Class Project in Two Parts: Central Limit Theorem & Inferential Data Analysis

Paul Clark

February 5, 2017

Contents

This document addresses a project assignment with two parts: Part 1, which is a simulation and investigation of the law of large numbers and the Central Limit Theorem, and Part 2, which is an exercise in Inferential analysis.

1 Central Limit Theorem (Simulation)

1.1 Overview

We investigate the *exponential* distribution in R and compare it with the Central Limit Theorem. The distribution is simulated via `rexp(n, lambda)`, and both mean and standard deviation are λ. We set `lambda = 0.2` and investigate the distribution of 1000 samples of 40 exponentials each.

Concretely, the exponential distribution with 'rate' λ has the density $f(x) = \lambda e^{-\lambda x}$, so it is very right-skewed. Therefore this will make for a fairly robust test.

1.2 Summary Conclusions

Simulation is consistent with theory.

1. The statistics of the distribution of sample means are approximately equal to their theoretical values, consistent with the properties of means and variances. We find:

   a. $\mu_{mean} \approx \mu_{population}$
   b. $\sigma^2_{mean} \approx \sigma^2_{population(N)}/N$

2. The distribution of means is right-skewed but approximately normal, consistent with the Central Limit Theorem

   a. From a simple plot, the underlying population distribution (exponential) is extremely right skewed.
   b. From a Q-Q plot, the distribution of means of 40 observations is *slightly* right-skewed.
   c. Nevertheless, the distribution of means is approximately normal over a relatively wide range of cumulative density.

1.3 Preliminaries

We first load packages needed for the analysis.

```r
rqd_pkgs <- c("ggplot2", "tidyr") # ggplot(), gather()
pkgs_to_install <- rqd_pkgs[!rqd_pkgs %in% installed.packages()[,1]]
if (length(pkgs_to_install)) install.packages(pkgs_to_install)
lapply(rqd_pkgs, require, character.only = TRUE)
```

1.4 Simulations

We will show the properties of the distribution of the mean and variance of samples of 40 simulated random exponentials.

```r
# We simulate S = 1000 samples of N = 40 exponentials each, all with lambda = 0.2
S <- 1000
N <- 40
lambda <- 0.2
set.seed(1234)
sim_data <- matrix(rexp(n = S*N, rate = lambda), S, N)
means <- apply(sim_data, 1, mean)
variances <-  apply(sim_data, 1, function(x) {sd(x)^2}) # sample var, N-1 in denominator
sim <- data.frame(population = rexp(1000, rate = lambda), means, variances)
str(sim)
```

```
## 'data.frame':    1000 obs. of  3 variables:
##  $ population: num  4.166 2.465 0.782 2.598 3.058 ...
##  $ means     : num  4.6 6.02 5.46 4.18 7.14 ...
##  $ variances : num  11.8 36.4 35.9 16.8 56.7 ...
```

1.4.1 Sample Mean vs. Theoretical Mean

In ***Figure 1***, we show our simulated distribution of sample means and compare it to the theoretical mean of the population.

```r
# We plot a histogram of sample means
g <- ggplot(sim, aes(x = means)) + geom_histogram()
g <- g + geom_vline(xintercept = mean(means), size = 1, color = "black") +
     geom_vline(xintercept = median(means), size = 1, color = "red") +
     labs(title = paste0("The simulated mean lies just below the ",
          "theoretical mean of ", 1/lambda,".", " \n The mean (black line) and median ",
          "(red line) are approximately equal.")) +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

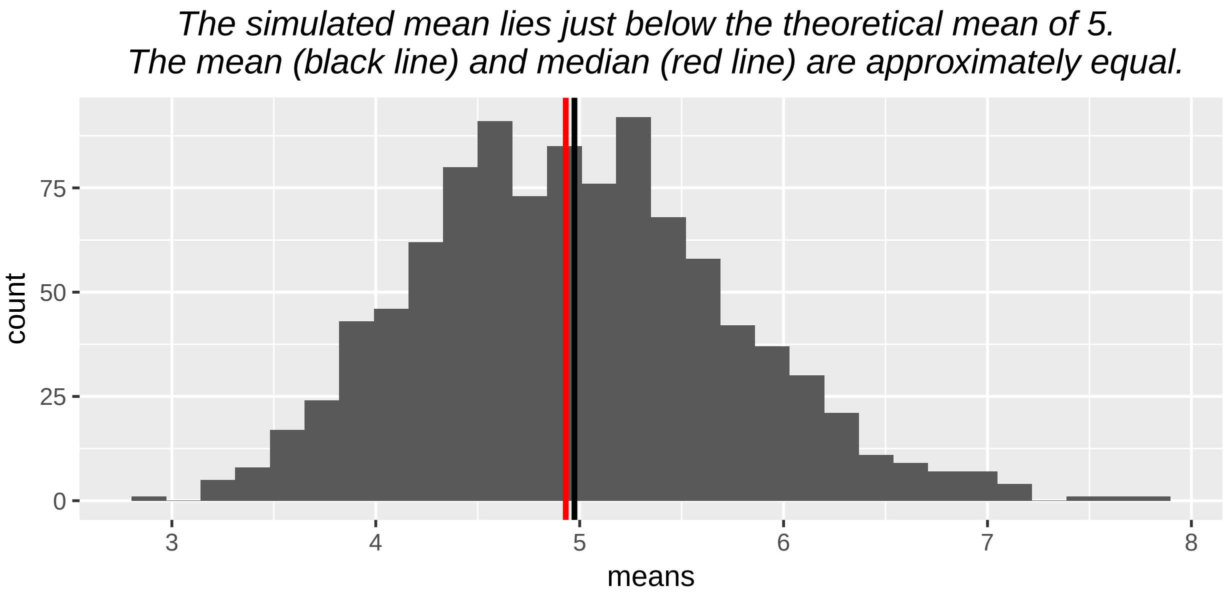

Figure 1: Histogram of means of 1000 samples of 40 exponentials with theoretical mean 5

The vertical line shows the position of the empirical mean. It almost exactly equals the theoretical value $1/\lambda$, or 5, for our lambda of 0.2. To be precise, the difference $\mu_{simulated} - \mu_{theory}$ is -0.03, or -0.52% from theoretical.

1.4.2 Sample Variance vs. Theoretical Variance

In ***Figure 2***, we show the sample variance and compare it to the theoretical variance of the population.

```
# Following slide 7 of lecture 5 of the statistical inference course, we plot a
# histogram of sample variances and show its mean
g <- ggplot(sim, aes(x = variances)) + geom_histogram()
g <- g + geom_vline(xintercept = mean(variances), size = 1) +
     labs(title=paste0("The mean of simulated variances is just below the ",
                       "theoretical value of ", 1/lambda^2)) +
     theme(plot.title = element_text(face="italic", hjust=0.5)) +
     scale_x_continuous(breaks = seq(from = 0, to = 90, by = 10))
print(g)
```

The vertical line shows the position of the empirical mean of the distribution of estimated population variances. The mean is again close to the theoretical value of $1/\lambda^2$, or 25. To be precise, the difference $\sigma^2_{simulated} - \sigma^2_{theory}$ is -0.65, or -2.59% from theoretical.

We can also estimate population variance using the variance of the simulated means. By the properties of variances, $\sigma^2_{population} = N\sigma^2_{mean(N)}$. This estimate gives 23.80, which differs by -1.20 from the theoretical, or -4.80%.

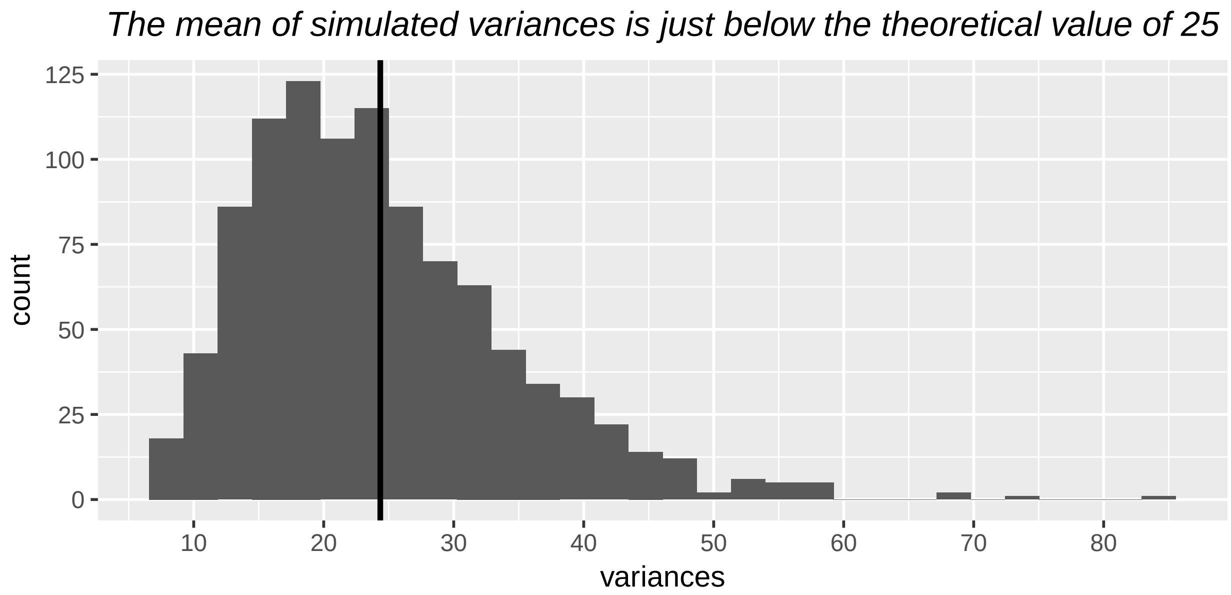

Figure 2: Histogram of estimated population variances of 1000 samples of 40 exponentials

1.5 Normality

We now show that the distribution of means is trending away from exponential toward normal. Our assessment will be primarily graphical.

```r
# First, we reformat the data for use in ggplot
long_sim <- gather(sim, key = Distribution, value = Measure, population:variances)
# We construct a two-facet plot contrasting the randomly sampled distribution
# of means with a randomly sampled distribution of the population
long_sim_sub <- with(long_sim, long_sim[!Distribution %in% "variances",])
g <- ggplot(data = long_sim_sub, aes(x = Measure)) + geom_histogram() +
     facet_grid(Distribution~.) + labs(title =
     "The means trend toward normal, while the population values are far from it") +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

By visual inspection of ***Figure 3***, the population distribution is severely right-skewed and not normal. The distribution of means is much closer to normal.

1.5.1 Equality of the Mean and Median

The Mean and Median of a normal distribution should be equal. This is approximately true for our distribution of means. In **Figure 1**, the black line represents the mean of means, whereas the red line represents the median.

1.5.2 Quantiles of sample distribution vs. quantiles of standard normal

Here we compare the quantiles (data values) of the distribution of means with those of standard normal to evaluate the normality.

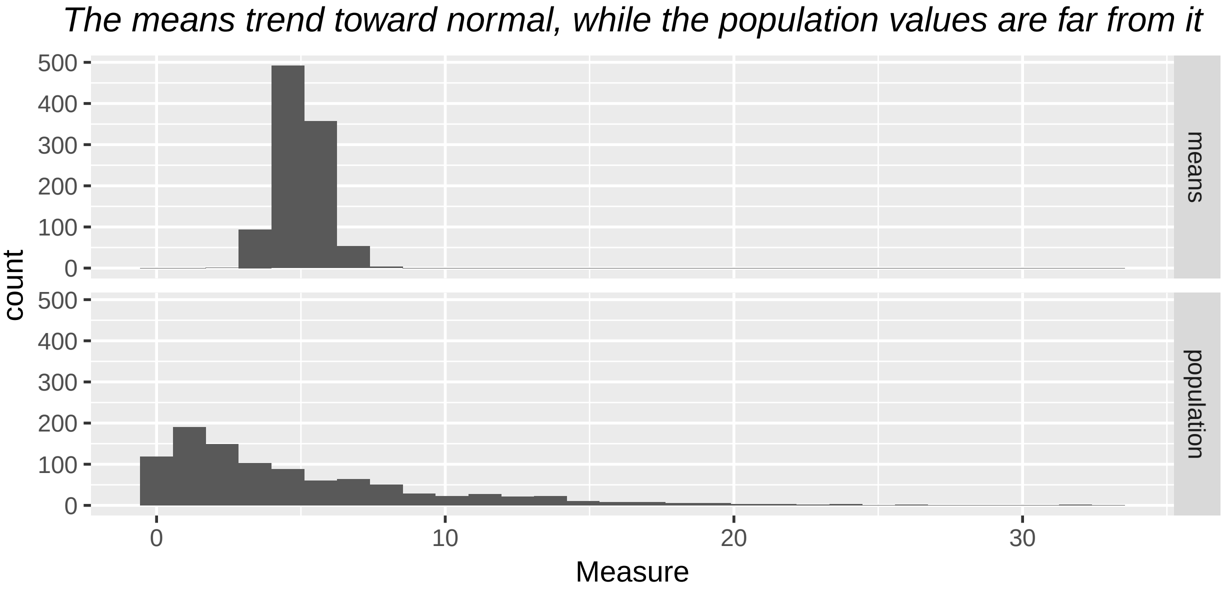

Figure 3: Histograms of 1000 Population Values and 1000 Means of random exponentials

1.5.2.1 Explanation of Q-Q plot

In **Figure 4**, our sample data is plotted on the vertical axis, and the associated normal quantiles (standard deviations from center) are plotted on the horizontal axis. The associated normal quantiles are calculated as follows: Empirical Cumulative Densities (ECD values) of the sample data are calculated, and those densities are used to calculate a normal quantile by inverting the normal cumulative density function (I.e, the calculated ECD probability and the normal cumulative density curve are used to obtain a normal quantile. This could be done in `r` via `qnorm(ECD)`.) Concretely, to obtain the cumulative densities, sample data points i are ascendingly ordered from $1\,to\,n$, then each point i is related to a cumulative probability associated with the mid-point between points i and $i-1$, calculated as $ECD = (i-0.5)/n$, where n is the number of sampled values. This way, for example, point $i = 500$, which in our data sample has sample quantile (data value) ≈ 5, is associated with cumulative probability `(500-0.5)/1000 = 0.4995` $= ECD = Normal\,Cumulative\,Density\,Function(q)$ (e.g., `pnorm(q)` in `r`). That function is inverted (e.g., via `qnorm(ECD)` in `r`) to obtain the q, or standard normal quantile value (in this example ≈ 0).

1.5.2.2 Q-Q plot for assessment of normality of the distribution of means

```r
qqnorm(sim$means, main = paste0("Our data is verging on normal, with some",
        "\nleft compression and right skew"), ylab = "Data Values in the Sample",
       xlab = "Normal Quantiles for Empirical Cumulative Density of Data (qnorm(ECD))")
qqline(sim$means)
```

The plot of sample data vs. corresponding quantiles of standard normal would be completely linear if our sample data was completely normal. However, we see from the Q-Q plot in **Figure 4** that as you move down from center (sample quantile $\approx 1/\lambda = 5$), values of the data sample below 4 are traversed more slowly than linear with corresponding standard normal quantiles, indicating left compression (more probability density spread across a smaller range of values). And as we move up from center, the data traverses values faster than linear with the normal quantiles, indicating probability mass more right skewed than normal. In summary, the distribution is not exactly normal, but retains relics of the underlying population distribution, which strongly exhibits this compression and skew (see bottom plot in *Figure 3*).

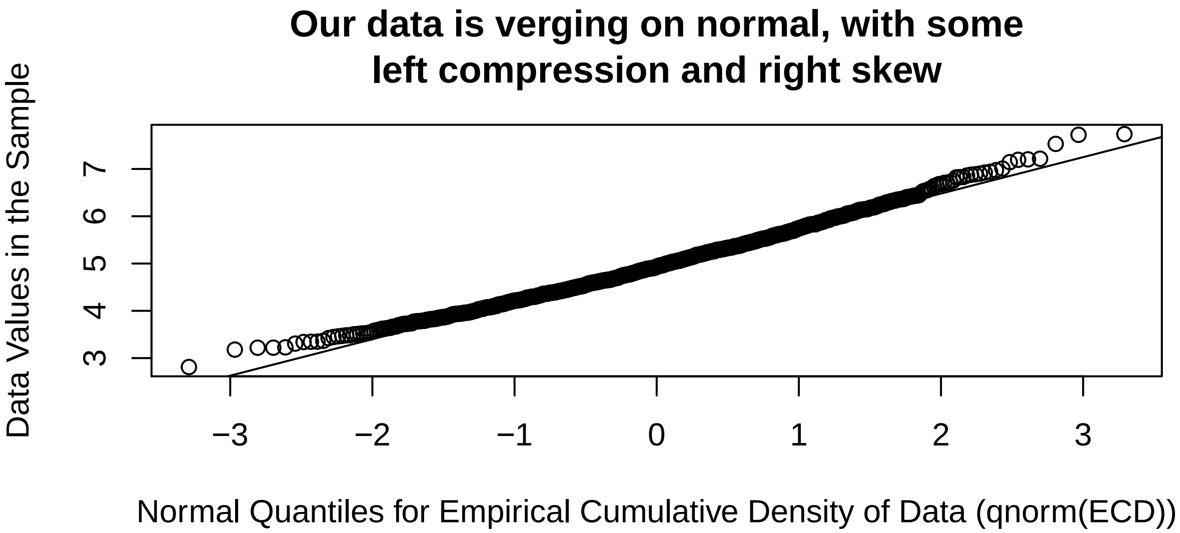

Figure 4: Plot of data against corresponding quantiles of standard normal

1.5.2.3 Number of standard deviations spanned by two normal quantiles of probability

The linear part of the Q-Q plot between -1 and +1 normal quantiles attests to the fact that the data is approximately normal. To demonstrate this, we compute the number of standard deviations of the mean between -1 and +1 standard normal quantiles from the center of the distribution. This should be 2 for a normal distribution.

```
data_quantiles_1_stdnorm <- quantile(sim$means, probs = c(pnorm(q=-1),pnorm(q=1)))
(med_plus_minus <- (data_quantiles_1_stdnorm - median(sim$means))/sqrt(var(sim$means)))
```

```
##  15.86553%  84.13447%
## -0.9296552  1.0526858
```

The number of standard deviations within one normal quantile of the median of our distribution (i.e., the number that spans the middle 68% of the distribution) differs from 2 by just -0.88%. However, note the right skew: the center of the distribution is shifted right of normal by approximately 0.06 standard deviations.

2 Basic Inferential Data Analysis

2.1 Overview

We now analyze the `ToothGrowth` data in the R `datasets` package via basic exploratory analyses, summaries, and statistical tests to compare tooth growth by `supp` and `dose`.

2.2 Summary Conclusions

1. Increased dose of vitamin C is associated with increased tooth length, whether it is delivered via orange juice (OJ) or ascorbic acid (VC).
2. At lower doses (0.5 mm and 1 mm), delivery via orange juice appears more effective than ascorbic acid. However, we see no such impact at the 2 mm dose.

2.3 Preliminaries

We first load a package and dataset needed in this analysis. Note that below we also make use of ggplot2, loaded for part I.

```
rqd_pkgs <- "datasets" # source for "ToothGrowth" data
pkgs_to_install <- rqd_pkgs[!rqd_pkgs %in% installed.packages()[,1]]
if (length(pkgs_to_install)) install.packages(pkgs_to_install)
lapply(rqd_pkgs, require, character.only = TRUE)
data("ToothGrowth") # load "ToothGrowth" data
```

2.4 Exploratory Data Analysis

We now perform basic summaries.

```
str(ToothGrowth)
```

```
## '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 ...
```

```
with(ToothGrowth, table(supp, dose))
```

```
##     dose
## supp 0.5  1  2
##   OJ  10 10 10
##   VC  10 10 10
```

The results of `table()` most clarify the structure of the data. There are 10 subjects for each unique combination of `supp` and `dose`. Guinea pigs were given one of 3 dosages of vitamin C daily (`dose`), in one of two forms (`supp`), and their tooth growth was measured.

Since we will just be performing T-tests between the various groups, it makes sense to re-cast `dose` as a factor.

```
ToothGrowth$dose <- factor(as.character(ToothGrowth$dose), levels = c("0.5", "1", "2"))
```

A box-plot reveals the data's major trends (**Figure 5**).

```
g <- ggplot(ToothGrowth, aes(x = supp, y = len, color = supp)) +
     geom_boxplot() + facet_grid(. ~ dose) +
     labs(title="A boxpot clearly depicts underlying trends from both variables") +
     theme(plot.title = element_text(face="italic", hjust=0.5))
print(g)
```

Teeth appear to grow longer on higher doses, and, except at `dose = 2`, they grow longer on OJ (orange juice) than on VC (ascorbic acid).

2.5 Tests of Significance of Inter-Group Differences

We will test the statistical significance of the above observations using one-sided T-tests. To verify the graphically observed trends, we will take the following approach:

1. Test two `dose` effects for each `supp` level:
    - Hypothesis that mean length is greater at dose 1 than dose 0.5
    - Hypothesis that mean length is greater at dose 2 than dose 1

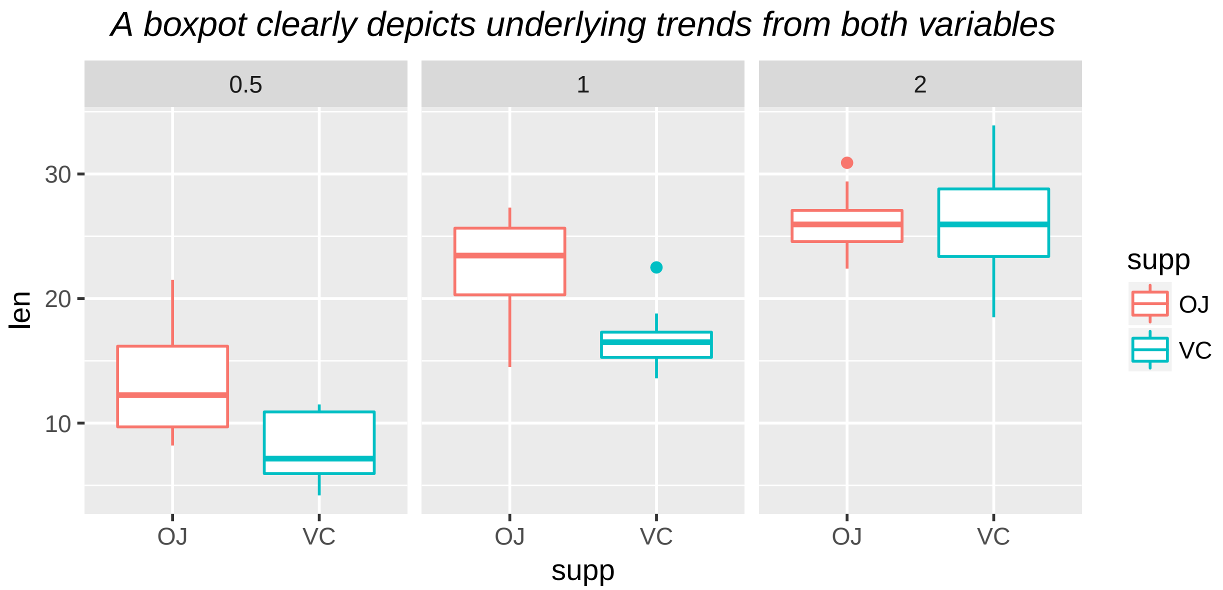

Figure 5: Boxplot of tooth length vs. vitamin delivery method at different dosings

2. Within each of the three `dose` levels, test the hypothesis that mean length of teeth is greater for guinea pigs given OJ than VC.

This implies we will be carrying out 7 tests. We will extract P-values for each test, and reject the null at level $\alpha = 0.025$.

First, we carry out the inter-dose tests…

```
# Function for t.tests
t.test_teeth <- function(dose2, dose1, supp2, supp1){
        t.test(with(ToothGrowth, len[dose == dose2 & supp == supp2]),
               with(ToothGrowth, len[dose == dose1 & supp == supp1]),
                          alt = "greater")$p.value
}
# Construct a data frame associated with the dosage effect
(dosage.effect <- data.frame(dose.change = c("0.5 to 1", "1 to 2", "0.5 to 1", "1 to 2"),
                            supp.choice = c("OJ", "OJ", "VC", "VC"),
                            pvalue = round(c(
                                    t.test_teeth("1","0.5", "OJ", "OJ"),
                                    t.test_teeth("2","1","OJ","OJ"),
                                    t.test_teeth("1","0.5", "VC", "VC"),
                                    t.test_teeth("2","1","VC","VC")
                                    ),4)
                            )
)
```

```
##   dose.change supp.choice pvalue
## 1    0.5 to 1          OJ 0.0000
## 2      1 to 2          OJ 0.0196
## 3    0.5 to 1          VC 0.0000
## 4      1 to 2          VC 0.0000
```

The upward trend in tooth length with increasing dosage appears significant, whether delivering vitamin C via orange juice or ascorbic acid.

Now, we test supplement changes (choice of OJ or VC) at fixed dosage.

```
# Construct and print a data frame associated with the VC to OJ effect
(VC_to_OJ.effect <- data.frame(dose.choice = c("0.5", "1", "2"),
                               pvalue = round(c(
                                     t.test_teeth("0.5","0.5", "OJ", "VC"),
                                     t.test_teeth("1","1","OJ","VC"),
                                     t.test_teeth("2","2","OJ","VC")
                                     ),4)
                               )
)
```

```
##   dose.choice pvalue
## 1         0.5 0.0032
## 2           1 0.0005
## 3           2 0.5181
```

Substitution of orange juice (OJ) for ascorbic acid (VC) appears significantly associated with increase in tooth length for dosing of 0.5 and 1, but not for 2.

Note that the hypothesis that mean OJ group length is greater than VC group length for dose of 2 is poorly formulated since the mean observed length for OJ subjects (26.06) is not greater than that for VC subjects (26.14). Therefore, we also perform a two-sided test for this case, testing the hypothesis that there is any difference in means when dose = 2.

```
attach(ToothGrowth)
t.test(len[dose == "2" & supp == "OJ"], len[dose == "2" & supp == "VC"],
       alt = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  len[dose == "2" & supp == "OJ"] and len[dose == "2" & supp == "VC"]
## t = -0.046136, df = 14.04, p-value = 0.9639
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.79807  3.63807
## sample estimates:
## mean of x mean of y
##     26.06     26.14
```

From the p-value and confidence interval above, we see that at dose of 2, there is no evidence to support a claim that mean tooth length differs for guinea pigs given OJ and guinea pigs given VC. For the highest dose, the impact of vitamin C delivery method drops to 0.

2.6 Key Assumptions

- For applicability of the t distribution, the underlying data should theoretically be normal and iid
- The guinea pigs are randomized across the various combinations of `supp` and `dose`
- The researchers are blind to the experimental conditions of each guinea pig