall our summary above, 18.81 is the mean of the population. Then, we'll set H_a, or our alternative hypothesis, to be my_OJ < 18.81. Given the null hypothesis is assumed to be true, we're stating that the assumed truth is that the average tooth length for OJ-intaking guinea pigs is higher than that of the population. We'll keep things simple with a 5% level of significance (alpha).

```r
# split into two groups
df_oj <- df %>%
  filter(supp == 'OJ') %>%
  select(len, supp)

# calculate means of each group
oj_mean <- mean(df_oj$len)
pop_mean <- mean(df$len)

# calculate the difference in means
oj_pop_mean <- oj_mean - pop_mean
paste('The difference between the OJ mean and the population mean is ', round(oj_pop_mean, 4),
'.', sep = '')
```

```
## [1] "The difference between the OJ mean and the population mean is 1.85."
```

```r
# calculate standard error of the mean
oj_sd <- sd(df_oj$len)
oj_n <- nrow(df_oj)
oj_se <- oj_sd / (sqrt(oj_n))

# calculate z-score
oj_z <- oj_pop_mean / oj_se
paste('Our calculated z-score given our test is ', round(oj_z, 4), '.', sep = '')
```

```
## [1] "Our calculated z-score given our test is 1.534."
```

So, taking a look at our z-score, we can see it's below the cutoff of 1.645. Given our 5% level of significance, we fail to reject the null hypothesis.

Conclusion

In conclusion, we took a look at the ToothGrowth dataset which explores the tooth length of guinea pigs after receiving vitamin C via two different channels. This was further split by the dosage volume received. We tested whether the OJ grouping had a mean tooth length higher than the population. Given our assumptions and test as outlined above, we concluded that we failed to reject the null hypothesis.